

Statistics Flashcards

Erik Erhardt

September 20, 2004

Note: These flashcards are designed to be printed double-sided and cut.

Section	Page
Distributions	3
C&B Ch.6 Data Reduction	5
C&B Ch.7 Point Estimation	13
RC Simple Linear Regression	21

C&B refers to Casella and Berger's "Statistical Inference", 2nd ed.

RC refers to Ron Christensen's "Analysis of Variance, Design, and Regression: Applied Statistical Methods".

DISTRIBUTION

Normal(μ, σ^2)

Statistics Flashcards 0.1

DISTRIBUTION

Statistics Flashcards 0.2

DISTRIBUTION

Statistics Flashcards 0.3

DISTRIBUTION

Statistics Flashcards 0.4

DISTRIBUTION

Statistics Flashcards 0.5

DISTRIBUTION

Statistics Flashcards 0.6

pdf $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$
mean $E(X) = \mu$
variance $\text{Var}(X) = \sigma^2$
mgf $M_X(t) = \exp\{\mu t + \sigma^2 t^2 / 2\}$

Notes

C&B CH.6 DATA REDUCTION

The Sufficiency Principle

Statistics Flashcards 6.1

C&B CH.6 DATA REDUCTION

Sufficient statistic for θ
(Def 6.2.1 / Th 6.2.2)

Statistics Flashcards 6.2

C&B CH.6 DATA REDUCTION

The Factorization Theorem
(Th 6.2.6)

Statistics Flashcards 6.3

C&B CH.6 DATA REDUCTION

Exponential family sufficient statistic for $\underline{\theta}$
(Th 6.2.10)

Statistics Flashcards 6.4

C&B CH.6 DATA REDUCTION

Minimal Sufficient Statistic
(Def. 6.2.11)

Statistics Flashcards 6.5

C&B CH.6 DATA REDUCTION

Minimal Sufficient Statistic
(Th. 6.2.13)

Statistics Flashcards 6.6

A statistic $T(\underline{X})$ is a *sufficient statistic for θ* if the conditional distribution of the sample \underline{X} given the value of $T(\underline{X})$ does not depend on θ .

If $p(\underline{x}|\theta)$ is the joint pdf or pmf of \underline{X} and $q(t|\theta)$ is the pdf or pmf of $T(\underline{X})$, then $T(\underline{X})$ is a **sufficient statistic for θ** if, for every \underline{x} in the sample space, the **ratio $p(\underline{x}|\theta)/q(T(\underline{X})|\theta)$ is constant as a function of θ** .

It turns out that outside of the exponential family of distributions, it is rare to have a sufficient statistic of smaller dimension than the size of the sample, so in many cases it will turn out that the order statistics are the best that we can do (p.275).

Any one-to-one function of a sufficient statistic is a sufficient statistic (p.280).

If $T(\underline{X})$ is a sufficient statistic for θ , then any inference about θ should depend on the sample \underline{X} only through the value of $T(\underline{X})$. That is, if \underline{x} and \underline{y} are two sample points such that $T(\underline{x}) = T(\underline{y})$, then the inference about θ should be the same whether $\underline{X} = \underline{x}$ or $\underline{X} = \underline{y}$ is observed.

Let X_1, \dots, X_n be iid observations from a pdf or pmf $f(x|\theta)$ that belongs to an exponential family given by

$$f(x|\theta) = h(x)c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta)t_i(x)\right),$$

where $\theta = (\theta_1, \dots, \theta_d)$, $d \leq k$. Then

$$T(\underline{X}) = \left(\sum_{j=1}^n t_1(X_j), \dots, \sum_{j=1}^n t_k(X_j)\right)$$

is a sufficient statistic for θ .

Let $f(\underline{x}|\theta)$ denote the joint pdf or pmf of a sample \underline{X} . A statistics $T(\underline{X})$ is a sufficient statistics for θ if and only if there exist functions $g(t|\theta)$ and $h(\underline{x})$ such that, for all sample points \underline{x} and all parameter points θ ,

$$f(\underline{x}|\theta) = g(T(\underline{X})|\theta)h(\underline{x}).$$

Let $f(\underline{x}|\theta)$ be the pmf or pdf of a sample \underline{X} . Suppose there exists a function $T(\underline{x})$ such that, for every two sample points \underline{x} and \underline{y} , the ratio $f(\underline{x}|\theta)/f(\underline{y}|\theta)$ is constant as a function of θ if and only if $T(\underline{x}) = T(\underline{y})$. Then $T(\underline{X})$ is a minimal sufficient statistic for θ .

A sufficient statistic $T(\underline{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\underline{X})$, $T(\underline{x})$ is a function of $T'(\underline{x})$.

C&B CH.6 DATA REDUCTION

Ancillary Statistic
(Def. 6.2.16)

Statistics Flashcards 6.7

C&B CH.6 DATA REDUCTION

Complete Statistic
(Def. 6.2.21)

Statistics Flashcards 6.8

C&B CH.6 DATA REDUCTION

Basu's Theorem
(Th. 6.2.24)

Statistics Flashcards 6.9

C&B CH.6 DATA REDUCTION

Complete statistics in the exponential family
(Th. 6.2.25)

Statistics Flashcards 6.10

C&B CH.6 DATA REDUCTION

Complete and minimal sufficiency
(Th. 6.2.28)

Statistics Flashcards 6.11

C&B CH.6 DATA REDUCTION

Likelihood Function
(Def. 6.3.1)

Statistics Flashcards 6.12

Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\underline{X})$. The family of probability distributions is called *complete* if $E_{\theta}g(T) = 0$ for all θ implies $\Pr_{\theta}(g(T) = 0) = 1$ for all θ . Equivalently, $T(\underline{X})$ is called a *complete statistic*.

Notice that completeness is a property of a family of probability distributions, not of a particular distribution. A fundamental property of a complete statistic is that it is minimal (p. 289).

A statistic $S(\underline{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic*.

$R = X_{(n)} - X_{(1)}$ is an ancillary statistic for a location family. $\frac{X_1 + \dots + X_n}{X_n}$ is ancillary for a scale family.

In any location family, S^2 is ancillary.

Let X_1, \dots, X_n be iid observations from an exponential family with pdf or pmf of the form

$$f(x|\underline{\theta}) = h(x)c(\theta) \exp\left(\sum_{j=1}^k w(\theta_j)t_j(x)\right),$$

where $\underline{\theta} = (\theta_1, \dots, \theta_k)$. Then the statistic

$$T(\underline{X}) = \left(\sum_{i=1}^n t_1(X_i), \dots, \sum_{i=1}^n t_k(X_i)\right),$$

is complete as long as the parameter space Θ contains an open set in \mathbb{R}^k (dimension is number of parameters).

If $T(\underline{X})$ is a complete and minimal sufficient statistic, then $T(\underline{X})$ is independent of every ancillary statistic.

Allows us to deduce the independence of two statistics without ever finding the joint distribution of the two statistics.

In order to use Basu's Theorem, we need to show that a statistic is complete, but most problems are in the exponential family, so can use Th. 6.2.25.

Converse is not true!

Let $f(\underline{x}|\theta)$ denote the joint pdf or pmf of the sample $\underline{X} = (X_1, \dots, X_n)$. Then, given that $\underline{X} = \underline{x}$ is observed, the function of θ defined by

$$\begin{aligned} L(\theta|\underline{x}) &= f(\underline{x}|\theta) && \text{continuous} \\ &= \Pr_{\theta}(\underline{X} = \underline{x}) && \text{discrete} \end{aligned}$$

is called the *likelihood function*.

Same as pdf or pmf but consider \underline{x} fixed and θ to be the variable.

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

C&B CH.6 DATA REDUCTION

Likelihood Principle

Statistics Flashcards 6.13

C&B CH.6 DATA REDUCTION

Formal Sufficiency Principle
Using Evidence Function $Ev(E, x)$

Statistics Flashcards 6.14

C&B CH.6 DATA REDUCTION

Conditionality Principle

Statistics Flashcards 6.15

C&B CH.6 DATA REDUCTION

Likelihood Principle Corollary

Statistics Flashcards 6.16

C&B CH.6 DATA REDUCTION

Birnbaum's Theorem
(**Th. 6.3.6**)

Statistics Flashcards 6.17

C&B CH.6 DATA REDUCTION

Equivariance Principle

Statistics Flashcards 6.18

Consider experiment $E = (\underline{X}, \theta, \{f(\underline{x}|\theta)\})$ and suppose $T(\underline{X})$ is a sufficient statistic for θ . If \underline{x} and \underline{y} are sample points satisfying $T(\underline{x}) = T(\underline{y})$, then $\text{Ev}(E, \underline{x}) = \text{Ev}(E, \underline{y})$.

An *experiment* E is the triple $(\underline{X}, \theta, \{f(\underline{x}|\theta)\})$, where \underline{X} is a random vector with pmf $f(\underline{x}|\theta)$ for some θ in the parameter space Θ . An experimenter, knowing what experiment E was performed and having observed a particular sample $\underline{X} = \underline{x}$, will make some inference or draw some conclusion about θ . This conclusion we denote by $\text{Ev}(E, \underline{x})$, which stands for the *evidence about θ arising from E and \underline{X}* .

For example, $\text{Ev}(E, \underline{x}) = (\bar{x}, \sigma/\sqrt{n})$ is the evidence about μ for σ^2 known.

Belief in this principle necessitates belief in the model, something that may not be easy to do (p.295).

If \underline{x} and \underline{y} are two sample points such that $L(\theta|\underline{x})$ is proportional to $L(\theta|\underline{y})$, that is, there exists a constant $C(\underline{x}, \underline{y})$ such that

$$\frac{L(\theta|\underline{x})}{L(\theta|\underline{y})} = C(\underline{x}, \underline{y}) \quad \text{for all } \theta,$$

then the conclusions drawn from \underline{x} and \underline{y} should be identical.

Used to compare the plausibility of various parameter values, e.g., if $L(\theta_2|\underline{x}) = 2L(\theta_1|\underline{x})$, then, in some sense, θ_2 is twice as plausible as θ_1 (p. 291). (Not “probable” since we often think of θ as being a fixed (unknown) value, and there is no guarantee that $L(\theta|\underline{x})$, as a function of θ is a pdf.

Fiducial inference sometimes interprets likelihoods as probabilities for θ , by normalizing $L(\theta|\underline{x})$ by multiplying by $M(\underline{x}) = \left(\int_{-\infty}^{\infty} L(\theta|\underline{x}) d\theta\right)^{-1}$. Most statisticians do not subscribe to the fiducial theory of inference, but it is based on Fisher’s (1930) *inverse probability*.

If $E = (\underline{X}, \theta, \{f(\underline{x}|\theta)\})$ is an experiment, then $\text{Ev}(E, \underline{x})$ should depend on E and \underline{x} only through $L(\theta|\underline{x})$.

If one of two experiments is randomly chosen and the chosen experiment is done, yielding data \underline{x} , the information about θ depends only on the experiment performed. That is, it is the same information as would have been obtained if it were decided (nonrandomly) to do that experiment from the beginning, and the data \underline{x} had been observed. The fact that this experiment was performed, rather than some other, has not increased, decreased, or changed knowledge of θ .

For example (ex 6.3.5), if do Binomial(20, p) or do Neg.Bin.(7, p) in 20 tosses, the knowledge about p is the same.

If $\underline{Y} = g(\underline{X})$ is a change of measurement scale such that the model for \underline{Y} has the same formal structure as the model for \underline{X} , then an inference procedure should be both measurement equivariant and formally equivariant.

Measurement equivariance prescribes that the inference made should not depend on the measurement scale that is used.

Formal invariance states that if two inference problems have the same formal structure in terms of the mathematical model used, then the same inference procedure should be used in both problems. The elements of the model that must be the same are: Θ , the parameter space; $\{f(\underline{x}|\theta) : \theta \in \Theta\}$, the set of pdfs or pmfs for the sample; and the set of *allowable inferences and consequences of wrong inferences*, which can be taken to be Θ , that an inference is simply a choice of an element of Θ as an estimate or guess at the true value of θ .

It says that one inference procedure is appropriate *even if the physical realities are quite different*, an assumption that is sometimes difficult to justify.

The Formal Likelihood Principle follows from the Formal Sufficiency Principle and the Conditionality Principle. The converse is also true.

C&B CH.6 DATA REDUCTION

Group of Transformations
(Def. 6.4.2)

Statistics Flashcards 6.19

C&B CH.6 DATA REDUCTION

pdf is invariant under a group of transformations
(Def. 6.4.4)

Statistics Flashcards 6.20

C&B CH.6 DATA REDUCTION

Necessary Statistics
(Def. 6.6.3)

Statistics Flashcards 6.21

C&B CH.6 DATA REDUCTION

Minimal and Necessary Statistics
(Th. 6.6.4)

Statistics Flashcards 6.22

C&B CH.6 DATA REDUCTION

Minimal Sufficient Statistics
(Th. 6.6.5)

Statistics Flashcards 6.23

C&B CH.6 DATA REDUCTION

x
 $()$

Statistics Flashcards 6.24

Let $\mathcal{F} = \{f(\underline{x}|\theta) : \theta \in \Theta\}$ be a set of pdfs or pmfs for \underline{X} , and let \mathcal{G} be a group of transformations of the sample space \mathcal{X} . Then \mathcal{F} is *invariant under the group* \mathcal{G} if for every $\theta \in \Theta$ and $g \in \mathcal{G}$ there exists a unique $\theta' \in \Theta$ such that $\underline{T} = g(\underline{X})$ has the distribution $f(\underline{y}|\theta')$ if \underline{X} has the distribution $f(\underline{x}|\theta)$.

A set of functions $\{g(\underline{x}) : g \in \mathcal{G}\}$ from the sample space \mathcal{X} onto \mathcal{X} is called a *group of transformations of \mathcal{X}* if

Inverse For every $g \in \mathcal{G}$ there is a $g' \in \mathcal{G}$ such that $g'(g(\underline{x})) = \underline{x}$ for all $\underline{x} \in \mathcal{X}$.

Composition For every $g \in \mathcal{G}$ and $g' \in \mathcal{G}$ there exists $g'' \in \mathcal{G}$ such that $g'(g(\underline{x})) = g''(\underline{x})$ for all $\underline{x} \in \mathcal{X}$.

Identity The identity, $e(\underline{x})$, defined by $e(\underline{x}) = \underline{x}$ is an element of \mathcal{G} .

Identity is a consequence of the first two.

A statistic is a minimal sufficient statistic if and only if it is a necessary and sufficient statistic.

A statistic is said to be *necessary* if it can be written as a function of every sufficient statistic.

Suppose that the family of densities $\{f_0(\underline{x}, \dots, f_k(\underline{x})\}$ all have common support. Then

a. The statistic

$$T(\underline{X}) = \left(\frac{f_1(\underline{X})}{f_0(\underline{X})}, \frac{f_2(\underline{X})}{f_0(\underline{X})}, \dots, \frac{f_k(\underline{X})}{f_0(\underline{X})} \right)$$

is minimal sufficient for the family $\{f_0(\underline{x}, \dots, f_k(\underline{x})\}$.

b. If \mathcal{F} is a family of densities with common support, and

- $f_i(\underline{x}) \in \mathcal{F}, i = 0, 1, \dots, k,$
- $T(\underline{x})$ is sufficient for \mathcal{F} ,

then $T(\underline{x})$ is minimal sufficient for \mathcal{F} .

Can be used to show that for samples from distributions like the logistic or double exponential, the order statistics are minimal sufficient (p.309).

C&B CH.7 POINT ESTIMATION

Point Estimator
(Def. 7.1.1)

Statistics Flashcards 7.1

C&B CH.7 POINT ESTIMATION

Method of Moments

Statistics Flashcards 7.2

C&B CH.7 POINT ESTIMATION

Satterthwaite approximation (χ^2 rv's)

Statistics Flashcards 7.3

C&B CH.7 POINT ESTIMATION

Maximum Likelihood Estimator (MLE)
(Def. 7.2.4)

Statistics Flashcards 7.4

C&B CH.7 POINT ESTIMATION

Invariance Property of MLEs
(Th.7.2.10)

Statistics Flashcards 7.5

C&B CH.7 POINT ESTIMATION

MLE Numerical Stability

Statistics Flashcards 7.6

Equate the first k sample moments to the corresponding k population moments, and solve the resulting system of simultaneous equations.

That is,

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i^1 \quad \mu'_1 = EX^1,$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad \mu'_2 = EX^2,$$

The population moment μ'_j will typically be a function of $\theta_1, \dots, \theta_k$, say $\mu'_j(\theta_1, \dots, \theta_k)$.

The MM estimator $(\tilde{\theta}_1, \dots, \tilde{\theta}_k)$ of $(\theta_1, \dots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \dots, \theta_k)$ in terms of (m_1, \dots, m_k) :

$$m_1 = \mu'_1(\theta_1, \dots, \theta_k)$$

$$m_2 = \mu'_2(\theta_1, \dots, \theta_k)$$

A *point estimator* is any function $W(X_1, \dots, X_n)$ of a sample; that is, any statistic is a point estimator.

For each sample point \underline{x} , let $\hat{\theta}(\underline{x})$ be a parameter value at which $L(\theta|\underline{x})$ attains its maximum as a function of θ , with \underline{x} held fixed. A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample \underline{X} is $\hat{\theta}(\underline{X})$.

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \dots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} \ln L(\theta|\underline{x}) = 0, \quad i = 1, \dots, k.$$

$$\frac{\partial^2}{\partial \theta_i^2} \ln L(\theta|\underline{x})|_{\theta=\hat{\theta}} < 0.$$

Check these interior points for the **global** maximum – **and check the endpoints** of the range of parameter space (range may be restricted). May need to maximize numerically (carefully) if it can not be done analytically. The MLE may be numerically sensitive to small changes in the data.

If $Y_i, i = 1, \dots, k$, are ind. $\chi_{r_i}^2$ rv's, then $\sum Y_i$ is chi-squared with $\nu = \sum r_i$.

$$\text{For } \sum a_i Y_i, \hat{\nu} = \frac{(\sum a_i Y_i)^2}{\sum \frac{a_i^2}{r_i} Y_i^2}.$$

Recall that the likelihood function is a function of the parameter, θ , with the data, \underline{x} , held constant. However, since the data are measured with error, we might ask how small changes in the data might affect the MLE. That is, we calculate $\hat{\theta}$ based on $L(\theta|\underline{x})$, but we might inquire what value we would get for the MLE if we based our calculations on $L(\theta|\underline{x} + \epsilon)$, for small ϵ . Intuitively, this new MLE, say $\hat{\theta}_1$, should be close to $\hat{\theta}$ if ϵ is small. But this is not always the case. (p.323).

If $\hat{\theta}$ is the MLE of θ , then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Define for $\tau(\theta)$ the *induced likelihood function* L^* ,

$$L^*(\eta|\underline{x}) = \sup_{\{\theta: \tau(\theta)=\eta\}} L^*(\theta|\underline{x}).$$

The value $\hat{\eta}$ that maximizes $L^*(\eta|\underline{x})$ will be called the MLE of $\eta = \tau(\theta)$, and the maxima of L^* and L coincide (p.320).

The invariance property of MLEs also holds in the multivariate case. If the MLE of $(\theta_1, \dots, \theta_k)$ is $(\hat{\theta}_1, \dots, \hat{\theta}_k)$, and if $\tau(\theta_1, \dots, \theta_k)$ is any function of the parameters, the MLE of $\tau(\theta_1, \dots, \theta_k)$ is $\tau(\hat{\theta}_1, \dots, \hat{\theta}_k)$ (p.321).

C&B CH.7 POINT ESTIMATION

Bayesian Conjugate Family
(Def. 7.2.15)

Statistics Flashcards 7.7

C&B CH.7 POINT ESTIMATION

The EM Algorithm
(Expectation-Maximization)

Statistics Flashcards 7.8

C&B CH.7 POINT ESTIMATION

The EM Algorithm
("Complete-data" and "Incomplete-data")

Statistics Flashcards 7.9

C&B CH.7 POINT ESTIMATION

EM (did not complete cards for this)
()

Statistics Flashcards 7.10

C&B CH.7 POINT ESTIMATION

x
()

Statistics Flashcards 7.11

C&B CH.7 POINT ESTIMATION

Monotonic EM sequense
(Th. 7.2.20)

Statistics Flashcards 7.12

The EM algorithm is guaranteed to converge to the MLE. It is particularly suited to “missing data” problems.

In using the EM algorithm we consider two different likelihood problems. The problem that we are interested in solving is the “incomplete-data” problem, and the problem that we actually solve is the “complete-data” problem. Depending on the situation, we can start with either problem (p.326).

Let \mathcal{F} denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a *conjugate family* for \mathcal{F} if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π , and all $x \in \mathcal{X}$.

We observe X_1, \dots, X_n and Y_1, \dots, Y_n , all mutually independent, where $X_i \sim f(\theta_{ix})$ and $Y_i \sim f(\theta_{iy})$. The joint pdf is therefore

$$\begin{aligned} f((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) | \theta_{1x}, \dots, \theta_{nx}, \theta_{1y}, \dots, \theta_{ny}) \\ = \prod_{i=1}^n f(\theta_{ix}) f(\theta_{iy}) \end{aligned}$$

This is the “complete-data” likelihood, and $((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ is called the complete data.

Missing data, which is a common occurrence, would make estimation more difficult. Suppose, for example, that the value of x_1 was missing. Discarding y_1 and proceeding with a sample of size $n - 1$ would ignore the information in y_1 . The pdf of the sample with x_1 missing is $\int_{x_1} f((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) | \theta_{1x}, \dots, \theta_{nx}, \theta_{1y}, \dots, \theta_{ny}) dx_1$. This is the “incomplete-data” likelihood, which we need to maximize.

The sequence $\{\hat{\theta}_{(r)}\}$ defined by

$$\theta^{(r+1)} = \text{the value that maximizes } E \left[\log L(\theta | \underline{y}, \underline{X}) | \theta^{(r)}, \underline{y} \right]$$

satisfies

$$L(\theta^{(r+1)}, \underline{y}) \geq L(\theta^{(r)}, \underline{y}),$$

with equality holding if and only if successive iterations yield the same value of the maximized expected complete-data log likelihood, that is

$$E \left[\log L(\hat{\theta}^{(r+1)} | \underline{y}, \underline{X}) | \hat{\theta}^{(r)}, \underline{y} \right] = E \left[\log L(\hat{\theta}^{(r)} | \underline{y}, \underline{X}) | \hat{\theta}^{(r)}, \underline{y} \right].$$

C&B CH.7 POINT ESTIMATION

Mean Squared Error (MSE)
(Def. 7.3.1)

Statistics Flashcards 7.13

C&B CH.7 POINT ESTIMATION

Bias
(Def. 7.3.2)

Statistics Flashcards 7.14

C&B CH.7 POINT ESTIMATION

Principles of
Measurement Equivariance and Formal Invariance
(p.333)

Statistics Flashcards 7.15

C&B CH.7 POINT ESTIMATION

Best estimators with same expected value
(p.334)

Statistics Flashcards 7.16

C&B CH.7 POINT ESTIMATION

Best Unbiased Estimator (BUE) or
Uniform Minimum Variance Unbiased Estimator (UMVUE)
(Def. 7.3.7)

Statistics Flashcards 7.17

C&B CH.7 POINT ESTIMATION

Cramér-Rao Inequality
(Th. 7.3.9)

Statistics Flashcards 7.18

The *bias* of a point estimator W of a parameter θ is the difference between the expected value of W and θ ; that is, $\text{Bias}_\theta W = E_\theta W - \theta$. An estimator whose bias is identically (in θ) equal to 0 is called *unbiased* and satisfies $E_\theta W = \theta$ for all θ .

If an estimator is unbiased, its MSE is equal to its variance.

Although many unbiased estimators are also reasonable from the standpoint of MSE, be aware that controlling bias does not guarantee that MSE is controlled. In particular, it is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in MSE (p.331).

For example, $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ has a smaller MSE than S^2 , thus by trading off variance for bias, the MSE is improved. However, $\hat{\sigma}^2$ is biased and will systematically underestimate σ^2 (p.332).

The *mean squared error* (MSE) of an estimator W of a parameter θ is the function of θ defined by $E_\theta(W - \theta)^2$.

$$\begin{aligned} E_\theta(W - \theta)^2 &= \text{Var}_\theta W + (E_\theta W - \theta)^2 \\ &= \text{Var}_\theta W + (\text{Bias}_\theta W)^2 \end{aligned}$$

The MSE measures the average squared difference between the estimator W and the parameter θ .

The MSE incorporates two components, one measuring the variability of the estimator (precision) and the other measuring its bias (accuracy).

While the MSE is a reasonable criterion for location parameters, it is not reasonable for scale parameters (p.332). This is so because the MSE penalizes equally for overestimation and underestimation, which is fine in the location case. In the scale case, however, 0 is a natural lower bound, so the estimation problem is not symmetric. Use of MSE in this case tends to be forgiving of underestimation (p.332).

Suppose that there is an estimator W^* of θ with $E_\theta W^* = \tau(\theta) \neq \theta$, and we are interested in investigating the worth of W^* . Consider the class of estimators

$$\mathcal{C}_\tau = \{W : E_\theta W = \tau(\theta)\}.$$

For any $W_1, W_2 \in \mathcal{C}_\tau$, $\text{Bias}_\theta W_1 = \text{Bias}_\theta W_2$, so

$$E_\theta(W_1 - \theta)^2 - E_\theta(W_2 - \theta)^2 = \text{Var}_\theta(W_1) - \text{Var}_\theta(W_2),$$

and MSE comparisons, within the class \mathcal{C}_τ , can be based on variances alone.

Measurement Equivariance: $W(\underline{x})$ estimates $\theta \implies \bar{g}(W(\underline{x}))$ estimates $\bar{g}(\theta) = \theta'$.

Formal Invariance: $W(\underline{x})$ estimates $\theta \implies W(g(\underline{x}))$ estimates $\bar{g}(\theta) = \theta'$.

Together these give $W(g(\underline{x})) = \bar{g}(W(\underline{x}))$.

Example 7.3.6 (MSE of equivariant estimators) gives a MSE that does not depend on θ , so the MSEs of these equivariant estimators are functions of θ . Thus, an estimator with smallest MSE can be found by finding the function W that minimizes the MSE (p.334).

Let X_1, \dots, X_n be a sample with pdf $f(\underline{x}|\theta)$, and let $W(\underline{X}) = W(X_1, \dots, X_n)$ be any estimator satisfying

$$\frac{d}{d\theta} E_\theta W(\underline{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} [W(\underline{x}) f(\underline{x}|\theta)] d\underline{x}$$

and

$$\text{Var}_\theta(W(\underline{X})) < \infty.$$

Then

$$\text{Var}_\theta(W(\underline{X})) \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\underline{X})\right)^2}{E_\theta \left(\left(\frac{\partial}{\partial \theta} \log f(\underline{X}|\theta)\right)^2\right)}.$$

See proof (p.336). Uses the Cauchy-Schwarz Inequality $[\text{Cov}(X, Y)]^2 \leq (\text{Var}X)(\text{Var}Y)$.

An estimator W^* is a *best unbiased estimator* (BUE) of $\tau(\theta)$ if it satisfies $E_\theta W^* = \tau(\theta)$ for all θ and, for any other estimator W with $E_\theta W = \tau(\theta)$, we have $\text{Var}_\theta W^* \leq \text{Var}_\theta W$ for all θ . W^* is also called a *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.

C&B CH.7 POINT ESTIMATION

$$\bar{x}$$
$$()$$

Statistics Flashcards 7.19

C&B CH.7 POINT ESTIMATION

$$\bar{x}$$
$$()$$

Statistics Flashcards 7.20

C&B CH.7 POINT ESTIMATION

$$\bar{x}$$
$$()$$

Statistics Flashcards 7.21

C&B CH.7 POINT ESTIMATION

$$\bar{x}$$
$$()$$

Statistics Flashcards 7.22

C&B CH.7 POINT ESTIMATION

$$\bar{x}$$
$$()$$

Statistics Flashcards 7.23

C&B CH.7 POINT ESTIMATION

$$\bar{x}$$
$$()$$

Statistics Flashcards 7.24

SIMPLE LINEAR REGRESSION

SLR Model

Statistics Flashcards 540.1

SIMPLE LINEAR REGRESSION

$$\hat{\beta}_0, \hat{\beta}_1, MSE \text{ from}$$

$$n, \bar{x}, \bar{y}, s_x, s_y, s_{xy}$$
Statistics Flashcards 540.2

SIMPLE LINEAR REGRESSION

E, Var and SE of $\hat{\beta}_0, \hat{\beta}_1, MSE$ *Statistics Flashcards 540.3*

SIMPLE LINEAR REGRESSION

Estimating β_1 *Statistics Flashcards 540.4*

SIMPLE LINEAR REGRESSION

Estimating β_0 *Statistics Flashcards 540.5*

SIMPLE LINEAR REGRESSION

Estimating $\beta_0 + \beta_1 x$ *Statistics Flashcards 540.6*

$$\begin{aligned}
 s_{xy} &= \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})} \\
 \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2} \\
 &= \frac{s_{xy}}{s_x^2} = \frac{\sum (x_i y_i) - n\bar{x}\bar{y}}{(n-1)s_x^2} \\
 \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
 MSE &= \frac{\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{n-2} \\
 &= \frac{1}{n-2} \left((n-1)s_y^2 - \hat{\beta}_1^2 (n-1)s_x^2 \right)
 \end{aligned}$$

Model:

$$\begin{aligned}
 y_i &= \beta_0 + \beta_1 x_{i1} + \varepsilon_i \\
 \varepsilon_i &\stackrel{iid}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n.
 \end{aligned}$$

x known and nonrandom, β_0, β_1 unknown nonrandom, y known but random.

The regression equation describes an observed relationship between y and x . It can be used to predict y from x , but does not imply causation.

Alternative Model:

$$\begin{aligned}
 y_i &= \alpha + \beta_1(x_{i1} - \bar{x}) + \varepsilon_i \\
 \varepsilon_i &\stackrel{iid}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n.
 \end{aligned}$$

$\hat{\alpha} = \bar{y}$, $(\bar{y}, \hat{\beta}_1)$ independent.

$$\begin{aligned}
 \text{Par} & \beta_1 \\
 \text{Est} & \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \\
 \text{SE(Est)} & \sqrt{\frac{MSE}{(n-1)s_x^2}} \\
 \text{Dist} & \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \sim t(n-2) \\
 \text{E}(\hat{\beta}_1) & = \beta_1 \\
 \text{Var}(\hat{\beta}_1) & = \frac{\sigma^2}{(n-1)s_x^2}
 \end{aligned}$$

$$\begin{aligned}
 \text{E}(MSE) &= \sigma^2 \\
 \text{E}(\hat{\beta}_1) &= \beta_1 \\
 \text{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2} \\
 \text{SE}(\hat{\beta}_1) &= \sqrt{\frac{MSE}{(n-1)s_x^2}} \\
 \text{E}(\hat{\beta}_0) &= \beta_0 \\
 \text{Var}(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right) \\
 \text{SE}(\hat{\beta}_0) &= \sqrt{MSE \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)}
 \end{aligned}$$

CI: $\text{Est} \pm \text{SE(Est)} t(1 - \frac{\alpha}{2}, df)$

For a fixed x ,

$$\begin{aligned}
 \text{Par} & \beta_0 + \beta_1 x \\
 \text{Est} & \hat{\beta}_0 + \hat{\beta}_1 x = (\bar{y} - \hat{\beta}_1 \bar{x}) + \frac{s_{xy}}{s_x^2} x \\
 \text{SE(Est)} & \sqrt{MSE \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right)} \\
 \text{Dist} & \frac{(\hat{\beta}_0 + \hat{\beta}_1 x) - (\beta_0 + \beta_1 x)}{\text{SE}(\hat{\beta}_0 + \hat{\beta}_1 x)} \sim t(n-2) \\
 \text{E}(\hat{\beta}_0 + \hat{\beta}_1 x) &= \beta_0 + \beta_1 x \\
 \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right)
 \end{aligned}$$

and

$$\text{SE(Pred)} \quad \sqrt{MSE \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right)}$$

$$\begin{aligned}
 \text{Par} & \beta_0 \\
 \text{Est} & \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\
 \text{SE(Est)} & \sqrt{MSE \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)} \\
 \text{Dist} & \frac{\hat{\beta}_0 - \beta_0}{\text{SE}(\hat{\beta}_0)} \sim t(n-2) \\
 \text{E}(\hat{\beta}_0) &= \beta_0 \\
 \text{Var}(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)
 \end{aligned}$$

SIMPLE LINEAR REGRESSION

ANOVA Table

Statistics Flashcards 540.7

SIMPLE LINEAR REGRESSION

Correlation, r , and Coefficient of Determination, R^2

Statistics Flashcards 540.8

SIMPLE LINEAR REGRESSION

\bar{x}
()

Statistics Flashcards 540.9

SIMPLE LINEAR REGRESSION

\bar{x}
()

Statistics Flashcards 540.10

SIMPLE LINEAR REGRESSION

\bar{x}
()

Statistics Flashcards 540.11

SIMPLE LINEAR REGRESSION

\bar{x}
()

Statistics Flashcards 540.12

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\hat{\beta}_1 s_x}{s_y}, \quad -1 \leq r \leq 1$$

From perfect decreasing linear relationship ($r = -1$) between x and y , to no linear relationship ($r = 0$), to perfect increasing linear relationship ($r = 1$).

$$\sqrt{n-2} \frac{r}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t(n-2)$$

Testing $\rho \neq 0$ with r is the same as $\beta_1 \neq 0$ with $\hat{\beta}_1$. Same test for slope or r different from 0.

$$R^2 = \frac{SSReg}{SSTot} \quad (= r^2 \text{ when one } x \text{ variable})$$

R^2 , the coefficient of determination, can be used as a measure of the model's predictive ability, but not as a measure of the correctness of a model in absolute terms.

Usually, the ANOVA table lacks the β_0 line, if including β_0 use values in brackets [...].

Source	df	SS	MS	F
β_0	n/a [1]	n/a [$n\bar{y}^2 \equiv C$]	n/a [$n\bar{y}^2/1$]	
β_1	1	<i>SSReg</i>	<i>MSReg</i>	$\frac{MSReg}{MSE}$
Error	$n - 2$	<i>SSE</i>	<i>MSE</i>	
Total	$n - 1$ [n]	<i>SSTot</i> [$\sum y_i^2$]		

where

SS	MS
$SSReg = \hat{\beta}_1^2 (n-1) s_x^2$	$MSReg = SSReg/1$
$SSE = (n-1) s_y^2 - \hat{\beta}_1^2 (n-1) s_x^2$	$MSE = SSE/(n-2)$
$SSTot = (n-1) s_y^2$	