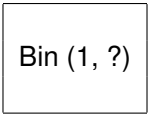
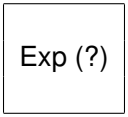
Lecture 22: Point Estimation

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

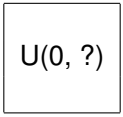
1. *The Election Example*



2. *The Computer Failure Time Example*



3. *The Random Number Example*



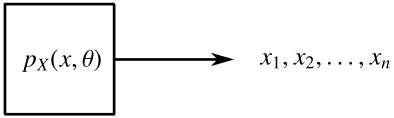
By convention the unknown parameter will be denoted ? So replace ? by θ in the three examples. So $\theta = p$ in example 1 and $\theta = \lambda$ in Example 2 and $\theta = B$ (so $U(0, B)$) in Example 3.

If the population X is discrete we will write its proof 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 known and "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.



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

Definition

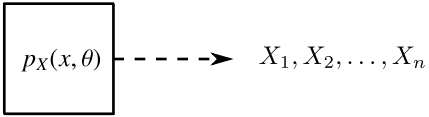
Any function $w = h(x_1, x_2, \ldots, x_n)$ we choose to estimate θ will he called an estimator for θ.
As first one might ask -

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

This is hopelessly naive. Let's try something else

The Main Problem (some what more precise)

Give quantitative criteria to decide whether one estimator $w = h_1(x_1, x_2, \ldots, x_n)$ 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 ofter words go back before an actual sample is taken or "go random".

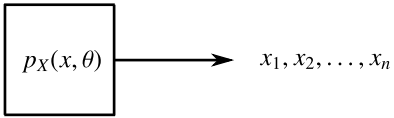


Now our *function h gives rise to a random variable (statistic)*

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

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 $\hat{\theta}$.

If we have an actual sample



then the *number* $h(x_1, x_2, \ldots, x_n)$ is called the observed value of the estimator statistic $W = h(X_1, X_2, \ldots, X_n)$ on the sample $x_1, x_2, \ldots, x_n$. Unfortunately if too is after denoted $\hat{\theta}$.

Remark

The estimator statistic should be denoted $\hat{\theta}$ and its observed value $\hat{\theta}$ but mathematical (and statistical) ?????? is for from consistent.

Main Problem (third version)

Find an estimator $h(x_1, x_n)$ 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 $P(h(x_1, x_n) = \theta)$" is stupid, idiotic, foolish...

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 surprising easily.

Unbiased Estimators Mam 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 then (**??**) but in fact it is surprising easy to achieve - in fact too easy. There are money W that satisfy $(***)$ so we will need further criteria.

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

Definition

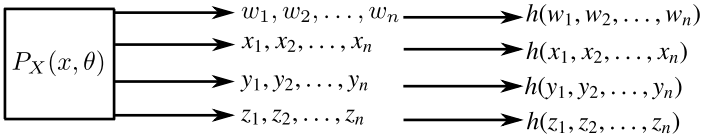
An estimator statistic
$W = h(X_1, X_2, \ldots, X_n)$ *is an unbiased estimator of the population parameter* θ *of*

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

Intuitively $(***)$ is a good idea but we can make this move precise Various theorems in probability e.g Cheston's requetity.

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 there to be near θ. So we want

$$E(W) = \theta$$



I have run out of letters. In the above there are four sample 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.

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

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

and

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

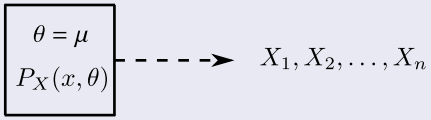
$$E(X) = \lambda$$

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



So we wont $h(x_1, x_2, \ldots, x_n)$ 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 the 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_1's are samples from the population so they have the same distribution as the population so

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

□

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

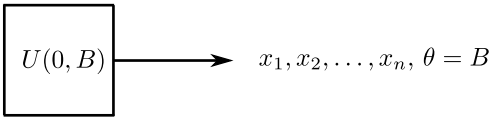
$$x_1 + x_2 + \ldots + x_n = \#\text{ of } 1's.$$

Thus $\overline{x}$ is the "common sense" estimator, the relative number of observed successes.

An Examples 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” problem 14, on page 242.

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



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 object if one of the x's is equal to B and

$$P\left(\bigcup_{1=1}^{n}(X_1 = B)\right) = \sum_{1=1}^{n} P(X_1 = B)$$
$$= 0 + 0 + \ldots + 0 = 0$$

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 page 252, problem 32 *you will see exactly by how much if undershoots the mark*.

Theorem

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

so $\left(\frac{n+1}{n}\right)$ *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 unbiased estimators of the population mean.
Common sense tells us that $\overline{X}$ is better.
What mathematical criterion separates them. We have

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

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

so

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

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 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?
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

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

so

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

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

So $MSE(\hat{\theta})$ is the squared error $(h(x_1, x_n) - \theta)^2$ 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 wont to minimize this squared error. Here is the point.

Theorem

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

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

This is amazingly easy to prove.

Proof.

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.

Definition (text page 238)

The standard error of the estimator $\hat{\theta}$, denoted $\sigma_{\hat{\theta}}$ is $\sqrt{V(\hat{\theta})}$.
It is of the denoted $s_{\hat{\theta}}$ (not quite true) see page 238.