Statistics (Notes)

Mark DiBattista

January 28, 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 hypothesized distributions or unknown parameters 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/Setting

The basic probabilistic setting for the study of parametrized models – a major branch of statistics – consists of the following items:

- Uncertainty in a physical quantity or process is modeled by a random variable, frequently governed by an unknown parameter, represented as $X|\theta$;
- A sequence of measurements of the underlying quantity or process is modeled by a random vector, $\mathbf{X} = (X_1, \cdots, X_n), X_i \sim X \equiv X|\theta$, for which each coordinate follows the common parametrized distribution. Should each coordinate measurement be assumed independent of the others – independently and identically distributed (IID) – then the covariance of distinct coordinates vanishes, $\mathbb{C}(X_i, X_j) = 0, i \neq j$.
- The specific vector of data points generated by measurement is represented in lower case, $\mathbf{x}$, and observations enter into statistical analyses through the condition, $\mathbf{X} = \mathbf{x}$;
- A **statistic** is any function of a random variable, usually a random vector, $T(\mathbf{X})$. A **test statistic** is one designed for quantitative inference about the values or properties of the underlying population model. The population and measured properties are linked through the test statistic, which is itself a random variable, and the probability distribution that it generates;
- Assertions on population properties are evaluated from the expected probabilities of observed test statistics, given the proposed underlying models.

2.2.2 Statistical Operations

The specification of an underlying population distribution, $X|\theta$, is a modeling choice. Sample data, $\mathbf{X} = \mathbf{x}$, is generated through physical measurement. Statistics, the mathematical branch, links measurement to model, using functions of sample data, $T(\mathbf{X})$, to assess proposed values for the parameter, or any other asserted property of the underlying population. There are two distinct types of analyses commonly run:

- **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 an assertion 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 and interval tests 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.

2.2.3 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.

2.2.4 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

Following the terminology introduced above in §2.2.1, a **point estimator** for a model parameter is a sample statistic derived from the IID joint distribution,

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

that serves to approximate the parameter. A point estimator is therefore a random variable, governed by a probability distribution, and with full rights to 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

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

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,

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

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

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

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

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,

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

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:

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

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

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

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

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

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),

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

3.3.3 Maximum-Likelihood Estimators

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

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

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**:

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

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,

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

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,

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

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

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

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

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

Iterative method

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

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

4 Interval Testing

Following the terminology introduced above in §2.2.1, an **interval test** for a model parameter is a quantitative measure applied to an assertion on point estimates, especially with respect to the range of values in which test statistics might fall.

As described above in §3, the point estimator is derived from an ensemble of points, usually sampled independently, $\mathbf{X} = (X_1, \cdots, X_n)$, $X_i \sim X \equiv X|\theta$, and is generated by the statistic, $T(\mathbf{X})$, that varies in a wll-know fashion with the parameter. Here, assertions of the interval estimator are defined through the probability measure that governs the test statistic, $\mathbb{P} \equiv \mathbb{P}_{T(\mathbf{X})}$, so that for any given subset of the parameter space, $\mathcal{A} \subset \Theta$, and any given data sample, $\mathbf{X} = \mathbf{x}$, the interval estimator is evaluated from the quantity,

$$\mathbf{X} = (X_1, \cdots, X_n),\ X_i \sim X \equiv X|\theta \Rightarrow \mathbb{P}\{(T(\mathbf{X}) \in \mathcal{A})|\mathbf{X} = \mathbf{x}\}. \tag{39}$$

4.1 Hypothesis Testing and Interval Estimators

4.1.1 Hypotheses

- A **hypothesis** is an assertion about the model distribution, $X|\theta$, usually with respect to the range in which the key parameter, θ, falls. If the set of possible parameter values, $\theta \in \Theta$, is split into two disjoint sets, Θ_0 and Θ_1, the complementary **null** and **alternative hypotheses**, H_0 and H_1, are the assertions,

$$\left.\begin{array}{l} \Theta_0 \cup \Theta_1 = \Theta \\ \Theta_0 \cap \Theta_1 = \emptyset \end{array}\right\} \Rightarrow \begin{cases} \text{null hypothesis:} & H_0 : \theta \in \Theta_0 \\ \text{alternative hypothesis:} & H_1 : \theta \in \Theta_1 \end{cases} \tag{40}$$

- A **test statistic** is a random variable that is a function of the data sample, $T(\mathbf{X})$. A realization of the test statistic given an observation, $\mathbf{x}$, is expressed in lower case symbols,

$$t(\mathbf{x}) \equiv T(\mathbf{X})|(\mathbf{X} = \mathbf{x}). \tag{41}$$

  The test statistic may be designed to approximate the population parameter, $T(\mathbf{X}) \approx \theta$, or it may take on a range of distributions that vary with the parameter in some planned fashion;

- The probability measure associated with the test statistic is parametrized by the population parameter,

$$\mathbb{P} \equiv \mathbb{P}_{T(\mathbf{X})|\theta} \tag{42}$$

- For any given parameter values, θ, only one of the hypotheses, H_0 or H_1, in (40) can be true. A **rejection region** is a subset of the range of the test statistic, usually a disjoint union of real intervals, $\mathcal{R} \subset \mathbb{R}$, constructed such that membership within is sufficiently improbable as to imply the likely falsity of the null hypothesis:

$$t(\mathbf{x}) \in \mathcal{R} \quad : \text{reject the null hypothesis} \tag{43}$$

$$t(\mathbf{x}) \in \mathcal{R}^C : \text{accept the null hypothesis} \tag{44}$$

- Given a decomposition of the parameter set, Θ, into disjoint subsets, Θ_0 and Θ_1, we may design a test statistic, $T_\theta(\mathbf{X})$, that determines the correct set to which any given parameter value belongs. In this case two kinds of errors may occur, for the two cases in which property and the inference disagree,

$$\text{Type I error: with probability } \mathbb{P}_{\theta|\theta\in\Theta_0}\{t(\mathbf{x}) \in \mathcal{R}\} \quad \Rightarrow \theta \in \Theta_0 \text{ and } t(\mathbf{x}) \in \mathcal{R} \tag{45}$$

$$\text{Type II error: with probability } \mathbb{P}_{\theta|\theta\in\Theta_1}\{t(\mathbf{x}) \in \mathcal{R}^C\} \Rightarrow \theta \in \Theta_1 \text{ and } t(\mathbf{x}) \in \mathcal{R}^C \tag{46}$$

- The **power function** of a hypothesis test shows the correctness of the inference:

$$\beta(\theta) = \mathbb{P}_{\theta|\theta\in\Theta_0}\{t(\mathbf{x}) \in \mathcal{R}\} = 1 - \mathbb{P}_{\theta|\theta\in\Theta_1}\{t(\mathbf{x}) \in \mathcal{R}^C\}. \tag{47}$$

  The power function is sharpest when there are no errors for any underlying parameter,

$$\beta(\theta) = \begin{cases} 0 \text{ for all } \theta \in \Theta_0 \\ 1 \text{ for all } \theta \in \Theta_1 \end{cases} \tag{48}$$

- A **level-α test** for a hypothesis test is for which the power function is bounded when the null hypothesis is true:

$$\arg\max_{\theta\in\Theta_0} \beta(\theta) \leq \alpha, \quad 0 \leq \alpha \leq 1. \tag{49}$$

- Designate the set of all level-α test as $\mathcal{C}_\alpha$. The **uniformly most powerful test** level-α test is one, $\beta \in \mathcal{C}_\alpha$, such that for any other level-α test, $\beta' \neq \beta, \beta' \in \mathcal{C}_\alpha$,

$$\beta(\theta) \geq \beta'(\theta) \text{ for all } \theta \in \Theta_1. \tag{50}$$

- A **p-value** is a test statistic, $p(\mathbf{X})$, whose realized values are bounded within the range of cumulative probabilities, $0 \leq p(\mathbf{X}) \leq 1$. A p-value is *valid*, provided that

$$\mathbb{P}_{\theta|\theta\in\Theta_0}\{p(\mathbf{x} \leq \alpha\} \leq \alpha. \tag{51}$$

  If the same bound applies to the test statistic and the probability measure that it generates, then the valid p-test serves as a level-α test. Small p-values, $p(\mathbf{x}) \leq \alpha \ll 1$, imply that the alternative hypothesis, H_1, is the correct one.

The skill in hypothesis test design is, given the parametrized model distribution, $X|\theta$, to craft the test statistic, $T(\mathbf{X})$, and acceptance region, $\mathcal{R}$, so that the resulting power function is sharp. It is common to use p-values as test statistics, and acceptance regions defined by the tails of the distribution.

4.1.2 Interval Tests

An **interval estimator** for a population parameter, θ, is a *pair* of test statistics, $A(\mathbf{X})$ and $B(\mathbf{X})$, such that one dominates the other, $A(\mathbf{x}) \leq B(\mathbf{x})$, for all samples, $\mathbf{x}$. The test statistics are selected to serve as bounds on the parameter, $A(\mathbf{x}) \leq \theta \leq B(\mathbf{x})$.

- The **coverage probability** for an interval test is probability that the population parameter falls within the bounds,
$$\mathbb{P}_\theta\{\theta \in [A(\mathbf{x}), B(\mathbf{x})]\}; \tag{52}$$
- The **confidence coefficient** is the smallest coverage probability over all population parameters for a given data sample,
$$\arg\min_\theta \mathbb{P}_\theta\{\theta \in [A(\mathbf{x}), B(\mathbf{x})]\}; \tag{53}$$
- A **confidence interval** is an interval estimator endowed with a confidence coefficient.

The confidence interval is a kind of hypothesis test: any given acceptance region – a set in the sample space – can be matched to a interval estimator defined in the parameter space. The function pair, $A(\mathbf{X})$ and $B(\mathbf{X})$, can be expressed as p-value – frequently in terms of their union – and interpreted as probability measures.

5 Statistical Tests

5.1 Common Interval Tests Based on Likelihood Functions

Three common interval tests for parametrized models, $X|\theta$, are derived from the likelihood function:

- Likelihood Ratio Test
- Wald Test
- Score Test

As with all interval tests, the null and alternative hypotheses are defined in terms of parameter membership in one of two disjoint sets, *i.e.*, $\theta \in \Theta$, for which $\Theta = \Theta_0 \cup \Theta_1$ and $\emptyset = \Theta_0 \cap \Theta_1$, so that

$$\text{null hypothesis:} \quad H_0 : \theta \in \Theta_0 \tag{54}$$
$$\text{alternative hypothesis:} \quad H_1 : \theta \in \Theta_1 \tag{55}$$

All three tests are designed for **single-point hypotheses** – the case for which $\Theta_0 = \{\theta_0\}$. Here, the three tests differ in the computational burden placed on likelihood functions: the Wald test requires information only at the global maximum, the score test requires information only at the hypothesized value, while the likelihood ratio test requires the evaluation at both.

The key insight behind all three tests is identical: should the null hypothesis be correct, then the maximum-likelihood estimator based on randomly sampled data would be close by, for which distance is measured in probabilistic terms. Although the details of each test differ for finite samples – and conclusions may differ in edge cases – asymptotically, all converge to the chi-square distribution.

Whereas the Wald and score tests handle only single-point hypotheses, the likelihood ratio test can be worked into the more general setting in which the null value falls within a specified range.

5.1.1 Likelihood Ratio Test

The **likelihood ratio test** is an interval test for a population parameter, $\theta \in \Theta$, given a model distribution, $X|\theta$. The test statistic for the likelihood ratio scales the likelihood of a proposed value, θ_0, or maximum of a set of values, against the *global* maximum, $\hat{\theta}$, which ensures the interpretion of the test statistic as a cumulative propability distribution. The test leverages well-known properties of likelihood functions, score functions and Fisher information (again, *cf. PN*, §10) to generate the machinery of the interval testing – such as rejection regions, level-α sets and probability measures – as well as to ensure the asymptotic normality of the estimator.

The hypothesized values of the population parameter for the model distribution might fall in one of two disjoint sets,

$$\text{null hypothesis:} \qquad H_0 : \theta \in \Theta_0 \tag{56}$$

$$\text{alternative hypothesis:} \quad H_1 : \theta \in \Theta_1 \tag{57}$$

Given the random variable for data sampled identically and independently from the common model distribution, $\mathbf{X} = (X_1, \cdots, X_n)$, the test statistic is formed from the ratio of likelihoods, one taken as the restricted maximizer of the set that comprises the null hypothesis, $\theta \in \Theta_0$, and the other as the global maximizer over all parameters, $\theta \in \Theta$,

$$\lambda(\mathbf{X}) = \frac{\underset{\theta \in \Theta_0}{\arg\max}\, L(\theta|\mathbf{X})}{\underset{\theta \in \Theta}{\arg\max}\, L(\theta|\mathbf{X})}. \tag{58}$$

For any given data sample, $\mathbf{x}$, the likelihood function has a single global maximum, with curvature that is everywhere positive. The likelihood ratio test therefore takes unit value at the global maximizer, $\theta_0 = \hat{\theta}$, and 'high values' at positions close by. The likelihood ratio test links a distance metric to probabilities generated by random sampling, which justifies acceptance or rejection of the asserted hypothesis.

- The range of the likelihood ratio is the unit interval, $0 \leq \lambda(\mathbf{x}) \leq 1$, and the null hypothesis is satisfied precisely at the endpoint, $\lambda = 1$;
- The null hypothesis is supported by sample data that lie in sets close to the global maximum, which is codified in rejection sets defined by threshold values, $0 \leq c \leq 1$,

$$\mathcal{R} = \{\mathbf{x} \in \mathbb{R}^n : \lambda(\mathbf{x}) \leq c\}; \tag{59}$$

- Given the probability measure, $\mathbb{P}_\theta$, generated by the test statistic, the level-α test is specified by choosing the threshold, c, such that

$$\underset{\theta \in \Theta_0}{\arg\min}\, \mathbb{P}_\theta\{\lambda(\mathbf{x}) \leq c\} = \alpha. \tag{60}$$

  Since the threshold, c, is itself an element of the unit interval, the likelihood ratio is a p-value;
- The curvature of the likelihood function is positive everywhere, and provided by the Fisher information, $\mathbb{I}_F X_\theta$, at the maximal point. The curvature at nearby points is approximately the same, and can be used to estimate the probability that the most-likely parameter values estimated from randomly generated data points differ from the true maximum value.

The special case of the **single-point hypothesis** occurs when the null hypothesis consists of a single value,

$$\Theta_0 = \{\theta_0\}, \tag{61}$$

for which the likelihoods of the restricted and global maximizers are given by

$$\left.\begin{array}{ll} \text{restricted maximizer:} & \theta_0 \\ \text{global maximizer:} & \hat{\theta} \end{array}\right\} \Rightarrow \lambda(\mathbf{X}) = \frac{L(\theta_0|\mathbf{X})}{L(\hat{\theta}|\mathbf{X})} \tag{62}$$

The asymptotic properties of the single-point hypothesis test are derived from a Taylor expansion of the likelihood in the numerator about the global maximum, for which the logarithm of the result, truncated to the second-order term, is the product,

$$\begin{aligned} -2\ln\lambda(\mathbf{X}) &\approx -2\left(\ln L(\hat{\theta}|\mathbf{X}) - \ln L(\hat{\theta}|\mathbf{X}) + (\theta_0-\hat{\theta})S(\hat{\theta}|\mathbf{X}) + \frac{1}{2}(\theta_0-\hat{\theta})^2\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X})\right) \\ &\approx -(\theta_0-\hat{\theta})^2\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X}). \end{aligned} \tag{63}$$

The derivation of the asymptotic distribution follows the same arguments as presented above in §3.3.3 for maximum-likelihood point estimators, since we have

$$\left.\begin{array}{l} -\frac{1}{n}\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X}) \xrightarrow{p} \mathbb{I}_F\mathbf{X}_\theta \\ \sqrt{n}(\theta_0-\hat{\theta}) \xrightarrow{d} \mathrm{N}\left(0,\mathbb{I}_F\mathbf{X}_\theta\right) \end{array}\right\} \Rightarrow \sqrt{n}(\theta_0-\hat{\theta})\sqrt{-\frac{1}{n}\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X})} \xrightarrow{d} \mathrm{N}(0,1) \tag{64}$$

The logarithm of the likelihood ratio is the square of the result in (64), so that we immediately derive the chi-square distribution in the asymptotic limit,

$$-2\ln\lambda(\mathbf{X}) \approx -(\theta_0-\hat{\theta})^2\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X}) = \left(\sqrt{n}(\theta_0-\hat{\theta})\sqrt{-\frac{1}{n}\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X})}\right)^2 \xrightarrow{d} \chi^2. \tag{65}$$

5.1.2 Wald Test

The **Wald Test** is an interval test for the single-point hypothesis formed from the ratio of the square distance between hypothesized and maximum values, and the variance of the global estimator. The asymptotic limit of these quantities can be derived from quantities expressed above in §§3.3.3.1 & 5.1.1. Assuming that the null hypothesis holds – the restricted maximizer, θ_0, is the true parameter value – then we have

$$\frac{(\theta_0-\hat{\theta})^2}{\mathbb{V}\hat{\theta}} = \mathbb{I}_F\mathbf{X}_\theta\,(\theta_0-\hat{\theta})^2 \xrightarrow{d} \chi^2. \tag{66}$$

Note that the Wald test only requires quantities evaluated at the unrestricted global maximum, $\hat{\theta}$.

5.1.3 Score Test

$$\begin{aligned} \frac{1}{\mathbb{I}_F\mathbf{X}_\theta}\left(S(\theta_0|\mathbf{X})\right)^2 &\approx \frac{1}{\mathbb{I}_F\mathbf{X}_\theta}\left(S(\hat{\theta}|\mathbf{X}) + (\theta_0-\hat{\theta})\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X})\right)^2 \\ &= \frac{1}{\mathbb{I}_F\mathbf{X}_\theta}\left(\sqrt{n}(\theta_0-\hat{\theta})\sqrt{-\frac{1}{n}\frac{\partial}{\partial\theta}S(\hat{\theta}|\mathbf{X})}\right)^2 \xrightarrow{d} \chi^2 \end{aligned} \tag{67}$$

Note that the score test only requires quantities evaluated at the null parameter, θ_0.

5.1.4 The Neyman-Pearson Lemma and Uniformly Most Powerful Tests

The **Neyman-Pearson Lemma**

5.2 Standard Tests

5.2.1 Discrete Models

5.2.1.1 Mean Frequencies of Multinomial Models: Pearson's Chi-square Test

5.2.1.2 Mean Frequencies of Multinomial Models: Fisher's Exact Test

5.2.2 Continuous Models

5.2.2.1 Mean of Gaussian Model: T-test

5.2.2.2 Common Means of Multiple Gaussian Models: F-Test