Thus far in the specialization, we have been emphasizing the importance of the **standard error** of a statistical estimate for making inference about the parameter being estimated. Recall from Week 4 of the first course of this specialization that this quantity provides us with an estimate of the standard deviation of the sampling distribution of estimates that would arise if we had drawn repeated samples of the same size and computed the same estimate for each random sample. In a simplified sense, the standard error gives us a sense of the uncertainty associated with our estimate. Estimates with smaller standard errors are thus consider more precise.

So what exactly impacts a standard error in terms of a study design? Below is a list of design features that would generally affect the standard error of an estimate. There is clearly a cost component associated with this list, as some of the design features would certainly require more financial resources.

1) The variance of the variables of interest that are used to compute the estimate.

In general, the more variability that is associated with a given variable being measured, the more imprecise estimates based on that variable will be. This makes careful and precise measurement of the variables of interest very important for any given study.

2) The size of the sample.

Larger samples will tend to produce sampling distributions with less variability (or, in other words, estimates with smaller standard errors). The more sample that can be measured, the better, but we also have to think carefully about the first point above. Just because we have a set of “big data” does not mean that we have a collection of precise measurements. Very unusual measures (outliers) could have strong influence on the variance of a given variable, and this requires careful descriptive assessment.

3) The amount of dependence in the observations collected, possibly due to cluster sampling.

In studies were clusters of units with similar characteristics are measured, the data collected will not be entirely independent with a given cluster (neighborhood, clinic, school, etc.). This is because units coming from the same cluster will generally have similar values on the variables of interest, and this

could happen for a variety of reasons. This lack of independence in the observations collected reduces our **effective sample size**; we don't have as much unique information as the size of our sample would suggest. We can account for this dependence within clusters by using specialized statistical procedures to estimate standard errors in a way that accounts for cluster sampling. The same problem arises in longitudinal studies, where we collect repeated measurements from the same individuals over time. While it may look like we have a large sample of observations, many of these observations will be strongly correlated with each other, and we need to account for this. In general, with these types of clustered data, standard errors will tend to be much larger, because the estimates computed across different studies will entirely depend on what clusters are under study. If the clusters tend to vary substantially in terms of the measures of interest, the variability of the sampling distribution will increase! Furthermore, the larger the sample size selected from each cluster (and thus the smaller the sample of clusters), the larger the standard errors will tend to be.

4) The stratification of the target sample.

If we select a **stratified sample** from a target population (see Week 4 of Course 1), we will tend to produce estimates with increased precision, because we are removing between-stratum variance from the variability of our estimates by design! Stratification of samples is always an important consideration, for this reason.

5) The use of sampling weights to compute our estimates.

While sampling weights are often necessary to compute unbiased population estimates, the use of weights in estimation can inflate the variance of our estimates. We can use specialized statistical procedures to make sure that our standard errors reflect the uncertainty in our estimates due to weighting. In general, the higher the variability in our weights, the more variable our estimates will be.

These five features are generally the main drivers of standard errors, but other design features may also ultimately affect standard errors (e.g., imputation of missing data). We will touch on these throughout the specialization.

To make these points clear, the figure below simulates nine sampling distributions for a population mean based on combinations of sample size (n = 500, 1000, and 5000) and the size of the clusters

(no clusters, clusters with 10 units sampled from each, and clusters with 50 units sampled from each) in a cluster sample design. The effects of these design decisions on the variability of the sampling distributions is clear: with larger sample sizes (going down the columns), the spread of the sampling distribution shrinks (lower standard errors!). With larger clusters, the spread of the sampling distribution increases (higher standard errors!).