Lecture 22

Point Estimation

Today we start Chapter 6 and with it the statistics part 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

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

2. The Computer Failure Time Example

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

3. The Random Number Example

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

By convention the unknown parameter will be denoted ? so replace ? by θ in the three examples.

so $\theta = p$ in Example 1 and $\theta = \lambda$ in Example 2 and $\theta = B$ (so $U(0,B)$) 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 point of [illegible] and "the unknown x".

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

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

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

So the big problem is

The Main Problem (Vague Version)

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

Definition

Any function $w = h(x_1, x_2, \ldots, x_n)$ we choose to estimate θ will be called an estimator for θ.

As first one might ask –

"find h so that for every sample $x_1, x_2, \ldots, x_n$ we have

$$h(x_1, x_2, \ldots, x_n) = \theta. \quad (*)$$

This is hopelessly naive. Let's try something else

The Main Problem (somewhat more precise)

Give quantitative criteria to decide whether one estimator $W = h_1(X_1, X_2, \ldots, X_n)$ for Θ is better than another estimator $h_2(X_1, X_2, \ldots, X_n)$.

The above version, though better, is still not useful. In order to pose the problem correctly we need to consider random samples from X, in other words go back before an actual sample is taken or "go random".

$$\boxed{P_X(x,\theta)} \dashrightarrow X_1, X_2, \dots, X_n$$

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

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

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

If we have an actual sample

$$\boxed{P_X(x,\theta)} \longrightarrow x_1, x_2, \dots, x_n$$

then the <u>number</u> $h(x_1, x_2, \dots, x_n)$ is called the observed value of the estimator statistic $W = h(X_1, X_2, \dots, X_n)$ on the sample $x_1, x_2, \dots, x_n$. Unfortunately it too is often denoted $\hat{\theta}$.

Remark

The estimator statistic should be denoted $\hat{\Theta}$ and its observed value $\hat{\theta}$ but mathematical (and statistical) [illegible] is far from consistent

Now we can formulate

Main Problem (third version)

Find an estimator $h(x_1, \ldots, x_n)$ 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 $P(h(x_1, \ldots, x_n) = \theta)$"

is stupid, idiotic, foolish, etc.....

because $h(x_1, \ldots, x_n)$ is a number and θ is a number so the above statement amounts to the hopelessly naive criterion from page 5

— choose h so that $h(x_1, x_2, \ldots, x_n) = \theta$.

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 satisfies

$$E(W) = \theta \qquad (***)$$

At first glance $(***)$ doesn't look much easier to achieve than $(**)$ but in fact it is surprisingly easy to achieve — in fact too easy. There are many W that satisfy $(**)$ so we will need further criteria.

Let's give estimator statistics that satisfy (***) a name.

Definition

An estimator statistic $W = h(X_1, X_2, \ldots, X_n)$ is an **unbiased** **estimator** of the population parameter θ if

$$E(W) = \theta.$$

Intuitively (***) is a good idea but we can make this more 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 W on each one to obtain many observed values of W then the resulting numbers will be near $E(W)$. But we want them to be near θ. So we want

$$E(W) = \theta$$

$$
\boxed{P_X(x;\theta)} \quad
\begin{array}{l}
\longrightarrow w_1, w_2, \ldots, w_n \longrightarrow h(w_1, w_2, \ldots, w_n) \\
\longrightarrow x_1, x_2, \ldots, x_n \longrightarrow h(x_1, x_2, \ldots, x_n) \\
\longrightarrow y_1, y_2, \ldots, y_n \longrightarrow h(y_1, y_2, \ldots, y_n) \\
\longrightarrow z_1, z_2, \ldots, z_n \longrightarrow h(z_1, z_2, \ldots, z_n)
\end{array}
$$

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 $h(z_1 \ldots z_n)$.

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 would be close to θ.

Examples of Unbiased Estimators

The most important estimation problem

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

$$\boxed{Bin(1,p)} \longrightarrow X_1, X_2, \ldots, X_n \quad (p = \theta)$$

$$\boxed{Exp(\lambda)} \longrightarrow X_1, X_2, \ldots, X_n \quad (\lambda = \theta)$$

Facts – for a Bernoulli random variable $X \sim Bin(1,p)$ we have $E(X) = p$

and

for an exponential random variable $X \sim Exp(\lambda)$

$$E(X) = \lambda.$$

So in both cases the unknown parameter is the population mean $E(X) = \mu$.

We have

Problem

Find an unbiased estimator for the population mean μ

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

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

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

$= $ the population mean

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

Theorem

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

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

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

$$E(\overline{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right)$$

$$= \frac{1}{n}\left(E(X_1) + \cdots + E(X_n) \right)$$

But $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$

because all the X_i's are samples from the population so they have the same distribution as the population so

$$E(\bar{X}) = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ times}})$$

$$= \frac{1}{n}(n\mu)$$

$$= \mu$$

$\square$

For the problem of estimating p in $Bin(1, p)$ we have

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

Since each of $x_1, x_2, \dots, x_n$ is either 1 or 0 so

$x_1 + x_2 + \dots + x_n = \#$ of 1's.

Thus $\overline{X}$ is the "common sense" estimating the relative number of observed successes.

An Example Where the "Common Sense" Estimator is Biased

Once we have a <u>mathematical</u> 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" problem 14, on page 242.

Another very similar problem occurs in Example 3 - estimating B in choosing a random number from $U(0,B)$.

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

The "common sense" estimator for B is $w = \max(x_1, x_2, \dots, x_n)$, 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

and $$P\left(\bigcup_{i=1}^{n}(X_i = B)\right) = \sum_{i=1}^{n} P(X_i = B)$$

$$= 0 + 0 + \cdots + 0 = 0$$

So the common sense estimator $W = \max(x_1, x_2, \ldots, x_n)$ is biased.

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

Amazingly, if you do page 252, problem 32 you will see exactly by how much it undershoots the mark

Theorem

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

So $\left(\frac{n+1}{n}\right) \mathrm{Max}(X_1, X_2, \ldots, X_n)$ is unbiased.

Mathematics trumps common sense

Minimum Variance Unbiased Estimators

We have seen that $\overline{X}$ and X_1 are unbiased estimators of the population mean. Common sense tells us that $\overline{X}$ is better. What mathematical criterion seperates them. We have

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

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

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

We will see later why this is good. First we state

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

$\overline{X}$ 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?

The following is treated incompletely on page 252 #34

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

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

so

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

or

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

So $MSE(\hat{\theta})$ is the squared error $(h(x_1, \ldots, x_n) - \theta)^2$ of the estimate of θ by $h(x_1, x_2, \ldots, x_n)$ averaged over all $x_1, x_2, \ldots, x_n$.

Obviously we want to minimize this squared error. Here is the point

<u>Theorem</u>

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

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

This is amazingly easy to prove

<u>Proof</u> If $\hat{\theta}$ is unbiased then

$E(\hat{\theta}) = \theta$ so

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

By by definition the RHS is

$V(\hat{\theta})$ $\square$

Here is an important definition used a lot in the text.

<u>Definition</u> (text page 238)

The standard error of the estimator $\hat{\theta}$, denoted $\sigma_{\hat{\theta}}$ is $\sqrt{V(\hat{\theta})}$.

It is often denoted $s_{\hat{\theta}}$ (not quite true) see page 238