AP Central

10% Assumption for Inference

by Robert W. Hayden
Plymouth State University
Plymouth, New Hampshire

A Puzzling Rule

A common question about the assumptions for inference procedures in AP Statistics is why there is attached to certain inference procedures an assumption that the sample is less than 10 percent of the population. If you are among those who find this puzzling, you are not alone! This "assumption" really is different from most of the other inference assumptions. Most of the formulae and procedures in statistics are exact under certain circumstances. These circumstances are the hypotheses of mathematical theorems. If the hypotheses are satisfied, the conclusion of the theorem is true. As in mathematics, we have to check that the hypotheses (often called "assumptions" in statistics) are true before applying the conclusion. Unfortunately, in real life the hypotheses rarely turn out to be exactly true. For mathematical results to be useful in the real world, they need to be robust; that is, the conclusions need to still be approximately true when the hypotheses are approximately true. Experience can tell us whether a technique is robust in general, but in a particular situation we rarely know how far we are from meeting the hypotheses, or how large an impact that will have on the conclusions.

The "10 percent rule" is not that kind of assumption. Instead, it is a numerical shortcut that simplifies certain computations. However, we have the option to do the exact computation and see just how far off the approximation is. In that sense it is like the "assumption" that θ must be "small" for $\sin\theta$ to "equal" θ. Here, if we restrict θ to be between, say, -1 and +1, there is only one value of θ for which equality holds exactly. Only on the assumption that $\theta = 0$ will $\theta = \sin\theta$ exactly. However, if θ is "close to" 0, $\sin\theta$ will be, too. How close to 0 must θ be? That depends on what sort of accuracy you want. For θ = 10% = 0.1 in radian measure, $\sin\theta = 0.0998$, and we are off by about two-tenths of 1 percent. If this is not good enough, we use a value for $\sin\theta$ from a table or calculator rather than approximating it with θ. Similarly, if a sample is more than 10 percent of the population, we just use the exact formula (given below) rather than use an approximation.

While the 10 percent rule applies to both means and proportions, I will deal only with means here. The same principles apply to proportions, but there the situation is further complicated by two additional approximations. They are old friends (or old enemies): the issue of n versus n -1, and the continuity correction when we approximate the binomial distribution with the normal. For all the gory details, see William Cochran's *Sampling Techniques* [1].

Inference for both means and proportions is based on theoretical results about the sampling distributions of those statistics. The basic theoretical result for means[1, 2] is that:

$\frac{s^2}{n}(1-\frac{n}{N})$ is an unbiased estimator of the variance of the sampling distribution. For inference we use the square root of this (the *standard error*):

$\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$ which is a (slightly) biased estimator of the standard deviation of the sampling distribution. For our purposes, we can look at three parts of this expression. The quantity $f = \frac{n}{N}$ is called the *sampling fraction*. The quantity:

$\sqrt{1-\frac{n}{N}}$ is called the *finite population correction* or *fpc* , and of course $\frac{s}{\sqrt{n}}$ is the more familiar approximation to the standard deviation of the sampling distribution. The approximation is exact when *fpc* = 1 and close when *fpc* is close to 1. Mathematically, we have *fpc* = 1 when the

exact when *fpc* = 1 and close when *fpc* is close to 1. Mathematically, we have *fpc* = 1 when the sampling fraction $\frac{n}{N} = 0$. This happens only when *n* = 0 or in the limit as *N* goes to ∞. The case *n* = 0 is of little practical importance as samples of size 0 contain no information; the limit as *N* goes to ∞ reflects the case of sampling from an infinite population. Early statistical theory dealt only with this limiting case. When later results were obtained for a population of *any* size, the quantity $\sqrt{1-\frac{n}{N}}$ was thought of as a correction to the earlier results. Hence the name *finite population correction*.

For situations in between *n* = 0 and *N* = ∞, we have to ask what will make the sampling fraction $\frac{n}{N}$ small. That happens when *n* is small compared to *N*. When *n* = 10% *N* = 0.1 *N*, the *fpc* = 0.94868..., and $\frac{s}{\sqrt{n}}$ will overestimate the standard error by about 5 percent, which many people consider "close enough."

f	*fpc*	usual	error%	1 - *f*/2	error%
0.00	1.00000	1	0.00	1.000	0.00
0.05	0.97468	1	2.60	0.975	0.03
0.10	0.94868	1	5.41	0.950	0.14
0.15	0.92195	1	8.47	0.925	0.33
0.20	0.89443	1	11.80	0.900	0.62
0.25	0.86603	1	15.47	0.875	1.04
0.30	0.83666	1	19.52	0.850	1.59
0.35	0.80623	1	24.03	0.825	2.33
0.40	0.77460	1	29.10	0.800	3.28
0.45	0.74162	1	34.84	0.775	4.50
0.50	0.70711	1	41.42	0.750	6.07
0.55	0.67082	1	49.07	0.725	8.08
0.60	0.63246	1	58.11	0.700	10.68
0.65	0.59161	1	69.03	0.675	14.10
0.70	0.54772	1	82.57	0.650	18.67
0.75	0.50000	1	100.00	0.625	25.00
0.80	0.44721	1	123.61	0.600	34.16
0.85	0.38730	1	158.20	0.575	48.46
0.90	0.31623	1	216.23	0.550	73.93
0.95	0.22361	1	347.21	0.525	134.79
1.00	0.00000	1	*******	0.500	*******

Table 1: The *fpc* and its approximations for various sampling fractions *f*

Table 1 allows us to get a feel for the overall picture. The first column gives the sampling fraction, *f*, for values every 0.05 throughout its range of applicability. The second column gives the corresponding *fpc* . Note the extreme cases. An *f* of 0 is the limiting case for an infinite population; an *f* of 1 means the sample is the entire population. When we make the usual approximation of ignoring the *fpc*, we are acting as though the *fpc* took on the constant value 1 (column 3), i.e., we are acting as though the population were infinite. Column 4 gives the percent error due to this approximation. Two rows of the table are especially noteworthy. For an *f* of 0.10 = 10%, our cutoff point for the 10 percent rule, the error is about 5 percent. When *f*= 1 = 100%, the *fpc* is 0. Remembering that the *fpc* is a multiplier in the standard error formula, this correctly predicts *no* error due to sampling when the sample is the entire population. This is intuitive, but not apparent when we use the approximate expression $\frac{s}{\sqrt{n}}$ for the standard error. The case represents the maximum error in using the approximation *fpc* =1.

The impact of errors made in estimating the standard error are clearer if we go on to do inference. Imagine that we take a random sample of 36 observations from a population and find a sample mean of 50 and a sample standard deviation of 18. The usual standard error approximation $\frac{s}{\sqrt{n}}$ gives $\frac{18}{\sqrt{36}} = 3$

$\sqrt{36}$

for the standard error. Since *t* is close to 2 for a 95 percent confidence interval, such an interval extends approximately from 50 - 6 = 44 to 50 + 6 = 56. However, if we knew the entire population included only 100 observations, then the *fpc* would be $\sqrt{1-\frac{36}{100}} = 0.8$, and the true standard error would be 6 x 0.8 = 4.8 and our confidence interval 45.2 to 54.8. We see that ignoring the *fpc* gives us confidence intervals that are too wide. For hypothesis tests, it makes it harder to reject the null. For the example with *n* =36, hypothesized population means between 44 and 45.2 (and also between 54.8 and 56) will not be rejected if we ignore the *fpc* but *will* be rejected if we do not ignore it. Of course, the correct decision comes from *not* ignoring the *fpc*.

Mathematical Theory Versus the Real World

I said earlier that the 10 percent rule is unlike other statistical assumptions. An important point to recognize about Table 1 is that no such table would be possible for most of the other assumptions we make in statistical inference. We cannot make a table showing how far off we might be if the population is not normally distributed or the sample is not random. Nor is it possible to correct for those problems by using a more precise formula. That is because those assumptions concern the fit between the mathematical theory and the real world. The 10 percent rule, on the other hand, is a numerical approximation that takes place entirely within the mathematics of the situation. We can calculate the exact errors due to this approximation, and correct for them if we consider them too large.

Another point to appreciate is that the choice of whether and how to make an approximation is often a subjective one. Frequently, the reasons are more historical than rational. To drive that home, let us consider just one alternative approximation to the *fpc*. You may remember Taylor's theorem from calculus. It says that "nice" functions have power series expansions that converge to the function. Often this allows us to approximate the function with a polynomial that is the first few terms of that power series expansion. We can consider *fpc* as a function of *f* and expand it in a Taylor series. The first term is just the constant 1. The approximation *fpc* = 1 amounts to a one-term Taylor approximation to the *fpc* as a function of *f*. You can see from the table that this is a pretty crude estimate. If we were to use the two terms *fpc* = 1 - ½*f* (a two-term Taylor polynomial), we would get a much more reasonable approximation (column 5 in Table 1). For *f* = 10%, for example, this polynomial would approximate the *fpc* with 0.95, compared to the exact value *fpc* = 0.94868.... The two-term approximation reduces the error by more than a factor of 10 here. If we used this approximation, and still wanted to limit our error to about 5 percent, the 10 percent rule could become the 48 percent rule, because this two-term approximation keeps the error under 5 percent until *f* is somewhere between 45 and 50 percent.

Another option would be to make no approximation at all; simply present the correct standard error formula including the *fpc* . Then the 10 percent rule would vanish and be replaced by nothing. The correct formula also has the advantage of shedding light on how the population size impacts sampling error. Users of $\frac{s}{\sqrt{n}}$ sometimes take this approximation too literally and say that the standard error does *not* depend on the population size. Clearly, that is not true, as the population's size, *N*, appears in the correct formula:

$\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$ What is true is that the degree of dependence is much less than we might have expected.

Now that we know the truth about the mysterious 10 percent, what do we tell the children? Since the *fpc* is not part of the AP curriculum, you may just wish to ignore it unless students ask about the 10 percent rule. Then your first line of defense can be to say that the 10 percent rule is there because the formula they see is an approximation that only works when the sample is a "tiny" fraction of the population. The value of 10 percent is just there because, "If we didn't put it there, you would ask, 'How tiny is "tiny"?'" It's an arbitrary cutoff point, like the age of consent or the age at which you can first work outside the home, vote, serve in the military, or buy cigarettes or alcoholic beverages -- some real-life cutoffs that may have had young people's recent attention. Just as with the ages, there is no right value. Cochran [1] uses *f* = 5% , for example. There are, however, ridiculous values. Most everyone would find 0.1 percent or 50 percent ridiculous cutoff for *f*, just as choosing 1 or 101 for the ages I mentioned would be considered ridiculous too. If you choose to ignore the *f*, and hence teach the 10 percent rule, it is very important to be sure your students understand that having a sample that is more than 10 percent of the population is not a BAD THING. This is in contrast to having a nonrandom sample (BAD THING!) or sampling from a population that is far from normal (BAD THING!). These are problems you cannot correct for. Having a sample that is more than 10 percent of the population is actually a GOOD THING. It just doesn't happen very often and is beyond the scope of the AP syllabus.

Another option is to tell students the truth. I teach in an environment where the students buy their books, and I just tell them to insert the *fpc* everywhere it is needed in their textbooks. I start them off with problems where the population size is given. Then I give them situations where they have to make a rough estimate. Perhaps for one poll the population is registered voters in the United States. Along the way they discover that Mary's estimate of 200,000,000 gives essentially the same outcome as Johnny's estimate of 150,000,000, even though they differ by 50,000,000.

The next step is to note that when the population is very large compared to the sample, we can ignore the *fpc* entirely. On exams, they have to address this issue explicitly, either plugging in a

value for *N*, or explaining why they think the population is much larger than the sample. (This is not an extra step compared to stating and checking the 10 percent assumption, and I think it is a more meaningful one.) One thing I like about teaching them about the *fpc* is that it avoids setting an arbitrary cutoff like 10 percent -- and the resulting questions. Another thing I like about it is that it provides students with an answer to the next question: "What do we do if the sample *is* more than 10 percent of the population?" It does make the formula for standard error a bit more complicated, but I have not had problems with that, and I am teaching in a college environment with students who are probably somewhat less able, less motivated, and less mathematically inclined than students in an AP class.

Finally, in the course of estimating the *size* of the population, my students have to think about what the population *is*. So for a Gallup poll on approval of the current U.S. president, one good response might be, "The population is U.S. voters of which I think there are about 200,000,000," and the student plugs that estimate into the *fpc*. I would also accept, "The population is U.S. voters, and the sample is but a tiny fraction of the population, so I will ignore the *fpc*." I would accept "U.S. adults" as the population and would be very lenient with size estimates -- order of magnitude accuracy usually suffices. Some of my students get into the humor of the situation, giving ridiculously exact estimates for U.S. voters, such as 178,230,514, and then dramatically X-ing out the *fpc* . In any event, the identification of the population to which inferences might apply is much more important than the size estimate.

Other Approximations

Earlier I said that what approximations are made and taught is often a matter of tradition. Let me leave you with a couple of examples of other approximations. In survey sampling, it is traditional to estimate margin of error based on a 95 percent confidence interval, and to use a generic critical value of 2. I did exactly this in my example with a sample of size 36 above. The error in using $t = 2$ was small compared to the error I was trying to illustrate. Indeed, the error is less than 5 percent for all sample sizes above 20. So, at the expense of having to assume "sample size must be greater than 20," we could do away with the distinction between *t* and *z* , forget about degrees of freedom for *t*, and dispense with tables or calculators for looking up critical values for *t* (assuming students can remember the constant "2"). This seems to me a *much* greater simplification than ignoring the *fpc*. While we are at it, we could also use *n* as a divisor in computing variances, and eliminate another bugaboo. We will be off by only about 5 percent for samples of size 20, and less for larger samples. Now I'm not seriously suggesting that you do that in your AP Statistics class, but I do want you to understand that these approximations and the 10 percent rule are just that: only approximations -- always optional, usually traditional, and never to be taken too seriously.

References

[1] William G. Cochran, *Sampling Techniques*, New York: Wiley, 2nd ed., 1963.

[2] Richard L. Scheaffer, William Mendenhall III, and R. Lyman Ott, *Elementary Survey Sampling*, Belmont, California: Duxbury, 1996.

Bob Hayden has been teaching at the college level since 1973, and teaching statistics since 1982. Prior to his teaching career he spent 5 years in engineering. He holds a B.S. in mathematics from M.I.T.; an M.S. in mathematics from the University of Connecticut; and a Ph.D. with a joint major in mathematics and education from Iowa State University, where he also obtained his training in statistics. He has taught at the University of Connecticut, Iowa State University, Southwest State University and Winona State University in Minnesota, and Plymouth State University in New Hampshire. Bob has authored a number of papers on teaching statistics over the past dozen years in places like The American Statistician *and the* MAA Notes *series. He has been an active member of the AP Statistics EDG for many years.*

See also...

- AP Statistics Course Home Page
- Teachers' Resources