Sampling

If you have a large population and a variable that you want to explore, you can think of picking an individual at random out of the population as a probability experiment and the variable that you measure as a random variable. This is *not* what we are going to do here. Instead, we will take a *sample.* That is, we will consider a probability experiment which is to pick n individuals at random from the population, so that our sample space is in fact the set of all possible samples. Our random variable will be the average value of the quantity of interest on all the individuals, or sometimes some other statistic for the sample. The distribution of this random variable is called the “sampling distribution.

1 The Mean

If we take a sample of n individuals chosen at random from the population and average the value of some variable X for these n individual, where X has a mean of μ and a standard deviation of σ, the resulting random variable $\overline{X}$

- has mean $\mu_{\overline{X}} = \mu$
- has standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$
- looks more like a normal curve than the distribution of X, and approaches the normal curve as n gets larger.

So if n is large enough we can assume that the sampling distribution is normal. How large is large enough?

Rules of Thumb to Assume Normality

The distribution $\overline{X}$ may be assumed to be approximately normally distributed if either

- X itself is normally distributed OR
- the distribution of X is symmetric and $n \geq 15$ OR
- $n \geq 40$.

Some people prefer $n \geq 30$ in the last case, and standards for judging a distribution normal vary widely.

2 Proportion

Suppose instead of a random variable, we have an event. That is, instead of a quantity that we can measure for each individual in the population, we have a something which is either true or false for each individual. Then we can take a sample of n individuals and count how many of them have this property. Or we could divide that number by n to get the proportion of individuals in the sample with that property. The distribution of the counts as we run over all samples is a binomial distribution. If p is the proportion *of the whole population* with this property, we have the count of successes in the sample

- has mean np
- has standard deviation $\sqrt{p(1-p)n}$
- approaches a normal distribution as n gets larger.

The distribution of $\hat{p}$, the *proportion of successes in each sample* then

- has mean p
- has standard deviation $\sqrt{p(1-p)/n}$
- approaches a normal distribution as n gets larger.

How big does n need to be? Not very if p is close to 0.5, but larger if p is close to 1 or 0 the usual rule is

Rules of Thumb to Assume Normality

The distribution $\hat{p}$ may be assumed to be approximately normally distributed if both

- $np > 5$ AND
- $n(1-p) > 5$.

3 Finite Population Correction Factor

All of the above discussion assumed that when you sample you do it with replacement. That is, once an individual is chosen to be in the sample once, they can just as well be chosen again later. This would be like dealing out a poker hand by dealing a card, writing its value down, putting it back in and shuffling, dealing another card, writing .... This is not what generally happens in real sampling, which is usually done without replacement. Just like when you draw an Ace it lowers your chance that the next card will be an Ace, when you pick one individual out of a population without replacement it changes the probabilities for the next choice. In fact the only effect this has on the above formulas is that we multiply the sample standard deviation in each case by

$$\sqrt{\frac{N-n}{N-1}}$$

where N is the size of the population. If the population is much larger than the sample size (say 100 times as large) this ha no noticable effect. This is the situation we will almost always be in. If the population is small however, you should add this factor into your formulas.