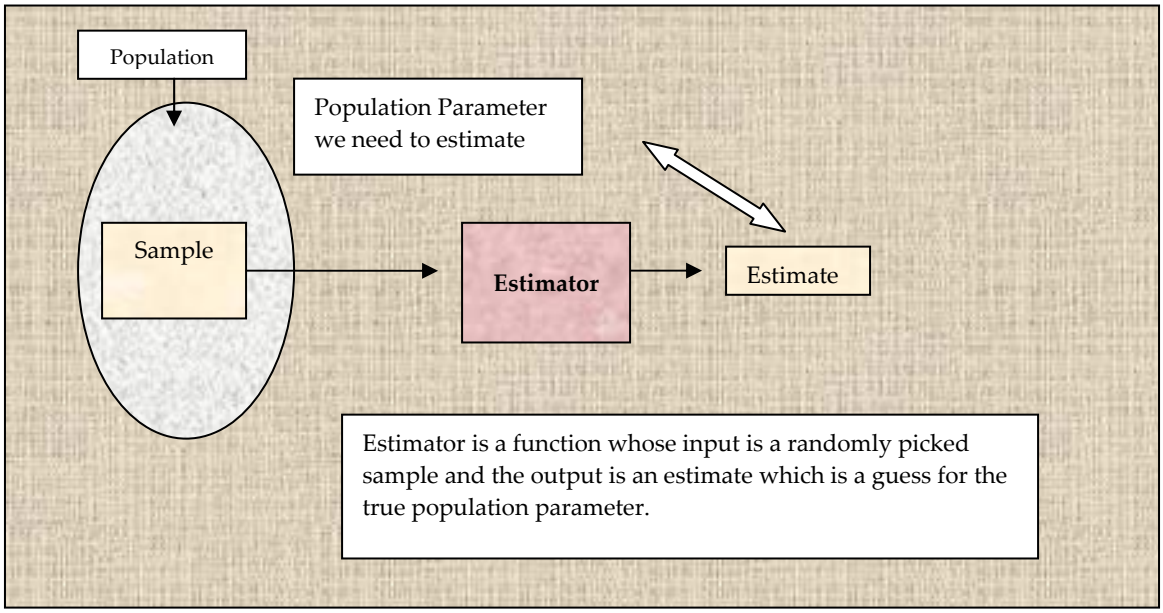
Estimation Theory

Make everything as simple as possible, but not a bit simpler- Albert Einstein

-Akshat Shankar, FRM

In statistics, an estimator is a a function of the observable sample data *(statistic)* that is used to estimate an unknown population parameter.
The resulting estimate can be a point estimate or an interval. Here we are concerned only with point estimates (the exact number)



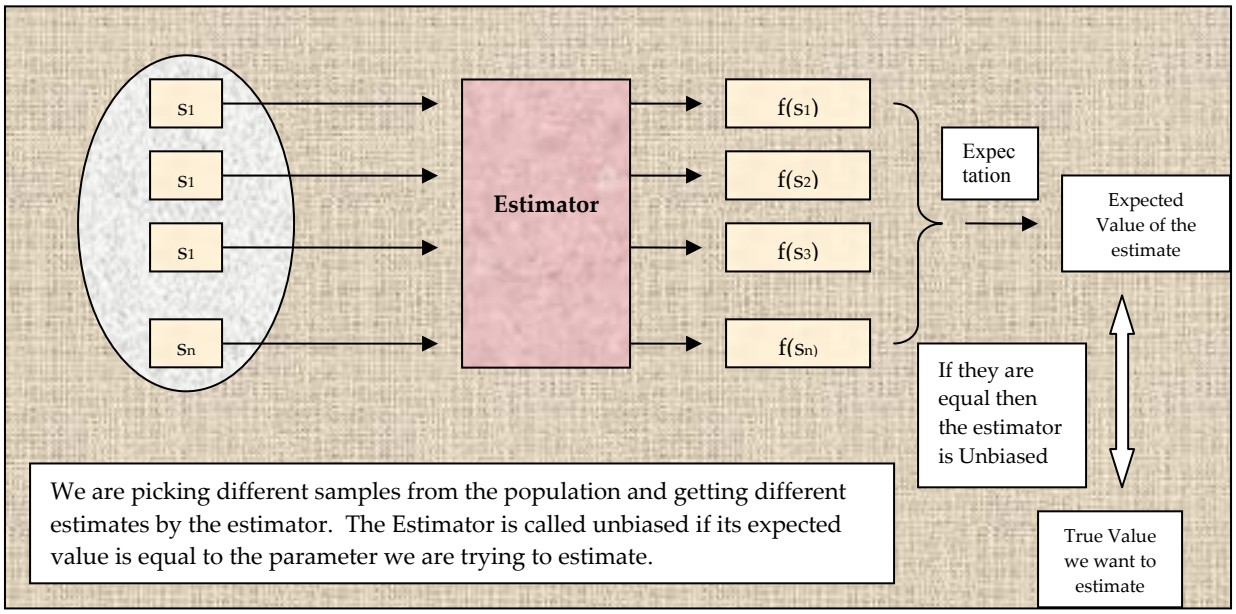
Suppose there is a coin with probability p of showing up heads. The only way of finding the actual p is either to study mechanics of the coin (and I have no clue if that can be done!) or to flip it infinite times. (and I have every clue that it cannot be done!) So we need to find a guess for p. And this guess can be anything.
One very simple thing to do is that we flip the coin 100 times and count how many times head comes. (say 20) . Then we can conjecture that p = 20/100= 0.2.
Of course it may happen that p was actually 0.5 and we were plain unlucky to get so less heads! *
But if our 'estimator' is good then it would behave properly in 'normal' circumstances. So what are some of the properties of a 'good' estimator.
Primarily, there are three desirable qualities of an estimator:

1. Unbiased: On an average the estimator should be right
2. Efficient: Our estimator should not throw very different results depending on the different samples drawn i.e the variance of the estimator should be low. The one which is unbiased and has the lowest possible variance is called efficient.
3. Consistent: As sample size increases (and hence the sample becomes closer to the population) our estimate should become closer to the actual value. If this is happening then we call it to be consistent.

*. The distribution of number of heads in 100 trials is binomially distributed and we can actually see that if p =0.5 then the probability of getting 20 or less heads is very low.

1. Unbiasedness

The bias of an estimator $\hat{\theta}$ is Bias($\hat{\theta}$) = E($\hat{\theta}$) – θ. If the bias of an estimator is 0, then it is said to be unbiased.



Question 1: Is the estimator we have defined above for finding probability p of a coin unbiased?

Answer 1: We toss the coin N times (N is our choice). Say the outcomes are $x_1, x_2 \ldots x_N$. (where $x_i = 1 \Rightarrow heads$, $x_i = 0 \Rightarrow tails$) Hence $S = x_1 + x_2 + .. + x_N$ =no′ of heads in N tosses. Our estimator = $\frac{S}{N}$. So the expected value = $E(\frac{S}{N}) = \frac{1}{N} E(x_1 + x_2 + .. + x_N)$

$= \frac{1}{N}(p + p + \ldots p) = \frac{N.p}{N} = p$. Hence our estimator is unbiased.

Question 2: Suppose $x_1, x_2 \ldots$ be iid's which are Normally distributed with mean μ and standard deviation σ. Why do we use n-1 instead of n in the estimator of standard deviation.

$$s_1 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad s_2 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Answer 2: $E(s_1) = \sigma$ while $E(s_2) = \sigma * \sqrt{\frac{n-1}{n}}$. So s_1 is an unbiased estimator of σ while s_2 is not.

Question 3: Suppose one is an Allied intelligence analyst during World War II, and one has some serial numbers of captured German tanks. Further, assume that the tanks are numbered sequentially from 1 to N. How does one estimate the total number of tanks?

Hint 3: Mathematically speaking, we have a population of numbers from $1,2...N$ **.**where N is unknown which we need to estimate. The numbers on the tanks we have captured can be viewed as drawing a sample from the population. From this sample we need to estimate N .

Answer 3: Suppose you have captured 4 tanks numbered {7, 3, 32, 18}. The unknown parameter N is definitely $\geq \max\{7,3,32,18\} = 32$ **.** But if we estimate N by 32, it would be a biased estimator as the true value would always be equal or more than the estimate and hence the expected value of the estimate would be less (and not equal) than true value.

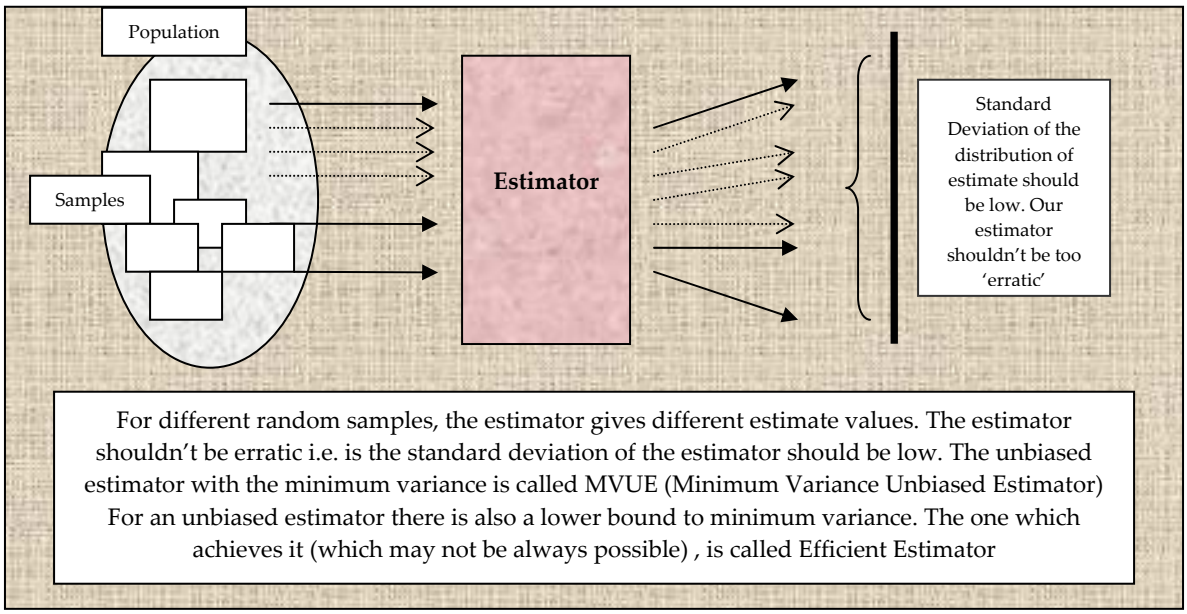
The unbiased estimator of N would be:

$\hat{N} = m + \frac{m}{k} - 1$ where m is the largest serial number observed k is the number of tanks observed. So in this case it is 39. Prove it!

2.Efficiency

If $\hat{\theta}$ is an unbiased estimator whose variance achieves equality in the Cramer Rao lower bound (for all θ), it is called efficient.

Comments: Our estimator should not be very erratic depending on the sample drawn. i.e the variance of the estimator should be low. Of course we can always come up with a 0 variance estimator. Like for the problem of finding probability *p* of a biased coin, the estimator can be a constant 0.4 regardless of the sample drawn. (It would be stupidity though!) It would have 0 variance but then it is a nonsense estimator. If the estimator is unbiased (which brings some sense), then we can talk about minimum variance unbiased estimator. In fact the result is co-proven by an Indian and known as Cramer Rao inequality.



Question 4: Suppose we want to compute the population mean. And we observe the value (i.e. we pick random sample of size 1) on two consecutive days. (we assume that the two observations are independent) Suppose we give a weight of 2/3 to the recent observation while 1/3 to the older observation. So the estimator of μ becomes

$\hat{\mu} = \frac{1}{3}X_1 + \frac{2}{3}X_2$**.** Is this unbiased? Is this efficient?

Answer 4: $E[\hat{\mu}] = E[\frac{1}{3}X_1 + \frac{2}{3}X_2] = \frac{1}{3}E[X_1] + \frac{2}{3}E[X_2] = \frac{1}{3}\mu + \frac{2}{3}\mu = \mu$

So the estimator is unbiased.

$$V[\hat{\mu}] = V[\frac{1}{3}X_1 + \frac{2}{3}X_2] = \frac{1}{9}V[X_1] + \frac{4}{9}V[X_2] = \frac{1}{9}\sigma^2 + \frac{4}{9}\sigma^2 = \frac{5}{9}\sigma^2$$

In case we give equal weights to the observation, it remains unbiased and we get variance as

$$V[\hat{\mu}] = V[\frac{1}{2}X_1 + \frac{1}{2}X_2] = \frac{1}{4}V[X_1] + \frac{1}{4}V[X_2] = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{1}{2}\sigma^2 (< \frac{5}{9}\sigma^2)$$

Clearly the variance of an equal weighted estimator is less. In fact it can be easily shown that an equal weighted estimator of μ is an efficient estimator.

Cramer Rao Inequality

Suppose θ is an unknown deterministic parameter which is to be estimated from measurements x, distributed according to some probability density function f(x;θ). The variance of any unbiased estimator $\hat{\theta}$ of θ is then bounded by

$$\operatorname{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where the Fisher information I(θ) is defined by

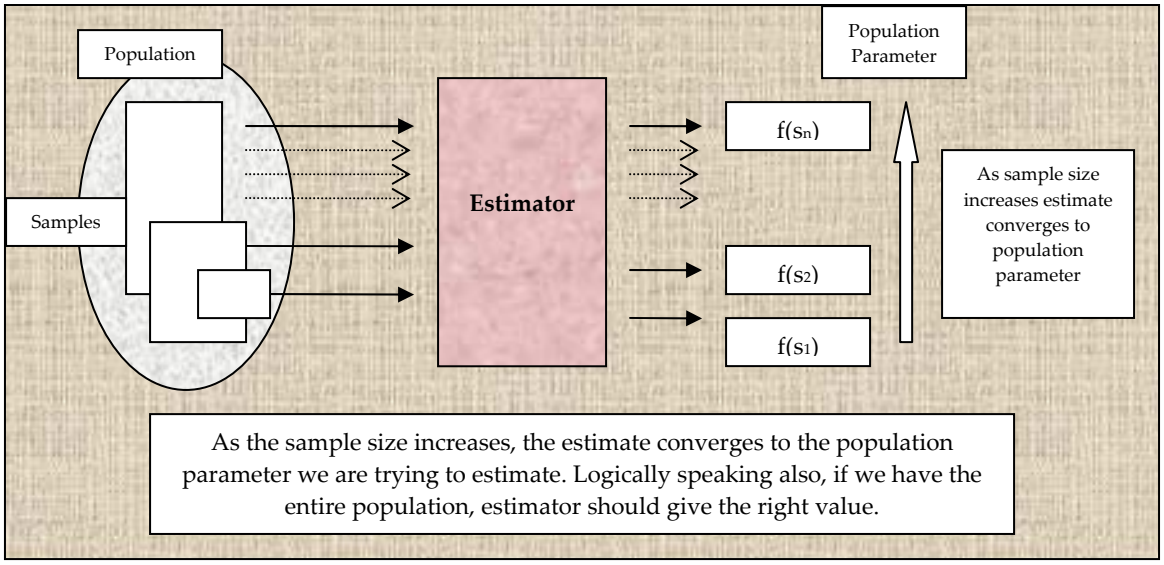
$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial \ell(x;\theta)}{\partial \theta}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 \ell(x;\theta)}{\partial \theta^2}\right]$$

Minimum Variance Unbiased Estimator

A uniformly minimum-variance unbiased estimator or minimum-variance unbiased estimator (UMVU or MVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter. (but may still not satisfy the CR bond)
If it satisfies the Cramer Rao bound then it is also called the Best Estimator.

3.Consistency

Let $X_1, X_2, \ldots X_n$ ***… be a sequence of iid random variables drawn from a distribution with parameter*** θ ***and*** $\hat{\theta}$ ***be an estimator of*** θ ***. We say that*** $\hat{\theta}$ ***is consistent as an estimator of*** $\hat{\theta}$ ***if*** $\hat{\theta} \xrightarrow{p} \theta$ ***^ or*** $Lim_n P(|\hat{\theta}(X_1, X_2 \ldots) - \theta| \leq \in \ \forall \in > 0$



Comments: It is satisfactory to know that an estimator θ will perform better and better as we obtain more examples. If the limit at $n \to \infty$ the estimator tends to be always right, it is said to be consistent. It is a relatively weak property but any nonsense estimator wouldn't be consistent

Question 5: Let there be a population (not necesaarily normal) with mean μ and we want to estimate μ. Is $\hat{\mu} = \frac{X_1 + X_2 + \ldots X_n}{n}$ a consistent estimator? (where $X_1, X_2, \ldots X_n$ randomly drawn samples.

Hint 5: Just remember Law of Large Numbers. It states that the sample average converges almost surely to the expected value.

Answer 5: By Law of Large Numbers, $\hat{\mu} = \frac{X_1 + X_2 + \ldots X_n}{n} \to \mu$ for $n \to \infty$ where $X_1, X_2, \ldots X_n$ is an infinite sequence of i.i.d. random variables with finite expected value E(X1) = E(X2) = ... = μ < ∞.Hence as $n \to \infty$, the estimator would give the correct value. Hence consistent.

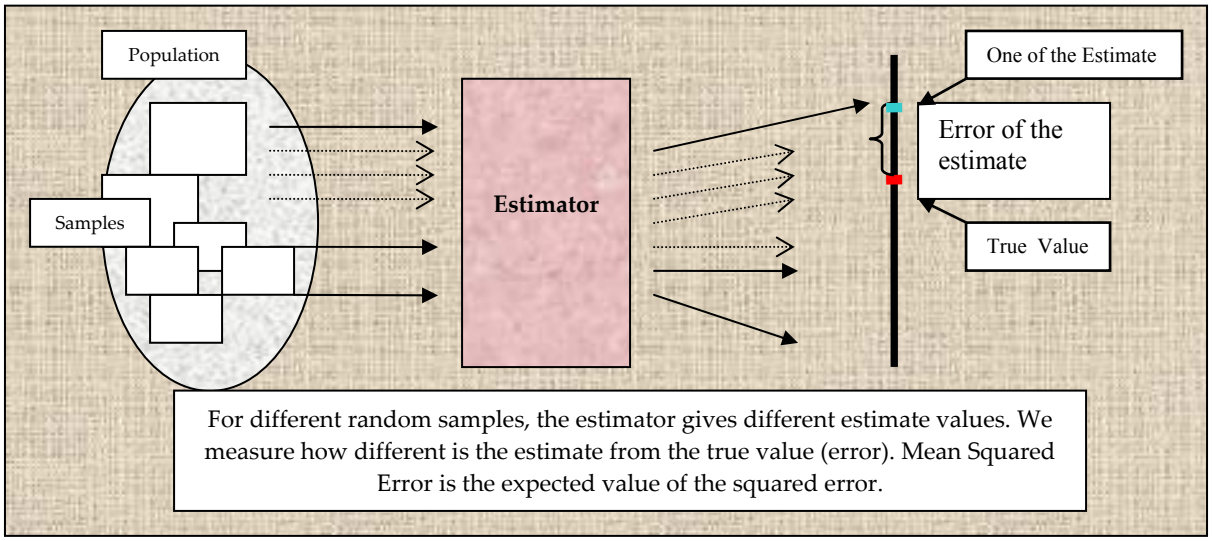
4. Mean Squared Error

The mean squared error (MSE) of an estimator is $E[(\theta-\hat{\theta})^2]$

*Comments :*Apart from the three properties which have been described, it would be great to come up with a metric that measures how good the estimator is. The most simple an intuitive way is to take the expected value of the squared errors (so that th errors don't cancel out) which is called Mean Squared Error. The lower the MSE the better the estimator

Result: The mean squared error of an estimator equals sum of variance and squared bias.

MSE = $E[(\theta-\hat{\theta})^2] = Var(\hat{\theta}) + E[(\theta-\hat{\theta})]^2 = Var(\hat{\theta}) + Bias(\hat{\theta})^2$



Question 6 : Suppose there is a population which is Normally Distributed with parameters μ and σ^2. We draw a sample from it and build an estimater of μ which is $\hat{\mu} = \bar{x}$. What is the Mean Squared Error of this estimator?

***Answer 6*:** $E[(\mu-\bar{x})^2 = Var(\bar{x}) + E[(\mu-\bar{x})]^2 = Var(\bar{x}) + 0$

$$Var(\frac{1}{n}(x_1+x_2+...x_n)) = \frac{1}{n^2}(Var(x_1)+Var(x_2)+...Var(x_n)) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

As x_i's are independent so Covariance term is 0.

Standard Error: Standard Error is the standard deviation of the estimator. $\sqrt{Var(\hat{\theta}} = \sqrt{E(\hat{\theta}-E(\hat{\theta}))}$

Comments: You would have heard this infinite times, lets try to understand it. We are trying to estimate $\hat{\theta}$. As $\hat{\theta}$ depends on the sample, it is a random variable. The standard deviation of this random variable is called Standard Error.

Question 7: What is the standard error of the above estimator (Question 6)

Answer 7: As it is an unbiased estimator, MSE = Var($\bar{x}$) Hence the Standard Error is $\sigma/\sqrt{n}$.

5. Maximium Likelihood Estimation

Definition. Let $X_1, X_2, \ldots X_n$ ***be sampled from a distribution with parameter*** θ***. The maximum likelihood estimator (MLE)*** $\hat{\theta}_{MLE}$ ***is the*** θ ***that maximizes the likelihood function*** $L(\theta) = L(X_1, X_2 \ldots X_n \mid \theta)$ ***if it exists***

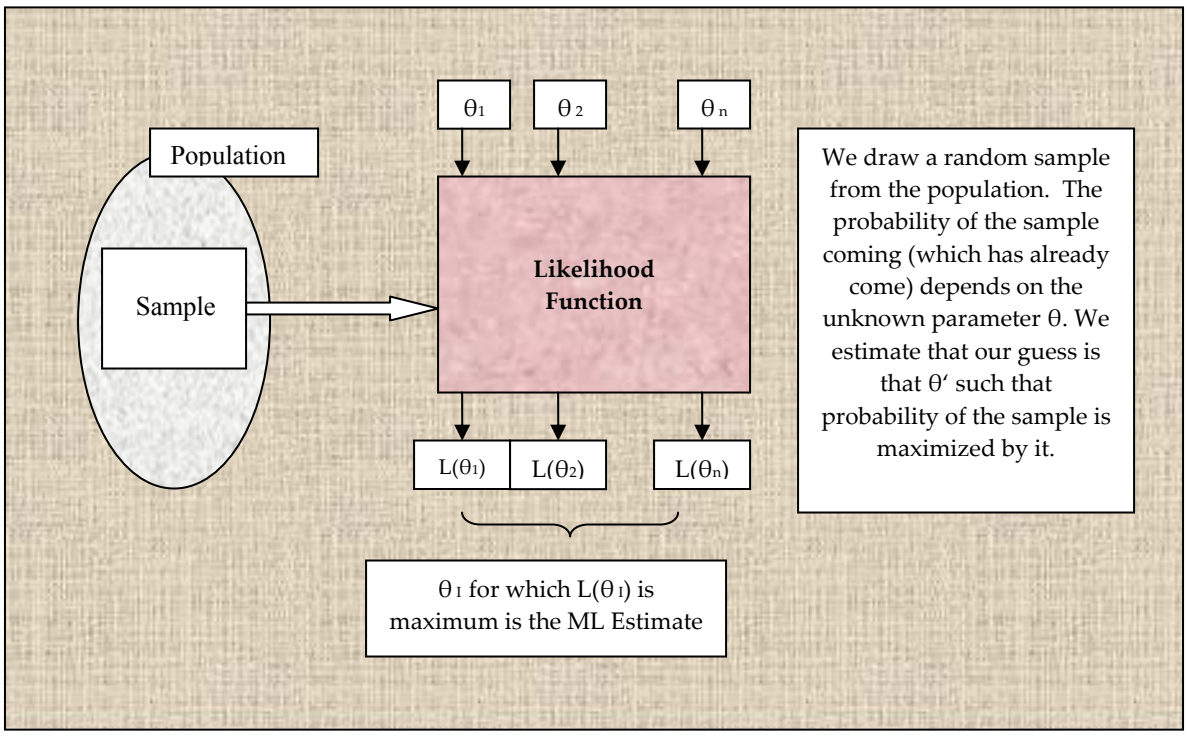
Likelihood Function: Let $X_1, X_2, \ldots X_n$ be iid RV's whose distribution is the pdf (pmf) $f_i, f_2 \ldots f_n$. The Likelihood function is the product of the pdf's.

$L(\theta) = L(X_1, X_2 \ldots X_n \mid \theta)$ $L(\theta) = L(X_1, X_2 \ldots X_n \mid \theta) = f_1(x_1).f_2(x_2)\ldots f_n(x_n)$

In case random variables are discrete. Then likelihood is the probability of the outcome $Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$.

Comments: If the random variables are continuous, then the value of likelihood function would always be 0 (as it is density not the mass) so how do we maximize the probability but it can be resolved if we consider intervals around the point (observation) rather than the point and then make the interval size tend to zero. Or in Orwell's words-All zeroes are born equal but some are more equal than others!

So the way to proceed is to differentiate the likelihood function (for most of us, this is only what MLE means!) as it would give the maximum value of θ which would be our answer.



Lets think about the problem in a different way. Suppose there is a girl you admire and you are interested in knowing whether she likes you or not. If we formulate the 'problem' (pun intended) mathematically.

We need to estimate the unknown parameter θ which can be 0 or 1. One means that she also likes you while zero means that (sadly) she doesn't! And you want to estimate θ. You can of course pluck rose petals and randomly (and foolishly) get 0 or 1 as the answer.
The Maximum Likelihood way of this problem would be this. You have been talking to her for sometime and so getting some hints. This is in essence a sample of her true thoughts about you. And her thoughts about you depend on the fact that she likes you or not, so the population is indeed a function of θ.
You think that if she loves you (θ = 1) then what is the probability that she would say all the things that she has said. You then compute the likelihood of the sample if she doesn't likes you. If the first thing is greater then you can estimate that she likes you.
So if the only hint is that she woke up till 12 to wish you happy birthday. (sample: say $X_1 = 1$ denotes that she did). Then supposing she loves you (θ = 1), the probability that $X_1 = 1$ ($\Pr(X_1 = 1 \mid \theta = 1)$ = p1. (I guess it would be almost 1 except if she is a philosopher!) Suppose she doesn't love you (θ = 0), the probability that $X_1 = 1$ ($\Pr(X_1 = 1 \mid \theta = 0)$)= p2. We all know that p1 is greater than p2 so if $X_1 = 1$ then Maximim Likelihood Estimate is (θ = 1)!
Your guess may still be wrong from the reality (as was mine some years back ;)) as the sample size may be small (it was just one in the above example) or it may not be random enough etc…but then it is an educated guess at the end of the day.

Question 8: Suppose I give you a biased coin and ask you to find the probability of heads. To make your life simple (or difficult if you have crammed MLE!) , I also tell that the answer is either 1/3 or ½ or 2/3. You toss the coin 100 times and get 60 heads and 40 tails. What is the true probability of heads coming if we use Maximum likelihood approach.
Hint 8 : We have been blindly trained to mark 60/100 as the answer but sadly as the true probability has to be from the set {1/3, ½, 2/3} your answer is wrong.
Answer 8 : Lets once again think of what MLE means. We draw some sample and see what value of the parameter had made this outcome more probable.

Case 1: p =1/3 $\Pr(H = 60 \mid p = 1/3) = {}^{100}C_{60}(\frac{1}{3})^{60}(\frac{2}{3})^{40} = 2.93 \cdot 10^{-8}$

Case 2: p = ½ $\Pr(H = 60 \mid p = 1/3) = {}^{100}C_{60}(\frac{1}{2})^{60}(\frac{1}{2})^{40} = 0.0108$

Case 3: p = 2/3 $\Pr(H = 60 \mid p = 1/3) = {}^{100}C_{60}(\frac{2}{3})^{60}(\frac{1}{3})^{40} = 0.03$

Case 3 has the maximum probability, so the Maximum Likelihood Estimate of p is 2/3.

Question 9– Suppose I give you another biased coin and ask you to find the probability of heads. This time I also don't know the answer so *p* can be any number from 0 to 1.
You toss the coin 100 times and get 70 heads and 30 tails. What is the true probability of heads coming if we use Maximum likelihood approach.
Answer 9: Lets once again think of what MLE means. We draw some sample and see what value of the parameter had made this outcome more probable.

Probabilty of what we got (70 heads) is

$L(p) = \Pr(H = 70 \mid prob = p) = {}^{100}C_{70}\, p^{70}(1-p)^{30}$

We want to find p such that *L(p)* is maximized. And we can differentiate *L(p)* w.r.t *p* and equate it to 0 to find such *p*. (do check the double derivative also)

$${}^{100}C_{70}(70p^{69}(1-p)^{30} + 30.(-1).p^{70}(1-p)^{29} = {}^{100}C_{70}\, p^{69}(1-p)^{29}(70(1-p) - 30.p) = 0$$

$$70(1-p) - 30.p = 0 \Rightarrow 70 - 100p = 0 \Rightarrow p = 0.7$$

Hence the Maximum Likelihood Estimate is 0.7. (similar to our intuition)

***Question 10*:** Suppose there is a population which is Normally Distributed with parameters μ and σ^2. We draw a sample from it and need to estimate through MLE.

a) μ

b) σ^2

Hint: We have been trained to answer this that population mean is equal to sample mean and population variance is equal to sample variance. (denominator has n or n-1 ?) But is there any mathematical genesis to this argument. Lets try to methodically work out through the algorithm of MLE.

Answer:

a) The sample we get is $(x_1, x_2, ...x_n)$

The probability density function of getting x_1 is

$\Pr(X_1 = x_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right)}$ (X_1 is the r.v. which denotes the outcome of the first pick)

We need to maximimize the Likelihood function = $\Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$

As $(X_1, X_2, ...X_n)$ are independent (we can draw the sample in such a way)

$$= \Pr(X_1 = x_1).\Pr(X_2 = x_2)...\Pr(X_n = x_n)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp^{\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right)} . \frac{1}{\sqrt{2\pi}\sigma} \exp^{\left(-\frac{(x_2-\mu)^2}{2\sigma^2}\right)} .... \frac{1}{\sqrt{2\pi}\sigma} \exp^{\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp^{\left(\left(-\sum_{i=1}^{n}(x_i-\mu)^2\right)/2\sigma^2\right)} ...(A)$$

Of course we can differentiate (A) and equate it to 0 to get the answer.

Or as log(x) is a monotonically increasing function, so x which maximizes $f(x)$ is same as the one which maximizes $\log(f(x))$ (which is sometimes called LogLikelihood function). And it eases the computational difficulty of the problem many times.

Though here we would be slightly smart and re-arrange the equation.

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp^{\left(\left(-\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right)/2\sigma^2\right)}$$

And when $\mu = \bar{x} = \frac{1}{n}.\sum_{i=1}^{n} x_i$, the likelihood function is maximized. So $\bar{x}$ is the MLE of μ .

b) If we differentiate (A) w.r.t σ^2 and equate it to 0, we get

$$\sigma^2 = \frac{1}{n}.\sum_{i=1}^{n}(x_i - \mu)^2$$

Note that the MLE estimator of σ^2 has a division by n rather than n-1. So MLE estimator is not an unbiased estimator.

. "Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius -- and a lot of courage -- to move in the opposite direction." : Albert Einstein