Lecture 22: Point Estimation

Today we start Chapter 6 and with it the statistics port of the course. We saw in Lecture 20 (Random Samples) that it frequently occurs that we know a probability distribution except for the value of a parameter.
In fact we had three examples

1. *The Election Example*

$$\boxed{\text{Bin}\ (1,\ ?)}$$

2. *The Computer Failure Time Example*

$$\boxed{\text{Exp }(?)}$$

3. *The Random Number Example*

$$\boxed{\text{U}(0,\ ?)}$$

By convention the unknown parameter will be denoted θ. So replace ? by θ in the three examples. So $\theta = p$ in example 1 and $\theta = \lambda$ in Example 2 and $\theta = B$ (so $U(0, B)$) in Example 3.

If the population X is discrete we will write its pmf as $p_X(x, \theta)$ to emphasize that it depends on the unknown parameter θ and if X is continuous we will write its pdf as $f_X(x, \theta)$ again to emphasize the dependence on θ.

Important Remark

θ is a fixed number, it is just that we don't know it. But we are allowed to make calculations with a number we don't know, that is the first thing we learn to do in high-school algebra, compute with "the unknown x".

Now suppose we have on actual sample $x_1, x_2, \ldots, x_n$ from a population X whose probability distribution is known except for an unknown parameter θ. For convenience we will assume X is discrete.

$$\boxed{p_X(x, \theta)} \longrightarrow x_1, x_2, \ldots, x_n$$

The idea of point estimation is to develop a theory of making a guess for θ ("estimating θ") in terms of $x_1, x_2, \ldots, x_n$.
So the big problem is

The Main Problem (Vague Version)

What function $h(x_1, x_2, \dots, x_n)$ of the items $x_1, x_2, \dots, x_n$ in the sample should we pick to estimate θ ?

Definition

Any function $w = h(x_1, x_2, \dots, x_n)$ we choose to estimate θ will be called an estimator for θ.
As first one might ask -

$$\left.\begin{array}{c}\text{find } h \text{ so that for every sample} \\ x_1, x_2, \dots, x_n \text{ we have} \\ h(x_1, x_2, \dots, x_n) = \theta.\end{array}\right\} \qquad (*)$$

This is hopelessly naive. Let's try something else

The Main Problem (some what more precise)

Give quantitative criteria to decide whether one estimator $w_1 = h_1(x_1, x_2, \ldots, x_n)$ for θ is better than another estimator $w_2 = h_2(x_1, x_2, \ldots, x_n)$ for θ.
The above version, though better, is not precise enough.
In order to pose the problem correctly we need to consider random samples from X, in ofter words go back before an actual sample is taken or "go random".

$$\boxed{p_X(x,\theta)} \dashrightarrow X_1, X_2, \ldots, X_n$$

Now our *function h gives rise to a random variable (statistic)*

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

which I will call (for a while) an estimator *statistic*, to distinguish if from the estimator (*number*) $w = h(x_1, x_2, \ldots, x_n)$. Once we have chosen h the corresponding estimator statistic will ofter be denoted $\hat{\theta}$.

Main Problem (third version)

Find an estimator $h(x_1, x_2, \ldots, x_n)$ so that

$$P(h(X_1, X_2, \ldots, X_n) = \theta) \qquad (**)$$

is maximized
This is what we want but it is too hard to implement - after all we don't know θ.

Important Remark

We have made a huge gain by "going random". The statement "maximize $P(h(x_1, x_2, \ldots, x_n) = \theta)$" does not make sense because $h(x_1, x_2, \ldots, x_n)$ is a fixed real number so either it is equal to θ or it is not equal to θ. But $P(h(X_1, X_2, \ldots, X_n)) = \theta$ does make sense because $h(X_1, X_2, \ldots, X_n)$ is a random variable.

Now we weaken $(**)$ to something that can be achieved, in fact achieved surprisingly easily.

Unbiased Estimators Main Problem (fourth version)

Find an estimator $w = h(x_1, \ldots, x_n)$ so that the expected value $E(W)$ of the estimator statistic $W = h(X_1, X_2, \ldots, X_n)$ is equal to θ.

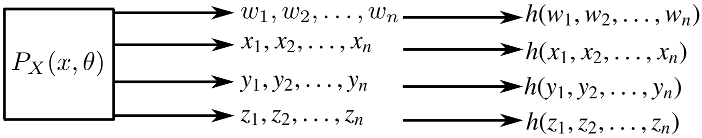
Definition

If an estimator W for an unknown parameter θ satisfies W satisfies $E(W) = \theta$ then the estimator W is said to be unbiased.

Intuitively, requiring $E(W) = \theta$ is a good idea but we can make this move precise. Various theorems in probability e.g Chebyshev's inequality, tell us that if Y is a random variable and $y_1, y_2, \ldots, y_n$ are observed values of Y then the numbers $y_1, y_2, \ldots, y_n$ will tend to be near $E(Y)$.

Applying this to our statistic W- if we take many samples of size n and compute the value of our estimator h on each one to obtain many observed values of W then the resulting numbers will be near $E(W)$. But we want these to be near θ. So we want

$$E(W) = \theta$$



I have run out of letters. In the above there are four samples of size n and four corresponding estimates $h(w_1, \ldots, w_n)$, $h(x_1, \ldots, x_n)$, $h(y_1, \ldots, y_n)$ and $h(z_1, \ldots, z_n)$ for θ.

Imagine that instead of four we have one hundred estimates of size n and one hundred estimates. Then if $E(W) = \theta$ most of these estimates will be close to θ.

Examples of Unbiased Estimators

Let's take another look at Problems 1 and 2 (pages 1 and 2)

For a Bernoulli random variable $X \sim \mathrm{Bin}(1, p)$ we have

$$E(X) = p.$$

Hence for the election example, **we are trying to estimate the mean in a Bernoulli distribution**.

For an exponential random variable $X \sim$ Exp (λ) we have

$$E(X) = \frac{1}{\lambda}.$$

Hence for the Dell computer failure time example , **we are trying to estimate the reciprocal of the mean in an exponential distribution**. One approach is to choose an estimator for the mean, compute it then takes its reciprocal. If we use this approach then the problem again amount estimating the mean.

So in both cases we are trying to estimate the population mean $E(X) = \mu$

However, in the second case we have to invert the estimate for μ to get an estimate for λ.

In fact many other estimation problems amount to estimating the mean in some probabiity distribution. Accordingly we state this as a general problem.

Problem

Find an unbiased estimator for the population mean μ

$$\boxed{\begin{array}{c}\theta = \mu \\ P_X(x,\theta)\end{array}} \dashrightarrow \quad X_1, X_2, \ldots, X_n$$

So we want $h(x_1, x_2, \ldots, x_n)$ *so that*

$$E\left(h\left(X_1, X_2, \ldots, X_n\right)\right) = \mu$$

= the population mean.

Amazingly there is a very simple solution to this problem no matter what the underlying distribution is

Theorem

The sample mean $\bar{X}$ is an unbiased estimator of the population mean μ; that is

$$E(\bar{X}) = \mu$$

Proof

The proof is so simple, deceptively simple because the theorem is so important.

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{X_1 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n}\left(E(X_1) + \ldots + E(X_n)\right) \end{aligned}$$

Proof (Cont.)

But $E(X_1) = E(X_2) = \ldots = E(X_n) = \mu$ because all the X_i's are samples from the population so they have the same distribution as the population so

$$\begin{aligned} E(\overline{X}) &= \frac{1}{n}\underbrace{(\mu + \mu + \ldots \mu)}_{n \text{ times}} \\ &= \frac{1}{n}(n\mu) \\ &= \mu \end{aligned}$$

$\square$

There is lots of other unbiased estimators of μ for any population. It is X_1, the first sample item (or any $X_i, 1 \leq i \leq n$). This is because, as noted above,

$$E(X_1) = E(X_i) = E(X) = \mu, 1 \leq i \leq n.$$

For the problem of estimating p in $\mathrm{Bin}(1, p)$ we have

$$\overline{x} = \frac{\text{number of observed successes}}{n}$$

Since each of $x_1, x_2, \ldots, x_n$ is either 1 on 0 so

$$x_1 + x_2 + \ldots + x_n = \# \text{ of } 1's.$$

is the number of "successes" (voters who say "Trump" in 2020 (I am joking)) so

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

is the the relative number of observed successes. This is the "common sense" estimator.

An Example Where the "Common Sense" Estimator is Biased

Once we have a *mathematical* criterion for an estimator to be good we will often find to our surprise that "common sense" estimators do not meet this criterion.
We saw an example of this in the "Pandemonium jet fighter" Section 6.1, problem 14,(on page 263).
Another very similar problem occurs in Example 3 - estimate B from the uniform distribution $U(0, B)$.

$$\boxed{U(0,B)} \longrightarrow x_1, x_2, \ldots, x_n,\ \theta = B$$

The “common sense” estimator for B is $w = \max(x_1, x_2, \ldots, x_n)$, the biggest number you observe. But it is intuitively clear that this estimate will be too small since it only gives the right answer if one of the x_i's is equal to B
So the common sense estimator $W = \max(x_1, x_2, \ldots, x_n)$ is biased.

$$E\left(\text{Max}\ (X_1, \ldots, X_n)\right) \underset{\neq}{<} B$$

Amazingly, if you do problem 32, page 274 *you will see exactly by how much if undershoots the mark*. We did this in class.

Theorem

$$E\left(Max(X_1, X_2, \ldots, X_n)\right) = \frac{n}{n+1}B$$

so $\left(\dfrac{n+1}{n}\right)$ *Max* $(X_1, X_2, \ldots, X_n)$ *is unbiased.*

Mathematics trumps common sense.

Minimum Variance Unbiased Estimators

We have seen that $\overline{X}$ and X_1 are both unbiased estimators of the population mean for any distribution. Common sense tells us that $\overline{X}$ is better since it uses all the elements of the sample whereas X_1 just uses one element of the sample (the first).

What mathematical criterion separates them. We have

$$V(X_1) = \sigma^2 = \text{ the population variance}$$

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

so if n is large then

$$V(\overline{X}) \text{ is a lot smaller than } V(X_1).$$

We will are now going to see why small variance is good. First we state this as a general principle.

The Principle of Minimum Variance Unbiased Estimation

Among all estimators of θ that are unbiased, choose one that has minimum variance.

The resulting estimator is called a minimum variance unbiased estimator, MVUB.

Theorem 1

$\overline{X}$ is a minimum variance unbiased estimator for the problems of

1. *Estimating p in Bin $(1, p)$*
2. *Estimating μ in $N(\mu, \sigma^2)$*

Why is it good to minimize the variance?
We will now see why, assuming the estimator $\widehat{\theta}$ is unbiased.

Suppose $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ is an estimator statistic for an unknown parameter θ.

Definition

The mean squared error $MSE(\hat{\theta})$ of the estimator $\hat{\theta}$ is defined by

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

so

$$MSE(\hat{\theta}) = \int \ldots \int_{\mathbb{R}^n} (h(x_1, \ldots, x_n) - \theta)^2 f_{X_1}(x_1) \ldots f_{X_n}(X_n) dx_1 dx_2, \ldots, d_{x_n}.$$

$$or \;= \sum_{all\; x_1, \ldots, x_n} (h(x_1, \ldots, x_n) - \theta)^2 \, P(X_1 = x_1) \ldots P(X_n = x_n)$$

So $MSE(\hat{\theta})$ is the square of the error $h(x_1, x_2, \ldots, x_n) - \theta)$ of the estimate of θ by $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ averaged over all $x_1, x_2, \ldots, x_n$.
Obviously we want to minimize the mean squared error (after all it does measure an error). Here is the point - **if $\hat{\theta}$ is unbiased this is the same minimizing the variance** $V(\hat{\theta})$. We now prove the last statement.

Theorem

If $\hat{\theta}$ is unbiased then

$$MSE(\hat{\theta}) = V(\hat{\theta})$$

This is amazingly easy to prove.

Proof.

By definition

$$MSE(th\hat{e}ta) = E\left((\hat{\theta} - \theta)^2\right).$$

But if $\hat{\theta}$ is unbiased then $E(\hat{\theta}) = \theta$ so

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - E(\theta)^2)\right)$$

By definition the RHS is $V(\hat{\theta})$. □

Here is on important definition used a lot in the text. I essentially copied the definition that is in the text, on page 259.

Definition (text page 259)

The standard error of the estimator $\hat{\theta}$, denoted $\sigma_{\hat{\theta}}$ is $\sqrt{V(\hat{\theta})}$. If the standard error itself contains unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields the **estimated standard error** *denoted $s_{\hat{\theta}}$*