1 Motivation: Random samples

Suppose that you are a pollster, and like in Lectures 10 and 11, you would like to predict the turnout of American citizens at an election. As we have seen, you cannot afford to survey the entire population, and instead contact only n citizens, with n fairly small as compared to the total number of people allowed to vote.

You then model the response of citizen i as a random variable X_i. The n random variables X_1, X_2,...,X_n thus constructed are called a **random sample**.

The corresponding observed values of the random sample (i.e. the answers you got on the phone after contacting the n people) is a *univariate* data set written with lowercase letters: x_1, x_2,...,x_n.

In this class, we will focus on the special case in which the X_i can be reasonably assumed to be mutually independent, and have the same probability distribution. If F is that probability distribution, we say that X_1, X_2,...,X_n is a **random sample from** F.

In this lecture, we will learn how to use a data set $x_1, x_2, \ldots, x_n$ to estimate a feature of F we are interested in. For our polling example, we are interested in $E[X_i]$, as you remember.

2 Estimators

2.1 Estimate

Let θ be the feature of the distribution F we are interested in. θ is called the **parameter of interest**.

We want to estimate θ with our data set $x_1, x_2, \ldots, x_n$. This motivates the following formal definition of an estimate in statistics.

Definition: An **estimate** is a value t which **only depends on the data set** x_1, x_2,...,x_n:

$$t = h(x_1, x_2, \ldots, x_n)$$

where h is some function from $\mathbb{R}^n$ to $\mathbb{R}$.

As an illustration, let us return to our polling example. The law of large numbers suggests that

$$\overline{x_n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad \text{Sample mean}$$

would provide a good estimate for $\mu = E[X_i]$. So in this case,

$$h(x_1, x_2, \ldots, x_n) = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

is an estimate.

Now, observe that there are many possible ways of constructing an estimate. For example, for our polling example,

$$h(x_1, x_2, \ldots, x_n) = x_n$$

is also an estimate (a bad one!).

Likewise,

$$h(x_1, x_2, \ldots, x_n) = \frac{x_{n-2} + x_{n-1} + x_n}{3}$$

is also an estimate (slightly better than the previous one).

A natural question thus is: *how to choose an estimate, and is there an optimal one?*

To answer this question, we need to study the properties of the random variables $h(X_1, X_2, \ldots, X_n)$, which are called estimators.

2.2 Estimators

Definition: Let $t = h(x_1, x_2, \ldots, x_n)$ be an estimate based on the dataset $x_1, x_2, \ldots, x_n$. Then t is a realization of the random variable

$$T = h(X_1, X_2, \ldots, X_n)$$

The random variable T is called an **estimator**. The probability distribution of T is called the **sampling distribution of** T.

As an illustration, let us return once more to our polling example. Let us take the case

$$T = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{S_n}{n}$$

We know that S_n is a binomial distribution with parameters n and μ. Hence, T takes the values $\frac{k}{n}$, with $k = 0, 1, \ldots, n$, according to

$$p_T\left(\frac{k}{n}\right) = P(T = \frac{k}{n}) = P(\frac{S_n}{n} = \frac{k}{n}) = P(S_n = k) = \binom{n}{k}\mu^k(1-\mu)^{n-k}$$

This is the sampling distribution of T.

2.3 Unbiasedness

If an estimator T is chosen reasonably, one expects its outcomes t to fluctuate about the value of the parameter of interest θ. A desirable feature of an estimator is to have these fluctuations in such a way that $E[T] = \theta$.

As an example, the estimator T we just saw has that desirable property:

$$E[T] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n E[X_1] = \mu$$

In other words, T has no *systematic* tendency to produce estimates t which are larger than μ, nor to produce estimates which are smaller than μ. T can be said to be *unbiased*.

Definition: An estimator T is called an **unbiased estimator** for the parameter of interest θ if

$$E[T] = \theta \qquad \text{Unbiased estimator } T \tag{1}$$

For a *general* estimator T for the parameter of interest θ, the **difference** $E[T] - \theta$ **is called the bias of** T. If this difference is **not equal to zero,** T **is said to be a biased estimator**.

We have seen that

$$T = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is an unbiased estimator.

What about

$$\tilde{T} = \frac{X_1 + X_2 + \ldots + X_n}{n+1} \quad ?$$

We have

$$E[\tilde{T}] = \frac{n}{n+1}\mu$$

We see that for all $n \in \mathbb{N}^*$,

$$E[\tilde{T}] < \mu$$

so $\tilde{T}$ has *negative bias*. However, this bias goes to zero as $n \to +\infty$.

2.4 Unbiased estimator for expectation and variance

In this subsection, we focus on the particular cases when the parameter of interest is the expected value μ or the variance σ^2 of the distribution.

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite expectation μ and finite variance σ^2. Then

$$\boxed{\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}} \tag{2}$$

is an **unbiased estimator for** μ, and

$$\boxed{S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2} \tag{3}$$

is an **unbiased estimator for** σ^2.

In the previous page, we gave a simple proof of the first result, which is intuitive. Let us now prove the second result, which is less intuitive, because of the presence of the $1/(n-1)$ factor, as we already briefly discussed in Lecture 12.

$$\begin{aligned}
E[S_n^2] &= \frac{1}{n-1}\sum_{i=1}^{n} E[(X_i - \overline{X_n})^2] \\
&\underset{E[X_i]=E[\overline{X_n}]}{=} \frac{1}{n-1}\sum_{i=1}^{n}\left[E[(X_i - \overline{X_n})^2] - (E[X_i] - E[\overline{X_n}])^2)\right] \\
&\underset{\text{Lin. of exp.}}{=} \frac{1}{n-1}\sum_{i=1}^{n}\left[E[(X_i - \overline{X_n})^2] - (E[X_i - \overline{X_n}])^2)\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{Var}(X_i - \overline{X_n})
\end{aligned}$$

Now,

$$X_i - \overline{X_n} = X_i - \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{n-1}{n}X_i - \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_j$$

We observe that the two different terms in the last expression above can be interpreted as two different random variables, $(n-1)/nX_i$ and $1/n\sum_{\substack{j=1\\ j\neq i}}^{n} X_j$, which are independent random variables since the X_i's are mutually independent. Therefore, we can write

$$\begin{aligned}
\mathrm{Var}(X_i - \overline{X_n}) &= \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(X_i) + \frac{1}{n^2}\mathrm{Var}(\underbrace{\sum_{\substack{j=1\\ j\neq i}}^{n} X_j}_{\text{all independent}}) \\
&= \left(\frac{n-1}{n}\right)^2 \sigma^2 + \frac{1}{n^2}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathrm{Var}(X_j) \\
&= \frac{n^2 - 2n + 1 + n - 1}{n^2}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

Thus,

$$\underline{E[S_n^2] = \frac{1}{n-1}\sum_{i=1}^{n}\frac{n-1}{n}\sigma^2 = \sigma^2}$$

With this derivation, we now understand why the unbiased estimator for σ^2 has a factor $1/(n-1)$ in front of the sum, and not $1/n$ as we may have naïvely expected. A factor of $1/n$ would lead to an estimator with negative bias. This also explains why in Lecture 12 we constructed the sample variance as $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2$.

A cautionary tale about unbiasedness: unbiasedness is not propagated in general.

Indeed, let T be an unbiased estimator for a parameter of interest θ. Then, if $g : \mathbb{R} \to \mathbb{R}$ is a function, then $g(T)$ is in general **not** an unbiased estimator for the parameter of interest $g(\theta)$.

The typical example for this fact concerns the standard deviation. We have just seen that S_n^2 is an unbiased estimator for σ^2: $E[S_n^2] = \sigma^2$. Is S_n an unbiased estimator for the standard deviation σ?

By Jensen's inequality, which we saw in Lecture 6,

$$(E[S_n])^2 \leq E[S_n^2] \quad \Leftrightarrow \quad (E[S_n])^2 \leq \sigma^2 \quad \Leftrightarrow \quad E[S_n] \leq \sigma$$

S_n is an estimator of the standard deviation with *negative bias*.

2.5 Variance of an estimator

At the start of this lecture, we saw the following estimators for our voter turnout:

$$T_1 = \frac{X_1 + X_2 + \ldots + X_n}{n} \qquad , \qquad T_2 = X_n$$

Both estimators are *unbiased* since $E[T_1] = E[T_2] = \mu$.

However, from the law of large numbers, it is clear that T_1 is more desirable than T_2, in the sense that T_1 produces estimates that are more concentrated around the parameter of interest θ than T_2. Not all unbiased estimators are created equal! This is motivates the following definition.

Definition (Efficiency of estimators): Let T_1 and T_2 be two **unbiased estimators** for the same parameter of interest θ. Then estimator T_1 is called **more efficient** than estimator T_2 if $\mathrm{Var}(T_1) < \mathrm{Var}(T_2)$.

For the situation we have chosen to motivate this definition, we have

$$\mathrm{Var}(T_1) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n} \qquad , \qquad \mathrm{Var}(T_2) = \sigma^2$$

Thus, for $n > 1$, T_1 is more efficient than T_2.

The **relative efficiency** of T_1 with respect to T_2 is defined as

$$\boxed{\frac{\mathrm{Var}(T_2)}{\mathrm{Var}(T_1)} \qquad \text{Relative efficiency}} \tag{4}$$

In our example, we find that the relative efficiency is

$$\frac{\mathrm{Var}(T_2)}{\mathrm{Var}(T_1)} = \frac{\sigma^2}{\frac{\sigma^2}{n}} = n$$

2.6 Mean squared error

We just considered a measure that can be used to compare the desirability of two unbiased estimators. While it is often convenient to deal with unbiased estimators, biased estimators can sometimes be favored, if they are easier to define and calculate, for example, and if the bias can be shown to not be large.

To compare two estimators, whether they are unbiased or not, the intuition remains the same: the most desirable estimator is the one whose spread about the parameter of interest θ is the smallest. This leads to the following definition regarding the relevant measure of desirability of an estimator.

Definition: Let T be an estimator for a parameter of interest θ. The **mean squared error (MSE)** of T is the quantity

$$\boxed{\text{MSE}(T) = E[(T-\theta)^2]} \tag{5}$$

We say that **an estimator T_1 performs better than an estimator T_2 in the mean squared sense** if $\text{MSE}(T_1) < \text{MSE}(T_2)$.

Link with previous measure of performance and efficiency

Let us write $\text{MSE}(T)$ in a different way.

$$\begin{aligned}\text{MSE}(T) &= E[(T-\theta)^2] = E[T^2 - 2\theta T + \theta^2] \\ &= \text{Var}(T) + (E[T])^2 - 2\theta E[T] + \theta^2\end{aligned}$$

Hence,

$$\boxed{\text{MSE}(T) = \text{Var}(T) + (E[T]-\theta)^2} \tag{6}$$

In the previous subsection, we considered estimates T_1 and T_2 which were unbiased: $E[T_1] = \theta = E[T_2]$. In that case,

$$MSE(T_1) = \text{Var}(T_1) \qquad , \qquad MSE(T_2) = \text{Var}(T_2)$$

which means that our new measure of efficiency of an estimator, based on the MSE, extends the previous measure, and agrees with it for the case of unbiased estimators.

The formula $MSE(T) = \text{Var}(T) + (E[T]-\theta)^2$ highlights the two sources of spread with respedct to the parameter of interest θ: the variance of the estimator $\text{Var}(T)$, and the squared bias, $(E[T]-\theta)^2$, which is nonzero for biased estimators.

We now understand why a biased estimator with small variance can be preferred to an unbiased estimator with large variance.

Example: Let X_1, X_2, ...,X_n be independent and identically distributed normal random variables with distribution $N(\mu, \sigma^2)$.

- As we have seen in subsection 4,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

  is an unbiased estimator of σ^2:

$$\text{MSE}(S_n^2) = \text{Var}(S_n^2)$$

  It can be shown that $\text{Var}(S_n^2) = \frac{2\sigma^4}{n-1}$, so

$$\text{MSE}(S_n^2) = \frac{2\sigma^4}{n-1}$$

- An alternative estimator for σ^2 is

$$\hat{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{n-1}{n}S_n^2$$

  We have

$$E[\hat{S}_n^{\,2}] = \frac{n-1}{n}E[S_n^2] = \frac{n-1}{n}\sigma^2$$

  so $\hat{S}_n^2$ is an estimator of σ^2 with negative bias.

Now,

$$\mathrm{Var}(\hat{S}_n^2) = \frac{(n-1)^2}{n^2}\mathrm{Var}(S_n^2) = \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2(n-1)}{n^2}\sigma^4$$

Therefore,

$$\mathrm{MSE}(\hat{S}_n^2) = \frac{2(n-1)}{n^2}\sigma^4 + (E[\hat{S}_n^2] - \sigma^2)^2 = \frac{2(n-1)}{n^2}\sigma^4 + \frac{\sigma^4}{n^2} = \frac{2n-1}{n^2}\sigma^4$$

We thus have the following inequalities:

$$\mathrm{MSE}(\hat{S}_n^2) = \frac{2n-1}{n^2}\sigma^4 < \frac{2}{n}\sigma^4 < \frac{2\sigma^4}{n-1} = \mathrm{MSE}(S_n^2)$$

The smaller variance of $\hat{S}_n^2$ leads to a smaller MSE, despite the negative bias.

This means that on average, $\hat{S}_n^2$ will be closer to σ^2 than S_n^2. However, $\hat{S}_n^2$ will underestimate σ^2 on average, unlike S_n^2.