Math 203, Fall 2008:

Statistical Inference: How reliable is a survey?

Consider a survey with a single question, to which respondents are asked to give an answer of yes or no.

Suppose you pick a random sample of n people, and you find that the proportion that answered yes is $\hat{p}$.

Question: How close is $\hat{p}$ to the actual proportion p of people in the whole population who would have answered yes?

In order for there to be a reliable answer to this question, the sample size, n, must be big enough so that the sample distribution is close to a bell shaped curve (i.e., close to a normal distribution). But even if n is big enough that the distribution is close to a normal distribution, usually you need to make n even bigger in order to make sure your margin of error is reasonably small.

Thus the first thing to do is to be sure n is big enough for the sample distribution to be close to normal. The industry standard for being close enough is for n to be big enough so that

$$n > 9\frac{1-p}{p} \quad \text{and} \quad n > 9\frac{1-p}{p}$$

both hold. When p is about 50%, n can be as small as 10, but when p gets close to 0 or close to 1, the sample size n needs to get bigger. If p is 1% or 99%, then n must be at least 892, for example. (Note also that n here depends on p but not on the size of the whole population.) See Figures 1 and 2 showing frequency histograms for the number of yes respondents if $p = 1\%$ when the sample size n is 10 versus 1000 (this data was obtained by running a computer simulation taking 10000 samples). When $n = 10$ the distribution does not look very close to being a normal distribution, but you get a nice bell shaped curve when $n = 1000$.

Consider for example a presidential poll. If you're trying to measure support for one of the major candidates, p is likely reasonably close to 50% (as is true for both McCain and Obama right now, a week before the election), so even fairly small samples are more or less normally distributed. But if you want to be able to accurately analyze the support for a minor candidate, say Nader, you'll need a much bigger sample. Of course, you don't know p or $\hat{p}$ before you do your survey, but usually you have some idea of about what the value of p is, so you can use that to get an idea of how big n will have to be. If it turns out that your expectation as to what the value of p was going to be was way off, you might have to run your survey over with a bigger sample size.

Once you know how big n must be for the distribution to be normal, you usually have to make it even bigger in order to achieve whatever margin of error you want. The industry standard is to use a 95% confidence level; i.e., for the margin of error to be big enough so that 95% of randomly chosen samples will have $\hat{p}$ fall within the margin of error. This means that the margin of error must be $\pm 2\sigma$. (A 68% confidence level would use a $\pm\sigma$ margin of error and a 99.7% confidence level would use a $\pm 3\sigma$ margin of error. This makes sense; the bigger you set your range to be, the more confident you can be that $\hat{p}$ will end up in that range, since the bigger a target is the easier it is to hit it.) Since the industry standard is $\pm 2\sigma$, that's what you're stuck with, but what you can control is how big σ is. To make σ smaller, and thus to have a smaller margin of error, you need to make n bigger. This is because in the formula for σ,

you divide by n:

$$\sigma = \sqrt{\frac{p(1-p)}{n}}$$

Note that the size of the whole population does not come into it. In particular, a bigger population doesn't mean you have to make n bigger.

In short, regardless of how big the whole population is, for 19 samples out of 20 (i.e., 95% of the time) the sample proportion $\hat{p}$ will be within 2σ of the population proportion, p. So in fact it is not true that polls show "only the answers of the tiny percentage represented by those who were polled". Nothing is 100% certain, but (assuming the samples are chosen randomly) you can be 95% sure that the value $\hat{p}$ you get with a poll is no more than 2σ from the value p you would get if you checked the whole population.

Of course, you don't know p (and hence you don't know σ) since p is what you're trying to measure, so you use $\hat{p}$ to get the sample standard deviation $\hat{s}$ (also called the standard error in this situation) using the formula

$$\hat{s} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and you use $\pm 2\hat{s}$ as your margin of error.

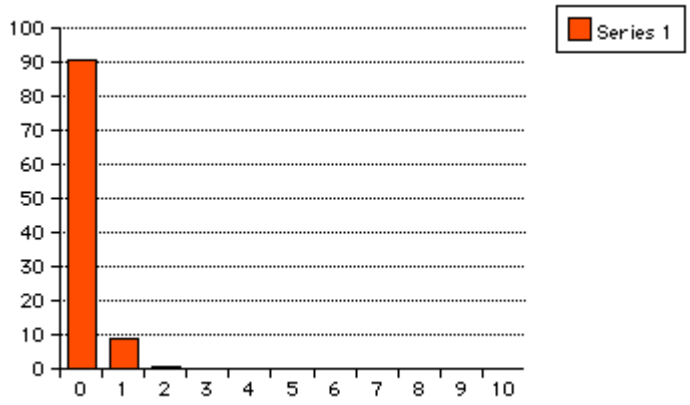

FIGURE 1. Histogram for distribution of samples with $n = 10$ when $p = 0.01$.

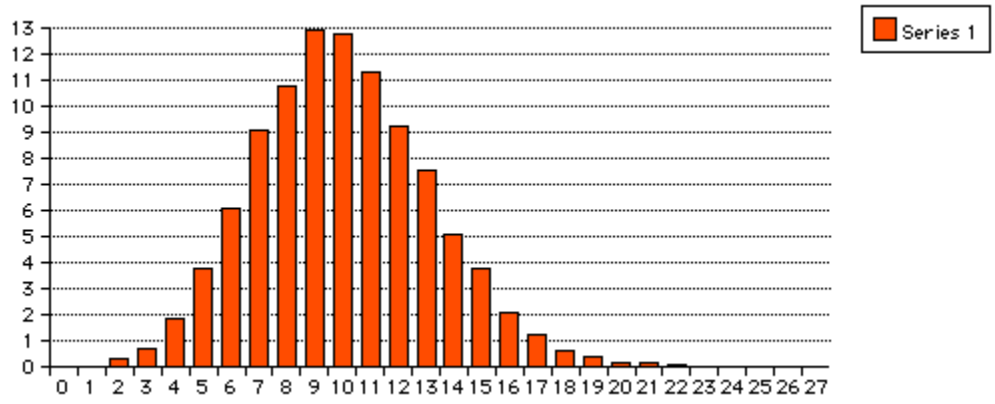

FIGURE 2. Histogram for distribution of samples with $n = 1000$ when $p = 0.01$.

(Note: the horizontal axis in the figures above gives the number x of people in the surveyed sample of size n who respond "Yes." The vertical axis is the percentage of samples for each given x-value. Thus in Figure 2, there is a bar of height 9 when $x = 7$. This means that in 9% of the samples there are 7 respondents who answer "Yes.")