Sampling and Surveys

Vocabulary

- Population is the entire group of individuals we want information about.
- Sample group (or just sample) is a subset of the population.
- Convenience sample is a sample was found in a manner that is easy (convenient) for the person gathering the sample.
- Voluntary response sample consists of a sample which is found by issuing a general appeal.
- A statistical study is biased if it systematically favours certain outcomes.
- A simple random sample (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

Observational Studies & Experiments

Vocabulary

- An observational study observes individuals and measures some behaviour, but does not attempt to influence the behaviour of the sample being studied.
- An experiment deliberately imposes some treatment on the sample group in order to observe the responses.
- A control group is a group of experimental subjects who are not given any treatment.
- Variables are said to be confounded when their effects cannot be distinguished from each other.

Statistical Significance

Understanding 95% Confidence Interval and Margin of Error

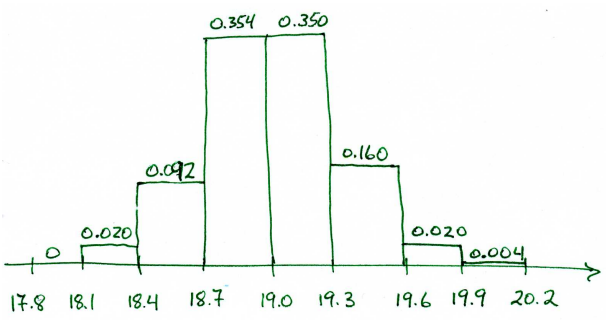
Example In 2011 in the USA, 19.0% of the adult population are smokers with a reported 95% confidence interval of 18.4%-19.6%. This was arrived at in a survey of $n = 33,014$ people.

http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6144a2.htm?s_cid=%20mm6144a2.htm_w

- $\hat{p} = 19.0\% = 0.190$ is the sample proportion of smokers.
- If we were to take many samples (of a given size) from the population of US adults, then few samples would have exactly 19.0% smokers. Most would be close to 19.0%, but they would differ by varying small amounts.
- The 95% confidence interval reported tells us that if we were to take 100 random samples of the population, 95 of those samples would have a percentage of smokers between 18.4% and 19.6%. Five of the samples would have a percentage of smokers less than 18.4% or greater than 19.6%.
- Results like these are very often reported as "the percentage of adults who are smokers in the USA is 19.0% with a margin of error of 0.6%."

This is related to the 68-95-99.7 rule for a normal distribution. If we were to take (for example) 500 simple random samples (SRS) of size 500 from the population, we might find the proportion of smokers is the following (I am making these numbers up to illustrate what's going on):

Percentage of Smokers	Number of Samples	Percentage of Samples
17.8-18.1	0	0
18.1-18.4	10	0.020
18.4-18.7	46	0.092
18.7-19.0	177	0.354
19.0-19.3	175	0.350
19.3-19.6	80	0.160
19.6-19.9	10	0.020
19.9-20.2	2	0.004



The confidence interval is related to the number of respondents, but not the size of the population

- Margin of error does NOT include errors due to undercoverage and nonresponse.
- Errors from undeercoverage and nonresponse are more serious and harder to quantify than the random sampling error (which is all that margin of error is concerned with).
- Actual error may be much larger than margin of error.
- Statistical conclusions are approximations to a complicated truth, not mathematical certainty.
- The size of the population being studied—provided it is much bigger than the samples and provided that the sample is truly random—does not matter.
- To halve the margin of error at a given confidence level, quadruple the sample size.
- The margin of error and the level of confidence are related. From the 68-95-99.7 rule, a smaller margin of error comes at the expense of a narrower confidence interval, or a higher level of confidence may be obtained by tolerating a larger margin of error.