

Sample Size and Population Size

- A large random sample almost always gives an estimate that is close to the parameter.
- **What fundamentally matters for the variability of a statistic from a random sample is the sample size**, not the population size: The variability of a statistic from a random sample does not notably depend on the size of the population.
- According to Moore/McCabe, this is true, strictly speaking, as long as the population is at least 100 times larger than the sample.
 - According to the other “state of the art” introductory statistics book, Freedman et al.’s *Statistics*: “When estimating percentages, it is the absolute size of the sample which determines accuracy, not the size relative to the population. This is true if the sample is only a small part of the population, which is the usual case” (p. 367).
 - There is a *marginal* difference, which the **finite population correction factor** (fpc) can compensate for, if the sample is a large portion of the population:
 - *Perhaps* use fpc when the sample is a large portion (say, 30-40+%) of the population - *but using it can cause uncertainty for inferring the sample’s results to a wider population* (Stats <http://www.childrens-mercy.org/stats/size/population.asp>).
 - So, even when the sample is a large portion of the population, use fpc *only when descriptive precision*, rather than *inference*, is the priority.
 - See Freedman et al., pp. 367-370:

fpc = square root of $(N - n/N - 1)$
N=population size n=sample size

- Here’s what the UCLA-Stata Resources web site has to say (http://www.ats.ucla.edu/STAT/stata/seminars/svy_stata_intro/default.htm):

FPC: This is the **finite population correction**. This is used when the sampling fraction (the number of elements or respondents sampled relative to the population) becomes large. The FPC is used in the calculation of the standard error of the estimate. If the value of the FPC is close to 1, it will have little impact and can be safely ignored. In some survey data analysis programs, such as SUDAAN, this information will be needed if you specify that the data were collected without replacement (see below for a definition of "without replacement"). The formula for calculating the FPC is $((N-n)/(N-1))^{1/2}$, where N is the number of elements in the population and n is the number of elements in the sample. To see the impact of the FPC for samples of various proportions, suppose that you had a population of 10,000 elements.

Sample size (n)	FPC
1	1.0000
10	.9995
100	.9950
500	.9747
1000	.9487
5000	.7071
9000	.3162

Sampling with and without replacement

Most samples collected in the real world are collected "without replacement". This means that once a respondent has been selected to be in the sample and has participated in the survey, that particular respondent cannot be selected again to be in the sample. Many of the calculations change depending on if a sample is collected with or without replacement. Hence, programs like SUDAAN request that you specify if a survey sampling design was implemented with or without replacement, and an FPC is used if sampling without replacement is used, even if the value of the FPC is very close to one.

http://www.ats.ucla.edu/STAT/stata/seminars/svy_stata_intro/default.htm

- In Stata: 'help svy' & see Stata manual for svy commands.

- Returning to the general issue: Freedman et al. acknowledge that the relationship of sample size and accuracy to population size is counterintuitive.
- Here's a helpful analogy: *"Every cook knows that it only takes a single sip from a well-stirred soup to determine the taste"* (Stats <http://www.childrens-mercy.org/stats/size/population.asp>).