STA437/2005 Methods for Multivariate Data

Gun Ho Jang

Lecture on 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.

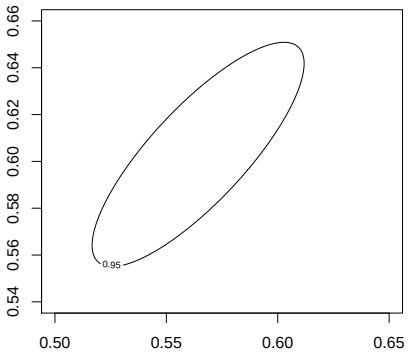
Example. Radiation example in text book. Let y_{i1} and y_{i2} be measure radiation with door closed and open, respectively. Both Y_1 and Y_2 are not normally distributed. Box-Cox transformation is applied with $\lambda = 1/4$ for both of them. Let $x_{ij} = (y_{ij})^{1/4}$. Then sample mean and variance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

$$R_{0.95} = R_{0.95}(\bar{\mathbf{x}}, S) = \{\mu = (\mu_1, \mu_2)^\top : n(\bar{\mathbf{x}}-\mu)^\top S^{-1}(\bar{\mathbf{x}}-\mu) \leq F_{0.95}(p, n-p)(n-1)p/(n-p)\}.$$

Which is given by



Confidence Regions of Marginal Parameters

In many cases, a linear combination of mean vector is of interest rather than the full parameter, that is,

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

is the parameter of interest. 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 $\zeta = \mathbf{a}^\top\Sigma\mathbf{a}$. Hence the sample mean and unbiased variance are given by

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

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

By noting that

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

a γ-confidence region (or interval) for ψ can be obtained by

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

Generally $\psi = A\mu \in \mathbb{R}^k$ is parameter of interest, then $\mathbf{z}_j = A\mathbf{x}_j \sim N_k(A\mu, A\Sigma A^\top)$. Hence a γ-confidence region becomes

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

Confidence intervals vary as linear combinations changes due to the existence of correlations. Naturally a question arises "Is it possible to have a simple form of simultaneous γ-confidence intervals?" Luckily the

answer is yes. The shapes of confidence intervals are $n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)^\top(\bar{\mathbf{x}}-\mu)\mathbf{a}/\mathbf{a}^\top S\mathbf{a} \le c^2$. The simultaneous confidence intervals satisfy, for $\mathbf{a}_1, \ldots, \mathbf{a}_k$,

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

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

$$\max_{\mathbf{a}} \frac{n\mathbf{a}^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}}{\mathbf{a}^\top S\mathbf{a}}$$

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

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

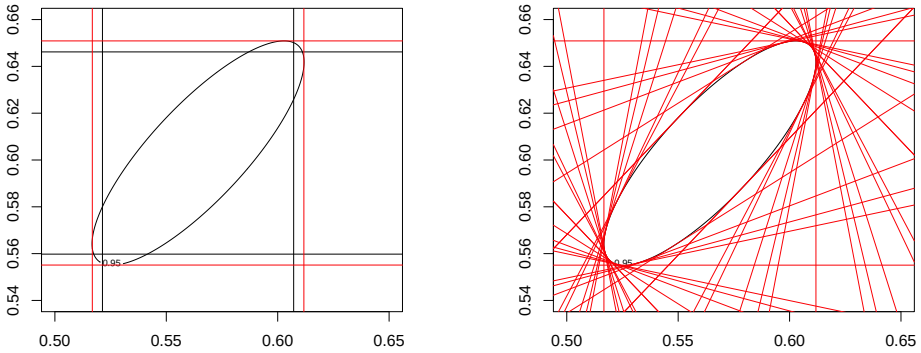
Hence the maximum is obtained when $\mathbf{b}$ is proportional to $S^{-1/2}(\bar{\mathbf{x}}-\mu)$, that is, $\mathbf{a}$ is proportional to $S^{-1/2}S^{-1/2}(\bar{\mathbf{x}}-\mu) = S^{-1}(\bar{\mathbf{x}}-\mu)$. Therefore, the simultaneous confidence interval is a γ-confidence region, that is,

$$P(n\mathbf{a}_j^\top(\bar{\mathbf{x}}-\mu)(\bar{\mathbf{x}}-\mu)^\top\mathbf{a}_j/(\mathbf{a}_j^\top S\mathbf{a}_j) \le F_\gamma(p,n-p)(n-1)p/(n-p), j=1,\ldots,k) \ge P(n(\bar{\mathbf{x}}-\mu)^\top S^{-1}(\bar{\mathbf{x}}-\mu) \le F_\gamma(p,n-p)(n$$

The simultaneous confidence intervals

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

has confidence at least γ.



Bonferroni Correction

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

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

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

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

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

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.

EM algorithm

Consider a complete data set $Y_c = (Y_o, Y_m)$ with parameter θ where Y_o, Y_m are sets of observed/missed data. The maximum likelihood estimator $\hat{\theta}$ can be obtained using the following steps.

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

E-step Compute the conditional log likelihood

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

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

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

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

When $\mathbf{x}_j \sim N_p(\mu, \Sigma)$ with some missing, the $Q(\mu, \Sigma \,|\, \hat{\mu}, \hat{\Sigma})$ 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 $(x_{i1}, x_{i2}, x_{i3}, \mathbb{E}((x_{i4}, x_{i5}) \,|\, \hat{\mu}, \hat{\Sigma}))$.