Statistics (Notes)

Mark DiBattista

December 11, 2019

Abstract

Statistics is a specific application of probability theor...

1 Suggested Resource Materials

Useful source texts:

- Probability/Statistics, intermediate (probability sections are better than statistics): *Statistical Inference*, Casella & Berger
- Probability, advanced: *Probability and Measure*, Billingsley

Throughout the text the acronyms refer to companion writeups,

LAN	*Linear Algebra (Notes)*
LAA	*Linear Algebra (Applications)*
PN	*Probability Notes*

within which information is referenced by chapter and/or numbered equation.

2 Statistics Preliminaries

2.1 Nomenclature

Given random variables, X and Y

Probability	$\mathbb{P}X$	*PN* §2.1.6
Expectation	$\mathbb{E}X$	*PN* §2.1.10
Variance	$\mathbb{V}X$	*PN* §3.1
Covariance	$\mathbb{C}(X,Y)$	*PN* §3.2
Fisher Information	$\mathbb{I}_F X$	*PN* §10.4.2

2.2 Terms and Definitions

2.2.1 Basic Definitions

Given a random variable, usually a random vector, $\mathbf{X}$, a **statistic** is any function of the random variable, $T(\mathbf{X})$. Typically, the coordinates of the random vector, $\mathbf{X} = (X_1, \cdots, X_n)$, are sampled from a common *parametrized* distribution, $X_1, \cdots, X_n \sim X \equiv X|\theta$, for which the value of the parameter, θ, is unknown. The symbol, $\mathbf{X}_\theta$, will be used in cases for which dependence of the random sample upon the parameter is to be stressed. There are two common types of tests that are run against sample statistics:

- **Point Estimation:** Given a random sample, $\mathbf{X} = \mathbf{x}$, and an explicit metric, provide the best estimate for the parameter, θ, that governs the distribution;
- **Interval Testing:** Given a hypothetical value for the parameter, θ, or a hypothesis that the random coordinates are independent, estimate the probability that the random sample, $\mathbf{X} = \mathbf{x}$, falls within a specified range. Hypotheses may be rejected if the realized measurement is deemed sufficiently improbable.

The **bias** of a point estimator is given by the mean of the difference between true value and test statistic:

$$\mathbb{B}T(\mathbf{X}) = \mathbb{E}\left(T(\mathbf{X}) - \theta\right). \tag{1}$$

The **mean-squared error** of a point estimator is given by the variance of the difference between true value and test statistic:

$$\mathbb{M}T(\mathbf{X}) = \mathbb{E}\left(T(\mathbf{X}) - \theta\right)^2 = \mathbb{V}T(\mathbf{X}) + (\mathbb{B}T(\mathbf{X}))^2. \tag{2}$$

There are two standard kinds of optimality for point estimators for a given parameter, expressed in terms of bias and mean-squared error:

- The **Minimum-Variance Unbiased Estimator (MVUE)** has the least mean-squared error among all estimators with no bias;
- The **Minimum Mean-Squared Error (MMSE)** has the least mean-squared error among *all* estimators.

Frequently, for point estimators small increases in bias can be traded for significant reduction in mean-squared error.

2.2.2 Properties of Point Estimators

Statistics are described by a number of properties:

- An **unbiased statistic** is one for which the bias is zero,

$$\mathbb{B}T(\mathbf{X}) = 0 \tag{3}$$

- A **sufficient statistic** is one that contains *all* information necessary to estimate the parameter, θ. A sufficient condition for the sufficition statistic is given by the **Fisher-Neyman Theorem**:

$$p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = h(\mathbf{X})g(\theta, T(\mathbf{X})) \Rightarrow T(\mathbf{X}) \text{ is a sufficient statistic} \tag{4}$$

- A **complete statistic** is one that ensure the

2.2.3 Asymptotic Properties of Statistics

Given an increasing vector of random variables, $\mathbf{X}_n = (X_1, \cdots, X_n), X_i \sim X$, sampled from a parametrized distribution, $X \equiv X|\theta$, and a sequence of statistics, $T(\mathbf{X}_n)$, we have the following statistical analogs to the law of large numbers and central limit theorem:

- A **consistent statistic** converges in probability to the true parameter value,

$$T(\mathbf{X}_n) \xrightarrow{p} \theta \tag{5}$$

- An **asymptotically normal statistic** converges in distribution to a Gaussian with variance decreasing in proportion to the square-root of the number of samples,

$$\sqrt{n}(T(\mathbf{X}_n) - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma^2) \tag{6}$$

3 Point Estimation

A **point estimator** is a sample statistic of parametrized distribution whose value provides an estimate for the parameter,

$$\mathbf{X} = (X_1, \cdots, X_n), X_i \sim X \equiv X|\theta \Rightarrow T(\mathbf{X}) \approx \theta. \tag{7}$$

A point estimator is therefore a random variable, governed by a probability distribution, and with full rights all other properties afforded by probability theory.

The **efficiency** of a point estimator is a measure of its variance in comparison to all other point estimators of the same parameter. The most efficient point estimator has least variance.

3.1 Cramer-Rao Lower Bound

The theoretical minimum variance for unbaised point estimators is provided by the **Cramer-Rao lower bound**: given a point estimator for the parameter,

$$\left.\begin{array}{l} \mathbf{X}_\theta = (X_1, \cdots, X_n), X_i \sim X \equiv X|\theta \\ \mathbb{B}T(\mathbf{X}_\theta) = 0 \end{array}\right\} \Rightarrow \mathbb{V}T(\mathbf{X}_\theta) \geq \frac{1}{\mathbb{I}_F \mathbf{X}_\theta}, \tag{8}$$

the theoretical minimum possible variance across all possible statistics is provided by the reciprocal of the Fisher information (*cf. PN*, §10.4.2). The Cramer-Rao lower bound provides the variance for the MVUE.

For arbitrary statistics, not necessarily unbiased, the Cramer-Rao lower bound is proportional to a function of the expectation of the statistic:

$$\mathbb{V}T(\mathbf{X}_\theta) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}T(\mathbf{X}_\theta)\right)^2}{\mathbb{I}_F \mathbf{X}_\theta}. \tag{9}$$

A derivation of the Cramer-Rao theorem follows directly from an application of the Cauchy-Schwarz inequality (*cf. PN*, §4.1) to the covariance of the score function (again *cf. PN*, §10.4.2), $S(\theta|\mathbf{X})$, and the statistic, $T(\mathbf{X})$. Given the covariance,

$$\begin{aligned} \mathbb{C}(S(\theta|\mathbf{X}_\theta), T(\mathbf{X}_\theta)) &= \mathbb{E}S(\theta|\mathbf{X}_\theta)T(\mathbf{X}_\theta) - \mathbb{E}S(\theta|\mathbf{X}_\theta) \cdot \mathbb{E}T(\mathbf{X}_\theta) = \mathbb{E}S(\theta|\mathbf{X}_\theta)T(\mathbf{X}_\theta) \\ &= \int_D p(\mathbf{x}|\theta)T(\mathbf{x})\frac{\partial}{\partial\theta}\ln p(\mathbf{x}|\theta)\, d\mathbf{x} = \int_D T(\mathbf{x})\frac{\partial}{\partial\theta}p(\mathbf{x}|\theta)\, d\mathbf{x} = \frac{\partial}{\partial\theta}\mathbb{E}T(\mathbf{X}_\theta), \end{aligned} \tag{10}$$

the Cauchy-Schwarz inequality yields

$$\mathbb{V}(S(\theta|\mathbf{X}_\theta)\cdot\mathbb{V}T(\mathbf{X}_\theta)\geq\left(\mathbb{C}(S(\theta|\mathbf{X}_\theta),T(\mathbf{X}_\theta))\right)^2\Rightarrow\mathbb{V}T(\mathbf{X}_\theta)\geq\frac{\left(\mathbb{C}(S(\theta|\mathbf{X}_\theta),T(\mathbf{X}_\theta))\right)^2}{\mathbb{V}(S(\theta|\mathbf{X}_\theta)}=\frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}T(\mathbf{X}_\theta)\right)^2}{\mathbb{I}_F\mathbf{X}_\theta}. \tag{11}$$

3.2 Rao-Blackwell Theorem

3.3 Methods for Generating Point Estimators

Point estimation is the

3.3.1 Point Estimators for Moments

$$\mathbf{X}=(X_1,\cdots,X_n),X_i\sim X\Rightarrow\begin{cases}\vdots\\ S_n^j=\frac{1}{n}\sum_{i=1}^n X_i^j\\ \vdots\end{cases} \tag{12}$$

$$\begin{aligned}&\vdots\\ \mu_i=\mathbb{E}X^i\ &=\mathbb{E}S_n^i\\ &\vdots\end{aligned} \tag{13}$$

$$\begin{aligned}\mathbb{C}(S_n^i,S_n^j)=\mathbb{E}(S_n^i-\mathbb{E}S_n^i)(S_n^j-\mathbb{E}S_n^j)=\mathbb{E}S_n^iS_n^j-\mathbb{E}S_n^i\mathbb{E}S_n^j=\mathbb{E}S_n^{i+j}-\mathbb{E}S_n^i\mathbb{E}S_n^j\\ =\mu_{i+j}-\mu_i\mu_j\end{aligned} \tag{14}$$

3.3.2 Method of Moments

$$\begin{aligned}\mu_1=\mathbb{E}X\ &=g_1(\theta_1,\cdots,\theta_k)\\ &\vdots\\ \mu_k=\mathbb{E}X^k&=g_k(\theta_1,\cdots,\theta_k)\end{aligned} \tag{15}$$

$$s_n^j=\frac{1}{n}\sum_i x_i^j\Rightarrow\begin{cases}\mathbb{E}X\ \approx s_n^1=g_1(\hat\theta_1,\cdots,\hat\theta_k)\\ \qquad\vdots\\ \mathbb{E}X^k\approx s_n^k=g_k(\hat\theta_1,\cdots,\hat\theta_k)\end{cases}$$

$$\mathbf{s}_n=\mathbf{g}(\hat{\boldsymbol\theta})\Rightarrow\hat{\boldsymbol\theta}=\mathbf{g}^{-1}(\mathbf{s}_n) \tag{16}$$

If $\mathbf{g}^{-1}$ is differentiable at the point, $\boldsymbol\mu=(\mu_1,\cdots,\mu_k)$, and $\mathbb{E}|X|^{2k}<\infty$, then $\hat{\boldsymbol\theta}$ is an *asymptotically normal estimator* for $\boldsymbol\theta$, since by the delta method (cf. *PN* §9.2.2.2),

$$\sqrt{n}(\hat{\boldsymbol\theta}-\boldsymbol\theta)\xrightarrow{d}\mathrm{N}(\mathbf{0},(\nabla\mathbf{g}^{-1}(\boldsymbol\mu))^\top\boldsymbol\Sigma(\nabla\mathbf{g}^{-1}(\boldsymbol\mu)),\ \ \Sigma_{ij}=\mu_{i+j}-\mu_i\mu_j \tag{17}$$

3.3.3 Maximum-Likelihood Estimators

Given a set of

$$\hat{\theta} \leftarrow \max_{\theta \in \Theta} \ln L_n(\theta|\mathbf{x}) = \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ln p(x_i|\theta) \tag{18}$$

The maximum-likelihood estimator is

- a *consistent estimator* for θ;
- an *asymptotically normal estimator* that satisfies the Cramer-Rao lower bound.

Together these statements imply that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\mathbb{I}_F \mathbf{X}_\theta}\right) \tag{19}$$

The proof follows from the asymptotic properties of the score function (*cf. PN*, §10.4.2):

$$S(\hat{\theta}|\mathbf{X}) \approx S(\theta|\mathbf{X}) + (\hat{\theta} - \theta)\frac{\partial}{\partial \theta} S(\theta|\mathbf{X}) \Rightarrow \sqrt{n}(\hat{\theta} - \theta) \approx \frac{\frac{1}{\sqrt{n}} S(\theta|\mathbf{X})}{-\frac{1}{n}\frac{\partial}{\partial \theta} S(\theta|\mathbf{X})} \tag{20}$$

Using the property that the expectation of the score function vanishes, we apply the central limit theorem (*cf. PN*, §9.2.2) to the numerator,

$$\frac{1}{\sqrt{n}} S(\theta|\mathbf{X}) = \sqrt{n}\left(\frac{1}{n} S(\theta|\mathbf{X}) - \mathbb{E}S(\theta|\mathbf{X})\right) \xrightarrow{d} \mathrm{N}(\mathbb{E}S(\theta|\mathbf{X}), \mathbb{V}S(\theta|\mathbf{X})) = \mathrm{N}\left(0, \mathbb{I}_F \mathbf{X}_\theta\right), \tag{21}$$

we apply the law of large numbers (*cf. PN*, §9.2.1) to the denominator,

$$-\frac{1}{n}\frac{\partial}{\partial \theta} S(\theta|\mathbf{X}) \xrightarrow{p} -\mathbb{E}\frac{\partial}{\partial \theta} S(\theta|\mathbf{X}) = \mathbb{I}_F \mathbf{X}_\theta \tag{22}$$

and conclude via Slutsky's theorem (*cf. PN*, §9.1.5) that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \frac{1}{\mathbb{I}_F \mathbf{X}_\theta} \mathrm{N}(0, \mathbb{I}_F \mathbf{X}_\theta) = \mathrm{N}\left(0, \frac{1}{\mathbb{I}_F \mathbf{X}_\theta}\right). \tag{23}$$

3.3.3.1 Multidimensional Maximum-Likelihood Estimators

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (\mathbb{I}_F \mathbf{X}_{\boldsymbol{\theta}})^{-1}) \tag{24}$$

3.3.4 Bayesian Estimators

Bayes Risk

$$\mathbb{E}_\pi \Lambda(\theta, \hat{\theta}) \tag{25}$$

$$\hat{\theta}^* \leftarrow \max_{\hat{\theta}^* \in \Theta} \mathbb{E}_\pi \Lambda(\theta, \hat{\theta}) \tag{26}$$

$$\Lambda(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2 \Rightarrow 0 = \frac{\partial}{\partial\hat{\theta}}\mathbb{E}_\pi(\theta-\hat{\theta})^2\Big|_{\hat{\theta}=\hat{\theta}^*} = -2\mathbb{E}_\pi(\theta-\hat{\theta}^*) \Rightarrow \hat{\theta}^* = \mathbb{E}_\pi\theta \tag{27}$$

Iterative method

$$\pi(\theta) \equiv \pi(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\pi(\theta)}{p(\mathbf{x})} \tag{28}$$

$$\cdots \mathbf{x}_i, \theta_i \Rightarrow \hat{\theta}_{i+1} : \mathbf{x}_{i+1}, \theta_{i+1} \Rightarrow \hat{\theta}_{i+2}, \cdots \tag{29}$$