Sampling

PSE 6.1-6.3

This material comes from Hansen's *Probability and Statistics for Economists* (PSE) and Len Goff's lecture notes along with some of my own comments.

This set of notes begins our discussion about statistics.

We will start by putting some structure on what type of data, or **sample**, we have access to. There are alternative sampling schemese, but we'll start with the simplest (and arguably most common)

Definition. A collection of random vectors $\{X_1, \ldots, X_n\}$ is a **(simple) random sample** from the population F if X_i are **independent and identically distributed (i.i.d)**

Here, X_i is a particular observation and there are n of them total. Sometimes I may use notation like $\{X_i\}_{i=1}^n$. The key condition is independent and identically distributed. Independent means that X_i is independent from X_j for $i \neq j$. Identically distributed means that all of the X_i are draws from the same distribution F. The distribution F is called the **population** or **population distribution**. The book discusses two different "metaphors" to think about the population that are worth mentioning. One is that there is an actual large population with N observations where N is much larger than n (the sample size) and where random sampling amounts to drawing a subset of n of these observations each with equal probability. The other metaphor is of a **data generating process (DGP)** where there is a process by which each observation is created, and the population is the probability model which generates the observations.

It is reasonable to think of much data as being a random sample. For example, it is likely reasonable to think of large datasets like the Current Population Survey or American Community Survey as being (at least approximately) random samples. In these data, if X_i is a person's income, it is reasonable to think that individual i's income is independent from individual j's and that they are drawn from the same distribution.

It is helpful to give a couple of examples that are not random samples.

- *Stratified random sampling*: The population is divided into groups, and then simple random sampling occurs within each group (e.g. I run my sampling algorithm separately for men and women, so that I can ensure equal representation of each). This is a common sampling procedure and many samples include "sampling weights" to "adjust back" to the overall population from a stratified random sample.
- *Clustered Sampling* After defining groups, we randomly select some of the groups. Then all individuals from those groups are included in the sample. An example of clustered sampling would be the case where a researcher is interested in studying the effect of some policy on student's test scores, and, instead of randomly sampling students, the researcher randomly

Sample **X**

row **i**	$\omega_{\mathbf{i}}$	$\mathbf{age}_i$	$\mathbf{married}_i$	$\mathbf{college}_i$
1	1	25	0	0
2	4	37	1	1
3	5	54	0	1

Population I

individual **i**	$\mathbf{age}_i$	$\mathbf{married}_i$	$\mathbf{college}_i$
1	25	0	0
2	74	1	1
3	8	0	0
4	37	1	1
5	54	0	1

Figure 1: An example of simple random sampling, in which $n = 3$ and $N = 5$. Each row of the dataset on the left is a realization of random vector $X = (age, married, college)$, which chooses a row at random from the population matrix on the right. We can conceptualize this sampling process as a probability space with outcomes $\omega = (\omega_1, \omega_2, \omega_3)$, where ω_i yields the index of the randomly selected individual in I. The random vectors $X_i = X_i(\omega_i)$ and $X_j = X_j(\omega_j)$ are independent for $i \neq j$, but the random variables within a row are generally not independent, e.g. age_i and $college_i$ are positively correlated.

selects classrooms and collects data about students within those classrooms. This sort of sample likely violates the independence condition of random sampling as test scores for students within the same classroom are likely to be correlated due to having the same teacher.

- *Time Series* Examples of time series data are a country's GDP over time or a company's stock price over time. Time series data is often serially correlated (if a country is in a recession this quarter, they are more likely than usual to be in a recession next quarter); this violates independence. Many time series have trends (e.g., the GDP of the United States has trended up over time). Thus, the distribution of GDP was different in, say, 2020 relative to, say 1970. This violates identically distributed.

The main goal of statistics is to conduct **inference** – that is, to learn something about the underlying population using the sample that we have access to.

Statistics, Parameters, and Estimators

PSE 6.4-6.6

Next, we'll introduce some additional important concepts. First, as discussed above, the main goal of statistics is to learn about features of the population. These features are called **parameters**.

Definition. A **parameter** θ is any function of the population F.

θ is a generic notation for a parameter, but it is common to use other Greek letters such as μ or β to represent population parameters or even to just use the population quantity itself. A simple example of a parameter is $\mu = \mathbb{E}[X]$.

Definition. A **statistic** is a function of the sample $\{X_i\}_{i=1}^n$.

One thing that is worth pointing out is that a parameter is non-random while a statistic is random. This second part is not immediately obvious. The textbook has a good discussion of this point which I paste here

> Recall that there is a distinction between random variables and their realizations. Similarly there is a distinction between a statistic as a function of a random sample – and is therefore a random variable as well – and a statistic as a function of the realized sample, which is a realized value. When we treat a statistic as random we are viewing it as a function of a sample of random variables. When we treat it as a realized value we are viewing it as a function of a set of realized values. One way of viewing the distinction is to think of "before viewing the data" and "after viewing the data." When we think about a statistic "before viewing" we do not know what value it will take. From our vantage point it is unknown and random. After viewing the data and specifically after computing and viewing the statistic the latter is a specific number is therefore a realization. It is what it is and it is not changing. The randomness is the process by which the data was generated – and the understanding that if this process were repeated the sample would be different and the specific realization would be therefore different.

Definition. An **estimator** $\hat{\theta}$ for a parameter θ is a statistic intended as a guess about θ.

As in the above definition, we will typically put a "hat" (that is, $\hat{}$) to indicate the estimator of the parameter θ. Once we have a specific sample (so that we can $\hat{\theta}$ is a particular number that we can calculate), we will refer to $\hat{\theta}$ as an **estimate** of θ. This distinction is similar to the above discussion of thinking of a statistic as being random but also being a number that can be calculated given for a particular sample.

A natural way to construct estimators is to use the **analogy principle**; that is, if we can write θ as a function of the population F, to have $\hat{\theta}$ be the same function of the sample. The previous sentence may sound complicated, but this is actually a very simple/straightforward idea.

The most common parameter that we'll estimate is the population mean of a random variable, $\mu = \mathbb{E}[X]$; we'll see that a large number of more complicated/interesting parameters are equal to functions of population means. The idea of the analogy principle is to estimate the population mean using the sample average:

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

The sample average is random in the sense that it would be different across different random samples from the same underlying population.

Other random variables can be written as the expected value of a transformation of X. For example, the second moment $\mathbb{E}[X^2]$, or, more generally, $\theta = \mathbb{E}[g(X)]$. The analog estimator of θ is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Again, all we are doing here is estimating the population parameter by replacing the population mean with the sample average.

Functions of Parameters

PSE 6.7

More generally, a large number of parameters can be written as transformations of population moments. We can write these generally as

$$\beta = h\Big(\mathbb{E}[g(X)]\Big)$$

where h and g are functions. Taking $\theta = \mathbb{E}[g(X)]$, so that $\beta = h(\theta)$, the natural way to estimate β is by

$$\hat{\beta} = h\Big(\hat{\theta}\Big) = h\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right)$$

This type of estimator is called a **plug-in estimator** because we plug in $\hat{\theta}$ in place of θ in order to estimate β.

An example comes from estimating $\sigma^2 = \text{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Replacing population expectations with sample averages and then applying the same function to them, we obtain plug-in estimator for σ^2 which is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

Sampling Distribution

PSE 6.8

Because statistics are random variables, they have a distribution. The distribution of a statistic is called its **sampling distribution**. We will be interested in the accuracy of our estimates which, in turn, depends on the sampling distribution of our estimator. The sampling distribution depends on the population distribution F, the sample size n, and the mapping from the data to the statistic.

Estimation Bias

PSE 6.9

An important feature of the sampling distribution of $\hat{\theta}$ is its mean, $\mathbb{E}[\hat{\theta}]$.

Definition. The **bias** of an estimator $\hat{\theta}$ of a parameter θ is

$$\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

This is the mean difference (over repeated samples) between $\hat{\theta}$ and θ. An estimator is said to be **unbiased** if $\text{bias}(\hat{\theta}) = 0$.

Let us return to estimating $\mu = \mathbb{E}[X]$ using the sample average $\bar{X}$. Notice that

$$\begin{aligned}
\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu
\end{aligned}$$

where the third equality holds because the X_i are identically distributed. This shows that $\bar{X}$ is unbiased for $\mathbb{E}[X]$.

Example: One alternative estimator for $\mathbb{E}[X]$ is X_1 (i.e., just use the first observation int he data). Is this estimator unbiased for $\mathbb{E}[X]$? Another estimator is $c\bar{X}$ for some constant c. Is this estimator unbiased for $\mathbb{E}[X]$?

A useful discussion from the textbook about the importance of unbiasedness as a property of an estimator:

> This last example shows that "reasonable" and "unbiased" are not necessarily the same thing. While unbiasedness seems like a useful property for an estimator, it is only one of many desirable properties. In many cases it may be okay for an estimator to possess some bias if it has other good compensating properties.

Sampling variance

PSE 6.10

The next feature of the sampling distribution for us to consider is the **sampling variance**, that is, the variance of the estimator $\hat{\theta}$ across repeated samples. For notation, we will typically just write $\text{var}(\hat{\theta})$ for the sampling variance of $\hat{\theta}$.

Let us calculate the sampling variance of $\bar{X}$.

$$\begin{aligned}\text{var}(\bar{X}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

where the second equality holds because $1/n$ is a constant and can come outside the variance after squaring it, the third equality crucially uses that the X_i are independent (otherwise, there would be a large number of additional covariance terms to deal with), and the last equality holds because $\text{var}(X_i) = \sigma^2$ which uses that X_i are identically distributed.

It is interesting to compare the above calculation of $\text{var}(\bar{X})$ to our previous calculation of $\mathbb{E}[\bar{X}]$. Notice that, similar to the expection, the $\text{var}(\bar{X})$ depends on $\text{var}(X)$ – this should make sense: the variance of the sample average depends on the variance of the random variable itself (and is smaller when the the variance of X itself is smaller, etc.). More interestingly, and distinct from the expectation, is that $\text{var}(\bar{X})$ depends on the sample size n. For larger values of n, the sampling variance becomes smaller. To get some intuition along these lines, think about rolling a die n times and calculating $\bar{X}$ over and over. When $n = 1$, the distribtion of $\bar{X}$ is the same as X, and, for example, it is not very uncommon for $\bar{X}$ to be equal to extreme values like 1 or 6 (the probability that $\bar{X} = 6$ is $1/6$ for example). Now consider the case where $n = 2$, in order for $\bar{X} = 6$, you would need to roll two 6's in a row, this happens with probability $1/2$, but values of $\bar{X}$ close to the population mean become more common. As you keep increasing n, it starts to become exceedingly unlikely that $\bar{X} = 6$ (and similar arguments apply to other "large" values).

All else equal, one would prefer an estimator with low sampling variance.

Mean Squared Error

PSE 6.11

Definition. The **mean squared error** of an estimator $\hat{\theta}$ for θ is

$$\text{mse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

The mean squared error is the mean distance between $\hat{\theta}$ and θ over repeated samples.

Proposition:

$$\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{var}(\hat{\theta})$$

This says that the mean squared error of an estimator is fully determined by its bias and variance, and gives an explicit formula.

Proof:

$$\begin{aligned}
\text{mse}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}\Big[\Big((\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta}] - \theta)\Big)^2\Big] \\
&= \underbrace{\mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\Big]}_{A} + \underbrace{\mathbb{E}\Big[(\mathbb{E}[\hat{\theta}] - \theta)^2\Big]}_{B} + 2\underbrace{\mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\Big]}_{C}
\end{aligned}$$

Let's consider each of these three terms individually. For the first one, it immediately holds that

$$A = \text{var}(\hat{\theta})$$

from the definition of variance. Next,

$$\begin{aligned}
B &= (\mathbb{E}[\hat{\theta}] - \theta)^2 \\
&= \text{bias}(\hat{\theta})^2
\end{aligned}$$

where the first equality holds since $\mathbb{E}[\hat{\theta}]$ and θ are both nonrandom (so the expected value of their difference squared is also non-random) implying that we can get rid of the outside expectation, and the second equality holds by the definition of $\text{bias}(\hat{\theta})$. Finally,

$$\begin{aligned}
C &= (\mathbb{E}[\hat{\theta}] - \theta) \\
\underbrace{\mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])\Big]}_{=0} &= 0
\end{aligned}$$

where the first equality holds because $\mathbb{E}[\hat{\theta}]$ and θ are non-random. Combining the expressions for Terms A, B, and C completes the proof.

Finally, notice that when an estimator is unbiased, then its mean squared error is equal to its variance.

Best Linear Unbiased Estimator

PSE 6.12

Next, we want to consider what is the "best" linear unbiased estimator (BLUE) of $\mu = \mathbb{E}[X]$. What we mean here is that we are looking for the lowest variance ("best") among linear estimators (in the sense of being a linear function of the X_i) that are unbiased. An estimator with these properties is also referred to as being **efficient**.

To be more precise about the class of linear estimators that we are considering, we will write them generically as

$$\tilde{\mu} = \sum_{i=1}^{n} w_i X_i$$

where w_i are non-random weights. An implication of unbiasedness is that, we must have that

$$\mathbb{E}[\tilde{\mu}] = \mu$$

This implies certain properties of the weights; in particular, notice that

$$\begin{aligned}
\mathbb{E}[\tilde{\mu}] &= \mathbb{E}\left[\sum_{i=1}^{n} w_i X_i\right] \\
&= \sum_{i=1}^{n} w_i \mathbb{E}[X_i] \\
&= \sum_{i=1}^{n} w_i \mu \\
&= \mu \sum_{i=1}^{n} w_i
\end{aligned}$$

Thus, an implication of unbiasedness is that

$$\sum_{i=1}^{n} w_i$$

Now, let's consider the variance of $\tilde{\mu}$.

$$\begin{aligned}
\text{var}(\tilde{\mu}) &= \text{var}\left(\sum_{i=1}^{n} w_i X_i\right) \\
&= \sum_{i=1}^{n} w_i^2 \text{var}(X_i) \\
&= \sigma^2 \sum_{i=1}^{n} w_i^2
\end{aligned}$$

Since $\tilde{\mu}$ is unbiased, we are interested in choosing weights that minimize $\text{var}(\tilde{\mu})$. This amounts to minimizing the variance subject to the unbiasedness constraint that the weights sum to 1. We can solve this problem using constrained optimization techniques. In particular, we can minimize the

Lagrangian

$$L(w_1, \ldots, w_n) = \sigma^2 \sum_{i=1}^{n} w_i^2 - \lambda \left(\sum_{i=1}^{n} w_i - 1 \right)$$

The first order condition of this problem with respect to w_i is

$$2\sigma^2 w_i - \lambda = 0$$

which we can re-write as

$$w_i = \frac{\lambda}{2\sigma^2}$$

Notice that this condition implies that the weights all must be equal to each other. In order to satisfy that they sum to 1, it therefore must be that $w_i = 1/n$. This implies that the best linear unbiased estimator of $\mathbb{E}[X]$ is given by

$$\sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X}$$

In other words, the sample average is BLUE.

Estimation of Variance

6.13

Standard Error

6.14

Multivariate Means

6.15

Monte Carlo Simulations