Sampling distribution of the sample proportion

In the same way that we were able to find a sampling distribution for the sample mean (SDSM), we can find a sampling distribution for the sample proportion (SDSP).

In other words, if we're dealing with a population proportion p, instead of a population mean μ, then we'll be trying to create a sampling distribution of the sample proportion, as opposed to a sampling distribution of the sample mean.

Sampling distribution of the sample proportion

Often we'll want to calculate a **population proportion** p, which is the number of subjects in our population that meet a certain condition.

For example, maybe we want to know how many students in our school have brown hair. If there are $5{,}000$ students who attend our school, it might not be possible to survey everybody. So instead we could take a random sample of 100 students and see how many of them have brown hair. This is the **sample proportion**, since it's the proportion of students in the sample with brown hair, which is given by

$$\hat{p} = \frac{x}{n}$$

where $\hat{p}$ is the sample proportion, x is the number of students in the sample with brown hair (the number of "successes"), and n is the sample size (we surveyed 100 students for our sample).

Just like for the SDSM, the **sampling distribution of the sample proportion (SDSP)** is created when we take every possible sample from our population, calculate the sample proportion for each sample, and then plot all of those sample proportions into a probability distribution.

In other words, the SDSP is the probability distribution of all possible sample proportions $\hat{p}$.

Central Limit Theorem

Remember that, whenever we're dealing with a proportion, the distribution is a binomial distribution, since every outcome is either a "success" or a "failure." So in the case of a population proportion, the original population will be modeled by a binomial distribution, not a normal distribution.

But even though the population follows a binomial distribution, we can still use the Central Limit Theorem to create a sampling distribution of the sample proportion. Just like the SDSM, the CLT tells us that the SDSP is only guaranteed to be normally distributed when we use a sample size of at least $n = 30$.

Mean, variance, and standard deviation

The mean of the sampling distribution of the sample proportion $\mu_{\hat{p}}$ will be equal to the population proportion p.

$$\mu_{\hat{p}} = p$$

The standard deviation of the sampling distribution of the sample proportion $\sigma_{\hat{p}}$, also called the **standard error of the proportion**, will be

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

where p is the population proportion and n is the sample size. We use this formula for standard error of the proportion if our population is infinite, or if the population is finite but large in comparison to our sample size (if sample size is no more than $5\,\%$ of the population, $n/N \leq 0.05$).

Of course, based on this formula for standard error, we can say that the variance of the sampling distribution of the sample proportion is

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$$

If the population proportion is unknown, we estimate the population proportion p using the next best thing, the sample proportion $\hat{p}$, and the formulas for standard error and variance become

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

$$\sigma_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n}$$

Finite population correction factor

In the case where the population is finite and $n/N > 0.05$, and assuming we're sampling without replacement, we have to apply the finite population correction factor, and in that case the correct formula for the standard error of the proportion is then

$$SE = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

where p is the population proportion, n is the sample size, and N is the size of the population. If we're applying the FPC, then the formula for variance of the SDSP is

$$\sigma_{\hat{p}}^2 = \left(\frac{p(1-p)}{n}\right)\left(\frac{N-n}{N-1}\right)$$

And again, if the population proportion is unknown, we approximate it with the sample proportion, and our formulas for standard error and variance with the FPC are

$$SE = \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}} \qquad \sigma_{\hat{p}}^2 = \left(\frac{\hat{p}(1-\hat{p})}{n}\right)\left(\frac{N-n}{N-1}\right)$$

But remember that, even if the population is finite and we're taking samples that are larger than 5% of the population, we still don't have to use the finite population correction factor if we're sampling with replacement.

Let's do an example so that we can see how these formulas work.

Example

A group of 4 people have hair colors brown, brown, brown, and blonde. Find all possible random samples of size 2 if we're sampling with replacement. If we define brown hair as a “success,” find the sample proportion for every sample, determine the probability distribution of the sample proportion, the mean of the SDSP $\hat{p}$, and the standard error $\sigma_{\hat{p}}$.

The total number of possible samples, given $N = 4$ and $n = 2$, is $N^n = 4^2 = 16$. The complete sample space, and the proportion for each sample, is

Sample	Sample proportion
brown, brown	1
brown, brown	1
brown, brown	1
brown, blonde	1/2
brown, brown	1
brown, brown	1
brown, brown	1
brown, blonde	1/2
brown, brown	1
brown, brown	1
brown, brown	1
brown, blonde	1/2
blonde, brown	1/2
blonde, brown	1/2
blonde, brown	1/2
blonde, blonde	0

Build a table for the probability distribution of the sample proportion. Because there are 16 total samples, the probability of each sample proportion will be given by the number of times that sample proportion occurs, divided by the total number of possible samples, so "count/16."

Sample proportion	P(pi)
0	1/16
1/2	6/16
1	9/16

Now we can calculate the mean of the sampling distribution of the sample proportion, $\mu_{\hat{p}}$, where $\hat{p}_i$ is a given sample proportion, $P(\hat{p}_i)$ is the probability of that particular sample proportion occurring, and N is the number of samples.

$$\mu_{\hat{p}} = \sum_{i=1}^{N} \hat{p}_i P(\hat{p}_i)$$

$$\mu_{\hat{p}} = 0\left(\frac{1}{16}\right) + \frac{1}{2}\left(\frac{6}{16}\right) + 1\left(\frac{9}{16}\right)$$

$$\mu_{\hat{p}} = \frac{3}{16} + \frac{9}{16}$$

$$\mu_{\hat{p}} = \frac{12}{16}$$

$$\mu_{\hat{p}} = \frac{3}{4}$$

Because we're sampling with replacement, we would expect this mean of the SDSP to be equivalent to the population proportion, $\mu_{\hat{p}} = p$, and we can see that it is if we calculate the population proportion.

$$p = \frac{\text{3 people with brown hair}}{\text{4 people in the population}} = \frac{3}{4}$$

Both proportions are $\mu_{\hat{p}} = p = 3/4$. The variance of the SDSP would be

$$\sigma^2_{\hat{p}} = \sum_{i=1}^{N} (\hat{p}_i - p)^2 P(\hat{p}_i)$$

$$\sigma^2_{\hat{p}} = \left(0 - \frac{3}{4}\right)^2 \left(\frac{1}{16}\right) + \left(\frac{1}{2} - \frac{3}{4}\right)^2 \left(\frac{6}{16}\right) + \left(1 - \frac{3}{4}\right)^2 \left(\frac{9}{16}\right)$$

$$\sigma^2_{\hat{p}} = \left(-\frac{3}{4}\right)^2 \left(\frac{1}{16}\right) + \left(-\frac{1}{4}\right)^2 \left(\frac{6}{16}\right) + \left(\frac{1}{4}\right)^2 \left(\frac{9}{16}\right)$$

$$\sigma^2_{\hat{p}} = \frac{9}{16}\left(\frac{1}{16}\right) + \frac{1}{16}\left(\frac{6}{16}\right) + \frac{1}{16}\left(\frac{9}{16}\right)$$

$$\sigma^2_{\hat{p}} = \frac{9}{256} + \frac{6}{256} + \frac{9}{256}$$

$$\sigma^2_{\hat{p}} = \frac{24}{256}$$

$$\sigma^2_{\hat{p}} = \frac{3}{32}$$

and then the standard error would be

$$\sigma_{\hat{p}} = \sqrt{\frac{3}{32}}$$

$$\sigma_{\hat{p}} = \frac{\sqrt{3}}{4\sqrt{2}}$$

$$\sigma_{\hat{p}} = \frac{\sqrt{6}}{8}$$

$$\sigma_{\hat{p}} \approx 0.31$$