STAT 3690 Lecture 10

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Feb 14, 2022

Testing on $\boldsymbol{\mu}$ (J&W Sec. 5.2 & 5.3)

- $\mathbf{X}_1,\ldots,\mathbf{X}_n \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu},\Sigma)$ $n>p$
- Hypotheses: $H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0$ v.s. $H_1:\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$
- Recall the univariate case ($p=1$)
  - The model reduces to $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\mu,\sigma^2)$
  - Hypotheses reduces to $H_0:\mu=\mu_0$ v.s. $H_1:\mu\neq\mu_0$
  - $\bar X$ and s^2 are sample mean and sample variance, respectively
  - Known σ^2
    * Name of approach: Z-test (also LRT)
    * Test statistic: $\sqrt{n}(\bar X-\mu_0)/\sigma\sim N(0,1)$ under H_0
      · OR $n(\bar X-\mu_0)^2/\sigma^2\sim\chi(1)$ under H_0
    * Rejction region at level α: $R=\{x_1,\ldots,x_n:\sqrt{n}|\bar x-\mu_0|/\sigma\geq\Phi^{-1}_{1-\alpha/2}\}=\{x_1,\ldots,x_n: n(\bar x-\mu_0)^2/\sigma^2\geq\chi^2_{1-\alpha,1}\}$
      · $\Phi^{-1}_{1-\alpha/2}$: the $(1-\alpha/2)$-quantile of $N(0,1)$
      · $\chi^2_{1-\alpha,1}$: the $(1-\alpha)$-quantile of $\chi^2(1)$
  - Unknown σ^2
    * Name of approach: t-test (also LRT)
    * Test statistic: $\sqrt{n}(\bar X-\mu_0)/s\sim t(n-1)$ under H_0
      · OR $n(\bar X-\mu_0)^2/s^2\sim F(1,n-1)$ under H_0
    * Rejction region at level α: $R=\{x_1,\ldots,x_n:\sqrt{n}|\bar x-\mu_0|/s\geq t_{1-\alpha/2,n-1}\}=\{x_1,\ldots,x_n: n(\bar x-\mu_0)^2/s^2\geq F_{1-\alpha,1,n-1}\}$
      · $t_{1-\alpha/2,n-1}$: the $(1-\alpha/2)$-quantile of $t(n-1)$
      · $F_{1-\alpha,1,n-1}$: the $(1-\alpha)$-quantile of $F(1,n-1)$

- Multivariate case (with known $\boldsymbol{\Sigma}$)
  - Name of approach: LRT
  - Test statistic: $n(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)\sim\chi^2(p)$ under H_0
  - Rejction region at level α: $R=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n: n(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)\geq\chi^2_{1-\alpha,p}\}$
  - p-value: $p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)=1-F_{\chi^2(p)}\{n(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)\}$
    * $F_{\chi^2(p)}$: the cdf of $\chi^2(p)$

```
options(digits = 4)
install.packages(c("dslabs"))
library(dslabs)
data("gapminder")
```

```r
head(gapminder)
dataset = as.matrix(gapminder[
  !is.na(gapminder$infant_mortality),
  c("infant_mortality", "life_expectancy", "fertility")])

# Assume we know Sigma
Sigma <- matrix(c(555, -170, 30,
                  -170, 65, -10,
                  30, -10, 2), ncol = 3)

(mu_hat <- colMeans(dataset))

# Test mu = mu_0
mu_0 <- c(25, 50, 3)
n = nrow(dataset)
p = ncol(dataset)
(test.stat <- drop(
   n * t(mu_hat - mu_0) %*% solve(Sigma) %*% (mu_hat - mu_0)
))
test.stat >= qchisq(0.95, df=p)
(p.val = 1-pchisq(test.stat, df=p))
```

- Report: Testing hypotheses $H_0 : \boldsymbol{\mu} = [25, 50, 3]^\top$ v.s. $H_1 : \boldsymbol{\mu} \neq [25, 50, 3]^\top$, we carried on the LRT and obtained 450477 as the value of test statistic. The corresponding p-value (resp. rejection region) was 0 (resp. $[7.815, \infty)$). So, at the .05 (significance) level, there was a strong statistical evidence implying the rejection of H_0, i.e., we believed that the mean vector is not $[25, 50, 3]^\top$.

- Multivariate case (with unknown $\boldsymbol{\Sigma}$)
  - Name of approach: LRT
  - Test statistic: $n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \sim T^2(p, n-1) = \frac{(n-1)p}{n-p} F(p, n-p)$ under H_0
  - Rejction region at level α: $R = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n : \frac{n(n-p)}{p(n-1)} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \geq F_{1-\alpha, p, n-p}\}$
  - p-value: $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 1 - F_{F(p,n-p)}\{\frac{n(n-p)}{p(n-1)} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\}$
    - $F_{F(p,n-p)}$: the cdf of $F(p, n-p)$

Σ is known:

$$\sup_{(\mu,\Sigma)\in\Theta} L(\mu,\Sigma) = L(\bar{X},\Sigma) = (2\pi)^{-np/2}(\det\Sigma)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^\top\Sigma^{-1}(X_i-\bar{X})\right\}$$

$$\begin{aligned}\sup_{(\mu,\Sigma)\in\Theta_0} L(\mu,\Sigma) = L(\mu_0,\Sigma) &= (2\pi)^{-np/2}(\det\Sigma)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_0)^\top\Sigma^{-1}(X_i-\mu_0)\right\}\\ &= (2\pi)^{-np/2}(\det\Sigma)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu_0)^\top\Sigma^{-1}(X_i-\bar{X}+\bar{X}-\mu_0)\right\}\\ &= (2\pi)^{-np/2}(\det\Sigma)^{-n/2}\exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}(X_i-\bar{X})^\top\Sigma^{-1}(X_i-\bar{X}) + n(\bar{X}-\mu_0)^\top\Sigma^{-1}(\bar{X}-\mu_0)\right\}\right]\end{aligned}$$

$$\Rightarrow \lambda = \frac{L(\mu_0,\Sigma)}{L(\bar{X},\Sigma)} = \exp\left\{-\frac{1}{2}\underbrace{n(\bar{X}-\mu_0)\Sigma^{-1}(\bar{X}-\mu_0)}_{=T^2}\right\}$$

$\Rightarrow$ Rejection region $\mathcal{R} = \{\lambda \le c_1\} = \{T^2 \ge c_2\}$ for some $c_1, c_2 \in \mathbb{R}$

Let $Pr_{H_0}(T^2 \ge c_2) = \alpha \Rightarrow c_2 = \chi^2_{1-\alpha,p}$

Σ is unknown:

Under H_0, Σ is estimated by $\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)(X_i-\mu_0)^\top$

So.

$$\begin{aligned}\sup_{(\mu,\Sigma)\in\Theta_0} L(\mu,\Sigma) = L(\mu_0,\hat{\Sigma}_0) &= (2\pi)^{-np/2}(\det\hat{\Sigma}_0)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_0)^\top\hat{\Sigma}_0^{-1}(X_i-\mu_0)\right\}\\ &= (2\pi)^{-np/2}(\det\hat{\Sigma}_0)^{-n/2}\exp\left[-\frac{1}{2}tr\left\{\hat{\Sigma}_0^{-1}\sum_{i=1}^{n}(X_i-\mu_0)(X_i-\mu_0)^\top\right\}\right]\\ &= (2\pi)^{-np/2}(\det\hat{\Sigma}_0)^{-n/2}\exp\left\{-\frac{1}{2}tr(nI_p)\right\}\\ &= 2\pi^{-np/2}(\det\hat{\Sigma}_0)^{-n/2}\exp(-np/2)\end{aligned}$$

Similarly,

$$\begin{aligned}\sup_{(\mu,\Sigma)\in\Theta} L(\mu,\Sigma) = L(\bar{X},\hat{\Sigma}_{ML}) &= (2\pi)^{-np/2}(\det\hat{\Sigma}_{ML})^{-n/2}\exp\left[-\frac{1}{2}tr\left\{\hat{\Sigma}_{ML}^{-1}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^\top\right\}\right]\\ &= (2\pi)^{-np/2}(\det\hat{\Sigma}_{ML})^{-n/2}\exp(-np/2)\end{aligned}$$

$$\Rightarrow \lambda = \left(\frac{\det\hat{\Sigma}_0}{\det\hat{\Sigma}_{ML}}\right)^{-n/2} = \left(\frac{\det(n\hat{\Sigma}_0)}{\det(n\hat{\Sigma}_{ML})}\right)^{-n/2} = \left(\frac{\det\{n\hat{\Sigma}_{ML}+n(\bar{X}-\mu_0)(\bar{X}-\mu_0)^\top\}}{\det(n\hat{\Sigma}_{ML})}\right)^{-n/2} = \left\{1+n(\bar{X}-\mu_0)^\top(\overbrace{n\hat{\Sigma}_{ML}}^{(n-1)S})^{-1}(\bar{X}-\mu_0)\right\}^{-n/2} = \left(1+\frac{T^2}{n-1}\right)^{-n/2}$$

$(\because n\hat{\Sigma}_0 = n\hat{\Sigma}_{ML} + n(\bar{X}-\mu_0)(\bar{X}-\mu_0)^\top$, $\det(A+vv^\top) = \det(A)(1+v^\top A^{-1}v)$ & $T^2 = n(\bar{X}-\mu_0)^\top S^{-1}(\bar{X}-\mu_0)$ $)$

$\Rightarrow \mathcal{R} = \{\lambda \le c_1\} = \{T^2 \ge c_2\}$.

```r
dataset = as.matrix(gapminder[
  !is.na(gapminder$infant_mortality),
  c("infant_mortality", "life_expectancy", "fertility")])

(mu_hat <- colMeans(dataset))

# Test mu = mu_0
mu_0 <- c(25, 50, 3)
n = nrow(dataset)
p = ncol(dataset)
(test.stat <- drop(
  n * t(mu_hat - mu_0) %*% solve(cov(dataset)) %*% (mu_hat - mu_0)
))
(cri.point = (n-1)*p/(n-p)*qf(.95, p, n-p))
test.stat >= cri.point
(p.val = 1-pf((n-p)/(n-1)/p*test.stat, p, n-p))
```

- Report: Testing hypotheses $H_0 : \boldsymbol{\mu} = [25, 50, 3]^\top$ v.s. $H_1 : \boldsymbol{\mu} \neq [25, 50, 3]^\top$, we carried on the LRT and obtained 249718 as the value of test statistic. The corresponding p-value (resp. rejection region) was 0 (resp. $[7.819, \infty)$). So, at the .05 (significance) level, there was a strong statistical evidence implying the rejection of H_0, i.e., we believed that the mean vector is not $[25, 50, 3]^\top$.