Hotellingov T^2 test za dva uzorka

Primijenjena statistika

Doris Đivanović

Sadržaj

Teorijski opis testa

Zadatak a)

Opišite **test omjera vjerodostojnosti** za hipotezu o **jednakosti očekivanja** (TMS 8.9, Problem 1, str. 137 - 138) dva normalna uzorka.

TMS 8.9, Problem 1, str. 137 - 138:

Neka su $\mathbf{x_1}, \ldots, \mathbf{x_n}$, slučajni uzorak duljine n iz p-dimenzionalnog normalnog modela $N_p(\boldsymbol{\mu}, \Sigma)$ gdje su $\mathbf{x_i}$ nezavisni i jednako distribuirani, i $\mathbf{y_1}, \ldots, \mathbf{y_m}$, slučajni uzorak duljine m iz p-dimenzionalnog normalnog modela $N_p(\boldsymbol{\tau}, \Sigma)$ gdje su $\mathbf{y_i}$ nezavisni i jednako distribuirani, dva nezavisna uzorka.

Definiramo uzoračke varijance

$$S_x = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^\tau,$$
$$S_y = \frac{1}{(m-1)} \sum_{i=1}^{m} (y_i - \overline{y})(y_i - \overline{y})^\tau,$$

i

$$S_{pool} = \frac{1}{(n+m-2)} \left[(n-1)S_x + (m-1)S_y \right].$$

Odredite:

i) **distribuciju varijable** S_{pool},

ii) **distribuciju Hotellingove statistike**

$$T^2 = \left(\frac{1}{n} + \frac{1}{m} \right)^{-1} (\overline{x} - \overline{y})^\tau S_{pool}^{-1} (\overline{x} - \overline{y})$$

korištene za testiranje hipoteza

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\tau}$$
$$H_1 : \boldsymbol{\mu} \neq \boldsymbol{\tau}.$$

i) Distribucija varijable S_{pool}

S predavanja znamo da su

$$(n-1)S_x = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^\tau = G_x$$

i

$$(m-1)S_y = \sum_{i=1}^{m}(y_i - \overline{y})(y_i - \overline{y})^\tau = G_y$$

pogreškovne statistike, te da imaju Wishartovu distribuciju:

$$G_x \sim w_p(n-1, \Sigma), \qquad G_y \sim w_p(m-1, \Sigma).$$

Budući da je

$$\begin{aligned} S_{pool} &= \frac{1}{(n+m-2)}\left[(n-1)S_x + (m-1)S_y\right] \\ &= \frac{1}{(n+m-2)}\left(G_x + G_y\right), \end{aligned}$$

pokazat ćemo da i S_{pool} ima Wishartovu distribuciju.

Po definiciji Wishartove distribucije, postoje nezavisni normalni slučajni vektori $X_i \sim N_p(\mathbf{0}, \Sigma),\ i = 1, \ldots, n-1$ s istom kovarijacijskom matricom Σ takvi da vrijedi

$$G_x = \sum_{i=1}^{n-1} X_i X_i^\tau.$$

Analogno, postoje nezavisni normalni slučajni vektori $Y_i \sim N_p(\mathbf{0}, \Sigma),\ i = 1, \ldots, m-1$ s istom kovarijacijskom matricom Σ takvi da vrijedi:

$$G_y = \sum_{i=1}^{m-1} Y_i Y_i^\tau.$$

Stoga, definiranjem

$$X_{(n-1)+i} = Y_i, \quad i = 1, \ldots, m-1$$

dobivamo

$$G_x + G_y = \sum_{i=1}^{n+m-2} X_i X_i^\tau,$$

gdje su $X_i \sim N_p(\mathbf{0}, \Sigma),\ i = 1, \ldots, n+m-2$, nezavisni normalni slučajni vektori s istom kovarijacijskom matricom Σ.

Dakle, po definiciji Wishartove distribucije, vrijedi

$$(n+m-2)S_{pool} \sim w_p(n+m-2, \Sigma).$$

ii) Distribucija Hotellingove statistike T^2

Statistika

$$T^2 = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} (\overline{x} - \overline{y})^\tau S_{pool}^{-1} (\overline{x} - \overline{y})$$

koristi se za testiranje hipoteza

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\tau}$$
$$H_1 : \boldsymbol{\mu} \neq \boldsymbol{\tau}.$$

Zbog ovakve definicije statistike, **njezinu bismo distribuciju mogli odrediti korištenjem sljedeće tvrdnje dokazane na predavanju.**

Propozicija 1.21 Ako su $Z \sim N_p(\boldsymbol{\delta}, I)$ i $W \sim w_p(d)$ nezavisne slučajne veličine, i prirodni brojevi d i p su takvi da je $d \geq p$, tada je

$$\frac{d-p+1}{p} Z^\tau W^{-1} Z \sim F(p, d-p+1; \frac{1}{2}|\boldsymbol{\delta}|^2).$$

Sjetimo se da smo u ***i)*** odredili distribuciju slučajne varijable

$$(n+m-2)S_{pool} \sim w_p(n+m-2, \Sigma).$$

Stoga, da bismo na desnoj strani dobili inverz varijable

$$(n+m-2)S_{pool},$$

čiju (upravo Wishartovu) distribuciju poznajemo, jednakost

$$T^2 = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} (\overline{x} - \overline{y})^\tau S_{pool}^{-1} (\overline{x} - \overline{y})$$

zapisat ćemo u sljedećem ekvivalentnom obliku (tj. pomnožiti je s $(n+m-2)^{-1}$):

$$\frac{1}{n+m-2} T^2 = (n+m-2)^{-1} \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} (\overline{x} - \overline{y})^\tau S_{pool}^{-1} (\overline{x} - \overline{y}),$$

te na desnoj strani iskoristiti kvaziasocijativnost i asocijativnost matričnog množenja, i dobiti

$$\frac{1}{n+m-2} T^2 = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} (\overline{x} - \overline{y})^\tau [(n+m-2)S_{pool}]^{-1} (\overline{x} - \overline{y}).$$

Da bismo mogli primijeniti propoziciju, moramo moći pisati

$$\frac{1}{n+m-2}T^2 = \frac{d-p+1}{p}Z^\tau W^{-1}Z,$$

gdje je

$$W \sim w_p(d) \equiv w_p(d, I_p), \quad d \geq p.$$

Po definiciji Wishartove razdiobe, to bi vrijedilo ukoliko bi postojao niz nezavisnih normalnih slučajnih vektora $Y_i \sim N_p(\mathbf{0}, I_p),\ i = 1, \ldots, d$, s istom kovarijacijskom matricom I_p, takvih da vrijedi

$$W = \sum_{i=1}^{d} Y_i Y_i^\tau.$$

Naš "kandidat" za W trenutno je matrica

$$\tilde{W} := (n+m-2)S_{pool} \sim w_p(n+m-2, \Sigma).$$

Iz rasprave u dijelu ***i)***, sjetimo se da vrijedi

$$\tilde{W} = (n+m-2)S_{pool} = \sum_{i=1}^{n+m-2} X_i X_i^\tau,$$

gdje je

$$X_i \sim N_p(\mathbf{0}, \Sigma), \quad i = 1, \ldots, n+m-2$$

niz nezavisnih normalnih slučajnih vektora s istom kovarijacijskom matricom Σ.

Primijetimo sada da bi

$$Y_i := \Sigma^{-\frac{1}{2}} X_i \sim N_p(\mathbf{0}, I_p), \quad i = 1, \ldots, n+m-2,$$

bio niz nezavisnih normalnih slučajnih vektora s istom kovarijacijskom matricom I_p.

Dakle, bili bismo zadovoljni ukoliko bi bilo

$$W = \sum_{i=1}^{n+m-2} \left(\Sigma^{-\frac{1}{2}} X_i\right)\left(\Sigma^{-\frac{1}{2}} X_i\right)^\tau.$$

To možemo pisati na sljedeće načine:

$$\begin{aligned} W = \sum_{i=1}^{n+m-2} \left(\Sigma^{-\frac{1}{2}} X_i\right)\left(\Sigma^{-\frac{1}{2}} X_i\right)^\tau &= \sum_{i=1}^{n+m-2} \left(\Sigma^{-\frac{1}{2}} X_i\right)\left(X_i^\tau \Sigma^{-\frac{1}{2}}\right) \\ &= \sum_{i=1}^{n+m-2} \Sigma^{-\frac{1}{2}} \left(X_i X_i^\tau\right) \Sigma^{-\frac{1}{2}} \\ &= \Sigma^{-\frac{1}{2}} \left(\sum_{i=1}^{n+m-2} X_i X_i^\tau\right) \Sigma^{-\frac{1}{2}} \\ &= \boxed{\Sigma^{-\frac{1}{2}}(n+m-2)S_{pool}\Sigma^{-\frac{1}{2}} = W} \end{aligned}$$

gdje smo u prvoj jednakosti primijenili pravilo transponiranja umnoška te $\left(\Sigma^{-\frac{1}{2}}\right)^{\tau} = \Sigma^{-\frac{1}{2}}$, tj. simetričnost matrice $\Sigma^{-\frac{1}{2}}$, u drugoj asocijativnost množenja matrica, a u trećoj distributivnost množenja matrica prema zbrajanju, slijeva i zdesna.

Za ovakav W vrijedi

$$W \sim w_p(n+m-2).$$

Po pravilu invertiranja umnoška matrica, slijedi

$$\begin{aligned} W^{-1} &= \left(\Sigma^{-\frac{1}{2}}(n+m-2)S_{pool}\Sigma^{-\frac{1}{2}}\right)^{-1} \\ &= \Sigma^{\frac{1}{2}}[(n+m-2)S_{pool}]^{-1}\Sigma^{\frac{1}{2}}. \end{aligned}$$

Dakle, bili bismo zadovoljni ukoliko bismo statistiku

$$\frac{1}{n+m-2}T^2 = \left(\frac{1}{n}+\frac{1}{m}\right)^{-1}(\overline{x}-\overline{y})^{\tau}[(n+m-2)S_{pool}]^{-1}(\overline{x}-\overline{y})$$

mogli moći zapisati kao

$$\begin{aligned} \frac{1}{n+m-2}T^2 &= \frac{d-p+1}{p}Z^{\tau}W^{-1}Z \\ &= \frac{d-p+1}{p}Z^{\tau}\left(\Sigma^{\frac{1}{2}}[(n+m-2)S_{pool}]^{-1}\Sigma^{\frac{1}{2}}\right)Z. \end{aligned}$$

Budući da vrijedi

$$\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}} = \Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}} = I_p,$$

možemo pisati

$$\begin{aligned} \frac{1}{n+m-2}T^2 &= \left(\frac{1}{n}+\frac{1}{m}\right)^{-1}(\overline{x}-\overline{y})^{\tau}\left(\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\right)[(n+m-2)S_{pool}]^{-1}\left(\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right)(\overline{x}-\overline{y}) \\ &= \left(\frac{1}{n}+\frac{1}{m}\right)^{-1}(\overline{x}-\overline{y})^{\tau}\Sigma^{-\frac{1}{2}}\left(\Sigma^{\frac{1}{2}}(n+m-2)S_{pool}]^{-1}\Sigma^{\frac{1}{2}}\right)\Sigma^{-\frac{1}{2}}(\overline{x}-\overline{y}) \\ &= \left(\frac{1}{n}+\frac{1}{m}\right)^{-1}(\overline{x}-\overline{y})^{\tau}\Sigma^{-\frac{1}{2}}W^{-1}\Sigma^{-\frac{1}{2}}(\overline{x}-\overline{y}). \end{aligned}$$

Primijetimo sada, ako nismo i ranije, da vrijedi

$$\left(\frac{1}{n}+\frac{1}{m}\right)^{-1} = \frac{1}{\frac{1}{n}+\frac{1}{m}} = \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\cdot\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}},$$

pa, koristeći (kvazi)asocijativnost množenja, možemo pisati

$$\frac{1}{n+m-2}T^2 = \left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}(\overline{x}-\overline{y})^{\tau}\Sigma^{-\frac{1}{2}}\right)W^{-1}\left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}(\overline{x}-\overline{y})\right).$$

Ukoliko se sjetimo svojstava transponiranja matrica, i $\left(\Sigma^{-\frac{1}{2}}\right)^{\tau} = \Sigma^{-\frac{1}{2}}$, vrijedi

$$\frac{1}{n+m-2}T^2 = \left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}(\overline{x}-\overline{y})\right)^{\tau} W^{-1} \left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}(\overline{x}-\overline{y})\right).$$

Sada je očito da ćemo definirati

$$\boxed{Z := \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}(\overline{x}-\overline{y})}$$

pa imamo da vrijedi

$$\frac{1}{n+m-2}T^2 = Z^{\tau}W^{-1}Z, \qquad W \sim w_p(n+m-2).$$

Preostaje nam još odrediti distribuciju od Z.

Sjetimo se sada da je $\mathbf{x_1}, \ldots, \mathbf{x_n}$ iz $N_p(\boldsymbol{\mu}, \Sigma)$, a $\mathbf{y_1}, \ldots, \mathbf{y_m}$ iz $N_p(\boldsymbol{\tau}, \Sigma)$.

Model za taj uzorak je

$$\boldsymbol{x} = \mathbf{1}_n\boldsymbol{\mu}^{\tau} + E_x$$
$$\boldsymbol{y} = \mathbf{1}_m\boldsymbol{\tau}^{\tau} + E_y,$$

gdje su

$$\boldsymbol{x} \in M_{n,p} \quad \text{i} \quad \boldsymbol{y} \in M_{m,p},$$

matrice dizajna i matrice parametara su

$$\mathbf{1}_n \in M_{n,1} \quad \text{i} \quad \boldsymbol{\mu}^{\tau} \in M_{1,p},$$
$$\mathbf{1}_m \in M_{m,1} \quad \text{i} \quad \boldsymbol{\tau}^{\tau} \in M_{1,p},$$

a matrice pogrešaka

$$E_x \in M_{n,p} \quad \text{i} \quad E_y \in M_{m,p},$$

gdje su

$$\mathcal{E}_i \sim N_p(\mathbf{0}, \Sigma)$$

nezavisni slučajni vektori, pa slijedi

$$\text{vec}\,(E_x) \sim N_{np}(\mathbf{0}, \Sigma \otimes I_n) \quad \text{i} \quad \text{vec}\,(E_y) \sim N_{mp}(\mathbf{0}, \Sigma \otimes I_m).$$

Parametre modela, $\boldsymbol{\mu}$ i $\boldsymbol{\tau}$, procjenjujemo metodom najmanjih kvadrata:

$$\hat{\boldsymbol{\mu}}^\tau = (\mathbf{1}_n^\tau \mathbf{1}_n)^{-1}\mathbf{1}_n^\tau x = \frac{1}{n}\mathbf{1}_n^\tau x = \frac{1}{n}\mathbf{1}_n^\tau (\mathbf{1}_n\boldsymbol{\mu}^\tau + E_x) = \frac{1}{n}\cdot n\boldsymbol{\mu}^\tau + \frac{1}{n}\mathbf{1}_n^\tau E_x = \boldsymbol{\mu}^\tau + \frac{1}{n}\mathbf{1}_n^\tau E_x = \overline{x}^\tau,$$

i, analogno,

$$\hat{\boldsymbol{\tau}}^\tau = (\mathbf{1}_m^\tau \mathbf{1}_m)^{-1}\mathbf{1}_m^\tau y = \frac{1}{m}\mathbf{1}_m^\tau (\mathbf{1}_m\boldsymbol{\tau}^\tau + E_y) = \boldsymbol{\tau}^\tau + \frac{1}{m}\mathbf{1}_m^\tau E_y = \overline{y}^\tau.$$

Iz ovoga slijedi

$$\begin{aligned}\overline{x} &= \boldsymbol{\mu} + \frac{1}{n}E_x^\tau \mathbf{1}_n \\ \overline{y} &= \boldsymbol{\tau} + \frac{1}{m}E_y^\tau \mathbf{1}_m,\end{aligned}$$

pa imamo

$$\begin{aligned}\overline{x} &= \boldsymbol{\mu} + \mathcal{E}_x \\ \overline{y} &= \boldsymbol{\tau} + \mathcal{E}_y.\end{aligned}$$

Sada, Z možemo zapisati kao

$$\begin{aligned}Z &= \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}(\overline{x}-\overline{y}) \\ &= \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}\left[(\boldsymbol{\mu}+\mathcal{E}_x)-(\boldsymbol{\tau}+\mathcal{E}_y)\right] \\ &= \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}\left[(\mathcal{E}_x-\mathcal{E}_y)+(\boldsymbol{\mu}-\boldsymbol{\tau})\right].\end{aligned}$$

Pogledajmo pomaže li nam ovo u određivanju distribucije varijable Z.

Po pretpostavci zadatka, E_x i E_y su normalno distribuirani, pa vrijedi

$$\begin{aligned}\mathcal{E}_x &= \frac{1}{n}E_x^\tau \mathbf{1}_n \sim N_p\left(\mathbf{0}, \frac{1}{n}\Sigma\right) \\ \mathcal{E}_y &= \frac{1}{m}E_y^\tau \mathbf{1}_m \sim N_p\left(\mathbf{0}, \frac{1}{m}\Sigma\right).\end{aligned}$$

Nadalje, $\mathcal{E}_x$ i $-\mathcal{E}_y$ su nezavisni, $\Sigma > 0$ je regularna, pa vrijedi

$$\mathcal{E}_x - \mathcal{E}_y \sim N_p\left(\mathbf{0}, \left(\frac{1}{n}+\frac{1}{m}\right)\Sigma\right).$$

Iz ovoga slijedi

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\left[(\mathcal{E}_x-\mathcal{E}_y)+(\boldsymbol{\mu}-\boldsymbol{\tau})\right]\sim N_p\left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}(\boldsymbol{\mu}-\boldsymbol{\tau}),\Sigma\right).$$

Konačno, zaključujemo da za

$$Z=\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}\left[(\mathcal{E}_x-\mathcal{E}_y)+(\boldsymbol{\mu}-\boldsymbol{\tau})\right]$$

vrijedi

$$Z\sim N_p\left(\Sigma^{-\frac{1}{2}}\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}(\boldsymbol{\mu}-\boldsymbol{\tau}),I_p\right).$$

Izvedimo sada zaključak.

Dakle, statistiku

$$\frac{1}{n+m-2}T^2=\left(\frac{1}{n}+\frac{1}{m}\right)^{-1}(\overline{x}-\overline{y})^\tau[(n+m-2)S_{pool}]^{-1}(\overline{x}-\overline{y})$$

možemo zapisati kao

$$\frac{1}{n+m-2}T^2=Z^\tau W^{-1}Z,$$

gdje je

$$W=\Sigma^{-\frac{1}{2}}(n+m-2)S_{pool}\Sigma^{-\frac{1}{2}}\sim w_p(n+m-2)$$

i

$$Z=\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\Sigma^{-\frac{1}{2}}\left[(\mathcal{E}_x-\mathcal{E}_y)+(\boldsymbol{\mu}-\boldsymbol{\tau})\right]\sim N_p\left(\Sigma^{-\frac{1}{2}}\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}(\boldsymbol{\mu}-\boldsymbol{\tau}),I_p\right).$$

Stoga, za $d:=n+m-2\geq p$, primjenom propozicije 1.21 zaključujemo da vrijedi

$$\begin{aligned}\frac{(n+m-2)-p+1}{p}\frac{1}{n+m-2}T^2&=\frac{(n+m-2)-p+1}{p}Z^\tau W^{-1}Z\\&\sim F(p,(n+m-2)-p+1;\frac{1}{2}|\boldsymbol{\delta}|^2),\end{aligned}$$

odnosno

$$\begin{aligned}\frac{n+m-p-1}{p}\frac{1}{n+m-2}T^2&=\frac{n+m-p-1}{p}Z^\tau W^{-1}Z\\&\sim F(p,n+m-p-1;\frac{1}{2}|\boldsymbol{\delta}|^2),\end{aligned}$$

gdje je

$$\boldsymbol{\delta}=\Sigma^{-\frac{1}{2}}\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}(\boldsymbol{\mu}-\boldsymbol{\tau}).$$

Test omjera vjerodostojnosti

Promatrajmo najprije jedan $\mathbf{x_1}, \ldots, \mathbf{x_n}$ slučajni uzorak duljine n iz p-dimenzionalnog normalnog modela $N_p(\boldsymbol{\mu}, \Sigma)$ gdje su $\mathbf{x_i}$ nezavisni i jednako distribuirani, i test omjera vjerodostojnosti za testiranje hipoteza

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1,$$

gdje je $\{\Theta_0, \Theta_1\}$ particija parametarskog prostora

$$\Theta = \{(\boldsymbol{\mu}, \Sigma) : \ \boldsymbol{\mu} \in M_{p,1}, \ \Sigma \in M_{np}, \ \Sigma > 0\}.$$

Uzoračka varijanca je

$$S = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^{\tau},$$

a

$$G = (n-1)S$$

je pogreškovna statistika.

Sjetimo se s predavanja da je funkcija vjerodostojnosti vektora parametara $\theta = (\boldsymbol{\mu}, \Sigma)$

$$L(\boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{np}{2}} (\det\Sigma)^{-\frac{n}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(G\Sigma^{-1}\right) - \frac{n}{2}(\overline{x}-\boldsymbol{\mu})^{\tau}\Sigma^{-1}(\overline{x}-\boldsymbol{\mu})}.$$

Omjer vjerodostojnosti je

$$\Lambda = \frac{\max_{\theta \in \theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)}.$$

Za test omjera vjerodostojnosti p-vrijednost iznosi $\mathbb{P}(\Lambda \leq c)$.

Kako je funkcija vjerodostojnosti strogo pozitivna, a eksponencijalna funkcija strogo rastuća, procjenitelj maksimalne vjerodostojnosti $\hat{\theta} = (\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ za vektor parametara $\theta = (\boldsymbol{\mu}, \Sigma)$ našli smo traženjem maksimuma tzv. *log*-vjerodostojnosti:

$$\begin{aligned} l(\boldsymbol{\mu}, \Sigma) &= \log L(\boldsymbol{\mu}, \Sigma) \\ &= -\frac{np}{2}(2\pi) - \frac{n}{2}\log(\det\Sigma) - \frac{1}{2}\mathrm{tr}\left(G\Sigma^{-1}\right) - \frac{n}{2}\left(\overline{x} - \boldsymbol{\mu}\right)^{\tau} \Sigma^{-1} \left(\overline{x} - \boldsymbol{\mu}\right), \end{aligned}$$

što smo dobili primjenom svojstava logaritamske funkcije.

Vrijedi

$$\hat{\boldsymbol{\mu}} = \overline{x} \quad \text{i} \quad \hat{\Sigma} = \frac{1}{n} G,$$

te

$$L(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = (2\pi)^{-\frac{np}{2}} \left(\det\hat{\Sigma}\right)^{-\frac{n}{2}} e^{-\frac{np}{2}},$$

odnosno,

$$l\left(\hat{\boldsymbol{\mu}}, \hat{\Sigma}\right) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\left(\det\hat{\Sigma}\right) - \frac{np}{2}.$$

Vratimo se sada na zadatak, i **promatrajmo situaciju za dva nezavisna uzorka**.

Sjetimo se da je $\mathbf{x_1}, \ldots, \mathbf{x_n}$ iz $N_p(\boldsymbol{\mu}, \Sigma)$, a $\mathbf{y_1}, \ldots, \mathbf{y_m}$ iz $N_p(\boldsymbol{\tau}, \Sigma)$.

Tražimo omjer vjerodostojnosti za testiranje hipoteza

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\tau} \ (\theta \in \Theta_0)$$
$$H_1 : \boldsymbol{\mu} \neq \boldsymbol{\tau} \ (\theta \in \Theta_1),$$

gdje je $\{\Theta_0, \Theta_1\}$ particija parametarskog prostora

$$\Theta = \{(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) : \ \boldsymbol{\mu}, \boldsymbol{\tau} \in M_{p,1}, \ \Sigma \in M_q, \ \Sigma > 0\}.$$

Tražimo

$$\Lambda = \frac{\max_{\theta\in\theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)}.$$

Maksimizirajmo najprije funkciju L na Θ.

Neka je $\theta \in \Theta$ proizvoljan. Budući da su $\boldsymbol{x}$ i $\boldsymbol{y}$ **nezavisni uzorci**, vrijedi

$$L(\theta) = L(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) = L(\boldsymbol{\mu}, \Sigma) L(\boldsymbol{\tau}, \Sigma),$$

pa slijedi

$$\begin{aligned} L(\theta) = L(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) = (2\pi)^{-\frac{np}{2}} (\det\Sigma)^{-\frac{n}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(G_x\Sigma^{-1}\right) - \frac{n}{2}(\overline{x}-\boldsymbol{\mu})^\tau \Sigma^{-1}(\overline{x}-\boldsymbol{\mu})} \\ \cdot (2\pi)^{-\frac{mp}{2}} (\det\Sigma)^{-\frac{m}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(G_y\Sigma^{-1}\right) - \frac{m}{2}(\overline{y}-\boldsymbol{\tau})^\tau \Sigma^{-1}(\overline{y}-\boldsymbol{\tau})}. \end{aligned}$$

Sada, primjenom svojstava eksponencijalne funkcije, konkretno $a^{x+y} = a^x a^y$, dobivamo

$$L(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) = (2\pi)^{-\frac{(n+m)p}{2}} (\det\Sigma)^{-\frac{n+m}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(G_x\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(G_y\Sigma^{-1}\right) - \frac{n}{2}(\overline{x}-\boldsymbol{\mu})^\tau \Sigma^{-1}(\overline{x}-\boldsymbol{\mu}) - \frac{m}{2}(\overline{y}-\boldsymbol{\tau})^\tau \Sigma^{-1}(\overline{y}-\boldsymbol{\tau})}$$

Budući da je

$$(2\pi)^{-\frac{(n+m)p}{2}} > 0$$

i da ne ovisi o $(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma)$, dovoljno je maksimizirati funkciju

$$g(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) := (\det\Sigma)^{-\frac{n+m}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(G_x\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(G_y\Sigma^{-1}\right) - \frac{n}{2}(\overline{x}-\boldsymbol{\mu})^\tau \Sigma^{-1}(\overline{x}-\boldsymbol{\mu}) - \frac{m}{2}(\overline{y}-\boldsymbol{\tau})^\tau \Sigma^{-1}(\overline{y}-\boldsymbol{\tau})}.$$

Nadalje, budući da je eksponencijalna funkcija strogo rastuća, a i g je strogo pozitivna funkcija, dovoljno je maksimizirati funkciju

$$h(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) := \log(g(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma)),$$

odnosno,

$$\begin{aligned} h(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) &= -\frac{n+m}{2}\log\left(\det\Sigma\right) - \frac{1}{2}\mathrm{tr}\left(G_x\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(G_y\Sigma^{-1}\right) \\ &\quad - \frac{n}{2}\left(\overline{x} - \boldsymbol{\mu}\right)^{\tau}\Sigma^{-1}\left(\overline{x} - \boldsymbol{\mu}\right) - \frac{m}{2}\left(\overline{y} - \boldsymbol{\tau}\right)^{\tau}\Sigma^{-1}\left(\overline{y} - \boldsymbol{\tau}\right) \\ &= -\frac{n+m}{2}\log\left(\det\Sigma\right) - \frac{1}{2}\left[\mathrm{tr}\left(G_x\Sigma^{-1}\right) + \mathrm{tr}\left(G_y\Sigma^{-1}\right)\right] \\ &\quad - \frac{n}{2}\left(\overline{x} - \boldsymbol{\mu}\right)^{\tau}\Sigma^{-1}\left(\overline{x} - \boldsymbol{\mu}\right) - \frac{m}{2}\left(\overline{y} - \boldsymbol{\tau}\right)^{\tau}\Sigma^{-1}\left(\overline{y} - \boldsymbol{\tau}\right), \end{aligned}$$

što smo dobili primjenom svojstava logaritamske funkcije.

Sada, primijetimo da je dovoljno minimizirati funkciju

$$k(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) := -2h(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma),$$

tj. funkciju

$$\begin{aligned} k(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) &= (n+m)\log\left(\det\Sigma\right) + \mathrm{tr}\left(G_x\Sigma^{-1}\right) + \mathrm{tr}\left(G_y\Sigma^{-1}\right) \\ &\quad + n\left(\overline{x} - \boldsymbol{\mu}\right)^{\tau}\Sigma^{-1}\left(\overline{x} - \boldsymbol{\mu}\right) + m\left(\overline{y} - \boldsymbol{\tau}\right)^{\tau}\Sigma^{-1}\left(\overline{y} - \boldsymbol{\tau}\right). \end{aligned}$$

Sjetimo se da je trag linearni funkcional, pa je, specijalno, aditivan, pa možemo pisati

$$\mathrm{tr}\left(G_x\Sigma^{-1}\right) + \mathrm{tr}\left(G_y\Sigma^{-1}\right) = \mathrm{tr}\left(G_x\Sigma^{-1} + G_y\Sigma^{-1}\right) = \mathrm{tr}\left[\left(G_x + G_y\right)\Sigma^{-1}\right],$$

gdje smo u drugoj jednakosti primijenili distributivnost množenja matrica prema zbrajanju zdesna.

Dakle, možemo pisati

$$\begin{aligned} k(\boldsymbol{\mu}, \boldsymbol{\tau}, \Sigma) &= (n+m)\log\left(\det\Sigma\right) + \mathrm{tr}\left[\left(G_x + G_y\right)\Sigma^{-1}\right] \\ &\quad + n\left(\overline{x} - \boldsymbol{\mu}\right)^{\tau}\Sigma^{-1}\left(\overline{x} - \boldsymbol{\mu}\right) + m\left(\overline{y} - \boldsymbol{\tau}\right)^{\tau}\Sigma^{-1}\left(\overline{y} - \boldsymbol{\tau}\right). \end{aligned}$$

Budući da je $\Sigma^{-1} > 0$, vrijedi

$$n\left(\overline{x} - \boldsymbol{\mu}\right)^{\tau}\Sigma^{-1}\left(\overline{x} - \boldsymbol{\mu}\right) \geq 0,$$

pa se minimum te funkcije ($\min = 0$) postiže za

$$\overline{x} - \hat{\boldsymbol{\mu}} = 0,$$

odnosno za

$$\hat{\boldsymbol{\mu}} = \overline{x}.$$

Analogno, minimum funkcije

$$m\left(\overline{y} - \boldsymbol{\tau}\right)^{\tau} \Sigma^{-1} \left(\overline{y} - \boldsymbol{\tau}\right)$$

postiže se za

$$\hat{\boldsymbol{\tau}} = \overline{y}.$$

Konačno, minimum funkcije

$$(n+m)\log\left(\det\Sigma\right) + \operatorname{tr}\left[\left(G_x + G_y\right)\Sigma^{-1}\right]$$

postiže se za isti $\hat{\Sigma}$ kao i minimum funkcije

$$\log\left(\det\Sigma\right) + \frac{1}{n+m}\operatorname{tr}\left[\left(G_x + G_y\right)\Sigma^{-1}\right],$$

što je, koristeći homogenost traga,

$$\log\left(\det\Sigma\right) + \operatorname{tr}\left[\frac{1}{n+m}\left(G_x + G_y\right)\Sigma^{-1}\right],$$

pa je

$$\hat{\Sigma} = \frac{1}{n+m}\left(G_x + G_y\right).$$

Dakle, odredili smo procjenitelj maksimalne vjerodostojnosti (MLE), $\hat{\theta} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\tau}}, \hat{\Sigma})$. Primijetimo da je rezultat analogan onome u slučaju jednog uzorka.

Vrijedi

$$\boxed{\max_{\theta\in\Theta} L(\theta) = L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\tau}}, \hat{\Sigma}) = L(\overline{x}, \overline{y}, \hat{\Sigma}) = (2\pi)^{-\frac{(n+m)p}{2}} \left(\det\hat{\Sigma}\right)^{-\frac{n+m}{2}} e^{-\frac{(n+m)p}{2}}}$$

Maksimizirajmo sada funkciju L na Θ_0.

Uz pretpostavku da vrijedi hipoteza H_0, neka je $\theta \in \Theta_0$. Budući da su $\boldsymbol{x}$ i $\boldsymbol{y}$ **nezavisni uzorci**, vrijedi

$$L(\theta) = L(\boldsymbol{\mu}, \boldsymbol{\mu}, \Sigma_0) = L(\boldsymbol{\mu}, \Sigma_0) L(\boldsymbol{\mu}, \Sigma_0),$$

pa slijedi

$$\begin{aligned} L(\theta) = L(\boldsymbol{\mu}, \boldsymbol{\mu}, \Sigma_0) = {} & (2\pi)^{-\frac{np}{2}} \left(\det\Sigma_0\right)^{-\frac{n}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(G_x\Sigma_0^{-1}\right) - \frac{n}{2}(\overline{x}-\boldsymbol{\mu})^{\tau}\Sigma_0^{-1}(\overline{x}-\boldsymbol{\mu})} \\ & \cdot (2\pi)^{-\frac{mp}{2}} \left(\det\Sigma_0\right)^{-\frac{m}{2}} e^{-\frac{1}{2}\operatorname{tr}\left(G_y\Sigma_0^{-1}\right) - \frac{m}{2}(\overline{y}-\boldsymbol{\mu})^{\mu}\Sigma_0^{-1}(\overline{y}-\boldsymbol{\mu})}, \end{aligned}$$

odnosno

$$L(\boldsymbol{\mu},\boldsymbol{\mu},\Sigma_0)=(2\pi)^{-\frac{(n+m)p}{2}}\left(\det\Sigma_0\right)^{-\frac{n+m}{2}}e^{-\frac{1}{2}\mathrm{tr}\left(G_x\Sigma_0^{-1}\right)-\frac{n}{2}(\overline{x}-\boldsymbol{\mu})^\tau\Sigma_0^{-1}(\overline{x}-\boldsymbol{\mu})-\frac{1}{2}\mathrm{tr}\left(G_y\Sigma_0^{-1}\right)-\frac{m}{2}(\overline{y}-\boldsymbol{\mu})^\mu\Sigma_0^{-1}(\overline{y}-\boldsymbol{\mu})}.$$

Potpuno analogno kao maloprije, zaključujemo da je dovoljno minimizirati funkciju

$$\begin{aligned}k(\boldsymbol{\mu},\boldsymbol{\mu},\Sigma_0):=&(n+m)\log\left(\det\Sigma_0\right)+\mathrm{tr}\left[\left(G_x+G_y\right)\Sigma_0^{-1}\right]\\&+n\left(\overline{x}-\boldsymbol{\mu}\right)^\tau\Sigma_0^{-1}\left(\overline{x}-\boldsymbol{\mu}\right)+m\left(\overline{y}-\boldsymbol{\mu}\right)^\mu\Sigma_0^{-1}\left(\overline{y}-\boldsymbol{\mu}\right).\end{aligned}$$

Budući da je $\Sigma_0^{-1}>0$, vrijedi

$$n\left(\overline{x}-\boldsymbol{\mu}\right)^\tau\Sigma_0^{-1}\left(\overline{x}-\boldsymbol{\mu}\right)+m\left(\overline{y}-\boldsymbol{\mu}\right)^\tau\Sigma_0^{-1}\left(\overline{y}-\boldsymbol{\mu}\right)\geq 0,$$

a minimum te funkcije kao funkcije u $\boldsymbol{\mu}$ postiže se u njenoj stacionarnoj točki

$$\hat{\boldsymbol{\mu}}=\frac{n\overline{x}+m\overline{y}}{n+m}$$

i iznosi

$$n\left(\overline{x}-\hat{\boldsymbol{\mu}}\right)^\tau\Sigma_0^{-1}\left(\overline{x}-\hat{\boldsymbol{\mu}}\right)+m\left(\overline{y}-\hat{\boldsymbol{\mu}}\right)^\tau\Sigma_0^{-1}\left(\overline{y}-\hat{\boldsymbol{\mu}}\right)=\frac{nm}{n+m}\left(\overline{x}-\overline{y}\right)^\tau\Sigma_0^{-1}\left(\overline{x}-\overline{y}\right).$$

Sada je dovoljno minimizirati funkciju

$$k(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\mu}},\Sigma_0)=(n+m)\log\left(\det\Sigma_0\right)+\mathrm{tr}\left[\left(G_x+G_y\right)\Sigma_0^{-1}\right]+\frac{nm}{n+m}\left(\overline{x}-\overline{y}\right)^\tau\Sigma_0^{-1}\left(\overline{x}-\overline{y}\right).$$

Primjenom aditivnosti traga, a zatim distributivnosti množenja matrica prema zbrajanju zdesna, možemo pisati

$$k(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\mu}},\Sigma_0)=(n+m)\log\left(\det\Sigma_0\right)+\mathrm{tr}\left[\left(G_x+G_y+\frac{nm}{n+m}\left(\overline{x}-\overline{y}\right)^\tau\left(\overline{x}-\overline{y}\right)\right)\Sigma_0^{-1}\right].$$

Iz ovog zapisa, analogno kao maloprije, zaključujemo da je

$$\hat{\Sigma}_0=\frac{1}{n+m}\left[G_x+G_y+\frac{nm}{n+m}\left(\overline{x}-\overline{y}\right)^\tau\left(\overline{x}-\overline{y}\right)\right].$$

Dakle, odredili smo procjenitelj maksimalne vjerodostojnosti (MLE), $\hat{\theta}=(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\mu}},\hat{\Sigma}_0)$, kada je $\theta\in\Theta_0$.

Vrijedi

$$\boxed{\max_{\theta\in\Theta_0}L(\theta)=L(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\mu}},\hat{\Sigma}_0)=(2\pi)^{-\frac{(n+m)p}{2}}\left(\det\hat{\Sigma}_0\right)^{-\frac{n+m}{2}}e^{-\frac{(n+m)p}{2}}}$$

Konačno, izračunajmo omjer vjerodostojnosti.

Imamo

$$\Lambda = \frac{\max_{\theta\in\theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)} = \frac{(2\pi)^{-\frac{(n+m)p}{2}} \left(\det\hat{\Sigma}_0\right)^{-\frac{n+m}{2}} e^{-\frac{(n+m)p}{2}}}{(2\pi)^{-\frac{(n+m)p}{2}} \left(\det\hat{\Sigma}\right)^{-\frac{n+m}{2}} e^{-\frac{(n+m)p}{2}}} = \left(\frac{\det\hat{\Sigma}_0}{\det\hat{\Sigma}}\right)^{-\frac{n+m}{2}}.$$

Primijetimo da je ovaj rezultat analogan onome u slučaju jednog uzorka.

Sada, uvrštavanjem

$$\hat{\Sigma}_0 = \frac{1}{n+m}\left[G_x + G_y + \frac{nm}{n+m}(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y})\right] \quad \text{i} \quad \hat{\Sigma} = \frac{1}{n+m}(G_x+G_y),$$

dobivamo

$$\Lambda = \left(\frac{\det\left[\frac{1}{n+m}\left(G_x + G_y + \frac{nm}{n+m}(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y})\right)\right]}{\det\left[\frac{1}{n+m}(G_x+G_y)\right]}\right)^{-\frac{n+m}{2}}.$$

Budući da je determinanta linearni funkcional, možemo pisati

$$\Lambda = \left(\frac{\frac{1}{n+m}\det(G_x+G_y) + \frac{1}{n+m}\det\left[\frac{nm}{n+m}(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y})\right]}{\frac{1}{n+m}\det(G_x+G_y)}\right)^{-\frac{n+m}{2}},$$

tj.

$$\Lambda = \left(1 + \frac{\det\left[\frac{nm}{n+m}(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y})\right]}{\det(G_x+G_y)}\right)^{-\frac{n+m}{2}}.$$

Sada bismo mogli pokušati, kao na predavanju, pokazati da je Λ funkcija Hotellingove statistike T^2 koja se koristi za testiranje naših hipoteza.

Sjetimo se da je

$$\begin{aligned} T^2 &= \left(\frac{1}{n}+\frac{1}{m}\right)^{-1}(\overline{x}-\overline{y})^\tau S_{pool}^{-1}(\overline{x}-\overline{y}) \\ &= \left(\frac{n+m}{nm}\right)^{-1}(\overline{x}-\overline{y})^\tau S_{pool}^{-1}(\overline{x}-\overline{y}) \\ &= \frac{nm}{n+m}(\overline{x}-\overline{y})^\tau S_{pool}^{-1}(\overline{x}-\overline{y}). \end{aligned}$$

Iz dijela **i)**, sjetimo se da je

$$S_{pool} = \frac{1}{(n+m-2)}(G_x+G_y),$$

pa je

$$S_{pool}^{-1} = (n+m-2)(G_x+G_y)^{-1}.$$

Dakle, imamo

$$T^2 = \frac{nm}{n+m}(\overline{x}-\overline{y})^\tau (n+m-2)(G_x+G_y)^{-1}(\overline{x}-\overline{y}),$$

što bismo vjerojatno mogli dobiti u izrazu za Λ.

Ukoliko se sjetimo da je

$$\frac{1}{\det(G_x+G_y)} = \det\left[(G_x+G_y)^{-1}\right],$$

te

$$1 = \det I,$$

možemo pisati

$$\begin{aligned}
\Lambda &= \left(1 + \frac{\det\left[\frac{nm}{n+m}(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y})\right]}{\det(G_x+G_y)}\right)^{-\frac{n+m}{2}} \\
&= \left(\det I + \det\left[(G_x+G_y)^{-1}\right]\det\left[\frac{nm}{n+m}(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y})\right]\right)^{-\frac{n+m}{2}} \\
&= \left(\det I + \det\left[(G_x+G_y)^{-1}\frac{nm}{n+m}(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y})\right]\right)^{-\frac{n+m}{2}}.
\end{aligned}$$

Sjetimo se sada da vrijedi

$$(\overline{x}-\overline{y})^\tau(\overline{x}-\overline{y}) = (\overline{x}-\overline{y})(\overline{x}-\overline{y})^\tau,$$

pa možemo pisati

$$\begin{aligned}
\Lambda &= \left(\det I + \det\left[(G_x+G_y)^{-1}\frac{nm}{n+m}(\overline{x}-\overline{y})(\overline{x}-\overline{y})^\tau\right]\right)^{-\frac{n+m}{2}} \\
&= \left(\det I + \det\left[(G_x+G_y)^{-1}(\overline{x}-\overline{y})\frac{nm}{n+m}(\overline{x}-\overline{y})^\tau\right]\right)^{-\frac{n+m}{2}} \\
&= \left(\det\left[I + (G_x+G_y)^{-1}(\overline{x}-\overline{y})\frac{nm}{n+m}(\overline{x}-\overline{y})^\tau\right]\right)^{-\frac{n+m}{2}}.
\end{aligned}$$

Sjetimo se sada sljedeće leme s predavanja.

Lema 1.16 Ako su $A \in M_{p,q}$ i $B \in M_{q,p}$, tada vrijedi

$$\det(I_p + AB) = \det(I_q + BA).$$

Dakle, ukoliko označino

$$A = (G_x + G_y)^{-1}(\overline{x} - \overline{y}) \quad \text{i} \quad B = \frac{nm}{n+m}(\overline{x} - \overline{y})^\tau,$$

vrijedi

$$\Lambda = \left(\det\left[I + \frac{nm}{n+m}(\overline{x} - \overline{y})^\tau (G_x + G_y)^{-1}(\overline{x} - \overline{y})\right]\right)^{-\frac{n+m}{2}}$$

Sada, budući da je

$$T^2 = \frac{nm}{n+m}(\overline{x} - \overline{y})^\tau (n+m-2)(G_x + G_y)^{-1}(\overline{x} - \overline{y}),$$

imamo

$$\begin{aligned}\Lambda &= \left(\det\left[I + \frac{1}{n+m-2}T^2\right]\right)^{-\frac{n+m}{2}} \\ &= \left(\det I + \det\frac{1}{n+m-2}T^2\right)^{-\frac{n+m}{2}} \\ &= \left(1 + \frac{1}{n+m-2}T^2\right)^{-\frac{n+m}{2}}.\end{aligned}$$

Dakle, omjer vjerodostojnosti je, očito strogo padajuća, funkcija statistike T^2 koja se koristi za testiranje hipoteza:

$$\boxed{\Lambda = \left(1 + \frac{1}{n+m-2}T^2\right)^{-\frac{n+m}{2}} =: f\left(T^2\right)}$$

Konačno, **izračunajmo *p*-vrijednost za test omjera vjerodostojnosti**.

Sjetimo se najprije distribucije testne statistike koju smo odredili u dijelu **ii)**:

$$\begin{aligned}\frac{n+m-p-1}{p}\frac{1}{n+m-2}T^2 &= \frac{n+m-p-1}{p}Z^\tau W^{-1}Z \\ &\sim F(p,\ n+m-p-1;\ \frac{1}{2}|\boldsymbol{\delta}|^2),\end{aligned}$$

gdje je

$$\boldsymbol{\delta} = \Sigma^{-\frac{1}{2}} \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} (\boldsymbol{\mu} - \boldsymbol{\tau}).$$

Uz pretpostavku da je istinita nulta hipoteza

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\tau},$$

očito vrijedi

$$\boldsymbol{\delta} = \mathbf{0},$$

odnosno vrijedi

$$\begin{aligned} \frac{n+m-p-1}{p} \frac{1}{n+m-2} T^2 &\sim F(p,\ qn+m-p-1;\ \mathbf{0}) \\ &\equiv F(p,\ n+m-p-1). \end{aligned}$$

Sada, budući da za test omjera vjerodostojnosti p-vrijednost iznosi $\mathbb{P}(\Lambda \leq c)$, računamo

$$\begin{aligned} \mathbb{P}\left(\Lambda \leq c\right) &= \mathbb{P}\left(f\left(T^2\right) \leq c\right) \\ &= \mathbb{P}\left(T^2 \geq f^{-1}(c)\right) \\ &= 1 - \mathbb{P}\left(T^2 \leq f^{-1}(c)\right) \\ &= 1 - \mathbb{P}\left(\frac{n+m-p-1}{(n+m-2)p} T^2 \leq \frac{n+m-p-1}{(n+m-2)p} f^{-1}(c)\right), \end{aligned}$$

gdje smo u drugoj jednakosti iskoristili činjenicu da je f^{-1} strogo padajuća funkcija, budući da je funkcija f strogo padajuća.

Provedba testa na konkretnim podacima iz prakse

ALM, Exercise 1.8.3, str. 68. i 70.

Dani su podaci o težinama za tri grupe štakora. Jednoj grupi u vodu za piće stavljan je tyroxin, drugoj je grupi stavljan thiouracil, a treća je grupa bila kontrolna. Težine su mjerene u gramima u tjednim intervalima. Podaci su dani u sljedećim tablicama.

Time 0	Time 1	Time 2	Time 3	Time 4
59	85	121	156	191
54	71	90	110	138
56	75	108	151	189
59	85	116	148	177
57	72	97	120	144
52	73	97	116	140
52	70	105	138	171

Tablica 1: Grupa Thyroxin

Time 0	Time 1	Time 2	Time 3	Time 4
61	86	109	120	129
59	80	101	111	122
53	79	100	106	133
59	88	100	111	122
51	75	101	123	140
51	75	92	100	119
56	78	95	103	108
58	69	93	114	138
46	61	78	90	107
53	72	89	104	122

Tablica 2: Grupa Thiouracil

Ove sam podatke spremila u Microsoft Excel dokument, a zatim ih iz njega učitavala u R-u.

Slijedi pripadni R-kod.

```
#postavljam se u radni direktorij gdje se nalazi
#dokument s podacima
> setwd("C:/Users/Djivo1/Desktop/APS")

> library(readxl)

#ucitavam podatke za skupinu koja je uzimala thyroxin
#iz MS Excel dokumenta
> thyroxin <- read_xlsx("Podaci.xlsx", col_names = TRUE, range = "A2:E9")

#pretvaram u matricu
> thyroxin = as.matrix(thyroxin)

#ispis radi provjere
> thyroxin
     Time 0 Time 1 Time 2 Time 3 Time 4
[1,]     59     85    121    156    191
[2,]     54     71     90    110    138
[3,]     56     75    108    151    189
[4,]     59     85    116    148    177
[5,]     57     72     97    120    144
[6,]     52     73     97    116    140
[7,]     52     70    105    138    171

#ucitavam podatke za skupinu koja je uzimala thiouracil
#iz MS Excel dokumenta
> thiouracil <- read_xlsx("Podaci.xlsx", col_names = TRUE, range = "G2:K12")

#pretvaram u matricu
> thiouracil = as.matrix(thiouracil)

#ispis radi provjere
> thiouracil
      Time 0 Time 1 Time 2 Time 3 Time 4
 [1,]     61     86    109    120    129
 [2,]     59     80    101    111    122
 [3,]     53     79    100    106    133
 [4,]     59     88    100    111    122
 [5,]     51     75    101    123    140
 [6,]     51     75     92    100    119
 [7,]     56     78     95    103    108
 [8,]     58     69     93    114    138
 [9,]     46     61     78     90    107
[10,]     53     72     89    104    122
```

Zadatak b)

Ispitajte normalnost podataka.

Iz pretpostavke i izlaganja u zadatku a), znamo da test omjera vjerodostojnosti za testiranje hipoteze o jednakosti očekivanja dva nezavisna uzorka možemo provesti isključivo ako uzorci dolaze iz **normalne** razdiobe.

Stoga, ukoliko opisani test želimo provesti na ovim konkretnim podacima, moramo se najprije uvjeriti da oni dolaze iz normalne razdiobe, dakle provesti neki test za testiranje nulte hipoteze da podaci dolaze iz normalne razdiobe nasuprot alternativnoj hipotezi da ne dolaze.

Ovdje ćemo najprije provesti jedan grafički test, a zatim jedan takav statistički test.

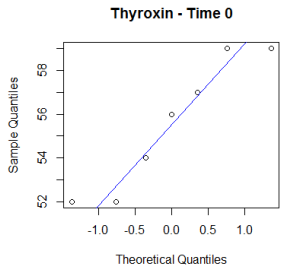
Normalnost podataka - grafički test

Grafički test normalnosti podataka provest ćemo crtanjem normalnih vjerojatnosnih grafova svih komponenti uzoraka.

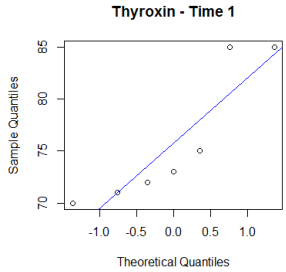
Slijedi kod u R-u za uzorak "Thyroxin".

```
#normalni vjerojatnosni grafovi za svaki stupac uzorka "Thyroxin"

> qqnorm(thyroxin[,1], main = "Thyroxin - Time 0")
> qqline(thyroxin[,1], col = "blue")

> qqnorm(thyroxin[,2], main = "Thyroxin - Time 1")
> qqline(thyroxin[,2], col = "blue")

> qqnorm(thyroxin[,3], main = "Thyroxin - Time 2")
> qqline(thyroxin[,3], col = "blue")

> qqnorm(thyroxin[,4], main = "Thyroxin - Time 3")
> qqline(thyroxin[,4], col = "blue")

> qqnorm(thyroxin[,5], main = "Thyroxin - Time 4")
> qqline(thyroxin[,5], col = "blue")
```

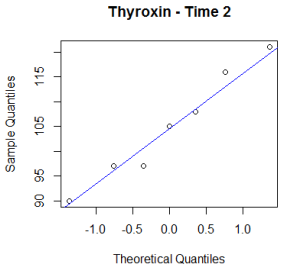
Slijede dobiveni rezultati.



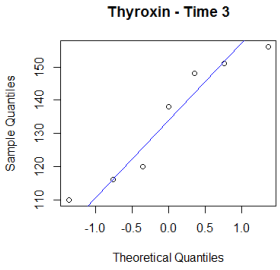
(a) Time 0



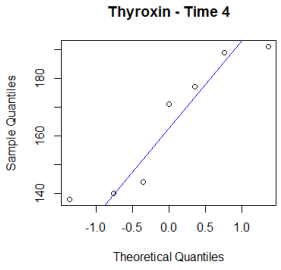
(b) Time 1



(c) Time 2



(d) Time 3



(e) Time 4

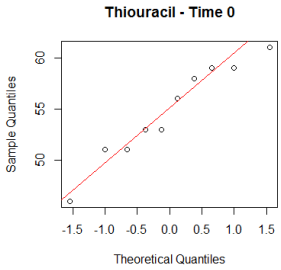
Slika 1: Normalni vjerojatnosni grafovi za stupce uzorka Thyroxin

Uočavamo da se za svaki stupac uzorka "Thyroxin" podaci dobro grupiraju oko pravca na normalnom vjerojatnosnom grafu, pa **nećemo zaključiti da ti podaci ne dolaze iz normalne distribucije**.

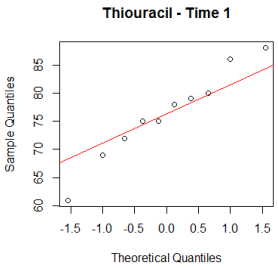
Analogno, slijedi kod u R-u za uzorak "Thiouracil".

```
#normalni vjerojatnosni grafovi za svaki stupac uzorka "Thiouracil"

> qqnorm(thiouracil[,1], main = "Thiouracil - Time 0")
> qqline(thiouracil[,1], col = "red")

> qqnorm(thiouracil[,2], main = "Thiouracil - Time 1")
> qqline(thiouracil[,2], col = "red")

> qqnorm(thiouracil[,3], main = "Thiouracil - Time 2")
> qqline(thiouracil[,3], col = "red")

> qqnorm(thiouracil[,4], main = "Thiouracil - Time 3")
> qqline(thiouracil[,4], col = "red")

> qqnorm(thiouracil[,5], main = "Thiouracil - Time 4")
> qqline(thiouracil[,5], col = "red")
```

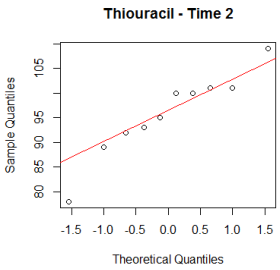
Slijede dobiveni rezultati.



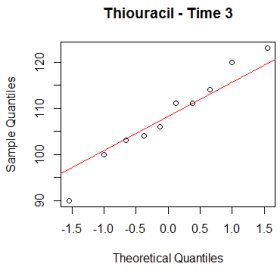
(a) Time 0



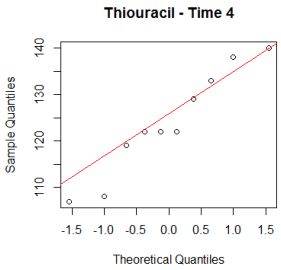
(b) Time 1



(c) Time 2



(d) Time 3



(e) Time 4

Slika 2: Normalni vjerojatnosni grafovi za stupce uzorka Thiouracil

Uočavamo da se za svaki stupac uzorka "Thiouracil" podaci dobro grupiraju oko pravca na normalnom vjerojatnosnom grafu, pa **nećemo zaključiti da ti podaci ne dolaze iz normalne distribucije**.

Normalnost podataka - statistički test

Statistički test normalnosti podataka koji ćemo provesti jest Lillieforsov test. Taj je test posebna varijanta (za normalnu distribuciju) Kolmogorov-Smirnovljeva testa za testiranje hipoteze da podaci dolaze iz neke distribucije.

Slijedi kod u R-u i dobiveni rezultati za uzorak "Thyroxin".

```
> library(nortest)

#provodim Lillieforsov test za svaki stupac uzorka "Thyroxin"

> lillie.test(thyroxin[,1])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thyroxin[, 1]
D = 0.1694, p-value = 0.7868

> lillie.test(thyroxin[,2])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thyroxin[, 2]
D = 0.26723, p-value = 0.1363

> lillie.test(thyroxin[,3])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thyroxin[, 3]
D = 0.18911, p-value = 0.6303

> lillie.test(thyroxin[,4])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thyroxin[, 4]
D = 0.20478, p-value = 0.5021

> lillie.test(thyroxin[,5])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thyroxin[, 5]
D = 0.23784, p-value = 0.2696
```

Uočavamo da su za svaki stupac uzorka "Thyroxin" p-vrijednosti testa normalnosti velike, pa **ne možemo odbaciti nultu hipotezu da ti podaci dolaze iz normalne distribucije**.

Sada, analogno, slijedi kod u R-u i dobiveni rezultati za uzorak "Thiouracil".

```
#provodim Lillieforsov test za svaki stupac uzorka "Thiouracil"

> lillie.test(thiouracil[,1])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thiouracil[, 1]
D = 0.15909, p-value = 0.6719

> lillie.test(thiouracil[,2])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thiouracil[, 2]
D = 0.13478, p-value = 0.8732

> lillie.test(thiouracil[,3])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thiouracil[, 3]
D = 0.18948, p-value = 0.3916

> lillie.test(thiouracil[,4])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thiouracil[, 4]
D = 0.11301, p-value = 0.9735

> lillie.test(thiouracil[,5])

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  thiouracil[, 5]
D = 0.17052, p-value = 0.5629
```

Uočavamo da su za svaki stupac uzorka "Thiouracil" p-vrijednosti testa normalnosti velike, pa **ne možemo odbaciti nultu hipotezu da ti podaci dolaze iz normalne distribucije**.

Zadatak c)

Sprovedite test iz a) na usporedbu skupina "Thyroxin" i "Thiouracil".

Budući da smo u zadatku b) pokazali da uzorci "Thyroxin" i "Thiouracil" dolaze iz normalne razdiobe, uistinu možemo provesti test. Provest ćemo ga točno onako kako smo opisali i zaključili u zadatku a).

Ako podatke za skupinu koja je dobivala thyroxin označimo s $\boldsymbol{x}$, u skladu s oznakama iz zadatka a), zaključujemo da je $n = 7$, a da je dimnenzija $p = 5$. Ako podatke za skupinu koja je dobivala thiouracil označimo s $\boldsymbol{y}$, u skladu s oznakama iz zadatka a), zaključujemo da je $m = 10$, a da je dimnenzija $p = 5$.

Najprije ćemo izračunati aritmetičke sredine uzoraka, $\overline{x}$ i $\overline{y}$, kao vektore aritmetičkih sredina svih stupaca od $\boldsymbol{x}$ i $\boldsymbol{y}$, da bismo mogli izračunati uzoračke varijance:

$$S_x = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^\tau$$
$$S_y = \frac{1}{(m-1)} \sum_{i=1}^{m} (y_i - \overline{y})(y_i - \overline{y})^\tau$$

Zatim računamo

$$S_{pool} = \frac{1}{(n+m-2)} \left[(n-1)S_x + (m-1)S_y \right].$$

Nadalje, računamo Hotellingovu statistiku

$$T^2 = \left(\frac{1}{n} + \frac{1}{m} \right)^{-1} (\overline{x} - \overline{y})^\tau S_{pool}^{-1} (\overline{x} - \overline{y}).$$

Iz T^2 računamo omjer vjerodostojnosti

$$\Lambda = \left(1 + \frac{1}{n+m-2} T^2 \right)^{-\frac{n+m}{2}} =: f\left(T^2\right),$$

te, prema zaključku iz zadatka a), testnu statistiku

$$\frac{n+m-p-1}{(n+m-2)p} T^2.$$

Konačno, računamo p-vrijednost

$$1 - \mathbb{P}\left(\frac{n+m-p-1}{(n+m-2)p} T^2 \leq \frac{n+m-p-1}{(n+m-2)p} f^{-1}(c) \right).$$

Slijedi pripadni kod u R-u.

```
 # unosim konkretne duljine uzoraka i dimenziju modela
> n = 7
> m = 10
> p = 5

#racunam vektore aritmetickih sredina stupaca uzoraka
> x_mean = c(mean(thyroxin[,1]), mean(thyroxin[,2]), mean(thyroxin[,3]),
    mean(thyroxin[,4]), mean(thyroxin[,5]))
> y_mean = c(mean(thiouracil[,1]), mean(thiouracil[,2]), mean(thiouracil[,3]),
    mean(thiouracil[,4]), mean(thiouracil[,5]))

> x_mean
[1]  55.57143  75.85714 104.85714 134.14286 164.28571
> y_mean
[1]  54.7  76.3  95.8 108.2 124.0

#racunam uzoracku varijancu S_x
> S_x = matrix(0, nrow = p, ncol = p)
> for(i in 1 : n)
+ {
+ S_x = S_x + (thyroxin[i,] - x_mean) %*% (t(thyroxin[i,] - x_mean))
+ }
> S_x = S_x / (n - 1)

> S_x
        Time 0   Time 1    Time 2    Time 3    Time 4
[1,]  8.952381 15.92857  22.42857  33.07143  36.80952
[2,] 15.928571 41.47619  61.80952  85.52381  95.88095
[3,] 22.428571 61.80952 123.14286 195.52381 232.04762
[4,] 33.071429 85.52381 195.52381 346.80952 427.45238
[5,] 36.809524 95.88095 232.04762 427.45238 537.23810


#racunam uzoracku varijancu S_y
> S_y = matrix(0, nrow = p, ncol = p)
> for(i in 1 : m)
+ {
+ S_y = S_y + (thiouracil[i,] - y_mean) %*% (t(thiouracil[i,] - y_mean))
+ }
> S_y = S_y / (m - 1)

> S_y
       Time 0   Time 1   Time 2   Time 3    Time 4
[1,] 22.01111 27.87778 29.71111 28.40000  15.00000
[2,] 27.87778 62.67778 58.51111 43.15556  18.55556
[3,] 29.71111 58.51111 72.17778 69.04444  53.00000
[4,] 28.40000 43.15556 69.04444 95.06667  87.55556
[5,] 15.00000 18.55556 53.00000 87.55556 126.66667

#racunam S_pool
```

```
> S_pool = 1 / (n + m - 2) * ((n - 1)*S_x + (m - 1)*S_y)

> S_pool
       Time 0   Time 1    Time 2    Time 3    Time 4
[1,] 16.78762 23.09810  26.79810  30.26857  23.72381
[2,] 23.09810 54.19714  59.83048  60.10286  49.48571
[3,] 26.79810 59.83048  92.56381 119.63619 124.61905
[4,] 30.26857 60.10286 119.63619 195.76381 223.51429
[5,] 23.72381 49.48571 124.61905 223.51429 290.89524

#racunam Hotellingovu statistiku T^2
> T2 = (1/n + 1/m)^(-1) * (t(x_mean - y_mean)) %*% (solve(S_pool)) %*% (x_mean -
    y_mean)

> T2
         [,1]
[1,] 30.76503

#racunam omjer vjerodostojnosti kao funkciju od T^2
> Lambda = (1 + (1/(n + m - 2)*T2))^(-(n + m) / 2)

> Lambda
              [,1]
[1,] 7.624986e-05

#racunam testnu statistiku
> TS = (n + m - p - 1) / ((n + m - 2)*p) * T2

> TS
         [,1]
[1,] 4.512204

#racunam p-vrijednost
> p_value = 1 - pf(TS, p, m + n - p - 1)

> p_value
           [,1]
[1,] 0.01754365
```

Dakle, dobili smo da je p-vrijednost $= 0.01754365$, pa na razini značajnosti od 5% **odbacujemo nultu hipotezu o jednakosti očekivanja dvaju nezavisnih normalnih uzoraka, u korist alternativne hipoteze**.