ORF 524: Statistical Theory and Methods **Fall 2016**

Lecture 10: 10/17/2016

Lecturer: Prof. Samory Kpotufe *Scribe: Kiran Vodrahalli*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

10.1 Confidence Sets with Gaussian Assumptions, Continued

Example 1. Regression (confidence bands).
Let $y = Xw + \eta$, $X \in \mathbb{R}^{n \times d}$ fixed, $w \in \mathbb{R}^d$. Let $\eta \sim \mathcal{N}(0, I_n)$, $\widehat{w} = (X^T X)^{-1} X^T y$ is well-defined. Let $f(x) = x^T w$, $\widehat{f}(x) = x^T \widehat{w}$, and we have

$$\begin{aligned} x^T(\widehat{w} - w) &= x^T\left((X^T X)^{-1} X^T (y - Xw)\right) \\ &= A\eta \sim \mathcal{N}(0, AA^T) \end{aligned} \tag{10.1}$$

where $A = x^T(X^T X)^{-1} X^T \implies AA^T = x^T (X^T X)^{-1} x = \sigma_x^2$.

Thus we have

$$\Pr\left(-z_{\alpha/2} \le \frac{x^T \widehat{w} - f(x)}{\sigma_x} \le z_{\alpha/2}\right) = 1 - \alpha \tag{10.2}$$

which we can rephrase as $f(x) \in S(y) = x^T \widehat{w} \pm z_{\alpha/2}\sigma_x$ with probability $1 - \alpha$.

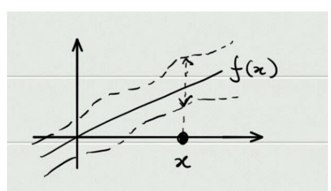

Figure 10.1: Confidence Band for Linear Regression

Exercise: Let Φ denote the c.d.f. of $\mathcal{N}(0, 1)$ and let $z_\alpha \in \mathbb{R}$ denote the critical value satisfying $\Phi(z_\alpha) = 1 - \alpha$. Consider the above fixed design.

1. Derive an ellipsoidal $(1 - \alpha)$-confidence set for $w \in \mathbb{R}^d$ using z_α values. (Remember $\{x : \|Ax\| \le r\}$ for $A \succeq 0$ is an ellipsoid.)

2. Derive a hypercubic $(1 - \alpha)$-confidernce set for $w \in \mathbb{R}^d$ of the form $\{w : \|w - c\|_\infty < r\}$.

3. Suppose now that the design matrix X is also random. Are the confidence sets in the first two parts still of level $1 - \alpha$?

Exercise: (Note that this draws upon the notation from the example above). Suppose w is k-sparse; i.e. has exactly k non-zero coordinates $i_1, \cdots, i_k$ with $X^TX = nI_d$. Design a procedure that identifies $\text{supp}(w) = \{i_1, \cdots, i_k\}$. How large should n be for the procedure to be successful with probability $\geq 1-\alpha$?

Exercise: Let $x \sim \text{Unif}^n([0,\theta])$. Find a $(1-\alpha)$-S for θ. Suppose each $P \in \mathcal{P}$ has median $m = m(P)$, $x \sim P^n$. What is the confidence coefficient of the intervals $[x_{(1)}, \infty), (-\infty, x_{(n)}], [x_{(1)}, x_{(n)}]$?

Remark 1. Smallest confidence sets are hard to obtain, and known in some cases by restricting attention to particular sets (see example below).

Proposition 2. *Let $x \sim \mathcal{N}^n(\mu, \sigma^2)$ with σ^2 known. Then $S(x) = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is shortest among $S'(x) = [\bar{x} - a_1, \bar{x} + a_2]$ with coverage $1-\alpha$.*

Proof. For any such S', $\Pr_\mu(\mu \in S'(x)) = \Pr_Z(a \leq Z \leq b)$ for some fixed a, b proportional to a_1, a_2: $(a = \frac{a_1}{\sigma_{\bar{x}}}, b = \frac{a_2}{\sigma_{\bar{x}}}, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}})$. So we just need to show that if $\Pr_Z(a \leq Z \leq b) = 1-\alpha$ then $b - a \geq 2z_{\alpha/2}$.

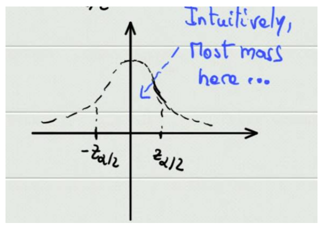

Figure 10.2: Intuition for Optimal Interval

First consider $a < b \leq -z_{\alpha/2}$. By the Mean-Value theorem (MVT),

$$\int_a^b f_z = (b-a)f_z(\alpha_1) = 2z_{\alpha/2}f_z(\alpha_2) = \int_{-z_{\alpha/2}}^{z_{\alpha/2}} f_z \tag{10.3}$$

where $\alpha_1 \in (a,b)$, $\alpha_2 \in (-z_{\alpha/2}, z_{\alpha/2})$. Therefore $f_z(\alpha_1) < f_z(\alpha_2) \implies (b-a) \geq 2z_{\alpha/2}$.

Now consider $a < -z_{\alpha/2} < b < z_{\alpha/2}$. We have

$$\int_a^b f_z = \int_{-z_{\alpha/2}}^{z_{\alpha/2}} f_z + \int_a^{-z_{\alpha/2}} f_z - \int_b^{z_{\alpha/2}} f_z = 1-\alpha \tag{10.4}$$

Then, since $\int_a^{-z_{\alpha/2}} f_z = \int_b^{z_{\alpha/2}}$, the Mean-Value theorem implies $(-z_{\alpha/2} - a) \geq (z_{\alpha/2} - b)$, or $b - a \geq 2z_{\alpha/2}$.

Now we're done since all other cases are solved by symmetry. $\square$

Remark 3. More generally "size" might denote "volume". Other notions of optimality (e.g. Probability of False Coverage) are related to UMPs in hypothesis testing (see 9.3.2 in Casella-Berger).

10.2 Hypothesis Testing

Definition 4. Given $\mathcal{P}$ and some parameter $\theta \in \Theta$, a hypothesis "test" is a procedure to decide between hypotheses of the form $\theta \in \Theta$ (disjoint subsets of Θ) based on observations $x \sim P \in \mathcal{P}$.

Example 2. Let p be the proportion of voters supporting candidate 1 out of 2. Let H_0 be the case where $p \leq 1/2$, i.e. candidate 1 will lose and H_1 be the case where $p > 1/2$, i.e. candidate 1 will win.

Remark 5. We will focus on the case with 2 hypotheses which will illustrate the main ideas.

Definition 6. H_0 is known as the **null hypothesis** and is chosen as a sort of default belief. H_1 is the **alternative hypothesis**.

Definition 7. A **test** T for a family $\mathcal{P}$ defines a **rejection region** $R(T) \subset \mathcal{X}$ (assume supp$(P) \subset \mathcal{X}\ \forall P \in \mathcal{P}$) such that if the observation $x \in R(T)$, then H_0 is rejected; otherwise, if $x \in \mathcal{X} \setminus R(T) = A(T)$, H_1 is rejected (i.e., the null hypothesis is accepted). $A(T)$ is called the **acceptance region**.

T can be viewed as the following function:

$$T(x) = \begin{cases} 1 & x \in R(T) \\ 0 & x \in A(T) \end{cases}$$

1 means "accepting" 1 and 0 means "accepting" 0.

Definition 8. The **error** of a test T for $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$ is err$(T) = \Pr_\theta\left(()\,T = 1\right)\mathbf{1}\{\theta \in \Theta_0\} + \Pr_\theta\left(T = 0\right)\mathbf{1}\{\theta \in \Theta_1\}$.

Type I error is $\Pr_\theta\left(T = 1\right)$ for any $\theta \in \Theta_0$ and Type II error is $\Pr_\theta\left(T = 0\right)$ for any $\theta \in \Theta_1$. Note that these are also called "probability" of Type I/II error.

Remark 9. While err(T), the probability that T chooses the wrong hypothesis, appears to be a natural quantity measure, it has some undesirable properties. For example;

Example 3. Let $x = \{x_i\} \sim \mathcal{N}(\mu, \sigma^2)$, $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$. Consider tests of the form $T_i(x) = 1$ if $\bar{x} \geq t_i$. Consider any $t_1 < t_2$. The difference err$(T_1) -$ err(T_2) is Pr$\left(t_1 \leq x < t_2\right)$ when $\mu \leq \mu_0$ and $-$Pr$\left(t_1 \leq x < t_2\right)$ when $\mu > \mu_0$.

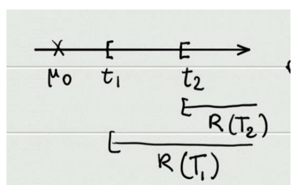

Figure 10.3: Rejection Regions for T_1 and T_2

In other words, none of the two is guaranteed to be better in terms of error. The problem is that our error function treates both Type I and Type II error equally, when in particular settings in practice, we may prefer to minimize one over the other.

The usual solution to this problem is to fix Type I error and aim for a small Type II error.

Definition 10. A test T has significance level α for the problem $H_0 : P \in \mathcal{P}_0$, $H_1 : P\ in \mathcal{P}_1$, $\mathcal{P}_i = P \in \mathcal{P}$ such that $\theta(P) \in \mathcal{P}_i$.

$$\text{size}(T) = \sup_{P \in \mathcal{P}_0} P(R(T)) \leq \alpha \tag{10.5}$$

for $0 < \alpha < 1$. We call $P(R(T))$ the **power** of T, and is denoted $\beta(P)$ and is viewed as a function of P. If $P \in \mathcal{P}_1$, then $\beta(P) = 1-$(Type II error for P).

Example 4. Let $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$. Let $x = \{x_i\} \sim \mathcal{N}^n(\mu, \sigma^2)$ for a known σ^2. Suppose we want $T(x) := 1$ iff $\bar{x} \geq t$. Then notice that

$$\Pr_{\mu_0}(\bar{x} \geq t) \geq \Pr_{\mu}(\bar{x} \geq t) \tag{10.6}$$

for $\mu \leq \mu_0$. So we only need to reconsider μ_0 to get level $1 - \alpha$:

$$\Pr\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \geq z_\alpha\right) = \alpha \tag{10.7}$$

Therefore $t = \mu_0 + \frac{z_\alpha \cdot \sigma}{\sqrt{n}}$.

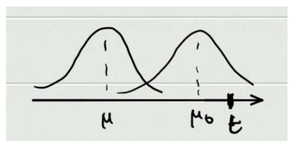

Figure 10.4: H_0 and H_1

Remark 11. By the above, we get the same test for $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$. Tests of the form $H_0 : \theta = \theta_0$ often yield general insight.

Example 5. Same $\mathcal{P} = \mathcal{N}^n(\mu, \sigma^2)$ as before, $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$. $T(x) := 1$ iff $\left|\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma}\right| \geq z_{\alpha/2}$ has size $1 - \alpha$.

Remark 12. Suppose we don't know σ. We can just use $\frac{\sqrt{n}(\bar{x}-\mu_0)}{S_{n-1}} \sim$ Student's t-distribution to get $R(t)$ since the distribution is then known.

Exercise: What is the power $\beta(\mu)$ for the above tests in terms of Φ, the standard normal c.d.f.? Draw an outline of this function. How does $\beta(\mu)$ behave as $n \to \infty$? Note that we want $\beta(\mu)$ to grow quickly towards 1, as $|\mu - \mu_0| \to \infty$. (See Figure 10.5).

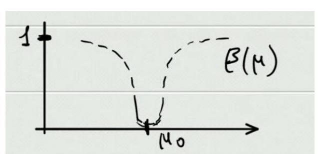

Figure 10.5: The Power