

WHERE DOES THE 10% CONDITION COME FROM?

The text has mentioned “The 10% Condition” (at least) twice so far:

p. 407 “Bernoulli trials must be independent. If that assumption is violated, it is still okay to proceed as long as the sample is smaller than 10% of the population.”

p. 435 One of the conditions to be checked before using the normal model for sample proportions is, “The sample size, n , must be no larger than 10% of the population.”

As suggested in the first quote, this condition arises because sampling without replacement (as is usually done in surveys and many other situations) from a finite population *does not give independent* Bernoulli trials.

Contrasting two similar situations can help show the different consequences of sampling with and without replacement: Suppose we have N balls and M of them are red, the rest blue. If we mix them up in a container and randomly draw out a ball, the probability that it is red is $p = M/N$.

Situation 1 (Sampling *with* replacement): If we put the ball back in the container and repeat the mixing and random drawing, *the probability that the ball we get on this second draw is red is still $p = M/N$* . We have Bernoulli trials, since *the draws are independent* (because the probability of drawing a red ball is not affected by the result of previous trials).

Situation 2 (Sampling *without* replacement): If we do *not* replace the first ball before mixing and drawing again, and if the first ball is red, the second ball is now drawn from a population with $N-1$ balls, of which $M-1$ are red, so the probability that it is red is $(M-1)/(N-1)$, which is different from the probability $p = M/N$ that the first ball drawn was red. If the first ball drawn is blue, the probability that the second ball is red is $M/(N-1)$, which is also not the same as p . If the first two balls drawn are red, the probability that the third ball drawn is red is $(M-2)/(N-2)$ (assuming we do not replace the first two balls, and we mix and randomly draw), and so forth. *The successive draws are **not** independent*, because the probability of drawing a red ball on the next draw changes depending on what we have previously drawn.

Now suppose that (in each of these situations) we do n draws, and let the random variable Y_i ($i = 1$ for Situation 1 and $i = 2$ for Situation 2) be the total number of red balls we obtain in the n draws.

We know that in Situation 1 (sampling *with* replacement), Y_1 is a binomial random variable, with distribution $P(Y_1 = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

In Situation 2 (sampling *without* replacement), Y_2 is called a *hypergeometric* random variable, with distribution $P(Y_2 = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$.

You may have seen the hypergeometric distribution in M362K. If you haven't, you might like to try to derive the formula for $P(Y_2 = k)$. [Hint: Think of a sequence of k draws in order as an outcome. The numerator counts the number of outcomes with exactly k red draws, and the denominator counts the total number of outcomes.]

Note: If we had an infinite population, with proportion p of “successes” (for example, consider tosses of a die with one side red, so that $p = 1/6$ for getting red, a “success”), then we would also get the binomial random variable Y_1 , since in this case there is no distinction between sampling with and without replacement.

The mean and variance of the hypergeometric random variable Y_2 are

$$E(Y_2) = n \frac{M}{N} \quad (1)$$

$$\text{Var}(Y_2) = \frac{N-n}{N-1} n \frac{M}{N} \left(1 - \frac{M}{N}\right) \quad (2)$$

Here are a couple of sources for the proofs of equations (1) and (2), in case you're interested:

S. Ross, *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006. Example 8j in Section 4.8.3 (p. 181) gives a proof using definitions and combinatorial identities.

M. Iltis, Week 4 Notes for Statistics 224, available at <http://www.stat.wisc.edu/courses/st224-iltis/notes4.pdf>. This gives a proof (pp. 12 – 13) that starts out in a way similar to the arguments in in the Math Box on p. 410 of the textbook, by writing Y_2 as a sum of Bernoulli random variables. However, the Bernoulli random variables in this case are not independent, which makes the calculation of the variance a little more complicated.

We will compare these formulas with the corresponding formulas for the expected value of the binomial random variable Y_1 in Situation 1. Recall (from p. 409 or 410 of the textbook) that

$$E(Y_1) = np$$

$$\text{Var}(Y_1) = np(1-p)$$

Since $p = \frac{M}{N}$, *the binomial random variable Y_1 and the corresponding hypergeometric random variable Y_2 have the same mean.*

Comparing the formulas for the variances of the binomial random variable Y_1 and the corresponding hypergeometric random variable Y_2 , (remembering that $p = \frac{M}{N}$), we see that

$$\text{Var}(Y_2) = \frac{N-n}{N-1} \text{Var}(Y_1) \quad (3)$$

The factor $\frac{N-n}{N-1}$ is called a *finite population correction factor*, since (by the comment toward the top of p. 2 above), if the population were infinite, we would have the random variable Y_1 .

Consequences of Equation (3):

1. Since $N - n < N - 1$, the finite population correction factor is less than 1. In other words, the variance when sampling without replacement is smaller than the variance when sampling with replacement.
2. If the 10% condition is satisfied, then $\frac{n}{N} \leq 0.1$, so

- i. $\frac{N-n}{N-1} > \frac{N-n}{N} = 1 - \frac{n}{N} \geq 1 - 0.1 = .9$

Taking square roots of both sides of equation (3) and using this inequality now gives

$$\text{SD}(Y_2) = \sqrt{\frac{N-n}{N-1}} \text{SD}(Y_1) \geq \sqrt{.9} \text{SD}(Y_1) \geq 0.9486 \text{SD}(Y_1)$$

3. In other words, if the 10% condition is satisfied, the standard deviation of number of successes when sampling with replacement is within a factor of about .05 of the standard deviation when sampling
4. Although we have used the variables Y_1 and Y_2 that measure counts, these results for means and standard deviations apply to proportions as well:

Let \hat{p}_1 and \hat{p}_2 denote the sample proportions when sampling with and without replacement, respectively. In other words,

$$\hat{p}_1 = Y_1/n \text{ and } \hat{p}_2 = Y_2/n$$

Exercise: Use properties of means and variances plus the equations above to show that

$$E(\hat{p}_2) = p \text{ (which equals } E(\hat{p}_1)), \text{ and}$$

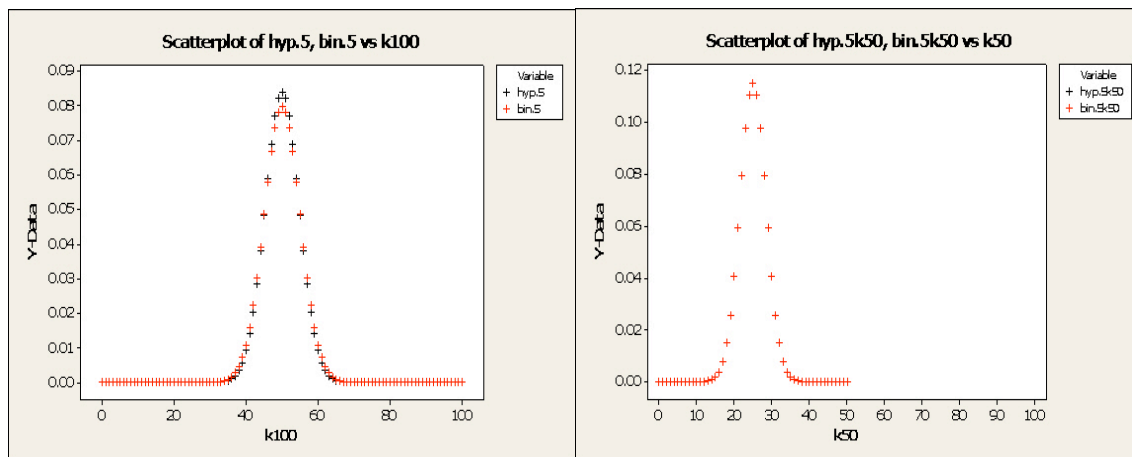
$$\text{Var}(\hat{p}_1) = \frac{N-n}{N-1} \text{Var}(\hat{p}_2)$$

5. As we will see in Chapters 19 and 20, the statistical techniques applied to sample proportions just depend on the means and standard deviations, so we really just need to worry about whether or not they are the same, or close enough, under the two sampling schemes. Items 3 and 4 above tell us exactly this.
6. Not everyone agrees that the 10% condition is good enough. Some practitioners advocate using the **5% condition**: It's okay to use the normal model when sampling without replacement provided the sample size, n , is no larger than 5% (that is, one twentieth) of the population size, N . Under the 5% condition, a calculation such as that in item 2 above will show that

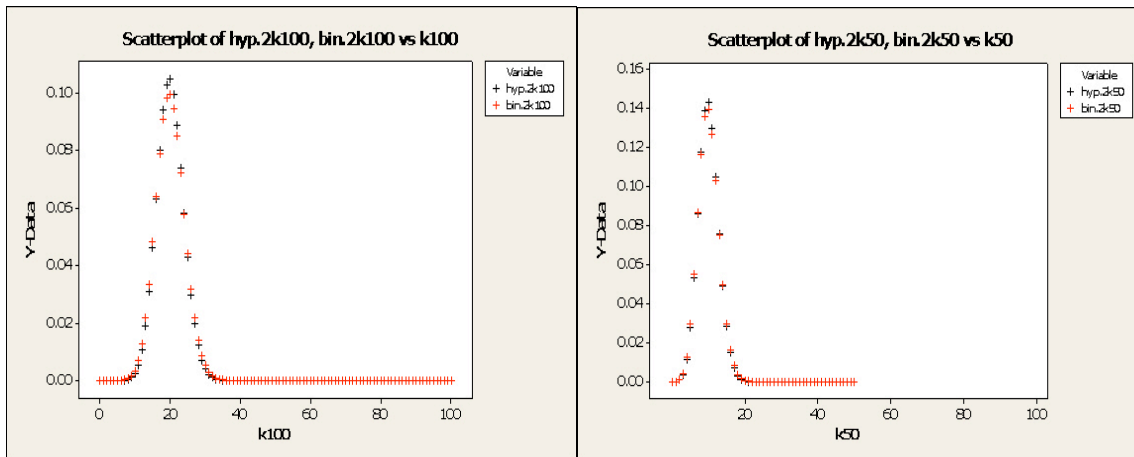
$$\text{SD}(Y_2) \geq \sqrt{.95} \text{SD}(Y_1) \geq 0.9746 \text{SD}(Y_1)$$

In fact, when the 10% or 5% conditions are satisfied, the binomial and hypergeometric distributions are quite close, as illustrated by the following graphs.

a. The *left* graph shows the probability distributions of the hypergeometric distribution with $N = 1000$, $M = 500$, and $n = 100$ (black) and the corresponding binomial distribution with $p = 0.5$ ($= M/N$) and $n = 100$ (red). Both distributions have mean $np = 50$. Note that in this case, $N/n = 0.1$, so the 10% condition is just barely satisfied. The graph shows that the hypergeometric and binomial distributions with these parameters are very similar, as we would hope when the 10% condition is satisfied. The *right* graph shows the probability distributions of the hypergeometric distribution with $N = 1000$, $M = 500$, and $n = 50$ (black) and the corresponding binomial distribution with $p = 0.5$ ($= M/N$) and $n = 50$ (red). Both distributions have mean $np = 25$. Note that in this case, $N/n = 0.05$, so the 5% condition is barely satisfied. In this graph, the two distributions are so similar that the symbols for the hypergeometric distribution are completely hidden by those for the binomial distribution.



b. The *left* graph below shows the probability distributions of the hypergeometric distribution with $N = 1000$, $M = 200$, and $n = 100$ (black) and the corresponding binomial distribution with $p = 0.2$ ($= M/N$) and $n = 100$ (red). Both distributions have mean $np = 20$. Note that in this case, $N/n = 0.1$, so the 10% condition is just barely satisfied. Again, the graph shows that the hypergeometric and binomial distributions with these parameters are very similar, as we would hope when the 10% condition is satisfied. The *right* graph shows the probability distributions of the hypergeometric distribution with $N = 1000$, $M = 200$, and $n = 50$ (black) and the corresponding binomial distribution with $p = 0.2$ ($= M/N$) and $n = 50$ (red). Both distributions have mean $np = 10$. Note that in this case, $N/n = 0.05$, so the 5% condition is barely satisfied. In this graph, the two distributions are not quite as similar as in the right graph in part (a). Why should this not be surprising? (Hint: Consider the success/failure condition.)



Note: p. 441 of the text also states a 10% condition for the Central Limit Theorem: “When the sample is drawn without replacement (as is usually the case), the sample size, n , should be no more than 10% of the population.” The reasoning behind this is similar to the reasoning behind the 10% condition for proportions:

- If the random variable Y has mean μ and standard deviation σ , then
 - The expected value of the sampling distribution of means for samples of size n is also μ , whether or not sampling is with or without replacement.
 - When sampling is *with* replacement, the standard deviation of the sampling distribution of means of size n is $\frac{\sigma}{\sqrt{n}}$.
 - When sampling is *without* replacement, from a finite population of size N , the standard deviation of the sampling distribution of means of size n is $\sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}$.
- Reasoning as above now establishes the 10% and 5% conditions.
- For more details, see Section 8.7 of Berry and Lindgren, *Statistics: Theory and Methods*, 1990, Brooks/Cole.