BIOS 6244 Analysis of Categorical Data
December 5, 2005 Lecture

Sample Size and Power for Logistic Regression (Sec. 5.6)

Actually, this section could be called "Sample Size and Power for Proportions and Odds Ratios."

Studies in which the sample size N is not sufficiently large have no hope of detecting an association between the explanatory variable(s) and the binary outcome Y, *regardless of how strong this association might be.* Even though statisticians are not always given the opportunity to participate in the planning of a study, we should still be proactive in emphasizing the importance of sample size determination *prior to beginning data collection.* At my first meeting with a potential client, I make a point of reinforcing the need for them to consult with me during the planning phase of any study in which they expect me to participate.

Sample Size for a Single Proportion

For confidence interval estimation, this was covered in Exercise 1.9, p. 14 (Assignment 1).

For hypothesis testing, to achieve power of 1-β for detecting a value of π_1 when the hypothesized value under H_0 is π_0 using a significance level of α, the normal approximation to the binomial can be used to derive the following formula for the required sample size:

$$N = \frac{(z_{\alpha/2} + z_{\beta})^2 \pi_1 (1-\pi_1)}{(\pi_1 - \pi_0)^2},$$

where z_γ = upper γ-percentage point of the standard normal distribution.

(Recall that β is the probability of a Type II error and that the power 1-β is therefore the probability of concluding that $\pi \neq \pi_0$ when in fact $\pi = \pi_1$.)

Sample Size for Comparing Two Proportions (Sec. 5.6.1)

If the explanatory variable X is also binary, then the problem of examining the association between X & Y is equivalent to comparing the true proportions of "success" (i.e., Y = 1) between the 2 groups defined by the values of X (π_1 and π_2, respectively). That is, we wish to test H_0: $\pi_1 = \pi_2$ vs. H_a: $\pi_1 \neq \pi_2$.

We know that using Pearson's χ^2 method to test for the independence of X & Y is equivalent to the "usual" method for comparing 2 proportions that is based on the normal

approximation to the binomial. (See pp. 10-12 of the lecture notes – October 3, 2005 lecture).

The formula given for ASE($\hat{\pi}_1 - \hat{\pi}_2$) on p. 11 of these notes can be used to derive the *approximate* sample size required to detect a true difference π_1 - π_2 with power 1-β using a two-tailed test with significance level α. The formula for the sample size required in each group (N_1 and N_2) is as follows:

$$N_1 = N_2 = \frac{(z_{\alpha/2} + z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2}, \qquad (27)$$

where z_γ = upper γ-percentage point of the standard normal distribution.

Note that we must be able to specify *a priori* one of the 2 population proportions being compared (either π_1 or π_2) *and* the true difference we are trying to detect (π_1 - π_2).

In clinical research, we always use α = .05 and, in most applications, we use β = .20 (or, power = 80%). That is, we want to choose sample sizes N_1 and N_2 so that there is an 80% chance of detecting a true difference of π_1 - π_2 using a significance level of α = .05. (Some authors, especially in the context of clinical trials, have suggested that the sample sizes should be chosen so that the power is at least 90%. If sufficient study subjects are available, then the sample size calculation can certainly be based on 90% power. However, in my experience, sample sizes that achieve 80% power are often difficult to obtain. If an adequate # of study subjects (enough to achieve at least 80% power) cannot be found, then *the study should not be done.*

Example (antibiotic cure rates)

Suppose it is known that the cure rate for a certain infectious disease using the standard antibiotic is 70%. A new antibiotic has been developed and is expected to improve this cure rate by at least 10% (i.e., it will have a cure rate of at least 80%). A randomized trial is planned, with half of the study subjects being assigned to the standard antibiotic and the other half assigned to the new one. What sample sizes in the 2 groups will be required to achieve 80% power for detecting this level of improvement in cure rate, using α = .05? (This is equivalent to asking: What sample sizes will be required to yield an 80% chance of finding $p < .05$, given that $\pi_1 = .70$ and $\pi_1 - \pi_2 = .10$?

Substituting into Equation (27), we obtain

$$N_1 = N_2 = \frac{(1.96 + .84)^2[.7(1-.7) + .8(1-.8)]}{(.7-.8)^2} = 290.08 \rightarrow 291.$$

Notes

(1) We always round up to the next highest integer, no matter how small the decimal fraction of the calculated sample size is.

(2) The power of any test for comparing 2 population proportions is maximized when the sample sizes in the 2 groups are equal. If the group sizes are required to be different (e.g., in a case-control study where it is known that there are 4 times as many controls as cases to select from), this can be incorporated into the sample size determination, but it will require a larger total sample size.

(3) As pointed out in Assignment 1 (Exercise 1.9 in Agresti), many sample size calculations are based on approximate methods for finding CI's or performing hypothesis tests. Since the calculated sample size is usually only a "best guess" anyway, the sample sizes calculated using approximate methods should be adequate even if you ultimately apply the exact method when analyzing the study data. StatXact can perform sample size calculations for both approximate and exact methods, and sample size calculations for exact tests will be incorporated into PROC POWER in future releases of SAS. By way of comparison, the required sample size if Fisher's exact test is used in planning the antibiotic study is $N_1 = N_2 = 311$ (vs. 291 for the "usual" method, which yields 77% power for using Fisher's exact test).

Sample Size for a Confidence Interval for the Difference of Two Proportions

One can also derive a formula for finding the sample size required to produce a 95% CI(π_1 - π_2) of a desired width W. (Note that this means that we can be 95% sure that the point estimate of (π_1 - π_2) will differ from the true (π_1 - π_2) by no more than $\frac{W}{2}$.) This sample size formula is as follows:

$$N_1 = N_2 = \frac{4z^2_{\alpha/2}[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}{W^2}. \qquad (28)$$

So, suppose that in the antibiotic example above,we want to be 95% sure that our estimate ($\hat{\pi}_1 - \hat{\pi}_2$) is within .10 of the true difference (π_1 - π_2) (i.e., W = .20). Then

$$N_1 = N_2 = \frac{4(1.96)^2[.7(1-.7)+.8(1-.8)]}{(.2)^2} = 142.1 \rightarrow 143.$$

Some sample size calculations have been incorporated into PROC POWER in SAS beginning with Version 9.1 and more are added with each new release. The SAS code required for performing the above sample size calculation for the test of H_0: $\pi_1 = \pi_2$ is as follows:

(SAS Code given on following page.)

```
proc power;
     twosamplefreq test=pchi
     alpha = .05
     sides = 2
     grouppproportions = (.7 .8)
     nullproportiondiff = 0
     npergroup = .
     power = .80;
     run;
```

The SAS code above produces the following output (thanks to Joe Hagan):

The SAS System

The POWER Procedure
Pearson Chi-square Test for Two Proportions

Fixed Scenario Elements

Distribution	Asymptotic normal
Method	Normal approximation
Number of Sides	2
Null Proportion Difference	0
Alpha	0.05
Group 1 Proportion	0.7
Group 2 Proportion	0.8
Nominal Power	0.8

Computed N Per Group

Actual Power	N Per Group
0.801	294

Note that the results produced by SAS agrees to within 3 with our hand calculation.

Sample Size for a Simple Logistic Regression Model

As we have seen previously, the null hypothesis H_0: $\pi_1 = \pi_2$ is equivalent to the hypothesis H_0: $\beta^* = 0$ in the simple logit model

$$\text{logit } \pi = \alpha^* + \beta^* X, \qquad (29)$$

where X = 1 for Group 1 and X = 0 for Group 2 in the comparison of π_1 & π_2.

[Agresti uses α^* & β^* for the parameters in the logit model so as not to confuse them with the significance level (α) and the probability of Type II error (β).]

Recall that we also know that the null hypothesis H_0: $\beta^* = 0$ in the model in Equation (29) is equivalent to H_0: OR = 1. Thus, determining the required sample size for comparing 2 population proportions is equivalent to determining the sample size required for testing an hypothesis about an odds ratio.

Example

Suppose that in an Epi study, we wish to determine the sample size required to detect a true odds ratio of 3.0 with 80% power when testing the null hypothesis H_0: OR = 1 using $\alpha = .05$. In order to do this, we must be able to specify a reasonable value for the baseline probability of disease in the unexposed group, π_1. Using the value OR = 3 specified in the alternative hypothesis, we can then "work backwards" to determine the corresponding value of π_2. For example, suppose that it is known that 5% of the unexposed subjects eventually get the disease. Then OR = 3 would require that $\pi_2 = .136$. Once we have the values of π_1 & π_2, we can substitute into the formula in Equation (27) to obtain the desired sample size.

To illustrate how one might describe the sample size calculation in a "real-world" Epi study, consider the following excerpt from the Tox & Trauma article provided on the course website (p. 563):

> "Determination of sample size for this study focused on achieving sufficient statistical power to detect important associations between positive screens for alcohol or drugs and injury characteristics or outcomes. Previous clinical experience with trauma patients such as those in the current study suggested that about 25% would have positive screen results for alcohol. Assuming that an OR of 2 or greater is clinically important, and that the injury characteristic or outcome of interest has a prevalence of 5% or greater in the group that has negative screen results for alcohol, a sample of 351 patients with positive screen results for alcohol and 1,054 with negative screen results would be sufficient to achieve 80% power for detecting an OR of this magnitude."

Sample Size with a Quantitative Predictor (Sec. 5.6.2)

Suppose that we wish to fit a simple LR model of the form

$$\text{logit } \pi = \alpha^* + \beta^* X,$$

where X is a quantitative variable (continuous or ordinal). We wish to determine the sample size needed to achieve power of 1-β for testing H_0: $\beta^* = 0$ (or, equivalently, H_0: OR = 1) using significance level α. We know that e^{β^*} is the odds ratio corresponding to a 1-unit increase in X. Thus, we need to know the sample size required to detect a true

odds ratio of $OR^* = e^{\beta^*}$. Hsieh (*Statistics in Medicine* 1989, Vol. 8, pp. 795-802) proposed that one consider the odds ratio OR^* for comparing the probability of "success" at $\bar{x}$ (denoted $\bar{\pi}$) with the probability of "success" at $\bar{x} + 1\sigma$, where σ denotes the true standard deviation of X. Letting $\lambda = \log(OR^*)$, Hsieh obtained the following sample size formula for a 1-sided test of H_0: $\beta^* = 0$:

$$N = \frac{[z_\alpha + z_\beta \exp(-\lambda^2/4)]^2 (1 + 2\bar{\pi}\delta)}{\bar{\pi}\lambda^2}, \qquad (30)$$

where $$\delta = \frac{1 + (1+\lambda^2)e^{(5\lambda^2/4)}}{1 + e^{(-\lambda^2/4)}}.$$

Example (cholesterol & severe heart disease)

In this example, Y = presence of severe heart disease (Yes/No) and X = serum cholesterol, measured as a continuous variable. The study subjects will be from the age group 35-55.

Suppose we wish to test H_0: OR = 1 vs. H_0: OR > 1. Previous studies have suggested that the probability of severe heart disease at the average level of cholesterol for subjects aged 35-55 is about .08. Suppose we wish to detect a true OR of 2.0 corresponding to a 1 standard deviation increase in cholesterol above the mean with 80% power ($\alpha = .05$). Then, $\lambda = \log(2.0) = .693$, and substituting into Equation (30) above, we obtain

$$\delta = \frac{1 + (1 + 0.693^2)e^{[5(.693)^2/4]}}{1 + e^{[-(.693)^2/4]}} = 1.960 \text{ and}$$

$$N = \frac{[1.645 + .84\exp(-(.693)^2/4)]^2 [1 + 2(.08)(1.960)]}{(.08)(.693)^2} = 195.3 \rightarrow 196.$$

Note that Hsieh's formula is based on the assumption that X is asymptotically normally distributed.

Sample Size in Multiple Logistic Regression (Sec. 5.6.3)

A multiple LR model requires larger sample sizes to detect adjusted OR's of a certain magnitude. The sample size required depends on the inter-correlation among the explanatory variables. Let R^2 denote the coefficient of determination obtained when X is regressed on the other explanatory variables in the model. A rough approximation to the total sample size required for fitting a multiple LR model can be found by dividing the sample size N obtained from the formula in Equation (30) by $1\text{-}R^2$. Note that $\bar{\pi}$ in Equation (30) is replaced by the probability of "success" calculated at the mean value of *all* of the explanatory variables and that the odds ratio we are now testing is the OR for the effect of increasing the predictor of interest by 1 unit while holding each of the other explanatory variables fixed at their mean levels.

Cholesterol example, cont.

Suppose we are now interested in examining cholesterol as a risk factor for severe heart disease, while controlling for the effects of systolic and diastolic blood pressure (bp). Suppose also that when we regress cholesterol level on systolic & diastolic bp, we obtain an R^2 of .16. Then the sample size of 196 obtained from Equation (29) would be inflated by a factor of $\frac{1}{1-.16} = 1.190$ in order to account for the presence of the bp variables in the LR model. Thus, (1.19)(196) = 234 subjects would now be required.

Notes

(1) In most Epi studies, the risk factor is dichotomous. In the example above, the principal investigator (PI) probably would have dichotomized serum cholesterol into "normal" and "high" using the "normal range." While we do lose information by dichotomizing in this way, we gain in interpretability of the model coefficients. For this reason, most logistic regression analyses published in clinical research journals have a dichotomized variable for the primary risk factor. The models may include continuous (e.g., age) and/or categorical (e.g., gender) confounding variables as well. Categorical confounders are usually also dichotomized (e.g., ethnic origin would probably be dichotomized as "white/non-white").

(2) In order to obtain a reasonable value of R^2 for use in the sample size calculation for a multiple LR model, we would have to rely on either pilot data or previously published research. I always encourage the PI to conduct a small pilot study (5-10 subjects in both the "exposed" and "non-exposed" groups) prior to beginning the "real" study if at all possible. This will give the statistician some data on which to base their estimate of R^2 (and perhaps other quantities needed for the sample size calculation) and will also give the PI an opportunity to "debug" their study protocol. In lieu of pilot data, one would have to rely on a previously published study to find an approximate value for R^2 and it is unlikely that such a study could be found. (Most published studies would not provide this level of detail from their analyses.) A possible alternative would be to perform a "sensitivity" analysis by trying several different values of R^2 to see what the net effect would be on the calculation of N. Ultimately, of course, the choice of N depends on the resources available to the PI for conducting the study. If adequate resources are not available to support a sufficiently large sample size, then the study *should not be conducted.*

Exact Inference for Logistic Regression (Sec. 5.7)

If the maximum likelihood estimation procedure for 1 or more of the parameters in the multiple LR model does not converge, then alternative estimation methods must be considered. One that seems to work well in most circumstances is *conditional maximum likelihood.* Unfortunately, we will not have time to discuss this method in our course.

However, it is available in LogXact and in SAS by using the EXACT statement within PROC LOGISTIC. (See Computer Lab 3, p. 36 – November 16.)