Statistics (Notes)

Mark DiBattista

January 5, 2020

Abstract

Statistics is a branch of probability theory that addresses questions of probabilistic modeling, usually the suitability of a hypothesized distribution or the presence/absence of a specific distributional property for a given set of measurements. Since statistics are simply functions of random variables, they are themselves random variables to which the tools of probability theory can be applied. There is however, a difference in emphasis between the two: probability theory is mainly concerned with analytical properties of parametrized functions, whereas statistics, with its attendant specialized vocabulary, mainly addresses consistency of distributions with randomly sampled data.

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 independently 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 statistics-based calculations that are run against sample data to infer properties of governing parameters:

- **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.

Point estimators are based on *test statistics*, $T(\mathbf{X})$, which, as functions of random variables, are themselves random variables. The first and second moments of the distribution associated with the test statistic provide quantitative information on the estimator, and are assigned characteristic terminology:

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

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.

It is frequently the case that, for point estimators, small increases in bias can be traded for significant reduction in mean-squared error.

2.2.2 Properties of Point Estimators

Point estimators based on test statistics are differentiated 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. In quantitative terms this is expressed as absence of the parameter in the distribution of data conditioned on the test statistic:

$$p_{\mathbf{X};\theta|T(\mathbf{X})}(\mathbf{x};\theta|t(\mathbf{x})) = h(\mathbf{x}). \tag{4}$$

  Given the value of the statistic, $T(\mathbf{X})$, no additional information is afforded by the parameter, θ. A necessary and sufficient condition is given by the **Fisher-Neyman Theorem**:

$$p(\mathbf{X}|\theta) = h(\mathbf{X})g(\theta, T(\mathbf{X})) \Leftrightarrow T(\mathbf{X}) \text{ is a sufficient statistic,} \tag{5}$$

  which is proven by executing a change of variable from the space of data into a space of statistics, only one of which (the test statistic) depends on the parameter. The partition of the joint distribution, $p_{\mathbf{X}}|\theta$, into marginal and conditional components with the required properties follows immediately. Note that the form of the exponential family of distributions – *cf. PN*, §7.4 – explicitly references a sufficient statistic, and is one for which the parameter and statistic are simple multiples that form the exponential argument.

- A **complete statistic** is one for which distinct parameters govern different probability distributions:

$$\mathbb{E}_\theta g(T(\mathbf{X}) = 0 \Rightarrow \mathbb{P}_\theta\{g(T(\mathbf{X}) = 0\} = 1. \tag{6}$$

  Completeness of a statistic is essentially identical to the *identifiability* of a statistical model.

These notions are united in the **Lehmann-Scheffé theorem**: an unbiased, sufficient and complete statistic is the unique MVUE for the parameter.

2.2.3 Asymptotic Properties of Statistics

Given an increasing vector of random variables, $\mathbf{X}_n = (X_1, \cdots, X_n), X_i \sim X$, for which each element in the series is sampled from the same parametrized distribution, $X \equiv X|\theta$, and a matching 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,

$$T(\mathbf{X}_n) \xrightarrow{p} \theta \tag{7}$$

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

$$\sqrt{n}(T(\mathbf{X}_n) - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma^2) \tag{8}$$

3 Point Estimation

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

$$\mathbf{X} = (X_1, \cdots, X_n), X_i \sim X \equiv X|\theta \Rightarrow T(\mathbf{X}) \approx \theta. \tag{9}$$

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,

$$\left.\begin{array}{l}\mathbf{X}_\theta = (X_1, \cdots, X_n), X_i \sim X \equiv X|\theta \\ \mathbb{B}T(\mathbf{X}_\theta) = 0\end{array}\right\} \Rightarrow \mathbb{V}T(\mathbf{X}_\theta) \geq \frac{1}{\mathbb{I}_F \mathbf{X}_\theta}, \tag{10}$$

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:

$$\mathbb{V}T(\mathbf{X}_\theta) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}T(\mathbf{X}_\theta)\right)^2}{\mathbb{I}_F \mathbf{X}_\theta}. \tag{11}$$

The derivation of the Cramer-Rao theorem follows directly from the 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,

$$\begin{aligned}\mathbb{C}(S(\theta|\mathbf{X}_\theta), T(\mathbf{X}_\theta)) &= \mathbb{E}S(\theta|\mathbf{X}_\theta)T(\mathbf{X}_\theta) - \mathbb{E}S(\theta|\mathbf{X}_\theta) \cdot \mathbb{E}T(\mathbf{X}_\theta) = \mathbb{E}S(\theta|\mathbf{X}_\theta)T(\mathbf{X}_\theta) \\ &= \int_D p(\mathbf{x}|\theta)T(\mathbf{x})\frac{\partial}{\partial\theta}\ln p(\mathbf{x}|\theta)\, d\mathbf{x} = \int_D T(\mathbf{x})\frac{\partial}{\partial\theta}p(\mathbf{x}|\theta)\, d\mathbf{x} = \frac{\partial}{\partial\theta}\mathbb{E}T(\mathbf{X}_\theta),\end{aligned} \tag{12}$$

the Cauchy-Schwarz inequality yields

$$\mathbb{V}(S(\theta|\mathbf{X}_\theta) \cdot \mathbb{V}T(\mathbf{X}_\theta) \geq \left(\mathbb{C}(S(\theta|\mathbf{X}_\theta), T(\mathbf{X}_\theta))\right)^2 \Rightarrow \mathbb{V}T(\mathbf{X}_\theta) \geq \frac{\left(\mathbb{C}(S(\theta|\mathbf{X}_\theta), T(\mathbf{X}_\theta))\right)^2}{\mathbb{V}(S(\theta|\mathbf{X}_\theta)} = \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}T(\mathbf{X}_\theta)\right)^2}{\mathbb{I}_F \mathbf{X}_\theta}. \tag{13}$$

3.2 Rao-Blackwell Theorem

Given a distribution parametrized by θ, the variance of any arbitrary estimator, $U(\mathbf{X})$, does not increase – and often improves – by conditioning the estimator with a sufficient statistic, $T(\mathbf{X})$. Defining the estimators,

$$\hat{\theta} = \mathbb{E}_\theta(U(\mathbf{X})) \tag{14}$$

$$\theta^* = \mathbb{E}_\theta(\hat{\theta}|T(\mathbf{X})) \tag{15}$$

we can show that the variance of the conditioned estimator is never larger,

$$\mathbb{E}_\theta(\theta^* - \theta)^2 = \mathbb{E}_\theta(\mathbb{E}_\theta(\hat{\theta}|T(\mathbf{X}) - \theta)^2 = \mathbb{E}_\theta(\mathbb{E}_\theta(\hat{\theta} - \theta|T(\mathbf{X}))^2 \leq \mathbb{E}_\theta(\mathbb{E}_\theta((\hat{\theta} - \theta^2|T(\mathbf{X})) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2. \tag{16}$$

3.3 Methods for Generating Point Estimators

There are three common bases upon which point estimators are constructed for distribution parameters given a sequence of data points:

- the *sample moments* of the data points, for which the moments of the distribution are linked through a system of equations;
- the *maximum likelihood* of the sample distribution given the data points, for which the Shannon information is leveraged;
- the *Bayes' risk* of the sequence of sample distributions, for which an informed prior can be assigned.

3.3.1 Point Estimators for Moments

Given a random vector of data points, each sampled independently from the same distribution, we can generate the random vector of moments by simple averages of samples:

$$\mathbf{X} = (X_1, \cdots, X_n), X_i \sim X \Rightarrow \begin{cases} \vdots \\ S_n^j = \frac{1}{n}\sum_{i=1}^n X_i^j \\ \vdots \end{cases} \tag{17}$$

which shows an arbitrary element of the random moment vector, $\mathbf{S}_n = (S_n^1, \cdots, S_n^n)$, constructed from the first n data points.

Assigning the subscripted symbol to the expectation of the sample moment

$$\begin{aligned} &\vdots \\ \mu_i = \mathbb{E}X^i &= \mathbb{E}S_n^i \\ &\vdots \end{aligned} \tag{18}$$

the equivalence of moments is guaranteed by linearity of the averaging operator and the law of large numbers (*cf. PN*, §9.2.1), and the covariance matrix of sample moments is expressed as simple combinations of sample expectations,

$$\begin{aligned} \mathbb{C}(S_n^i, S_n^j) = \mathbb{E}(S_n^i - \mathbb{E}S_n^i)(S_n^j - \mathbb{E}S_n^j) = \mathbb{E}S_n^i S_n^j - \mathbb{E}S_n^i \mathbb{E}S_n^j = \mathbb{E}S_n^{i+j} - \mathbb{E}S_n^i \mathbb{E}S_n^j \\ = \mu_{i+j} - \mu_i \mu_j. \end{aligned} \tag{19}$$

3.3.2 Method of Moments

Given a random variable, X, parametrized by a vector, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^\top$, the information in the distribution is reduced to an independent system of equations of moments that are functions of the parameters alone:

$$\begin{aligned} \mu_1 &= \mathbb{E}X &&= g_1(\theta_1, \cdots, \theta_k) \\ &\vdots \\ \mu_k &= \mathbb{E}X^k &&= g_k(\theta_1, \cdots, \theta_k) \end{aligned} \tag{20}$$

And given a set of data points, $\mathbf{x} = (x_1, \cdots, x_n)$, the true moments are replaced with sample values, $\mathbf{s}_n = (s_n^1, \cdots, s_n^k)$,

$$s_n^j = \frac{1}{n}\sum_i x_i^j \Rightarrow \begin{cases} \mathbb{E}X &\approx s_n^1 = g_1(\hat{\theta}_1, \cdots, \hat{\theta}_k) \\ &\vdots \\ \mathbb{E}X^k &\approx s_n^k = g_k(\hat{\theta}_1, \cdots, \hat{\theta}_k) \end{cases} \tag{21}$$

from which estimators, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)^\top$, are inferred through the inverse relation,

$$\mathbf{s}_n = \mathbf{g}(\hat{\boldsymbol{\theta}}) \Rightarrow \hat{\boldsymbol{\theta}} = \mathbf{g}^{-1}(\mathbf{s}_n). \tag{22}$$

If $\mathbf{g}^{-1}$ is differentiable at the point, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_k)^\top$, 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),

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (\nabla \mathbf{g}^{-1}(\boldsymbol{\mu}))^\top \boldsymbol{\Sigma} (\nabla \mathbf{g}^{-1}(\boldsymbol{\mu})), \;\; \Sigma_{ij} = \mu_{i+j} - \mu_i \mu_j. \tag{23}$$

3.3.3 Maximum-Likelihood Estimators

The method of maximum likelihood is based on properties of Shannon information, which is covered in *PN*, §10. The method is based on the log-likelihood function,

$$\mathbb{I}_S \mathbf{X}_\theta \equiv \ln L(\theta|\mathbf{X}) \to \ln p(\mathbf{x}|\theta), \tag{24}$$

for which emphasis is shifted from the probability domain of unknown data points with known parameter to the statistics domain of known data points and unknown parameter. The gradient of the log-likelihood function with respect to the parameter is a random variable known as the score function, whose expectation vanishes and whose variance is defined as Fisher information:

$$S(\theta|\mathbf{X}) = \frac{\partial}{\partial \theta} \ln L(\theta|\mathbf{X}) \Rightarrow \begin{cases} \mathbb{E}_\theta S(\theta|\mathbf{X}) = 0 \\ \mathbb{V}_\theta S(\theta|\mathbf{X}) = \mathbb{I}_F \mathbf{X}_\theta \end{cases} \tag{25}$$

The method of maximum likelihood asserts that the most-probable estimate for the parameter given the measured data is the extremum of the score function. Asymptotic properties, as well as specific numeric methods for calculating estimates, are determined by the moments defined in (25).

3.3.3.1 Univariate Maximum-Likelihood Estimators

Given an admissible set of parameters, $\theta \in \Theta$, the maximum-likelihood estimator, $\hat{\theta}$, is one that, given the sample data, $\mathbf{x}$, maximizes the log-likelihood function,

$$\hat{\theta} \leftarrow \max_{\theta \in \Theta} \ln L_n(\theta|\mathbf{x}) = \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ln p(x_i|\theta). \tag{26}$$

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

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\mathbb{I}_F \mathbf{X}_\theta}\right) \tag{27}$$

The proof follows from the asymptotic properties of the score function (*cf. PN*, §10.4.2), whose Taylor expansion takes the form

$$S(\hat{\theta}|\mathbf{X}) \approx S(\theta|\mathbf{X}) + (\hat{\theta} - \theta) \frac{\partial}{\partial \theta} S(\theta|\mathbf{X}) \Rightarrow \sqrt{n}(\hat{\theta} - \theta) \approx \frac{\frac{1}{\sqrt{n}} S(\theta|\mathbf{X})}{-\frac{1}{n} \frac{\partial}{\partial \theta} S(\theta|\mathbf{X})} \tag{28}$$

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,

$$\frac{1}{\sqrt{n}}S(\theta|\mathbf{X}) = \sqrt{n}\left(\frac{1}{n}S(\theta|\mathbf{X}) - \mathbb{E}S(\theta|\mathbf{X})\right) \xrightarrow{d} \mathrm{N}(\mathbb{E}S(\theta|\mathbf{X}), \mathbb{V}S(\theta|\mathbf{X})) = \mathrm{N}(0, \mathbb{I}_F\mathbf{X}_\theta), \tag{29}$$

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

$$-\frac{1}{n}\frac{\partial}{\partial\theta}S(\theta|\mathbf{X}) \xrightarrow{p} -\mathbb{E}\frac{\partial}{\partial\theta}S(\theta|\mathbf{X}) = \mathbb{I}_F\mathbf{X}_\theta \tag{30}$$

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

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \frac{1}{\mathbb{I}_F\mathbf{X}_\theta}\mathrm{N}(0, \mathbb{I}_F\mathbf{X}_\theta) = \mathrm{N}\left(0, \frac{1}{\mathbb{I}_F\mathbf{X}_\theta}\right). \tag{31}$$

3.3.3.2 Multidimensional Maximum-Likelihood Estimators

The extension of maximum-likelihood estimation to the multidimensional setting is quite straightforward: the univariate parameter is replaced by a vector, $\theta \to \boldsymbol{\theta}$, the first and second moments of the score function are replaced by vector- and matrix-valued quantities, $\mathbf{0}$ and $\mathbb{I}_F\mathbf{X}_{\boldsymbol{\theta}}$, respectively, and the solution to the extremal equation,

$$\hat{\boldsymbol{\theta}} \leftarrow \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ln L_n(\boldsymbol{\theta}|\mathbf{x}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^{n} \ln p(x_i|\boldsymbol{\theta}). \tag{32}$$

The Taylor expansion and asymptotic arguments in (28) – (31) are extended to cover the vector-matrix quantities for first and second moments, with the analogous result for asymptotic normality,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (\mathbb{I}_F\mathbf{X}_{\boldsymbol{\theta}})^{-1}). \tag{33}$$

3.3.3.3 Numerical Methods for Maximum-Likelihood Estimators

3.3.4 Bayesian Estimators

Bayes Risk

$$\mathbb{E}_\pi\Lambda(\theta, \hat{\theta}) \tag{34}$$

$$\hat{\theta}^* \leftarrow \max_{\hat{\theta}^*\in\Theta} \mathbb{E}_\pi\Lambda(\theta, \hat{\theta}) \tag{35}$$

$$\Lambda(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 \Rightarrow 0 = \frac{\partial}{\partial\hat{\theta}}\mathbb{E}_\pi(\theta - \hat{\theta})^2\bigg|_{\hat{\theta}=\hat{\theta}^*} = -2\mathbb{E}_\pi(\theta - \hat{\theta}^*) \Rightarrow \hat{\theta}^* = \mathbb{E}_\pi\theta \tag{36}$$

Iterative method

$$\pi(\theta) \equiv \pi(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\pi(\theta)}{p(\mathbf{x})} \tag{37}$$

$$\cdots \mathbf{x}_i, \theta_i \Rightarrow \hat{\theta}_{i+1} : \mathbf{x}_{i+1}, \theta_{i+1} \Rightarrow \hat{\theta}_{i+2}, \cdots \tag{38}$$