Statistics 210B, Spring 1998

Class Notes

P.B. Stark

stark@stat.berkeley.edu

www.stat.berkeley.edu/~stark/index.html

January 22, 1998

First Set of Notes

1 Hypothesis Testing and Confidence Sets

1.1 Set-up

We are to collect a vector of data $X \in \mathcal{X}$, which has probability distribution $\mathbf{P}_\theta$, with (possibly infinite-dimensional) parameter θ unknown, except that $\theta \in \Theta$, where Θ is a known set. Typically, $\mathcal{X} = \mathbf{R}^n$, but it might instead be a more general measurable space of possible observations. We are interested in making statistical inferences about $\tau(\theta)$, which might be θ itself, or a function of θ (for example, for a univariate normal we might have $\theta = (\mu, \sigma^2)$, and be interested in $\tau(\theta) = \mu$). Let

$$\mathbf{T} \equiv \tau(\Theta) = \{\gamma : \exists \eta \in \Theta \text{s.t.} \gamma = \tau(\eta)\}, \tag{1}$$

and

$$\mathbf{P}_\Theta = \{\mathbf{P}_\eta : \eta \in \Theta\}. \tag{2}$$

We wish to test the *null hypothesis* $H : \tau(\theta) \in \mathbf{T}_H \subset \mathbf{T}$ against an alternative K not yet specified. In a deliberate "overloading" of notation, let H also stand for $\{\mathbf{P}_\eta, \eta \in \Theta : \tau(\eta) \in$

$\mathbf{T}_H\}$ (the set of probability distributions for which the null hypothesis H is true), and let K also stand for $\{\mathbf{P}_\eta, \eta \in \Theta : \tau(\eta) \in \mathbf{T}_K\}$ (the set of probability distributions for which the alternative hypothesis K is true). We shall typically assume that $H \cup K = \mathbf{P}_\Theta$.

Definition 1 *If $\{\mathbf{P}_\eta \in H\}$ be a singleton set (just one distribution), we say the null hypothesis H is* simple. *If the alternative K be a singleton set, we say K is* simple. *If an hypothesis is not simple, it is* composite.

Definition 2 *A (significance) level α test of the hypothesis $\tau(\theta) \in \mathbf{T}_H$ is a (possibly random) measurable decision rule $\delta(X) : \mathcal{X} \to \{$ accept, reject$\}$ such that*

$$\sup_{\{\mathbf{P}_\eta \in H\}} \mathbf{P}_\eta\{\delta(X) = \textit{reject}\} \leq \alpha. \tag{3}$$

The constant α is (an upper bound on) the probability of a false rejection.

The most common decision rules (deterministic rules) reject when the data X fall outside a set $A = A_H$ that satisfies

$$\sup_{\{\mathbf{P}_\eta \in H\}} \mathbf{P}_\eta\{X \notin A_H\} \leq \alpha, \tag{4}$$

The set A_H is called the *acceptance region* of the test; A_H^C is the *rejection region* of the test. Under the Neyman-Pearson paradigm, the term "acceptance region" is a misnomer—one never "accepts" the null hypothesis; one merely fails to reject it given certain data (evidence) X. I shall often blur the notational distinction between a test and its acceptance region.

Another family of decision rules performs a random experiment that depends on the observed value of X, such that for each x, the null hypothesis is rejected with probability $\phi(x)$ and not rejected with probability $1 - \phi(x)$. To have a significance level α randomized test, we need

$$\sup_{\{\mathbf{P}_\eta \in H\}} E_\eta \phi(X) = \int \phi(x) d\mathbf{P}_\eta(x) \leq \alpha. \tag{5}$$

Deterministic rules correspond to decision functions ϕ that take only the values 0 (do not reject, with probability 1) and 1 (reject, with probability 1).

Typically, the set A_H is defined in two steps: first, one selects a statistic $T(X)$ (a function of X that is $\mathbf{P}_\gamma$-measurable for all $\gamma \in \Theta$, and that does not depend on θ), then one defines a subset A_{T_H} of the range of T, with the property that

$$\sup_{\{\mathbf{P}_\eta \in H\}} \mathbf{P}_\eta\{T(X) \notin A_{T_H}\} = \alpha. \tag{6}$$

Thus A_H, a subset of $\mathcal{X}$, is the pre-image under T of A_{T_H}, a subset of the range of T. (In symbols, $A_H = T^{-1}(A_{T_H})$.)

Suppose that the range $\mathcal{X}$ of X is endowed with a distance

$$\begin{aligned} d(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbf{R}^+ \\ (x, y) &\mapsto d(x, y), \end{aligned} \tag{7}$$

where $\mathbf{R}^+$ are the nonnegative reals. (Recall that a distance $d(\cdot,\cdot)$ on a set $\mathcal{X}$ must satisfy

1. $0 \le d(x, y) \le \infty$; $d(x, y) = 0 \iff x = y$ (positive definiteness)

2. $d(x, y) = d(y, x)$ (symmetry)

3. $d(x, z) \le d(x, y) + d(y, z)$ (triangle inequality)

for all x, y, z in $\mathcal{X}$.)

Definition 3 *The* diameter *of a set A on which a metric d is defined is*

$$|A| \equiv \sup_{x,y \in A} d(x, y). \tag{8}$$

The radius of A relative to the point x *is*

$$|A|_\theta \equiv \sup_{y \in A} d(x, y). \tag{9}$$

One natural criterion of optimality of an acceptance region is that its diameter be minimal. This is related to (but not equivalent to) the power of the test against a family of alternatives; *vide infra.*

Definition 4 *A* family of tests for $\tau \in \mathbf{T}$ *is a set-valued function A_γ such that for each $\gamma \in \mathbf{T}$, A_γ is the acceptance region for a level α test of the hypothesis $H : \tau = \gamma$.*

Examples.

1. Suppose that $\mathbf{P}_\theta$ is the normal distribution with mean θ and unit variance, that $\Theta = \mathbf{R}$, $\tau(\theta) = \theta$, and that we observe $X \sim \mathbf{P}_\theta$. Let z_λ be the λ critical value of the standard normal distribution; that is,

$$\mathbf{P}_0\{X \geq z_\lambda\} = \lambda. \tag{10}$$

Then

$$A_\gamma \equiv (\gamma - z_{\alpha/2}, \gamma + z_{\alpha/2}) \tag{11}$$

is a family of level α tests for $\tau(\theta) = \theta \in \mathbf{R}$.

2. Suppose $\mathbf{P}_\Theta$ is the family of distributions on $\mathbf{R}$ that are continuous with respect to Lebesgue measure. Let $\tau(\theta)$ be the 90th percentile of the distribution parametrized by θ. We observe $X = \{X_j\}_{j=1}^n$ i.i.d. P_θ. Let $T_\gamma : \mathbf{R}^n \rightarrow \mathbf{N}$ equal $\#\{X_j \geq \gamma\}$. ($\mathbf{N}$ are the nonnegative integers). For all ν such that $\tau(\nu) = \gamma$, the probability distribution of $T_\gamma(X)$ is binomial with parameters n and $p = 0.1$. Thus for any γ, we can find integers $a_- = a_-(\gamma, n, \alpha)$ and $a_+ = a_+(\gamma, n, \alpha)$ such that

$$\mathbf{P}_\nu\{T_\gamma(X) \notin [a_-, a_+]\} \leq \alpha \;\; \forall \nu \text{s.t.} \tau(\nu) = \gamma. \tag{12}$$

Such a pair of mappings defines a family of level α tests for $\tau(\theta) \in \mathbf{R}$.

3. Suppose that $\mathbf{P}_\Theta$ is the set of probability distributions on $\mathbf{R}$ that are continuous with respect to Lebesgue measure; let θ be the distribution function of the "true" measure, and suppose we are interested in $\tau(\theta) = \theta$. We observe $X = \{X_j\}_{j=1}^n$ i.i.d. $\mathbf{P}_\theta$. Let $\hat{\theta}_n$ denote the empirical distribution

$$\hat{\theta}_n\{(-\infty, x]\} \equiv \frac{1}{n} \sum_{j=1}^{n} 1_{x \geq X_j}, \tag{13}$$

where 1_B is the indicator function of the event B. For any two probability distributions $\mathbf{P}_1$, $\mathbf{P}_2$, on $\mathbf{R}$, define the Kolmogorov-Smirnov distance

$$d_{KS}(\mathbf{P}_1, \mathbf{P}_2) \equiv \|\mathbf{P}_1 - \mathbf{P}_2\|_{KS} \equiv \sup_{x \in \mathbf{R}} |\mathbf{P}_1\{(-\infty, x]\} - \mathbf{P}_2\{(-\infty, x]\}|. \tag{14}$$

There exist universal constants χ_α so that for every continuous (w.r.t. Lebesgue measure) distribution θ,

$$\mathbf{P}_\theta\left\{\|\theta - \hat{\theta}_n\|_{KS} \geq \chi_n(\alpha)\right\} = \alpha. \tag{15}$$

This is the Dvoretzky-Kiefer-Wolfowitz indquality. Moreover, Massart (*Ann. Prob., 18*, 1269–1283, 1990) showed that the constant

$$\chi_n(\alpha) \leq \sqrt{\frac{\ln \frac{2}{\alpha}}{2n}} \tag{16}$$

is *tight*. For $y = (y_1, \cdots, y_n) \in \mathbf{R}^n$, let $\hat{y}_n$ be the probability measure on $\mathbf{R}$ whose distribution function is $1/n \sum_{j=1}^{n} 1_{x \geq y_j}$. Then

$$A_\gamma \equiv \{y \in \mathbf{R}^n : \|\gamma - \hat{y}_n\|_{KS} \leq \chi_\alpha\} \tag{17}$$

is a family of level α tests for $\theta \in \Theta$.

1.2 Most Powerful Tests

Definition 5 *The* power β of the test δ of H against the alternative K *is*

$$\beta = \beta(\delta, K) \equiv \inf_{\mathbf{P}_\nu \in K} \mathbf{P}_\nu\{\delta(X) = \mathit{reject}\}. \tag{18}$$

That is, $\beta(\delta, K)$ is the smallest probability of rejecting the null hypothesis when the value of the parameter of interest, $\tau(\theta)$, is in the alternative set $\mathbf{T}_K$.

In the Neyman-Pearson paradigm for hypothesis testing, one is concerned with the probabilities of two kinds of errors: rejecting the null hypothesis H when it is in fact true (a Type I error), and failing to reject the null hypothesis when it is in fact false (a Type II error). The significance level of a test is a bound on the probability of a Type I error; the power of the test against the alternative K is $1 - \sup_{\mathbf{P}_\nu \in K} \mathbf{P}_\nu\{\text{Type II error}\}$.

For a given bound α on the chance of a Type I error, one is naturally led to maximize the power $\beta(K)$. This can be thought of as a more general statistical decision problem with two zero-one loss functions: Define

$$L_1(\theta, \text{reject}) = \begin{cases} 0, & \mathbf{P}_\theta \notin H \\ 1, & \mathbf{P}_\theta \in H \end{cases} \tag{19}$$

$$L_1(\theta, \text{accept}) = 0, \forall \theta \in \Theta, \tag{20}$$

and

$$L_2(\theta, \text{reject}) = 0, \forall \theta \in \Theta, \tag{21}$$

$$L_2(\theta, \text{accept}) = \begin{cases} 0, & \mathbf{P}_\theta \in H \\ 1, & \mathbf{P}_\theta \notin H \end{cases} \tag{22}$$

Then the problem of finding the most powerful test is to find the decision rule δ that minimizes $EL_2(\theta, \delta(X))$ subject to the constraint $EL_1(\theta, \delta(X)) \leq \alpha$.

For the case H and K are simple, let $\mathbf{P}_H = H$ and $\mathbf{P}_K = K$. Considering first nonrandomized tests, one wants to find A_H to maximize

$$\beta = \int_{x \notin A_H} d\mathbf{P}_K(x) \tag{23}$$

subject to

$$\int_{x \notin A_H} d\mathbf{P}_H(x) \leq \alpha. \tag{24}$$

Subject to a bound on the chance of a Type I error, the best points to exclude from A_H are those that are most probable under K relative to their probability under H. Let $r(x) = d\mathbf{P}_K(x)/d\mathbf{P}_H(x)$. Then the most powerful nonrandomized level α test δ has

$$A_H = \{x : r(x) > c\}, \tag{25}$$

where c solves

$$\mathbf{P}_H\{X \notin A_H\} = \int_{x: r(x) > c} d\mathbf{P}_H(x) = \alpha. \tag{26}$$

If $\mathbf{P}_H$ contains atoms, it can happen that for some values of α, the most powerful deterministic decision rule δ that attains exactly level α is not given by the likelihood ratio region 25 for some special values of α (for a given value of c, the level would be too large, while for infinitesmaly larger c, the level would be too small). If one allows randomized decisions, that problem does not occur; one makes a deterministic decision when $r < c$ or $r > c$, and makes a random decision for $r = c$, with probability of rejection chosen s.t. the overall level is α. A more common approach (essentially ubiquitous in practice) is to choose α to avoid such pathology.

Theorem 1 *Fundamental Lemma of Neyman and Pearson (See Lehmann, TSH, 3.2, Theorem 1.) Suppose* $\mathbf{P}_H$ *and* $\mathbf{P}_K$ *have densities* p_H *and* p_K *relative to a measure* μ (e.g., $\mathbf{P}_H + \mathbf{P}_K$). *Then*

1. *There is a decision function* ϕ *and a constant* c *such that*

$$E_H\phi(X) = \alpha, \tag{27}$$

$$\phi(x) = \begin{cases} 1, & p_K(x) > cp_H(x) \\ 0, & p_K(x) < cp_H(x). \end{cases} \tag{28}$$

(The value of ϕ *for* $p_K(x) = cp_H(x)$ *is adjusted to give* $E_H\phi(X) = \alpha$*; depending on* α*,* H*, and* K*, this can result in a randomized decision rule.)*

2. *If a decision function* ϕ *satisfies 27 and 28 for some* c*, it is most powerful for testing* H *against* K *at level* α*.*

3. *If* ϕ *is the most powerful decision function for testing* H *against* K*, then for some* c *it satisfies 28 a.e.(*μ*), and it satisfies 27 unless there is a level* $< \alpha$ *test of* H *against* K *with* $\beta = 1$*.*

The fundamental lemma of Neyman and Pearson applies just to simple null and alternative hypotheses. One might hope that when H and K were composite, the same test would be most powerful for all $\mathbf{P}_\eta \in H$ against all $\mathbf{P}_\eta \in K$; unfortunately, that is not typically the case. Such a test, when it exists is called *uniformly most powerful* (UMP).

There is an important class of distributions with real parameters for which UMP tests exist. Suppose $\mathbf{P}_\eta$, $\eta \in \Theta = \mathbf{R}$ has density $p_\eta(x)$.

Definition 6 *The set of densities* p_η *has* monotone likelihood ratio (in $T(x)$) *if there exists a function* $T : \mathcal{X} \to \mathbf{R}$ *such that for* $\nu < \eta$

1. $\mathbf{P}_\nu \neq \mathbf{P}_\eta$*, and*

2. $p_\eta(x)/p_\nu(x)$ *is a monotone non-decreasing function of* $T(x)$*.*

Theorem 2 *(See Lehmann, TSH, 3.3, Theorem 2.) Suppose* $\theta \in \Theta = \mathbf{R}$ *and* X *has density* $p_\theta(x)$ *with monotone likelihood ratio in* $T(x)$*. Let* $H = \{\mathbf{P}_\eta : \eta \leq \eta_H\}$ *and* $K = \{\mathbf{P}_\eta : \eta > \eta_H\}$*. (Such a* K *is called a* one-sided alternative*.) Then*

1. *A UMP level* α *test of* H *against* K *exists.*

2. The decision function ϕ for the UMP test is

$$\phi(x) = \begin{cases} 1, & T(x) > c \\ b & T(x) = c \\ 0, & T(x) < c, \end{cases} \tag{29}$$

with b and c chosen to satisfy

$$E_{\mathbf{P}_{\eta_H}} \phi(X) = \alpha. \tag{30}$$

3. For this test, the power

$$\beta(\mathbf{P}_\theta) = E_{\mathbf{P}_\theta} \phi(X) \tag{31}$$

is a strictly increasing function of θ at all points for which $0 < \beta(\theta) < 1$.

4. For all γ, this test is UMP for testing $\theta \leq \gamma$ against $\theta > \gamma$ at level $\beta(\gamma)$.

5. For any $\theta < \eta_H$, the test minimizes $\beta(\theta)$ among all level α tests.

Definition 7 *Let $\mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbf{R}$ have density*

$$p_\theta(x) = C(\theta) e^{Q(\theta)T(x)} h(x) \tag{32}$$

relative to some measure μ, with $Q(\cdot)$ strictly monotone. Then $\{\mathbf{P}_\theta : \theta \in \Theta\}$ is a one parameter exponential family*.*

Remark. The one-parameter exponential families have monotone likelihood ratio in $T(x)$.

Remark. Lehmann refers to a converse due to Pfanzagl (1968) that under weak regularity conditions, if there exist level α UMP tests against one-sided alternatives for all sample sizes, $\mathbf{P}_\Theta$ is an exponential family.

1.3 Confidence Regions.

Definition 8 *A $1 - \alpha$ confidence region for $\tau(\theta)$ is a random set $S(X) \subset \mathbf{T}$ satisfying*

$$\mathbf{P}_\theta\{S(X) \ni \tau(\theta)\} \geq 1 - \alpha. \tag{33}$$

The most common way to construct a $1 - \alpha$ confidence region for $\tau(\theta)$ is by "inverting" a family of tests for the hypotheses $\tau(\theta) = \gamma$:

Theorem 3 *Duality between Tests and Confidence Regions. (See Lehmann, TSH, 3.5, Theorem 4). Let A_γ be a family of acceptance regions for level α tests of the hypotheses $\tau(\theta) = \gamma$. For each value of $x \in \mathbf{R}^n$, define*

$$S(x) = \{\gamma \in \mathbf{T} : x \in A_\gamma\}. \tag{34}$$

Then $S(X)$ is a confidence region for $\tau(\theta)$ with confidence level $1 - \alpha$.

Theorem 4 *The Ghosh-Pratt Identity. (See Pratt, J.W., 1961. Length of confidence intervals,* JASA, 56, *549–567; Ghosh, J.K., 1961. On the relation among shortest confidence intervals of different types,* Calcutta Stat. Assoc. Bull., *147–152.) For a set $S(x) \subset \Theta$, let*

$$\mu(S(x)) \equiv \int_{\gamma \in S(x)} d\mu(\gamma), \tag{35}$$

for some measure μ on Θ. Then

$$E_{\mathbf{P}_\eta} \mu(S(X)) = \int \mathbf{P}_\eta\{S(X) \ni \gamma\} d\mu(\gamma). \tag{36}$$

The Ghosh-Pratt identity relates the expected "volume" (w.r.t. the measure μ) of a confidence set to the probability that points other than the true parameter are in the set: the right hand side is the integral of the "false coverage" probability. That is in turn related to the power of the tests to which S is dual against the alternative with respect to which the expectation and the probability are calculated. For example, suppose that $\Theta = \mathbf{R}^m$, that μ is Lebesgue measure (so the expectation on the left is the "ordinary" expected volume of the confidence set) and that S is the dual of a family of tests that are most powerful against the alternative $\theta = \mathbf{0}$. That is, the sets A_ν minimize $\mathbf{P}_\nu\{\mathbf{0} \ni A_\nu\}$. Then the confidence set $S(X)$ has minimal expected volume when the true value of θ is $\mathbf{0}$ among all confidence sets.

Brown, Casella and Huang (Optimal Confidence Sets, Bioequivalence, and the Limacon of Pascal, Brown Univ. Tech. Rept. BU-1205-M, 1993, rev.1994) use this result to develop confidence sets for assessing bioequivalence. In the case $X \sim N(\theta, \mathrm{I})$, the acceptance regions of tests with optimal power against $\mathbf{0}$ can be derived from the likelihood ratio; Brown and Huang obtain closed-form expressions for the shape of the resulting confidence sets.

Problem. Find a formula for a $1 - \alpha$ confidence set for the mean of a Poisson distribution from n i.i.d. observations, with minimal expected volume when the true mean $\theta = 1$. Is the set always an interval? Give the confidence set that results when $X = 2$. It might help to read Brown and Huang.