Elements of statistics (MATH0487-1)

Prof. Dr. Dr. K. Van Steen

University of Liège, Belgium

October 15, 2012

Outline

Probability is the Science of Uncertainty

1. Rules $\rightarrow$ data: Given the rules, describe the likelihoods of various events occurring.
2. Probability is about prediction - looking forward.
3. Probability is mathematics.

Statistics is the Science of Data

1. Rules $\leftarrow$ data: Given only the data, try to guess what the rules were. That is, some probability model controlled what data came out, and the best we can do is guess - or approximate - what that model was. We might guess wrong; we might refine our guess as we get more data.
2. Statistics is about looking backward.
3. Statistics is an art. It uses mathematical methods, but it is more than maths.
4. Once we make our best *statistical guess* about what the probability model is (what the rules are), based on looking backward, we can then use that *probability model* to predict the *future* $\rightarrow$ The purpose of statistics is to make inference about unknown quantities from samples of data

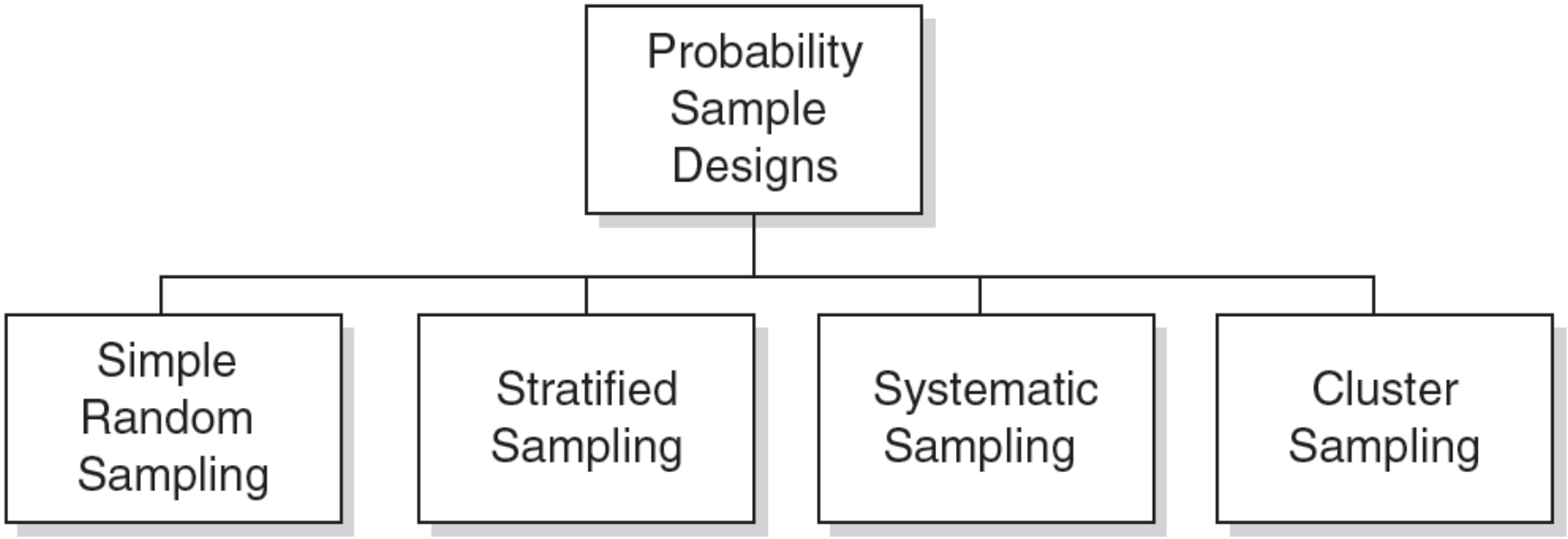
Statistics is the Science of Data

- **Sampling and experimentation**: Clarifying the question, deciding on methods of collection and analysis to produce valid information.
- **Exploring data**: Using graphical and numerical techniques to study patterns and departures from patterns (in order to interpreting data)
- **Anticipating patterns**: Exploring random phenomena using probability and simulation. Probability is our tool for anticipating distributions . . .
- **Statistical Inference**: Estimating population parameters and testing hypothesis.

Statistics is the Science of Data

- **Sampling and experimentation**: Clarifying the question, deciding on methods of collection and analysis to produce valid information.
- **Exploring data**: Using graphical and numerical techniques to study patterns and departures from patterns (in order to interpreting data)
- **Anticipating patterns**: Exploring random phenomena using probability and simulation. Probability is our tool for anticipating distributions . . .
- **Statistical Inference**: Estimating population parameters and testing hypothesis.

Major Types of Probability Sampling



Simple Random Sampling

- **Simple random sampling** is a probability sampling procedure that gives every element in the target population, and each possible sample of a given size, an equal chance of being selected. As such, it is an equal probability selection method (EPSEM).
  - *Sampling without replacement* tends to be *more efficient* than sampling with replacement in producing representative samples.
  - Sampling done without replacement is *no longer independent*, but still satisfies exchangeability: i.e., any order of a finite number of samples is equally likely.

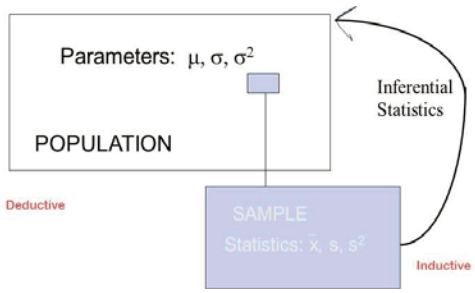
How important is exchangeability?

- For a small sample from a large population, sampling without replacement is approximately the same as sampling with replacement, since the odds of choosing the same individual twice is very low.

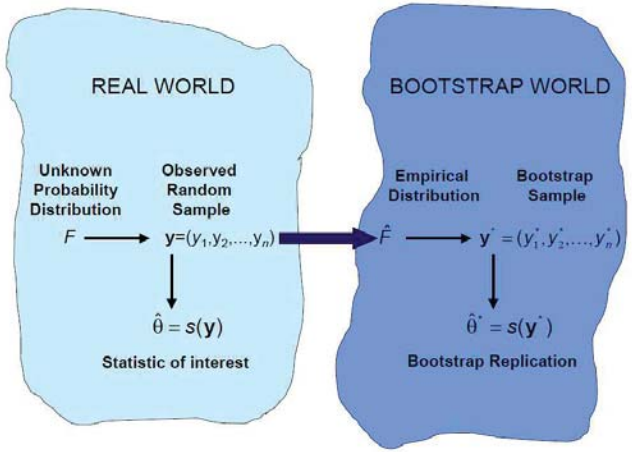
What about sampling from a finite population?

Using a Sample to Estimate the Truth

- Truth = Population parameter



- Truth = Population distribution function

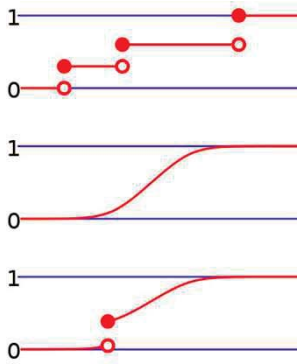


  - Use the empirical cdf $\hat{F}_n(y) = \frac{\text{nr of observations} \leq y}{n}$, or more formally
  - use $\hat{F}_n(y) = 1/n \sum_{i=1}^{n} I(y_i \leq y)$ as estimate of F, $I()$ denoting the indicator function, n reminding us that it is based on sample size n.

Comparing Distribution Functions

- Hence, the empirical cumulative distribution function is a step function that jumps for $1/n$ at each of the n data points.

  Shapes of cumulative distribution functions: From top to bottom, the cumulative distribution function of a discrete probability distribution, continuous probability distribution, and a distribution which has both a continuous part and a discrete part.



- The Kolmogorov-Smirnov (KS) test is based on quantifying a distance between cumulative distribution functions and can be used to test to see whether two empirical distributions are different or whether an empirical distribution is different from an ideal distribution (i.e. a reference distribution).
- The KS test is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

Sampling Distibutions

Statistic	Mean	Variance
$\overline{X}$	μ	$\frac{\sigma^2}{n}$
$\overline{X_1} - \overline{X_2}$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
p	P	$\frac{P(1-P)}{n}$
np	nP	$nP(1-P)$
$p_1 - p_2$	$P_1 - P_2$	$\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}$

- Why do we need sampling distributions?
- Answer: Sampling distributions allow us to make statements about the *unobserved true population parameter* in relation to the *observed sample statistic* $\longrightarrow$ **statistical inference**

Data Analysis Procedures

- For **classical analysis**, the data collection is followed by proposing a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.
- For a **Bayesian analysis**, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model: the analysis formally combines both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.
- For **EDA**, the data collection is not followed by a model imposition: it is followed immediately by analysis with a goal of inferring what model would be appropriate

EDA

- Exploratory Data Analysis was named by Tukey(1977) as an alternative to Confirmatory Data Analysis. It is an attitude or philosophy about how data analysis should be carried out, instead of being a fixed set of techniques.
- Important tasks within an EDA context include:
  - **Checking assumptions**: E.g., Is there multicollinearity in the data?
  - **Spotting outliers**: E.g., Are there "extreme" observations driving the results?
  - **Data transformations**: E.g., Which transformations ensure a better adherence to model assumptions, without loosing interpretability?
  - **Transparency and interpretability**: E.g., While "hypothesis" testers submit the data to complicated algorithms without understanding how exactly the "p-values" are computed, data visualizers could directly see the pattern on the graph.
  - **Resampling and validation**: E.g., EDA focuses on pattern recognition using the data at hand. How can you go beyond the initial sample to validate findings?

Approach to Estimating a Parameter I

- Consider the random experiment which consists of picking 3 people at random from the 2012 electoral register for Liege.
- The outcome of such an experiment will be a collection of 3 human beings and the set of all possible outcomes of this experiment Ω consists of all subsets of 3 human beings which may be formed from the set of all human beings whose names are on the register.
- Consider the random vector $\mathbf{X} = (X_1, X_2, X_3)$ where for $i = 1, 2, 3$, $X_i = 0$ if the ith person chosen is a male and $X_i = 1$ if the ith person chosen is a female.
- We furthermore assume that X_1, X_2, X_3 are *independent and identically distributed* or i.i.d. with $P(X_i = 1) = \theta$

  [i.e., in an infinite sequence of independent repetitions of the experiment the proportion of outcomes which produce, for instance, a value of $\mathbf{X} = (1, 1, 0)$ is given by $\theta^2(1 - \theta)$.]

Approach to Estimating a Parameter II

- Suppose that the value of θ is unknown and we propose to estimate it by the estimator $\hat{\theta}$ whose value is given by the proportion of females in the sample of size 3.
- Since $\hat{\theta}$ depends on the value of $\mathbf{X}$ we sometimes write $\hat{\theta}(\mathbf{X})$ to emphasise this fact.
- We can work out the probability distribution of $\hat{\theta}$ as follows :

$\mathbf{X}$	$P(\mathbf{X} = \mathbf{x})$	$\hat{\theta}(\mathbf{x})$
$(0,0,0)$	$(1-\theta)^3$	0
$(0,0,1)$	$\theta(1-\theta)^2$	$1/3$
$(0,1,0)$	$\theta(1-\theta)^2$	$1/3$
$(1,0,0)$	$\theta(1-\theta)^2$	$1/3$
$(0,1,1)$	$\theta^2(1-\theta)$	$2/3$
$(1,0,1)$	$\theta^2(1-\theta)$	$2/3$
$(1,1,0)$	$\theta^2(1-\theta)$	$2/3$
$(1,1,1)$	θ^3	1

How good is an Estimator? I

- If $\theta = 0$ we have that $P(\hat{\theta} = \theta) = P(\hat{\theta} = 0) = 1$ which is good.
- Likewise if $\theta = 1$ we also have that $P(\hat{\theta} = \theta) = P(\hat{\theta} = 1) = 1$.
- If $\theta = 1/3$ then $P(\hat{\theta} = \theta) = P(\hat{\theta} = 1/3) = 3(1/3)(1 - 1/3)^2 = 4/9$.
- However if the value of θ lies outside the set $\{0, 1/3, 2/3, 1\}$ we have that $P(\hat{\theta} = \theta) = 0$.
- Since $\hat{\theta}$ is a random variable we might try to calculate its expected value $\mathrm{E}(\hat{\theta})$ i.e. the average value we would get if we carried out an infinite number of independent repetitions of the experiment.
- We have that

$$\begin{aligned} \mathrm{E}(\hat{\theta}) &= 0P(\hat{\theta} = 0) + (1/3)P(\hat{\theta} = 1/3) + (2/3)P(\hat{\theta} = 2/3) + 1P(\hat{\theta} = 1) \ , \\ &= 0(1-\theta)^3 + (1/3)3\theta(1-\theta)^2 + (2/3)3\theta^2(1-\theta) + 1\theta^3 \ , \\ &= \theta. \end{aligned}$$

How good is an Estimator? II

- Thus if we carried out an infinite number of independent repetitions of the experiment and calculate the value of $\hat{\theta}$ for each repetition the average of the $\hat{\theta}$ values would be exactly θ, the true value of the parameter! This is true no matter what the actual value of θ is. Such an estimator is said to be **unbiased**.

How good is an Estimator?

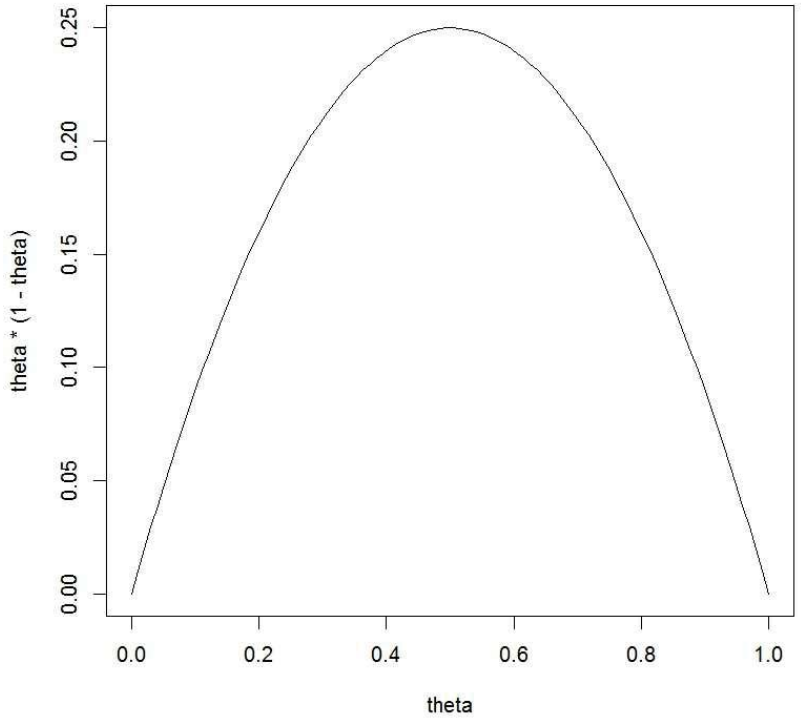
- Consider the quantity $L = (\hat{\theta} - \theta)^2$ which might be regarded as a measure of the error or loss involved in using $\hat{\theta}$ to estimate θ. The possible values for L are $(0-\theta)^2$, $(1/3-\theta)^2$, $(2/3-\theta)^2$ and $(1-\theta)^2$.
- We can calculate the expected value of L as follows:

$$\begin{aligned} \mathrm{E}(L) &= (0-\theta)^2 P(\hat{\theta}=0) + (1/3-\theta)^2 P(\hat{\theta}=1/3) \\ &\qquad + (2/3-\theta)^2 P(\hat{\theta}=2/3) + (1-\theta)^2 P(\hat{\theta}=1) \\ &= \theta^2(1-\theta)^3 + (1/3-\theta)^2 3\theta(1-\theta)^2 + (2/3-\theta)^2 3\theta^2(1-\theta) + (1-\theta)^2\theta^3 \ , \\ &= \theta(1-\theta)/3 \ . \end{aligned}$$

- The quantity $\mathrm{E}(L)$ is called the **mean squared error** (MSE) of the estimator $\hat{\theta}$.

How good is an Estimator?

- Since the quantity $\theta(1-\theta)$ attains its maximum value of $1/4$ for $\theta = 1/2$, the largest value $\mathrm{E}(L)$ can attain is $1/12$ which occurs if the true value of the parameter θ happens to be equal to $1/2$.
- For all other values of θ the quantity $\mathrm{E}(L)$ is less than $1/12$.
- If somebody could invent a different estimator $\tilde{\theta}$ of θ whose MSE was less than that of $\hat{\theta}$ for *all* values of θ then we would prefer $\tilde{\theta}$ to $\hat{\theta}$.



Summarizing our Thoughts about Estimation

- The **basic frequentist principle** is that statistical procedures should be judged in terms of their *average* performance in an infinite series of independent repetitions of the experiment which produced the data.
- Using different methods of estimation can lead to different estimators.
- An important point to note is that the parameter values are treated as fixed (although unknown) throughout this infinite series of repetitions.
- We should be happy to use a procedure which performs well on the average and should not be concerned with how it performs on any one particular occasion.

Frequentists are Unbiased Estimators

- **The frequentist philosophy**: to evaluate the usefulness of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{x})$ of θ, examine the properties of the random variable $\hat{\theta} = \hat{\theta}(\mathbf{X})$.
- Criteria for deciding which are good estimators are required:
  - As a first condition it seems reasonable to ask that the distribution of the estimator be centered around the parameter it is estimating. If not it will tend to overestimate or underestimate θ.
  - A second property an estimator should possess is precision. An estimator is precise if the dispersion of its distribution is small.
- The two concepts above are incorporated in the definitions of "unbiasedness" and "efficiency".

Frequentists are Unbiased Estimators

- **Unbiased estimators**: An estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is said to be unbiased for a parameter θ if it equals θ in expectation:

$$\mathrm{E}[\hat{\theta}(\mathbf{X})] = \mathrm{E}(\hat{\theta}) = \theta.$$

  Intuitively, an unbiased estimator is "right on target".

- **Bias of an estimator**: The bias of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ of θ is defined as

$$\mathrm{bias}(\hat{\theta}) = \mathrm{E}[\hat{\theta}(\mathbf{X}) - \theta].$$

  There may be large number of unbiased estimators of a parameter for any given distribution and a further criterion for choosing between all the unbiased estimators is needed.

- **Bias corrected estimators**: If $\mathrm{bias}(\hat{\theta})$ is of the form $c\theta$, then (obviously) $\tilde{\theta} = \hat{\theta}/(1+c)$ is unbiased for θ. Likewise, if $\mathrm{bias}(\hat{\theta}) = \theta + c$, then $\tilde{\theta} = \hat{\theta} - c$ is unbiased for θ. In such situations we say that $\tilde{\theta}$ is a biased corrected version of $\hat{\theta}$.

Frequentists are Unbiased Estimators

- **Unbiased functions**: More generally $\hat{g}(\mathbf{X})$ is said to be unbiased for a function $g(\theta)$ if $\mathrm{E}[\hat{g}(\mathbf{X})] = g(\theta)$.
- Even if $\hat{\theta}$ is an unbiased estimator of θ, $g(\hat{\theta})$ will generally not be an unbiased estimator of $g(\theta)$ unless g is linear or affine.
- This limits the importance of the notion of unbiasedness.
- Also, it might be at least as important that an estimator is accurate in the sense that its distribution is highly concentrated around θ.

Is Unbiasedness a Good Thing?

- Unbiasedness is important when combining estimates, as averages of unbiased estimators are unbiased.
- Example:
  - When combining standard deviations $s_1, s_2, \ldots, s_k$ with degrees of freedom $\mathrm{df}_1, \ldots, \mathrm{df}_k$ we always average their squares

$$\bar{s} = \sqrt{\frac{\mathrm{df}_1 s_1^2 + \cdots + \mathrm{df}_k s_k^2}{\mathrm{df}_1 + \cdots + \mathrm{df}_k}}$$

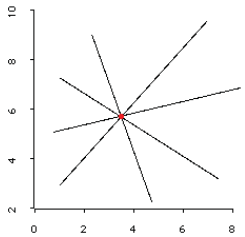
    as s_i^2 are unbiased estimators of the variance σ^2, whereas s_i are not unbiased estimators of σ.

Therefore, be careful when averaging biased estimators! It may well be appropriate to make a bias-correction before averaging.

A Note Aside

Degrees of freedom (df) in terms of sample size:

- Toothaker (1986) explained df as the number of independent components (n) minus the number of parameters (r) estimated (or more vaguely, the number of "relationships").
- In the scatterplot below where there is only one datum point, the analyst cannot do any estimation of the regression line because the line can go in any direction.
- When the degree of freedom is zero ($\text{df} = n - r = 1 - 1 = 0$), there is no way to affirm or reject the model! In this sense, the data have no "freedom" to vary and you don't have any "freedom" to conduct research with this data set.



What is the total df for a general linear regression problem with n independent data points?

A Note Aside

Degrees of freedom (df) in terms of dimensionality:

- According to I. J. Good (1973) [What are degrees of freedom? American Statisticians, 27, 227-228], degrees of freedom can be expressed as

$$D(K) - D(H),$$

whereas

- $D(K)$ = the dimensionality of a broader hypothesis, such as a full model in regression
- $D(H)$ = the dimensionality of the null hypothesis, such as a restricted or null model

Types of Estimators

- Focus on finding **point estimators** first, i.e. for which the true value of a (function of a) parameter is assumed to be a point.
- Several methods exist to compute point estimators, including the "methods of moments" and "maximum likelihood", but also the "method of least squares" (see Regression Analysis chapter), etc.
- Second, focus on finding **interval estimators**, i.e. acknowledge the utility for some interval about the point estimate together with some measure of accurance that the true value of the parameter lies within the interval.

Inference choices:
1) making the inference of estimating the true value of the parameter to be a point,
2) making the inference of estimating that the true value of the parameter is contained in some interval.

Sample Moments as Estimators

- For a random variable X, the rth moment about the origin 0, or the rth moment of its corresponding density function is defined as $\mu'_r = E(X^r)$.
- For a random sample $X_1, X_2, \ldots, X_n$, the rth sample moment about the origin is defined by

$$M_r = \sum_{i=1}^{n} X_i^r / n, r = 1, 2, 3, \ldots$$

and its observed value is denoted by

$$m_r = \sum_{i=1}^{n} x_i^r / n.$$

- The following property of sample moments holds:

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample of X. Then

$$E(M_r) = \mu'_r, r = 1, 2, 3, \ldots$$

Definitions

- The sample moments, $M_1, M_2, \ldots$, are random variables whose means are $\mu'_1, \mu'_2, \ldots$
- Since the population moments depend on the parameters of the distribution, estimating them by the sample moments leads to estimation of the parameters.
- When using sample moments to estimate population parameters, the resulting estimators are called **method of moments estimators** or MMEs.

The MME Procedure

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $F(x : \theta_1, \ldots, \theta_k)$. Hence, suppose that there are k parameters to be estimated.
- Let μ'_r, m_r $(r = 1, 2, \ldots, k)$ denote the first k population and sample moments respectively.
- Suppose that each of these population moments are certain *known* functions of the parameters:

$$\begin{aligned} \mu'_1 &= g_1(\theta_1, \ldots, \theta_k), \\ \mu'_2 &= g_2(\theta_1, \ldots, \theta_k), \\ &\vdots \\ \mu'_k &= g_k(\theta_1, \ldots, \theta_k). \end{aligned}$$

- Solving simultaneously the set of equations,

$$g_r(\overline{\theta_1}, \ldots, \overline{\theta_k}) = m_r, r = 1, 2, \ldots, k$$

gives the required estimates $\overline{\theta_1}, \ldots, \overline{\theta_k}$.

The MME Procedure Applied

- For the normal distribution, we know that $E(X) = \mu$ and $\sigma^2 = E(X^2) - \mu^2$.
- The unknown parameters to estimate are μ and σ^2
- Let μ'_r, m_r $(r = 1, 2)$ denote the first k population and sample moments. The population moments are known functions of these population parameters.
- Equate:

$$E(X) \to \mu'_1 = \mu,$$
$$E(X^2) \to \mu'_2 = \sigma^2 + \mu^2.$$

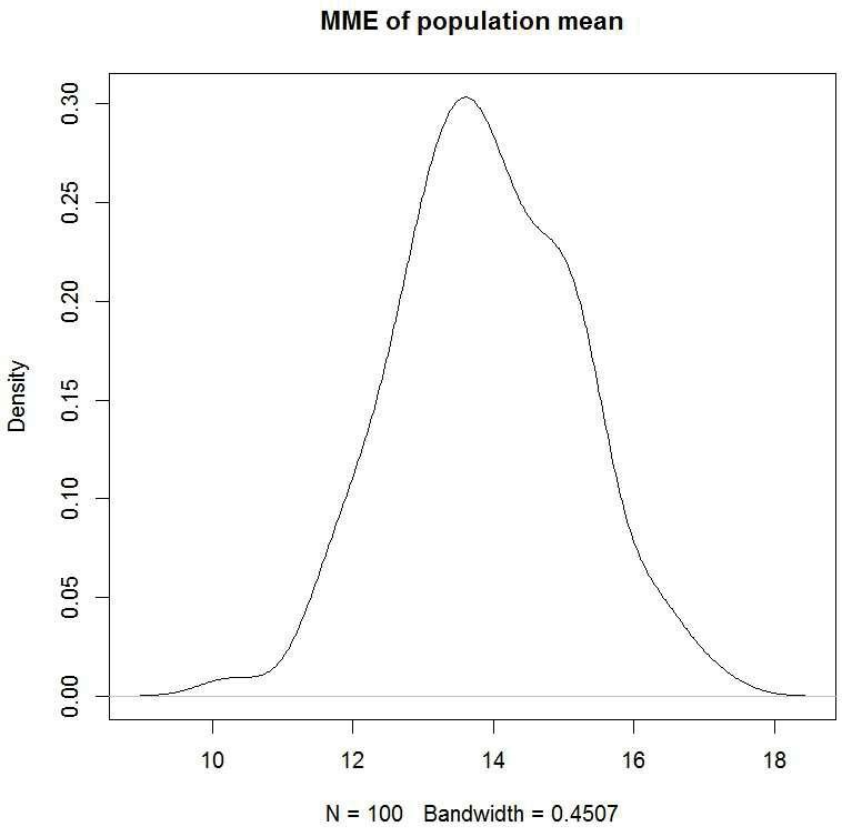
- Solving simultaneously the set of equations, gives

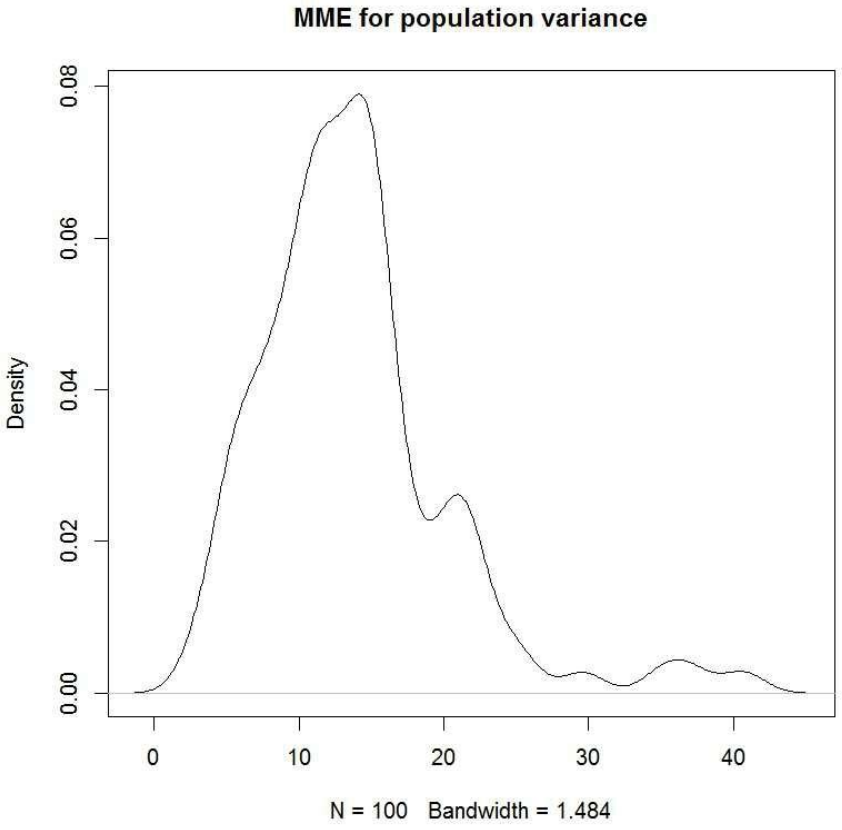
$$\overline{\mu} = \overline{x}, \text{ and } \overline{\sigma}^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2.$$

The MME Procedure Applied

```
#_____________NormalMoments.R ___________
set.seed(69)
mu <- 14
sigma <- 4
sampsz <- 10
nsimulations <- 100
mu.estimates <- numeric(nsimulations)
var.estimates <- numeric(nsimulations)
for (i in 1:nsimulations){
rn <- rnorm(mean=mu,sd=sigma,n=sampsz)

## computing MMEs
mu.estimates[i] <- mean(rn)
var.estimates[i] <- mean( (rn -mean(rn))^2 )

} # end of i loop

plot(density(mu.estimates),main="MME of population mean")
plot(density(var.estimates),main="MME for population variance")
```

The MME Procedure Applied





Generalized Methods of Moments

- The MM only works when the number of moment conditions equals the number of parameters to estimate
- If there are more moment conditions than parameters, the system of equations is algebraically over identified and cannot be solved [E.g., when estimating the slope of a regression line through the origin]
- Generalized method-of-moments (GMM) estimators choose the estimates that minimize a quadratic form of the sample moment conditions
  - GMM gets as close to solving the over-identified system of sample moment equations as possible
  - GMM reduces to MM when the number of parameters equals the number of moment conditions
- Hansen (1982) produced many of the key results; Wooldridge (2002); Cameron and Trivedi (2005) provide good introductions

Desirable Properties of an Estimator

Unbiasedness		ν
Trading off Bias and Variance	MSE MVUE	
Efficiency		
Consistency		
Sufficiency		

Mean-Squared Error

- Although desirable from a frequentist standpoint, unbiasedness is not a property that helps us choose between estimators.
- To do this we must examine some measure of loss like the mean squared error.
- The **mean squared error** of the estimator $\hat{\theta}$ is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2].$$

- Given the same set of data, $\hat{\theta}_1$ is "better" than $\hat{\theta}_2$ if

$$MSE(\hat{\theta}_1) \leq \mathrm{MSE}(\hat{\theta}_2)$$

  or *uniformly better* if true $\forall\, \theta$.
- The problem of finding minimum MSE estimators cannot be solved uniquely ...

Mean-Squared Error

Lemma (The MSE variance-bias tradeoff)

The MSE decomposes as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2.$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}(\hat{\theta} - \theta)^2 \\
&= \mathrm{E}\{\,[\,\hat{\theta} - \mathrm{E}(\hat{\theta})\,] + [\,\mathrm{E}(\hat{\theta}) - \theta\,]\}^2 \\
&= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]^2 + \mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta]^2 \\
&\qquad +2\underbrace{\mathrm{E}\left\{[\hat{\theta} - \mathrm{E}(\hat{\theta})][\mathrm{E}(\hat{\theta}) - \theta]\right\}}_{=0} \\
&= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]^2 + \mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta]^2 \\
&= \mathrm{Var}(\hat{\theta}) + \underbrace{[\mathrm{E}(\hat{\theta}) - \theta]^2}_{\mathrm{Bias}(\hat{\theta})^2}
\end{aligned}$$

□

Mean-Squared Error Applied

Problem

Consider $X_1, \ldots, X_n$ where $X_i \sim \mathrm{N}(\theta, \sigma^2)$ and σ is known. Three estimators of θ are $\hat{\theta}_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\hat{\theta}_2 = X_1$, and $\hat{\theta}_3 = (X_1 + \bar{X})/2$. Pick one.

- All three estimators are unbiased.
- For a class of estimators that are unbiased, the mean squared error will be equal to the estimation variance.
- Calculate the corresponding variances:
  - $\mathrm{Var}(\hat{\theta}_1) = \frac{1}{n^2}[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)] = \frac{1}{n^2}[\sigma^2 + \cdots + \sigma^2] = \frac{1}{n^2}[n\sigma^2] = \frac{1}{n}\sigma^2$.
  - $\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}(X_1) = \sigma^2$.
  - $\mathrm{Var}(\hat{\theta}_3) = (\sigma^2/n + \sigma^2)/4 + \mathrm{Cov}(\bar{X}, X_1)/2$.

  [Note that for any two random variables X and Y, $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.]
- Therefore $\bar{X}$ appears "best" in the sense that $\mathrm{Var}(\hat{\theta})$ is smallest among these three unbiased estimators.

Mean-Squared Error Applied

Problem

Consider $X_1, \ldots, X_n$ to be independent random variables with means $\mathrm{E}(X_i) = \mu + \beta_i$ and variances $\mathrm{Var}(X_i) = \sigma_i^2$.

Such a situation could arise when X_i are estimators of μ obtained from independent sources and β_i is the bias of the estimator X_i.

Consider pooling the estimators of μ into a common estimator using the linear combination $\hat{\mu} = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n$.

What is the most optimal linear combination?

Mean-Squared Error Applied

- If the estimators are unbiased, $\hat{\mu}$ is unbiased if and only if $\sum w_i = 1$.

Proof.

$\mathrm{E}(\hat{\mu}) = \mathrm{E}(w_1 X_1 + \cdots + w_n X_n) = \sum_i w_i \mathrm{E}(X_i) = \sum_i w_i \mu = \mu \sum_i w_i$ so $\hat{\mu}$ is unbiased if and only if $\sum_i w_i = 1$. □

- If the estimators are unbiased, $\hat{\mu}$ has minimum variance when the weights are inversely proportional to the variances σ_i^2.

Proof.

The variance of our estimator is $\mathrm{Var}(\hat{\mu}) = \sum_i w_i^2 \sigma_i^2$, which should be minimized subject to the constraint $\sum_i w_i = 1$. Differentiating the Lagrangian $\mathcal{L} = \sum_i w_i^2 \sigma_i^2 - \lambda \left(\sum_i w_i - 1 \right)$ with respect to w_i and setting equal to zero yields $2 w_i \sigma_i^2 = \lambda \Rightarrow w_i \propto \sigma_i^{-2}$ so that $w_i = \sigma_i^{-2} / (\sum_j \sigma_j^{-2})$. □

Mean-Squared Error Applied

- The variance of $\hat{\mu}$ for optimal weights w_i is $\mathrm{Var}(\hat{\mu}) = 1/\sum_i \sigma_i^{-2}$.

Proof.

For optimal weights we get
$\mathrm{Var}(\hat{\mu}) = \sum_i w_i^2 \sigma_i^2 = (\sum_i \sigma_i^{-4} \sigma_i^2)/(\sum_i \sigma_i^{-2})^2 = 1/(\sum_i \sigma_i^{-2})$. □

- With the optimal linear combination obtained above, and assuming $\sigma_i^2 = \sigma^2$, the bias tends to dominate the variance as n gets larger, which is very unfortunate ...

Proof.

When $\sigma_i^2 = \sigma^2$ we have that $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$ which tends to zero for $n \to \infty$ whereas $\mathrm{bias}(\hat{\mu}) = \sum \beta_i/n = \bar{\beta}$ is equal to the average bias and $\mathrm{MSE}(\hat{\mu}) = \sigma^2/n + \bar{\beta}^2$. □

Uniformly minimum-variance unbiased estimator - UMVUE

- When no bias-variance trade-off can be made, one approach is to restrict ourselves to the subclass of estimators that are *unbiased* and *minimum variance*
- Let $X_1, \ldots, X_n$ be a random sample from $f(.;\theta)$. An estimator $T^* = t^*(X_1, \ldots, X_n)$ of $\tau(\theta)$ is said to be a **uniformly minimum-variance unbiased estimator** of $\tau(\theta)$ if and only if
  - $\mathrm{E}(T^*) = \tau(\theta)$
  - $\mathrm{Var}_\theta(T^*) \leq \mathrm{Var}_\theta(T)$, for any other estimator $T = t(X_1, \ldots, X_n)$ of $\tau(\theta)$ which satisfies $\mathrm{E}(T) = \tau(\theta)$
- In words: If an unbiased estimator of $g(\theta)$ has minimum variance among all unbiased estimators of $g(\theta)$ it is called a minimum variance unbiased estimator (MVUE).

How to find MVUE when it exists?

- For the *(possibly vector valued)* observation $X = x$ to be informative about θ, the density must vary with θ. If $f(x|\theta)$ is smooth and differentiable, this change is quantified to first order by the **score function**

$$S(\theta) = \frac{\partial}{\partial\theta} \ln f(x|\theta) \equiv \frac{f'(x|\theta)}{f(x|\theta)}.$$

- Under suitable regularity conditions (differentiation wrt θ and integration wrt x can be interchanged), we have

$$\begin{aligned} \mathrm{E}\{S(\theta)\} &= \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)\mathrm{d}x = \int f'(x|\theta)\mathrm{d}x \,, \\ &= \frac{\partial}{\partial\theta}\left\{\int f(x|\theta)\mathrm{d}x\right\} = \frac{\partial}{\partial\theta} 1 = 0. \end{aligned}$$

[. . . average across all possible samples . . .]

How to find MVUE when it exists?

- Variance measures lack of knowledge.
- So it is reasonable to say that the reciprocal of the variance should be defined as the amount of information carried by the observation x about θ.

How to find MVUE when it exists? I

Lemma (Fisher information)

The variance of $S(\theta)$ is the expected Fisher information about θ

$$\mathrm{E}(\mathcal{I}(\theta)) = \mathrm{E}\{S(\theta)^2\} \equiv \mathrm{E}\left\{\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2\right\}$$

How to find MVUE when it exists? II

Expected Fisher Information.

Using the chain rule

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\ln f &= \frac{\partial}{\partial\theta}\left[\frac{1}{f}\frac{\partial f}{\partial\theta}\right]\\ &= -\frac{1}{f^2}\left[\frac{\partial f}{\partial\theta}\right]^2+\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\\ &= -\left[\frac{\partial \ln f}{\partial\theta}\right]^2+\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\end{aligned}$$

If integration and differentiation can be interchanged

$$\mathrm{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\right]=\int_X\frac{\partial^2 f}{\partial\theta^2}\mathrm{d}x=\frac{\partial^2}{\partial\theta^2}\int_X\mathrm{d}x=\frac{\partial^2}{\partial\theta^2}\,1\;=\;0,$$

thus

$$-\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right]=\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2\right]=\mathrm{E}(\mathcal{I}(\theta)).$$

How to find a lower bound for MVUE?

Theorem (Cramér Rao Lower Bound - CRLB)

Let $\hat{\theta}$ be an unbiased estimator of θ. Then

$$\mathrm{Var}(\hat{\theta}) \geq \{ \mathrm{E}(\mathcal{I}(\theta)) \}^{-1}.$$

- Remarks:
  - The proof requires that the limits of the integrals do not depend on θ.
  - This condition is violated for many density functions, i.e. the CRLB is not valid for the uniform distribution.
  - We can have absolute assessment for unbiased estimators by comparing their variances to the CRLB. Biased estimators will be considered good, if their variances are lower than the CRLB.

How to find a lower bound for MVUE?

- Remarks:
  - In some textbooks you will find that $\mathrm{E}\{S(\theta)^2\}$ is called Fisher information instead of expected Fisher information.
  - Throughout this course, we will call $S(\theta)^2$ **observed Fisher Information** $\mathcal{I}(\theta)$ and $\mathrm{E}\{S(\theta)^2\}$ (logically) **expected Fisher information** $\mathrm{E}(\mathcal{I}(\theta))$.

Proof of the CRLB

Theorem (Cramér Rao Lower Bound - CRLB)

Let $\hat{\theta}$ be an unbiased estimator of θ. Then

$$\mathrm{Var}(\hat{\theta}) \;\geq\; \{\, \mathrm{E}(\mathcal{I}(\theta)) \,\}^{-1}.$$

Proof.

Unbiasedness, $\mathrm{E}(\hat{\theta}) = \theta$, implies

$$\int \hat{\theta}(x) f(x|\theta) \mathrm{d}x = \theta.$$

Assume we can differentiate wrt θ under the integral, then

$$\int \frac{\partial}{\partial \theta} \left\{ \hat{\theta}(x) f(x|\theta) \mathrm{d}x \right\} = 1, \textit{and}$$

$$\int \hat{\theta}(x) \frac{\partial}{\partial \theta} \left\{ f(x|\theta) \mathrm{d}x \right\} = 1,$$

since the estimator $\hat{\theta}(x)$ cannot depend on θ.

Proof of the CRLB

Proof.

TRICK: For any pdf f,

$$\frac{\partial f}{\partial \theta} = f \frac{\partial}{\partial \theta} (\ln f),$$

so that now

$$\int \hat{\theta}(x) f \frac{\partial}{\partial \theta} (\ln f) \, \mathrm{d}x = 1,$$

and

$$\mathrm{E}\left[\hat{\theta}(x) \frac{\partial}{\partial \theta} (\ln f)\right] = 1.$$

This implies that, with new random variables $U = \hat{\theta}(x)$ and $S = \frac{\partial}{\partial \theta} (\ln f)$

$$\mathrm{E}(US) = 1.$$

Proof of the CRLB

Proof.

We already know that the score function has expectation zero, $\mathrm{E}(S) = 0$.
Consequently $\mathrm{Cov}(U, S) = \mathrm{E}(US) - \mathrm{E}(U)\mathrm{E}(S) = \mathrm{E}(US) = 1$.

$$\{\mathrm{Corr}(U, S)\}^2 = \frac{\{\mathrm{Cov}(U, S)\}^2}{\mathrm{Var}(U)\mathrm{Var}(S)} \leq 1$$

Setting $\mathrm{Cov}(U, S) = 1$ we get

$$\mathrm{Var}(U)\mathrm{Var}(S) \geq 1,$$

which implies

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathrm{E}(\mathcal{I}(\theta))}.$$

General form of the CRLB - function of a parameter

Theorem (General Cramér Rao Lower Bound - CRLB)

Let $H = h(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$, then under the appropriate regularity conditions (see below),

$$\mathrm{Var}(H) \geq \frac{(\tau'(\theta))^2}{\mathrm{E}(\mathcal{I}(\theta))}.$$

- In the above expression, it should be noted that the expected Fisher information $\mathrm{E}(\mathcal{I}(\theta))$ is computed on a sample at hand: $x_1, \ldots, x_n$.

Regularity Conditions for CRLB

- Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$, where $X_1, \ldots, X_n$ represents a random sample from the density $f(.;\theta)$ and θ belongs to the parameter space, a subset of the real line.
- In listing the regularity conditions for CRLB, we consider the case of a (probability) density function, although the development for discrete density functions is similar:
  - $\frac{\partial}{\partial\theta} \ln f(x;\theta)$ exists for all x and all θ
  - $\frac{\partial}{\partial\theta} \int \ldots \int \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n = \int \ldots \int \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n$
  - $\frac{\partial}{\partial\theta} \int \ldots \int t(x_1, \ldots, x_n) \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n =$

    $\int \ldots \int t(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n$
  - $0 < \mathrm{E}[(\frac{\partial}{\partial\theta} \ln f(X;\theta))^2] < \infty$ for all acceptable θ

CRLB by Example I

- Consider i.i.d. random variables $X_i,\ i = 1, \ldots, n$, with

$$f_{X_i}(x_i|\mu) = \frac{1}{\mu} \exp\left(-\frac{1}{\mu} x_i\right),$$

[i.e., exponential distribution with parameter $\frac{1}{\mu}$]

- Denote the joint distribution of $X_1, \ldots, X_n$ by

$$f = \prod_{i=1}^{n} f_{X_i}(x_i|\mu) = \left(\frac{1}{\mu}\right)^n \exp\left(-\frac{1}{\mu} \sum_{i=1}^{n} x_i\right),$$

so that

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

CRLB by Example II

- Then the score function is

$$S(\mu) = \frac{\partial}{\partial \mu} \ln f = -\frac{n}{\mu} + \frac{1}{\mu^2} \sum_{i=1}^{n} x_i$$

and

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

- For $X \sim \mathrm{Exp}(1/\mu)$, we have $\mathrm{E}(X) = \mu$ implying $\mathrm{E}(X_1 + \cdots + X_n) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = n\mu$ and $\mathrm{E}\{S(\mu)\} = 0$ as required.

CRLB by Example III

- The variance of the score function is

$$\begin{aligned}
\mathrm{E}(\mathcal{I}(\mu)) &= \text{“} - \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right]'' \\
&= -\mathrm{E}\left\{\frac{\partial}{\partial\mu}\left(-\frac{n}{\mu}+\frac{1}{\mu^2}\sum_{i=1}^{n}x_i\right)\right\} \\
&= -\mathrm{E}\left\{\frac{n}{\mu^2}-\frac{2}{\mu^3}\sum_{i=1}^{n}x_i\right\} \\
&= -\frac{n}{\mu^2}+\frac{2}{\mu^3}\mathrm{E}\left\{\sum_{i=1}^{n}x_i\right\} \\
&= -\frac{n}{\mu^2}+\frac{2n\mu}{\mu^3} = \frac{n}{\mu^2}
\end{aligned}$$

- Hence, $\mathrm{CRLB} = \frac{\mu^2}{n}$.

CRLB by Example

- Now propose $\hat{\mu} = \bar{X}$ as an estimator of μ.
- For $X \sim \mathrm{Exp}(1/\mu)$ we have $\mathrm{E}(X) = \mu$, so

$$\mathrm{E}(\hat{\mu}) = \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} = \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n} X_i\right\} = \mu,$$

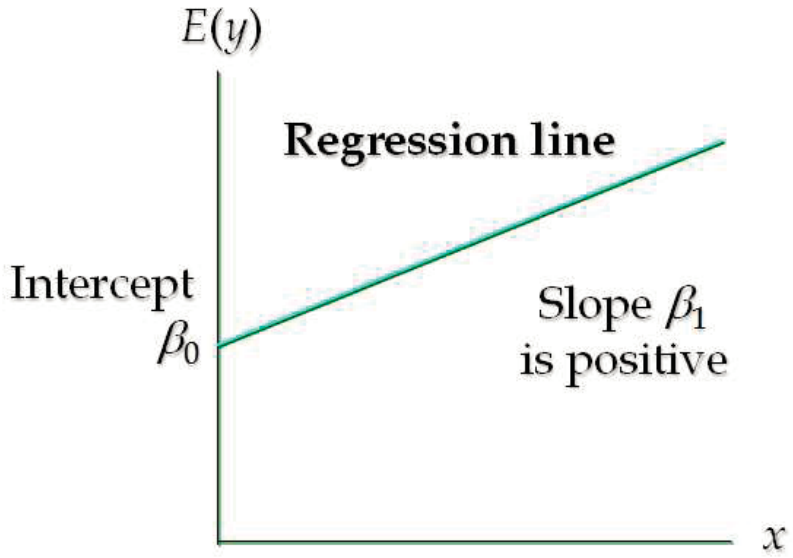
  verifying that $\hat{\mu} = \bar{X}$ is indeed an unbiased estimator of μ.
- For $X \sim \mathrm{Exp}(1/\mu)$ we also have $\mathrm{Var}(X) = \mu^2$, implying

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{n\mu^2}{n^2} = \frac{\mu^2}{n}.$$

- Hence, $\mathrm{Var}(\hat{\mu}) = \{\ \mathrm{E}(\mathcal{I}(\theta))\ \}^{-1}$, and our unbiased estimator $\hat{\mu} = \bar{x}$ achieves its CRLB.

Variance-Bias Decomposition for Regression

- Assume that $Y = \beta_0 + \beta_1 X + \epsilon$, with ϵ a random variable called the error-term, and β_0, β_1 parameters, is a **simple linear regression model**.
- In such a model, it is assumed that the expectation of Y given X is $E(Y) = \beta_0 + \beta_1 X$ (see later).



Variance-Bias Decomposition for Regression

- We can rewrite the aforementioned model as

$$y = F(x) + \epsilon,$$

  where ϵ is additive white noise with variance σ^2.
- In a general context, noise does not have to be Gaussian, but does have to be white: for any x_0, $F(x_0) = E_{[y|x]}(y0|x0)$.
- The "expected loss" with a predictor $\hat{f}$:

$$\begin{aligned} E_{\mathbf{x},y}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] &= E_{\mathbf{x},y}\left[\left(y_0 - F(\mathbf{x}_0) + F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] \\ &= E_{\mathbf{x},y}\left[(y_0 - F(\mathbf{x}_0))^2\right] \\ &\quad + E_{\mathbf{x},y}\left[\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] \\ &\quad + 2E_{\mathbf{x},y}\left[(y_0 - F(\mathbf{x}_0))\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)\right] \end{aligned}$$

- Note that the predictor $\hat{f}$ depends on the data at hand, and hence it makes sense to take $E_X[]$ of it.

Variance-Bias Decomposition for Regression

- For the last term:

$$
\begin{aligned}
E_{\mathbf{x},y}\left[(y_0 - F(\mathbf{x}_0))\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)\right] & \\
&= \int\int (y_0 - F(\mathbf{x}_0))\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right) p(y_0|\mathbf{x}_0)p(\mathbf{x}_0)dy_0 d\mathbf{x}_0 \\
&= \int \left\{E_{y|\mathbf{x}}\left[(y_0 - F(\mathbf{x}_0))\right]\right\}\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right) p(\mathbf{x}_0)d\mathbf{x}_0 \\
&= 0,
\end{aligned}
$$

- Using the notation $\bar{f}(x_0) = E_X[\hat{f}(x_0)]$:

$$
\begin{aligned}
\left(F(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2 &= \left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0) + \bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2 \\
&= \left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)^2 \\
&+ \left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2 \\
&+ 2\left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)\left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)
\end{aligned}
$$

Variance-Bias Decomposition for Regression

- Taking the expections $E_X[]$ of the terms on the previous slide, involves taking the expectation wrt X of the last term on the right-hand side:

$$\begin{aligned} E_X\left[2\left(F(\mathbf{x}_0)-\bar{f}(\mathbf{x}_0)\right)\left(\bar{f}(\mathbf{x}_0)-\hat{f}(\mathbf{x}_0)\right)\right] &= 2\left(F(\mathbf{x}_0)-\bar{f}(\mathbf{x}_0)\right)E_X\left[\left(\bar{f}(\mathbf{x}_0)-\hat{f}(\mathbf{x}_0)\right)\right] \\ &= 0, \end{aligned}$$

- and thus:

$$\begin{aligned} E_X\left[\left(F(\mathbf{x}_0)-\hat{f}(\mathbf{x}_0)\right)^2\right] &= \underbrace{\left(F(\mathbf{x}_0)-\bar{f}(\mathbf{x}_0)\right)^2}_{\text{bias}^2} \\ &\quad + \underbrace{E_X\left[\left(\bar{f}(\mathbf{x}_0)-\hat{f}(\mathbf{x}_0)\right)^2\right]}_{\text{variance}}. \end{aligned}$$

Variance-Bias Decomposition for Regression

- Putting it all together, we have the following decomposition in "regression context":

$$E_{X,\mathbf{x}_0,y_0}\left[\left(y_0 - \hat{f}(\mathbf{x}_0, X)\right)^2\right] = \sigma^2 \qquad \text{noise variance}$$

$$+ \int \left(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0)\right)^2 p(\mathbf{x}_0) d\mathbf{x}_0 \qquad \text{expected squared bias}$$

$$+ \int E_X\left[\left(\bar{f}(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0)\right)^2\right] p(\mathbf{x}_0) d\mathbf{x}_0 \qquad \text{expected variance.}$$

Efficiency

- Define the **efficiency** of the unbiased estimator $\hat{\theta}$ as

$$\text{eff}(\hat{\theta}) = \frac{\text{CRLB}}{\text{Var}(\hat{\theta})},$$

  where $\text{CRLB} = \{ \text{E}(\mathcal{I}(\theta)) \}^{-1}$. Clearly $0 < \text{eff}(\hat{\theta}) \leq 1$.
- An unbiased estimator $\hat{\theta}$ is said to be **efficient** if $\text{eff}(\hat{\theta}) = 1$.
- The **asymptotic efficiency** of an unbiased estimator $\hat{\theta}$ is the limit of the efficiency as $n \to \infty$.

Efficiency

- An unbiased estimator $\hat{\theta}$ is said to be **asymptotically efficient** if its asymptotic efficiency is equal to 1.
- Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be 2 unbiased estimators of θ with variances $Var(\widehat{\theta_1})$, $Var(\widehat{\theta_2})$ respectively. We can then say that $\widehat{\theta_1}$ is *more efficient* than $\widehat{\theta_2}$ if

$$Var(\widehat{\theta_1}) < Var(\widehat{\theta_2}).$$

  That is, $\widehat{\theta_1}$ is more efficient than $\widehat{\theta_2}$ if it has a smaller variance.
- The **relative efficiency** of $\widehat{\theta_2}$ with respect to $\widehat{\theta_1}$ is defined as $Var(\widehat{\theta_1})/Var(\widehat{\theta_2})$

Consistency: Increasing Sample Sizes

- Let $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_n, \ldots$ be a sequence of estimators of θ, where each estimator $\widehat{\theta}_n$ based on a sample of size n. This sequence of estimators is defined to be a *mean-squared-error* consistent sequence of estimators of θ if and only if

$$\lim_{n \to \infty} E[(\widehat{\theta}_n - \theta)^2] = 0, \text{for all } \theta$$

- Mean-squared-error consistency implies that both the bias and the variance of $\widehat{\theta}_n$ approaches 0, since the mean-squared-error in the definition can be decomposed in a variance and squared-bias component.

Consistency: Increasing Sample Sizes

- $\widehat{\theta}_n$ is a **consistent** estimator of θ if

$$\lim_{n\to\infty} P(\left|\widehat{\theta}_n - \theta\right| > \epsilon) = 0 \text{ for all } \epsilon > 0.$$

  We then say that $\widehat{\theta}_n$ **converges in probability** to θ as $n \to \infty$.
- Equivalently,

$$\lim_{n\to\infty} P(\left|\widehat{\theta}_n - \theta\right| < \epsilon) = 1 \text{ for all } \epsilon > 0.$$

- If an estimator is a mean-squared-error consistent estimator, it is also a simple consistent estimator (proof by Chebyshev inequality), but not necessarily vice versa.

Theorem

Let X be a random variable and $g()$ a non-negative function with domain the real line, then $P[g(X) \geq k] \leq \frac{\mathrm{E}(g(X))}{k}$, for every $k > 0$.

Consistency: Increasing Sample Sizes

- Note that these definitions of consistency involve large-sample or *asymptotic* properties.
- Consistency has to do only with the *limiting behaviour of an estimator* as the sample size increases without limit and does not imply that the observed value of $\widehat{\theta}$ is necessarily close to θ for any specific size of sample n.
- If only a relatively small sample is available, it would seem immaterial whether a consistent estimator is used or not.
- In the context of large samples, the following theorem (without proof) is useful:

Theorem

If $\lim_{n\to\infty} E(\widehat{\theta}_n) = \theta$ *and* $\lim_{n\to\infty} Var(\widehat{\theta}_n) = 0$, *then* $\widehat{\theta}_n$ *is a consistent estimator of* θ.