STA 532: Theory of Statistical Inference

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 Models & Inference

Introduction

In this introductory lecture we get a brief glimpse of most of the topics to be covered in the course STA 532, as a sort of preview.

In Probability Theory, we begin with a known probability distribution (perhaps one with a famous name, like Binomial or Poisson, or perhaps just the CDF $F(x) = \mathsf{P}[X \le x]$ in some convenient form) and we try to make predictions about properties of one or many random variables $\{X_j\}$ with that distribution.

Statistics addresses the **Inverse Problem**: we begin with observations $\{X_j\}$ of a collection of random variables, and then we try to guess (properties of) their distribution, like the CDF $F(\cdot)$.

Parametric & Nonparametric Models

Two Principal Paradigms

Fisherian

A central 90% interval estimate for a real-valued parameter $\theta \in \Theta \subseteq \mathbb{R}$ is a pair of statistics (*i.e.*, functions of the data) $L(X)$ and $R(X)$ with the properties that

$$(\forall\theta)\mathsf{P}_\theta[\theta < L(X)] \le 0.05 \qquad\qquad (\forall\theta)\mathsf{P}_\theta[\theta > R(X)] \le 0.05$$

from which it follows that

$$(\forall\theta)\mathsf{P}_\theta\left\{\theta \in [L(X),\ R(X)]\ \right\} \ge 0.90$$

where "P_θ" means we compute the probability of different possible values of X for the specified value of the parameter θ. This is in some sense a separate statement for every possible $\theta \in \Theta$, and the probability bounds must hold for all of them. The parameter θ is a fixed real number in each statement, while X and hence $L(X)$ and $R(X)$ are random variables.

Bayesian

In the Bayesian perspective (Bayes, 1763; Laplace, 1774) both θ *and* X are "random", with a joint distribution that may have a density function of the form $\pi(\theta)\, f(x \mid \theta)$. A central 90% interval for

θ is now a pair of statistics $L(X)$, $R(X)$ with the property that

$$\mathsf{P}[\theta < L(X) \mid X] \le 0.05 \qquad \mathsf{P}[\theta > R(X) \mid X] \le 0.05$$

and hence

$$\mathsf{P}[\theta \in [L(X),\ R(X)] \mid X] \ge 0.90$$

where, here, "P" represents the joint probability distribution for X and θ.

Comparing These

The difference is just what we treat as *known* (we "condition" on these) and what we treat as *random* (we "quantify" over these). The Fisherian treats the parameter θ as known, and quantifies over possible values of the observable data X; the Bayesian treats the observed data X as known, and quantifies over the uncertain parameter values.

In this course we will present both views, and note where they agree and where they do not. We will sometimes encounter other related paradigms, like "Neyman-Pearson" and "fiducial" (also due to Fisher) and "likelihoodist", but none of these is as commonly used and scientifically important in this century as the Bayesian and Fisherian. The Fisherian and Neyman-Pearson approaches are sometimes lumped together and called called "frequentist" or "sampling based" or, a bit ironically in view of their mid-twentieth century introduction (Fisher, 1922, 1925, 1935; Neyman and Pearson, 1933), "classical".

Fundamental Concepts

A *Statistical Model* for an observation vector X is a family of probability distributions on the "outcome space" $\mathcal{X}$ of all possible values that X might take.

The outcome space $\mathcal{X}$ might be discrete (typically finite or countable) or continuous; it may be scalar or vector or even infinite-dimensional. The cases we will see most often are finite sets $\mathcal{X}$, the integers ($\mathbb{N} = \{1, 2, ...\}$ or $\mathbb{Z}_+ = \{0, 1, ...\}$ or $\mathbb{Z} = \{..., -1, 0, 1, ...\}$), the reals ($\mathbb{R}_+ = [0, \infty)$ or $\mathbb{R} = (-\infty, \infty)$), or intervals (often $[0, 1]$), or Cartesian powers or products of any of these.

Recall that a *probability distribution* on $\mathcal{X}$ is an assignment to subsets $E \subset \mathcal{X}$ of numbers $\mathsf{P}(E) \in [0, 1]$ such that the probability of the union of disjoint subsets is the sum of the individual probabilities. Typically (always, in this class) these arise from *probability mass functions* (pmfs) $p(\cdot)$ for discrete $\mathcal{X}$ and *probability density functions* (pdfs) $f(\cdot)$ for continuous $\mathcal{X}$ via the relations

$$\mathsf{P}[E] = \sum \{p(x) :\ x \in E\} \qquad \text{or} \qquad \mathsf{P}[E] = \int_E f(x)\,dx$$

The CDF of X and the expectation of any function $g(X)$ can be written

$$F(x) = \sum \{p(t) :\ t \le x\} \qquad F(x) = \int_{-\infty}^{x} f(x)\,dx$$

$$\mathsf{E}[g(X)] = \sum \{g(x)\,p(x) :\ x \in \mathbb{R}\} \qquad \mathsf{E}[g(X)] = \int_{\mathbb{R}} g(x)\,f(x)\,dx$$

$$= \int_{\mathbb{R}} g(x)\,dF(x) \qquad = \int_{\mathbb{R}} g(x)\,dF(x)$$

The Riemann-Stieltjes notation for these expectations (shown in the last line) are identical for discrete, continuous, or indeed any other distribution (such as mixtures of discrete and continuous, which arise with censored observations). One way to define them for continuous $g()$ is

$$\int_{\mathbb{R}} g(x)\,dF(x) := \lim_{n\to\infty} \sum_{i=-n^2}^{n^2} g(i/n)\left[F\big((i+1)/n\big) - F\big(i/n\big)\right]$$

provided that the limit exists. The measure-theoretic notation $\int_{\mathbb{R}} g(x)\,F(dx)$ is also used for this same expression, simplifying and clarifying expressions when F might depend on another variable (perhaps y) as well and extending to non-continuous $g()$. To avoid writing everything twice, once for continuous and once for discrete distributions, we will typically use this Stieltjes notation whenever it is unspecified whether a distribution with CDF F might be continuous or discrete.

Statistical models come in two varieties:

- **Parametric**, in which the CDFs $F(x \mid \theta)$ (or, equivalently, the pmfs $p(x \mid \theta)$ or pdfs $f(x \mid \theta)$) are indexed by a low-dimensional set $\theta \in \Theta \subset \mathbb{R}^d$ for some modest $d \in \mathbb{N}$; or
- **Nonparametric**, in which they aren't. In this case tools from functional analysis, like Sobolev and Besov spaces, are commonly used.

We'll spend about a week looking at some nonparametric ideas, but most of this course will focus on familiar parametric families, like the Normal, Gamma, Exponential, Geometric, Negative Binomial, Weibull, *etc.*, often with $\Theta \subset \mathbb{R}$ or $\Theta \subset \mathbb{R}^2$ so the distribution is determined by just one or two numbers. A list of common parametric distributions along with their notation, pdf or pmf, mean, variance, *etc.* is available at course PDF sheet (also available from the course syllabus page). Often the components X_i of the observation vector X will be modeled as independent and identically distributed from one of these distributions.

Typically we will assume that X is a realization of a random vector from a distribution with pdf $f(x \mid \theta)$ or pmf $p(x \mid \theta)$ for some particular parameter value $\theta \in \Theta$ that is unknown to us.

Point Estimates

One possible objective of inference would be to determine *which* value $\theta \in \Theta$ gave rise to $X \in \mathcal{X}$. An *estimator* is some function $T : \mathcal{X} \to \Theta$ intended to have a value $T(X)$ that is equal to, or at least close to, θ. Often we consider sequences $\{X_i\}$ of iid random variables all with the same distribution, *i.e.*, the same pdf $f(x \mid \theta)$, and corresponding sequences $T_n : \mathcal{X}^n \to \Theta$ of estimators in the hope that with more and more data we will have a better and better estimate in the sense that T_n gets closer and closer to θ as n increases. For example, if $X_i \overset{\text{iid}}{\sim} \mathsf{No}(\theta, 1)$ are all independent and normally distributed with known variance $\sigma^2 > 0$ and uncertain mean $\theta \in \mathbb{R}$, we know from the Law of Large Numbers that the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ converges to θ as $n \to \infty$. So does the sample median $\mathring{X}_n$ and many other estimators we will compare later.

A "statistic" is any function of the data, *i.e.*, any function on the space $\mathcal{X}$; thus an "estimator" is just a Θ-valued statistic. There are a variety of ways to quantify how close the random variable $T(X)$ is to θ, typically:

- The **bias** of an estimator $T(X)$ is its expected error,

$$\mathsf{bias}(\theta) := \mathsf{E}_\theta\big[\ T(X) - \theta\ \big],$$

  or (by the LLN) the long-term average discrepancy. If $\mathsf{bias}(\theta)$ is identically zero, T is called "unbiased". In this case the errors could still be quite large, but the positive and negative errors cancel out on average.

- The **mse** or Mean Squared Error of an estimator is the long-term average of its *squared* error,

$$\mathsf{mse}(\theta) := \mathsf{E}_\theta\big[|T(X) - \theta|^2\big].$$

- The **se** or Standard Error of an estimator is the square root of its variance,

$$\mathsf{se}(\theta) := \mathsf{V}_\theta\big[T(X)\big]^{1/2}.$$

- By elementary probability,

$$\mathsf{mse}(\theta) = \mathsf{bias}(\theta)^2 + \mathsf{se}(\theta)^2$$

- Corollary: A family $\{T_n\}$ of estimators is "mean square consistent", *i.e.*, has mse_n converging to zero, if and only if (1) It is asymptotically unbiased, $\mathsf{bias}_n(\theta) \to 0$; and (2) Its standard error (or, equivalently, its variance) shrinks to zero, $\mathsf{se}_n(\theta) \to 0$.

- Example: if $X_j \overset{\text{iid}}{\sim} \mathsf{Po}(\lambda)$ and $T_n(X) := \overline{X}_n = n^{-1}\sum_i X_i$, then $\mathsf{E}[T_n(X)] = \lambda$ so T_n is unbiased. The variance $\mathsf{V}(T_n) = \lambda/n$ converges to zero, so $\{T_n\}$ is MS consistent.

- Example: if $X_j \overset{\text{iid}}{\sim} \mathsf{Po}(\lambda)$ and $S_n(X) := (\beta+n)^{-1}\big[\alpha + \sum_i X_i\big]$ for any fixed $\alpha, \beta \ge 0$, then $\mathsf{E}[S_n(X)] = \frac{\alpha+n\lambda}{\beta+n} = (\lambda + \alpha/n)/(1+\beta/n) \to \lambda$ so S_n is asymptotically unbiased. The bias $\mathsf{bias}_n(\lambda) = \frac{\alpha-\beta\lambda}{\beta+n}$ and variance $\mathsf{V}(S_n) = \frac{n\lambda}{(n+\beta)^2} \le \lambda/n$ converges to zero, so $\{S_n\}$ is MS consistent.

- Example: if $X_j \overset{\text{iid}}{\sim} \mathsf{Ex}(\lambda)$ and $U_n(X) := n\big/\sum_i X_i$ (the "MLE"— more on these later), then $\mathsf{E}[U_n(X)] = n\lambda/(n-1)$ so U_n is asymptotically unbiased. The bias $\mathsf{bias}_n(\lambda) = \lambda/(n-1)$ and variance $\mathsf{V}(U_n) = \lambda^2 n^2/(n-1)^2(n-2)$ converge to zero, so $\{U_n\}$ is MS consistent.

Interval Estimates

One way to quantify how close an estimator is in a one-dimensional problem (where $\Theta \subseteq \mathbb{R}$) is to give a plus-or-minus range with a probability bound— or, equivalently, to give *two* statistics, a lower one $L(X)$ and an upper one $U(X)$, with the intention that typically $L(X) < \theta < U(X)$. Commonly a probability bound γ is specified (95% is common) and bounds L, U are found for which, in some sense,

$$\mathsf{P}[L(X) \le \theta \le U(X)] \ge \gamma.$$

We have already seen that Frequentist and Bayesian approaches to inference will both want an inequality of this sort, but the probabilities will have different interpretations— in the Frequentist approach the quantity θ is fixed and the statement is about the random values of $L(X)$ and $R(X)$,

while in the Bayesian approach the observed value of X is fixed and the statement is about the random value of θ. In either case one seeks interval estimates $[L(X), U(X)]$ with the two conflicting goals that (1) $\mathsf{P}\{\theta \in [L(X), U(X)]\}$ is large and (2) that the interval $[L(X), U(X)]$ is short.

In problems with vector values of $\theta \in \Theta$ one can still offer interval estimates for individual components θ_j, or one can construct *set*-valued statistics $C : \mathcal{X} \to 2^\Theta$ with the property that $\mathsf{P}[\theta \in C(X)] \geq \gamma$. Again one seeks sets with $\mathsf{P}[\theta \in C(X)]$ large yet $C(X)$ small.

Frequently a family of estimators $T_n(X)$ is both asymptotically unbiased and *asymptotically normal* with variance proportional to $1/n$, *i.e.*, $\sqrt{n}[T_n(X) - \theta]$ has approximately a $\mathsf{No}(0, \sigma^2)$ normal distribution in one dimension, or multivariate normal $\mathsf{No}(0, \Sigma)$ in $d > 1$ dimensions, for some constant $\sigma > 0$ or positive-definite matrix Σ. In that case the confidence interval

$$L(X) = T_n(X) - \frac{\sigma z_\gamma}{\sqrt{n}}, \qquad U(X) = T_n(X) + \frac{\sigma z_\gamma}{\sqrt{n}}$$

will be an approximate $100\gamma\%$ "confidence interval" if $\Phi(z_\gamma) = (1+\gamma)/2$, in one dimension, or

$$C(X) = \{\theta : \ (\theta - T_n(X))' \Sigma^{-1} (\theta - T_n(X)) \leq \zeta/n\}$$

will be a $100\gamma\%$ confidence set (an ellipsoid) with $\Theta \subset \mathbb{R}^d$, if ζ is the γ'th quantile of the χ^2_d distribution with d degrees of freedom, *i.e.*, the $\Gamma(d/2, 1/2)$ distribution. We'll see more about that in a few weeks.

Example

Let $X \sim \mathsf{Ex}(\lambda)$ be a single observation from the exponential distribution with rate $\lambda > 0$. How can we find a one-sided confidence interval satisfying

$$(\forall \lambda > 0)\ \mathsf{P}\left\{\lambda \in [L(X), \infty)\right\} \geq \gamma$$

for fixed $0 < \gamma < 1$? The mean and median of the $\mathsf{Ex}(\lambda)$ distribution are $1/\lambda$ and $\log(2)/\lambda$, respectively, so we should anticipate that $L(X)$ may be a multiple of $1/X$— if X is a distance measured in centimeters cm for example then λ has units cm^{-1}. A switch to kilometers km would re-scale X by a factor of 10^{-5} and hence re-scale λ by 10^5, so $L(X)$ should also be rescaled by 10^5. Sometimes dimension arguments can help us guess the form functions must have, and help us discover our own errors when the dimensions don't work out.

We're looking for a statistic $L(X)$ that satisfies $\mathsf{P}[L(X) \leq \lambda] = \gamma$. One approach would begin by noting that, for any $x > 0$,

$$\mathsf{P}[X > x] = \exp(-\lambda x)$$

For this probability to be $\gamma > 0$, take $x = -\log(\gamma)/\lambda$; then

$$\mathsf{P}\left\{X > \frac{-\log\gamma}{\lambda}\right\} = \exp\left\{-\lambda \frac{-\log\gamma}{\lambda}\right\} = \gamma, \qquad \textit{i.e.},$$
$$\mathsf{P}[\lambda > -\log(\gamma)/X] = \gamma.$$

SO, the interval $[L(X), \infty)$ contains λ with probability *exactly* γ for the statistic $L(X) := -\log(\gamma)/X$. For $\gamma = 0.90$, for example, the one-sided interval is about $\left[\frac{0.105}{X}, \infty\right)$.

A similar two-sided interval can be constructed:

$$\mathsf{P}\left\{\frac{-\log(\frac{1+\gamma}{2})}{X} \le \lambda \le \frac{-\log(\frac{1-\gamma}{2})}{X}\right\} = \gamma$$

For $\gamma = 0.90$, for example, the two-sided interval is about $\left[\frac{0.05}{X}, \frac{3.0}{X}\right]$.

Hypothesis Tests

Another traditional approach to inference is to consider whether or not some assertion about θ is true. This is equivalent to the question of whether or not θ lies in some subset $H_0 \subset \Theta$, namely, the set H_0 of all those $\theta \in \Theta$ for which the assertion *is* true. Such a subset is called an "hypothesis", or "Null Hypothesis", and its complement $H_1 = \{\theta \in \Theta : \ \theta \notin H_0\}$ is called the "alternate".

The Frequentist and Bayesian approaches will have different ways of quantifying how plausible H_0 is after observing $X \in \mathcal{X}$. The Bayesian approach is simply to report the probability $\mathsf{P}[\theta \in H_0 \mid X]$. In the Frequentist approach, where θ isn't treated as random, this probability can take only two values— one, if in fact $\theta \in H_0$, and zero if $\theta \notin H_0$, but the investigator can't tell which of these is true. Instead, the Frequentist approach is to identify a region $\mathcal{R} \subset \mathcal{X}$ of possible outcome values that are *not* particularly likely for $\theta \in H_0$, called the "rejection region", and to report whether or not X lies in this region. The "size" of such a test,

$$\alpha = \sup\left\{\mathsf{P}_\theta[X \in \mathcal{R}] : \ \theta \in H_0\right\},$$

is a measure of the strength of evidence against H_0 represented by an observation $X \in \mathcal{R}$: if α is small, then an observation $X \in \mathcal{R}$ is a near-miracle if $\theta \in H_0$, while it may be rather expected if $\theta \notin H_0$. Since miracles are rare, it seems reasonable to "reject H_0" if we observe $X \in \mathcal{R}$. We will see this in much more detail soon, and see some variations (like "P-values").

Regression

Sometimes we observe *pairs* of quantities (Y_i, X_i) from some product space $\mathcal{Y} \times \mathcal{X}$ and hope to discover how they are related. Commonly the "explanatory variables" $\{X_i\}$ are treated as known with certainty (often they are specified by the investigator, and are not random at all) while some mystery surrounds the distributions of the "response variables" $\{Y_i\}$, which may depend on the corresponding $\{X_i\}$. The simplest case is to imagine that $Y_i \approx g(X_i)$ for some "regression function" $g(\cdot) : \mathcal{X} \to \mathcal{Y}$ that is either entirely unknown or, more often, is thought to be from some low-dimensional family of functions $\mathcal{G} = \{g_\beta(\cdot) : \ \beta \in B\}$, like linear functions $g_\beta(X) = X'\beta$. Commonly the approximation errors $e_i := \big(Y_i - g(X_i)\big)$ are taken to be iid with mean zero from some small parametric family, usually $\mathsf{No}(0, \sigma^2)$. More generally one may model

$$Y_i \overset{\text{ind}}{\sim} f\big(y \mid g_\beta(X_i)\big)$$

for some family of pdfs or pmfs $f(y \mid \theta)$ and some family of regression functions $g_\beta : \mathcal{X} \to \Theta$ indexed by an uncertain "regression vector" $\beta \in B$. For example, the Y_i might be integer counts with Poisson distributions whose means $g_\beta(X_i) = \exp(X_i'\beta)$ have a log-linear dependence on the explanatory variables. In the most common example the $Y_i \overset{\text{ind}}{\sim} \mathsf{No}(X_i'\beta, \sigma^2)$ are normal with mean $g_\beta(X_i) = \exp(X_i'\beta)$ and constant unknown variance σ^2.

Now interest may center on point or interval estimates for the parameter vector β, or on prediction of future observations $\{Y_j^* : \ j \in J\}$ for specified vectors at new "design" locations $\{X_j^* : \ j \in J\}$, or on the *selection* of those design points (the "design of experiments") in an effort to learn as much as possible about β from as few new observations $\{(Y_j^*, X_j^*) : \ j \in J\}$ as possible.

Last edited: October 20, 2017

References

Bayes, T. (1763), "An essay toward solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society*, pp. 370–418.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222, 309–368.

Fisher, R. A. (1925), "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.

Fisher, R. A. (1935), "Statistical Tests," *Nature*, 136, 474.

Laplace, P. S. (1774), *A philocophical essay on probabilities*, New York, NY: Dover, translated from 6th French edn (1774) by Frederick Wilson Truscott and Frederick Lincoln Emory; Dover edition, 1952.

Neyman, J. and Pearson, E. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society. Series A*, 231, 289–337.