Eötvös Loránd Tudományegyetem
Természettudományi Kar

Véletlenített tesztek a többváltozós varianciaanalízisben

Szakdolgozat

Halápi Gergely
Matematika BSc,
elemző szakirány

Témavezető: Prőhle Tamás
Valószínűségelméleti és Statisztikai Tanszék

Budapest
2018

Tartalomjegyzék

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Prőhle Tamásnak a végtelen türelemért, rengeteg segítségért, konzultációs időpontért, ötletért és tudásért, amit megosztott velem.

Szeretném még megköszönönni családomnak, barátnőmnek és barátaimnak a folyamatos támogatást, bíztatást, nagyon sokat jelentett.

1. Bevezetés

A varianciaanalízist a 20. században kezdték el használni, de főbb alkotóelemeinek alkalmazása, mint például a hipotézisvizsgálaté- már jóval régebbre tehető. A legkisebb négyzetes módszerek továbbfejlesztése (Gauss és Laplace, 1800 körül) által lehetőség nyílt arra, hogy a megfigyelések kombinálásának módszerei magasabb szintre emelkedjenek. Ugyanakkor a négyzetösszegek további tanulmányozása után Laplace rájött, hogyan becsülhet szórást reziduális négyzetösszegből. 1827-re pedig a legkisebb négyzetes módszereket Laplace ANOVA-problémák megoldására használta.

Az ANOVA és a MANOVA a statisztikában nagyon gyakran használt módszerek, melyek a mindennapi ember számára is érthető feladatok megoldására ugyanúgy használható, mint egy bonyolultabb problémáéra.

A szakdolgozatom első részében egy módszert mutatok be, mely alkalmas korreláció mátrixok generálására, ennek a módszernek egy kismértékben egyszerűsített változatát fogom használni a dolgozat végén szereplő egyik példában. A második részben két becslési eljárást írok le, a variancia-kovariancia mátrixokra vonatkozóan, melyek széles körben elterjedtek a valószínűségi vektorváltozók analízisének területén. A harmadik részben bemutatom az ANOVA és MANOVA módszereket, a negyedikben pedig az ezek egy speciális esetére használt véletlenített próbákat, a bootstrap és jackknife becslési eljárásokat. Az utolsó részben szemléltető példák szerepelnek.

2. Korreláció mátrix generálás

2.1. Bevezetés

Ebben a részben pozitív szemidefinit korreláció mátrixok generálására mutatok egy módszert. Ehhez először néhány szükséges alapdefiníciót írok le, ezek után következik a generáló módszer. A valószínűségi változó, illetve valószínűségi vektorváltozó itt nem kerül definiálásra, alapfogalomnak tekintjük őket, csakúgy, mint a várható érték és a szórás fogalmát. Az X valószínűségi változó várható értékét EX-el, szórását DX-el jelöljük.

2.1. Definíció. Kovariancia: Legyen X és Y két valószínűségi változó, ekkor X és Y kovarianciája: $Cov(X,Y) = E((X - EX)(Y - EY))$

2.2. Definíció. Korreláció: Legyen X és Y két valószínűségi változó, ekkor X és Y korrelációja: $\rho(X,Y) = \frac{Cov(X,Y)}{DXDY}$

Most térjünk át valószínűségi vektorváltozókra. Itt a várható érték helyett várható érték vektort veszünk, jele: $\underline{EX}$, a szórás többdimenziós megfelelője a variancia-kovariancia mátrix, amit Σ-val fogunk jelölni, a korreláció mátrix jele pedig legyen P. Legyen most $\underline{X}$ egy n-dimenziós valószínűségi vektorváltozó, $\underline{X}$ i. komponensét jelöljük $\underline{X}_i$-vel:

- A várható érték vektor i. komponense a valószínűségi vektorváltozó i. komponensének várható értékével lesz egyenlő, azaz: $\underline{EX}_i = EX_i$
- A variancia-kovariancia mátrix i. sorának j. oszlopában lévő elem egyenlő lesz X i. és j. komponensének kovarianciájával, azaz: $\Sigma_{ij} = Cov(X_i, X_j)$.
- A korreláció mátrix i. sorának j. oszlopában lévő elem egyenlő lesz X i. és j. komponensének korrelációjával, azaz: $P_{ij} = \rho(X_i, X_j)$.

A fentiekből látszik, a variancia-kovariancia mátrix három legfontosabb tulajdonsága:

1. A variancia-kovariancia mátrix diagonálisában a komponensek szórásnégyzetei szerepelnek, azaz: $\Sigma_{ii} = D^2X_i$, ez nyilvánvaló, hiszen:

$$\Sigma_{ii} = Cov(X_i, X_i) = E((X_i - EX_i)(X_i - EX_i)) =$$
$$E(X_i^2 - 2EX_i^2 + E^2X_i) = EX_i^2 - 2EX_i^2 + E^2X_i = E^2X_i - EX_i^2 = D^2X_i$$

2. A variancia-kovariancia mátrix szimmetrikus, ez a kovariancia tulajdonságaiból következik:

$$Cov(X, Y) = Cov(Y, X)$$

3. A tapasztalati és elméleti variancia-kovariancia mátrix pozitív szemidefinit.

- Először nézzük a **3.** tulajdonság bizonyítását az elméleti esetre:

  - Az $n \times n$-es Σ mátrix pozitív szemidefinit, ha minden nem nulla, $n \times 1$-es u vektorra $u^T\Sigma u \geq 0$

  - Mivel az $\underline{X} = (X_1, X_2, X_3, \ldots, X_n)^T$ valószínűségi vektorváltozó variancia kovariancia mátrixa kiszámolható, mint:
  $\Sigma = E[(\underline{X} - \underline{EX})(\underline{X} - \underline{EX})^T]$, innen:

$$\begin{aligned} u^T\Sigma u &= u^T E[(\underline{X} - \underline{EX})(\underline{X} - \underline{EX})^T]u \\ &= E[u^T(\underline{X} - \underline{EX})(\underline{X} - \underline{EX})^T u] \\ &= E[Z^2] \\ &= \sigma_Z^2 \ , \end{aligned}$$

ahol a centrálás miatt Z egy nulla várható értékű, σ_Z^2 szórásnégyzetű valószínűségi változó. Mivel tudjuk, hogy egy valószínűségi változó szórásnégyzete nemnegatív, készen vagyunk a bizonyítással.

$$u^T\Sigma u = \sigma_Z^2 \geq 0$$

- Bizonyítás a tapasztalati esetre:

  - Legyen az $\underline{X}$ n-dimenziós vektorváltozó i. megfigyelése $i = 1 \dots n$-re: $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{in})^T$. Ekkor a mintaátlag vektor: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

  - A mintához tartozó tapasztalati variancia-kovariancia mátrixot a következőképp számoljuk: $\Sigma_{tap} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$

  - Innen az elméleti esethez hasonlóan, legyen u egy tetszőleges, nem nulla, $n \times 1$-es vektor, ekkor:

$$\begin{aligned} u^T\Sigma u &= u^T \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T \right) u \\ &= \frac{1}{n}\sum_{i=1}^{n} u^T(x_i - \overline{x})(x_i - \overline{x})^T u \\ &= \frac{1}{n}\sum_{i=1}^{n}((x_i - \overline{x})^T u)^T((x_i - \overline{x})^T u) \\ &= \frac{1}{n}\sum_{i=1}^{n}((x_i - \overline{x})^T u)^2 \geq 0 \end{aligned}$$

  Beláttuk tehát, hogy a tapasztalati variancia-kovariancia mátrix is pozitív szemidefinit.

A variancia-kovariancia mátrixhoz hasonlóan a korreláció mátrix is rendelkezik négy fő tulajdonsággal:

1.A korreláció mátrix diagonálisában csupa egyes szerepel.

2.A korreláció mátrix nem diagonálisbeli elemei mind valós számok a $[-1, 1]$ intervallumból.

3.A korreláció mátrix szimmetrikus.

4.A tapasztalati és elméleti korreláció mátrix pozitív szemidefinit.

2.2. Korreláció mátrixok generálása

Az algoritmus alapjai

Egy minden fent említett követelménynek megfelelő korreláció mátrix (P) konstruálható trigonometrikus függvények segítségével. A korreláció mátrix ekkor szögek ($\vartheta_{i,j}$) egy függvényévé válik, ami egy hatásos módszert ad a határok kiszámítására. A fenti követelményeket kielégítő korrelációs mátrix leírható, mint:

$$P = BB^T, \tag{1}$$

ahol

$$b_{ij} = \begin{cases} cos(\vartheta_{i,j}), & j = 1\text{-re} \\ cos(\vartheta_{i,j}) \prod_{k=1}^{j-1} sin(\vartheta_{i,k}), & 2 \leq j \leq n-1\text{-re} \\ \prod_{k=1}^{j-1} sin(\vartheta_{i,k}), & j = n\text{-re} \end{cases} \tag{2}$$

B egy n-dimenziós, négyzetes mátrix, aminek elemei a (2)-ben meghatározott b_{ij}-k. [7] alapján (2)-t egyszerűsíthetjük, ha minden i-re ϑ_{ii}-t 0-nak vesszük, ekkor B egy alsó háromszög mátrixá redukálódik:

$$b_{ij} = \begin{cases} 1, & i = 1,\ j = 1\text{-re} \\ cos(\vartheta_{i,j}), & i \geq 2\text{-től},\ j = 1\text{-re} \\ \prod_{k=1}^{j-1} sin(\vartheta_{i,k}), & i = j,\ 2 \leq i,\ j \leq n\text{-re} \\ cos(\vartheta_{i,j}) \prod_{k=1}^{j-1} sin(\vartheta_{i,k}), & 2 \leq j \leq i-1\text{-re} \\ 0, & i+1 \leq j \leq n\text{-re} \end{cases} \tag{3}$$

Ekkor a B mátrix az alábbi módon fog kinézni:

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ cos(\vartheta_{2,1}) & sin(\vartheta_{2,1}) & 0 & \dots & 0 \\ cos(\vartheta_{3,1}) & cos(\vartheta_{3,2})sin(\vartheta_{3,1}) & sin(\vartheta_{3,2})sin(\vartheta_{3,1}) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cos(\vartheta_{n,1}) & cos(\vartheta_{n,2})sin(\vartheta_{n,1}) & cos(\vartheta_{n,3})sin(\vartheta_{n,2})sin(\vartheta_{n,1}) & \dots & \prod_{k=1}^{n-1}\sin(\vartheta_{n,k}) \end{pmatrix} \tag{4}$$

Innen látható, hogy a B mátrix erősen függ $\vartheta_{i,j}$-től, ezt a szöget korrelatív szögnek nevezzük. A korrelatív szögek négyzetes mátrixát (ϑ) a következőképp definiáljuk:

$$\vartheta = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vartheta_{2,1} & 0 & 0 & \dots & 0 & 0 & 0 \\ \vartheta_{3,1} & \vartheta_{3,2} & 0 & \dots & 0 & 0 & 0 \\ \vartheta_{4,1} & \vartheta_{4,2} & \vartheta_{4,3} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vartheta_{n,1} & \vartheta_{n,2} & \vartheta_{n,3} & \dots & \vartheta_{n,n-2} & \vartheta_{n,n-1} & 0 \end{pmatrix} \tag{5}$$

Tehát egy minden feltételt kielégítő korreláció mátrixot számolhatunk, ha ismerjük a korrelatív szögek mátrixát.

1. Példa. Most nézzünk egy példát az eddig leírt számolási módszerre: Először is tegyük fel, hogy adott a korrelatív szögek mátrixa, ami ebben az esetben négy dimenziós:

$$\vartheta = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \vartheta_{2,1} & 0 & 0 & 0 \\ \vartheta_{3,1} & \vartheta_{3,2} & 0 & 0 \\ \vartheta_{4,1} & \vartheta_{4,2} & \vartheta_{4,3} & 0 \end{pmatrix} \tag{6}$$

A B mátrix pedig kifejezhető, mint:

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ b_{2,1} & b_{2,2} & 0 & 0 \\ b_{3,1} & b_{3,2} & b_{3,3} & 0 \\ b_{4,1} & b_{4,2} & b_{4,3} & b_{4,4} \end{pmatrix} = \tag{7}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ cos(\vartheta_{2,1}) & sin(\vartheta_{2,1}) & 0 & 0 \\ cos(\vartheta_{3,1}) & cos(\vartheta_{3,2})sin(\vartheta_{3,1}) & sin(\vartheta_{3,2})sin(\vartheta_{3,1}) & 0 \\ cos(\vartheta_{4,1}) & cos(\vartheta_{4,2})sin(\vartheta_{4,1}) & cos(\vartheta_{4,3})sin(\vartheta_{4,2})sin(\vartheta_{4,1}) & sin(\vartheta_{4,3})sin(\vartheta_{4,2})sin(\vartheta_{4,1}) \end{pmatrix}$$

Végül a korreláció mátrix:

$$P = BB^T = \begin{pmatrix} 1 & c_{2,1} & c_{3,1} & c_{4,1} \\ c_{2,1} & 1 & c_{3,2} & c_{4,2} \\ c_{3,1} & c_{3,2} & 1 & c_{4,3} \\ c_{4,1} & c_{4,2} & c_{4,3} & 1 \end{pmatrix} \tag{8}$$

Ahol,

$$\begin{aligned} c_{2,1} &= b_{2,1} \\ c_{3,1} &= b_{3,1} \\ c_{4,1} &= b_{4,1} \\ c_{3,2} &= b_{2,1}b_{3,1} + b_{2,2}b_{3,2} \\ c_{4,2} &= b_{2,1}b_{4,1} + b_{2,2}b_{4,2} \\ c_{4,3} &= b_{3,1}b_{4,1} + b_{3,2}b_{4,2} + b_{3,3}b_{4,3} \end{aligned} \tag{9}$$

Ezek az elemek, viszont leírhatók a korrelatív szögek függvényében is:

$$\begin{aligned} c_{2,1} &= cos(\vartheta_{2,1}) \\ c_{3,1} &= cos(\vartheta_{3,1}) \\ c_{4,1} &= cos(\vartheta_{4,1}) \\ c_{3,2} &= cos(\vartheta_{2,1})cos(\vartheta_{3,1}) + sin(\vartheta_{2,1})cos(\vartheta_{3,2})sin(\vartheta_{3,1}) \\ c_{4,2} &= cos(\vartheta_{2,1})cos(\vartheta_{4,1}) + sin(\vartheta_{2,1})cos(\vartheta_{4,2})sin(\vartheta_{4,1}) \\ c_{4,3} &= cos(\vartheta_{3,1})cos(\vartheta_{4,1}) + cos(\vartheta_{3,2})sin(\vartheta_{3,1})cos(\vartheta_{4,2})sin(\vartheta_{4,1}) \end{aligned} \tag{10}$$

A korrelációs együtthatók határai

Ahogy azt (6)-tól (10)-ig láttuk, egy minden feltételt teljesítő korreláció mátrixot konstruálhatunk a B mátrix segítségével, B elemeit pedig a korrelatív szögek határozzák meg, ahogy azt szintén fent láthattuk. Pontosan ezért meghatározhatjuk, hogy B mely elemeire van hatással a korrelatív szögek megváltoztatása a ϑ mátrixban, amiből két fontos következtetés vonható le:

1:Az első oszlopbeli korrelációs együtthatók ($c_{i,1}$) erősen függenek $\vartheta_{i,1}$-től.

2:A többi korrelációs együttható ($c_{i,j}$, ha $j \geq 2$) is számolható, ha $\vartheta_{p,q}$ adott, $p \leq i$ és $q \leq j < i$-re.

Mivel minden $\vartheta_{i,j}$ a $[0, \pi]$ zárt intervallumon van értelmezve, a sinus függvény nemnegatív értékeket fog adni ezekre a szögekre, míg a cosinus függvény a $[-1, 1]$ intervallumon vehet fel értékeket. Példának használva a (10)-ben kapott korrelációs együtthatókat, tisztán látszik, hogy bármely korrelációs együttható ($c_{i,j}$) határait ki tudjuk számolni, ha a $cos(\vartheta_{i,j})$-t (-1)-re, vagy 1-re rögzítjük. Továbbá látható az is, hogy a határok számolásához csak $\vartheta_{p,q}$-ra van szükség, az $p \leq i$ és $q \leq j < i$ esetekben, kivéve a $p = i$ és $q = j$ esetekben. Fontos hozzátenni azt is, hogy a korrelációs együttható határainak számolása során nem feltétlenül használjuk fel az összes $\vartheta_{p,q}$-t, ahogy az a korábbi, négy dimenziós példánkra vonatkozó határok táblázatában is látszik:

$c_{i,j}$	alsó határ	felső határ	szükséges $\vartheta_{i,j}$
$c_{2,1}$	-1	1	egyik sem
$c_{3,1}$	-1	1	egyik sem
$c_{4,1}$	-1	1	egyik sem
$c_{3,2}$	$cos(\vartheta_{2,1})cos(\vartheta_{3,1}) - sin(\vartheta_{2,1})sin(\vartheta_{3,1})$	$cos(\vartheta_{2,1})cos(\vartheta_{3,1}) + sin(\vartheta_{2,1})sin(\vartheta_{3,1})$	$\vartheta_{2,1}, \vartheta_{3,1}$
$c_{4,2}$	$cos(\vartheta_{2,1})cos(\vartheta_{4,1}) - sin(\vartheta_{2,1})sin(\vartheta_{4,1})$	$cos(\vartheta_{2,1})cos(\vartheta_{4,1}) - sin(\vartheta_{2,1})sin(\vartheta_{4,1})$	$\vartheta_{2,1}, \vartheta_{4,1}$
$c_{4,3}$	$cos(\vartheta_{2,1})cos(\vartheta_{4,1}) + cos(\vartheta_{3,2})$ $sin(\vartheta_{3,1}cos(\vartheta_{4,2}sin(\vartheta_{4,1}) - sin(\vartheta_{3,2})$ $sin(\vartheta_{3,1})sin(\vartheta_{4,2})sin(\vartheta_{4,1})$	$cos(\vartheta_{2,1})cos(\vartheta_{4,1}) + cos(\vartheta_{3,2})$ $sin(\vartheta_{3,1}cos(\vartheta_{4,2}sin(\vartheta_{4,1}) + sin(\vartheta_{3,2})$ $sin(\vartheta_{3,1})sin(\vartheta_{4,2})sin(\vartheta_{4,1})$	$\vartheta_{2,1}, \vartheta_{3,1}$ $\vartheta_{4,1}, \vartheta_{3,2}$ $\vartheta_{4,2}$

Ezek eredményeképp, ha a $c_{i,j}$ a számolt határain belül helyezkedik el és a számoláshoz szükséges $\vartheta_{p,q}$-k adottak, akkor $\vartheta_{i,j}$-t ki tudjuk számolni (11) alapján.

$$
\begin{aligned}
\vartheta_{2,1} &= arccos(c_{2,1}) \\
\vartheta_{3,1} &= arccos(c_{3,1}) \\
\vartheta_{4,1} &= arccos(c_{4,1}) \\
\vartheta_{3,2} &= arccos\left(\frac{c_{3,2} - b_{2,1}b_{3,1}}{b_{2,2}\sin(\vartheta_{3,1})}\right) \\
\vartheta_{4,2} &= arccos\left(\frac{c_{4,2} - b_{2,1}b_{4,1}}{b_{2,2}\sin(\vartheta_{4,1})}\right) \\
\vartheta_{4,3} &= arccos\left(\frac{c_{4,3} - (b_{3,1}b_{4,1} + b_{3,2}b_{4,2})}{b_{3,3}\sin(\vartheta_{4,2})\sin(\vartheta_{4,1})}\right)
\end{aligned} \tag{11}
$$

Ugyanezt a módszert könnyen lehet alkalmazni magasabb dimenziós korreláció mátrixok esetén is, habár ez hosszabb formulákra vezet és több számolást igényel.

Az algoritmus

Ebben a részben egy algoritmust ismertetek a korreláció mátrix konstruálására, úgy, hogy egymás után számoljuk minden korrelációs együttható határait a bevezető részben leírtak alapján, és véletlen változókat generálunk egyenletes eloszlásból, ezeken a határokon belül.
Legyen U szigorúan alsó háromszög mátrixa a véletlen, egyenletes változóknak amik a $[0, 1]$ intervallumon vesznek fel valamilyen értéket, legyen ϑ a korrelatív szögek szigorúan alsó háromszög mátrixa, és Y és Z a korrelációs együtthatók alsó és felső határainak szigorúan alsó háromszög mátrixai.

Az alábbiakban ismertetem azt a négylépéses algoritmust, ami egy $n \times n$ korreláció mátrixot generál:

1.lépés: Számoljuk ki az első oszlopbeli korrelációs együtthatókat a következő módon:
$i = 1 \ldots n$-re, legyen $c_{i,1} = -1 + 2 \times u_{i,1}$, $b_{i,1} = c_{i,1}$ és számoljuk ki $\vartheta_{i,1}$-et
$i = 2 \ldots n$-re, legyen $b_{i,j} = \sin(\vartheta_{i,1})$,ha $j = 1 \ldots i$.

2.lépés: Számoljuk ki a maradék korrelációs együtthatókat a harmadik sortól az utolsó sorig ($i = 3, \ldots, n$) és a második oszloptól az utolsó oszlopig ($j = 2, \ldots, i - 1$), minden sorban.

Ezután számoljuk ki minden korrelációs együttható alsó ($y_{i,j}$) és felső ($z_{i,j}$) határait. A határok kiszámolására az előző részben ismertetett módszert használjuk, az első táblázatban látható példa a határok kiszámolására egy négydimenziós korreláció mátrix esetén.

- Ha $z_{i,j} - y_{i,j} < K$, akkor legyen $c_{i,j} = y_{i,j} + (z_{i,j} - y_{i,j})/2$, egyébként legyen $c_{i,j} = y_{i,j} + (z_{i,j} - y_{i,j}) \times u_{i,j}$. Itt a K egy küszöbérték, ami segít az instabilitás csökkentésében, oly módon, hogy minden korrelációs együtthatót, amire az alsó és felső határ által meghatározott $[y_{i,j}, z_{i,j}]$ intervallum hossza kisebb, mint K bekényszerít a $[y_{i,j}, z_{i,j}]$ intervallum felezőpontjába. Nagyobb K érték stabilabb rendszert fog eredményezni, de ezzel csökken a véletlen hatás a $c_{i,j}$ együtthatókban.
- Számoljuk ki $\vartheta_{i,j}$-t a (11)-ben bemutatotthoz hasonlóan.
- Konstruáljuk meg a szimmetrikus korreláció mátrixunkat az előzőekben generált korrelációs együtthatók segítségével, a főátlóban csupa 1-es szerepeljen.

3.lépés: Véletlenszerűen rendezzük újra a korreláció mátrixot. Erre azért van szükség, hogy megbizonyosodhassunk arról, hogy minden korrelációs együttható azonos eloszlású.

4.lépés: Ellenőrizzük, hogy az így kapott korreláció mátrixunk minden követelménynek megfelel-e. Habár elméletileg a fent leírt lépések egy minden követelménynek megfelelő korreláció mátrixot adnak, néhány esetben az instabilitás feltűnése miatt előfordulhat, hogy rossz korreláció mátrixot kapunk. Az instabilitásnak két fő oka lehet:

- K túl kicsi a mátrix dimenziójához viszonyítva
- a generált korrelációs együtthatók túl közel helyezkednek el az alsó , vagy felső korlátukhoz

Bár annak a valószínűsége, hogy rossz korreláció mátrixot kapunk nagyon kicsi, azért mégsem 0. Ez a lépés azért fontos, mert így bizonyosodhatunk meg afelől, hogy nem fogunk rossz korreláció mátrixot használni további vizsgálódásaink során. Tehát az utolsó lépésünk két alapvető allépésből tevődik össze:

1. Keressük meg a legkisebb sajátértéket, ha ez negatív, a generált korrelációs mátrixunk rossz, egyébként jó lesz.

2. Ha rossz a generált korreláció mátrixunk, térjünk vissza az első lépéshez és generáljunk újat.

2. Példa. Most kreáljunk egy ötdimenziós korreláció mátrixot, úgy, hogy az algoritmus első lépésében szereplő egyenletes, véletlen mátrix (U) a következőképp néz ki:

$$U = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.6220 & 0 & 0 & 0 & 0 \\ 0.0751 & 0.8576 & 0 & 0 & 0 \\ 0.9668 & 0.6035 & 0.4107 & 0 & 0 \\ 0.6100 & 0.8478 & 0.7324 & 0.7571 & 0 \end{pmatrix}$$

Mielőtt még véletlenszerűen újrarendeznénk őket, az alsó korlát mátrix (Y), a felső korlát mátrix (Z), és a korreláció mátrix (X) a következőképpen generálható:

$$Y = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ -1 & -0.7185 & 0 & 0 & 0 \\ -1 & -0.1197 & -0.8946 & 0 & 0 \\ -1 & -0.8923 & -0.1893 & 0.0213 & 0 \end{pmatrix}$$

$$Z = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0.3038 & 0 & 0 & 0 \\ 1 & 0.5753 & -0.6363 & 0 & 0 \\ 1 & 0.8478 & 0.7324 & 0.7571 & 0 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & 0.2440 & -0.8498 & 0.9336 & 0.2200 \\ 0.2440 & 1 & 0.1582 & 0.2997 & 0.7117 \\ 0.-8498 & 0.1582 & 1 & -0.7885 & 0.1889 \\ 0.9336 & 0.2997 & -0.7885 & 1 & 0.3454 \\ 0.2200 & 0.7117 & 0.1889 & 0.3454 & 1 \end{pmatrix}$$

Mivel a kapott korreláció mátrix legkisebb sajátértéke 0.00510, a korreláció mátrix pozitív szemidefinit, ami igazolja, hogy X minden követelménynek megfelel.

3. Vektorváltozó variancia-kovariancia mátrixának becslése

3.1. Bevezetés

A gyakorlatban nagyon ritkán tudjuk meghatározni egy, vagy két valószínűségi vektorváltozó valódi (elméleti) variancia-kovariancia mátrixát, de valahogy szeretnénk becsülni azt. Ebben a részben két ilyen becslési módszert fogok ismertetni. A becsült variancia-kovariancia mátrixot $\hat{\Sigma}$-val jelöljük.

3.2. Az X^TX-módszer

Ez a legáltalánosabb módszer a variancia-kovariancia mátrix becslésére. A következő módszerrel ellentétben itt nem teszünk fel semmit a vizsgált vektorváltozó eloszlásáról.

Tegyük fel, hogy az X egy $n \times k$-s mátrix, ami az $\underline{X}$ k-dimenziós valószínűségi vektorváltozóra vonatkozóan tartalmaz n darab megfigyelést. $\underline{X}$ j. komponensének i. megfigyelését X_{ij}-vel jelöljük, ekkor a mátrixunk:

$$X = \begin{pmatrix} X_{11} & X_{12} & X_{13} & \dots & X_{1k} \\ X_{21} & X_{22} & X_{23} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & X_{n3} & \dots & X_{nk} \end{pmatrix}$$

Az X^TX-módszer a fent említett X mátrixból készíti el a variancia-kovariancia mátrix becslését.

1.lépés: Készítsük el az x mátrixot a $x = X - 1 \cdot 1^T \cdot X \cdot \frac{1}{n}$ képlettel (ez lesz az úgynevezett eltérésmátrix), ahol az 1 egy $n \times 1$-es oszlopvektor, aminek elemei csak egyesek, az x az $n \times k$-s eltérésmátrix és a X az adatokat tartalmazó mátrix. Az x mátrix különbségeket fog tartalmazni: az X mátrix minden eleméből kivonjuk az adott elem oszlopának átlagát. Ez a következő módon fog kinézni:

$$x = \begin{pmatrix} X_{11} - \sum_{i=1}^{n} \frac{X_{i1}}{n} & X_{12} - \sum_{i=1}^{n} \frac{X_{i2}}{n} & \dots & X_{1k} - \sum_{i=1}^{n} \frac{X_{ik}}{n} \\ X_{21} - \sum_{i=1}^{n} \frac{X_{i1}}{n} & X_{22} - \sum_{i=1}^{n} \frac{X_{i2}}{n} & \dots & X_{2k} - \sum_{i=1}^{n} \frac{X_{ik}}{n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \sum_{i=1}^{n} \frac{X_{i1}}{n} & X_{n2} - \sum_{i=1}^{n} \frac{X_{i2}}{n} & \dots & X_{nk} - \sum_{i=1}^{n} \frac{X_{ik}}{n} \end{pmatrix}$$

2.lépés: Kiszámoljuk a $k \times k$-s eltérés-négyzetösszeg mátrixot: $z = x^T x$

3.lépés: A z mátrix összes elemét elosztjuk n-el, így hozzuk létre a $k \times k$-s variancia-kovariancia mátrix becslését: $\hat{\Sigma} = z \cdot \frac{1}{n}$.

3. Példa. Becsüljük meg az $X^T X$-módszerrel az alábbi megfigyelések alapján a variancia-kovariancia mátrixot! Az alábbi mátrixban 4 tanuló dolgozatainak pontszámait látjuk, 3 különböző tárgyból:

$$X = \begin{pmatrix} 90 & 67 & 81 \\ 83 & 73 & 67 \\ 83 & 80 & 92 \\ 24 & 32 & 60 \end{pmatrix}$$

ebből az $x = X - 1 \cdot 1^T \cdot X \cdot \frac{1}{n}$ képlettel készítsük el az x mátrixot:

$$x = \begin{pmatrix} 20 & 4 & 6 \\ 13 & 10 & -8 \\ 13 & 17 & 17 \\ -46 & -31 & -15 \end{pmatrix}$$

az x-mátrixból számoljuk az eltérés-négyzetösszeg mátrixot, a $z = x^T x$ képlettel:

$$z = \begin{pmatrix} 2854 & 1857 & 927 \\ 1857 & 1366 & 698 \\ 927 & 698 & 614 \end{pmatrix},$$

z minden elemét leosztva 4-el ($n = 4$) kapjuk a variancia-kovariancia mátrix becslését:

$$\hat{\Sigma} = \begin{pmatrix} 713.50 & 464.25 & 231.75 \\ 464.25 & 341.50 & 174.50 \\ 231.75 & 174.50 & 153.50 \end{pmatrix}$$

Itt $\hat{\Sigma}$-n látszik, hogy szimmetrikus, legkisebb sajátértéke pedig: 22.82108, tehát pozitív definit is, ebből tudjuk, hogy ez egy minden követelménynek megfelelő variancia-kovariancia mátrix becslés.

3.3. Maximum likelihood becslés többdimenziós normális eloszlás esetén

A maximum likelihood becslés abszolút folytonos eloszlás esetén

Ha egy adott valószínűségi változóra, vagy vektorváltozóra vonatkozóan birtokában vagyunk többszöri megfigyelésnek, viszont valamelyik paraméterét nem ismerjük, akkor használjuk a maximum likelihood becslést. A becslés során azt a paramétert fogjuk keresni ami mellett a legnagyobb az adott mintarealizáció valószínűsége. Az ismeretlen paramétert ϑ-val, becslését $\hat{\vartheta}$-val jelöljük.

Legyen adott egy X valószínűségi változó f sűrűségfüggvénnyel, ami függ a ϑ paramétertől. Tegyük fel, hogy rendelkezésünkre áll n-darab megfigyelés az X-re vonatkozóan. Ha ez az n-elemű minta független, azonos eloszlású, az együttes sűrűségfüggvény a következőképp fejezhető ki:

$$f(x_1, x_2, \ldots, x_n; \vartheta) = \prod_{i=1}^{n} f_{X_i}(x_i; \vartheta) = L(\vartheta),$$

ezt nevezzük a mintához tartozó likelihood-függvénynek. A maximum likelihood becslés a likelihood-függvény maximuma az adott paraméter szerint. A számítások egyszerűsítése végett általában a likelihood-függvény természetes alapú logaritmusát vesszük, ezt nevezzük log-likelihood függvénynek:

$$l(\vartheta) = log\left(\prod_{i=1}^{n} f_{X_i}(x_i; \vartheta)\right) = \sum_{i=1}^{n} log\left(f_{X_i}(x_i; \vartheta)\right),$$

ezt szeretnénk maximalizálni ϑ szerint, ennek maximuma lesz $\hat{\vartheta}$. Ezt azért tehetjük meg, mert a *log* - függvény szigorúan monoton növekvő függvény, ennélfogva : $\text{argmax}(L(\vartheta)) = \text{argmax}(l(\vartheta))$.

A variancia-kovariancia mátrix maximum-likelihood becslése

Ebben az esetben feltételezzük a vizsgált vektorváltozóról, hogy többdimenziós normális eloszlású, ebből a feltételezésből kiindulva készítjük el a variancia-kovariancia mátrixának becslését. Tegyük fel az $\underline{X} \in \mathbb{R}^{p\times 1}$-es vektorváltozóról, hogy p-dimenziós normális eloszlású, $\Sigma \in \mathbb{R}^{p\times p}$-s, pozitív szemidefinit variancia-kovariancia mátrixszal, $\underline{X}$ sűrűségfüggvénye pedig a következő:

$$f(x) = (2\pi)^{-\frac{p}{2}}\ det(\Sigma)^{-\frac{1}{2}}\ exp\left(-\frac{1}{2}\ (\underline{x}-\mu)^T\ \Sigma^{-1}\ (\underline{x}-\mu)\right),$$

ahol a $\mu \in \mathbb{R}^{p\times 1}$ az $\underline{X}$ várható érték vektora.

- Tegyük fel, hogy az $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$-ek független, azonos eloszlású minták a fenti eloszlásból. Az ebből a mintából származó $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ megfigyelt értékeket tartalmazó vektorokkal szeretnénk becsülni Σ-t.

- A likelihood-függvény a következőképp fog kinézni:

$$\mathcal{L}(\mu, \Sigma) = (2\pi)^{-\frac{np}{2}}\ \prod_{i=1}^{n} det(\Sigma)^{-\frac{1}{2}}\ exp\left(-\frac{1}{2}\ (x_i-\mu)^T\ \Sigma^{-1}\ (x_i-\mu)\right)$$

- Tudjuk, hogy a μ várható érték vektor maximum likeklihood becslése a $p \times 1$-es mintaátlag vektor lesz:

$$\overline{x} = \frac{\underline{x}_1 + \underline{x}_2 + \cdots + \underline{x}_n}{n}$$

- Mivel $\overline{x}$ nem függ Σ-tól, beírhatjuk a μ helyére a maximum likelihood függvényben:

$$\mathcal{L}(\overline{x}, \Sigma) =\ det(\Sigma)^{-\frac{n}{2}}\ exp\left(-\frac{1}{2}\ \sum_{i=1}^{n}(x_i-\overline{x})^T\ \Sigma^{-1}\ (x_i-\overline{x})\right),$$

és most azt a Σ-t keressük, ami maximalizálja a likelihood függvényt.

- Értelmezzük a $(x_i-\overline{x})^T\ \Sigma^{-1}\ (x_i-\overline{x})$ skalárt,mint egy 1×1-es mátrix nyomát. Ez lehetővé teszi, hogy használjuk a $tr(AB) = tr(BA)$ azonosságot, ha A és B olyan mátrixok, melyekre létezik az AB és BA szorzat is, ekkor:

$$\begin{aligned}
\mathcal{L}(\overline{x},\Sigma) = &\ det(\Sigma)^{-\frac{n}{2}}\ exp\left(-\frac{1}{2}\ \sum_{i=1}^{n} tr\left((x_i-\overline{x})^T\ \Sigma^{-1}\ (x_i-\overline{x})\right)\right)\\
= &\ det(\Sigma)^{-\frac{n}{2}}\ exp\left(-\frac{1}{2}\ \sum_{i=1}^{n} tr\left((x_i-\overline{x})(x_i-\overline{x})^T\ \Sigma^{-1}\right)\right)\\
= &\ det(\Sigma)^{-\frac{n}{2}}\ exp\left(-\frac{1}{2}\ tr\left(\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^T\ \Sigma^{-1}\right)\right)\\
= &\ det(\Sigma)^{-\frac{n}{2}}\ exp\left(-\frac{1}{2}\ tr\left(S\Sigma^{-1}\right)\right),
\end{aligned}$$

ahol $S=\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^T \in \mathbb{R}^{p\times p}$, az úgynevezett szórásmátrix, ami pozitív definit. A lineáris algebrából jól ismert spektrálfelbontásból következik, hogy egy pozitív definit, szimmetrikus mátrix (ebben az esetben S) egy sajátos, pozitív definit, szimmetrikus négyzetgyökkel rendelkezik ($S^{\frac{1}{2}}$). A nyom fent említett tulajdonságát újra felhasználva kapjuk:

$$\mathcal{L}(\overline{x},\Sigma) = \ det(\Sigma)^{-\frac{n}{2}}\ exp\left(-\frac{1}{2}\ tr\left(S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}}\right)\right)$$

Most legyen $B=S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}}$, ekkor:

$$\mathcal{L}(\overline{x},\Sigma) = \ det(S)^{-\frac{n}{2}}\ det(B)^{\frac{n}{2}}\ exp\left(-\frac{1}{2}\ tr\left(B\right)\right)$$

A B mátrix pozitív definit mátrix, tehát diagonalizálható, így a problémánk módosul arra, hogy azt a B-t találjuk meg, ami maximalizálja az alábbi kifejezést:

$$det(B)^{\frac{n}{2}}\ exp\left(-\frac{1}{2}\ tr\left(B\right)\right)$$

Mivel egy négyzetes mátrix nyoma egyenlő a sajátértékeinek összegével, azokat a $\lambda_1, \lambda_2, \ldots, \lambda_p$-ket szeretnénk megtalálni amik maximalizálják az alábbi kifejezést:

$$\lambda_i^{\frac{n}{2}} \; exp\left(-\frac{\lambda_i}{2}\right)$$

Ha ezt a kifejezést deriváljuk λ szerint, $\lambda_i = n$-et kapunk, minden i-re. Tegyük fel, hogy Q a sajátvektorok mátrixa, a $p \times p$-s egységmátrixot pedig I_p-vel jelöljük, ekkor:

$$B = Q(nI_p)Q^{-1} = nI_p$$

Végül a $\Sigma \in \mathbb{R}^{p \times p}$-s variancia-kovariancia mátrixot az alábbi módon kapjuk:

$$\begin{aligned} \Sigma = S^{\frac{1}{2}} \; B^{-1} \; S^{\frac{1}{2}} = S^{\frac{1}{2}} \left(\frac{1}{n} I_p\right) S^{\frac{1}{2}} = \frac{S}{n} \\ = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^T \end{aligned}$$

Megmutatható a véletlen S mátrixról, hogy Wishart eloszlású $n-1$ szabadsági fokkal.

4. Varianciaanalízis

4.1. ANOVA-MANOVA

Bevezetés

A varianciaanalízis egy statisztikán belül nagyon sokszor és sokféle módon használt módszer, ami egy diszkrét és egy abszolút folytonos valószínűségi változó, vagy vektorváltozó közötti kapcsolatot mutatja ki. Az abszolút folytonos változót kategorizáljuk a diszkrét változó felvett értékei szerint, így osztályokat kapunk. Ezeknek az osztályoknak a várható értékeinek különbözőségéről, vagy egyenlőségéről szeretnénk dönteni, hipotézisvizsgálat segítségével. A hipotézisvizsgálat során a szórásokat fogjuk elemezni, hogy információt kapjunk a várható értékek egyenlőségéről. A várható értékek eltéréséből eredő "fölös" szórás tesztelése ad lehetőséget a döntésre.

Az ötlet, hogy kétféleképp becsüljük meg a szórást:
- A teljes mintából, az osztályonkénti szórásbecslések átlagaként
- Az osztályok átlagainak szórásaként (ezt még fel kell szoroznunk az elemszámmal, hogy össze tudjuk hasonlítani az előzővel)

Ha a várható értékek között nincs szignifikáns eltérés, a kétféle becslésnek szinte ugyanazt kellene adnia, eltérésük nem lehet szignifikáns. A két becslés egyezését független, normális eloszlású minta esetén, a Fisher-Cohran tételből következően F-próba segítségével ellenőrizhetjük:

Fisher-Cohran tétel: Ha $X_1, X_2, X_3, \ldots, X_n$ független, azonos eloszlású minta, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, azaz $\overline{X}$ a mintaátlag, és $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, a korrigált tapasztalati szórásnégyzet, akkor: $\overline{X}$ és S_n^{*2} pontosan akkor függetlenek, ha $(X_1, X_2, X_3, \ldots, X_n)$ normális eloszlású.

Ez esetünkben azt jelenti, hogy a teljes mintából számolt szórás becslése független az átlagoktól, tehát az átlagokból számolt szórás becslésétől is. Két független becslésünk van normális eloszlású mintából, így az összes feltétel adott az F-próba elvégzéséhez.

One-way ANOVA

A modell felírása: A One-way ANOVA esetén adott két valószínűségi változó, az abszolút folytonos Y, $f(Y)$ sűrűségfüggvénnyel és a diszkrét X, ami k-féle értéket vehet fel, ezek kapcsolatát szeretnénk tanulmányozni.

Jelöljük Y egy realizációját y-al. Az Y-ra vonatkozó $n \times k$ megfigyelés esetén az adatmátrix az alábbi módon fog kinézni:

1.osztály	**2.osztály**	**3.osztály**	...	**k.osztály**
$\mathcal{N}(\mu_1, \sigma^2)$	$\mathcal{N}(\mu_2, \sigma^2)$	$\mathcal{N}(\mu_3, \sigma^2)$	...	$\mathcal{N}(\mu_k, \sigma^2)$
$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$\cdots$	$y_{1,k}$
$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$\cdots$	$y_{2,k}$
$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$\cdots$	$y_{3,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$y_{n_1,1}$	$y_{n_2,2}$	$y_{n_3,3}$	$\cdots$	$y_{n_k,k}$

Legyen $\sum_{i=1}^{k} n_i = N$. Az N darab megfigyelést vehetjük egy véletlen mintának, ami az X szerinti osztályozás után, végül az $1., 2., 3., \ldots, k.$ osztályokra rendre $n_1,\ n_2,\ n_3, \ldots,\ n_k$ megfigyelésszámokat eredményez. Az osztályokban jelölje Y várható értékét : $\mu_1,\ \mu_2,\ \mu_3, \ldots, \mu_k$. Minden osztályban Y eloszlását normálisnak tekintjük azonos σ^2 szórásnégyzettel, legfeljebb az osztályok várható értéke különbözhet. Az ANOVA célja, hogy információt nyerjen a várható értékek egyenlőségéről, ezt hipotézisvizsgálat segítségével viszi végbe:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$
$$H_1 : \exists\ i, j,\ \text{hogy}\ i \neq j - re : \mu_i \neq \mu_j$$

A $j.$ osztály $i.$ megfigyelését jelöljük y_{ij}-vel, ahol $i = 1,\ 2,\ 3, \ldots,\ n_j$ és $j = 1,\ 2,\ 3, \ldots,\ k$. Most nézzük a szükséges képleteket:

- A $j.$ osztály mintaátlaga: $\overline{y}_{\bullet j} = \sum_{i=1}^{n_j} \frac{y_{ij}}{n_j}$
- Az N megfigyelés mintaátlaga, azaz a teljes átlag: $\overline{y}_{\bullet\bullet} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{y_{ij}}{n_j}$
- A $j.$ osztály mintából számolt szórása: $s_j^2 = \sum_{i=1}^{n_j} \frac{(y_{ij} - \overline{y}_{\bullet j})^2}{n_j - 1}$

- A megfigyelt Y értékek visekedését leíró modellt a következő formában fejezhetjük ki:

$$y_{ij} = \mu_j + \varepsilon_{ij}, \quad i = 1, 2, 3, \ldots, n_j, \quad j = 1, 2, 3, \ldots, k,$$

ahol az ε_{ij} egy véletlen valószínűségi változó 0 várható értékkel és σ^2 szórásnégyzettel. Feltételezzük, hogy az ε_{ij}-k függetlenek. Homogén szórás feltételezése mellett a μ_j legjobb torzítatlan becslése $\overline{y}_{\bullet j}$, így a modell a kövezkező alakban írható fel:

$$y_{ij} = \overline{y}_{\bullet j} + (y_{ij} - \overline{y}_{\bullet j}), \quad i = 1, 2, 3, \ldots, n_j, \quad j = 1, 2, 3, \ldots, k,$$

ahol $\overline{y}_{\bullet j}$-vel becsültük μ_j-t és $(y_{ij} - \overline{y}_{\bullet j}) = e_{ij}$-vel becsültük ε_{ij}-t. Ebből a modellből y_{ij} becsült értékeit az $\hat{y}_{ij} = \overline{y}_{\bullet j}$ egyenletekkel kapjuk meg.

Négyzetösszegek és a kétfajta szórásnégyzet becslés: A négyzetösszegekből készítjük el a kétfajta szórásbecslést. Y teljes szórását a mintán az SST-vel (sum of squares total) jelölt négyzetösszeg adja meg:

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_{\bullet\bullet})^2$$

Ezzel Y-nak a nagy átlag körüli szórását kapjuk meg.
Az osztályok átlagainak szórásából készült szórásbecslést MSA-val jelöljük és az SSA-val (sum of squares among) jelölt négyzetösszegből fogjuk számolni:

$$SSA = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet})^2 = \sum_{j=1}^{k} n_j \ (\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet})^2$$

Innen az átlagok szórásnégyzetének becslése:

$$MSA = \frac{SSA}{k-1} = \hat{\sigma}^2$$

Az osztályonkénti szórásbecslések átlagaként megkapott szórásnégyzet becslést MSW-vel fogjuk jelölni és az SSW-vel (sum of squares within) jelölt négyzetösszeg segítségével fogjuk kiszámolni.

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\overline{y}_{ij} - \overline{y}_{\cdot\cdot})^2$$

Innen kapjuk a második szórásnégyzet becslésünk:

$$MSW = \frac{1}{k} \sum_{j=1}^{k} \frac{\sum_{i=1}^{n_j} (y_{ij} - \overline{y}_{\bullet\bullet})^2}{n-1} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_{\bullet\bullet})^2}{k(n-1)} = \frac{SSW}{N-k} = \hat{\sigma}^2.$$

A négyzetösszegek számítási módjából következik a következő azonosság is:

$$SST = SSW + SSA$$

F-próba és az ANOVA-tábla: A szórásbecslések egyezését F-próbával ellenőrizzük. A próbastatisztikánk ekkor:

$$F = \frac{MSA}{MSW} = \frac{MSA/(k-1)}{MSW/(N-k)} \sim F_{k-1,N-k}$$

A próba elutasít, ha $F > F_{krit}$. Adott α szignifikancia szint mellett, F_{krit} az F eloszlás $1-\alpha$-s kvantilisét jelenti. Az ehhez a modellhez tartozó ANOVA-tábla az alábbi módon fog kinézni:

ANOVA-tábla				
Szórásbecslés típusa	Szabadsági fok	A megfelelő négyzetösszeg	Becslések szórásnégyzetre	F-próba
osztályok átlagainak szórása	k-1	SSA	MSA	$\frac{MSA}{MSW}$
osztályok szórásbecsléseinek átlaga	N-k	SSW	MSW	
teljes szórás a mintán	N-1	SST		

4. Példa. Nézzünk egy példát a fent leírtakra: Egy piaci kutató cég több városban is figyeli számos család tejtermékek vásárlására fordított pénzét. A családok minden héten leadnak egy jelentést, amiben többek között a héten tejtermékekre költött pénzösszeg is szerepel. Egy tanulmányt szeretnénk készíteni, ami négy tejtermék vásárlására ösztönző reklám hatékonyságát vizsgálja. A négy reklámot helyi TV-csatornákon, több különböző városban, öt földrajzi területen sugározták. Mind az öt földrajzi területen négy várost választottak ki, mind a négy város a többiekétől különböző reklámot kapott és mind a négy városból hat család adatait választották ki. Miután a reklámokat két hónapig adták, a hat család tejtermékfogyasztását lejegyezték dollárban, ez látható a 21. oldalon található táblázatban. A megfelelő számolások elvégzése után az ANOVA-tábla a következőképpen fog kinézni:

ANOVA-tábla				
Szórásbecslés típusa	Szabadsági fok	A megfelelő négyzetösszeg	Becslések szórásnégyzetre	F-próba
osztályok átlagainak szórása	3	4585.680	1528.56	3.16
osztályok szórásbecsléseinek átlaga	116	56187.444	484.375	
teljes szórás a mintán	119	60773.124		

A 3.16 értékű, 3 és 116 szabadsági fokokkal rendelkező F-statisztikához tartozó p-érték: 0.0275. Ha a szignifikancia szintet 5%-nak választjuk, akkor mivel $p < 0.05$ elvethetjük a nullhipotézist, azaz elfogadjuk, hogy van valamekkora különbség az átlagok között.

Régió	1.reklám	2.reklám	3.reklám	4.reklám	Család mérete
1	12.35	21.86	14.43	21.44	1
1	20.52	42.17	22.26	31.21	2
1	30.85	49.61	23.99	40.09	3
1	39.35	63.65	36.98	55.68	4
1	48.87	73.75	42.13	65.81	5
1	58.01	85.95	54.19	76.61	6
2	28.26	13.76	14.44	30.78	1
2	37.67	24.59	29.63	45.75	2
2	44.70	37.30	38.27	56.37	3
2	57.54	49.53	51.59	70.19	4
2	67.57	59.25	59.09	79.81	5
2	77.70	67.68	71.69	94.23	6
3	10.97	0.000	2.900	6.460	1
3	26.70	2.410	17.28	18.61	2
3	36.81	16.10	19.62	30.14	3
3	51.34	22.71	29.53	39.12	4
3	62.69	30.19	38.57	51.15	5
3	72.68	41.64	48.20	59.11	6
4	0.000	11.90	4.480	27.62	1
4	4.520	27.75	18.01	42.63	2
4	13.71	42.22	21.96	59.20	3
4	27.91	56.06	34.42	74.92	4
4	38.57	66.16	40.14	92.37	5
4	42.71	78.71	57.06	98.02	6
5	13.11	8.000	10.90	14.36	1
5	16.89	18.27	28.22	26.37	2
5	27.99	27.72	38.62	34.15	3
5	36.35	42.04	48.31	54.02	4
5	48.85	48.50	60.23	59.90	5
5	61.97	59.92	71.39	74.79	6

One-way MANOVA

A One-way MANOVA esetén minden mefigyelés adatok egy vektora lesz. Például a megfigyelt vektor komponensei lehetnek a levegőt alkotó gázok százalékos arányai. A levegő összetételét k különböző városban vizsgáljuk (k darab osztály), n napon keresztül (esetek). Az adatelemzés hasonlóan zajlik, mint a One-way ANOVA esetén, csak most vektorokkal kell dolgoznunk.

A következő adatmátrixban minden osztály megfigyelésszáma ugyanannyi, de a számolást úgy írjuk le, hogy különböző elemszámra is működjön.
Az adatmátrix az alábbi módon néz ki:

1.osztály	**2.osztály**	**3.osztály**	...	**k.osztály**
$\mathcal{N}(\underline{\mu}_1, \Sigma)$	$\mathcal{N}(\underline{\mu}_2, \Sigma)$	$\mathcal{N}(\underline{\mu}_3, \Sigma)$	...	$\mathcal{N}(\underline{\mu}_k, \Sigma)$
$\underline{y}_{1,1}$	$\underline{y}_{1,2}$	$\underline{y}_{1,3}$	$\cdots$	$\underline{y}_{1,k}$
$\underline{y}_{2,1}$	$\underline{y}_{2,2}$	$\underline{y}_{2,3}$	$\cdots$	$\underline{y}_{2,k}$
$\underline{y}_{3,1}$	$\underline{y}_{3,2}$	$\underline{y}_{3,3}$	$\cdots$	$\underline{y}_{3,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\underline{y}_{n,1}$	$\underline{y}_{n,2}$	$\underline{y}_{n,3}$	$\cdots$	$\underline{y}_{n,k}$

Minden osztályban $\underline{Y}$ eloszlását normálisnak tekintjük azonos Σ variancia-kovariancia mátrixszal, legfeljebb a osztályok várható érték vektora különbözhet. Ha az $i.$ várható érték vektor $k.$ komponensét $(\underline{\mu}_i)_k$-val jelöljük, a hipotézisek az alábbi módon alakulnak:

$$H_0 : \underline{\mu}_1 = \underline{\mu}_2 = \underline{\mu}_3 = \cdots = \underline{\mu}_k$$
$$H_1 : \exists\, i, j, k, \text{ hogy } i \neq j - re : (\underline{\mu}_i)_k \neq (\underline{\mu}_j)_k$$

A hipotézisvizsgálat menete: Egydimenzióban a szórást becsültük kétféleképp, itt a variancia-kovariancia mátrixal tesszük ugyanezt. Az első becsült mátrixot hibamátrixnak, a másodikat hipotézismátrixnak nevezzük és rendre E-vel és H-val jelöljük őket. Tegyük fel, hogy a j.osztály megfigyelésszáma n_j. A hibamátrix becslése (ez az egydimenziós esetbeli MSW-nek megfelelő becslés):

$$E = \begin{pmatrix} SP_{w,1,1} & SP_{w,1,2} & SP_{w,1,3} & \dots & SP_{w,1,k} \\ SP_{w,2,1} & SP_{w,2,2} & SP_{w,2,3} & \dots & SP_{w,2,k} \\ SP_{w,3,1} & SP_{w,3,2} & SP_{w,3,3} & \dots & SP_{w,3,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ SP_{w,n,1} & SP_{w,n,2} & SP_{w,n,3} & \dots & SP_{w,n,k} \end{pmatrix}$$

$$SP_{w,p,q} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left[\left(\underline{y}_{i,j} \right)_p - \left(\underline{\overline{y}}_j \right)_p \right] \left[\left(\underline{y}_{i,j} \right)_q - \left(\underline{\overline{y}}_j \right)_q \right].$$

Jelölések: $\underline{\overline{y}}$ az összes megfigyelt vektor átlagvektora, $\underline{\overline{y}}_j$ a j. osztály vektorainak átlagvektora.

A hipotézismátrix becslése (ez az egydimenziós esetbeli MSA-nak megfelelő becslés):

$$H = \begin{pmatrix} SP_{a,1,1} & SP_{a,1,2} & SP_{a,1,3} & \dots & SP_{a,1,k} \\ SP_{a,2,1} & SP_{a,2,2} & SP_{a,2,3} & \dots & SP_{a,2,k} \\ SP_{a,3,1} & SP_{a,3,2} & SP_{a,3,3} & \dots & SP_{a,3,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ SP_{a,n,1} & SP_{a,n,2} & SP_{a,n,3} & \dots & SP_{a,n,k} \end{pmatrix},$$

$$SP_{a,p,q} = \sum_{j=1}^{k} n_j \left[\left(\underline{\overline{y}}_j \right)_p - \left(\underline{\overline{y}} \right)_p \right] \left[\left(\underline{\overline{y}}_j \right)_q - \left(\underline{\overline{y}} \right)_q \right].$$

Wilks próbája a variancia-kovariancia mátrixok egyenlőségére: A többdimenziós próbastatisztikánk, Wilks-féle Λ eloszlású lesz, ez felel meg az egydimenziós F-próbának. Ebben az esetben, mivel két mátrix hányadosa nem definiált, az egyik becsült mátrix inverzét vesszük és megszorozzuk a másikkal. Tudjuk, hogy H_0 igaz volta esetén a két becsült mátrix nem térhet el szignifikánsan, ezért ennek a szorzatnak nem szabad nagyon eltérnie az egységmátrixtól. A próbastatisztika az alábbi módon számolható:

$$\Lambda = \frac{det(E)}{det(E+H)} = \frac{1}{I + E^{-1}H} \sim \Lambda_{k-1,N-k}$$

Itt I az egységmátrix, és ha $E^{-1}H$ rangja r, sajátértékeit pedig λ_i-vel jelöljük, akkor:

$$\Lambda = \prod_{i=1}^{r} \frac{1}{1+\lambda_i}$$

Ebben az esetben, viszont akkor utasítunk el, ha $\Lambda < \Lambda_{krit}$, tehát, ha a kapott próbastatisztika értékünk kicsi.

Roy teszt: A tesztstatisztika, ha $E^{-1}H$ legnagyobb sajátértékét λ_1-vel jelöljük:

$$\vartheta = \frac{\lambda_1}{1+\lambda_1}$$

Ezt már $\vartheta > \vartheta_{krit}$-re utasítjuk el.

Pillai-Bartlett teszt: Ha s az $E^{-1}H$ rangja:

$$V^{(s)} = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$$

Itt is $V^{(s)} > V_\alpha^{(s)}$-ra utasítjuk el a nullhipotézist, ahol $V_\alpha^{(s)}$ a Pillai eloszlás $1-\alpha$-s kvantilise.

Lawley-Hotelling teszt: Ha s az $E^{-1}H$ rangja:

$$U^{(s)} = \sum_{i=1}^{s} \lambda_i$$

A nullhipotézist elutasítjuk, ha $U^{(s)} > U_\alpha^{(s)}$.

5. Példa. Nézzünk egy példát a MANOVA-ra is. Az `R` programnyelv beépített `skulls` adatszettjét használtuk. Ebben az adathalmazban 150 koponya adatai szerepelnek, melyeket a következő korszakok szerint öt osztályba soroltak: Kr.e. 4000, Kr.e. 3300, Kr.e. 1850, Kr.e. 200 és Kr.u. 150 (a kódban a kor változó neve: `epoch`). Minden koponyához tartozik négy mérési adat, ezek különböző hosszmértékek. A programban az adathalmazt `data`-val, a hosszakat `mb`, `bh`, `bl`, `nh`-vel jelöljük. Azt szeretnénk megtudni, hogy az évek során változott-e a koponya mérete. Tehát nullhipotézisünk az lesz, hogy a koponya mérete nem változott a fent említett korokban, míg az alternatív hipotézis, hogy a koponya mérete változott. Az alábbiakban a MANOVA `R`-beli végrehajtását mutatjuk meg a fent ismertetett feladatra:

```
Y1=cbind(data$mb,data$bh,data$bl,data$nh)
skull.manova=manova(Y1 ~ data$epoch, data)
summary(skull.manova)
```

Az így kapott eredmény:

```
            Df  Pillai  approx F  num Df  den Df     Pr(>F)
data$epoch   4 0.35331     3.512     16     580     4.675e-06
Residuals  145
```

Az F-statisztikához tartozó P-érték nagyon kicsi, ezért elvethetjük H_0-t, vagyis mondhatjuk, hogy a koponya mérete különbözött az említett korokban. A kimeneten látható, hogy a Pillai-Bartlett tesztet használja alapértelmezettként az `R`. Az alábbiakban látható, hogy a többi teszt használata esetén sem kapunk más eredményt, minden teszt elutasítja a nullhipotézist.

```
vek=c("Wilks", "Roy", "Hotelling-Lawley")
for(i in 1:length(vek))
{print(summary(skull.manova, test=vek[i]))}
```

A másik három teszt eredményeit táblázatba foglaltuk, a szabadsági fokok minden teszt esetén 4 és 145.

	tesztstatisztika érték	approx F	num Df	den Df	Pr(>F)
Wilks	0.66359	3.9009	16	434.45	7.01e-07
Roy	0.4251	15.41	4	145	1.588e-10
Hotelling-Lawley	0.48182	4.231	16	562	8.278e-08

A táblázatból látható, hogy a P-értékek minden esetben nullához nagyon közeliek, tehát minden teszt elutasítja a nullhipotézist.

5. Véletlenített tesztek

Az eddig leírtakban feltettük, hogy az osztályok homogén szórással rendelkeznek és normális eloszlásúak. Ez a két feltevés azonban sok esetben nem teljesül. Ilyen esetekre is léteznek nem parametrikus alternatívái az F-próbának, továbbá léteznek módszerek, melyekkel elkészíthetjük a kovariancia mátrixok becsléseit osztályonként és ezeknek a becsléseknek különböző tulajdonságait is megkaphatjuk (pl. szórás, ferdeség).

5.1. Jackknife

Bevezetés:

A jackknife módszer egy úgynevezett újramintavételezési technika, amivel nem parametrikus becsléseket kaphatunk egy paraméterre, vagy a paraméterbecslés szórására és torzítására. Egy paraméterbecslés szórásának jackknife becslését úgy kapjuk, hogy szisztematikusan mindig kihagyunk egy megfigyelést az adatinkból és így számoljuk az adott paraméter becslését, majd miután ezt az összes lehetséges módon megtettük (például n megfigyelés esetén n db becslést számolunk)átlagoljuk ezeket a becsléseket, és innen számoljuk az átlagtól vett átlagos eltérést.

A módszer:

Legyenek $X_1, X_2, X_3, \ldots, X_n$ független, azonos eloszlású valószínűségi változók. Tegyük fel, hogy $\rho(X_1, X_2, X_3, \ldots, X_n)$ a statisztika, melynek szórását szeretnénk vizsgálni. Legyen $\rho_i = \rho(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, ez a statisztikánk értéke, ha X_i-t kihagyjuk a mintából. Jelöljük $\hat{\rho}$-al és $\hat{\rho}_i$-al a valószínűségi változók megfigyelt értékeiből számolt becslését a fenti statisztikáknak. A statisztikára is tudunk jackknife becslést számolni, ezt $\hat{\rho}_\bullet$-vel jelöljük és az alábbi módon számoljuk:

$$\hat{\rho}_\bullet = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}_i$$

A statisztika szórásának jackknife becslését $\hat{\sigma}_{\hat{\rho},J}$-vel jelöljük és a következőképpen fogjuk számolni:

$$\hat{\sigma}_{\hat{\rho},J} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\rho}_i - \hat{\rho}_\bullet)^2$$

Értelemszerűen, ha a fenti kifejezésből gyököt vonunk, megkapjuk a statisztika standard hibájának jackknife becslését. A jackknife módszerrel lehet egy becsült paraméter ($\hat{\rho}$) torzítását is becsülni a következő módon:

$$\widehat{bias}_{\hat{\rho},J} = (n-1)(\hat{\rho}_\bullet - \hat{\rho})$$

5.2. Bootstrap

Bevezetés

A bootstrap, a jackknife-hoz hasonlóan egy olyan újramintavételezési eljárás, melynek során egy paraméterbecslés különböző tulajdonságait (például a szórását) tudjuk becsülni, úgy , hogy ezeket a tulajdonságokat egy, az eredetit közelítő eloszlásból vett mintán mérjük. Az egyik leggyakrabban használt közelítő eloszlás, a megfigyelt értékekhez tartozó tapasztalati eloszlásfüggvény.

Nézzünk egy gyors példát, hogy jobban értelmezhessük az eljárást: Tegyük fel, hogy szeretnénk megtudni az összes ember átlagtömegét. Nyilvánvalóan ehhez nem tudunk több milliárd mérést végezni, ezért valami más ötletre van szükségünk. Ha ennyi mérés nem is áll rendelkezésünkre, azért tudunk méréseket végezni, csak kisebb nagyságrendben (száz mérést például). Ebből a százas mintából csak egy becslését kapjuk az átlagtömegnek, ami még mindig nem elegendő. Ahhoz, hogy a Föld egész népességére mondhassunk valamit, szükségünk van az előbb számolt átlag szóródásának mértékére.

A legegyszerűbb bootstrap módszer veszi a száz megfigyelt tömegértéket és visszatevéses mintavétellel ezekből az adatokból új százelemű mintát készít. Ez ugyanaz, mintha a száz megfigyelt érték által meghatározott tapasztalati eloszlásfüggvény alapján generálna egy véletlen mintát, ahol minden megfigyelt értéket ugyanannyi, 0.01-os valószínűséggel választana ki.

Az így készült új mintát bootstrap mintának nevezzük. Ezt a folyamatot sokszor elvégezve (itt a sok alatt több ezret értünk) minden újonnan kapott bootstrap mintára kiszámoljuk az átlagot, amit itt bootstrap becslésnek nevezünk. A rendelkezésre álló bootsrap becslésekből már tudunk egy szórásbecslést, vagy egyéb statisztika jellemzőket (pl. ferdeség, lapultság) adni az átlagra.

Fontos még hozzátenni, hogy ez az eljárás bármilyen statisztikára alkalmazható, nem csak a szórásra és az átlagra.

A módszer:

Legyen az adatunk az előbbihez hasonlóan egy n elemű véletlen minta egy ismeretlen F eloszlásból:

$$X_1, X_2, X_3, \ldots, X_n \sim F$$

A bootstrap módszer, tehát a következő négy lépésből tevődik össze:

1.lépés: Az X_i megfigyelt értékét jelöljük x_i-vel. Legyen $\hat{F}_n$ a megfigyelt értékekhez tartozó tapasztalati eloszlás, mely minden x_i-hez $(i = 1, 2, \ldots, n)$ azonos, $1/n$-es valószínűséget rendel. A megfigyelt értékekből konstruáljuk meg az $\hat{F}_n$ tapasztalati eloszlást.

2.lépés: Legyen $X_1^*, X_2^*, X_3^*, \ldots, X_n^*$ egy véletlen minta $\hat{F}_n$-ből, azaz:

$$X_1^*, X_2^*, X_3^*, \ldots, X_n^* \sim \hat{F}_n$$

Az $\hat{F}_n$ tapasztalati eloszlásból visszatevéses mintavétellel készítsük el az $X_1^*, X_2^*, X_3^*, \ldots, X_n^*$ bootstrap mintát. Ez azt jelenti, hogy n darab véletlen, egymástól független húzást végzünk visszatevéssel az $x_1, x_2, x_3, \ldots, x_n$ halmazból, úgy, hogy minden x_i kihúzási valószínűsége $1/n$. Ezután kiszámoljuk a bennünket érdeklő paraméter bootstrap becslését az előbb kapott bootstrap mintából:

$$\hat{\rho}_B = \hat{\rho}(X_1^*, X_2^*, X_3^*, \ldots, X_n^*).$$

3.lépés: Végezzük el a második lépést N-szer, ahol az N egy többezres nagyságrendű szám. Ily módon kapni fogunk N darab független bootstrap-becslést: $\hat{\rho}_{B,1}, \hat{\rho}_{B,2}, \hat{\rho}_{B,3} \ldots \hat{\rho}_{B,N}$. Ezek átlagaként számolható a szóban forgó paraméter bootstrap becslése, amit $\hat{\rho}_{B,\bullet}$-el jelöltünk:

$$\hat{\rho}_{B,\bullet} = \tfrac{1}{N} \sum_{i=1}^{N} \hat{\rho}_{B,i}$$

A $\hat{\rho}$ szórásának bootstrap becslése $(\hat{\sigma}_{\hat{\rho},B})$ innen:

$$\hat{\sigma}_{\hat{\rho},B} = \left(\sum_{i=1}^{N} (\hat{\rho}_{B,i} - \hat{\rho}_{B,\bullet})^2 \Big/ (N-1) \right)^{\frac{1}{2}}.$$

6. Példák

6.1. Korreláció mátrix generálása

Az első példánkban generálni fogunk egy négydimenziós korreláció mátrixot az első részben leírtak alapján. A folyamatot lépésenként fogjuk ismertetni, a számításokat az R programnyelvben végezzük. Miután a generálás kész, ellenőrizni fogjuk, hogy a kapott korreláció mátrix minden követelménynek megfelel-e.
Elsőként a korrelatív szögek négydimenziós, szigorúan alsó háromszög mátrixát generáljuk. Ezt a mátrixot ϑ-val jelöljük, elemei 0 és π közé eső korrelatív szögek:

```
theta=matrix(nrow = 4, ncol=4)
k=lower.tri(theta)
for (i in 1:4)
{for(j in 1:4)
  if(k[i,j]==TRUE) {theta[i,j] <- runif(1,0,pi)}
  else{theta[i,j]=0}}
```

Az ebből kapott ϑ mátrix:

$$\vartheta = \begin{pmatrix} 0.0000000 & 0.000000 & 0.000000 & 0 \\ 0.3147156 & 0.000000 & 0.000000 & 0 \\ 1.5447925 & 3.089837 & 0.000000 & 0 \\ 2.6889302 & 2.360183 & 1.374924 & 0 \end{pmatrix}$$

Ebben a mátrixban a vektorok komponenseinek korrelatív szögeit láthatjuk radiánban.

Most a korrelatív szögek mátrixából kiszámoljuk a 4×4-es, alsó háromszög B mátrixot, a (3)-ban ismertetett módon:

```
B=matrix(nrow=4,ncol = 4)
db1 <- db2 <- 1
for (i in 1:4)
{for (j in 1:4)
  {if(i==1 & j==1){B[i,j] <- 1}
   if(i>=2 & j==1){B[i,j] <- cos(theta[i,j])}
   if(i==j & i>=2 & j<=4)
   {for(k in 1:(j-1)){db1=db1*sin(theta[i,k])}
    B[i,j] <- db1}
   if(2<=j & j<=i-1)
   {for(k in 1:(j-1)){db2=db2*sin(theta[i,k])}
    B[i,j] <- cos(theta[i,j])*db2}
   if(i+1<=j & j<=4){B[i,j] <- 0}}}
```

Ekkor a B mátrix az alábbi módon fog kinézni:

$$B = \begin{pmatrix} 1.00000000 & 0.0000000 & 0.00000000 & 0.00000000 \\ 0.95088447 & 0.3095460 & 0.00000000 & 0.00000000 \\ 0.02600093 & -0.9983233 & 0.01600832 & 0.00000000 \\ -0.89928583 & -0.3103871 & 0.02621032 & 0.00483668 \end{pmatrix}$$

Innen a korreláció mátrix főátlóbeli elemeit 1-re rögzítve, majd a többi elemre a $P = BB^T$ képlet használatával megkapjuk a generált korreláció mátrixunkat:

$$P = \begin{pmatrix} 1.0000000 & 0.9508845 & 0.0260009 & -0.8992858 \\ 0.9508845 & 1.0000000 & -0.2843032 & -0.9511961 \\ 0.0260009 & -0.2843032 & 1.0000000 & 0.2869042 \\ -0.8992858 & -0.9511961 & 0.2869042 & 1.0000000 \end{pmatrix}$$

A mátrixról ránézésre látszik, hogy szimmetrikus, a főátlójában egyesek vannak, és a főátlón kívüli elemei mind a $[-1, 1]$ intervallumról vesznek fel értékeket. A pozitív definit tulajdonságot az `R`-beli `eigen` paranccsal ellenőriztük. A legkisebb sajátértéket vizsgáltuk, ami ebben az esetben: 0.0001272446, mivel ez pozitív $\Rightarrow$ a mátrix pozitív szemidefinit, tehát sikerült generálnunk egy minden követelménynek megfelelő korreláció mátrixot.

6.2. Jackknife módszer alkalmazása

A már ismert `skulls` adatszettel fogunk dolgozni. A koponyák becsült kora szerint alakítjuk ki az öt osztályt, és az egyes osztályokban a megfigyelt vektoraink négydimenziósak lesznek. Ezek szerint a becsült variancia-kovariancia mátrixok 4×4-esek lesznek minden osztályban. Az egyes osztályokhoz tartozó variancia-kovariancia mátrixok becslését úgy készítjük el, hogy az adott osztályban egyesével kihagyjuk a megfigyelt vektorokat és így számolunk egy becslést a variancia-kovariancia mátrixra. Mivel minden osztályban harminc megfigyelés áll rendelkezésre, osztályonként harminc becsült variancia-kovariancia mátrixunk lesz. Ezt a harminc becslést átlagolva kapjuk az osztályokhoz tartozó variancia-kovariancia mátrixok jackknife becslését:

```
pr.list<-function(L) lapply(L,function(x,d=2)round(x,d))
jVlist<-list()
csoport<-names(table(data[,1]))

for(idx in csoport)
 { A <- data[data[,1]==idx,]
   n <- nrow(A)
   V <- matrix(0,4,4)
   for(i in 1:n)
    V <- V+cov(A[-i,-1])*(n-1)/n
   V <- V/n
   jVlist<-c(jVlist,list(V))
  }
names(jVlist)<-csoport
pr.list(jVlist)
```

A kimenetben szereplő öt mátrixból a Kr.e.1850-es és a Kr.e.4000-es osztály variancia-kovariancia mátrixainak becsléseit láthatjuk alább:

Kr.e.1850:

$$\begin{pmatrix} 11.72 & 0.76 & -0.75 & 0.87 \\ 0.76 & 23.96 & 3.47 & -0.09 \\ -0.75 & 3.47 & 20.03 & 1.61 \\ 0.87 & -0.09 & 1.61 & 12.18 \end{pmatrix}$$

Kr.e.4000:

$$\begin{pmatrix} 25.43 & 4.01 & 0.44 & 7.00 \\ 4.01 & 19.31 & -0.77 & 0.38 \\ 0.44 & -0.77 & 33.47 & -1.86 \\ 7.00 & 0.38 & -1.86 & 7.38 \end{pmatrix}$$

Ezután következik a lényegi rész. Mivel ez a kettő csak egy-egy becslése a variancia-kovariancia mátrixokra és minden becslés más eredményeket fog adni, jó lenne tudni mekkora a becsült kovariancia mátrix elemeinek szórásainak becslései. A következő R-kód segítségével megkaphatjuk ezeket a szórásbecsléseket:

```
jVVlist<-list()
csoport<-names(table(data[,1]))

for(idx in csoport)
 {
   A <- data[data[,1]==idx,]
   n <- nrow(A)
   Vm <- matrix(0,0,16)
   for(i in 1:n)
    Vm <- rbind(Vm,as.numeric(cov(A[-i,-1])*(n-1)/n))
    mVV <- matrix(apply(Vm,2,sd),4,4)
    jVVlist<-c(jVVlist,list(mVV))
  }
names(jVVlist)<-csoport
pr.list(jVVlist)
```

A két fent bemutatott variancia-kovariancia mátrixbecslés szórásainak becslései:

Kr.e.1850:

$$\begin{pmatrix} 0.53 & 0.47 & 0.55 & 0.44 \\ 0.47 & 1.22 & 0.77 & 0.77 \\ 0.55 & 0.77 & 0.87 & 0.58 \\ 0.44 & 0.77 & 0.58 & 0.63 \end{pmatrix}$$

Kr.e.4000:

$$\begin{pmatrix} 1.20 & 0.73 & 0.95 & 0.66 \\ 0.73 & 1.24 & 0.85 & 0.42 \\ 0.95 & 0.85 & 1.68 & 0.63 \\ 0.66 & 0.42 & 0.63 & 0.35 \end{pmatrix}$$

6.3. Bootstrap alkalmazása

Ebben az esetben is az R-beli `skulls` adatszettel fogunk dolgozni. Az osztályok szétválasztása után elkészítjük a kovariancai mátrixok hagyományos becslését. Ezt követően elkészítjük az egyes osztályokhoz tartozó kovariancia mátrixok bootstrap becslését oly módon, hogy az adott osztály 30 darab négydimenziós megfigyelt vektorából húzunk 30-at, visszatevéssel (ez lesz a bootstrap mintánk), majd a bootstrap mintából kiszámoljuk az osztály kovariancia mátrixának becslését. Ezt az újramintavételezést végrehajtjuk 1000-szer, majd az 1000 darab kovariancia mátrixbecslést átlagoljuk. Így minden osztály kovariancia mátrixára kapunk egy bootstrap becslést.

Ezt követően kiszámoljuk a bootstrap módszerrel becsült kovariancia mátrixok elemenkénti szórását.

Végül kiszámoljuk a bootstrap-el becsült kovariancia mátrixok relatív hibáját a szórás mértékében (elemenként) úgy, hogy vesszük a hagyományos és bootstrap becslés abszolút értékben vett eltérését és ezt az eltérést elosztjuk az előzőekben kiszámolt elemenkénti szórással.

Az osztályokat a következőképp választottuk szét:

```
p <- 4
m <- 5
n <- 30

skulls.list <- list()
for(k in levels(skulls[,1]))
skulls.list <- c(skulls.list,list(skulls[skulls[,1]==k,(1:p)+1]))
names(skulls.list) <- levels(skulls[,1])
skulls.list ,
```

ahol `p` a minta dimenziója, `m` az osztályok száma, `n` az osztályok mérete. Ezután elkészítettük a kovariancia mátrixok hagyományos becslését, amit a `cov.est` nevű listában tároltunk:

```
cov.est<-lapply(skulls.list,cov)
```

A Kr.e 4000 és a Kr.e 1850-es osztályok kovariancia mátrixának hagyományos becslései:

Kr.e.1850:

$$\begin{pmatrix} 12.12 & 0.79 & -0.77 & 0.90 \\ 0.79 & 24.79 & 3.59 & -0.09 \\ -0.77 & 3.59 & 20.72 & 1.67 \\ 0.90 & -0.09 & 1.67 & 12.6 \end{pmatrix}$$

Kr.e.4000:

$$\begin{pmatrix} 26.31 & 4.15 & 0.45 & 7.25 \\ 4.15 & 19.97 & -0.79 & 0.39 \\ 0.45 & -0.79 & 34.63 & -1.92 \\ 7.25 & 0.39 & -1.92 & 7.64 \end{pmatrix}$$

Ezt követően definiáltunk egy függvényt `boot.cov` néven, melynek bemenete megfigyelt vektorok egy halmaza. Erre a vektorhalmazra kiszámolja a kovariancia mátrix bootstrap becslését és a becsült mátrixok elemenkénti szórásának becslését is.

```
N<-1000

boot.cov<-function(x)
{
  work <- list()
  for(k in 1:N)
  { y <- x[sample(nrow(x),rep=TRUE),]
    work <- c(work,list(cov(y)))      }
  M <- matrix(unlist(work),p^2)
  return(list(b.est=matrix(apply(M,1,mean),p),
              b.sd=matrix(apply(M,1,sd),p)    )) ,
}
```

ahol `N` adja meg, hogy hányszor végezzünk újramintavételezést.
Most, hogy az adatszettünk szét van bontva osztályokra és kész a szükséges függvény is, alkalmazzuk a függvényt az adatainkra:

```
set.seed(123)
cov.boot<-lapply(skulls.list,boot.cov)
```

A Kr.e 4000 és a Kr.e 1850-es osztályok kovariancia mátrixának bootstrap becslései, két tizedesjegyre kerekítve:

Kr.e.1850:

$$\begin{pmatrix} 11.84 & 0.75 & -0.79 & 0.74 \\ 0.75 & 23.55 & 3.21 & -0.04 \\ -0.79 & 3.21 & 19.96 & 1.58 \\ 0.74 & -0.04 & 1.58 & 12.07 \end{pmatrix}$$

Kr.e.4000:

$$\begin{pmatrix} 25.67 & 4.19 & 0.51 & 7.02 \\ 4.19 & 19.64 & -0.87 & 0.39 \\ 0.51 & -0.87 & 33.37 & -1.75 \\ 7.02 & 0.39 & -1.75 & 7.33 \end{pmatrix}$$

A fenti két becslés elemenkénti szórásbecslései, ugyanúgy a bootstrap módszerrel, szintén két tizedesjegyre kerekítve:

Kr.e.1850:

$$\begin{pmatrix} 2.74 & 2.48 & 2.82 & 2.29 \\ 2.48 & 6.35 & 3.70 & 3.94 \\ 2.82 & 3.70 & 4.42 & 3.02 \\ 2.29 & 3.94 & 3.02 & 3.20 \end{pmatrix}$$

Kr.e.4000:

$$\begin{pmatrix} 6.03 & 3.79 & 4.92 & 3.39 \\ 3.79 & 6.21 & 4.50 & 2.20 \\ 4.92 & 4.50 & 8.67 & 3.09 \\ 3.39 & 2.20 & 3.09 & 1.81 \end{pmatrix}$$

Most következik az elemenkénti relatív hiba számolása, a következő módon:

```
rel.err <- list()
for(k in 1:m)
{
  H <- cov.est[[k]]
  B <- cov.boot[[k]]
  rel.err<-c(rel.err, list(abs(H-B[[1]])/B[[2]]))
}
names(rel.err) <- levels(skulls[,1])
```

Ez után a következő függvénnyel kerekítjük három tizedesjegyre a relatív hiba mátrixokat:

```
pr.list <- function(L,ts) lapply(L,function(x,d=ts)round(x,d))
pr.list(rel.err,3)
```

A Kr.e 4000 és a Kr.e 1850-es osztályok elemenkénti relatív hibájának becslései, három tizedesjegyre kerekítve:

Kr.e.1850:

$$\begin{pmatrix} 0.1027 & 0.0168 & 0.0067 & 0.0704 \\ 0.0168 & 0.1940 & 0.1035 & 0.0126 \\ 0.0067 & 0.1035 & 0.1736 & 0.0314 \\ 0.0704 & 0.0126 & 0.0314 & 0.1641 \end{pmatrix}$$

Kr.e.4000:

$$\begin{pmatrix} 0.1058 & 0.0088 & 0.0117 & 0.0669 \\ 0.0088 & 0.0532 & 0.0167 & 0.0001 \\ 0.0117 & 0.0167 & 0.1446 & 0.0565 \\ 0.0669 & 0.0001 & 0.0565 & 0.1678 \end{pmatrix}$$

Ezeken a mátrixokon azt láthatjuk, hogy a bootstrap becslések elemenkénti relatív hibája a legtöbb esetben nem haladja meg a 10%-ot.

7. Irodalomjegyzék

1. J.D. Jobson (1991): Applied Multivariate Data Analysis pp. 399-533.

2. Bradley Efron & Gail Gong (1983): A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, The American Statistician, Vol. 37, No. 1, pp. 36-48

3. Kawee Numpacharoen, Amporn Atsawarungruangkit (2012): Generating Correlation Matrices Based on the Boundaries of Their Coefficients, PLoS ONE 7(11): e48902, doi:10.1371/journal.pone.0048902

4. Avery I. McIntosh: The Jackknife Estimation Method, https://arxiv.org/abs/1606.00497

5. Bradley Efron (1979): Bootstrap methods: another look at the jackknife, The annals of Statistics 1979, pp. 1-26

6. B.Effron & C.Stein (1981): The Jackknife estimate of variance, The Annals of Statistics 1981, vol 9 , No. 3, 586-596

7. Rebonat R. & Jäckel P. (2000): The most general methodology for creating a valid correlation matrix for risk management and option pricing purposes. J Risk 2: 17-27