Sample size for a single proportion

"How big should my sample size be?" Statisticians are often asked this question. When studies involve data in the form of counts or proportions, the best answer is probably, "As big as you can afford." The reason for this is that there is surprisingly little information in such data, even from quite big studies.

For example, in a 1991 poll of $1,000$ adult Americans, 45% said that they would be prepared to ask a child under 18 to give up a kidney for a transplant into a relative. The corresponding 95% confidence interval for the true percentage in the population who would do this is given by $[42\%, 48\%]$, an interval of width 6%. This is fairly imprecise, and yet it is based upon a sample of $1,000$ which is an unbelievably large sample size for researchers in most areas outside of survey sampling. And even with this study of $1,000$ people, if you wanted to look at subgroups, e.g. by age and sex, the sample sizes of the individual groups of interest soon become small, and the resulting confidence intervals are much wider than the 6% above.

1994 World Heath Organization figures rated Hepatitis B as the fourth biggest killer among the world's infectious diseases.[1] Suppose that we wanted to take a sample of the citizens of a particular city to determine the percentage of people who have, or are carrying, Hepatitis B. Suppose that the level of precision we require is such that a 95% confidence interval is no wider than 5% (i.e. 0.05).

Assuming n is large enough, a confidence interval for the true proportion p is given by $\widehat{p} \pm z \, \mathrm{se}(\widehat{p})$, where $\mathrm{se}(\widehat{p}) = \sqrt{\widehat{p}\,(1-\widehat{p})/n}$. The interval has

$$\mathit{Width} = 2 \times z \, \sqrt{\widehat{p}\,(1-\widehat{p})/n}.$$

which is twice the *margin of error*. The above expression involves $\widehat{p}$ which is unknown until the study is finished. We will postpone considering what to do about $\widehat{p}$ for the time being. Now, suppose that we want the *margin of error* to be no more than m. We therefore want

$$z \, \sqrt{\widehat{p}\,(1-\widehat{p})/n} \; \leq \; m.$$

or, on squaring both sides,

$$z^2 \times \frac{\widehat{p}\,(1-\widehat{p})}{n} \leq m^2 \, .$$

[1] 1-2 million deaths per year. Follows Acute Respiratory infections (4.3 million), Diarrheal diseases (3.2 million), and Tuberculosis (3 million). AIDS (550,000) was $10th$.

Solving this equation for n gives

$$n \geq \left(\frac{z}{m}\right)^2 \times \widehat{p}(1-\widehat{p})$$

The boxed equation still depends upon the eventual $\widehat{p}$, which is unknown when one is planning the survey. However, we see that the bigger $\widehat{p}(1-\widehat{p})$ is, the larger n has to be. Now, Fig. 1 graphs $\widehat{p}(1-\widehat{p})$ versus $\widehat{p}$, and we see that $\widehat{p}(1-\widehat{p})$ takes its biggest value when $\widehat{p} = \frac{1}{2}$. Thus, the "worst case" occurs when we use $\widehat{p} = \frac{1}{2}$ in the boxed equation as it leads to the biggest value of n. Use of this value of n guarantees that the interval will be no wider than w no matter what $\widehat{p}$ turns out to be.

$\widehat{p}$	0	.1	.2	.3	.4	.5
$\widehat{p}$	1	.9	.8	.7	.6	.5
$\widehat{p}(1-\widehat{p})$	0	.09	.16	.21	.24	.25

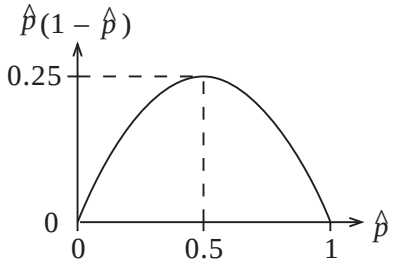

Figure 1 : $\widehat{p}(1-\widehat{p})$ versus $\widehat{p}$.

In the motivating example, we want to use a 95% confidence interval and so use $z = 1.96$. We require an interval no wider than $w = 0.05$ (since $5\% = 5/100 = 0.05$), or $m = 0.025$. Thus, we will have sufficient precision if we take

$$n \geq \left(\frac{1.96}{0.025}\right)^2 \times \frac{1}{2} \times \frac{1}{2} = 1536.64.$$

In the interests of using nice round numbers, we would probably end up taking a sample of around 1500, or 2000 depending upon the budget.

From Fig. 1, we see that $\widehat{p}(1-\widehat{p})$ can be quite a bit smaller than $\frac{1}{2} \times \frac{1}{2}$ if $\widehat{p}$ is close to "zero" or "one". Suppose that we knew that the prevalence of Hepatitis B in our city was no more than 10%. The sample proportion should end up roughly in the same sort of range. If we substitute $\widehat{p} = 0.1$ into the equation instead of $\widehat{p} = \frac{1}{2}$ we get $n \geq 553.19$. Thus, if we knew $p \leq 0.1$, we would know we could be confident of getting the required level of precision from a much smaller (and therefore cheaper) sample.

Rule for dealing with $\widehat{\boldsymbol{p}}$ in the sample size formula: ***Use*** $\widehat{\boldsymbol{p}} = \frac{\mathbf{1}}{\mathbf{2}}$, ***unless*** it is known that p belongs to an interval $a \leq p \leq b$ that does not include $\frac{1}{2}$, in which case substitute the interval endpoint nearer to $\frac{1}{2}$ for $\widehat{p}$.

To illustrate, if nothing is known about the true value of p, or it is known that $0.2 \leq p \leq 0.6$, or $0.4 \leq p \leq 0.9$ substitute $\widehat{p} = \frac{1}{2}$. If it is known that $0.02 \leq p \leq 0.10$ substitute $\widehat{p} = 0.1$, and if $0.7 \leq p \leq 0.9$, substitute $\widehat{p} = 0.7$.

Example Suppose we require a 95% confidence interval for p to be no wider than 0.02 (i.e. 2%). Using the formula above with $z = 1.96$, if nothing is known about the size of p, we should take

$$n \geq \left(\frac{1.96}{0.01}\right)^2 \times \frac{1}{2} \times \frac{1}{2} = 9604.$$

If we knew that $0.01 \leq p \leq 0.05$, we would substitute $\widehat{p} = 0.05$ to obtain

$$n \geq \left(\frac{1.96}{0.01}\right)^2 \times 0.05 \times 0.95 = 1824.76.$$ ∎

These calculations show that we need huge samples to estimate proportions very precisely – nearly $10,000$ observations in the first calculation. These calculations apply only to getting the required precision for estimates of single proportions. Even larger numbers are required to get this sort of precision for differences. However, survey organizations seldom sample more than one or two thousand people. Part of the reason for this is cost. Another reason is that, in practice, it is very hard to control the level of non-sampling errors in very large surveys (see Section 1.1.2). There is little point in reducing the sampling errors below the level of the non-sampling errors (or biases).

Exercises

Estimate the size of sample required to ensure that:

1. A 90% confidence interval for an unknown p is no wider than 0.04.

2. A 99% confidence interval for p expressed as a percentage is no wider than 4% if it is believed that $p < 0.2$.

3. A 95% confidence interval for an unknown p is no wider than 0.01.

4. A 90% confidence interval for p expressed a a percentage is no wider than 5% if it is believed that $p > 0.9$.