STA437/2005 - Methods for Multivariate Data

Lecture 5

Gun Ho Jang

September 29, 2014

Confidence Region

Definition

A random region $R(\mathbf{X})$ is called a γ-*confidence region* of a parameter θ if

$$P_\theta(\theta \in R(\mathbf{X})) \geq \gamma$$

for any $\theta \in \Theta$.

Example

If $\mathbf{x}_j \sim N_p(\mu, \Sigma)$, then Hotelling's T^2 statistic satisfies $T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \sim \frac{(n-1)p}{n-p} F(p, n-p)$. Thus $P(T^2 = n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \leq \frac{(n-1)p}{n-p} F_\gamma(p, n-p)) = \gamma$ regardless of μ and Σ. Then

$$R(\bar{\mathbf{x}}, S) = \{\mu : n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \leq \frac{(n-1)p}{n-p} F_\gamma(p, n-p)\}$$

is a γ-confidence region for μ.

Confidence Region

Example

- Radiation example in text book.
- Let y_{i1} and y_{i2} be measure radiation with door closed and open, respectively.
- Both Y_1 and Y_2 are not normally distributed.
- Box-Cox transformation is applied with $\lambda = 1/4$ for both of them. Let $x_{ij} = (y_{ij})^{1/4}$.
- sample mean and variance are

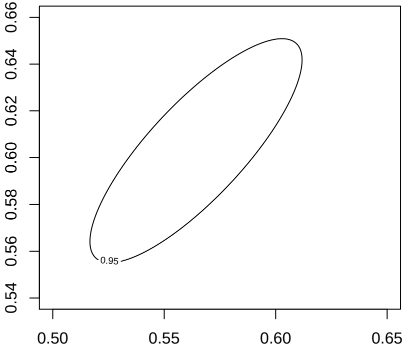
$$\bar{\mathbf{x}} = \begin{pmatrix} 0.5643 \\ 0.6030 \end{pmatrix} \quad S = \begin{pmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{pmatrix}$$

Then 95% confidence region of μ is

$$\begin{aligned} R_{0.95} &= R_{0.95}(\bar{\mathbf{x}}, S) \\ &= \{\mu = (\mu_1, \mu_2)^\top : n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \le F_{0.95}(p, n-p)\frac{(n-1)p}{n-p}\}. \end{aligned}$$

Confidence Region

The corresponding 95%-confidence region is



Confidence Regions of Marginal Parameters I

- a linear combination of mean vector is of interest rather than the full parameter, that is, parameter of interest is

$$\psi = a_1\mu_1 + \cdots + a_p\mu_p$$

- Let $\mathbf{x}_j \sim N_p(\mu, \Sigma)$ and $\mathbf{a} = (a_1, \ldots, a_p)^\top$.
- Define $z_j = \mathbf{a}^\top \mathbf{x}_j$ so that $z_j \sim N(\mathbf{a}^\top \mu, \mathbf{a}^\top \Sigma \mathbf{a}) \sim N(\psi, \zeta)$ where $\zeta = \mathbf{a}^\top \Sigma \mathbf{a}$.
- sample mean and unbiased variance are

$$\bar{z} = \frac{1}{n}\sum_{j=1}^{n} z_j = \frac{1}{n}\sum_{j=1}^{n} \mathbf{a}^\top \mathbf{x}_j = \mathbf{a}^\top \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j = \mathbf{a}^\top \bar{\mathbf{x}}$$

$$s_z^2 = \frac{1}{n-1}\sum_{j=1}^{n}(z_j - \bar{z})^2 = \frac{1}{n-1}\mathbf{a}^\top(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^\top \mathbf{a} = \mathbf{a}^\top S \mathbf{a}.$$

Confidence Regions of Marginal Parameters II

γ-confidence region

Note the t statistic

$$\frac{\bar{z} - \psi}{s_z/\sqrt{n}} = \frac{\sqrt{n}\mathbf{a}^\top(\bar{\mathbf{x}} - \mu)}{\sqrt{\mathbf{a}^\top S\mathbf{a}}} \sim t(n-1).$$

Then a γ-confidence region (or interval) for ψ is

$$\mathbf{a}^\top\bar{\mathbf{x}} \pm t_{(1+\gamma)/2}(n-1)\sqrt{\mathbf{a}^\top S\mathbf{a}}/\sqrt{n}.$$

Confidence Region for vector parameter

- Parameter of interest: $\psi = A\mu \in \mathbb{R}^k$
- $\mathbf{z}_j = A\mathbf{x}_j \sim N_k(A\mu, A\Sigma A^\top)$
- γ-confidence region becomes

$$\{\psi : n(A\bar{\mathbf{x}} - \psi)^\top(ASA^\top)^{-1}(A\bar{\mathbf{x}} - \psi) \le \frac{(n-1)k}{n-k}F_\gamma(k, n-k)\}.$$

Confidence Regions of Marginal Parameters III

- Confidence intervals vary as linear combinations changes
- because of the correlations.
- "Is it possible to have a simple form of simultaneous γ-confidence intervals?"
- Note CIs are $n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)^\top(\bar{\mathbf{x}}-\mu)\mathbf{a}/\mathbf{a}^\top S\mathbf{a} \le c^2$.
- For $\mathbf{a}_1, \ldots, \mathbf{a}_k$,

$$\begin{aligned}
&P(n\mathbf{a}_j^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}_j/(\mathbf{a}_j^\top S\mathbf{a}_j) \le c, j=1,\ldots,k) \\
&\ge P(\max_{\mathbf{a}} n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}/(\mathbf{a}^\top S\mathbf{a}) \le c) \\
&= P(n(\bar{\mathbf{x}}-\mu)^\top S^{-1}(\bar{\mathbf{x}}-\mu) \le c)
\end{aligned}$$

where the last equality can be obtained when $\mathbf{a}$ is proportional to $S^{-1}(\mathbf{x}-\mu)$, that is,

Confidence Regions of Marginal Parameters IV

- Let $\mathbf{b} = S^{1/2}\mathbf{a}$ or $\mathbf{a} = S^{-1/2}\mathbf{b}$

$$\max_{\mathbf{a}} \frac{n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}}{\mathbf{a}^\top S\mathbf{a}} = \max_{\mathbf{b}} \frac{n\mathbf{b}^\top S^{-1/2}(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top S^{-1/2}\mathbf{b}}{\mathbf{b}^\top\mathbf{b}}$$

$$= n\max_{\mathbf{b}} \frac{||(\bar{\mathbf{x}}-\mu)^\top S^{-1/2}\mathbf{b}||^2}{||\mathbf{b}||^2}$$

  Maximum is when $\mathbf{b} \propto S^{-1/2}(\bar{\mathbf{x}}-\mu)$ and $\mathbf{a} \propto S^{-1/2}S^{-1/2}(\bar{\mathbf{x}}-\mu) \propto S^{-1}(\bar{\mathbf{x}}-\mu)$.

- The simultaneous confidence interval is a γ-confidence region,

$$P\left(\frac{n\mathbf{a}_j^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}_j}{\mathbf{a}_j^\top S\mathbf{a}_j} \le \frac{(n-1)p}{n-p}F_\gamma(p,n-p), j=1,\dots,k\right)$$

$$\ge P\left(n(\bar{\mathbf{x}}-\mu)^\top S^{-1}(\bar{\mathbf{x}}-\mu) \le \frac{(n-1)p}{n-p}F_\gamma(p,n-p)\right) = \gamma.$$

Confidence Regions of Marginal Parameters V

Simultaneous confidence interval

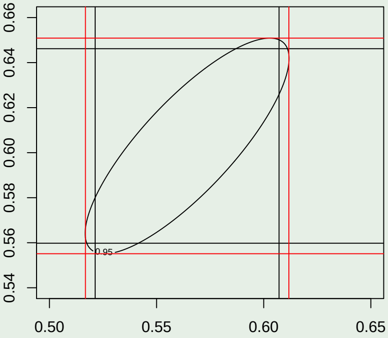
The simultaneous confidence intervals for any **a** is

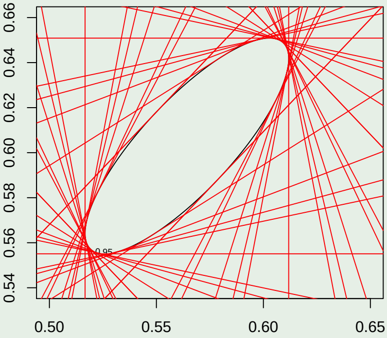
$$\mu_j \in \bar{x}_j \pm \sqrt{\frac{(n-1)p}{n-p} F_\gamma(p, n-p)} \sqrt{\frac{S_{jj}}{n}} \quad \text{for} \quad j = 1, \ldots, p$$

has confidence at least γ.

Confidence Regions of Marginal Parameters VI

Example (Radition Example)





Bonferroni Correction

If all coordinates are independent, simultaneous marginal confidence regions have confidence

$$P(\mu_j \in \bar{\mathbf{x}}_j \pm t_{(1+\gamma)/2}(n-1)\sqrt{S_{jj}/n}, j = 1, \dots, p) = \gamma^p \leq \gamma.$$

It becomes very conservative. To make the confidence close to nominate confidence take γ^* a bit bigger, that is,

$$P(\mu_j \in \bar{\mathbf{x}}_j \pm t_{(1+\gamma^*)/2}(n-1)\sqrt{S_{jj}/n}, j = 1, \dots, p) = 1 - P(\mu_j \notin \bar{\mathbf{x}}_j \pm t_{(1+\gamma^*)/2}(n$$

$$\geq 1 - \sum_{j=1}^{p} P(\mu_j \notin \bar{\mathbf{x}}_j \pm t_{(1+\gamma^*)/2}(n-1)\sqrt{S_{jj}/n}) = 1 - p(1-\gamma^*) \approx \gamma$$

Which gives $\gamma^* = 1 - (1-\gamma)/p \geq \gamma$.

Large Sample Confidence Intervals

When the sample size is large, Hotelling's T^2 statistic follows approximately a $\chi^2(p)$ distribution using the central limit theorem and the continuous mapping theorem. Hence the region

$$\{\mu : n(\bar{\mathbf{x}} - \mu)^\top S^{-1}(\bar{\mathbf{x}} - \mu) \leq \chi^2_\gamma(p)\}$$

has confidence approximately γ.
Similarly, for any vector $\mathbf{a}$, the confidence of the interval

$$\mathbf{a}^\top \bar{\mathbf{x}} \pm \sqrt{\chi^2_\gamma(p)}\sqrt{\mathbf{a}^\top S\mathbf{a}/n}$$

is approximately γ.

Inference with Missing Observations I

- Often there are missing values in practice.
- If the proportion of missing data is not big, then mean and variance matrix can be estimated very efficiently using expectation-maximization (EM) algorithm.

- complete data set $Y_c = (Y_o, Y_m)$ with parameter θ
- Y_o, Y_m are sets of observed/missed data.
- MLE $\hat{\theta}$ can be obtained using the following steps.

Initial step Set an initial parameter $\theta^{(0)}$

E-step Compute the conditional log likelihood

$$Q(\theta \mid \theta^{(l)}) = \mathbb{E}[\log \mathrm{pdf}_{Y_c}(y_o, y_m \mid \theta) \mid \theta^{(l)}]$$

given observed data and current parameter value $\theta^{(l)}$.

M-step Find new estimator $\theta^{(l+1)}$ maximizing $Q(\theta \mid \theta^{(l)})$.

Repeat Repeat E-step and M-step until the parameter converges.

Inference with Missing Observations II

Example (Multivariate Normal)

- $\mathbf{x}_j \sim N_p(\mu, \Sigma)$ with some missing.
- $Q(\mu, \Sigma \mid \hat{\mu}, \hat{\Sigma})$ function is the log likelihood function of complete data with missing values replaced by the conditional expectation given $\hat{\mu}, \hat{\Sigma}$.
- For example, if x_{i4}, x_{i5} are missing while x_{i1}, x_{i2}, x_{i3} are observed, the Q function is the likelihood function of $(x_{i1}, x_{i2}, x_{i3}, \mathbb{E}((x_{i4}, x_{i5}) \mid \hat{\mu}, \hat{\Sigma}))$.
- Plug in Y_m by $\mathbb{E}(Y_m \mid Y_o, \mu^{(l)}, \Sigma^{(l)})$
- Compute new sample mean and variance of (Y_o, Y_m).