Inferences about Mean Vectors

Univariate Case:

\[ X_1, X_2, \ldots, X_n \sim N_1(\mu, \sigma^2) \]

Hypothesis testing

\[ H_0 : \mu = \mu_0 \text{ and } H_1 : \mu = \mu_0 \]

\[ \bar{X}, s \Rightarrow t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \]

Under \( H_0 \), \( t \) follows the student's \( t \)-distribution with \( n - 1 \) degrees of freedom

Decision rule: reject \( H_0 \), if \( |t| > t_{n-1}(\alpha/2) \)

which is equivalent to: don’t reject \( H_0 \), if \( |t| \leq t_{n-1}(\alpha/2) \)

\[ |t| \leq t_{n-1}(\alpha/2) \Leftrightarrow \left| \frac{\bar{X} - \mu}{s/\sqrt{n}} \right| \leq t(\alpha/2) \]

\[ \Leftrightarrow \bar{X} - t(\alpha/2) \frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t(\alpha/2) \frac{s}{\sqrt{n}} \]

Decision rule based on \( [\bar{X} - t(\alpha/2) \frac{s}{\sqrt{n}}, \bar{X} + t(\alpha/2) \frac{s}{\sqrt{n}}] \)

Don’t reject \( H_0 \), if \( \mu_0 \in [\bar{X} - t(\alpha/2) \frac{s}{\sqrt{n}}, \bar{X} + t(\alpha/2) \frac{s}{\sqrt{n}}] \).

Confidence Interval:

Random interval: \( [\bar{X} - t(\alpha/2) \frac{s}{\sqrt{n}}, \bar{X} + t(\alpha/2) \frac{s}{\sqrt{n}}] \), denoted by \( I \).

The probability that \( I \) does contain the true population mean:

\[ P[I \text{ contains the true mean}] \]

\[ = P[\bar{X} - t(\alpha/2) \frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t(\alpha/2) \frac{s}{\sqrt{n}} | \mu \text{ is true}] \]
\[ P\left[ \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t(\alpha/2) \mid \mu \text{ is true} \right] = 1 - \alpha. \]

Hence the Confidence level is 100(1 - \alpha)\%

Questions:

What does the confidence level mean? If you are given with a confidence interval [170, 180] with confidence level 95\%, can you say that the probability that the interval contains the true mean is 95\%? How to construct a confidence interval with a given confidence level?

**Multivariate Case**

\[ H_0 : \bar{\mu} = \bar{\mu}_0 \text{ and } H_1 : \bar{\mu} \neq \bar{\mu}_0 \]

Recall the univariate case:

Reject \( H_0 \), when \( t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} > t(\alpha/2) \)

\( \iff t^2 = \left( \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right)^2 > t^2(\alpha/2) \)

\( \iff t^2 = n((X) - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0) > t^2(\alpha/2) \)

By analogy, generalize to multivariate case:

\[ T^2 = n(\bar{X} - \bar{\mu}_0)(S)^{-1}(\bar{X} - \bar{\mu}_0) \]

(Hotelling’s \( T^2 \) statistic)

Under \( H_0 \), what is the distribution of \( T^2 \):

\[ T^2 \sim \frac{(n - 1)p}{(n - p)} F_{p, n - p} \]
Decision rules at significance level $\alpha$:

$$\text{Reject } H_0, \text{ if } T^2 = n(\bar{X} - \bar{\mu_0})(S)^{-1}(\bar{X} - \bar{\mu_0}) > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

**Likelihood Ratio Test**

$H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$

$\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n$ is a random sample (sample). The likelihood function is:

$$L(\mu, \Sigma) = \left\{ \text{joint density of } \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n \right\}$$

**MLEs:**

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})'$$

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} \bar{X}_i$$

The maximum likelihood is:

$$L(\hat{\mu}, \hat{\Sigma}) = \max_{\mu, \Sigma} L(\mu, \Sigma) = (2\pi)^{-np/2}e^{-n\mu'/2} | \hat{\Sigma} |^{-n/2}$$

Now under $H_0$: $\mu_0$ is fixed

$$\hat{\mu}_0 = \mu_0$$

$$\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} (\bar{X}_i - \mu_0)(\bar{X}_i - \mu_0)'$$

and the maximum likelihood is

$$L(\hat{\mu}_0, \hat{\Sigma}_0) = \max_{\mu_0, \Sigma} L(\mu_0, \Sigma) = (2\pi)^{-np/2}e^{-n\mu'/2} | \hat{\Sigma}_0 |^{-n/2}$$
Clearly \( L(\mu_0, \Sigma_0) \leq L(\hat{\mu}, \hat{\Sigma}) \). So how different are they? The ratio between them is called the likelihood ratio

\[
\Lambda = \frac{L(\mu_0, \Sigma_0)}{L(\hat{\mu}, \hat{\Sigma})} = \frac{|\Sigma_0|^{-n/2}}{|\hat{\Sigma}|^{-n/2}} = \left( \frac{|\Sigma|}{|\Sigma_0|} \right)^{n/2}
\]

\( \Lambda^{2/n} \) is called the Wilk’s lambda statistic. The decision rule is

Reject \( H_0 \), if \( \Lambda < c_\alpha \)

**Results:**

\[
\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1}\right)^{-1}
\]

Decision rule:

Reject \( H_0 \), if \( \Lambda < c_\alpha \)

\[
\Leftrightarrow \Lambda^{2/n} < c_\alpha \Leftrightarrow \left(1 + \frac{T^2}{n-1}\right)^{-1} < (c_\alpha)^{2/n}
\]

\[
\Leftrightarrow (c_\alpha)^{-2/n} > \frac{T^2}{n-1} \Leftrightarrow T^2 > (n-1)[(c_\alpha)^{-2/n} - 1]
\]

\[
\Leftrightarrow T^2 > c^*_\alpha
\]

Conclusion: Hotelling’s \( T^2 \) test is equivalent to likelihood ratio test.

**Confidence Regions**

100(1 - \( \alpha \))% confidence region \( R(X) \):

\[
P[R(X) \text{ will cover the true parameter } \theta] = 1 - \alpha
\]
Recall: under $H_0 : \mu = \mu_0$

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

Now, define

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

$R(X)$ covers the true mean $\mu$ is equivalent to

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

Since

$$P[n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu) \leq \frac{(n - 1)p}{n - p} F_{p,n-1}(\alpha) | \mu \text{ is the true mean} ] = 1 - \alpha$$

$R(X)$ is a $100(1 - \alpha)$% confidence interval for $\mu$.

**Simultaneous Confidence Intervals**

$$\bar{\mu} = (\mu_1, \mu_2, \ldots, \mu_p), \ \bar{X}' = (X_1, X_2, \ldots, X_p)$$

Confidence interval for $a'\mu = a_1\mu_1 + a_2\mu_2 + \ldots + a_p\mu_p$. Let $Z = a_1X_1 + a_2X_2 + \ldots + a_pX_p$. $\mu_Z = a'\mu, \ \sigma_Z^2 = a'\Sigma a$, and $Z \sim N_1(a'\mu, a'\Sigma a)$. The sample mean and sample variance for $Z$ are $\bar{Z} = a'\bar{X}$ and $S_Z^2 = a'Sa$. And,

$$t = \frac{\bar{Z} - \mu_Z}{S_Z/\sqrt{n}} = \frac{\sqrt{n}(a'\bar{X} - a'\mu)}{\sqrt{a'Sa}}$$

For a specific vector $a$, the $100(1 - \alpha)$% confidence interval for $a'\mu$ is given as follows,

$$a'\bar{X} - t_{n-1}(\alpha/2)\frac{\sqrt{a'Sa}}{\sqrt{n}} \leq a'\bar{X} - t_{n-1}(\alpha/2)\frac{\sqrt{a'Sa}}{\sqrt{n}}$$

$$\Leftrightarrow \{ a'\mu : t^2 = \frac{n(a'\bar{X} - a'\mu)^2}{a'Sa} = \frac{n(a'(\bar{X} - \mu))^2}{a'Sa} \leq \frac{\hat{\rho}^2}{n-1}(\alpha/2) \}$$
In order to make the confidence statement for all possible \( a^t\mu \), we need to determine \( c \) such that
\[
1 - \alpha = P[\text{for all } a, t^2 \leq c^2]
\]
\[
= P[\text{for all } a, \frac{n(a^t(\bar{X} - \mu))^2}{a^tSa} \leq c^2]
\]
\[
= P[\max_a \frac{n(a^t(\bar{X} - \mu))^2}{a^tSa} \leq c^2]
\]
We have the following result
\[
\max_a \frac{n(a^t(\bar{X} - \mu))^2}{a^tSa} = n(\bar{X} - \mu)^tS(\bar{X} - \mu) = T^2
\]
Since \( T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p} \),
\[
1 - \alpha = P[T^2 \leq c^2] \Rightarrow c^2 = \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)
\]
Hence simultaneous confidence interval for all \( a \) is given by
\[
\frac{n(a^t\bar{X} - a^t\mu)^2}{a^tSa} \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)
\]
\[
\Leftrightarrow a^t\bar{X} - \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p} \frac{\sqrt{a^tSa}}{n}} \leq a^t\mu \leq a^t\bar{X} + \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p} \frac{\sqrt{a^tSa}}{n}}
\]
**Result:** Let \( \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n \) be a random sample from \( N_p(\mu, \Sigma) \) population with \( \Sigma \) positive definite. Then, simultaneously for all \( a \), the above intervals will contain \( a^t\mu \) with probability \( 1 - \alpha \).

**Bonferroni Method for Multiple Comparison**

One-at-a-time intervals for \( \mu_1, \mu_2, \ldots, \mu_m \):
\[
\bar{X}_1 - t_{n-1}(\alpha/2) \sqrt{\frac{s^2_1}{n}} \leq \mu_1 \leq \bar{X}_1 + t_{n-1}(\alpha/2) \sqrt{\frac{s^2_1}{n}}
\]
\[
\bar{X}_2 - t_{n-1}(\alpha/2) \sqrt{\frac{s^2_2}{n}} \leq \mu_2 \leq \bar{X}_2 + t_{n-1}(\alpha/2) \sqrt{\frac{s^2_2}{n}}
\]
6
\[ X_m - t_{n-1}(\alpha/2) \sqrt{\frac{\hat{s}_{mm}}{n}} \leq \mu_1 \leq X_m + t_{n-1}(\alpha/2) \sqrt{\frac{\hat{s}_{mm}}{n}} \]

But, the statement that all the $t$ intervals contain the $\mu_i$’s do not the confidence level $1 - \alpha$.

In fact,

\[ P[\text{all } t \text{ intervals contain the } \mu_i \text{'s }] < 1 - \alpha \]

Bonferroni Inequality:

Suppose $P[C_i \text{ is true}] = 1 - \alpha_i$, $i = 1, 2, \ldots, m$, then

\[ P[\text{all } C_i \text{ are true}] \geq 1 - \sum_{i=1}^{m} P[C_i \text{ is false}] = 1 - (\alpha_1 + \alpha_2 + \ldots + \alpha_m) \]

If $\alpha_1 + \alpha_2 + \ldots + \alpha_m = \alpha$, $P[\text{all } C_i \text{'s are true}] \geq 1 - \alpha$.

Bonferroni Intervals for $\mu_1, \ldots, \mu_m$:

\[ X_i \pm t_{n-1}(\alpha/2m) \sqrt{\frac{\hat{s}_{ii}}{n}}, i = 1, 2, \ldots, m \]

So that

\[ P[X_i \pm t_{n-1}(\alpha/2m) \sqrt{\frac{\hat{s}_{ii}}{n}} \text{ contain } \mu_i, \text{ for all } i] \geq 1 - \alpha \]

Similarly, for any given $m$ vectors, $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m$, we can construct the Bonferroni confidence intervals for $\bar{a}_i^\prime\bar{\mu}$ where $i = 1, 2, \ldots, m$.

Large Sample Inference

The population model is not assumed to be normal. When the sample size is large,

\[ n(\bar{X} - \mu)^\prime S^{-1}(\bar{X} - \mu) \sim \chi^2_p \text{ approximately} \]

So,

\[ P[n(\bar{X} - \mu)^\prime S^{-1}(\bar{X} - \mu) \leq \chi^2_p(\alpha)] = 1 - \alpha \text{ approximately} \]

Hypothesis testing:
\[ H_0 : \mu = \mu_0 \text{ and } H_1 : \mu \neq \mu_0 \]

Decision rule:

Reject \( H_0 \) if \( n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu) > \chi^2_p(\alpha) \)

Simultaneous confidence interval for all \( a'\mu \):

\[ a'\bar{X} \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{a'Sa}{n}} \]

**Inferences with Missing Observations**

missing by random mechanism

EM algorithm:

Step 1 (Prediction step):

Given some estimate \( \hat{\theta} \) of the unknown parameters, predict the contribution of any missing observation to the complete-data sufficient statistics.

Step 2 (Estimation step):

Use the predicted sufficient statistics to compute a revised estimate of the parameters