Survey Sampling

Kosuke Imai
Department of Politics, Princeton University

February 19, 2013

Survey sampling is one of the most commonly used data collection methods for social scientists. We begin by describing the simplest method of survey sampling, called *simple random sampling*. Suppose that we are conducting election polling and are interested in estimating the proportion of voters who support Obama in a battle ground state, say Florida. There are N voters in this state and we call this population of voters $\mathcal{P}$. Thus, N denotes the *population size*. We assume that the complete list of N voters is available, allowing us to sample a subset of them from this list. Such a list is called a *sampling frame*. Before we begin our sampling, we determine the total number of voters we are going to interview. This number represents the *sample size* and is denoted by n.

Simple random sampling refers to the procedure in which a researcher randomly samples n voters from the list of N voters with equal probability. There are two key characteristics of this procedure. First, we are sampling exactly n voters without replacement. That is, every voter gets sampled at most once. Secondly, each voter has an equal probability of being sampled. Clearly, this *sampling probability* is equal to n/N for every voter on the sampling frame.

Below, we introduce two inferential approaches, one design-based and the other model-based, to survey sampling. In the simple case we are considering, both approaches give substantially similar estimation procedures. For example, from both perspectives, the sample proportion of Obama's supporters in a poll is a good estimate of the population proportion. However, the two approaches are conceptually quite different. The basic idea of the design-based approach is that all statistical properties of one's estimator are based solely on the actual data collection procedure employed by the researcher (here, simple random sampling). Thus, this approach faithfully follow the research design. In contrast, the model-based approach requires the researcher to specify a probability model. Such a model represents an approximation to the actual data generating process. As described in more detail below, each approach has its advantages and disadvantages.

1 Design-based Inference

Statistical inference is about learning what we do not observe, called parameters, from what we do observe, called data. What sets apart statistical inference from other mathematical (or non-mathematical) method of inference is that it can calibrate the inferential uncertainty using formal quantities such as standard errors and confidence intervals. But where do these uncertainty measures come from? The design-based inference is an approach where all statistical properties of estimation procedures are derived solely from the key features of research design imposed by the researcher. In the current context, the fact that a researcher collect the data via simple

random sampling allows us to characterize the degree of accuracy for the resulting estimate. The design-based inference is attractive because it faithfully respects the key features of data collection methods employed by researchers. The approach also provides intuition about how statistical properties are related to actual research.

1.1 Setup

To formalize the idea, let X_i represents a binary variable indicating whether or not voter i in the population $\mathcal{P}$ supports Obama. We define this quantity for each voter in the population. That is, for $i = 1, 2, \ldots, N$, $X_i = 1$ ($X_i = 0$) implies that voter i supports (does not support) Obama. We now introduce the *sampling indicator* variable Z_i for each voter in the population, representing whether or not the voter is sampled. Thus, $Z_i = 1$ ($Z_i = 0$) indicates that voter i is sampled (not sampled). According to our simple random sampling procedure, we sample exactly n voters with equal probability and without replacement. This implies that $\sum_{i=1}^{N} Z_i = n$ and $\Pr(Z_i = 1) = n/N$.

How many unique samples of voters can this simple random sampling produce? To answer this question, it is helpful to consider a scenario where we sample and interview one voter at a time. For the first interview, we have N choices, and we choose the second interviewee from $N-1$ voters, and finally the nth voter will be selected from $N - n + 1$ remaining voters. Thus, all together, we have $N \times (N-1) \times \cdots (N-n+1)$ ways of interviewing n voters from a population of N voters. What we just did is to count all *permutations* by assuming that the order of sampling matters. However, we actually do not care about the order of sampling and hence need to remove double-counting. How many duplicates do we have? We have a total of n voters in our sample, and using the same logic of permutations above, there are a total of $n! = n \times (n-1) \times \cdots 1$ ways to arrage these n voters of any distinct sample. Thus, all together we arrive at the standard formula for combinations

$$\binom{N}{n} \;=\; \frac{N \times (N-1) \times \cdots (N-n+1)}{n!} \;=\; \frac{N!}{n!(N-n)!} \tag{1}$$

Since each distinct sample is equally likely, the probability that a particular combination of n voters is selected equals $1/\binom{N}{n} = n!(N-n)!/N!$.

1.2 Estimation

Our quantity of interest, or *estimand*, is the proportion of Obama supporters in the population, and this quantity can be written as the population mean of X_i, i.e., $\overline{X} = \sum_{i=1}^{N} X_i/N$, because the population $\mathcal{P}$ consists of N voters. Our goal here is to estimate the population proportion of Obama supporters using our sample of n voters. The sample proportion appears to be a reasonable estimator for the population proportion. In the current notation, we can write the sample proportion (or sample mean) as $\bar{x} = \sum_{i=1}^{N} Z_i X_i/n$ since those who are not in the sample and have $Z_i = 0$ will not contribute to this quantity.

What are the statistical properties of this estimator? We consider how this estimator performs over hypothetical repeated sampling. We will never know how close the proportion of Obama supporters in our actual sample to the population proportion, but we can say something about how our estimator behave if we were to repeat the same sampling process many times. Of course, we can never administer the same survey to the same population at the same point of time

more than once. Nevertheless, by knowing how one's estimator performs under this hypothetical scenario, we can gain a better understanding of how reasonable our estimator is.

To derive the statistical properties of the sample mean under the design-based framework, we take the position that the only source of randomness comes from the procedure of simple random sampling. Formally, for each voter i, Z_i is regarded as a random variable with probability distribution defined by simple random sampling whereas X_i is assumed to be a fixed (but possibly unobserved depending on the realization of Z_i) quantity. The idea is that each voter i is either a supporter of Obama, in which case we have $X_i = 1$, or not, i.e,. $X_i = 0$, and this quantity is fixed. While X_i is a fixed variable for any given voter, it has a distribution over the population of voters because a certain number of voters are Obama supporters while others are not. Now, suppose that X_i represents one's income. Income is a fixed quantity for each voter at the time of survey, but we can still consider a distribution of income over the population of voters.

Given this distinction between fixed and random variables, we compute the expected value of our estimator over repeated sampling. What does this mean? As discussed above, there exist a total of $\binom{N}{n}$ possible samples one can end up with when applying simple random sampling. Since each sample contains different combinations of voters with different realizations of Z_i, its sample mean may be different and may not even be close to the population mean, which we would like to estimate. However, the following calculation shows that on average we get the right answer,

$$\mathbb{E}(\bar{x}) \;=\; \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}(Z_i X_i) \;=\; \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}(Z_i)X_i \;=\; \frac{1}{n}\sum_{i=1}^{N}\frac{n}{N}X_i \;=\; \overline{X} \tag{2}$$

where the first equality follows from the fact that the expectation is a linear operator, the second equalitiy holds because X_i is a fixed quantity, and the third equality comes from the binary nature of the sampling indicator variable, i.e., $\mathbb{E}(Z_i) = \Pr(Z_i = 1)$. Note that the expectation is taken with respect to Z_i, which is the only random variable in this setting. We have just shown that the sample mean is *unbiased* for the population mean. That is, even though the sample mean we compute from our actual sample may differ from the population parameter for any particular draw, the average value of this estimator over (hypothetical) repeated sampling equals exactly the estimand of interest.

The unbiasedness of the sample mean over repeated sampling is helpful, but how close is my actual estimate to the population mean? While one cannot know exactly how far off one's estimate is from the truth, we can characterize the average distance from the truth, again, over repeated sample. One such measure is the variance of sampling distribution of the estimator. Under the design-based framework, this variance is given by,

$$\mathbb{V}(\bar{x}) \;=\; \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \tag{3}$$

where $(1-n/N)$ is the finite sample correction and $S^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2/(N-1)$ is the population variance. The finite sample correction term adjusts the variance by taking into account the sample size relative to the population size. If the sample size equals the population size, then the sample mean equals the population mean and hence the variance is zero. In most cases, the sample size is small relative to the population size, which means that this term is close to one, resulting in a little impact on the variance.

The variance expression given in equation (3) is a theoretical result because it is a function of the population variance of X_i, which is an unknown quantity. However, we can estimate S^2 in the

exactly same way in which we estimated the population mean using the sample mean. In fact, we can use the sample variance of X_i, $s^2 = \sum_{i=1}^{N} Z_i(X_i - \bar{x})^2/(n-1)$, as an unbiased estimator for its population variance, S^2. An unbiased estimator for the variance of the sampling distribution, then, is given by,

$$\hat{\sigma}^2 \;=\; \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \tag{4}$$

where $\mathbb{E}(\hat{\sigma}^2) \;=\; \mathbb{V}(\bar{x})$.

We have seen that the sample mean is a good estimator of the population mean and the sample variance is a good estimator of the population variance. Did you notice any pattern here? Indeed, this is called *plug-in principle*: a sample analogue of a population parameter is a good estimator because it usually equals the population parameter in expectation.

How do we derive the variance expression given in equation (3)? The calculation required for derivation is quite similar to what we did for showing the unbiasedness of the sample mean, though of course it is much more involved (the slides provide details about each step of the derivation). The basic idea is still the same, however. Recall that the variance represents the average square distance from its mean, i.e., $\mathbb{V}(\bar{x}) = \mathbb{E}\{\bar{x} - \mathbb{E}(\bar{x})\}^2$ where in this case $\mathbb{E}(\bar{x}) = \overline{X}$. Thus, the derivation amounts to taking the expectation with respect to the sampling indicator variable Z_i, which again is the only source of randomness under the design-based approach.

1.3 Unequal Probability Sampling and Sampling Weights

So far, we have focused on the simplest probability sampling methods, simple random sampling. However, in practice, this procedure may not be appropriate and researchers may wish to sample respondents with different sampling probabilities. For example, it is often a common practice to *over-sample* minorities in order to ensure that the resulting sample contains enough minority respondents. Moreover, even if one applies simple random sampling, it is often the case that the refusal to participate in survey (i.e., unit non-response) results in unequal sampling probability across different individuals. This means that when making inference about the population, we must make additional statistical adjustments. The sample mean is no longer unbiased for the population mean because the sample is not representative of the population.

The most common way to make an adjustment is through weighting. We will weight each observation by the inverse of sampling probability. This makes sense because, for example, over-sampled minorities need to be underweighted so that their sample proportion will resemble their population proportion. For example, if an African American voter is twice more likely to be sampled when compared to other voters, then the sample proportion of African American voters is, on average, twice as large as their population proportion. Thus, the African American voters in a sample need to be downweighted by a factor of two in order for the sample to be representative of the population. This simple idea is the basis of the following classical weighting estimator by Horvitz and Thompson,

$$\tilde{x} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i X_i}{\pi_i} \tag{5}$$

where $\pi_i = \Pr(Z_i = 1)$ is the sampling probability for each voter i in the population. This probability is indexed by i in order to allow for the possibility that the sampling probability varies

from one voter to another. In the case of simple random sampling, everyone has the same sampling probability so that $\pi_i = n/N$ for all i.

Showing that this inverse probability weighting leads to an unbiased estimator can be done in the exactly same manner as before. Taking the expectation with respect to Z_i and treating X_i as a fixed quantity, we have,

$$\mathbb{E}(\tilde{x}) \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{E}(Z_i)X_i}{\pi_i} \;=\; \overline{X}. \tag{6}$$

where $\mathbb{E}(Z_i) = \pi_i$. The derivation of the variance of this estimator is similar except that, again, it is much more involved than deriving its expectation.

Although we have so far assumed that the sampling probability, whether it is constant or varying across units, is known, in practice we often must estimate it from data. As mentioned earlier, a main reason for this is the existence of unit nonresponse, which results in different and unknown sampling probabilities across units. We note that survey weights typically available in many social science surveys are constructed using the method of post-stratification. Under this methodology, we first obtain the population distribution of certain covariates (e.g., demographics from the census). Then, survey weights are calculated such that after weighting the observations of a sample the sample distribution of these covariates approximates the population distribution. While this approach guarantees that the sample is representative of the population with respect to these covariates, it is not necessarily the case that there may be significant differences in other observed and unobserved covariates between the sample and the population. Later in this course, we will have a more complete discussion of the issues that arise from such sample selection.

2 Model-based Inference

The design-based inference introduced above is clean and intuitive, and does not require any strong assumptions. This approach has limited applicability, however, since the derivation of variance of the sample mean is challenging even under simple random sampling. It is easy to imagine that in more complex but realistic situations, the design-based approach may face analytical and other difficulties.

To address this issue, we introduce the model-based inference. This approach requires researchers to directly model the data generating process by specifying a probability distribution that characterize the population we are interested in. The advantage is that under this approach all the useful tools of probability theory can be used to derive the statistical properties of one's estimator. The model-based inference gives much greater analytical flexibility than the design-based approach because the latter is not allowed to deviate from the actual data collection mechanisms. A major disadvantage of the model-based approach, however, is that in some cases a model selected by researchers does not approximate the actual data generating process well. Constructing a good model requires a careful balancing act. As famously put by the statistician George Box, "All models are false but some are useful." Therefore, the credibility of model-based inference relies upon the critical question of whether one's model is a useful approximation of the real world.

2.1 Estimation

Let's go back to the polling example. A model-based approach may assume that the number of Obama supporters in one's sample, denoted by $X^* = \sum_{i=1}^n X_i$, follows a Binomial distribution with probability p and size n. Is this a reasonable model? In substantive terms, the Binomial model assumes that n independent draws are sampled from an infinite population of voters where the proportion of Obama supporters is exactly p. This model is a good approximation to the simple random sampling of voters in Florida for two reasons. First, the number of voters in Florida is quite large and hence assuming that the population size is infinite is reasonable. Secondly, using p as the proportion of Obama supporters in the population of Florida voters, obtaining n independent draws from the Binomial distribution resembles the simple random sampling procedure of n voters.

Given this setup, we estimate the population proportion of Obama supporters p using the sample proportion $\hat{p} = X^*/n$. Now, recall that the mean and variance of a Binomial random variable are $\mathbb{E}(X^*) = np$ and $\mathbb{V}(X^*) = np(1-p)$, respectively. Using this fact, it is straightforward to show the unbiasedness of this estimator and its variance as follows,

$$\mathbb{E}(\hat{p}) \;=\; \frac{np}{n} \;=\; p \quad \text{and} \quad \mathbb{V}(\hat{p}) \;=\; \frac{\mathbb{V}(X^*)}{n^2} \;=\; \frac{p(1-p)}{n} \tag{7}$$

Again, the variance is a theoretical quantity and must be estimated from data because p is unknown. This is easily done by following the plug-in principle introduced above: we replace p with its unbiased estimate $\hat{p}$ in the variance expression $\mathbb{V}(\hat{p})$ given in equation (7).

In addition to the mean and variance of the sampling distribution, the model-based approach also enables inference based on *asymptotic approximation* where the sample size is assumed to approach infinity in the limit. Of course, in practice, we will never have infinite amount of data. Asymptotic approximation is useful because it allows us to use powerful theorems of probability theory to derive the statistical properties of various estimators in the limit as the sample size tends to infinity. Such a calculation is often difficult in finite sample settings. Thus, when the sample size is large enough, one can apply these asymptotic approximations.

One important asymptotic theorem is the (weak) *law of large numbers*. The theorem states that the sample mean approaches the population mean as the sample size increases. More formally, if $\{X_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with mean μ and finite variance σ^2, then letting $\overline{X}_n = \sum_{i=1}^n X_i/n$ denote the sample mean, we have

$$\overline{X}_n \;\xrightarrow{p}\; \mu \tag{8}$$

where "$\xrightarrow{p}$" denotes the *convergence in probability*, which is formally defined as $\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| > \epsilon) = 0$ for any $\epsilon > 0$. Thus, convergence in probability states that for any arbitrarily small positive number ϵ, the probability that the deviation of the sample mean deviates from the population mean is greater than ϵ becomes negligible as the sample size approaches infinity.

In the case of survey sampling, this implies that the sample mean is not only unbiased for the population mean, i.e., $\mathbb{E}(\overline{X}_n) = \mu$, but it is also *consistent*, becoming arbitrarily close to the population mean as the sample size increases. Note that unbiased and consistency are different properties. For example, consider another estimator of the population mean $\overline{X}_n + 1/n$. This (rather silly) estimator is biased because $\mathbb{E}(\overline{X}_n + 1/n) = \mu + 1/n \neq \mu$. However, it is still consistent for the population mean as the bias term $1/n$ tends to zero as the sample size approches infinity. One could also say that this estimator is asymptotically unbiased (though it is biased in a finite sample) because the bias disappears asymptotically. In fact, consistency implies asymptotic unbiasedness, while the converse is true so long as the variance goes to zero as the sample size increases.

2.2 Confidence Intervals

Another important probability theorem used for asymptotic approximation is the *central limit theorem*. This theorem enables us to derive the (asymptotic) sampling distribution of the sample mean, going beyond its mean and variance given in equation (7). Using the knowledge of sampling distribution, we can make probabilistic statements about the properties of estimators. Specifically, the theorem states that the sample mean is distributed (over repeated sampling) according to a normal distribution. Formally, if $\{X_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with mean μ and finite variance σ^2, then we can write the central limit theorem as,

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1) \tag{9}$$

where "$\xrightarrow{d}$" represents the *convergence in distribution*. We say that a sequence of random variable, X_n, converges to another random variable X, if $\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$ for all x with $P(X \leq x)$ being continuous at every x. In words, the distribution function of X_n converges to that of X as the sample size approaches infinity.

What is the intuition behind equation (9)? Under the model-based approach, we can derive the variance of the sample mean as $\mathbb{V}(\overline{X}_n) = \sum_{i=1}^n \mathbb{V}(X_i)/n^2 = \sigma^2/n$ using the assumption that X_i is an i.i.d. random variable. Given this result, it is clear that the left hand side of equation (9) represents the standardized version of the sample mean $\overline{X}_n$ where the mean of the sampling distribution, i.e., μ, is first subtracted from it and then this difference is divided by the standard deviation, i.e., $\sigma/\sqrt{n}$. This standardized version of the sample mean, according to the theorem, has the standard normal distribution in the limit as the sample size tends to infinity.

We present another way to understand the central limit theorem. First, rewrite the theorem as $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by multiplying the random variable in the left hand side of equation (9) with a constant σ. We know that by law of large numbers, $\overline{X}_n$ approaches μ as the sample size increases, implying that $\overline{X}_n - \mu$ in the numerator becomes smaller and smaller as n tends to infinity. This is off set exactly by the increase of $\sqrt{n}$ term and as the sample size increases the distribution of $\sqrt{n}(\overline{X}_n - \mu)$ becomes the normal distribution with mean zero and variance σ^2. For this reason, $\sqrt{n}$ is called the *rate of convergence*.

How can we use the central limit theorem under the model-based inference? The theorem provides a direct large-sample (i.e., asymptotic) approximation to the sampling distribution of the sample mean. To do this, we assume that for large enough n, equation (9) holds approximately, i.e., $\sqrt{n}(\overline{X}_n - \mu)/\sigma \overset{\text{approx.}}{\sim} \mathcal{N}(0,1)$. Dividing this random variable by $\sigma/\sqrt{n}$ and then adding μ, we have the following approximate sampling distribution of $\overline{X}_n$,

$$\overline{X}_n \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{10}$$

Thus, in addition to its mean and variance, we now know the sampling distribution in the limit as the sample size increases.

In typical situations, we do not know σ^2 and hence estimate it from the sample. Again, we apply the plug-in principle and use the sample variance, which is an unbiased and consistent estimator for the population variance σ^2. Does the above approximation still hold when we replace the population variance with its estimate? The answer turns out to be yes. That is, we

have $\sqrt{n}(\overline{X}_n - \mu)/\hat{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$. This result is based on another important result in probability theory, called the *Slutzky Theorem*, which states that if $X_n \xrightarrow{p} x$ and $Y_n \xrightarrow{d} Y$, then

$$X_n + Y_n \xrightarrow{d} x + Y \quad \text{and} \quad X_n Y_n \xrightarrow{d} xY. \tag{11}$$

Specifically, we apply the Slutzky Theorem in the following manner,

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\hat{\sigma}} \;=\; \underbrace{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}_{\xrightarrow{d}\mathcal{N}(0,1)} \times \underbrace{\frac{\sigma}{\hat{\sigma}}}_{\xrightarrow{p}1} \;\xrightarrow{d}\; \mathcal{N}(0,1) \tag{12}$$

where the first term converges to the standard normal random variable because of the central limit theorem and the second term converges in probability to one because $\hat{\sigma}^2$ is a consistent estimator of σ^2. Together, the result directly implies that the so-called *z-score* follows the standard normal distribution in a large sample,

$$z\text{−score} \;=\; \frac{\overline{X}_n}{\text{s.e.}} \;=\; \frac{\overline{X}_n}{\hat{\sigma}/\sqrt{n}} \;\xrightarrow{\text{approx.}}\; \mathcal{N}(0,1) \tag{13}$$

Finally, we derive the asymptotic confidence intervals, which are often used as a way to express the uncertainty of an estimator. The basic idea is to construct an interval such that over repeated sampling it covers the true value, say, 95% of time. Note that what is random here is the confidence interval rather than the truth. The probability that any given confidence interval reported by the researcher contains the true value is either zero or one since the truth is fixed. However, over hypothetical repeated sampling, the procedure should produce confidence intervals (which vary from one sample to another) that contain the truth for the pre-specified proportion. In the current context, this means that our 95% confidence interval from polling data may or may not contain the true proportion of Obama supporters in Florida but if we were to conduct polling in the exactly same manner, over this hypothetical repeated sampling, 95% of such confidence intervals will contain the true proportion of Obama supporters.

To formally construct such confidence intervals, we use the result about z-score given in equation (13) and the definition of convergence in distribution.

$$\lim_{n\longrightarrow\infty} \Pr\left(-z \;\leq\; \frac{\overline{X}_n - \mu}{\text{s.e.}} \;\leq\; z\right) \;=\; \Pr(-z \leq Z \leq z) \tag{14}$$

where Z is the standard normal random variable and z is a pre-determined quantity called *critical value*. Now, for 95% confidence intervals, choose $z_{\alpha/2} \approx 1.96$ such that $\Pr(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = \Phi(z_{\alpha/2}) = 1 - \alpha = 0.95$ where $\alpha = 0.05$. By rearranging the left hand side of equation (14), we obtain,

$$\begin{aligned}
\Pr\left(-z_{\alpha/2} \;\leq\; \frac{\overline{X}_n - \mu}{\text{s.e.}} \;\leq\; z_{\alpha/2}\right) \;&=\; \Pr\left(-z_{\alpha/2} \times \text{s.e.} \;\leq\; \overline{X}_n - \mu \;\leq\; z_{\alpha/2} \times \text{s.e.}\right) &(15)\\
&=\; \Pr\left(\overline{X}_n - z_{\alpha/2} \times \text{s.e.} \;\leq\; \mu \;\leq\; \overline{X}_n + z_{\alpha/2} \times \text{s.e.}\right) &(16)
\end{aligned}$$

where the second line is obtained by subtracting $\overline{X}_n$ from each of the three terms and then multiplying them by negative one (and therefore switching the direction of inequalities). Thus, in

a large sample, over repeated sampling, the confidence intervals, $[(\overline{X}_n - z_{\alpha/2} \times \text{s.e.},\ \overline{X}_n + z_{\alpha/2} \times \text{s.e.}]$, contains the true value $100 \times (1-\alpha)\%$ of time.

To sum up, the model-based approach allows us to derive the mean and variance of the estimator by using all available rules of probability theory. In addition, two powerful asymptotic theorems, i.e., law of large numbers and central limit theorem, enables large sample approximation with a large class of data generating processes. This asymptotic approximation tells us the sampling distribution of our estimator, which in turn lets us characterize the estimation uncertainty with confidence intervals.