IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: MSc

Subject : Probability and Statistics -2

Chapter: Unit 3 Chapter 3

Chapter Name: Theory of Estimation 1

Index

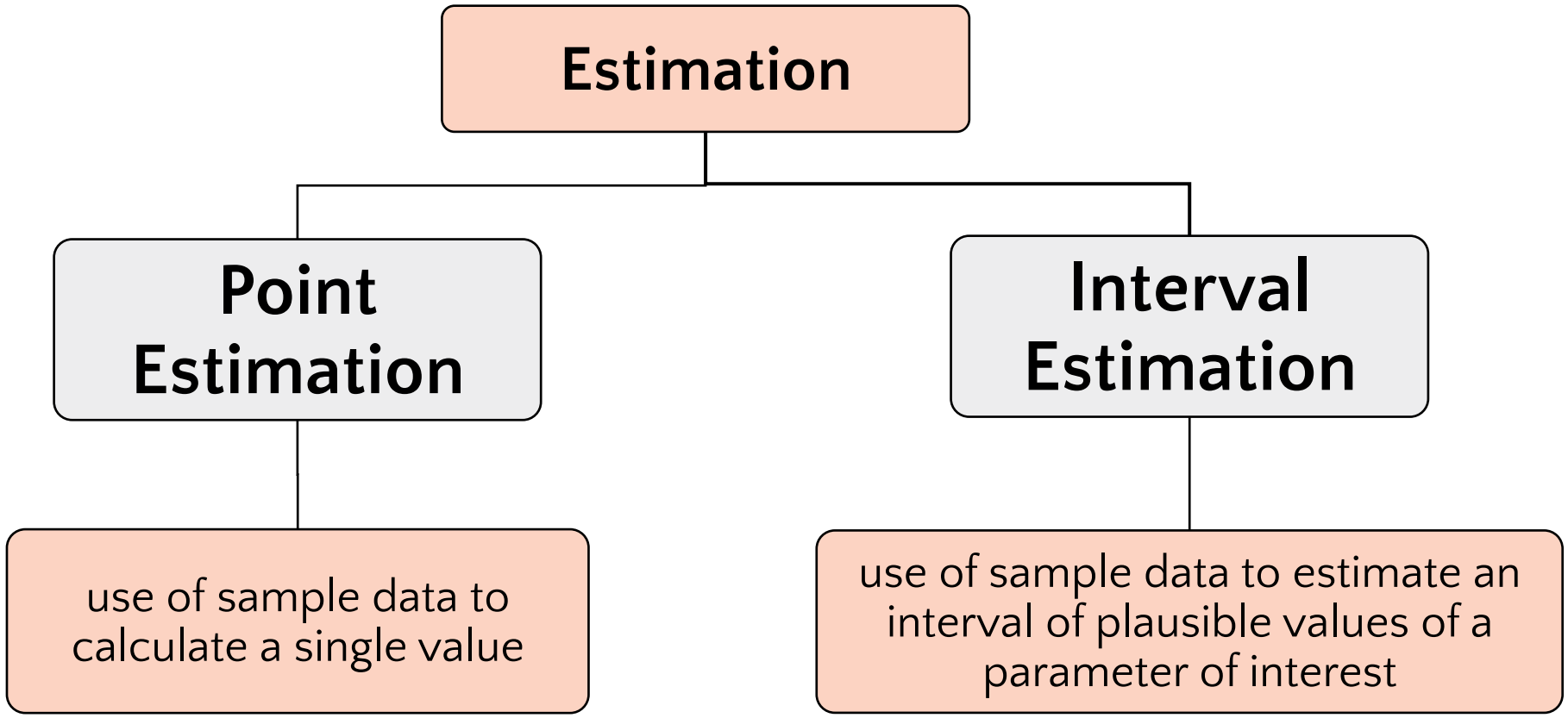
Index

1 Estimation

Estimation (or estimating) is the process of finding an estimate, or approximation, which is a value that is usable for some purpose even if input data may be incomplete, uncertain, or unstable.

- The value is nonetheless usable because it is derived from the best information available.
- Typically, estimation involves "using the value of a statistic derived from a sample to estimate the value of a corresponding population parameter". The sample provides information that can be projected, through various formal or informal processes, to determine a range most likely to describe the missing information.
- An estimate that turns out to be incorrect will be an overestimate if the estimate exceeded the actual result, and an underestimate if the estimate fell short of the actual result.

1 Estimation



2 Method of Moments

- The basic principle is to equate population moments (ie the means, variances, etc of the theoretical model) to corresponding sample moments (ie the means, variances, etc of the sample data observed) and solve for the parameter(s).

The one-parameter case

- This is the simplest case: to equate population mean, E(X) , to sample mean, $\bar{x}$, and solve for the parameter, ie:

$$\boldsymbol{E[X] = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

2 Method of Moments

- We can apply this method to a number of different single parameter distributions. For example, the method works well with a random sample from a Poisson distribution.

- Note: For some populations the mean does not involve the parameter, such as the uniform on $(-\theta, \theta)$ or the normal $N(0,\sigma^2)$, in which case a higher-order moment must be used.
- However such cases are rarely of practical importance.

- The **estimator** is written in upper case as it is a random variable and will have a sampling distribution. The **estimate** is written in lower case as it comes from an actual sample of numerical values.

2 Method of Moments

The two-parameter case

- With two unknown parameters, we will require two equations.
- This involves equating the first and second-order moments of the population and the sample, and solving the resulting pair of equations.
- Moments about the origin can be used but the solution is the same (and often more easily obtained) using moments about the mean – apart from the first-order moment being the mean itself.

2 Method of Moments

- The first-order equation is the same as in the one-parameter case:

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The second-order equation is:

$$E[X^2] = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

- or equivalently:

$$E[(X-\mu)^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 1/n \sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$or \quad var(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

- Note that we are not equating sample and population variances here; we are using a denominator of n on the right hand side of the final equation, whereas the sample variance uses a denominator of n-1.

4 Maximum Likelihood Estimate(MLE)

- In most cases taking logs greatly simplifies the determination of the maximum likelihood estimator (MLE) $\hat{\theta}$.

- Differentiating the likelihood or log likelihood with respect to the parameter and setting the derivative to zero gives the maximum likelihood estimator for the parameter.

- It is necessary to check, either formally or through simple logic, that the turning point is a maximum. Generally the likelihood starts at zero, finishes at or tends to zero, and is non-negative. Therefore if there is one turning point it must be a maximum.

- The formal approach would be to check that the second derivative is negative. For the above example we get:

$$\frac{d^2}{d\lambda^2} logL(\lambda) = -\frac{n}{\lambda^2} < 0 \Rightarrow max$$

Question 1

- The sample mean and sample variance for a large random sample from a Gamma(α, λ) distribution are 10 and 25, respectively. Use the method of moments to estimate α and λ .

Solution

- Equating the mean and variance, we get:

$$\frac{\hat{\alpha}}{\hat{\lambda}} = 10 \quad \text{and} \frac{\hat{\alpha}}{\hat{\lambda}^2} = 25$$

- Dividing the first equation by the second gives:

$$\hat{\lambda} = \frac{10}{25} = 0.4 \Rightarrow \quad \hat{\alpha} = 10 \text{ x } 0.4 = 4$$

2 Method of Moments

- For cases with more than two parameters, moments about zero should be used.
- For example, if you had 3 parameters to estimate, you would use the set of equations:

$$E[X] = \frac{1}{n}\sum x_i \qquad E[X^2] = \frac{1}{n}\sum x_i^2 \qquad E[X^3] = \frac{1}{n}\sum x_i^3$$

- This approach can be extended in an obvious way for more than three parameters.

3 Likelihood vs Probability

- Probability corresponds to finding the chance of something given a sample distribution of the data, while on the other hand, Likelihood refers to finding the best distribution of the data given a particular value of some feature or some situation in the data.

Probability

- Consider a dataset containing the heights of the people of a particular country. Let's say the mean of the data is 170 & the standard deviation is 3.5.

$$\mathbf{P}(\boldsymbol{height} > \mathbf{170} | \boldsymbol{\mu} = \mathbf{170}, \boldsymbol{\sigma} = \mathbf{3.5})$$

- While calculating probability, feature value can be varied, but the characteristics(mean & Standard Deviation) of the data distribution cannot be altered.

Likelihood

- Consider the exactly same dataset example as provided above for probability, if their likelihood of height > 170 cm has to be calculated then it will be done using the information shown below:

$$\mathbf{Likelihood}(\boldsymbol{\mu} = \mathbf{170}, \boldsymbol{\sigma} = \mathbf{3.5} | \boldsymbol{height} > \mathbf{170})$$

- The likelihood in very simple terms means to increase the chances of a particular situation to happen/occur by varying the characteristics of the dataset distribution.

4 Maximum Likelihood Estimate(MLE)

- The method of maximum likelihood is widely regarded as the best general method of finding estimators. In particular maximum likelihood estimators have excellent and usually easily determined asymptotic properties and so are especially good in the large-sample situation.

- Prerequisites and Assumptions:
  - Knowing the underlying distribution.
  - Assuming that the sample points are independent and identically distributed.

4 Maximum Likelihood Estimate(MLE)

The one-parameter case

- The most important stage in applying the method is that of writing down the likelihood:

$$\boldsymbol{L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)}$$

- for a random sample $x_1, x_2, \ldots, x_n$ from a population with density or probability function $f(x; \theta)$.

- $\prod$ means product, so $\prod_{i=1}^{n} f(x_i)$ would mean $f(x_1)$ x $f(x_2)$ x $f(x_3)$ x ... x $f(x_n)$. The above statement is saying that the likelihood function is the product of the densities (or the probability functions in the case of discrete distributions) calculated for each sample value.

- Remember that θ is the parameter whose value we are trying to estimate.

- The likelihood is the probability of observing the sample in the discrete case, and is proportional to the probability of observing values in the neighbourhood of the sample in the continuous case.

Example

- $f(X_i|P) = P^{X_i}(1-P)^{1-X_i}$
- $X_i = \begin{cases} 1, male \\ 0, female \end{cases}$

$$f(1|P) = P^1(1-P)^{1-1} = P$$
$$f(0|P) = P^0(1-P)^{1-0} = 1-P$$

$$f(X_1, X_2, \dots, X_N|P) = P^{X_1}(1-P)^{1-X_1} * P^{X_2}(1-P)^{1-X_2} * \cdots * P^{X_N}(1-P)^{1-X_N}$$

$$P(X_1 = x_1, X_2 = x_2, \dots, X_N = x_n) = \prod_{i=1}^{N} P^{X_i}(1-P)^{1-X_i} = L$$

Example

- $L = \prod_{i=1}^{N} P^{X_i}(1-P)^{1-X_I}$
- $\frac{\partial L}{\partial P} = 0 \Rightarrow \hat{P}$
- $logL = \log\left(\prod_{i=1}^{N} P^{X_i}(1-P)^{1-X_I}\right) = \sum_{i=1}^{N} \log(P^{X_i}(1-P)^{1-X_i}) = \sum_{i=1}^{N}(X_i logP + (1-X_I)\log(1-P))$
- $logL = logP \sum_{i=1}^{N} X_i + \log(1-P)\sum_{i=1}^{N}(1-X_i)$
- But $\frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}$
- $logL = N\bar{X} log\hat{P} + N(1-\bar{X})\log(1-\hat{P})$
- $\frac{\partial logL}{\partial \hat{P}} = \frac{N\bar{X}}{\hat{P}} - \left(\frac{\mathrm{N}(1-\bar{X})}{1-\hat{P}}\right) = 0$
- $\hat{P} = \bar{X}$

4 Maximum Likelihood Estimate(MLE)

The two-parameter case

- This is straightforward in principle and the method is the same as the one-parameter case, but the solution of the resulting equations may be more awkward, perhaps requiring an iterative or numerical solution.

- The only difference is that a partial derivative is taken with respect to each parameter, before equating each to zero and solving the resulting system of simultaneous equations for the parameters.

4 Maximum Likelihood Estimate(MLE)

So in summary, the steps for finding the maximum likelihood estimator in straightforward cases are:

- Write down the likelihood function, L.
- Find $\ln L$ and simplify the resulting expression.
- Partially differentiate $\ln L$ with respect to each parameter to be estimated.
- Set the derivatives equal to zero.
- Solve these equations simultaneously.

5 A special case – the uniform distribution

- For populations where the range of the random variable involves the parameter, care must be taken to specify when the likelihood is zero and non-zero. Often a plot of the likelihood is helpful.

- We look at this in the next example – note how we specify when the likelihood is zero (ie it does not exist for the specified values of the parameter) and non-zero (ie where it does exist for the specified values of the parameter).

- The second important feature about this example is that the usual route for finding the maximum using differentiation breaks down.

Example

Question:

- Derive the maximum likelihood estimate of θ for U(0,θ) based on a random sample of values $x_1, x_2, \ldots, x_n$.

Example

Solution:

- For a sample from the $U(0, \theta)$ distribution we must have $0 \leq x_1, \ldots, x_n \leq \theta$. Hence max $x_i \leq \theta$.
- Thus the likelihood for a sample of size n is:

$$L = \begin{cases} \dfrac{1}{\theta^n} & \text{if } \theta > \max x_i \\ 0 & \text{otherwise} \end{cases}$$

- Differentiation doesn't work because $\frac{d}{d\theta} lnL(\theta) = -\frac{n}{\theta}$ which gives a turning point of $\theta \to \infty$.
- The second derivative shows the problem $\frac{d^2}{d\theta^2} lnL(\theta) = \frac{n}{\theta^2} > 0$. We have a minimum as $\theta \to \infty$.
- So using common sense, we must find the θ that maximises $L(\theta) = \frac{1}{\theta^n}$. We want θ to be as small as possible subject to the constraint that $\theta \geq \max x_i$. Hence $\hat{\theta} = max\ x_i$.

6 Incomplete samples

- The method of maximum likelihood can be applied in situations where the sample is incomplete. For example, truncated data or censored data in which observations are known to be greater than a certain value, or multiple claims where the number of claims is known to be two or more.

- Censored data arise when you have information about the full range of possible values but it's not complete (eg you only know that there are, say, 6 values greater than 500). Truncated data arise when you actually have no information about part of the range of possible values (eg you have no information at all about values greater than 500).

- In these situations, as long as the likelihood (the probability of observing the given information) can be written as a function of the parameter(s), then the method can be used. Again in such cases the solution may be more complex, perhaps requiring numerical methods.

6 Incomplete samples

- For example, suppose a sample yields n observations $(x_1, x_2 \ldots, x_n)$ and m observations greater than the value y , then the likelihood is given by:

$$\boldsymbol{L(\theta) = \left[\prod_{i=1}^{n} f(X_i, \theta)\right] \mathrm{x}\, [P(X > y)]^m}$$

- Our estimate will be as accurate as possible if we use all the information that we have available. For incomplete samples, we don't know what the values above y are. All we know is that they are greater than y.

- Since the values above y are unknown we cannot use $L(\theta) = \prod_{i=1}^{n+m} f(X_i, \theta)$ .We instead use the formula given.

- If the information is more detailed than 'greater than y ' we can use a more detailed likelihood function. For example, if we have m observed values between y and z , and p observed values above z , in addition to the n known values, then we would use:

$$\mathbf{L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)\, x\, [P(y < X < z)]^m\, x\, [P(X > z)]^P}$$

7 Independent samples

- For independent samples from two populations which share a common parameter, the overall likelihood is the product of the two separate likelihoods.

Example

Question:

- The number of claims, X , per year arising from a low-risk policy has a Poisson distribution with mean μ . The number of claims, Y , per year arising from a high-risk policy has a Poisson distribution with mean 2μ .
- A sample of 15 low-risk policies had a total of 48 claims in a year and a sample of 10 high-risk policies had a total of 59 claims in a year. Determine the maximum likelihood estimate of μ based on this information.

Example

Solution:

- The likelihood for these 15 low-risk and 10 high-risk policies is:

$$L(\mu) = \prod_{i=1}^{15} P(X = x_i) \text{ x } \prod_{j=1}^{10} P(Y = y_i) = \prod_{i=1}^{15} \frac{\mu^{X_i}}{X_j!} e^{-\mu} \text{ x } \prod_{j=1}^{10} \frac{(2\mu)^{y_j}}{y_j!} e^{-2\mu}$$

$$= constant \text{ x } \mu^{\sum_{i=1}^{15} X_i} e^{-15\mu} \text{ x } \mu^{\sum_{j=1}^{10} y_j} e^{-20\mu}$$

$$= constant \text{ x } \mu^{48} e^{-15\mu} \text{ x } \mu^{59} e^{-20\mu} = constant \text{ x } \mu^{107} e^{-35\mu}$$

- The log-likelihood is:

$$\ln L(\mu) = contant + 107 \ln \mu - 35\mu$$

- Differentiating and setting equal to zero gives:

$$\frac{d}{d\mu} \ln L(\mu) = \frac{107}{\mu} - 35 \Rightarrow \hat{\mu} = \frac{107}{35} = 3.057$$

- Differentiating again to check we get a maximum:

$$\frac{d^2}{d\mu^2} \ln L(\mu) = -\frac{107}{\mu^2} < 0 \Rightarrow max$$

8 Unbiasedness

- Consideration of the sampling distribution of an estimator can give an indication of how good it is as an estimator. Clearly the aim is for the sampling distribution of the estimator to be located near the true value and have a small spread.

- If we have a random sample $X = (X_1, X_2, \dots, X_n)$ from a distribution with an unknown parameter θ and g($\underline{X}$) is an estimator of θ , it seems desirable that $E\left[g\left(\underline{X}\right)\right] = \theta$

- This is the property of unbiasedness.

- You can think of an unbiased estimator as one whose mean value equals the true parameter value.

9 Bias

- If an estimator is biased, its bias is given by $E\left[g\left(\underline{X}\right)\right] - \theta$, ie it is a measure of the difference between the expected value of the estimator and the parameter being estimated.

- If the bias is greater than zero, the estimator is said to be positively biased ie it tends to overestimate the true value. Alternatively, the bias could be less than zero, leading to a negatively biased estimator that would tend to underestimate the true value.

8 Unbiasedness

- The property of unbiasedness is not preserved under non-linear transformations of the estimator/parameter.
- So, for example, the fact that S^2 is an unbiased estimator of the population variance does not mean that S is an unbiased estimator of the population standard deviation.
- As indicated earlier unbiasedness seems to be a desirable property. However it is not necessarily an essential property for an estimator. There are many common situations in which a biased estimator is better than an unbiased one, and, in fact, better than the best unbiased estimator.
- The importance of unbiasedness is secondary to that of having a small mean square error.
- An unbiased estimator is simply one that for different samples will give the true value on average. However, it could be that some of the estimates are too large and some are too small – but on average they give the true value. So we need some way of measuring the 'spread' of the estimates obtained for different samples. That measure is the mean square error and is covered in the next slides.
- Therefore a biased estimator whose value does not deviate very far from the true value (ie has a small spread) would be preferable to an unbiased one whose values are 'all over the place' – as the biased estimator would be more reliable (ie no matter what sample we had, the estimate is still likely to be closer to the true value).

Example

Question:

- The following are estimators for the variance of a distribution having mean μ and variance σ^2 .
- Obtain the bias for each estimator:

i. $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

ii. $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Example

Solution:

- The formula for the bias of S^2 is:

$$bias(S^2) = E(S^2) - \sigma^2$$

- Consider $E(S^2)$:

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = E[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\Big]$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)\right)$$

- But since:

$$E(X_i^2) = var\ (X_i) + E^2(X_i) = \sigma^2 + \mu^2$$

$$E(\bar{X}^2) = var(\bar{X}) + E^2(\bar{X}) = \frac{\sigma^2}{n} + \mu^2$$

Example

Solution:

- So we get:

$$E(S^2) = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n}(\sigma^2+\mu^2)\right) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right)$$
$$= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)$$
$$= \frac{1}{n-1}(n-1)\sigma^2$$
$$= \sigma^2$$

- So the bias is:

$$bias(S^2) = E(S^2) - \sigma^2 = \sigma^2 - \sigma^2 = 0$$

- This means that S^2 is an unbiased estimator of σ^2

Example

Solution:

ii.

- Since $\sigma^{\mathbf{2}} = \frac{\boldsymbol{n-1}}{\boldsymbol{n}} S^2$ we can use the result from part (i) to get:

$$E(\hat{\sigma}^2) = E\left[\frac{n-1}{n} S^2\right] = \frac{n-1}{n} E(S^2) = \frac{n-1}{n}\sigma^2$$

- So the bias is:

$$bias(\hat{\sigma}^{\mathbf{2}}) = \boldsymbol{E}(\hat{\sigma}^{\mathbf{2}}) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

10 Mean Square Error

- As biased estimators can be better than unbiased ones a measure of efficiency is needed to compare estimators generally. That measure is the mean square error.
- The mean square error (MSE) of an estimator g($\underline{X}$) for θ is defined by:

$$MSE\big((g(\underline{X}))=E[(g(\underline{X})-\theta\big)^2]$$

- Note that this is a function of θ .
- Thus the mean square error is the second moment of g($\underline{X}$) about θ and an estimator with a lower MSE is said to be more efficient.
- The MSE of a particular estimator can be worked out directly as an integral using the density of the sampling distribution of g($\underline{X}$) , or using the density of $\underline{X}$ itself.

10 Mean Square Error

- However it is usually much easier to use the alternative expression:

$$MSE = Variance + bias^2$$

- as this makes use of quantities that are already known or can easily be obtained.

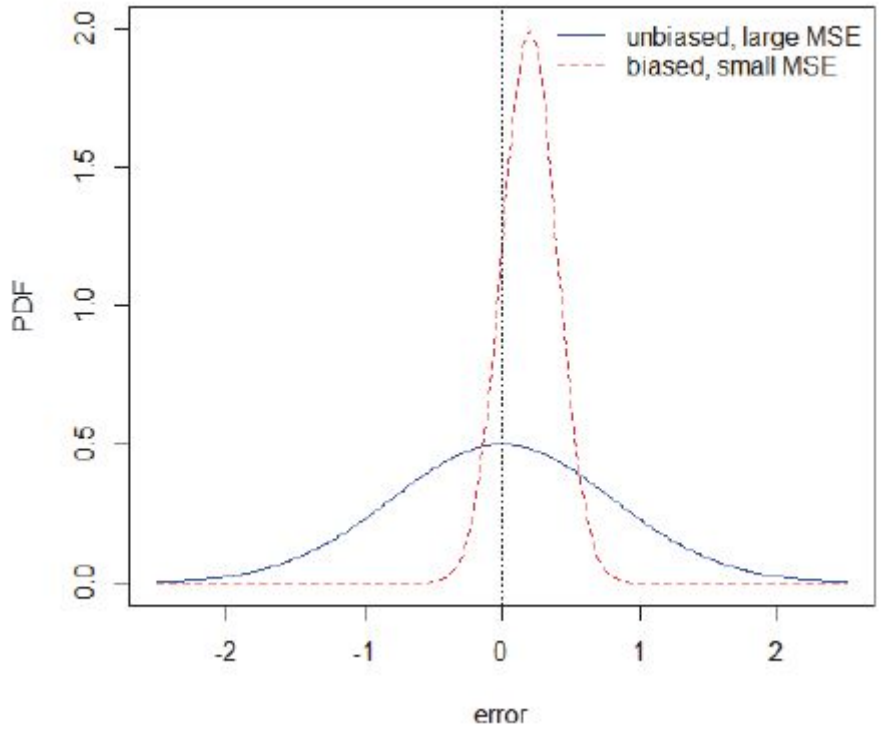
- This expression can be proved as follows:
- (Simplifying things by dropping the $\underline{X}$ and writing simply g.)

$$\begin{aligned} MSE(g) &= E[(g-\theta)^2] \\ &= E[\{(g-E[g])+(E[g]-\theta)\}^2] \\ &= E[(g-E[g])^2] + 2(E[g]-\theta)E\big[g-E[g]\big] + [E[g]-\theta]^2 \\ &= var[g] + 0 + bias^2[g] \quad as\ required \end{aligned}$$

- Note: If the estimator g($\underline{X}$) is unbiased, then MSE = variance.

11 Consistency

- The following diagram gives the sampling distributions of two estimators: one is unbiased but has a large variance, the other is biased with a much smaller variance. This illustrates a situation in which a biased estimator is better than an unbiased one.



11 Consistency

- It is clear that an estimator with a 'small' MSE is a good estimator. It is also desirable that an estimator gets better as the sample size increases. Putting these together suggests that it is desirable that MSE $\rightarrow 0$ as $n \rightarrow \infty$. This property is known as consistency.

Example

Question

- The estimator, $\hat{\sigma}^2$, is used to estimate the variance of a $N(\mu, \sigma^2)$ distribution based on a random sample of n observations:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

i. Determine the mean square error of $\hat{\sigma}^2$

ii. Determine whether $\hat{\sigma}^2$ is consistent.

12 Asymptotic distribution of MLEs

- Given a random sample of size n from a distribution with density (or probability function in the discrete case) $f(x;\theta)$, the maximum likelihood estimator $\hat{\theta}$ is such that, for large $n, \hat{\theta}$ is approximately normal, and is unbiased with variance given by the Cramér-Rao lower bound, that is:

$$\hat{\theta} \quad N(\theta, CRLB)$$

$$where\ CRLB = \frac{1}{nE\left\{\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right\}}$$

- The MLE can therefore be called asymptotically efficient in that, for large n , it is unbiased with a variance equal to the lowest possible value of unbiased estimators.

12 Asymptotic distribution of MLEs

- CRLB gives a lower bound for the variance of an unbiased estimator of a parameter (which is the same as its mean square error). So no unbiased estimator can have a smaller variance than the CRLB.

- This is potentially a very useful result as it provides an approximate distribution for the MLE when the true sampling distribution may be unknown or impossible to determine easily, and hence may be used to obtain approximate confidence intervals.

- Confidence intervals will be covered going further.

- The result holds under very general conditions with only one major exclusion: it does not apply in cases where the range of the distribution involves the parameter, such as the uniform distribution.

- This is due to a discontinuity, so the derivative in the formula doesn't make sense.

12 Asymptotic distribution of MLEs

- There are two useful alternative expressions for the CRLB based on the likelihood itself.
- Noting that $L(\theta)$ is really $L(\theta, \underline{X})$, these are:

$$\boldsymbol{CRLB = \frac{1}{E\left\{\left[\frac{\partial}{\partial\theta} logL(\theta, \underline{X})\right]^2\right\}}} \quad \boldsymbol{and} \quad \boldsymbol{CRLB = \frac{1}{-E\left[\frac{\partial^2}{\partial\theta^2} logL(\theta, \underline{X})\right]}}$$

Example

Question

- Derive the CRLB for estimators of the variance of a $N(\mu, \sigma^2)$ distribution, where μ is known, based on a random sample of n observations.

13 Comparing the method of moments with MLE

- We now compare the method of moments and the method of maximum likelihood.
- Essentially maximum likelihood is regarded as the better method.
- In the usual one-parameter case the method of moments estimator is always a function of the sample mean $\bar{X}$ and this must limit its usefulness in some situations. For example in the case of the uniform distribution on $(0,\theta)$ the method of moments estimator is $2\bar{X}$ and this can result in inadmissible estimates which are greater than θ.
- For example, supposing we had the following data from U$(0,\theta)$:

$$4.5, 1.8, 2.7, 0.9, 1.3$$

- This gives $\bar{x} = 2.24$. Since the method of moments estimator is $\hat{\theta} = 2\bar{X}$, we have $\hat{\theta} = 4.48$. But this estimate for the upper limit is inadmissible as one of the data values is greater than this.

13 Comparing the method of moments with MLE

- Nevertheless in many common applications such as the binomial, Poisson, exponential and normal cases both methods yield the same estimator.

- In some situations such as the gamma with two unknown parameters the simplicity of the method of moments gives it a possible advantage over maximum likelihood which may require a complicated numerical solution.

- To obtain the MLE of α from a gamma distribution requires the differentiation of $\Gamma(\alpha)$, which will require numerical methods.