Multivariate Statistics

Sudipta Das

Assistant Professor,
Department of Data Science,
Ramakrishna Mission Vivekananda University, Kolkata
Source: Jhonson & Winchern

Outline I

1. Multivariate Normal Inference
   - Hypothesis & Testing
   - Interval Estimation
     - Simultaneous Confidence Intervals
     - One-at-a-Time Confidence Intervals
     - Boneferroni Confidence Intervals
     - Large Sample Confidence Intervals

Inference about Mean (Univariate) I

- Univariate Normal Distribution: $X_1, X_2, \ldots, X_n$ denote a random samle from a normal population with mean μ and varaince σ^2.
- Then

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

follows student's t-distribution with $n-1$ d.f., where $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$

and $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$.

Inference about Mean (Univariate) II

- Hypotheses and Testing

$$H_0 : \mu = \mu_0 \text{ and } H_1 : \mu \neq \mu_0$$

  - Reject H_0, in favor of H_1, at significance level α, if

$$t^2 = (\bar{X} - \mu_0) \left(\frac{1}{n} s^2 \right)^{-1} (\bar{X} - \mu_0) > t^2_{n-1}(\alpha/2),$$

    where $t_{n-1}(\alpha/2)$ denotes the upper $100(\alpha/2)$th percentile of the t-distribution with $n-1$ d.f.
  - Note that equivalently,
    - t^2 follows $F_{1,n-1}$ distribution
    - Thus, reject H_0 if $t^2 > F_{1,n-1}(\alpha)$

Inference about Mean (Univariate) III

- Interval estimation
  - The $100(1-\alpha)\%$ confidence interval of mean (μ) is

$$\bar{X} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}.$$

Inference about Mean vector (Multivariate) I

- Multivariate Normal Distribution: $X_1, X_2, \ldots, X_n$ denote a multivariate random samle from a normal population with mean μ and varaince Σ.
- Then

$$T^2 = (\bar{X} - \mu_0)' \left(\frac{1}{n} S \right)^{-1} (\bar{X} - \mu_0)$$

follows $\frac{(n-1)p}{n-p} F_{p,n-p}$ distribution, where

  - $\bar{X}_{p\times 1} = \frac{1}{n} \sum_{j=1}^{n} X_j$ and
  - $S_{p\times p} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})'(X_j - \bar{X})$.

- Note that

$$
\begin{aligned}
T^2 &= \sqrt{n}(\bar{X}-\mu_0)' \left(\frac{(n-1)S}{n-1}\right)^{-1} \sqrt{n}(\bar{X}-\mu_0) \\
&= [N_p(0,\Sigma)]' \left[\frac{W_{p,n-1}(\Sigma)}{n-1}\right] [N_p(0,\Sigma)], \text{ where}
\end{aligned}
$$

  - $\sqrt{n}(\bar{X}-\mu_0)$ follows $N_p(0,\Sigma)$ and
  - $(n-1)S$ follows $W_{p,n-1}(\Sigma)$, Wishart distribution of $(n-1)$ d.f.,
    - Wishart distribution of $(n-1)$ d.f. is the distribution of $\sum_{j=1}^{n-1} \mathbf{Z}_j\mathbf{Z}_j'$, where $\mathbf{Z}_j \sim N_p(0,\Sigma)$.
- Hotelling's T^2 statistics.
- Example 5.1 (Page 213)

Hypothesis & Testing on Mean vector I

- Hypotheses

$$H_0 : \mu = \mu_0 \text{ and } H_1 : \mu \neq \mu_0$$

- Reject H_0, in favor of H_1, at significance level α, if

$$T^2 = n(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha),$$

  where $F_{p,n-p}(\alpha)$ denotes the upper 100αth percentile of the F-distribution with p and $n - p$ d.f.

- Example 5.2 (Page 214)

Hypothesis & Testing on Mean vector II

- Likelihood ratio statistics:

$$\Lambda = \frac{\max_{\Sigma} L(\mu_0, \Sigma)}{\max_{\mu,\Sigma} L(\mu, \Sigma)} = \left[\frac{e^{-np/2}}{(2\pi)^{np/2}|\hat{\Sigma}_0|^{n/2}}\right]\left[\frac{(2\pi)^{np/2}|\hat{\Sigma}|^{n/2}}{e^{-np/2}}\right] = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|}\right)^{\frac{n}{2}},$$

where $\hat{\Sigma}_0 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_0)(x_j - \mu_0)'$ and $\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})'$.

  - If the observed value of the likelihood ratio is too small, the hypothesis $H_0 : \mu = \mu_0$ is rejected.
- When n is large, under the null hypothesis H_0, $-2\ln\Lambda$ is approximately $\chi^2_{\nu-\nu_0=p}$.
  - Unrestricted degrees of freedom: $\nu = p + p(p+1)/2$ and
  - Degrees of freedom under the null hypothesis: $\nu_0 = p(p+1)/2$.

- Connection between Hotelling T^2 Statistics and Likelihood ratio test
- Wilks' lambda: $\Lambda^{\frac{2}{n}} = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} = \left(1 + \frac{T^2}{n-1}\right)^{-1}$.

Interval Estimation of Mean vector I

- Confidence Region
- Let θ be a vector of unknown population parameters and Θ be the set of all possible values of θ.
- Goal is to find a region $R(\mathbf{X})$ such that

$$P[\theta \in R(\mathbf{X})] = 1 - \alpha.$$

Interval Estimation of Mean vector II

- A $100(1-\alpha)\%$ **confidence region for the mean vector** of a $p-$dimensional normal distribution is the ellipsoid determined by all μ such that

$$n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha).$$

- Axes of confidence interval and their relative lengths
  - The directions and lengths of the axes of

$$(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \leq \frac{(n-1)p}{(n-p)n}F_{p,n-p}(\alpha)$$

    are determined by lengths $\sqrt{\lambda_i}\sqrt{\frac{(n-1)p}{(n-p)n}F_{p,n-p}(\alpha)}$s along eigenvector $\mathbf{e_i}$s, respectively.
  - Note that, $S\mathbf{e_i} = \lambda_i \mathbf{e_i}$ for $i = 1, \ldots, p$.
- Beginning at the center $\bar{x}$, the axes of the confidence ellipsoids are $\pm\sqrt{\lambda_i}\sqrt{\frac{(n-1)p}{(n-p)n}F_{p,n-p}(\alpha)}\mathbf{e_i}$.
- Example 5.3 (Page 221)

Interval Estimation of Mean vector IV

- Problem of Interpretation of elliptical confidence range
  - Summary of statistical conclusions need confidence statements about individual component means.
  - One needs something of the form "$\mu_i \in [\bar{x} \pm \text{something}], \forall i = 1, \ldots, p$" rather than by saying that "by all μ such that $n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$."

Simultaneous Confidence Intervals I

- Create the intervals in the way, such that
  - *the confidence statement holds simultaneously, for all the individual components*
- Conservatively, for all the linear combinations (i.e. for any **a**) of the components

$$Z = \mathbf{a}'\mathbf{X},$$

the interval $[\bar{Z} \pm c \times \frac{s_Z}{\sqrt{n}}]$ will contain the μ_Z with probability $1 - \alpha$

  - In other words

$$P\left(\mathbf{a}'\mu \in \left[\mathbf{a}'\bar{\mathbf{X}} \pm c \times \sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}\right]\right) = 1 - \alpha$$

Simultaneous Confidence Intervals II

- Result: Let $X_1, X_2, \ldots, X_n$ be a random sample from an $N_p(\mu, \Sigma)$ population with Σ positive definite. Then, simultaneously for all $\mathbf{a}$, the interval

$$\left(\mathbf{a}'\bar{X} - \sqrt{\frac{(n-1)p}{(n-p)n}F_{p,n-p}(\alpha)\mathbf{a}'\mathbf{S}\mathbf{a}}, \mathbf{a}'\bar{X} + \sqrt{\frac{(n-1)p}{(n-p)n}F_{p,n-p}(\alpha)\mathbf{a}'\mathbf{S}\mathbf{a}}\right)$$

will contain $\mathbf{a}'\mu$ with probability $1-\alpha$.

Simultaneous Confidence Intervals III

- Sketch of proof:
  - We need a constant 'c' such that

$$P\left(\left\|\frac{\mathbf{a}'(\bar{\mathbf{X}}-\mu)}{\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}}\right\| \leq c\right) \geq 1-\alpha$$

    for all $\mathbf{a}$.
  - Equivalently,

$$P\left(\max_{a}\left[\frac{\mathbf{a}'(\bar{\mathbf{X}}-\mu)}{\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}}\right]^2 \leq c^2\right) \geq 1-\alpha$$

- Now, (by Maximization Lemma in Page 80),

$$
\begin{aligned}
\max_a \left(n \frac{[\mathbf{a}'(\bar{\mathbf{X}} - \mu)]^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} \right) &= n \max_a \left(\frac{[(\mathbf{S}^{\frac{1}{2}}\mathbf{a})'(\mathbf{S}^{-\frac{1}{2}}(\bar{\mathbf{X}} - \mu))]^2}{(\mathbf{S}^{\frac{1}{2}}\mathbf{a})'(\mathbf{S}^{\frac{1}{2}}\mathbf{a})} \right) \\
&= n(\mathbf{S}^{-\frac{1}{2}}(\bar{\mathbf{X}} - \mu))'(\mathbf{S}^{-\frac{1}{2}}(\bar{\mathbf{X}} - \mu)) \\
&= n(\bar{\mathbf{X}} - \mu)'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \mu) \\
&= T^2
\end{aligned}
$$

- We know,

$$P\left(T^2 \leq c^2\right) = 1 - \alpha,$$

for $c^2 = \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$.

- Hence, the result!

Simultaneous Confidence Intervals V

- For different choices of $\mathbf{a}' = [1, 0, \ldots, 0], \mathbf{a}' = [0, 1, \ldots, 0], \ldots, \mathbf{a}' = [0, 0, \ldots, 1]$, we can say,

$$\bar{x}_1 - \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{11}}{n}} \leq \mu_1 \leq \bar{x}_1 + \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{11}}{n}}$$

$$\bar{x}_2 - \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{22}}{n}} \leq \mu_2 \leq \bar{x}_2 + \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{22}}{n}}$$

$$\vdots$$

$$\bar{x}_p - \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{pp}}{n}} \leq \mu_p \leq \bar{x}_p + \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{pp}}{n}}.$$

- Example 5.4 (Page 226)
- Drawback: As a combination, the overall CI is larger than $(1 - \alpha)$.

One-at-a-Time Confidence Intervals

- Ignoring the covariance structure of multivariate data, we can give the individual CI as following,

$$\bar{x}_1 - t_{n-1}(\alpha/2)\sqrt{\frac{s_{11}}{n}} \leq \mu_1 \leq \bar{x}_1 + t_{n-1}(\alpha/2)\sqrt{\frac{s_{11}}{n}}$$

$$\bar{x}_2 - t_{n-1}(\alpha/2)\sqrt{\frac{s_{22}}{n}} \leq \mu_2 \leq \bar{x}_2 + t_{n-1}(\alpha/2)\sqrt{\frac{s_{22}}{n}}$$

$$\vdots$$

$$\bar{x}_p - t_{n-1}(\alpha/2)\sqrt{\frac{s_{pp}}{n}} \leq \mu_p \leq \bar{x}_p + t_{n-1}(\alpha/2)\sqrt{\frac{s_{pp}}{n}}.$$

- Drawback: As a combination, the overall CI is lesser than $(1-\alpha)$.
- Table 5.3 (Page 231)

Boneferroni Confidence Intervals I

- Boneferroni simultaneous Confidence Intervals

$$\bar{x}_1 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}} \leq \mu_1 \leq \bar{x}_1 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}}$$

$$\bar{x}_2 - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}} \leq \mu_2 \leq \bar{x}_2 + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}}$$

$$\vdots$$

$$\bar{x}_p - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}} \leq \mu_p \leq \bar{x}_p + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}}.$$

Boneferroni Confidence Intervals II

- Note:

$$
\begin{aligned}
P\left(\mu_i \in \left[\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}\right], \text{ for all } i\right) &= P\left(\bigcap_{i=1}^{p} \mu_i \in \left[\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}\right]\right) \\
&= 1 - P\left(\bigcup_{i=1}^{p} \mu_i \notin \left[\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}\right]\right) \\
&\geq 1 - \sum_{i=1}^{p} P\left(\mu_i \notin \left[\bar{x}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}\right]\right) \\
&= 1 - \sum_{i=1}^{p} \frac{\alpha}{p} = 1 - \alpha
\end{aligned}
$$

- Bonferroni simultaneous CI is also more than $(1-\alpha)$
  - but less than T^2 simultaneous CI.
- Example 5.6 (Page 233)

Large Sample Confidence Intervals I

- Large sample inference of the population mean vector
- Advantage: Departure from assumption of normal population is overcome by large sample size.
- A $100(1-\alpha)\%$ confidence region for the mean of a $p-$dimensional distribution is the ellipsoid determined by all μ such that

$$n(\bar{X}-\mu)S^{-1}(\bar{X}-\mu) \leq \chi_p^2(\alpha),$$

provided n and $n-p$ are large.

Large Sample Confidence Intervals II

- Similarly, $100(1-\alpha)\%$ confidence region

$$\bar{x}_1 - \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{11}}{n}} \le \mu_1 \le \bar{x}_1 + \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{11}}{n}}$$

$$\bar{x}_2 - \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{22}}{n}} \le \mu_2 \le \bar{x}_2 + \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{22}}{n}}$$

$$\vdots$$

$$\bar{x}_p - \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{pp}}{n}} \le \mu_p \le \bar{x}_p + \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{pp}}{n}}.$$

- Note:

$$\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha) \to \chi^2_p(\alpha)$$

as $n - p \to \infty$.