Többváltozós statisztikai módszerek

Bolla, Marianna

Krámli, András

Nagy-György, Judit

Többváltozós statisztikai módszerek

Bolla, Marianna
Krámli, András
Nagy-György, Judit

Publication date 2013

TÁMOP-4.1.2.A/1-11/1 MSc Tananyagfejlesztés

Interdiszciplináris és komplex megközelítésű digitális tananyagfejlesztés a természettudományi képzési terület mesterszakjaihoz

Tartalom

Előszó

A jelen digitális tananyag a TÁMOP-4.1.2.A/1-11/1-2011-0025 számú, "Interdiszciplináris és komplex megközelítésű digitális tananyagfejlesztés a természettudományi képzési terület mesterszakjaihoz" című projekt részeként készült el.

A projekt általános célja a XXI. század igényeinek megfelelő természettudományos felsőoktatás alapjainak a megteremtése. A projekt konkrét célja a természettudományi mesterképzés kompetenciaalapú és módszertani megújítása, mely folyamatosan képes kezelni a társadalmi-gazdasági változásokat, a legújabb tudományos eredményeket, és az info-kommunikációs technológia (IKT) eszköztárát használja.

1. Annotáció

Jelen elektronikus tananyag elsősorban alkalmazott matematikus szakos hallgatók számára készült, de mindazok számára hasznos segédanyag, akik valamelyik természettudományi szakot hallgatják, vagy már elvégezték azt, rendelkeznek a középiskolai tanyagot jelentősen nem meghaladó matematikai műveltséggel (a differenciál- és integrálszámítás elemeivel), munkájuk során szembetalalálják magukat olyan statisztikai feladatokkal, amelyek megoldásához valamilyen statisztikai programcsomagot kell alkalmazniuk, és ambícionálják az általuk használt statisztikai programcsomagok mögött álló elmélet alapelveinek megértését.

2. Bevezetés

Jelen elektronikus Tananyag célja a többváltozós statisztikai módszerek bemutatása, illusztrálása statikus ábrákkal és animációkkal, valamint számos - a megértést segítő és ellenőrző - feladattal.

A többváltozós statisztikai módszereket természetesen nem lehet megérteni a matematikai statisztika alapfogalmainak és a valószínűségszámítás elemeinek ismerete nélkül. A tananyag felhsználói munkájának megkönnyítése céljából az előzetes tudnivalókat függelékben valamint részletes fogalom- és képletgyűjteményben összefoglaltuk. Az általános statisztikai tudnivalókat is illusztráltuk ábrákkal, és számos e tárgykörbe tartozó feladatot is kitűztünk. A Tananyag összeállítása során szembesültünk azzal a ténnyel, hogy olyan látszólag nyilvánvaló fogalomnak mint pl. a *marginális* eloszlás kettőnél több valószínűségi változó együttes eloszlása esetén az egzakt definíciója már reménytelenül bonyolult. Ilyenkor az ábra sem segít: számpéldákkal illusztráltuk a fogalmat.

A többváltozós statisztika klasszikus módszereit (ilyenek a regresszióanalízis, a legkisebb négyzetek módszere, a varianciaanalízis és a diszkriminanciaanalízis) együttesen normális (Gauss) eloszlású valószínűségi változókra dolgozták ki a XX. század első felében.

Ezek a módszerek erősen építenek a lineáris algebrának azon eredményeire, amelyek talán látszólagos egyszerűségük miatt kisebb hangsúlyt kapnak a matematikai képzésben, pedig a legkiválóbb matematikusok is komoly munkát fektetnek a lineáris algebra modern módszereinek tankönyvekben való feldolgozására; csak egy példa a sok közül: Lax Péter Abel-díjas matematikus rendkívül élvezetes, és számos új matematikai eredményt tartalmazó, magyarul is olvasható könyvet írt e témakörről, A Tananyag feladatai között is számos statisztikai eredetű, a lineáris algebra segítségével megoldható feladat van. Már itt figyelmeztetjük a felhasználót, hogy ezen feladatok megoldásához fejlett térszemléletre van szükség.

A modern módszerek (pl. a klaszteranalízis) inkább épülnek a heurisztikára, noha ezek elméleti megalapozásának is nagy és mély matematikai eszköztárat igénylő irodalma van. Éppen emiatt ebben a tárgykörben gyakorlatilag nem lehet vonzó és elemi eszközökkel megoldható feladatokat kitűzni.

Vannak olyan új módszerek, amelyekkel jelen sorok írója nem tud mit kezdeni, ilyen a gyakoriságtáblák közelítése alacsonyabb rangú mátrixokkal (korrespondenciaanalízis), ugyanis a lineáris algebra módszereit mechanikusan alkalmazva negatív valószínűségeket is kaphatunk eredményként. Ugyanakkor számos statisztikus sikerrel alkalmazza ezt a módszert, mi sem hagyhattuk ki a Tananyagból.

Ezzel szemben a gyakoriságtáblák elemzésének információelméleti módszereit, amelyeknek a kidolgozásában nagy szerepe van a magyar matematikusoknak - elsősorban Csiszár Imrének - részletesen ismertetjük, és ebben a tárgykörben feladatokat is kitűzünk.

Egy másik általunk csak érintett módszer a rendkívül nagyméretű mátrixokkal kapcsolatas (spektrálfelbontási) feladatok véletlen kiválasztással történő közelítése. Itt az a probléma, hogy kisméretű bemutatható példát nem találtunk.

Zárszóként két megjegyzés:

1. A statisztika legnevesebb művelői, Kolmogorovtól a vezető magyar statisztikusokig egybehangzóan állítják, hogy *vakon nem lehet statisztikát csinálni*, azaz az adatok kritikus megszemlélése nélkül már értelmes hipotézist sem lehet föltenni. Erre nyújt lehetőséget az ún. többdimenziós skálázás, azaz az adatok optimalis beágyazása lehetőleg minél kisebb dimenziós euklideszi térbe.

2. Bármilyen látványos is egy elektronikus tananyag, csupán a képernyő nézésével és kattintásokkal nem lehet elmélyülni egyetlen tudományágban sem. Az nem várható el egy felhasználótól, hogy az elmélet részleteit megjegyezze, de nem hagyható ki a papírral-ceruzával, ha úgy nem megy kalkulátorral, esetleg formulakezelő programok használatával történő aktív részvétel a tanulási folyamatban.

A tesztek a matematika elemeit meg nem haladó ismerettel rendelkező hallgatóknak nyújtanak önellenőrzési lehetőséget.

Azon a hallgatók a számára akik nagyobb óraszámban (legalább 20 kredit) hallgattak matematikát, a tesztek nem jelentenek komoly önellenőrzést, erre a feladatok szolgálnak. Még a magukat digitális bennszülöttnek érző hallgatóknak is azt javasoljuk, hogy először minden segédeszköz nélkül, pusztán a tananyagban, illetve a feladathoz írt útmutatásokban található információk alapján kíséreljenek megoldást találni. Egy kellően képzett digitalis bennszülött az Interneten szinte minden feladathoz talál hasonló kidolgozottat. Azokhoz a fejezetekhez (9., 10., 11.) nem csatoltunk feladatokat, amelyek elsősorban heurisztikus eredményeket tartalmaznak, vagy az egzakt eredmények bizonyítása lényegesen meghaladja az egyetemi tananyagot.

Végül néhány szó a Tananyag forrásairól. A közvetlen statisztikai ismeretek forrása a két szerző (Bolla Marianna és Krámli András, A statisztikai következtetések elmélete, Typotex 2005) könyve, valamint - az irodalomjegyzékben idézett - néhány eredeti folyóiratcikk. Innen csak az alapvető definíciókat és tételeket vettük át, a hangsúly a feladatokon és az illusztációkon van. A feladatok nagy részét a harmadik szerző (Nagy-György Judit) tűzte ki a gyakorlatokon. A teljes ábra- és animacióanyagot is ő készítette. Ezek jelentős része ma már közkinccsé vált eredményeket illusztrál, néhány bonyolultabb ábra Bolla Marianna javaslatára készült, az eredeti dolgozatok alapján újraszerkesztve. A Tananyag csak a Feladatok megoldásában tartalmaz bizonyításokat. Ugyanakkor alkalmazott matematikus szakon a bizonyítások ismerete szükséges a vizsgán, ezért az adott tételeknél hivatkozást adunk a papíralapú Bolla- Krámli könyv megfelelő oldalszámára.

Az előszóhoz tartozik két videó is, az *első*n látható *animáció* a t (a matematikai statisztika alaptételét) szemlélteti, a *második*on felrajzolt ábra pedig a ben szereplő függvényt ábrázolja.

A tananyagban található animációk megjelenítésére a legtöbb képnéző és böngésző alkalmas, az interaktív ábrákhoz javasoljuk a Wolfram honlapjáról (www.wolfram.com) ingyenesen letölthető Mathematica Player programot.

Szeged, 2012. december 17. Krámli András

1. fejezet - Előismeretek 1.: valószínűsgelmélet

1. Elméleti háttér

Ebben a paragrafusban a valószínűségelméletKolmogorov-féle felépítését ismertetjük, különös kiemelve a feltéles várható érték Kolmogorov-féle definícióját és annak a statisztikában használatos tulajdonságait. Hangsúlyozzuk, hogy a feltételes várható érték (és a feltételes valószínűség is) valószínűségi változó, amely bizonyos optimum tulajdonsággal rendelkezik. A nem matematikus szakos hallgatóknak elegendő annyit tudni az alábbi absztrakt definíciók nagy részéről, hogy **léteznek**. Az alkalmazó természettudományi hallgatók számára is feltétlenül tudnivaló definíciókat és állításokat *-gal megjelöljük.

Mindenek előtt vezessük be a valószínűségimező Kolmogorov féle definícióját.

1.1.1. Definíció (Kolmogorov-féle $(\Omega, \mathcal{A}, \mathbb{P})$ valószínűségi mező). .

(i) Adva van egy nem üres Ω halmaz (eseménytér), Ω elemeit elemi eseményeknek nevezzük, és ω-val jelöljük.

(ii) Ki van tüntetve az Ω részhalmazainak egy $\mathcal{A}$ algebrája ($\Omega \in \mathcal{A}, A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$, $A \in \mathcal{A} B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$.)

(iii) $\mathcal{A}$σ-algebra, azaz $A_k \in \mathcal{A} (k = 1, 2, \dots) \Rightarrow \cup_{k=1}^{\infty} \in \mathcal{A}$.

(iv) minden $A \in \mathcal{A}$ eseményhez hozzá van rendelve egy $\mathbb{P}(A)$ nemnegatív szám, az A esemény valószínűsége.

(v) $\mathbb{P}(\Omega) = 1$.

(vi) Ha $A_k \in \mathcal{A}$, páronkent egymást kizáró események, akkor $\mathbb{P}(\cup_{k=1}^{\infty}) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$.

1.1.2. Állítás (szita-formula*).

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} S_k^{(n)},$$

$n \geq k$ *és*

$$S_k^{(n)} := \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}).$$

1.1.3. Definíció (események függetlensége*). *Az $A_1, \dots, A_n$ események páronként (illetve teljesen) függetlenek, ha minden $1 \leq j < k \leq n$ párra $\mathbb{P}(A_j \cap A_k) = \mathbb{P}(A_j) \cdot \mathbb{P}(A_k)$ (illetve minden $1 \leq k \leq n$ egészre és $i_1 < \dots < i_k \varepsilon \leq n$ idexsorozatra $\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \dots \cdot \mathbb{P}(A_{i_k})$). A teljes függtelenség implikálja a páronkénti függetlenséget. Fordítva ez nem igaz!*

1.1.4. Definíció (feltételes valószínűség*).

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

ha $\mathbb{P}(B) > 0$.

1.1.5. Definíció (teljes eseményrendszer*). $A_1, \dots, A_n \in \mathcal{A}, \mathbb{P}(A_i \cap A_j) = 0$, *ha* $\mathbb{P}(B) > 0$.

1.1.6. Állítás (Bayes tétele*). *Ha $A_1, \dots, A_n$ teljes eseményrendszer és $\mathbb{P}(B) > 0$, akkor*

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)}{\sum_{k=1}^{n} \mathbb{P}(B|A_k) \cdot \mathbb{P}(A_k)}$$

1.1.7. Definíció (valószínűségi változó*). *Az Ω halmazon értelmezett olyan $X(\omega)$ valós értékű függvény, amelyre $\{\omega\colon X(\omega) \le x\}$ minden valós x-re esemény. Ha X értékkészlete megszámlálható halmaz, akkor diszkrét valószínűségi változóról beszélünk.*

1.1.8. Definíció (valószínűségi változók függetlensége*). *Az $X_1, \dots, X_n$ valószínűségi változók páronként (illetve teljesen) függetlenek, ha az $\{X_1(\omega) \le x_1\}, \dots, \{X_n(\omega) \le x_n\}$ események páronként (illetve teljesen) függetlenek, $x_1, \dots, x_n$ minden értékére.*

1.1.9. Definíció (valószínűségi változók eloszlásfüggvénye*). *Az X valószínűségi változó eloszlásfüggvénye $F_X(x) := \mathbb{P}(X \le x)$. $F_X(x)$ monoton nemcsökkenő, jobbról folytonos függvény.*

$$\lim_{n\to-\infty} F_X(x) = 0 \quad \lim_{n\to\infty} F_X(x) = 1.$$

(i) Diszkrét eset. Ha az X valószínűségi változó értékkészlete $\{x_0, x_1, \dots\}$, akkor eloszlása:

$$p_j := \mathbb{P}(x_j)$$

(ii) Abszolút folytonos eset. Ha van olyan $f(t)$ függvény amelyre $F_X(x) = \int_{-\infty}^{x} f(t)dt$. Ekkor az $f(t)$ függvényt az X valószínűségi változó sűrűségfüggvényének nevezzük.

1.1.10. Definíció (valószínűségi változó momentumai, absztrakt definíció).

Az X valószínűségi változó várható értéke $\mathbb{E}(X) := \int_\Omega X(\omega)d\mathbb{P}$, ha ez az integrál létezik. Az X valószínűségi változó n-edik (abszolút) momentuma $M_n := \int_\Omega X(\omega)^n d\mathbb{P}$, ($:= \int_\Omega |X(\omega)|^n d\mathbb{P}$), ha a fenti integrálok léteznek.

Ha $\Psi(x)$ tetszőleges Borel-mérhető valós függvény (azaz a $\{x : \Psi(x) \le y\}$ halmaz minden $y \in \mathbb{R}$-re Borel-mérhető), akkor $\mathbb{E}(\Psi(X)) := \int_\Omega \Psi(X(\omega))d\mathbb{P}$.

Az X valószínűségi változó $\mathbb{D}^2$ szórásnégyzete $\mathbb{D}^2 := \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

1.1.11. Definíció (kovariancia, korreláció, absztrakt definíció). *Két valószínűségi változó, X és Y kovarianciája:*

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].$$

Két valószínűségi változó, X és Y korrelációja:

$$r_{X,Y} := \frac{\mathrm{Cov}(X, Y)}{\mathbb{D}(X) \cdot \mathbb{D}(Y)}$$

1.1.12. Definíció (valószínűségi változó várható értékének kiszámítása*).

(i) Diszkrét eset. Ha az X valószínűségi változó értékkészlete $\{x_0, x_1, \dots\}$, akkor várhtó értéke:

$$\mathbb{E}(X) := \sum_{j=0}^{\infty} x_j \mathbb{P}(x_j) = \sum_{j=0}^{\infty} x_j p_j,$$

amennyiben a fenti sor abszolút konvergens

(ii) Abszolút folytonos eset. Ha az X valószínűségi változó sűrűségfüggvénye $f(t)$ akkor várhtó értéke:

$$\mathbb{E}(X) := \int_{-\infty}^{\infty} x f(x) dx$$

amennyiben a fenti integrál létezik.

Ha ismerjük a várható érték kiszámítási módját, a magasabb momentumok és szórásnégyzet kiszámítási módja már könnyen adódik:

(i) n-edik momentum: $M_n := \mathbb{E}(M_n)$,

(ii) szórásnégyzet: $\mathbb{D}^2 := \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

Hasonlóan számítható ki két valószínűségi változó kovarianciája és korrelációja. Ez természetesen nem azt jelenti, hogy a tényleges számolás elvégzése is könnyű.

1.1. Feltételes várható érték

A fent ismertetett valószínűségelmélet alapismeretek már elegendőek a feltételes várható érték fogalmának bevezetéséhez, tulajdonságaik, valamint - diszkrét és abszolút folytonos esetben - kiszámítási módjuk ismertetéséhez.

1.1.1.1. Definíció (egy σ-algebrára nézve vett feltételes várható érték). *Az X valószínűségi változónak az $\mathcal{A}_1 \subseteq \mathcal{A}$ σ-algebrára nézve akkor vehető az $X_1 := \mathbb{E}(X|\mathcal{A}_1)$ feltételes várható értéke, ha $\mathbb{E}(X)$ létezik. X_1-et az alábbi két tulajdonság definiálja.*

(i) X_1 $\mathcal{A}_1$-mérhető, azaz minden valós x-re $\{\omega\colon X_1 \leq x\} \in \mathcal{A}_1$.

(ii) Minden $A \in \mathcal{A}_1$ halmazra $\mathbb{E}(1_A \cdot X) = \mathbb{E}(1_A \cdot X_1)$ vagy másképpen írva $\int_A X d\mathbb{P} = \int_A X_1 d\mathbb{P}$, ahol 1_A jelenti az A halmaz indikátorfüggvényét.

Bebizonyítható, hogy 1. es 2. feltételek teljesíthetők, és X_1 majdnem biztosan egyértelmű.

1.1.1.2. Megjegyzés. *Ha $\mathcal{A}_1$ valamely Y valószínűségi változó $\{Y(\omega) \leq x\}$ $x \in \mathbb{R}$ nívóhalmazai által generált σ-algebra, akkor van értelme az $\mathbb{E}(X|Y)$ feltételes várható értéknek.*

1.1.1.3. Állítás. *Felsoroljuk a feltételes várható érték alapvető tulajdonságait.*

(i) A feltételes várható érték vétel lineáris operáció, azaz

$$\mathrm{E}((a \cdot X + b \cdot Y)|\mathcal{A}_1) = a \cdot E(X|\mathcal{A}_1) + b \cdot E(Y|\mathcal{A}_1).$$

(ii) Ha az Y valószínűségi változó $\mathcal{A}_1$-mérhető, akkor

$$\mathbb{E}(Y \cdot X|\mathcal{A}_1) = Y\mathbb{E}(X|\mathcal{A}_1).$$

(iii) Ha az X valószínűségi változó független Y-tól, akkor

$$\mathbb{E}(X|Y) = \mathbb{E}(X).$$

(iiii) Toronyszabály: $\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|X)]$.

A statisztika egyik alapvető feladata az ún. regresszió, azaz egy $\mathbf{Y}$ valószínűségi változó egy $\mathbf{X}$ valószínűségi változó valamilyen Borel-mérhető $f(x)$ valós függvényével való optimális közelítése (az „optimális" szó jelentése különböző esetekben más és más lehet). Az alábbi állítás alapvető jelentősgű ennek a célnak a megvalósítása szempontjából.

1.1.1.4. Állítás. *Ha létezik $\mathbb{E}(Y)$ és Y mérhető az $\mathbf{X}$ valószínűségi változó $\{X(\omega) \leq x\}$ $x \in \mathbb{R}$ nívóhalmazai által generált $\mathcal{A}_x$ σ-algebrára, akkor akkor van olyan Borel-mérhető $t(x)$ valós függvény, hogy*

$$\mathbb{P}(Y(\omega)) = t(X(\omega)) = 1$$

A 16. Állítás egy közvetlen alkalmazása a következő

1.1.1.5. Állítás. *Ha $\mathbb{E}(Y^2) < \infty$, akkor*

$$\min_{t\colon t\mathcal{A}\text{-mérhető}} \mathbb{E}(Y - t(X))^2 = \mathbb{E}(Y - \mathbb{E}(Y|X))^2,$$

azaz az Y valószínűségi változó legjobb közelítése X Borel-mérhető függvényeivel éppen $\mathbb{E}(Y|X)$.

Most rátérünk a feltételes eloszlás (diszkrét eset), feltételes sűrűségfüggvény, valamint a feltételes várható érték kiszámítási módjára.

1.1.1.6. Definíció (feltételes eloszlás). *Legyen az X és Y valószínűségi változók értékkészlete $x_1, \dots, x_m$, illetve $y_1, \dots, y_n$, együttes eloszlásuk (p_{ij}), az X, illetve Y perem- (vagy marginális) eloszlásai legyenek $p_{i\cdot} = \sum_{j=1}^{n} p_{ij}$, illetve $p_{\cdot j} = \sum_{i=1}^{m} p_{ij}$. Ekkor a feltételes valószínűségdefiníciója alapján az Y valószínűségi változó $X = x_i$ melletti feltételes eloszlása:*

$$p_{j|i} = \frac{p_{ij}}{p_{i\cdot}}, \qquad j = 1, \dots, n.$$

1.1.1.7. Definíció (feltételes várható érték, diszkrét eset). *A fenti jelölésekkel az Y valószínűségi változó $X = x_i$ melletti feltételes várható értéke:*

$$\mathbb{E}(Y|X = x_i) = \sum_{j=1}^{n} y_j \cdot p_{j|i} = \frac{1}{p_{i\cdot}} \sum_{j=1}^{n} y_j \cdot p_{ij}.$$

1.1.1.8. Megjegyzés. *Vegyük észre, hogy sem a $(p_{j|i})$ feltételes eloszlás, sem az $\mathbb{E}(Y|X = x_i)$ feltételes várható érték nem függ az x_i konkrét értéktől!*

1.1.1.9. Definíció (feltételes sűrűségfüggvény). *Legyen $f(x, y)$ az X és Y valószínűségi változók együttes sűrűségfüggvénye, $f_1(x) := \int_{-\infty}^{\infty} f(x, y)dy$ pedig az X valószínűségi változó perem- (vagy marginális) sűrűsége. Az Y valószínűségi változó $X = x$ feltétel melletti feltételes sűrűsége:*

$$\begin{aligned} f_{2|1}(y|x) &= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{\mathbb{P}(X \in [x, x + \Delta x), Y \in [y, y + \Delta y))}{\mathbb{P}(X \in [x, x + \Delta x)) \cdot \Delta y} = \\ &= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{\mathbb{P}(X \in [x, x + \Delta x), Y \in [y, y + \Delta y))}{\frac{\mathbb{P}(Y \in [x, x + \Delta x))}{\Delta x} \cdot \Delta x \cdot \Delta y} = \qquad (1) \\ &= \frac{f(x, y)}{f_1(x)}. \end{aligned}$$

Most megfogalmazzuk a Bayes-tételnek a statisztikában rendkívül hasznos, abszolút folytonos eloszlásra érvényes alakját.

1.1.1.10. Tétel (Bayes-tétel). *Legyenek X, Y, $f(x, y)$, $f_1(x)$ és $f_{2|1}(y|x)$ ugyanazok, mint a fenti definícióban. Ekkor*

$$f_{1|2}(x|y) = \frac{f_{2|1}(y|x) f_1(x)}{\int_{-\infty}^{\infty} f_{2|1}(y|x) f_1(x) dx}.$$

1.1.1.11. Definíció (feltételes várható érték, abszolút folytonos eset). *A fenti jelölésekkel az Y valószínűségi változó $X = x$ feltétel melletti feltételes várható értéke:*

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y \cdot f_{2|1}(y|x)dx = \frac{1}{f_1(x)} \int_{-\infty}^{\infty} y \cdot f(x, y)dy. \qquad (1)$$

Az $\mathbb{E}(Y|X = x)$ feltételes várható érték - ellentétben a diszkrét esettel - függ az x értéktől; jelölje ezt a függést $t(x)$. A feltételes várható érték szemléletes jelentése: Az $\mathbb{E}(Y|X)$ nem más, mint az Y valószínűségi változó integrálközepe az X valószínűségi változó nívóhalmazain.

Végül definiáljuk a feltételes szórásnégyzetet, kovarianciát, és az ún. parciális korrelációt.

1.1.1.12. Definíció (feltételes szórásnégyzet). *Az Y valószínűségi változó feltételes szórásnégyzete az X valószínűségi változóra nézve:*

$$\mathbb{D}^2(Y|X) := \mathbb{E}[Y - \mathbb{E}(Y|X)^2|X].$$

1.1.1.13. Definíció (feltételes kovariancia). *Az Y és Z valószínűségi változók feltételes kovarianciája az X valószínűségi változóra nézve:*

$$\mathrm{Cov}(Y, Z|X) := \mathrm{Cov}(Y - \mathbb{E}(Y|X), Z - \mathbb{E}(Z|X)).$$

1.1.1.14. Definíció (parciális korreláció). *Az Y és Z valószínűségi változók feltételes kovarianciája az X valószínűségi változóra nézve:*

$$r_{Y,Y|X} := \frac{Cov(Y, Z|X)}{\mathbb{D}(Y - \mathbb{E}(Y|X)) \cdot \mathbb{D}(Z - \mathbb{E}(Y|Z))}.$$

Vegyük észre, hogy míg a feltételes szórásnégyzet és a feltételes kovariancia valószínűségi változók, amelyek függenek a feltételtől, a parciális korreláció szám, ami csak $r_{Y,Z}$-től, $r_{Y,X}$-től és $r_{Z,X}$-től függ; igaz az alábbi állítás.

1.1.1.15. Állítás.

$$r_{Y,Z|X} := \frac{r_{Y,Z} - r_{Y,X} \cdot r_{Z,X}}{\sqrt{(1 - r_{Y,X}^2)(1 - r_{Z,X}^2)}}.$$

A parciális korreláció szemléletesen azt a jelenséget írja le, hogy két valószínűségi változó (Y és Z) azért korreláltak erősen, mert mindketten erősen korreláltak egy harmadik valószínűségi változóval, nevezetesen X-szel. A fenti állítás bizonyítása azon az alapvető tényen múlik, hogy két valószínűségi változó kovarianciája két vektor skaláris szorzatának tekinthető, és ha ez a kovariancia zérus, akkor a két valószínűségi változó mint vektor merőleges egymásra.

1.2. A normális eloszlásból származtatott eloszlások

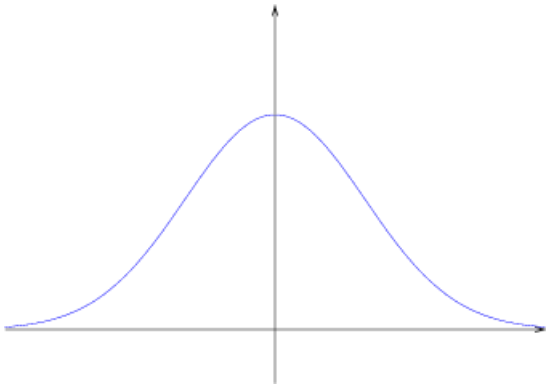
1.1.2.1. Definíció (normális eloszlás). *Az m várható értékű és σ^2 szórásnégyzetű X valószínűségi változó sűrűségfüggvénye*

$$f(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-m)^2}{2}\}. \qquad (1)$$

A $\Phi(x) := \int_{-\infty}^{x} f(s)ds$ eloszlásfügvény nem fejezhető ki elemi függvényekkel.

Az m várható értékű és σ^2 szórásnégyzetű normális eloszlás jelölése: $\mathcal{N}(m, \sigma^2)$.

Az alábbi ábra mutatja a standard normális eloszláshoz, azaz $\mathcal{N}(0, 1)$-hez tartozó sűrűségfüggvényt.



$\phi(x)$

1.1.2.2. Definíció (n szabadságfokú χ^2 eloszlás). *Ha $X_1, \dots, X_n$ független $\mathcal{N}(m, \sigma^2)$ valószínűségi változók, az*

$$Y_n := X_1^2 + \dots + X_n^2$$

valószínűségi változó definíció szerint Y_n szabadságfokú centrált χ^2 -eloszlású: $Y_n \sim \chi^2(n)$, melynek sűrűségfüggvénye

$$f_n(x) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)}, \qquad \text{ha} \, x > 0.$$

ahol $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x}$. Megjegyezzük, hogy $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(n) = (n-1)!$ és $Gamma(1/2) = \sqrt{\pi}$

(i) Az $\chi^2(n)$-eloszlás $\mathcal{G}(n/2, 1/2)$ Gamma-eloszlás.

(ii) A $\chi^2(n)$ eloszlás tetszőleges momentuma meghatározható, a számolás visszavezethető a normális eloszlás páros momentumainak meghatározására: $\mathbb{E}(Y_n) = n$, $\mathbb{D}^2(Y_n) = 2n$.

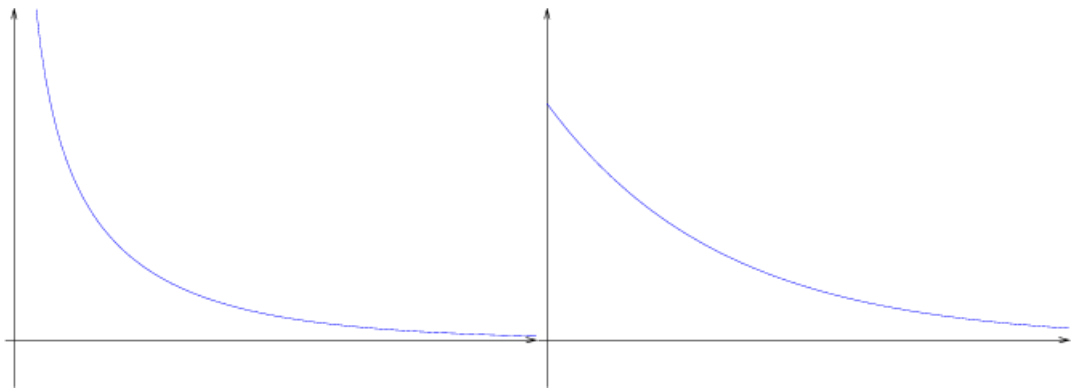
(iii) Ha $X \sim \mathcal{N}(0, \sigma^2)$, akkor minden n természetes számra

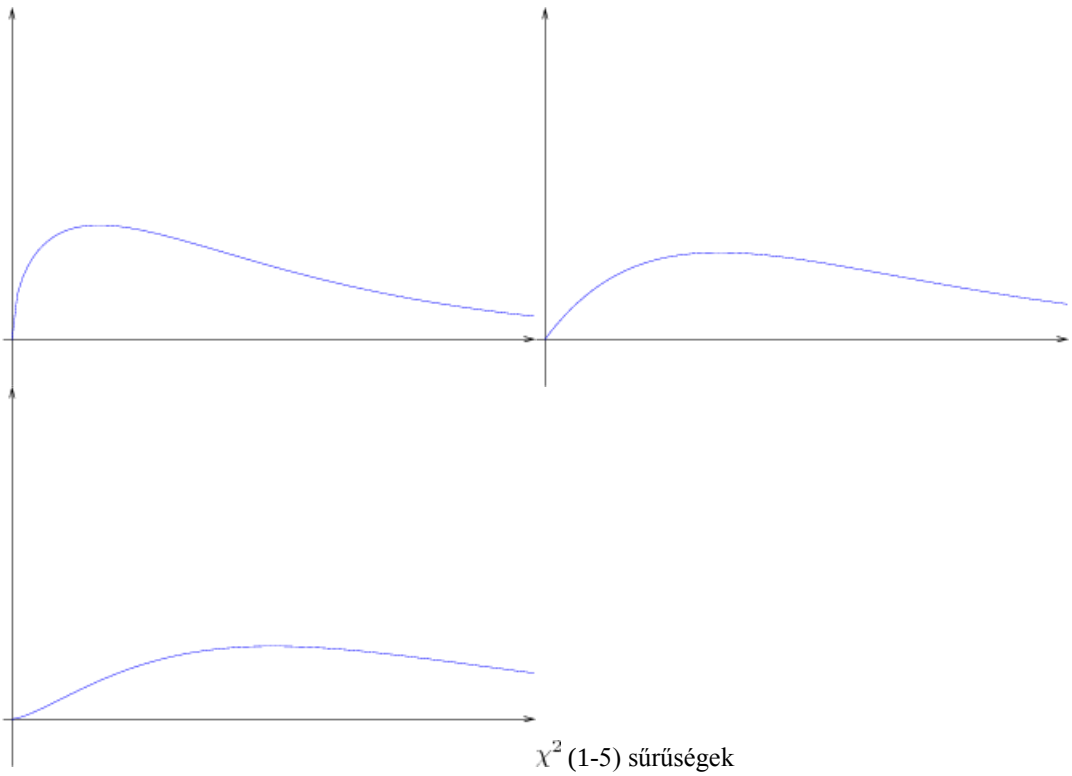
$$\mathbb{E}(X^{2n}) = \prod_{j=0}^{n-1} (2j+1)\sigma^{2n} \qquad (1)$$

???

(iiii) Ha $n \to \infty$, Y_n eloszlása $\mathcal{N}(n, 2n)$-nel közelíthető.

Az alábbi ábrák mutatják az 1, 2, 3, 4, és 5 szabadságfokú χ^2 eloszlásokhoz tartozó sűrűségfüggvényeket.





χ^2 (1-5) sűrűségek

1.1.2.3. Definíció (n szabadságfokú Student-féle eloszlás (t-eloszlás)). *Ha X standard normális eloszlású valószínűségi változó, és $Y_n \sim \chi(n)$ független X-től, akkor*

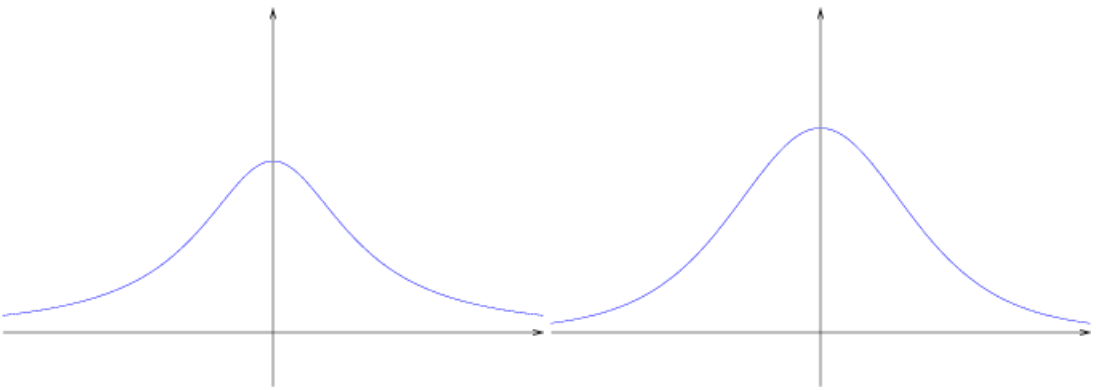
$$Z_n := \sqrt{n} \cdot \frac{X}{\sqrt{Y_n}} = \frac{X}{\sqrt{Y_n/n}}$$

definíció szerint n szabadsági fokú standard Student-eloszlású valószínűségi változó: $Z_n \sim t(n)$

1.1.2.4. Állítás. *A $t(n)$ eloszlás sűrűségfüggvénye:*

$$g_n(z) = \frac{2}{\sqrt{\pi n}\Gamma(\frac{n}{2})\left(1+\frac{z^2}{n}\right)^{\frac{n+1}{2}}} \int_0^\infty t^{\frac{n-1}{2}} e^{-t} dt = \\ = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1+\frac{z^2}{n}\right)^{-\frac{n+1}{2}}. \qquad (1)$$

Az alábbi ábrák mutatják az 1, és 5 szabadságfokú Student eloszlásokhoz tartozó sűrűségfüggvényeket.



t(1) és t(5) sűrűségek

$n \to \infty$ sűrűségfüggvény (1.5) alakjából leolvasható, hogy a Z_n eloszlásban tart a standard normális eloszláshoz, ha . Ezt az alábbi *animáció* szemlélteti.

Ugyancsak (1.5)-ből látható az is, hogy Z_n-nek csak $n-1$ momentuma véges. Az 1 szabadságfokú t-eloszlás a Cauchy-eloszlás.

1.1.2.5. Definíció ((n,m) szabadságfokú F-eloszlás). *Ha $X_n \sim \chi^2(n)$ és és $Y_m \sim \chi^2(m)$, akkor a*

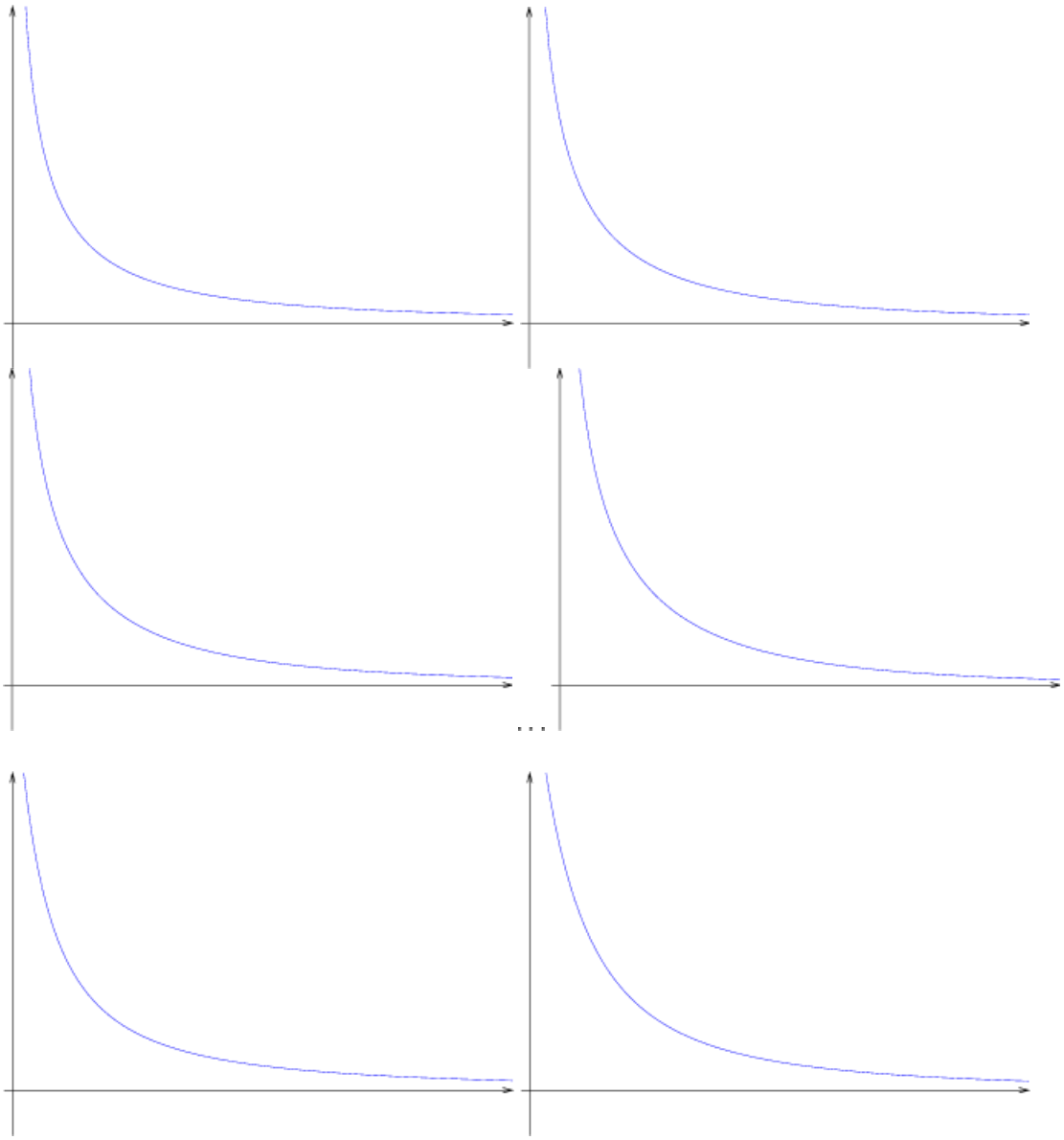
$$Z_{n,m} := \frac{\frac{X_n}{n}}{\frac{Y_m}{m}}$$

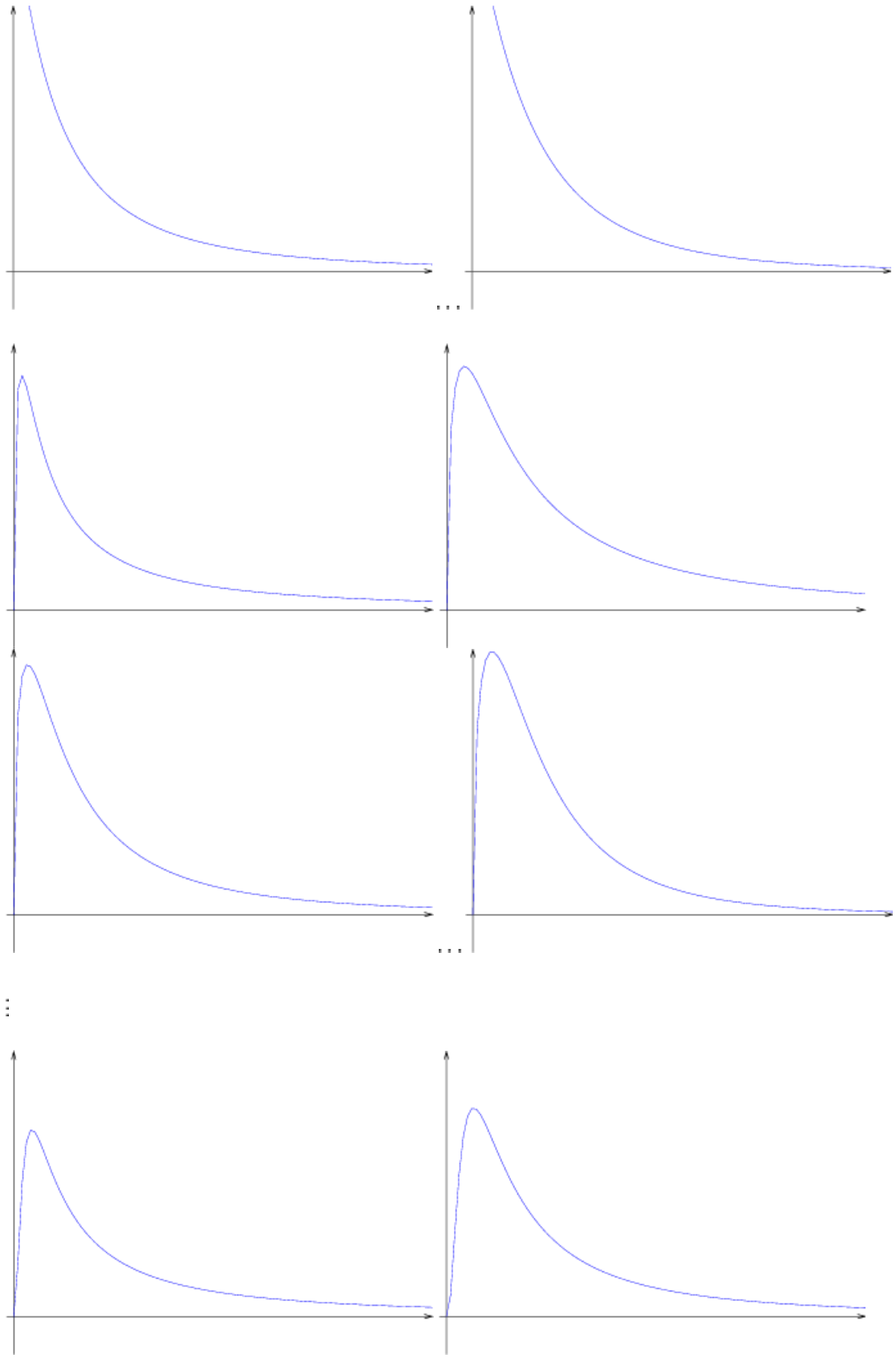
valószínűségi változó (n,m) szabadságfokú F-eloszlású: $Z_{n,m} \sim \mathcal{F}(n,m)$.

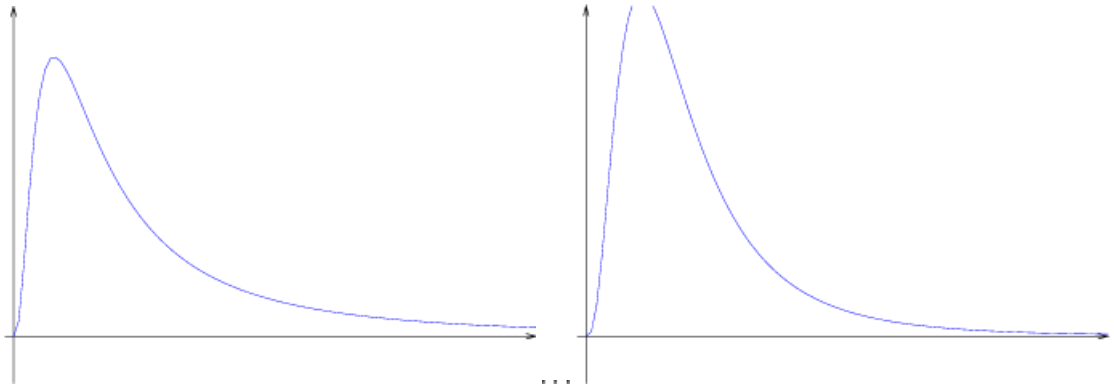
$Z_{n,m}$ változó sűrűségfüggvénye

$$f_{n,m}(z) = \frac{n\Gamma\left(\frac{n+m}{2}\right)}{m\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{\left(\frac{n}{m}z\right)^{\frac{n}{2}-1}}{\left(1+\frac{n}{m}z\right)^{\frac{n+m}{2}}}.$$

Az alábbi ábrák mutatják az (1,1), (1,2), (1,3), (1,9), (2,1), (2,2), (2,3), (2,9), (3,1), (3,2), (3,3), (3,9), (9,1), (9,2), (9,3) és (9,9) szabadságfokú F eloszlásokhoz tartozó sűrűségfüggvényeket.







F sűrűségek

1.1.2.6. Definíció (Béta-eloszlás). *Ha* $X_1, \dots, X_n, \dots, X_{n+m}$ *független* $\mathcal{N}(0,1)$*-változók, akkor a*

$$\tilde{Z}_{n,m} = \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n+m} X_i^2}$$

valószínűségi változó $\mathcal{B}(n/2, m/2)$*-eloszlású:* $Z_{n,m} \sim \mathcal{B}(n/2, m/2)$.

A $\tilde{Z}_{n,m}$ változó $f_{Z_{n,m}}(z)$ sűrűségfüggvénye

$$f_{Z_{m,n}}(z) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} z^{\frac{n}{2}-1}(1-z)^{\frac{m}{2}-1}, \qquad \text{ha } 0 < z < 1.$$

A fenti képletnek akkor is van értelme, ha a kitevőben szereplő $\frac{n}{2}$ illetve $\frac{m}{2}$ helyett tetszőleges a illetve b pozitív számok állnak. Ez az (a,b)-rendű*béta-eloszlás* sűrűségfüggvénye:

$$f_{a,b}(z) = \frac{1}{B(a,b)} \cdot z^{a-1}(1-z)^{b-1}, \qquad \text{ha } 0 < z < 1.$$

Vegyük észre, hogy a $\mathcal{B}(1,1)$-eloszlás megegyezik a $[0,1]$ intervallumon egyenletes $\mathcal{U}(0,1)$-eloszlással!

1.3. Többváltozós ismeretek

Eddig $X_1, \dots, X_n$ független $\mathcal{N}(\theta, \sigma^2)$ valószínűségi változókat jelentettek. Most kimondunk egy állítást megkönnyíti a normális eloszlású valószínűségi változók függetlenségének ellenőrzését.

1.1.3.1. Állítás. *Ha* $Y_1, \dots, Y_m$ *az* $X_1, \dots, X_n$ *független* $\mathcal{N}(\theta, \sigma^2)$ *valószínűségi változók lineáris kombinációi, akkor* $\operatorname{Cov}(Y_i, Y_j) = \delta_{ij}$ *maga után vonja az* $Y_1, \dots, Y_m$ *változók (teljes!) függetlenségét.*

Most már minden ismeret rendelkezésünkre áll ahhoz, hogy megfogalmazzunk egy, a becsléselméletben és a hipotézisvizsgálatban gyakran használt tételt, ami Lukács Jenő tételének speciális esete (l. [21]).

1.1.3.2. Tétel (Lukács Jenő). *Legyenek* $X_1, \dots X_n$ *független* $\mathcal{N}(\theta, \sigma^2)$ *valószínűségi változók, legyen továbbá* $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$, $S_n^{*2} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

(i) $\bar{X} \sim \mathcal{N}(\theta, \sigma^2/n)$,

(ii) $(n-1)S_n^{*2}/\sigma^2 \sim \chi^2(n-1)$,

(iii) $\bar{X}$ *és* S_n^{*2} *függetlenek.*

1.1.3.3. Következmény.

$$Y = \frac{\sqrt{n}(\bar{X} - \theta)}{\sqrt{S_n^{*2}}} \sim t(n-1).$$

1.1.3.4. Tétel. *Ha $X_1, \dots, X_n$ független $N(0, \vartheta)$ valószínűségi változók, akkor*

$$Z' := \frac{\sqrt{n} \cdot \bar{X}}{\sqrt{\sum_{j=1}^{n} X_j^2}} \quad \text{és} \quad S^2(\mathbf{X}) := \sum_{j=1}^{n} X_j^2$$

függetlenek.

1.1.3.5. Következmény. *A*

$$T = \frac{\sqrt{n}\bar{X}}{\sqrt{S_n^{*2}}}$$

Student-statisztika is független S^2-től, ugyanis egyszerű számolással adódik, hogy Z' a T monoton függvénye: $Z' = \frac{T}{\sqrt{T^2+n-1}}$.

($\bar{X}$ és S_n^* definícióit l. 35. tételben.)

A varianciaanalízis alapvető eszköze a következő meglepő tétel, amely a 35. tétel általánosításának is tekinthető.

1.1.3.6. Tétel (Fisher- Cochran-tétel). *Legyen $\mathbf{X} = (X_1, \dots, X_n)^T \sim \mathcal{N}_n(\mathbb{0}, \mathbf{I}_n)$ véletlen vektor (komponensei független $\mathcal{N}(0,1)$-változók) és legyenek a $Q = \mathbf{X}^T\mathbf{I}_n\mathbf{X} = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} X_i^2$ és a $Q_j = \mathbf{X}^T\mathbf{A}_j\mathbf{X}$ $(j = 1, \dots, k)$ $\mathbf{X}$-szel és a szimmetrikus, $n \times n$-es $\mathbf{A}_j$ mátrixokkal $(j = 1, \dots, k \leq n)$ képzett kvadratikus alakok olyanok, hogy rájuk*

$$Q = Q_1 + Q_2 + \dots + Q_k$$

teljesül. Legyen Q_j rangja: $\mathrm{rk}(A_j) = n_j$. A $Q_1, Q_2, \dots, Q_k$ kvadratikus alakok pontosan akkor független χ^2-eloszlásúak $n_1, n_2, \dots, n_k$ szabadságfokkal, ha

$$\sum_{j=1}^{k} n_j = n.$$

A Fisher- Cochran-tétel fontossága miatt kivételesen közöljük annak egy elemi bizonyítását. Az egyik irány a χ^2-eloszlás definícójának egyszerű következménye, a másik - meglepő - irány az alábbi lineáris algebrai állításból adódik.

1.1.3.7. Állítás. *Ha az n-dimenziós egységmátrix*

$$\mathbf{I}_n = \mathbf{A}_1 + \dots + \mathbf{A}_k \qquad (1)$$

alakú, ahol az $\mathbf{A}_1, \dots, \mathbf{A}_k$ valós szimmetrikus mátrixok és

$$\mathrm{rang}(\mathbf{A}_1) + \dots + \mathrm{rang}(\mathbf{A}_k) = n, \qquad (1)$$

akkor ezen mátrixok $\mathrm{rang}(\mathbf{A}_1), \dots, \mathrm{rang}(\mathbf{A}_k)$ dimenziós ortogonális alterekre való ortogonális projekciók mátrixai.

Az alábbi megjegyzés segít abban, hogy bonyolult számítások elvegzése nélkül is alkalmazzuk a Fisher- Cochran tételt.

1.1.3.8. Megjegyzés. *A kvadratikus alakok rangját az alábbi heurisztikus formulával számolhatjuk (Q itt is a kvadratikus alak rövidítése):*

$$\mathrm{rang}(Q) = \text{a Q-ban szereplő független azonos eloszlású valószínűségi változók száma mínuszaz ugyanezen valószínűségi változók alapján függetlenül becsült paraméterek száma.}$$

Végül kimondunk egy tételt, ami bizonyos értelemben indokolja, hogy első közelítésben miért veszünk mindig lineáris regressziót.

$\hat{Y} := \mathbb{E}(Y|X_1, \ldots, X_n)$ *Legyenek* $Y, X_1, \ldots, X_n$ $X_1, \ldots, X_n$ *normális eloszlású valószínűségi változók. Az feltételes várható érték az valószínűségi változók lineáris függvénye.*

Mivel a 17. állítás szerint Y feltételes várható értéke az $X_1, \ldots, X_n$ valószínűségi változókra éppen a négyzetes középben való legjobb közelítés a fenti állítás szerint ez a közelítés az $X_1, \ldots, X_n$ valószínűségi változók lineáris függvénye.

2. Feladatok

(i) Számítsuk ki a λ paraméterű Poisson eloszlás első négy momentumát!

Tipp: Alkalmazzuk a momentumoknak a deriváltjai alapján történő kiszámítási módját.

Válasz: $M_1 = \lambda$, $M_2 = \lambda^2 + \lambda$, $M_3 = \lambda^3 + 3\lambda^2 + \lambda$, $M_4 = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$.

(ii) Legyen X egy (r, p) paraméterű $(r > 1)$ negatív binomiális eloszlású valószínűségi változó. Számítsuk ki $\mathrm{E}(\frac{1}{X-1})$ várható értéket!

Tipp: Használjuk a definíciót .

Válasz: A definíció alapján $\frac{p}{r-1}$.

(iii) Számoljuk ki az n-edrendű λ paraméterű Gamma eloszlás $-k$-adik momentumát, ahol $k < n$.

Tipp: definíciót .

Válasz: A definíció alapján $\frac{\lambda^k (n-k-1)!}{(n-1)!}$.

(iiii) Legyenek X, Y független, azonos eloszlású, véges várható értékű valószínűségi változók. Határozzuk meg $\mathrm{E}(X + Y|X)$ és $\mathrm{E}(X|X + Y)$ feltételes várható értékeket!

Tipp: Alkalmazzuk tulajdonságait, és vegyük észre, hogy X és Y szerepe szimmetrikus!

Válasz: $X + \mathrm{E}(Y)$ ill. $\frac{X+Y}{2}$.

(iiiii) Legyen X és Y két független, 1/2 paraméterű Bernoulli-eloszlású valószínűségi változó. Adjuk meg $\mathrm{E}(X|X + Y)$ által generált σ-algebrát és $\mathrm{E}(X|X + Y)$ eloszlását!

Tipp: $X + Y$ által generált σ-algebrát.

Válasz: $Z := \mathrm{E}(X|X + Y)$, $P(Z = 0) = 1/4$, $P(Z = 1/2) = 1/2$, $P(Z = 1) = 1/4$.

(iiiiii) Legyen X nemnegatív valószínűségi változó. Tegyük fel, hogy léteznek az $\mathrm{E}(X^2)$ és $\mathrm{E}(\frac{1}{X})$ várható értékek!

(a) Határozzuk meg $\mathrm{E}(X^2|X)$-et!

(b) Határozzuk meg $\mathrm{E}(\frac{1}{X}|X)$-et!

Tipp: Egy X valószínűségi változó $f(X)$ függvényének feltételes várható értéke X-re $f(X)$, ha ez utóbbi várható értéke létezik.

Válasz:

(a) X^2,

(b) $\frac{1}{X}$.

(iiiiiii) Legyen X a $[-1, 1]$ intervallumon egyenletes eloszlású valószínűségi változó. Határozzuk meg $\mathrm{E}(X|X^2)$-t!

Tipp: Használjuk a definíciót és a tulajdonságait.

Válasz: A definíció alapján: 0.

(iiiiiiii) Legyenek X_1, X_2 a $[0,1]$ intervallumon egyenletes eloszlású független valószínűségi változók, továbbá $Y := \min\{X_1, X_2\}$, valamint $Z := \max\{X_1, X_2\}$. Határozzuk meg

(a) $\mathrm{E}(Y|Z)$,

(b) $\mathrm{E}(Z|Y)$,

(c) $\mathrm{E}(X_1|Z)$

feltételes várható értékeket!

Tipp: Használjuk a definícióját! Használjuk ki X_1 és X_2 szimmetriáját, valamint azt, hogy $X_1 + X_2 = Y + Z$!

Válasz:

(a) $Z/2$,

(b) $(Y+1)/2$,

(c) $\frac{3}{4}Z$.

(iiiiiiiii) Legyenek $X, Y \sim \mathcal{N}(0,1)$ független valószínűségi változók, továbbá $a, b, c \in \mathbb{R}$.

(a) Milyen eloszlású $aX + bY + c$?

(b) Adjuk meg $|X|$ sűrűségfüggvényét!

(c) Határozzuk meg X^2 sűrűségfüggvényét! Milyen eloszlást követ X^2?

(d) Milyen eloszlású $X^2 + Y^2$?

Tipp: (c) Alkalmazzuk a valószínűségi változó függvénye eloszlására vonatkozó képletét, valamint a nevezetes abszolút folytonos eloszlások felsorolását.

Válasz:

(a) $\mathcal{N}(c, a^2+b^2)$,

(b) $\frac{2}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$ ha $x \geq 0$ és 0 egyébként,

(c) $\frac{x^{-1/2}\exp(-x/2)}{\sqrt{2\pi}}$, azaz $\chi^2(1)$

(d) $\chi^2(2)$, ami megegyezik a $\lambda = 1/2$ paraméterű $\mathcal{E}xp(1/2)$ exponenciális eloszlással.

(iiiiiiiii) Legyenek $X, Y \sim \mathcal{E}xp(\lambda)$ független valószínűségi változók.

(a) Milyen eloszlású $X + Y$?

(b) Adjuk meg $Z = \frac{X}{Y}$ sűrűségfüggvényét!

Tipp:

(a) Alkalmazzuk a nevezetes abszolút folytonos eloszlások felsorolását.

(b) Alkalmazzuk a 2 valószínűségi változó hányadosának sűrűségfüggvényére eloszlására vonatkozó képletét, valamint a nevezetes abszolút folytonos eloszlások felsorolását.

Válasz:

(a) $\mathcal{G}(2,\lambda)$.

(b) $\frac{2}{(1+z)^2}$, ha $X \geq 0$ azaz $\mathcal{F}(2,2)$

(iiiiiiiiii) * Legyenek $N, X_1, X_2 \ldots$ független valószínűségi változók, ahol N egy p paraméterű geometriai eloszlású, $X_1, X_2, \ldots$ pedig λ paraméterű exponenciális eloszlásúak. Milyen eloszlású lesz $\sum_{i=1}^{N} X_i$?

Tipp: Alkalmazzuk a megfelelő formuláit és írjuk be az exponenciális eloszlás karakterisztikus függvényét az $1, 2, \ldots$ értékkészletű geometriai eloszlás generátorfüggvényébe.

Válasz: $\mathcal{E}xp(p\lambda)$

(iiiiiiiiiii) Mi a kapcsolat az alábbi eloszlásseregek között?

(a) Bernoulli, binomiális és Poisson;

(b) geometriai és negatív binomiális;

(c) exponenciális, χ^2 és Gamma;

(d) Student és Cauchy.

Tipp: Alkalmazzuk a t, és keressük meg hogy a felsoroltak között melyik eloszlás speciális esete, ill. határesete egy másik eloszlásnak.

Válasz:

(a) Bernoulli $\subset$ binomiális: a Poisson határesete;

(b) geometriai $\subset$ negatív binomiális;

(c) exponenciális: $\chi^2(2) \subset$ Gamma;

(d) Cauchy: $t(1)$.

(iiiiiiiiiiii) Legyen X egy (α,λ), Y pedig (β,λ) paraméterű Gamma eloszlású, egymástól független valószínűségi változó. Igaz-e, hogy X/Y egy (α,β) paraméterű másodfajú Béta eloszlású valószínűségi változó, amely sűrűségfüggvénye

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{x^{\alpha-1}}{(x+1)^{\alpha+\beta}}.$$

Tipp: 2 valószínűségi változó hányadosának sűrűségfüggvényére eloszlására vonatkozó képletét, valamint a nevezetes abszolút folytonos eloszlások felsorolását.

Válasz: Igaz.

(iiiiiiiiiiiii) * Legyen X egy (α,β) paraméterű másodfajú Béta eloszlású valószínűségi változó. Igazoljuk, hogy

(a) $\frac{1}{X}$ valószínűségi változó (β,α) paraméterű másodfajú Béta eloszlású!

(b) $\frac{X}{1+X}$ valószínűségi változó (α,β) paraméterű Béta eloszlású!

(c) $\frac{1}{1+X}$ valószínűségi változó (β,α) paraméterű Béta eloszlású!

Tipp: Keressük meg a ben a Fischer-féle $\mathcal{F}$ eloszlás képletét, vegyük észre, hogy az $n/2\,m/2$ paraméterű másodfajú Béta eloszlású valószínűségi változó az n, m szabadságfokokkal normált χ^2 eloszlású valószínűségi változók hányadosa. Továbbá alkalmazzuk a valószínűségi változó függvényének illetve valószínűségi változók hányadosának sűrűségére vonatkozó képletet.

Válasz: L. Tipp.

(iiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m} \sim \mathcal{E}xp(\lambda)$ független azonos eloszlású valószínűségi változók.

(a) Milyen eloszlású $\sum_{i=1}^{n} X_i$?

(b) Igazoljuk, hogy

$$Z = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=n+1}^{n+m} X_i}$$

statisztika (n, m) paraméterű másodfajú Béta eloszlású!

(c) Igazoljuk, hogy

$$\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n+m} X_i} = \frac{1}{1 + 1/Z} \sim Beta(n, m).$$

Tipp:

(a) Keressük meg a ben a megfelelő eloszlásokat.

(b) Alkalmazzuk a valószínűségi változók hányodosának eloszlására vonatkozó képletét.

(c) Alkalmazzuk a valószínűségi változók hányodosának eloszlására vonatkozó képletét.

Válasz:

(a) $\mathcal{G}(n, \lambda)$.

(b) L. Tipp.

(c) L. Tipp.

(iiiiiiiiiiiiiii) Mi a kapcsolat a Student, F és Béta eloszlásseregek között?

Tipp: Alkalmazzuk a t, és keressük meg, hogy a felsoroltak között melyik eloszlás speciális esete, ill. melyik eloszláshoz tartozó valószínűségi változó függvénye egy másik eloszláshoz tartozó valószínűségi változónak.

Válasz: Ha $X \sim t(n)$, akkor $X^2 \sim \mathcal{F}(1, n)$

Ha $Z_{m,n} \sim \mathcal{F}(m, n)$, akkor $Y = \frac{Z_{m,n}}{1+Z_{m,n}} \sim B(m/2 - 1, n/2 - 1)$

(iiiiiiiiiiiiiiii) Legyenek $X_1, \ldots, X_n \sim \mathcal{E}xp(\lambda)$ független azonos eloszlású valószínűségi változók. Definiáljuk $Y_1, \ldots, Y_n$ valószínűségi változóket a következő módon:

$$Y_1 = X_1, Y_2 = X_1 + X_2, \ldots, Y_{n-1} = X_1 + \cdots + X_{n-1}.$$

(a) Legyen $Z = X_1 + \cdots + X_n$. Határozzuk meg az $Y_1, \ldots, Y_n$ valószínűségi változók együttes feltételes sűrűségfüggvényét a $Z = z$ feltétel mellett.

(b) Határozzuk meg az $Y_1/Z, \ldots, Y_{n-1}/Z$ valószínűségi változók együttes sűrűségfüggvényét!

Tipp:

(a) Alkalmazzuk a valószínűségi változó függvénye eloszlására vonatkozó képletét, kihasznalva, hogy az X és Y valószínűségi változók közötti összefüggés lineáris és a Jakobi determináns értéke 1!

(b) Alkalmazzuk az előző alfeladat eredményét!

Válasz:

(a) $\frac{1}{(n-1)!}\frac{1}{z^n}$, azaz $n-1$ darab független azonos eloszlású a $[0,z]$ intervallumon egyenletes eloszlású valószínűségi változó együttes sűrűségfüggvénye.

(b) $n-1$ darab független azonos eloszlású a $[0,1]$ intervallumon egyenletes eloszlású valószínűségi változó együttes sűrűségfüggvénye.

(iiiiiiiiiiiiiiiii) Legyenek $X_1,\ldots,X_n \sim \mathcal{N}(0,1)$ és $Y_1,\ldots,Y_m \sim \mathcal{N}(0,1)$ független változók, továbbá $T_n^2 := X_1^2+\ldots+X_n^2$ és $T_m^2 := Y_1^2+\ldots+Y_m^2$.

(a) Határozzuk meg X_1^2 sűrűségfüggvényét!

(b) Milyen eloszlású a T_n^2 valószínűségi változó ?

(c) Milyen eloszlású a

$$Z_n := \frac{Y_1}{\sqrt{T_n^2/n}}$$

valószínűségi változó ?

(d) Milyen eloszlású a

$$Z_{n,m} := \frac{mT_n^2}{nT_m^2}$$

valószínűségi változó ?

Tipp:

(a) Határozzuk meg $|X_1|$ sűrűségét, majd alkalmazzuk a valószínűségi változó függvénye eloszlására vonatkozó képletét!

(b) Alkalmazzuk az előző pont eredményét és a ben található abszolút folytonos eloszlások felsorolását.

(c) Alkalmazzuk az előző két pont eredményét és a ben található abszolút folytonos eloszlások felsorolását.

(d) Alkalmazzuk a ben található abszolút folytonos eloszlások felsorolását.

Válasz:

(a) $\chi^2(1)$

(b) $\chi^2(n)$

(c) n szabadságfokú Student $(t(n))$ eloszlású.

(d) (n,m) szabadságfokú F eloszlású.

(iiiiiiiiiiiiiiiii) Legyen $X_1,\ldots,X_{n+1} \sim \mathcal{N}(0,1)$ független minta, továbbá legyen $Y_n = X_2^2+\cdots+X_{n+1}^2$ Milyen eloszlású a $Z_n = \sqrt{n}X_1/\sqrt{Y_n}$ valószínűségi változó

Tipp: Alkalmazzuk a ben található abszolút folytonos eloszlások felsorolását.

Válasz: n szabadságfokú Student $(t(n))$ eloszlású.

(iiiiiiiiiiiiiiiiii) Legyenek $X_n \sim \chi^2(n)$ és $Y_m \sim \chi^2(m)$ független valószínűségi változók. Milyen eloszlású a

$$\tilde{Z}_{n,m} := \frac{mX_n}{nY_m}$$

valószínűségi változó $(n/2,m/2)$ paraméterű béta eloszlású!

Tipp: Alkalmazzuk a ben a két valószínűségi változó hányadosa eloszlására vonatkozó képletet és az abszolút folytonos eloszlások felsorolását.

Válasz: $(n/2, m/2)$-paraméterű béta eloszlású.

(iiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_{n+m}$ független standard normális eloszlású változók. Milyen eloszlású a

$$\tilde{Z}_{n,m} := \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n+m} X_i^2}$$

valószínűségi változó $(n/2, m/2)$ paraméterű béta eloszlású!

Tipp: Alkalmazzuk a ben a két valószínűségi változó hányadosa eloszlására vonatkozó képletet és az abszolút folytonos eloszlások felsorolását.

Válasz: (n, m)-paraméterű F eloszlású.

(iiiiiiiiiiiiiiiiiiiii) Adjuk meg X_n határeloszlását ($n \to \infty$), ha X_n egy n szabadságfokú Stundent eloszlású valószínűségi változó!

Tipp: Elemi analízis.

Válasz: $\mathcal{N}(0,1)$

(iiiiiiiiiiiiiiiiiiiii) Adjuk meg $\frac{X_n - n}{\sqrt{n}}$ határeloszlását ($n \to \infty$), ha X_n egy n szabadságfokú χ^2 eloszlású valószínűségi változó.

Tipp: Alkalmazzuk a t! A szórásnégyzet kiszámításához alkalmazzuk a ben a normális eloszlás páros momentumaira adott formulát.

Válasz: $\mathcal{N}(0,2)$

(iiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$ független azonos eloszlású változók, továbbá $T := \sqrt{X_1^2 + \ldots + X_n^2}$.

(a) Legyen $Z_1 := X_1/T$. Bizonyítsuk be, hogy Z_1^2 és T^2 is függetlenek!

(b) Legyen $Z := \overline{X}/T$. Bizonyítsuk be, hogy Z és T^2 is függetlenek!

Tipp:

(a) A számoláshoz a Bayes-tételt alkalmazzuk. Először meghatározzuk a T^2 statisztika $f(t|y)$ feltételes sűrűségfüggvényét adott $Y_1^2 = y$ esetén. Ez nem más, mint a $\chi^2(n-1)$ eloszlás sűrűségfüggvénye a $t - y$ helyen.

Bayes tétele alapján határozzuk meg az Y_1^2 valószínűségi változó $g(y|t)$ sűrűségfüggvényét adott $T^2 = t$ helyen!

Vegyük észre, hogy a nevezőben a $\chi^2(n-1)$ és a $\chi^2(1)$ eloszlás sűrűségfüggvényeinek a konvolúciója áll, ami a $\chi^2(n)$ eloszlás sűrűségfüggvénye. Így adódik a

$$g(y|t) = C \cdot \frac{(t-y)^{\frac{n-1}{2}-1} y^{-\frac{1}{2}}}{t^{\frac{n}{2}-1}}$$

összefüggés (C normáló tényező).

A Z_1^2 tört $h(y|t)$ feltételes sűrűségfüggvénye adott $T^2 = t$ helyen:

$$h(y|t) = t \cdot g(ty|t) = C \cdot (1-y)^{\frac{n-1}{2}-1} y^{-\frac{1}{2}},$$

ami éppen a $\mathcal{B}(1/2, n/2)$-eloszlású Z^2 valószínűségi változó feltétel nélküli sűrűségfüggvénye.

(b) Először bizonyítsuk be hogy Z^2 és $TY_1^2, \ldots, Y_n^2$ek! Vezessünk $\sim \chi^2(1)$ változókat: $Y_1^2 = n(\overline{X})^2, Y_2^2, \ldots, Y_n^2$ $Y_1^2, \ldots, Y_n^2 = Z_1^2, \ldots, Z_n^2$t úgy, hogy független eloszlásúak legyenek és az egyenlőség teljesüljön. Ez mindig megtehető az

$$Y_2 = \sum_{j=1}^{n} u_{2j} X_j,\ Y_3 = \sum_{j=1}^{n} u_{3j} X_j, \ldots Y_n = \sum_{j=1}^{n} u_{nj} X_j$$

választással, ahol az u_{ij} valós számok ortonormált és az azonosan 1 sorvektorra ortogonális sorvektorok koordinátái. Ezután alkalmazzuk az előző feladat eredményét

Végül a Z^2 és T^2 valószínűségi változók függetlenségből kovetkeztethetünk Z és T valószínűségi változók függetlenségére, felhasználva hogy a számláló sűrűségfüggvénye páros.

Válasz: A fenti számolások valójában fölöslegesek, ha figyelembe vesszük a többdimenziós $\mathbf{I}_p$ kovariancia mátrixú normális eloszlás szimmetriatulajdonságát (l.)

3. Tesztek

(i) Határozzuk meg $E(1/X|X)$-et, ha X tetszőleges véletlen változó és a szükséges várható értékek léteznek.

(a) Nem feltétlenül létezik.

(b) X

(c) $1/X$

(d) $-1/X$

Válasz: (c)

(ii) Határozzuk meg $E(X^2|X)$-et, ha X tetszőleges véletlen változó és a szükséges várható értékek léteznek.

(a) Nem feltétlenül létezik.

(b) $\sqrt{X}$

(c) X

(d) X^2

Válasz: (d)

(iii) Ha X és Y független változók, akkor (ha a szükséges várható értékek léteznek) $E(X+Y|X) =$

(a) $X+Y$.

(b) $E(X+Y)$.

(c) $E(X)+Y$.

(d) $X+E(Y)$.

Válasz: (d)

(iiii) Legyenek $X_1, \ldots, X_n$ független standard normális eloszlású változók. Milyen eloszlású $X_1 + \ldots + X_n$?

(a) standard normális

(b) $N(0,n)$

(c) $N(0,n^2)$

(d) $t(n)$

Válasz: (b)

(iiiii) Legyenek $X_1, \ldots, X_n$ független $\chi^2(m)$ eloszlású változók. Milyen eloszlású $X_1 + \ldots + X_n$?

(a) F(n,m)

(b) F(m,n)

(c) $\chi^2(mn)$

(d) $\chi^2(n+m)$

Válasz: (c)

(iiiiii) Legyenek $X_1, \ldots, X_n$ független λ paraméterű exponenciális eloszlású változók. Milyen eloszlású $X_1 + \ldots + X_n$?

(a) exp($n\lambda$)

(b) Gamma(n, λ)

(c) Béta(n,λ)

(d) másodfajú Béta(n,λ)

Válasz: (b)

(iiiiiii) Melyik igaz?

(a) A különböző szabadságfokú χ^2 eloszlások családja (röviden χ^2 eloszlássereg) és exponenciális eloszlássereg a különböző α, λ paraméterű Gamma eloszlások családja (röviden Gamma eloszlássereg) részei.

(b) A Gamma és χ^2 eloszlásseregek az exponenciális eloszláscsalád részei.

(c) Az exponenciális és Gamma eloszlásseregek a χ^2 eloszlássereg részei.

(d) Egyik eloszlássereg sem része a többi.

Válasz: (a)

2. fejezet - Előismeretek 2.: statisztikai alapok

1. Elméleti háttér

1.1. Az egyváltozós statisztika alapfogalmai

Az alábbiakban röviden összefoglaljuk az egyváltozós statisztikai módszereknek a Tananyagban használt alapfogalmait.

Az egyváltozós statisztikai feladatokat kissé mesterségesen szokás becsléselméletre és hipotézisvizsgálatra osztani. Mindkét feladatkörben megkülönböztetnek paraméteres és nemparaméteres módszereket. A Tananyag ezek közül csak a paraméteres módszerek többváltozós analogonjait és más az egyváltozós statisztikában fel sem merülő módszereket tárgyal. A Tananyag formálisan nem támaszkodik a rendezett minták elméletére, de a rendezett minták implicit módon szinte minden statisztikai módszerben megjelennek, ezért röviden erre is kitérunk.

1.1.1. Alapstatisztikák és rendezett minták

Legyen $X_1, \dots, X_n$ független azonos eloszlású n-elemű minta.

2.1.1.1.1. Definíció. *Az*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

statisztikát mintaátlag*nak nevezzük. Ha hangsúlyozni szeretnénk a mintaelemszámot, akkor az* $\bar{X}_n$ *jelölést használjuk, ha pedig a konkrét realizációkkal számolunk, akkor* $\bar{x}$*-t vagy* $\bar{x}_n$*-t írunk.*

2.1.1.1.2. Definíció. *Az*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

statisztikát empirikus (tapasztalati) szórásnégyzet*nek nevezzük, az*

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

statisztikát pedig korrigált empirikus (tapasztalati) szórásnégyzet*nek. A fenti mennyiségek gyöke az* empirikus (tapasztalati) szórás *illetve a* korrigált empirikus (tapasztalati) szórás, *melyeket* S *illetve* S^* *jelöl.*

A szórásnégyzet, a második momentum és a várható érték közötti összefüggések az alábbi Állításból (mely a merev testek fizikájából jól ismert Steiner-tetel átfogalmazása) következnek

2.1.1.1.3. Állítás (Steiner-tétel). *Az* $x_1, \dots, x_n \in \mathbb{R}$ *rögzített értékekkel és tetszőleges* $c \in \mathbb{R}$ *valós számmal*

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + (\bar{x} - c)^2$$

teljesül.

2.1.1.1.4. Következmény. *A Steiner tételből* $c = 0$ *választással következik, hogy az empirikus szórásnégyzetet a következőképpen is számolhatjuk:*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \overline{X^2} - \bar{X}^2.$$

2.1.1.1.5. Definíció. *Legyen k rögzített pozitív egész. Az*

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

statisztikát k-adik empirikus (tapasztalati) momentum*nak nevezzük, az*

$$M_k^c = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

statisztika pedig a k-adik empirikus (tapasztalati) centrális momentum.

Nyilván $S^2 = M_2^c = M_2 - M_1^2$.

2.1.1.1.6. Definíció. *Legyen $(X,Y)^T$ 2-dimenziós valószínűségi változó, $(X_1,Y_1)^T,\dots,(X_n,Y_n)^T$ pedig vele azonos eloszlású független azonos eloszlású n-elemű minta. Jelölje S_X illetve S_Y a komponensek empirikus szórását! A*

$$C = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}\bar{Y}$$

statisztikát empirikus (tapasztalati) kovarianci*ának, az*

$$R = \frac{C}{S_X S_Y} = \frac{\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}}$$

statisztikát pedig empirikus (tapasztalati) korreláció*nak nevezzük.*

2.1.1.1.7. Definíció. *Az $X_1,\dots,X_n$ mintaelemek értékeit nem-csökkenő sorrendben felvevő $X_1^*, X_2^*, \dots, X_n^*$ valószínűségi változókat* n-elemű rendezett mintának *nevezzük, azaz*

$$X_1^*(\omega) \le X_2^*(\omega) \le \dots \le X_n^*(\omega), \qquad \forall \omega \in \Omega\times\Omega\times\dots\times\Omega = \Omega^n.$$

Tehát minden konkrét $x_1, x_2, \dots, x_n$ realizáció esetén ezt az n valós számot kell nagyság szerint nem csökkenő sorrendbe rendezni, és a nagyság szerint i-ediket x_i^*-gal jelölni. Természetesen az Ω különböző elemeire más és más lesz a mintaelemek sorrendje, és így a rendezés is. Nyilván a rendezett mintaelemek már nem függetlenek egymástól, és nem is azonos eloszlásúak.

2.1.1.1.8. Definíció. *Empirikus mediánon értjük páratlan n ($n = 2k+1$) esetén X_{k+1}^*-ot, páros n ($n = 2k$) esetén pedig $(X_k^* + X_{k+1}^*)/2$-t.*

Ez valójában a középső mintaelem, és amennyiben a realizációból számolt értékét m jelöli, ezzel teljesül a Steiner-tétel L_1- normában vett megfelelője:

2.1.1.1.9. Állítás.

$$\min_{c\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}|x_i - c| = \frac{1}{n}\sum_{i=1}^{n}|x_i - m|.$$

A fenti minimumot a minta *átlagos abszolút eltérés*ének is szokták nevezni.

A mediánnak több előnye is van a várható értékkel szemben.

* Olyan eloszlásoknak is létezik a mediánja, amelyeknek a várható értéke nem létezik.

* A minta mediánja (empirikus medián) az eltolási paraméternek a mintaátlagnál stabilabb becslése, érzeketlen egy-két kiugró adatra.

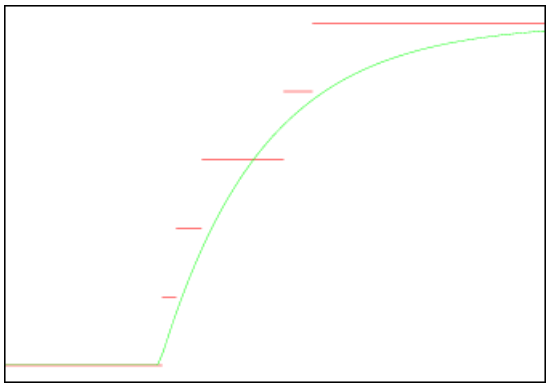
A következőkben egy n-elemű minta alapján kívánjuk közelíteni a háttéreloszlást, ezért megkonstruáljuk az ún. empirikus eloszlásfüggvényt, amiről belátjuk, hogy ``elég nagy" n -re jól rekonstruálja az ismeretlen

eloszlásfüggvényt, akármi is legyen a véletlen minta. Ezt a tényt fogalmazza meg precízen a Glivenko- Cantelli-tétel, melyet a statisztika egyik alaptételének is szoktak tekinteni.

2.1.1.1.10. Definíció (Empirikus (tapasztalati)). *eloszlásfüggvény alatt a következő véletlen függvényt értjük: tetszőleges $x \in \mathbb{R}$ számra legyen*

$$F_n^*(x) := \frac{\sum_{i=1}^n I(X_i < x)}{n} = \begin{cases} 0, & \text{ha} \quad x \le X_1^*, \\ \frac{k}{n}, & \text{ha} \quad X_k^* < x \le X_{k+1}^* \quad (k = 1, \dots, n-1) \\ 1, & \text{ha} \quad x > X_n^*. \end{cases}$$

Itt $I(\cdot)$ az argumentumban álló esemény indikátorváltozója. Könnyű látni, hogy az $I(X_i < x)$ indikátorváltozók független azonos eloszlású Bernoulli eloszlásúak $F(x)$ paraméterrel, ahol F az X háttérváltozó eloszlásfüggvénye.


empirikus eloszlásfüggvény

Megjegyezzük, hogy F_n^* az $x_1, \dots, x_n$ realizációra olyan, mint egy $Y \sim \mathcal{U}(x_1, \dots, x_n)$ diszkrét egyenletes eloszlású valószínűségi változó eloszlásfüggvénye. Nyilván $\mathbb{E}(Y) = \bar{X}$ és $\mathbb{D}^2(Y) = S^2$.

2.1.1.1.11. Tétel (Glivenko- Cantelli-tétel). *Legyen $F(x)$ az elméleti eloszlásfüggvény és $x \in \mathbb{R}$ rögzített. Akkor*

$$\mathbb{E}(F_n^*(x)) = F(x), \quad \mathbb{D}^2(F_n^*(x)) = \frac{F(x)(1-F(x))}{n},$$

és $\lim_{n\to\infty} F_n^*(x) = F(x)$, 1 valószínűséggel.

A bizonyítást ld. [5] 68. o (1.4. Tétel). A tételt animáció is szemlélteti.

Rendezett mintaelemek eloszlása és együttes sűrűsége Legyen most az X háttérváltozó abszolút folytonos eloszlású F eloszlás- és f sűrűségfüggvénnyel. A rendezett mintaelemekre

$$X_1^* < X_2^* < \dots < X_n^*, \qquad 1 \text{ valószínűséggel.}$$

Először határozzuk meg $X_k^* F_{n;k}$-val jelölt eloszlás-, és $f_{n;k}$-val jelölt sűrűségfüggvényét! Nyilván

$$F_{n;k}(x) = \mathbb{P}(X_k^* < x) = \mathbb{P}(\text{legalább } k \text{ db. mintaelem} < x) =$$
$$= \sum_{i=k}^{n} \binom{n}{i} \mathbb{P}(\text{pontosan } i \text{ db. mintaelem} < x) = \sum_{i=k}^{n} \binom{n}{i} [F(x)]^i [1-F(x)]^{n-i} \tag{1}$$

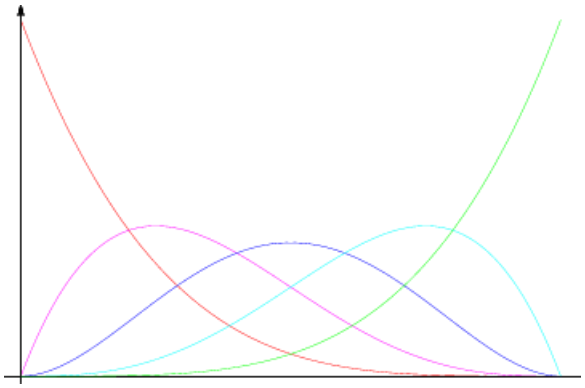
???A sűrűségfüggvényt nem ennek a deriválásával, hanem más meggondolással lehet egyszerűen kiszámolni, a végeredmény:

$$f_{n;k}(x) = n \binom{n-1}{k-1} [F(x)]^{k-1} [1-F(x)]^{n-k} f(x). \tag{1}$$

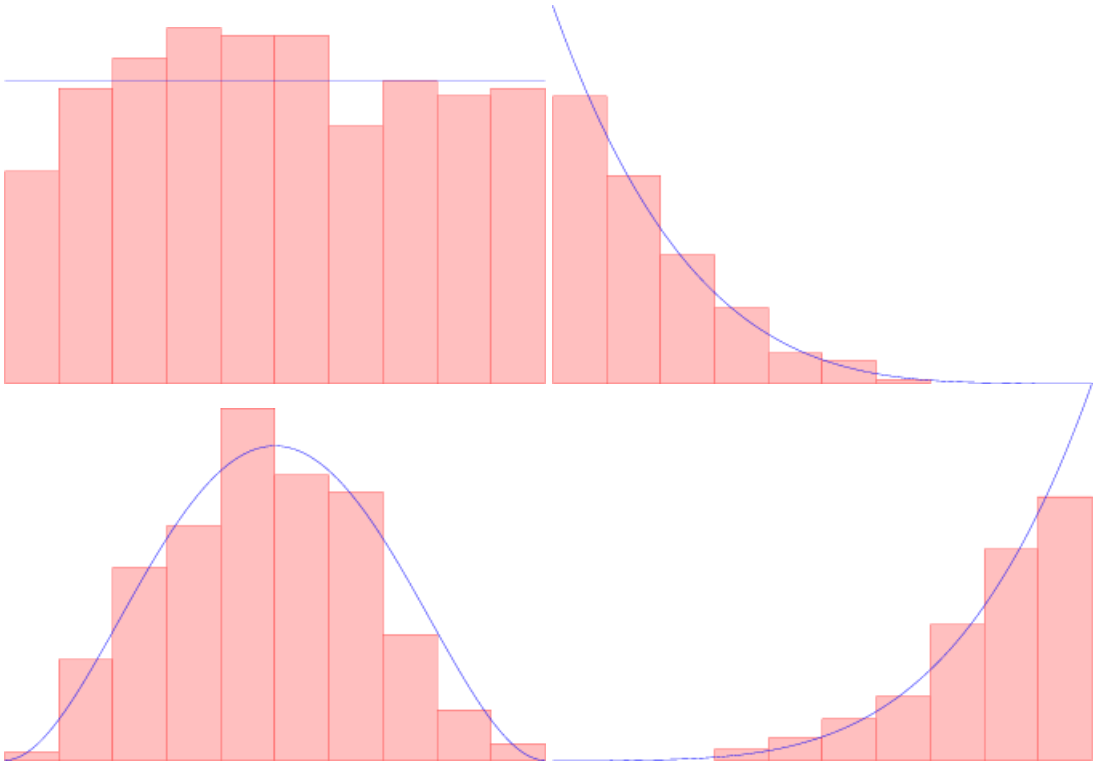
Az $\mathcal{U}[0,1]$ egyenletes eloszlásra alkalmazva a (2.1) formulát és (2.2) formula integrálját 0-tól y-ig a következő értékes összefüggést nyerjük:

$$\sum_{i=k}^{n} \binom{n}{i} y^i (1-y)^{n-i} = n \binom{n-1}{k-1} \int_0^y u^{k-1}(1-u)^{n-k} du.$$

Az egyenletes eloszlásból vett 5 elemű rendezett minta elemeinek sűrűségeit mutatják az alábbi ábrák.



5 elemű rendezett minta elemeinek sűrűségei



Egyenletes minta hisztogramja, 5 elemű rendezett minta 1.,3.,5. elemének hisztogramjai

A alapján látható, hogy az egyenletes eloszlásból vett n-elemű minta Y_k^* k-adik rendezett mintaeleme $\mathcal{B}(k, n-k+1)$ Béta-eloszlású. Ennek alapján meghatározhatók Y_k^* momentumai. Így:

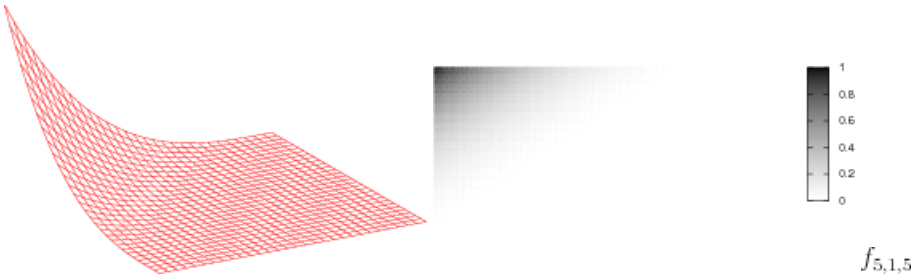
$$\begin{aligned} \mathbb{E}(Y_k^*) &= \frac{k}{n+1} \\ \mathbb{E}(Y_k^*)^2 &= \frac{k(k+1)}{(n+1)(n+2)} \qquad (1) \\ \mathbb{D}^2(Y_k^*) = \mathbb{E}(Y_k^*)^2 - \mathbb{E}^2(Y_k^*) &= \frac{k(n-k+1)}{(n+1)^2(n+2)} \qquad (k=1,\ldots,n). \end{aligned}$$

$X^*_{k_1}, X^*_{k_2}, \dots, X^*_{k_r}$ aka $(1 \le k_1 < k_2 < \dots < k_r \le n)$ együttes sűrűségfüggvényét. Legyenek ezek a mintaelemek: -ét .

$$f_{n;k_1,\dots,k_r}(x_1,\dots,x_r) = \frac{n!}{(k_1-1)!(k_2-k_1-1)!\cdots(k_r-k_{r-1}-1)!(n-k_r)!}\cdot$$
$$\cdot F(x_1)^{k_1-1}[F(x_2)-F(x_1)]^{k_2-k_1-1}\cdots[F(x_r)-F(x_{r-1})]^{k_r-k_{r-1}-1}[1-F(x_r)]^{n-k_r}\cdot$$
$$\cdot f(x_1)\cdots f(x_r), \qquad \text{ha } x_1 \le x_2 \le \dots \le x_r, \tag{1}$$

és nyilván 0 különben.

Az alábbi szürkeárnyalatos ábra $f_{5,1,5}$-öt mutatja egyenletes eloszlásból vett rendezett minta esetén.



$f_{5,1,5}$

Az $r = 1$ speciális esetben megkapjuk a (2.2) képletet. Az $r = n$ speciális esetben megkapjuk az összes rendezett mintaelem együttes sűrűségfüggvényét.

$$f_{n;1,\dots,n}(x_1,\dots,x_n) = \begin{cases} n!f(x_1)\cdots f(x_n), & \text{ha } x_1 \le x_2 \le \dots \le x_n \\ 0, & \text{különben.} \end{cases}$$

Az eredmény nem meglepő, hiszen az összes rendezett mintaelem együttes eloszlása olyan, mint az összes (független) mintaelem együttes eloszlása azzal a különbséggel, hogy a rendezés miatt az előbbi eloszlás $\mathbb{R}^n$-nek az $x_1 \le x_2 \le \dots \le x_n$ egyenlőtlenség által meghatározott, $1/n!$ részarányú szimplexére koncentrálódik.

1.1.2. Elégségesség, teljesség, exponenciális eloszláscsalád

Legyen $\Omega, \mathcal{AP}$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta \colon \theta \in \Theta\}$. Az $X_1, \dots, X_n$ független azonos eloszlású minta egy $T(X_1, \dots, X_n) = T(\mathbf{X})$ **statisztikájában** a mintaelemekben rejlő a θ paraméterre vonatkozó információt sűritjük ösze.

2.1.1.2.1. Definíció. *Likelihood-függvényen értjük a mintaelemek együttes valószínűség illetve sűrűségfüggvényét. Legyen* $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ *rögzített, és* $L_\theta(\mathbf{x})$ *a likelihood-függvény az* $\mathbf{x}$ *helyen. Ha a háttéreloszlás diszkrét* p_θ *valószínűségfüggvényel, akkor*

$$L_\theta(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^n p_\theta(x_i),$$

ha pedig abszolút folytonos f_θ *sűrűségfüggvénynyel, akkor*

$$L_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i).$$

2.1.1.2.2. Definíció. *Azt mondjuk, hogy a* $T(\mathbf{X})$ *statisztika elégséges a* θ *paraméterre, ha diszkrét esetben a*

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t) = \begin{cases} \dfrac{L_\theta(\mathbf{x})}{\mathbb{P}_\theta(T(\mathbf{X}) = t)}, & \text{ha} \quad T(\mathbf{x}) = t, \\ 0 \quad \text{különben} \end{cases} \tag{1}$$

feltételes valószínűség, abszolút folytonos esetben pedig az

$$f_\theta(\mathbf{x}|T(\mathbf{X})=t)=\begin{cases}\frac{L_\theta(\mathbf{x})}{f_\theta^T(t)}, & \text{ha} \quad T(\mathbf{x})=t,\\ 0 & \text{különben}\end{cases} \qquad (1)$$

feltételes sűrűség nem függ θ-tól, $\forall\theta\in\Theta$, ahol $f_\theta^T(t)$ jelöli a $T(\mathbf{X})$ statisztika sűrűségfüggvényét a t helyen.

A fenti definíció alapján látható, hogy az elegséges statisztika a mintaelemekben rejlő a θ paraméterre vonatkozó teljes információt tartalmazza.

Felmerül a kérdés: hogyan lehetne megsejteni egy elégséges statisztika alakját? A választ a következő tétel adja meg.

2.1.1.2.3. Tétel (Neyman- Fisher faktorizáció). *Egy $\mathbf{X}$ minta $T(\mathbf{X})$ statisztikája pontosan akkor elégséges, ha létezik olyan $g_\theta(t)(\theta\in\Theta, t\in\mathcal{T}$ (=T értékkészlete)) és $h(\mathbf{x})(\mathbf{x}\in\mathcal{X})$ mérhető függvény, hogy*

$$L_\theta(\mathbf{x})=g_\theta(T(\mathbf{x}))\cdot h(\mathbf{x})$$

teljesül minden $\theta\in\Theta$, $\mathbf{x}\in\mathcal{X}$ esetén.

Azaz a likelihood-függvény csak a T statisztikán keresztül függ a paramétertől. Bizonyítást ld. [5] 87. o. (3.1. Tétel).

Természetesen a teljes minta vagy a rendezett minta is elégséges statisztika, de mi minél egyszerűbbet szeretnénk kapni. Ezért bevezetünk a valamilyen paraméterre elégséges statisztikák között egy részben rendezést: azt mondjuk, hogy T_1 a T_2-nek *alárendelt statisztika*, ha van olyan mérhető v függvény, hogy $T_1=v(T_2)$. Ezt úgy jelöljük, hogy $T_1\leq T_2$, és a T_1 statisztika ``gazdaságosabb" T_2-nél. Ha T_1 és T_2 kölcsönösen alárendeltek a másiknak, akkor ekvivalenseknek mondjuk őket: $T_1=T_2$ (nyilván ekkor v invertálható függvény).

2.1.1.2.4. Definíció. *A T elégséges statisztikát minimális elégséges statisztikának nevezzük, ha alárendelt statisztikája bármely más elégséges statisztikának.*

2.1.1.2.5. Definíció. *A T statisztika* teljes, *ha a*

$$\mathbb{E}_\theta(g(T))=0, \quad \forall\theta\in\Theta$$

összefüggés a g függvényeknek egy elég gazdag (például folytonosan deriválható) osztályára teljesül, akkor

$$g=0, \quad \mathbb{P}_\theta^T(g=0)=1,$$

ahol $\mathbb{P}_\theta^T$ jelöli a T statisztika által generált mértéket.

Ennnek a tulajdonságnak a jelentősége az, hogy, ha a T statisztika elégséges és teljes akkor minimális elegséges. Ugyanakkor ezt a tulajdonságot nehéz ellnőrizni, de az alább definiált ún. exponenciális eloszláscsaládra teljesül.

2.1.1.2.6. Definíció. *Azt mondjuk, hogy az X háttérváltozó eloszlása tagja az* ***exponenciális eloszláscsalád****nak, ha diszkrét esetben a valószínűség-, abszolút folytonos esetben a sűrűségfüggvénye a következő alakban állítható elő:*

$$c(\theta)\cdot\exp\left[\sum_{j=1}^{k}a_j(\theta)\cdot T_j(x)\right]\cdot h(x), \qquad \forall\theta\in\Theta. \qquad (1)$$

Itt $k=\dim(\Theta)$, c és a_j-k véges, mérhető függvények Θ-n, T_j-k és h pedig véges, mérhető valós függvények.

(A $c>0$ ún. súlyfüggvény biztosítja , hogy a $\sum$ vagy $\int$ 1 legyen).

2.1.1.2.7. Tétel. *Vegyünk egy n-elemű $\mathbf{X} = (X_1, \dots, X_n)$ mintát a fenti eloszlásból. Akkor*

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} T_1(X_i), \dots, \sum_{i=1}^{n} T_k(X_i) \right) \qquad (1)$$

elégséges statisztika a θ paraméter-vektorra.

Ismeretes, hogy a normális-, exponeciális-, Poisson-, Bernoulli-, geometriai- Γ -eloszlások tagjai az exponenciális eloszláscsaládnak. A negatív binomiális (Pascal), binomiális, polinomiális eloszlások csak rögzített rend esetén azok (csak a valószínűség(ek) a paraméter(ek)). A diszkrét és folytonos egyenletes eloszlások viszont nem tagjai.

1.2. Becsléselmélet

1.2.1. Pontbecslések, torzítatlanság, hatásosság, konzisztencia

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. A θ paramétert vagy annak valamely $\psi(\theta)$ függvényét szeretnénk becsülni az $\mathbf{X} = (X_1, \dots, X_n)$ független azonos eloszlású minta alapján konstruált $T(\mathbf{X})$ statisztika segítségével. Jelölje $\hat{\theta}$ ill. $\hat{\psi}$ az így kapott becslést!

2.1.2.1.1. Definíció (Torzítatlanság). $T(\mathbf{X})$ torzítatlan becslés $\psi(\theta)$*-ra, ha*

$$\mathbb{E}_\theta(T(\mathbf{X})) = \psi(\theta), \qquad \forall \theta \in \Theta.$$

Ezt a fogalmat a legegyszerűbb példán szemléltetjük.

2.1.2.1.2. Állítás. $\bar{X}$ *mindig torzítatlan becslés* $m(\theta) = \mathbb{E}_\theta(X)$*-re, ha ez véges.*

2.1.2.1.3. Definíció (Aszimptotikus torzítatlanság). *A* $T(\mathbf{X}_n)$ *statisztikasorozat* aszimptotikusan torzítatlan becslés $\psi(\theta)$*-ra, ha*

$$\lim_{n\to\infty} \mathbb{E}_\theta(T(\mathbf{X}_n)) = \psi(\theta), \qquad \forall \theta \in \Theta.$$

A szórásnégyzet becslésén szemléltetjük mindkét fogalmat.

2.1.2.1.4. Állítás. *Legyen $X_1, \dots, X_n$ független azonos eloszlású minta egy tetszőleges olyan eloszlásból, melyre minden $\theta \in \Theta$ esetén $\sigma^2(\theta) = \mathbb{D}^2_\theta(X) < \infty$. Akkor*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2,$$

$S_n^{*2} := \frac{n}{n-1} S_n^2$ *pedig torzítatlan becslése a szórásnégyzetnek.*

Megjegyezzük, hogy az S_n^{*2} becslés torzítatlansága a Steiner-tétel következménye.

1.2.2. Hatásosság (efficiencia)

2.1.2.2.1. Definíció. *Legyen a T_1 és T_2 statisztika torzítatlan becslés a θ paraméterre, vagy annak valamely $\psi(\theta)$ függvényére. Azt mondjuk, hogy* T_1hatásosabb (efficiensebb) becslés, *mint T_2, ha*

$$\mathbb{D}^2_\theta(T_1) \le \mathbb{D}^2_\theta(T_2), \qquad \forall \theta \in \Theta,$$

és legalább egy $\theta_0 \in \Theta$ esetén (2)-ben $<$ teljesül.

2.1.2.2.2. Definíció. *Egy torzítatlan becslés* hatásos (efficiens) becslés*, ha bármely más torzítatlan becslésnél hatásosabb.*

A következő tétel azt állítja, hogy amennyiben van hatásos becslés, az egyértelmű.

2.1.2.2.3. Tétel ($\psi(\theta)$-Egyértelműségi). *Legyen a T_1 és T_2 statisztika egyaránt torzítatlan, hatásos becslés ugyanarra a paraméterfüggvényre. Akkor*

$$\mathbb{P}_\theta(T_1 = T_2) = 1, \qquad \forall \theta \in \Theta.$$

Tételek garantálják, hogy exponenciális eloszláscsalád esetén $\bar{X}$ a várható érték hatásos becslése. Nem minden eloszláscsalád esetén igaz ez. Az $\mathcal{U}[0,\theta]$ egyenletes eloszláscsalád esetén például legyen $\hat{\theta} X_n^*$ legnagyobb rendezett mintaelem $\frac{n+1}{2n}$-szerese, ez szintén várható érték torzítatlan becslése (l. (18)), és hatásosabb, mint $\bar{X}$

Konzisztencia

A konzisztencia azt jelenti, hogy a megfigyelések számának növelésével javul a becslés pontossága.

2.1.2.2.4. Definíció. *A $T(\mathbf{X}_n)$ statisztikasorozat* gyengén (erősen) konzisztens becslés *$\psi(\theta)$-ra, ha minden $\theta \in \Theta$-ra $n \to \infty$ esetén $T(\mathbf{X}_n) \to \psi(\theta)$ sztochasztikusan (1 valószínűséggel).*

A maga után vonja az alábbi Állítást.

2.1.2.2.5. Állítás. *Ha $X_1, \dots, X_n$ független azonos eloszlású minta X-re és $m(\theta) = \mathbb{E}_\theta(X)$ véges, akkor $\bar{X}_n$ erősen konzisztens becslés $m(\theta)$-ra.*

Ezt szemlélteti az alábbi *animáció*.

2.1.2.2.6. Definíció. *A $T(\mathbf{X}_n)$ statisztikasorozat a $\psi(\theta)$ paraméterfüggvény* négyzetes középben konzisztens becslése, *ha minden $\theta \in \Theta$-ra $\mathbb{E}_\theta(T^2(\mathbf{X}_n)) < \infty$ ($\forall n \in \mathbb{N}$) és*

$$\lim_{n\to\infty} \mathbb{E}_\theta(T(\mathbf{X}_n) - \psi(\theta))^2 = 0.$$

2.1.2.2.7. Állítás. *Ha a $T(\mathbf{X}_n)$ statisztikasorozat négyzetes középben konzisztens becslést ad $\psi(\theta)$-ra, akkor a becslés gyengén konzisztens is.*

A szórásnégyzet becslése konzisztenciájának bizonyításának eszköze az alábbi - önmagában is érdekes - Állítás.

2.1.2.2.8. Állítás.

$$\mathbb{D}^2(S_n^2) = \frac{(n-1)[(n-1)M_4^c - (n-3)\sigma^4]}{n^3},$$

és

$$\mathbb{D}^2(S_n^{*2}) = \frac{1}{n}\left(M_4^c - \frac{n-3}{n-1}\sigma^4\right).$$

Ha egy adott paraméterre nincs torzítatlan becslés, alkalmazó nem a $\hat{\theta}$ becslés szórásnégyzetét, hanem a valódi paraméterértéktől vett távolsága négyzetének $R_\theta(\hat{\theta}) = \mathbb{E}_\theta(|\hat{\theta} - \theta|^2)$ várható értékét, azaz a *négyzetes rizikót* kívánja minimalizálni.

Cramér- Rao-egyenlőtlenség Legyen $(\Omega, \mathcal{P}, \mathcal{P})$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Célunk az, hogy a θ paraméterre vagy annak valamely $\psi(\theta)$ függvényére konstruált torzítatlan becslések szórásnégyzetére alsó korlátot adjunk. Ha egy torzítatlan becslésre ez a korlát eléretik, akkor biztosak lehetünk abban, hogy hatásos becslésünk van, ami az 67 Tétel alapján egyértelmű.

Szükségünk lesz a következő, R. A. Fishertől származó fogalomra, l..

2.1.2.2.9. Definíció. *Legyen $\mathbf{X} = (X_1, \dots, X_n)$ független azonos eloszlású minta az X háttérváltozó eloszlásából, amely a θ paramétertől függ ($\theta \in \Theta$), itt csak a $\dim(\Theta) = 1$, Θ konvex esettel foglalkozunk. A fenti minta* Fisher-féle információ*ja az*

$$I_n(\theta) = \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta} l_\theta(\mathbf{X})\right)^2 \geq 0$$

mennyiséggel van definiálva, ahol

$$l_\theta(\mathbf{x}) = \ln L_\theta(\mathbf{x})$$

az ún. log-likelihood függvény-*t jelöli.*

Az információmennyiségtől elvárjuk, hogy független valószínűségi változók esetén additív legyen. Ez itt nem részletezett regularitási feltételek mellett - amelyek fennálnak az exponenciális eloszláscsaládokra, de például az egyenletes eloszláscsaládra nem állnak fenn - igaz is. Így a definícióban szereplő független azonos eloszlású valószínűségi változók esetén igaz az

$$I_n(\theta) = nI_1(\theta).$$

Ugyanezen regularitási feltételek mellett igaz az $I_1(\theta)$ egyszerűbb kiszámítási módját biztosító

$$I_1(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\ln L_\theta(X)\right)$$

összefüggés.

A következő állítás illusztrálja azt a tényt, hogy az elégséges statisztika tartalmazza a mintában lévő, a paraméterre vonakozó teljes információt.

2.1.2.2.10. Állítás. *Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta egy θ paramétertől függő eloszlásból ($\theta \in \Theta$), és tegyük fel, hogy $I_n(\theta) < \infty$. Akkor tetszőleges $T(\mathbf{X})$ elégséges statisztikára*

$$I_T(\theta) = I_n(\theta),$$

ahol $I_T(\theta)$ ugyanúgy számolható a T statisztika valószínűség ill. sűrűségfüggvényéből, mint ahogyan a teljes minta információja a mintaelemek együttes eloszlásából.

Bizonyítást ld. [5] 109-110. o. (2.2. Állítás).

Miután a Cramér- Rao egyenlőtlenségben szereplő valamennyi fogalmat definiáltunk, kimondhatjuk magát a tételt.

2.1.2.2.11. Tétel (Cramér- Rao-egyenlőtlenség). *Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ reguláris statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, $\dim(\Theta) = 1$. Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta a $\mathbb{P}_\theta$ eloszlásból, amiről most tegyük fel, hogy abszolút folytonos. Tegyük fel továbbá, hogy a $T(\mathbf{X})$ statisztika valamely deriválható ψ függvénnyel képzett $\psi(\theta)$ paraméterfüggvény torzítatlan becslése,*

$$\mathbb{D}^2_\theta(T) < +\infty, \qquad \forall\theta \in \Theta$$

továbbá teljesülnek az alábbi bederiválhatósági feltételek:

$$\frac{\partial}{\partial\theta}\int\cdots\int L_\theta(\mathbf{x})d\mathbf{x} = \int\cdots\int\frac{\partial}{\partial\theta}L_\theta(\mathbf{x})\,d\mathbf{x}, \qquad \forall\theta\in\Theta$$

és

$$\frac{\partial}{\partial\theta}\int\cdots\int T(\mathbf{x})L_\theta(\mathbf{x})d\mathbf{x} = \int\cdots\int T(\mathbf{x})\frac{\partial}{\partial\theta}L_\theta(\mathbf{x})\,d\mathbf{x}, \qquad \forall\theta\in\Theta,$$

ahol $\int\cdots\int$ n-dimenziós integrálást jelent a likelihood-függvény tartóján. Akkor

$$\mathbb{D}^2_\theta(T) \geq \frac{(\psi'(\theta))^2}{I_n(\theta)}, \qquad \forall\theta\in\Theta.$$

Bizonyítást ld. [5] 110-113. o. (2.3. Tétel).

Példaként megemlítjük, hogy az $\mathcal{N}(\theta, \sigma^2)$ normális eloszlásra ismert σ^2 esetén $I_1 = \sigma^{-2}$, és a $\hat{\theta} = \bar{X}$ átlagra az egyenlőtlenség helyett egyenlőség áll, azaz eléretik az információs határ, míg az $\mathcal{E}xp(\lambda)$ exponenciális

eloszlásra a torzítatlan $\hat{\lambda} = \frac{n-1}{nX}$ becslés a következő tétel miatt hatásos, de az információs határ nem éretik el. Ugyanakkor a $\mathcal{U}(0,\theta)$ egyenletes eloszlás

$$\hat{\theta} = X_n^* \quad (\text{a legnagyobb rendezett mintaelem } \frac{n+1}{n}\text{-szerese})$$

becslés szórásnégyzete $1/n$ nagyságrendű, azaz lényegesen **kisebb**, mint az információs határ, mert a bederiválhatósági feltételek nem teljesülnek.

Rizikó értelemben nem mindig a torzítatlan becslés a legjobb: A következő meghökkentő példát James ás Stein (1961) (l.) konstruálták. Legyen $\mathbf{X}k$-dimenziós véletlen vektor $(k > 2)$, melynek komponensei függetlenek és azonos szórásúak (az egyszerűség kedvéért legyenek 1 szórásúak). Vegyünk egyetlen mintát és konstruáljuk a

$$\mathbf{T}(\mathbf{X}) = \mathbf{X}\left(1 - \frac{k-2}{||X||^2}\right)$$

k-dimenziós statisztikát! Ez ugyan nem ad torzítatlan becslést az eloszlás $\underline{\theta}$ várható érték vektorára, de belátható, hogy $R_{\underline{\theta}}(\mathbf{T}) = \mathbb{E}_{\underline{\theta}}(\|\mathbf{T} - \underline{\theta}\|^2) < k$, míg $R_{\underline{\theta}}(\mathbf{X}) = \sum_{j=1^k} \mathbb{D}^2(X_i) = k$, tehát rizikó értelemben jobb becslést ad a fenti $\mathbf{T}$ statisztika, mint a mintaátlag ($R_{\underline{\theta}}(\cdot)$ a korábban bevezetett rizikó többdimenziós általánosítása). Ez azért meglepő, mert a komponensek függetlenek, tehát ha például a normális eloszlású testmagasság, a fénysebesség és egy árucikk árának a várható értékét akarnánk egyszerre becsülni, akkor a James-Stein becslés összehozza a három mintát, és így javít a becslésen.

2.1.2.2.12. Tétel (Rao- Blackwell- Kolmogorov-tétel). *Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}$. Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta valamely $\mathbb{P}_\theta$ eloszlásból. Legyen továbbá*

* *(a) $T(\mathbf{X})$ elégséges statisztika,*

* *(b) $S(\mathbf{X})$ torzítatlan becslés a $\psi(\theta)$ paraméterfüggvényre. Akkor T-nek van olyan $U = g(T)$ függvénye, amely*

* *(1) szintén torzítatlan becslése a $\psi(\theta)$ paraméterfüggvénynek: $\mathbb{E}_\theta(U) = \psi(\theta), \forall\theta \in \Theta$,*

* *(2) U legalább olyan hatásos becslése $\psi(\theta)$-nak, mint S: $\mathbb{D}^2_\theta(U) \leq \mathbb{D}^2_\theta(S), \forall\theta \in \Theta$.*

* *(3) U konstrukciója a következő: $U := \mathbb{E}_\theta(S|T) = g(T(\mathbf{X})), \forall\theta \in \Theta$ (ezt nevezzük ``blackwellizálásnak'').*

Bizonyítást ld. [5] 115-117. o. (3.1. Tétel).

A tétel üzenete: a hatásos becsléseket a minimális elégséges statisztika függvényei közt kell keresni.

1.2.3. Becslési módszerek

A paraméterek (akár többdimenziós paraméterek) becslésére számos ad hoc módszer ismertes, itt csak az ún. maximum-likelihood becslést ismertetjük elsősorban azért, mert általánosan alkalmazható, és az általa kapott eredmény közel esik a más becslések (például az ún. Bayes-becslés, vagy a momentum módszeren alapuló becslés) által kapott eredményhez. Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}$ (a paramétertér lehet többdimenziós és legyen konvex). Vegyünk egy $X_1, \ldots, X_n$ független azonos eloszlású mintát a $\mathbb{P}_\theta$ eloszlásból (θ ismeretlen). Az $x_1, \ldots, x_n$ realizáció birtokában a paraméter becslésének azt a $\hat{\theta}$-ot fogadjuk el, amely mellett annak a valószínűsége, hogy az adott realizációt kapjuk, maximális. Mivel ezt a valószínűséget a likelihood-függvény tükrözi, a módszer ezt maximalizálja. A maximumhely csak a realizációtól függ, tehát statisztikát kapunk becslésként.

2.1.2.3.1. Definíció. *Legyen $L_\theta(\mathbf{x})$: n-elemű mintához tartozó likelihood-függvény. A $\hat{\theta}$: $\hat{\theta}(x_1, \ldots, x_n)$ statisztikát a θ paraméter* maximum likelihood (ML-)becslés*ének nevezzük, ha $\hat{\theta}$ globális maximumhelye a likelihood-függvénynek, azaz*

$$L_{\hat{\theta}(x_1,\ldots,x_n)}(x_1, \ldots, x_n) \geq L_\theta(x_1, \ldots, x_n)$$

teljesül $\forall\theta \in \Theta$ és $(x_1, \ldots, x_n)$ esetén.

Megjegyzés. Ha létezik is L-nek globális maximuma minden realizáció esetén, az nem biztos, hogy a max. helyek egyértelműek. Ezesetben választanunk kell a max. helyek között. Általános tételek biztosítják, hogy $n \to \infty$ esetén a különböző maximum $\hat{\theta}_n$ paraméter valódi értéke θ^* körül $\sqrt{n}(\theta^* - \theta_n)$ gálnak. Tehát a M-L becslés aszimptotikusan torzítatlan, sőt $\mathcal{N}(01/I_1(\theta^*))$ -nel aszimptotikusan normális eloszlású, azaz aszimptotikusan efficiens.

1.2.4. Konfidencia intervallum szerkesztés

Az eddigiekben ún. *pontbecslés*ekkel foglalkoztunk, vagyis a becsülendő paramétert v. paraméterfüggvényt a mintaelemekből képzett egyetlen statisztikával becsültük. Most becslésként egy egész intervallumot - melynek határait természetesen statisztikák jelölik ki - fogunk használni. A köznapi beszédben úgy fogalmazunk, hogy „a $\psi(\theta)$ paraméterfüggvény P valószínűséggel a T_a és T_f statisztikák által meghatározott intervallumban van". Természetesen $\psi(\theta)$ nem valószínűségi változó. Az alábbi kijelentésnek mégis van értelme Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta a $\mathbb{P}_\theta$ sokaságból (θ ismeretlen)!

2.1.2.4.1. Definíció. *A $(T_a(\mathbf{X}), T_a(\mathbf{X}))$ statisztikapárral definiált intervallum legalább $1-\varepsilon$* szintű konfidenciaintervallum *a $\psi(\theta)$ paraméterfüggvényre, ha*

$$\mathbb{P}_\theta(T_a(\mathbf{X}) < \psi(\theta) < T_f(\mathbf{X})) \geq 1 - \varepsilon, \qquad (1)$$

ahol ε előre adott „kis" pozitív szám (például $\varepsilon = 0.05$, $\varepsilon = 0.01$, a hozzájuk tartozó szignifikanciaszint pedig $95\%, 99\%$).

Nem világos, hogy a definícióban szereplő $\mathbb{P}_\theta$ valószínűség milyen paraméterértékhez tartozik.

Egyes szerencsés esetekben az (2.9) beli valószínűség nem függ θ-tól.

Konfidenciaintervallum szerkesztése a normális eloszlás várható értékére ismert szórás esetén

Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$ független azonos eloszlású minta, ahol σ_0^2 ismert, μ (a várható érték) ismeretlen paraméter. $(\bar{X} - r_\varepsilon, \bar{X} + r_\varepsilon)$ szimmetrikus alakban:

$$\mathbb{P}_\mu(\bar{X} - r_\varepsilon < \mu < \bar{X} + r_\varepsilon) = \mathbb{P}_\mu(|\bar{X} - \mu| < r_\varepsilon) = \mathbb{P}_\mu(-r_\varepsilon < \bar{X} - \mu < r_\varepsilon) =$$
$$\mathbb{P}_\mu\left(\frac{-r_\varepsilon}{\sigma_0/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} < \frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) = \Phi\left(\frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{-r_\varepsilon}{\sigma_0/\sqrt{n}}\right),$$

ahol $\Phi(\cdot)$ standard normális eloszlásfüggvény, és r_ε-t úgy kell megválasztani, hogy $2\Phi\left(\frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) - 1 = 1 - \varepsilon$, teljesüljön. Így $r_\varepsilon = \frac{\Phi^{-1}(1-\frac{\varepsilon}{2})\sigma_0}{\sqrt{n}}$.

Vegyük észre, hogy a konfidenciaintervallum hossza n növelésével és a σ_0 szórás csökkentésével csökken.

Ismeretlen szórásnégyzet esetén a a standard normális eloszlást a megfelelő szabadságfokú Student-eloszlással helyettesítjük. Részleteket ld. [5] 129. o. 2. Példa.

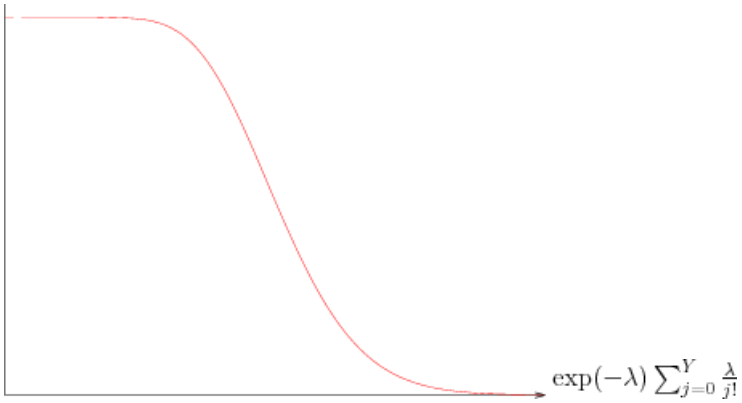
A fenti két esetben az (2.9) képletben $\mathbb{P}_\theta(T_a(\mathbf{X}) < \psi(\theta) < T_f(\mathbf{X})) \geq 1 - \varepsilon$ valószínűség nem függ θ-tól. Ha a feladatot nem lehet θ-tól független szimmetrikus eloszlás valószínűségeire visszavezetni, akkor monoton nem csökkenő $\psi(\theta)$ függvény esetén a következőképpen járunk el. Először önkényesen felbontjuk az (2.9) képletet $\mathbb{P}_{\theta_1}(T_a(\mathbf{X}) > \psi(\theta)) \leq \varepsilon/2$ -re és $\mathbb{P}_{\theta_2}(\psi(\theta) > T_f(\mathbf{X})) \leq \varepsilon/2$ -re. Szavakban kifejezve, ha $\psi(\theta_1)$ értékét csökkentjük, a minta θ_1 melletti valószínűsége, $1 - \varepsilon/$ fölé nő, míg ha $\psi(\theta_2)$ értékét növeljük, a minta θ_2 melletti valószínűsége, $\varepsilon/2$ alá csökken. Az eljárás akkor korrekt, ha a $\theta_a(\varepsilon)$ függvény monoton nem növekvő, míg a $\theta_f(\varepsilon)$ függvény monoton nem csökkenő.

A módszert a Poisson-eloszlás λ paraméterére szerkesztett konfidencia intervallummal illusztráljuk. Legyen $X_1, \ldots, X_n$ ismeretlen λ paraméterű Poisson eloszlásból vett független azonos eloszlású minta, ismeretes, hogy az $Y = X_1 + \cdots + X_n$ összeg elégséges statisztika, és eloszlása $n\lambda$ paraméterű Poisson.

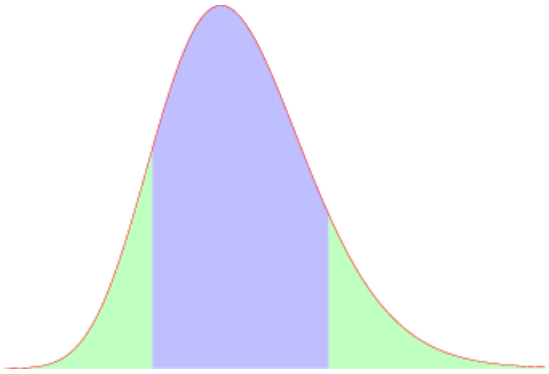
Számítsuk ki azt a λ_a értéket, amire $\exp(-\lambda_a)\sum_{j=0}^{Y}\frac{\lambda_a^j}{j!} = 1 - \varepsilon/2$, majd azt a λ_f értéket, amire $\exp(-\lambda_f)\sum_{j=0}^{Y}\frac{\lambda_a^j}{j!} = \varepsilon/2$,

Nyilván λ csökkentésével a definiáló összeg nő, és λ növelésével a definiáló összeg csökken.

Az alábbi ábra λ függvényében mutatja $\exp(-\lambda)\sum_{j=0}^{Y}\frac{\lambda}{j!}$-t.



A $[\lambda_a, \lambda_f]$ intervallumot tekinthetjük a λ paraméter $1-\varepsilon$ magbízhatósági szintű konfidencia intervallumának. Ezt az alábbi ábra illusztrálja (a kék terült $1-\varepsilon$).


Konfindencia intervallum a Poisson eloszlás λ paraméterére

Az alábbi *interakív ábra* a binomiális eloszlás p paramétere esetén szemlélteti a fenti eljárást.

1.3. Hipotézisvizsgálat

A Tananyagban csak ún. paraméteres hipotézisvizsgálatokkal foglalkozunk. Ez tekinthető a paraméterbecslési feladat egy speciális esetének, amikor előzetes információnk van a paraméter lehetséges értékeiről, és csak azt kell eldönteni, hogy melyik érték a valószínűbb. Valójában a hipotézisvizsgálat majdnem minden feladatát az egyszerű alternatívára vezetjük vissza. Tegyük fel, hogy a Θ paramétertér mindössze két elemből áll: $\Theta = \{\theta_0, \theta_1\}$. $\theta = \theta_0$ hipotézist szokás H_0-lal jelölni és *null-hipotézis*nek nevezni, míg a $H_1 : \theta = \theta_1\}$ az *ellen-hipotézis*. Mindkét hipotézis lehet összetett is: a Θ paramétertartományt két halmaz diszjunkt uniójára $(\Theta = \Theta_0 \cup \Theta_1$ és $\Theta_0 \cap \Theta_1 = \emptyset)$. Leggyakrabban a null-hipotézis egyszerű $\theta = \theta_0$, míg az ellenhipotézis $\theta \neq = \theta_0$ alakú.

Döntésünkkor kétféle hibát követhetünk el:

1. Elvetjük a null-hiptézist, pedig igaz; ezt nevezzük elsőfajú hibának, mert ennek a valószínűsége egyszerű null-hipotézis esetén null-hipotézishez tartozó eloszlás alapján kiszámolható. *A hipotézisvizsgálat a gyakorlatban legtöbbször úgy történik, hogy keresünk a mintaelemeknek egy olyan függvényét, amelynek eloszlása az egyszerű null-hipotézis fennállása esetén ismert. Ez a próbastatisztika. (ha szerencsénk van, az ellen-hipotézishez tartozó paraméterértékekre is ismert)*

2. Elfogadjuk a null-hiptézist, pedig nem igaz; ezt nevezzük másodfajú hibának, ennek a valószínűsége összetett H_1 hipotézis esetén függ a $\theta \in \Theta_1$ paramétertől.

Döntésünk valamely, az $\mathbf{X} = (X_1, \ldots, X_n)$ minta alapján lehet determinisztikus, $\mathcal{X}$'s (diszkrét értékkészlű $\mathcal{X}_e$ valószínűségi $\mathcal{X}_k$ltozók esetén) ún. randomizált. A determinisztikus döntéskor a mintateret felosztjuk *elfogadási-* és **kritikus tartomány**ra.

$$\mathcal{X}_e \cap \mathcal{X}_k = \emptyset, \qquad \mathcal{X}_e \cup \mathcal{X}_k = \mathcal{X}.$$

Az elsőfajú hiba valószínűsége egyszerű null-hipotézis esetén:

$$\mathbb{P}_{\theta_0}(\mathbf{X} \in \mathcal{X}_k).$$

A hipotézisvizsgálatban a döntést **próbá**nak nevezik.

A kritikus tartományt leggyakrabban ún. Ψ**próbafüggvénnyel** definiáljuk:

$$\begin{cases} \mathbf{X} \in \mathcal{X}_e \Leftrightarrow \Psi(\mathbf{X}) = 0, \\ \mathbf{X} \in \mathcal{X}_k \Leftrightarrow \Psi(\mathbf{X}) = 1. \end{cases}$$

Előfordulhat, hogy ilyen alakú próbafüggvénnyel még egyszerű alternatíva esetén sem lehet minden ε értékére pontosan beállítani az elsőfajú hibát, sőt a mintateret sem lehet két diszjunkt tartományra osztani úgy hogy az elsőfajú hiba adott ε legyen. Ilyenkor háromértékű (randomizált) próbafüggvényt alkalmazunk:

$$\Psi(\mathbf{X}) = \begin{cases} 0, \\ p, \\ 1, \end{cases}$$

Ha $\Psi(\mathbf{X}) = p$, akkor a nullhipotézist p valószínűséggel elfogadjuk.

Ha a null-hipotézis összetett a próba terjedelméről beszélünk.

2.1.3.1. Definíció. *A $\mathcal{X}_k$ kritikus próba **pontos terjedelme**:*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X} \in \mathcal{X}_k).$$

A pontos terjedelem diszkrét eloszlások esetén általában nem érhető el.

2.1.3.2. Definíció. *Az $\mathcal{X}_k$ kritikus tartománnyal értelmezett* próba ereje *a $\theta \in \Theta_1$ alternatívával szemben:*

$$\beta_n(\theta, \varepsilon) = 1 - \mathbb{P}_\theta(\mathbf{X} \in \mathbb{X}_e) = \mathbb{P}_\theta(\mathbf{X} \in \mathbb{X}_k), \qquad \theta \in \Theta_1$$

teljesül.

A próbák esetén is definiálható a torzítatlanság, nevezetesen, ha erőfüggvénye az ellen-hipotézishez tartozó paraméterértekre sem kisebb, mint a próba terjedelme. Precízen fogalmazva:

2.1.3.3. Definíció. *Az $\mathbb{X}_k$ kritikus tartománnyal definiált próba legfeljebb ε terjedelmű **torzítatlan**, ha*

$$\mathbb{P}_\theta(\mathbf{X} \in \mathbb{X}_k) \leq \varepsilon, \quad \text{ha} \quad \theta \in \Theta_0,$$

és

$$\mathbb{P}_\theta(\mathbf{X} \in \mathbb{X}_k) \geq \varepsilon, \quad \text{ha} \quad \theta \in \Theta_1.$$

Rögzített terjedelem esetén elvárható, hogy a mintaelemszám növelésével próba másodfajú hibája az ellen-hipotézishez tartozó minden paraméterértékre nullához tartson.

2.1.3.4. Definíció. *Az n elemű mintához tartozó $\mathbb{X}_k^{(n)}$ kritikus tartománnyal definiált próba ε terjedelmű **konzisztens**, ha*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X}_n \in \mathbb{X}X_k^{(n)}) = \varepsilon, \quad \forall n \in \mathbb{N}$$

és

$$\lim_{n\to\infty} \beta_n(\theta,\varepsilon) = \lim_{n\to\infty} \mathbb{P}_\theta(\mathbf{X}_n \in \mathbb{X}_k^{(n)}) = 1, \quad \forall\theta \in \Theta_1.$$

A hipotézisvizsgálat legalapvetőbb tétele az egyszerű alternatívára érvényes Neyman- Person-Lemma.

2.1.3.5. Tétel (Neyman- Pearson-Lemma). *A*

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1$$

egyszerű alternatívára tetszőleges $\varepsilon > 0$-ra létezik ε terjedelmű próba, amelynek másodfajú hibája minimális, amelynek (esetleg randomizált) próbafüggvénye

$$\psi(\mathbf{X}) = \begin{cases} 0, & \text{ha} \quad \frac{L_{\theta_1}(\mathbf{X})}{L_{\theta_0}(\mathbf{X})} < c, \\ p, & \text{ha} \quad \frac{L_{\theta_1}(\mathbf{X})}{L_{\theta_0}(\mathbf{X})} = c, \\ 1, & \text{ha} \quad \frac{L_{\theta_1}(\mathbf{X})}{L_{\theta_0}(\mathbf{X})} > c, \end{cases} \qquad (1)$$

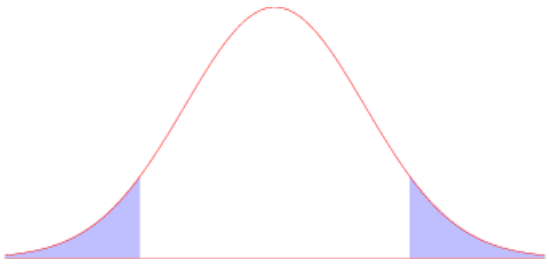
ahol a $L_{\theta_j}(\mathbf{X}) \quad j = 0,1$ és a $c = c_\varepsilon > 0$ és $p = p_\varepsilon$ számokat úgy választjuk meg, hogy a próba terjedelme ε legyen

Bizonyítást ld. [5] 144-149. o (1.1. Tétel).

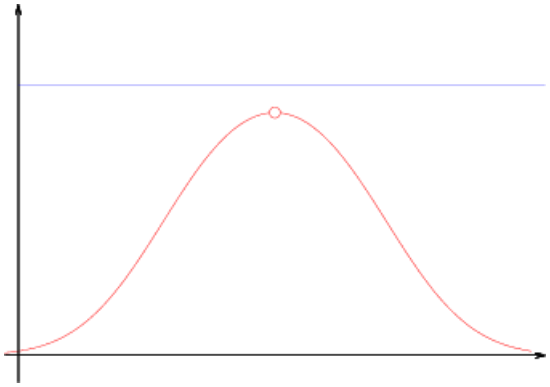
2.1.3.6. Megjegyzés. *Diszkrét eloszlás esetén általában nincs olyan c érték, amire a determinisztikus próba elsőfajú hibája pontosan ε ezért randomizált próbát alkalmazunk. Természetesen megtehetjük, hogy ,,szigorúak" vagyunk és szűkebb kritikus tartományt (kisebb c-t) választunk, vagy a kisebb elsőfajú hiba előnyosebb, és engedékenyebbek vagyunk.*

Az elméleti összefoglalóban egyetlen példát mutatunk arra az esetre, amikor a Neyman- Pearson-lemma alapján próba szerkeszthető. Ez az ún. **egymintás u-próba**.

Legyen $\mathbf{X}$: $X_1, \ldots, X_n$ független azonos eloszlású $\mathcal{N}(\theta, 1)$ eloszlású minta, θ lehetséges értékei θ_0 (null-hipotézis) és $\theta_1 > \theta_0$ (ellen-hipotézis). A normális eloszlás sűrűségfüggvényének alakjából kiolvasható, hogy a $\frac{L_{\theta_1}(\mathbf{X})}{L_{\theta_0}(\mathbf{X})} \geq c$ egyenlőtlenség pontosan akkor teljesül ha $\sqrt{n}\overline{\mathbf{X}} \geq c'$, ahol c'-t ugy kell megválasztani, hogy $\mathcal{P}(\sqrt{n}\overline{\mathbf{X}} > c') = \varepsilon$ teljesüljön. Mivel Mivel $\sqrt{n}\overline{\mathbf{X}}$ standard normális eloszlású, $c' = \Phi^{-1}(1-\varepsilon)$. A megfelelő kvantiliseket itt *interaktív ábra* segítségével határozhatjuk meg.

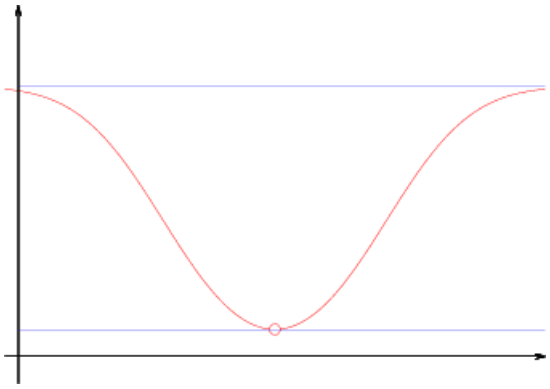


u próba elsőfajú hibája



u próba másodfajú hibája μ függvényében

Az erőfüggvény mutatja az u próba konzisztenciáját (az alsó kék vonal az elsőfajú hibánál, a felső 1-nél van).



u próba ereje ($1-$másodfajú hiba) μ függvényében

Az alábbi *animáció* az u próba konzisztenciáját mutatja. A Neyman- Pearson-lemma randomizált változata alapján szerkesztendő próba a feladatok között szerepel. Végül mutatunk egy általanosan használt módszert, amely számos módszer alapját képezi, és a többváltozós statisztikában más lehetőség híján mindig ezt alkalmazzuk.

1.3.1. A Likelihood-hányados próba

Ez a fajta próba olyan, viszonylag általános esetekben használható, mikor a null-hipotézis azt jelenti, hogy paraméterünk a véges dimenziós, konvex paramétertér valamely alacsonyabb dimenziós, összefüggő részsokaságába esik:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1,$$

ahol $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$, és a $\dim(\Theta_0) = r$, $\dim(\Theta) = k$ jelöléssel $r < k$ teljesül. Az n-elemű minta alapján konstruálandó próbastatisztika:

$$\lambda_n(\mathbf{X}) = \frac{\sup\limits_{\theta\in\Theta_0} L_\theta(\mathbf{X})}{\sup\limits_{\theta\in\Theta} L_\theta(\mathbf{X})}.$$

Tényleg statisztikát kapunk ($\lambda_n(\mathbf{X})$ nem függ θ-tól), amely 0 és 1 közötti értékeket vesz fel.

2.1.3.1.1. Állítás. *Bizonyos regularitási feltételek mellett $n \to \infty$ esetén*

$$-2\ln\lambda_n(\mathbf{X}) \to \chi^2(k-r)$$

eloszlásban, H_0 fennállása esetén. (l. [3] 3.10 paragrafus)

Ezért ε terjedelemhez a kritikus tartomány:

$$\mathbb{X}_k = \{\mathbf{x} : \lambda_n(\mathbf{x}) \le \lambda_\varepsilon\} = \{\mathbf{x} : -2\ln\lambda_n(\mathbf{x}) \ge c_\varepsilon\},$$

ahol a $c_\varepsilon = -2\ln\lambda_\varepsilon > 0$ konstans a $\chi^2(k-r)$ eloszlás $1-\varepsilon$ kvantilise.

1.3.2. A leggyakrabban használt próbák

t-próba (Student-próba). Normális eloszlás várható értékének tesztelésére vagy két normális várható érték összehasonlítására használják ismeretlen szórás(ok) esetén. A gyakorlatban kis mintákra alkalmazzák, a normális eloszlást fel kell tenni.

Egymintás t-próba. Legyen $X \sim \mathcal{N}(\mu, \sigma^2)$ háttérváltozó ismeretlen paraméterekkel. A

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

hipotézis vizsgálatára az n elemű $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$ független, azonos eloszlású mintából konstruált próbastatisztika:

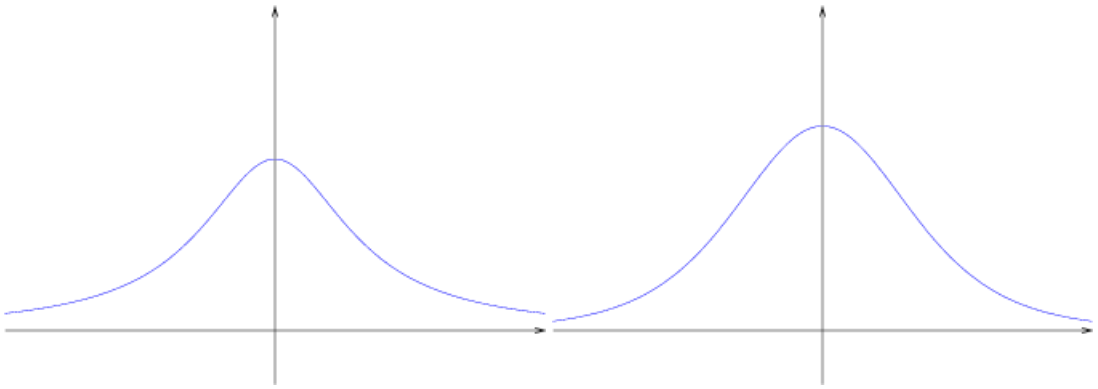
$$t(\mathbf{X}) = \frac{\bar{X} - \mu_0}{S_n^*}\sqrt{n},$$

az $1-\varepsilon$ szignifikanciaszinthez konstruált kritikus tartomány pedig

$$\mathcal{X}_k = \{\mathbf{x} : |t(\mathbf{x})| \geq t_{\varepsilon/2}(n-1)\},$$

ahol $t_{\varepsilon/2}(n-1)$ az $n-1$ szabadságfokú t-eloszlás $(1-\varepsilon/2)$-kvantilise.

Az alábbi ábrák mutatják az 1, és 5 szabadságfokú Student (t) eloszlásokhoz tartozó sűrűségfüggvényeket.



t(1) és t(5) sűrűségek

A t-eloszlások kvantiliset itt *interaktív ábra* segítségével tudjuk meghatározni.

Null-hipotézisünket $1-\varepsilon$ szinten elfogadjuk, ha a mintarealizációból számolt $|t(\mathbf{x})| < t_{\varepsilon/2}(n-1)$, és elutasítjuk különben.

Kétmintás t-próba. Legyen $X \sim \mathcal{N}(\mu_1, \sigma^2)$ és $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ két tetszőleges várható értékű, de azonos szórású háttérváltozó. Az összes paraméter ismeretlen. Még ebben a paragrafusban megmutatjuk, hogyan lehet ismeretlen szórások egyenlőségét tesztelni. A

$$H_0 : \mu_1 = \mu_2 \quad \text{vers.} \quad H_1 : \mu_1 \neq \mu_2$$

hipotézis vizsgálatára az n_1 elemű $X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ független, azonos eloszlású és az $Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$ független, azonos eloszlású, egymástól is független mintákból konstruált próbastatisztika:

$$t(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1-1){S_X^*}^2 + (n_2-1){S_Y^*}^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}}$$

az $1-\varepsilon$ szignifikanciaszinthez konstruált kritikus tartomány pedig

$$\mathcal{X}_k = \{(\mathbf{x}, \mathbf{y}) : |t(\mathbf{x}, \mathbf{y})| \geq t_{\varepsilon/2}(n_1 + n_2 - 2)\},$$

ahol most az $n_1 + n_2 - 2$ szabadsági fokú t-eloszlást használjuk. A t-eloszlások kvantiliset itt *interaktív ábra* segítségével tudjuk meghatározni.

F-próba. Két normális eloszlású változó szórásának összehasonlítására használják. Legyen $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ és $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ két ismeretlen paraméterű, normális eloszlású háttérváltozó. A szórások egyenlőségét szeretnénk tesztelni:

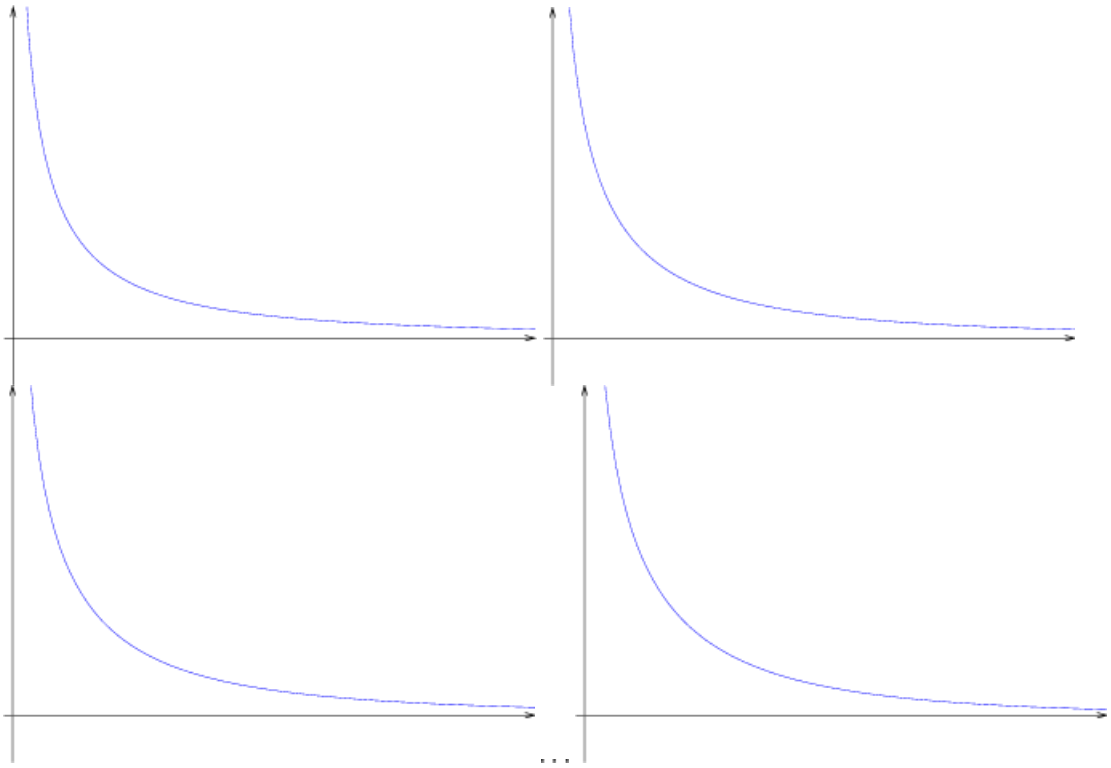
$$H_0 : \sigma_1 = \sigma_2 \quad \text{versus} \quad H_1 : \sigma_1 \neq \sigma_2.$$

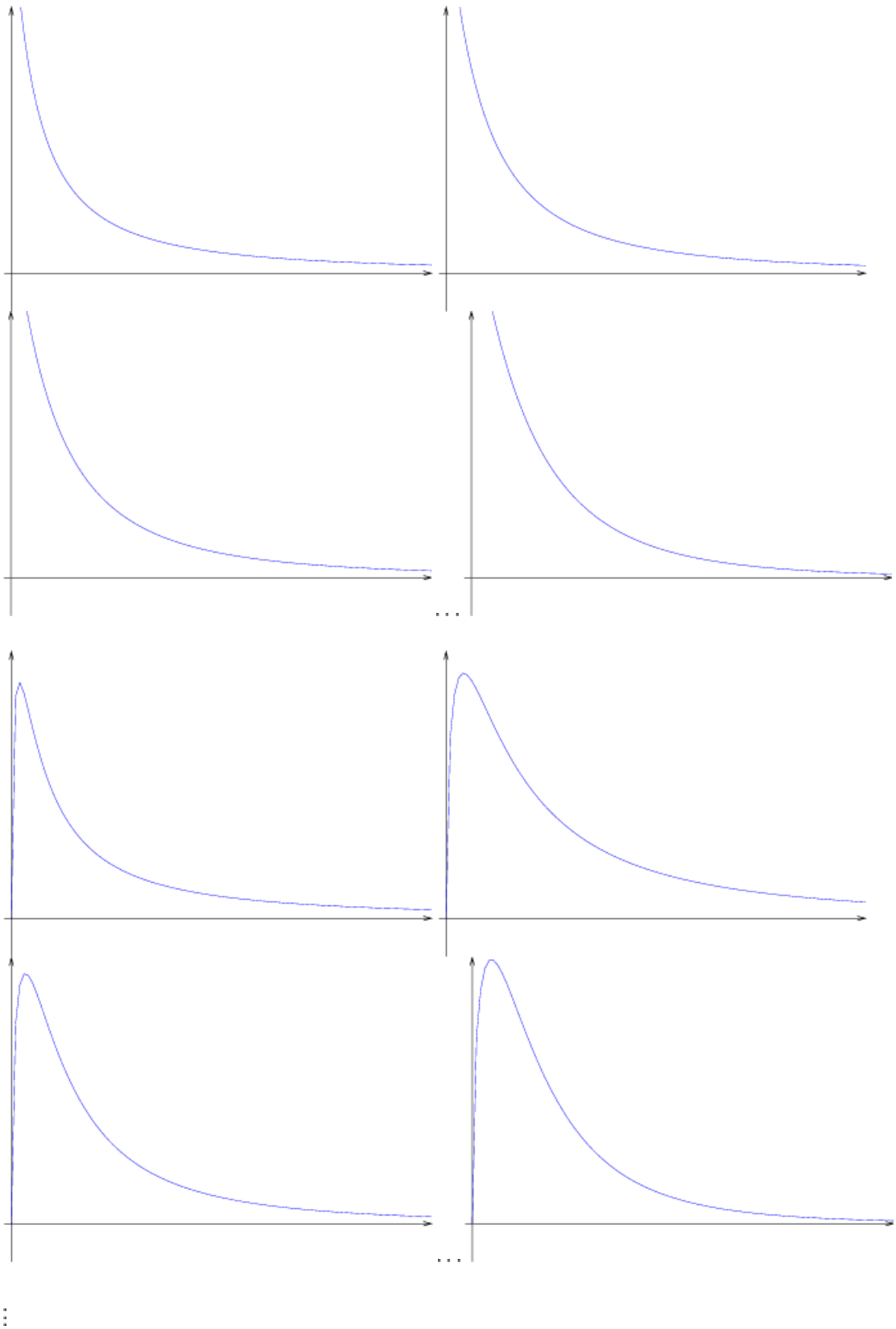
Az n_1 elemű $X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ független, azonos eloszlású és az $Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$ független, azonos eloszlású, egymástól is független minták alapján vizsgálódunk. Tudjuk, hogy $(n_1 - 1){S_X^*}^2/\sigma_1^2 \sim \chi^2(n_1 - 1)$ és $(n_2 - 1){S_Y^*}^2/\sigma_2^2 \sim \chi^2(n_2 - 1)$ függetlenek. Leosztva őket külön-kölön a saját szabadsági fokukkal, majd a hányadosukat véve $\mathcal{F}(n_1, n_2)$-eloszlású valószínűségi változót kapunk, ezt tekinthetjük egyben az (n_1, n_2) szabadsági fokú Fisher-eloszlás definíciójának. H_0 fennállása esetén a hányados

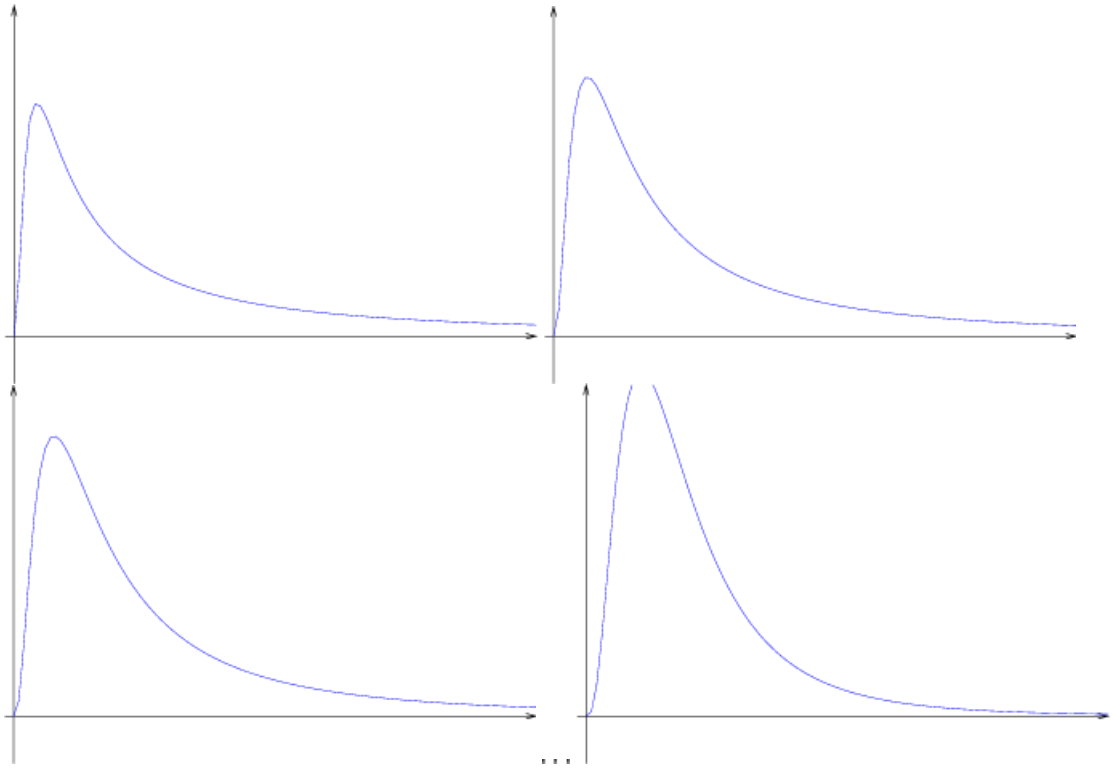
$$F(\mathbf{X}, \mathbf{Y}) = \frac{{S_X^*}^2}{{S_Y^*}^2},$$

így ezt a próbastatisztikát vezetjük be. Mivel egy $\mathcal{F}(f_1, f_2)$ eloszlású valószínűségi változó reciproka $\mathcal{F}(f_2, f_1)$ eloszlású lesz, az X, Y szereposztást úgy választhatjuk, hogy a konkrét realizáció alapján számolt ${s_X^*}^2 \geq {s_Y^*}^2$ legyen. Ezután $1 - \varepsilon$ szinten elutasítjuk H_0-t, ha $F(\mathbf{x}, \mathbf{y}) \geq F_{\varepsilon/2}(n_1 - 1, n_2 - 1)$, ahol a megfelelő szabadsági fokú F-eloszlás $(1 - \varepsilon/2)$-kvantilise a kritikus érték.

Az alábbi ábrák mutatják az (1,1), (1,2), (1,3), (1,9), (2,1), (2,2), (2,3), (2,9), (3,1), (3,2), (3,3), (3,9), (9,1), (9,2), (9,3) és (9,9) szabadságfokú F eloszlásokhoz tartozó sűrűségfüggvényeket.







F sűrűségek

Az F-eloszlások kvantiliset itt *interaktív ábra* segítségével tudjuk meghatározni.

A következő két próba ún. nemparaméteres próba, az első esetben a H_0 hipotézis az, hogy a minta egy adott diszkrét eloszlást követ, míg a második esetben a H_0 hipotézis az, hogy a minta egy adott folytonos eloszlást követ.

χ^2-próba. Legyen $A_1, \dots, A_r$ teljes eseményrendszer és

$$H_0 : \mathbb{P}(A_i) = p_i \quad (i = 1, \dots, r),$$

ahol a $p_i > 0$, $\sum_{i=1}^r p_i = 1$ valószínűségek adottak. Végezzünk n db. megfigyelést! Jelölje $\nu_1, \dots, \nu_r$ az $A_1, \dots, A_r$ esemény gyakoriságát ($\sum_{i=1}^r \nu_i = n$)! Akkor H_0 fennállása esetén a $(\nu_1, \dots, \nu_r)$ valószínűségi változó polinomiális eloszlású:

$$\mathbb{P}_{H_0}(\nu_1 = n_1, \dots, \nu_r = n_r) = \begin{cases} \frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \cdots p_r^{n_r}, & \text{ha } n_1 + \dots + n_r = n, \\ 0, & \text{különben.} \end{cases}$$

A alábbi tétel biztosítja, hogy a az $\sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i}$ próbafüggvény aszimptotikusan χ^2-eloszlású.

2.1.3.2.1. Tétel. *Ha $(\nu_1, \dots, \nu_r)$ polinomiális eloszlású n és $p_1, \dots, p_r$ ($p_i > 0$) paraméterekkel (vagyis a (3.1)-beli H_0 fennállása esetén), akkor $n \to \infty$ esetén*

$$\sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i} \to \chi^2(r-1)$$

eloszlásban.

A χ^2-eloszlások kvantiliset itt *interaktív ábra* segítségével tudjuk meghatározni.

Megjegyzés. A határeloszlás nem függ a p_i értékektől, csak r-től.

Kolmogorov- Szmirnov-próba. Ezt a próbát tiszta illeszkedésvizsgálat céljára használjuk olyan esetekben, mikor a háttéreloszlás folytonos. A próbastatisztika konstrukciójánál kihasználjuk a Kolmogorov- Szmirnov tételkört.

Egymintás eset (illeszkedésvizsgálat):

$$H_0 : \mathbb{P}(X < x) = F(x), \quad \forall x \in \mathbb{R}$$

(F adott folytonos eloszlásfüggvény).

$$H_1 : \text{ van olyan } x \in \mathbb{R}, \mathbb{P}(X < x) \neq F(x).$$

Jelölje F^* a tapasztalati eloszlást és legyen

$$D_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)|.$$

Amennyiben $x_1^* \leq \cdots \leq x_n^*$ az $\mathbf{x} = (x_1, \ldots, x_n)$ mintarealizáció rendezett alakja, akkor

$$D_n(\mathbf{x}) = \max_i \max\{|F_n^*(x_i^*) - F(x_i^*)|, |F_n^*(x_i^* + 0) - F(x_i^*)|\} =$$
$$= \max_i \max\{|\frac{i-1}{n} - F(x_i^*)|, |\frac{i}{n} - F(x_i^*)|\}.$$

Kolmogorov tétele alapján tudjuk, hogy H_0 fennállása esetén

$$\lim_{n \to \infty} \mathbb{P}(\sqrt{n} D_n < z) = K(z), \quad \forall z \in \mathbb{R},$$

ahol

$$K(z) = \begin{cases} 0, & \text{ha} \quad z \leq 0, \\ \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 z^2} = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2}, & \text{ha} \quad z > 0, \end{cases}.$$

Részleteket ld. [5] 73-79. o. A Kolmogorov-eloszlás kvantiliset itt *interaktív ábra* segítségével tudjuk meghatározni.

2. Feladatok

(i) Igaz-e, hogy a tapasztalati korreláció mindig -1 és 1 közé esik? Mikor teljesülhet valamelyik egyenlőség?

Tipp: Alkalmazzuk a véges dimenzós Cauchy- Schwarz-egyenlőtlenséget!

Válasz: Igaz.

$$\begin{cases} 1, & \text{ha a két minta egymás pozítív számszorosa,} \\ -1, & \text{ha a két minta egymás negatív számszorosa.} \end{cases}$$

(ii) Legyen $X_1, \ldots, X_n$ független, p paraméterű Bernoulli eloszlásból vett statisztikai minta.

(a) Milyen eloszlású $\sum_{i=1}^n X_i$?

(b) Adjuk meg a k-adik empirikus (tapasztalati) momentum eloszlását!

(c) Adjuk meg a második empirikus (tapasztalati) centrális momentum eloszlását!

Tipp:

(a) Elemi számolás.

(b) A diszkrét eloszlású valószínűségi változók függvénye eloszlásának számolása.

(c) Alkalmazzuk az előző 2 pont eredeményét $k = 1, 2$-re.

Válasz:

(a) $\mathcal{B}_n(p)$.

(b) Az $n^k/n, (n-1)^k/n, \dots, 1/n, 0$ számok valószínűségei ugyanazok, mint a $\mathcal{B}_n(p)$ eloszlásban az $n, n-1, \dots, 1, 0$ értékek valószínűségei.

(c) $\left(n - \frac{(n+1)}{2n}\right)^2, \dots, \left(-\frac{(n+1)}{2n}\right)^2$ számok valószínűségei ugyanazok, mint a $\mathcal{B}_n(p)$ eloszlásban az $n, n-1, \dots, 1, 0$ értékek valószínűségei.

(iii) Legyen $X_1, \dots, X_n$ független, $\lambda_1, \dots, \lambda_n$ paraméterű Poisson eloszlásból vett minta.

(a) Milyen eloszlású $\sum_{i=1}^n X_i$?

(b) Adjuk meg $\overline{X}$ eloszlását!

Tipp: Alkalmazzuk a t.

Válasz:

(a) $n\lambda$ paraméterű Poisson.

(b) A $\{0, 1/n, 2/n, \dots\}$ értékeket ugyanazzal a valószínűségel veszi fel, mint az $n\lambda$ paraméterű Poisson-eloszlás.

(iiii) Legyen $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független minta. Milyen eloszlású $\overline{X}$? (Adjuk meg a várható értéket és a szórásnégyzetet is!)

Tipp: l. .

Válasz: $\mathcal{N}(\mu, \sigma^2/n)$.

(iiiii) Legyen $X_1, \dots, X_n \sim \mathcal{U}(-1, 1)$ független minta. Aszimptotikusan milyen eloszlású $\sqrt{n} \cdot \overline{X}$?

Tipp: Számítsuk ki a $\mathcal{U}(-1, 1)$ eloszlás első két momentumát és alkalmazzuk a t.

Válasz: $\mathcal{N}(0, 1/3)$.

(iiiiii) Legyen $X_1, \dots, X_n$ független minta $f(x) = \frac{1}{2\sqrt{2}} e^{-\sqrt{2}|x|}$ sűrűségfüggvénnyel. Aszimptotikusan milyen eloszlású $\sqrt{n} \cdot \overline{X}$?

Tipp: A feladatban szereplő valószínűségi változók várható értéke 0, szórasnégyzetet jelölje σ^2, ez utóbbit az exponenciális eloszlás sűrűségfüggvényének és második momentumának ismeretében kiszámíthatjuk. Alkalmazzuk a t.

Válasz: Vegyük észre, hogy $f(x)$ a teljes számegyenesen van értelmezve! $\mathcal{N}(0, 1)$.

(iiiiiii) Legyen $X_1, \dots, X_n$ független, λ paraméterű exponenciális eloszlásból vett minta. Milyen eloszlású $\overline{X}$?

Tipp: keressük meg a ben a gamma eloszlás sűrűségfüggvényét.

Válasz: $\mathcal{G}(n, n\lambda)$.

(iiiiiiii) Számoljuk ki az n-edrendű λ paraméterű gamma eloszlás $-k$-adik momentumát, ahol $k < n$.

Tipp: Számitsuk ki az $\int_0^\infty X^{-k} f(x) dx$ integrált, ahol $f(x)$ a $\mathcal{G}(n, \lambda)$ eloszlás sűrűségfüggvénye. Használjuk ki azt a tényt, hogy $x^{-k} f(x) \mathcal{G}(n-k, \lambda)$ sűrűségfüggvényének konstansszorosa (l. abszolút folytonos eloszlások).

Válasz: $\frac{\lambda^k}{(n-1)\dots(n-k)}$

(iiiiiiiii)

(iiiiiiiii) Legyen $X_1^* < \ldots < X_n^*$ a $[0,1]$ intervallumon egyenletes eloszlásból vett rendezett minta.

(a) Igazoljuk, hogy $X_1^*, \ldots, X_n^*$ nem függetlenek!

(b) Igazoljuk, hogy $1 - X_n^*, \ldots, 1 - X_1^*$ szintén a $[0,1]$ intervallumon egyenletes eloszlásból vett rendezett minta!

(c) Milyen eloszlású $X_{k+1}^* - X_k^*$, ahol $1 \leq k < n$?

Tipp:

(a) Elemi logika.

(b) Hivatkozzunk a egyenletes eloszlás szimmetriájára.

(c) l. elemeinek együttes sűrűségfüggvénye.

Válasz:

(a) Ha például $X_1^* = 0,001$, akkor X_2^* felveheti a 0,002 értéket, míg ha $X_1^* = 0,99$, akkor X_2^* nem veheti fel a 0,002 értéket, azaz X_2^* feltételes eloszlása X_1^*-ra nézve függ X_1^* értékétől.

(b) Mivel az egyenletes eloszlás szimmetrikus az $1/2$ ponra, $1 - X_n, \ldots, 1 - X_1$ szintén egyenletes eloszlásból vett minta, igy a belőle képzett rendezett minta szintén az egyenletes eloszlásból vett rendezett minta.

(c) $X_{k+1}^* - X_k^*$ valószínűségi változók azonos eloszlású (de nem független!) valószínűségi változók, $X_{k+1}^* - X_k^*$ eloszlása azonos az X_1^* valószínűségi változóeloszlásával, ami $\mathcal{B}(1,n)$ Béta eloszlású.

(iiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, az $[a,b]$ intervallumon egyenletes eloszlásból vett minta, $X_1^* < \ldots < X_n^*$ pedig a belőle gyártott rendezett minta. Adjuk meg X_k eloszlás- és sűrűségfüggvényét, valamint várható értékét!

Tipp: l. a elemeinek eloszlását.

Válasz: Eloszlásfüggvény:

$$G_{n,k}(x) = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}$$

és a sűrűségfüggvény:

$$g_{n,k}(x) = n \binom{n-1}{k-1} [F(x)]^{k-1} [1 - F(x)]^{n-k} F'(x),$$

ahol F az $[a,b]$ intervallumon egyenletes eloszlás eloszlásfüggvénye. A várható érték $\frac{a+b}{2} \cdot \frac{k}{n+1}$.

(iiiiiiiiiii) ??? Legyen $X_1, \ldots, X_n$ független minta az $F(x) = \sqrt{x} \quad (0 < x < 1)$ eloszlásfüggvénnyel. Adjuk meg X_k^* sűrűségfüggvényét!

Tipp: Lásd az elő ző feladat megoldását!

Válasz:

$$1_{[0,1]} 1/2 \cdot g_{n,k}(x) = n \binom{n-1}{k-1} [\sqrt{x}]^{k-1} [1 - \sqrt{x}]^{n-k} x^{-1/2}$$

(iiiiiiiiiii) Legyen $X_1^* < \ldots < X_n^*$ a $[0,1]$ intervallumon egyenletes eloszlásból vett rendezett minta, és $Y_1^* < \ldots < Y_n^*$ az előzőtől független, szintén a $[0,1]$ intervallumon egyenletes eloszlásból vett rendezett minta. Adjuk meg $X_k^* - Y_k^*$ sűrűségfüggvényét $(1 \leq k \leq n)$!

Tipp: Két független $\mathcal{B}(k, n-k+1)$ eloszlású valószínűségi változó különbségének sűrűsége a kérdés, ami konvolúcióval meghatározható. Figyeljünk az integrálás tartományára!

Válasz:

(iiiiiiiiiiiii) Legyen $X_1^*, \ldots, X_n^*$ a λ paraméterű exponenciális eloszlásból vett rendezett minta.

(a) Adjuk meg a k-adik $(1 \leq k \leq n)$ mintaelem eloszlás- és sűrűségfüggvényét!

(b) Milyen eloszlású a $\delta_k := X_{k+1}^* - X_k^*$, ahol $1 \leq k < n$?

Tipp:

(a) Alkalmazzuk a 12 feladatot, $F(x)$ helyébe $1 - \exp(-\lambda x)$-et írva.

(b) Alkalmazzuk az exponenciális eloszlás örökifjú tulajdonságát.

Válasz:

(a)

$$f_{n,k}(x) = n\binom{n-1}{k-1}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

ahol $F(x) = 1 - \exp(-\lambda x)$-et és $f(x) = \lambda \exp(-\lambda x)$.

(b) $\delta_k \sim \mathcal{E}xp[(n-k)\lambda]$.

(iiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, a $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ intervallumon egyenletes eloszlású minta. Legyen

$$T(\mathbf{X}) = \frac{X_1^* + X_n^*}{2}.$$

Határozzuk meg $T(\mathbf{X})g(z)$ sűrűségfüggvényét!

Tipp: Lásd elemeinek együttes sűrűségfüggvényéről tanultakat!

Ha X és Y valószínűségi változók együttes sűrűségfüggvénye $f(x,y)$, akkor a konvolúcióhoz hasonlóan a $Z = X + Y$ valószínűségi változó sűrűségfüggvénye: $g(z) = \int f(x, z-x)dx$ Figyeljünk az integrálás tartományára, és használjuk fel azt a tényt, hogy a keresett sűrűségfüggvény szimmetrikus θ-ra!

Válasz:

$$g(z) = \begin{cases} n \cdot [1 + 2(z-\theta)]^{n-1}, & \text{ha } z < \theta, \\ n/2 \cdot [1 - 2(z-\theta)]^{n-1}, & \text{ha } z > \theta \end{cases}$$

(iiiiiiiiiiiiiii) Igazoljuk, hogy ha $n > 1$, és X_1 nem elfajult és sűrűségfüggvénye valóban függ a paramétertől, akkor $T(\mathbf{X}) = X_1$ semmilyen paraméterre sem elégséges!

Tipp: Használjuk fel definícióját!

Válasz: Legyen két mintánk: X_1 és X_2. A függetlenség miatt kettejük együttes sűrűségfüggvényének feltételes sűrűségfüggvénye X_1-re nézve éppen X_2 sűrűségfüggvénye, ami természetesen függ a paramétertől.

(iiiiiiiiiiiiiii) Igazoljuk, hogy a rendezett minta minden paraméterre elégséges statisztika!

Tipp: Legyen az $X_1, \ldots, X_n$ független azonos eloszlású valószínűségi változók közös $f_\theta(x)$ sűrűségfüggvénye, ahol θ egy paraméter. Legyenek $X_1^*, \ldots, X_n^*$ a fenti valószínűségi változókból készített elemei. Mutassuk meg hogy az eredeti $f(x_1, \ldots, x_n)$ sűrűségfüggvény rekonstruálható a rendezett minta $f^*(x_1^*, \ldots, x_n^*)$ sűrűségfüggvénye alapján!

Válasz:

$f\{x_1,\dots,x_n\} = 1_{(x_{\pi(1)}\le\dots\le x_{\pi(n)})} f^*(x_{\pi(1)},\dots,x_{\pi(n)})$

ahol π az a permutáció ami szerint az aktuális minta rendezetté válik.

Emögött az a heurisztikus tény húzódik meg, hogy ha van egy független mintánk valamely F eloszlásból, azt rendezzük, majd a rendezett mintából véletlenszerűen visszatevés nélkül kiválasztjuk a mintaelemeket, akkor ismét egy független mintát kapunk ugynabból az F eloszlásból.

(iiiiiiiiiiiiiiiii) Legyenek $X_1,\dots,X_n$ független, a $[0,\theta]$ intervallumon egyenletes eloszlásból vett minta! Igaz-e, hogy X_n^* a θ paraméterre elégséges statisztika?

Tipp: l. abszolút folytonos eloszlások és alkalmazzuk a Neyman-Fisher faktorizációt.

Válasz: Igen.

(iiiiiiiiiiiiiiiii) Tegyük fel, hogy T statisztika torzítatlan becslése θ paraméternek. Tekintsünk egy tetszőleges S statisztikát. Igaz-e, hogy $\mathrm{E}(T|S)$ is torzítatlan becslése θ-nak?

Tipp: Alkalmazzuk tulajdonságait,

Válasz: Igen, mert $E(E(T|S)) = E(T)$.

(iiiiiiiiiiiiiiiiii) Legyen X valószínűségi változó, amelynek létezik a szórása.

(a) Tegyük fel, hogy ismert az $\mathrm{E}(X) = \theta$ várható érték. Igazoljuk, hogy $S_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \theta)$ torzítatlan becslése a szórásnégyzetnek! Mit mondhatunk a konzisztenciáról?

(b) Az (a) pont segítségével igazoljuk, hogy az $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ empirikus szórásnégyzet nem torzítatlan becslése a szórásnégyzetnek! Készítsünk segítségével torzítatlan becslést!

Tipp:

(a) Közvetlen számolás. Alkalmazzuk a nagy számok törvényét (keressük meg a ben).

(b) Közvetlen számolás.

Válasz:

(a) Erősen konzisztens.

(b) Az $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ torzítatlan becslés.

(iiiiiiiiiiiiiiiiiii) Tekintsünk az alábbi eloszlásokból egy n elemű mintát! Adjunk elégséges statisztikát az ismeretlen paraméterre!

(a) p paraméterű geometriai eloszlás,

(b) $(5,p)$ paraméterű $\mathcal{B}_5(p)$ binomiális eloszlás,

(c) $(3,p)$ paraméterű negatív binomiális eloszlás,

(d) $\mathcal{G}(2,\lambda)$,

(e) $\mathcal{G}(\alpha,2)$,

(f) $\theta = (\alpha,\lambda)$ paraméterű Gamma eloszlás,

(g) $\mathcal{N}(\mu,1)$,

(h) $\mathcal{N}(0,\sigma^2)$,

(i) $\mathcal{N}(\mu, \sigma^2)$,

(j) m szabadságfokú χ^2 eloszlás,

(k) $\theta = (a, b)$ paraméterű Béta eloszlás,

(l) $[-\alpha, \alpha]$ intervallumon egyenletes eloszlás.

Tipp: l. nevezetes eloszlások, továbbá használjuk a Neyman-Fisher faktorizációt (l.).

Válasz:

(a) Pl. $X_1 + \ldots + X_n$,

(b) pl. $X_1 + \ldots + X_n$,

(c) pl. $X_1 + \ldots + X_n$,

(d) pl. $X_1 + \ldots + X_n$,

(e) pl. $X_1 \cdot \ldots \cdot X_n$,

(f) pl. $X_1 + \ldots + X_n, \quad X_1 \cdot \ldots \cdot X_n$,

(g) pl. $X_1 + \ldots + X_n$,

(h) pl. $X_1^2 + \ldots + X_n^2$,

(i) pl. $X_1 + \ldots + X_n, \quad X_1^2 + \ldots + X_n^2$,

(j) pl. $X_1 + \ldots + X_n, \quad X_1^2 + \ldots + X_n^2$,

(k) pl. $\prod_{i=1}^n X_i, \prod_{j=1}^n (1 - X_j)$,

(l) pl. $\max\{-X_1^*, X_n^*\}$.

(iiiiiiiiiiiiiiiiiiii) $X_1, \ldots, X_n$ független, $\theta = (r, p)$ paraméterű negatív binomiális eloszlásból vett minta. A θ paraméterre elégséges statisztika-e a mintaátlag?

Tipp: l. diszkrét eloszlások és Neyman-Fisher faktorizáció (l.).

Válasz: Nem, itt két paraméterre kell elégséges statisztikát adni!

(iiiiiiiiiiiiiiiiiiiii) Elégséges statisztika-e θ paraméterre $L_\theta(\mathbf{X})$ (ahol L_θ a likelihood-függvény)?

Tipp: Elemi logika.

Válasz: Nyilván nem, hiszen benne van a paraméter.

(iiiiiiiiiiiiiiiiiiiii) ??? Legyenek $X_1, \ldots, X_n$ független, λ paraméterű Poisson eloszlású valószínűségi változók.

(a) Igaz-e, hogy $\overline{X}$ statisztika a λ paraméterre!

(b) Adjunk a λ paraméterre a fentitől különböző elégséges statisztikát!

Tipp:

(a) l. diszkrét eloszlások és Neyman-Fisher faktorizáció

(b) L. statisztika tulajdonságait.

Válasz:

(a) Igaz.

(b) Pl. a teljes minta, a rendezett minta, a mintaösszeg és annak invertálható függvényei (utóbbiak a minimális megoldások).

(iiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \lambda$ paraméterű exponenciális eloszlásból vett független minta.

(a) Igaz-e, hogy $\sum_{i=1}^{n} X_i$ elégséges statisztika a λ paraméterre?

(b) Adjunk a λ paraméterre más elégséges statisztikákat!

Tipp:

(a) Írjuk fel a likelihood függvényt azaz az $X_1, \ldots, X_n$ együttes sűrűségfüggvényét (l. abszolút folytonos eloszlások)

(b) L. előző feladat.

Válasz:

(a) Igaz.

(b) Pl. a teljes minta, a rendezett minta, a mintaátlag, a mintaösszeg invertálható függvényei (utóbbiak a minimális megoldások).

(iiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, p paraméterű geometriai eloszlású minta.

(a) Adjuk meg a p paraméter Y maximum likelihood becslését!

(b) Alkalmasan transzformálva tegyük Y-t torzítatlan becsléssé!

Tipp:

(a) Közvetlen számolás.

(b) Keressük meg a ben a negatív binomiális eloszlást, és okoskodjunk az $\mathrm{E}(1/\overline{X})$ kiszámításához hasonló módon, ugyanis a negatív binomiális eloszlás éppolyan általánosítása a geometriai eloszlásnak, mint a gamma eloszlás az exponenciális eloszlásnak.

Válasz:

(a) $\frac{n}{Z}$, ahol $Z = X_1 + \ldots + X_n$

(b) $\frac{n-1}{Y}$.

Vegyük észre, hogy ez a képlet $n = 1$-re nincs értelmezve!

(iiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, a $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ intervallumon egyenletes eloszlású minta.

(a) $\overline{X}$ torzítatlan becslése-e θ-nak? Ha nem, készítsünk segítségével torzítatlan becslést!

(b) $X_n^* - \frac{1}{2}$ torzítatlan becslése-e θ-nak? Ha nem, készítsünk segítségével torzítatlan becslést!

(c) Igazoljuk, hogy $\overline{X}$ erősen és $X_n^* - \frac{1}{2}$ gyengén konzisztens becslései θ-nak!

Tipp:

(a) A mintaátlag torzítatlan becslése a várható értéknek.

(b) Számítsuk ki az $X_n^* - \frac{1}{2}$ valószínűségi változó várható értékjét (l. a -ról szóló paragrafust).

$E(Y_1-\theta)^2$ Y_1 $E(\overline{Y_2}-\theta)^2$z $Y_2=X_n^*-\frac{1}{2}$ becslések gyenge konzisztenciájának igazolásáhY_1 számítsuk ki és négyzetes rizikókat és alkalmaY_2uk Csebisev-egyenlőtlenséget. Az becslés a naY_1 számok erős törvénye miatt erősen konzisztens, míg az négyzetes rizikója kisebb nagyságrendű, mint az becslésé. (A szükséges információkat keressük meg a ben és a -ról szóló paragrafusban).

Válasz:

(a) Igen.

(b) Nem, de az $Y_2+1/(n+1)$ már torzítatlan.

(c) Az $\overline{X}$ erős konzisztenciája az Útmutatás alapján nyilvánvaló, míg az $X_n^*-\frac{1}{2}$ gyenge konzisztenciája nyilvánvaló az Útmutató alapján (az erős konzisztencia is igaz, de az (egyszerű) bizonyítás eszköze nem szerepel a Tananyagban).

(iiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1,\dots,X_n$ független, a $[0,\theta]$ intervallumon egyenletes eloszlású minta.

(a) Adjunk maximum likelihood becslést θ-ra!

(b) Igazoljuk, hogy $2\overline{X}$ torzítatlan becslés θ-ra!

(c) Mivel a $\theta/2$-re szimmetrikus az eloszlásunk, a medián egybeesik a várható értékkel. Tegyük fel, hogy n páratlan, és készítsünk a tapasztalati medián segítségébel torzítatlan becslést θ-ra!

(d) X_1 torzítatlan becslése-e θ-nak? Ha nem, készítsünk segítségével torzítatlan becslést!

(e) X_1^* torzítatlan becslése-e θ-nak? Ha nem, készítsünk segítségével torzítatlan becslést!

(f) X_n^* torzítatlan becslése-e θ-nak? Ha nem, készítsünk segítségével torzítatlan becslést!

(g) A fenti becslések közül melyik konzisztens?

(h) Számítsuk ki és hasonlítsuk össze a fenti torzítatlan becslések szórásnégyzetét! Melyik a leghatásosabb?

(i) Teljesül-e az $I_n(\theta)=nI_1(\theta)$ összefüggés? Teljesül-e minden esetben a Cramér-Rao egyenlőtlenség?

(j) Igazoljuk, hogy X_n^* elégséges statisztika θ -ra. Segítségével blackwellizáljuk a fenti torzítatlan becsléseket!

Tipp:

(a) Vigyázzunk, a linelihood-függvény nem mindenütt deriválható!

(b) A mintaátlag mindig torzítatlan becslése a várható értéknek, ami itt $\theta/2$.

(c) Legyen $n=2k+1$, mivel két egymást követő rendezett minta különbségének várható értéke $\frac{\theta}{2k+2}$.

(d) Nyilvánvaló.

(e) $\mathrm{E}(X_1^*)=\theta/(n+1)$.

(f) $\mathrm{E}(X_n^*)\theta n/(n+1)$

(g) Vizsgáljuk meg a szórásnégyzetüket!

(h) $\theta=1$ esetén ismert mindegyik, használjuk ki!

(i) A $2\overline{X}$ szórásnégyzete $\frac{\theta^2}{3n}$, $I_1(\theta)=\frac{1}{\theta^2}$.

(j) A rendezett mintákon alapuló becslésekre alkalmazzuk a következő heurisztikát: $E(X_k^*|x_n^*)=\frac{k}{n+1}|x_n^*$. Ami a $2\overline{X}$-ot illeti, hasonló heurisztika alapján: tetszőleges n-re $E(X_n|X_n^*)=\frac{n-1}{2n}X_n^*+frac1nX_n^*$.

Válasz:

(a) X_n^*

(b) $2\overline{X}$

(c) a tapasztalati medián kétszerese (jelölje ezt $\hat{\theta}_{0,5}$) θ torzítatlan becslése.

(d) $\hat{\theta}_1 = 2X_1$.

(e) $\hat{\theta}_2 = X_1^*(n+1)$.

(f) $\hat{\theta}_3 = X_n^*(n+1)/n$.

(g) $\hat{\theta}_1$

(h) $\hat{\theta}_2$ a leghatásosabb, de a $\hat{\theta}_{0,5}$ szórásnégyzetének is ugyanekkora a nagyságrendje $(\sim 2/n^2)$, elég nagy n-re ez is meghaladja az $nI_1(\theta) = \frac{n}{\theta^2}$ információs határt.

(i) A Cramér-Rao egyenlőtlenség n nagy értékeire csak a $2\overline{X}$ és a $\hat{\theta}_1$-re nem teljesül.

(j) Az X_n^* statisztika elégségessége következik a Neyman-Fisher szorzat-tételből, figyelembevéve, hogy a likelihood függvény alakja $L_\theta(\mathbf{x}) = \frac{1}{\theta} \cdot 1_{\{0 \le x_n^* \le \theta\}}$. Valamennyi blackwellizált: θ_2

(iiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, a $[-\theta, \theta]$ intervallumon egyenletes eloszlású minta.

(a) Adjunk θ-ra torzítatlan becslést $\overline{|X|}$ segítségével!

(b) Konzisztens-e a fenti becslés?

Tipp:

(a) Alkalmazzuk a következő heurisztikus meggondolást: az $X_1, \ldots, X_n$ független, a $[-\theta, \theta]$ intervallumon egyenletes eloszlású mintát ugy is kisorsolhatjuk, hogy a $[0, \theta]$ intervallumon kisorsolunk az $Y_1, \ldots, Y_n$ független mintát, valamint egy tőlük és egymástól is független $p = 1/2$ paraméterű $\varepsilon_1, \ldots, \varepsilon_n$ Bernoulli-mintát. Legyen $X_k(2\varepsilon - 1)Y_k$ minden k-ra. Ilymódon a feladatot visszavezettük az előző feladat (f) pontjára.

(b) Az előzőek alapján nyilvánvaló.

Válasz:

(a) $\hat{\theta} = 2\overline{|X|}$

(b) Igen.

(iiiiiiiiiiiiiiiiiiiiiiiiii) Legyenek X_1, X_2, X_3 rendre $\mathcal{N}(\mu, 1), \mathcal{N}(\mu, 4), \mathcal{N}(\mu, 1/4)$ eloszlású független mintaelemek.

(a) Milyen a, b, c értékekre lesz $aX_1 + bX_2 + cX_3$ torzítatlan becslése μ-nek?

(b) Milyen a, b, c választással kapjuk meg a leghatásosabb becslést a torzítatlanok közül?

Tipp: A becslés akkor lesz torzítatlan, ha $a + b + c = 1$. Az optimális becslést akkor kapjuk meg, ha az a, b, c súlyok fordítottan arányosak a valószínűségi változók szórásnégyzeteivel (pl. Lagrange multiplikátor modszerrel igazolható).

Válasz: $a = \frac{16}{273}$ $b = \frac{1}{273}$ $c = \frac{256}{273}$

(iiiiiiiiiiiiiiiiiiiiiiiiii) Tekintsük az $X_1, \ldots, X_n$ független, θ paraméterű Bernoulli eloszlású mintát és számítsuk ki a Fisher-információját! Tekintsük az $Y_1, \ldots, Y_n$ független mintát is, amely háttérváltozója θ

valószínűséggel $1, 1-\theta$ valószínűséggel -1 értéket vesz fel. Számítsuk ki ennek is a Fisher-információját és vessük össze az előbb meghatározott információval!

Tipp: Jelöljük $p_\theta(x)$-szel annak a valószínűségét, hogy $X = x$. Itt $x = 0, x = 1$, illetve $x = -1, x = 1$. Alkalmazzuk paragrafusban szereplő definíciót:

$$I_1(\theta) = \frac{\left(\frac{\partial}{\partial\theta}p_\theta(0)\right)^2}{p_\theta(0)} + \frac{\left(\frac{\partial}{\partial\theta}p_\theta(1)\right)^2}{p_\theta(1)},$$

illetve

$$I_1(\theta) = \frac{\left(\frac{\partial}{\partial\theta}p_\theta(-1)\right)^2}{p_\theta(-1)} + \frac{\left(\frac{\partial}{\partial\theta}p_\theta(1)\right)^2}{p_\theta(1)},$$

Válasz: Mindkét esetben $I_n(\theta) = \frac{n}{\theta(1-\theta)}$

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, p paraméterű Bernoulli eloszlású minta.

(a) Adjunk maximum likelihood becslést p-re!

(b) Számítsuk ki $\mathrm{D}^2_p(\overline{X})$-ot is! Mit mondhatunk a Cramér- Rao-egyenlőtlenség alapján?

(c) Szeretnénk p-re torzítatlan becslést adni. Mekkora legyen n, ha azt szeretnénk, hogy becslésünk szórása ne haladja meg 0,03-at p bármely értéke esetén sem?

Tipp:

(a) Az M-L becslés definícióját lásd a paragrafusban

(b) Közvetlen számolás, az informaciós határt illetően lásd az előző feladatot!

(c) Legyen ez a becslés a $(\hat{p} = \overline{X})$. Az előző pontban már kiszámítottuk $\mathrm{D}^2_p(\overline{X})$-ot Keressük meg a $\max_{0\le p\le 1} p(1-)p$ -t

Válasz:

(a) $\hat{p} = (\overline{X})$.

(b) $\mathrm{D}^2_p(\overline{X}) = \frac{p(1-p)}{n}$. A becslés hatásos, a Cramér-Rao egyenlőtlenségben itt egyenlőség all.

(c) A $\mathrm{D}^2_p(\overline{X})$ maximuma $\frac{1}{4n}$ Ennek alapján $n = \left(\frac{1}{0{,}06}\right)^2$.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, λ paraméterű exponenciális eloszlású minta.

(a) Adjunk maximum likelihood becslést λ-ra, majd ennek alapján szerkeszzünk torzítatlan becslést λ-ra.

(b) Számoljuk ki a minta Fisher-információját!

(c) $1/\overline{X}$ nem torzítatlan becslése a λ paraméternek. Készítsünk segítségével $\hat{\eta}$ torzítatlan becslést és számoljuk ki $\hat{\eta}$ szórásnégyzetét!

(d) Az $\overline{X}$ elégséges statisztika segítségével blackwellizáljuk a fenti torzítatlan becslést! (Ismert, hogy az így kapott becslés hatásos becslése λ-nak. Ellentmond-e ez a Cramér- Rao egyenlőtlenségnek?)

Tipp:

(a) Alkalmazzuk a definíciót (l. és).

(b) Alkalmazzuk a megfelelő formuláját.

(c) $1/\overline{X}$ nem torzítatlan becslése a λ paraméternek.

(d) A számoláshoz használjuk a Gamma eloszlást (l.), ennek alapján $\hat{\eta}$ az $1/\overline{X}$ statisztika alkalmas konstanszorosa lesz.

(e) Az $\overline{X}$ Lásd az előbbi észrevételt.

Válasz:

(a) $1/\overline{X}, (n-1)/(n\overline{X})$.

(b) $I_n(\lambda) = \frac{n}{\lambda^2}$

(c) $\hat{\eta} = \frac{n-1}{n\overline{X}}$, $D^2(\hat{\eta}) = \frac{\lambda^2 n^2}{(n-1)^2(n-2)}$

(d) Az $\hat{\eta}$ becslés blackwellizáltja önmaga.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, $(2, \lambda)$ paraméterű Gamma eloszlású minta.

(a) Adjunk maximum likelihood becslést λ-ra!

(b) Adjunk becslést λ-ra a momentumok módszerével!

(c) Torzítatlan becslése-e X_1 statisztika a $1/\lambda$-nak? Ha nem, készítsünk segítségével torzítatlan becslést!

(d) Torzítatlan becslése-e $1/X_1$ statisztika a λ paraméternek? Ha nem, készítsünk segítségével torzítatlan becslést!

(e) Torzítatlan becslése-e $1/\overline{X}$ statisztika a λ paraméternek? Ha nem, készítsünk segítségével torzítatlan becslést!

(f) Igazoljuk, hogy $\sum_{i=1}^n X_i$ elégséges statisztika a λ paraméterre! Segítségével blackwellizáljuk a fenti torzítatlan becsléseket!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ független minta.

(a) Igazoljuk, hogy X_1 torzítatlan, de nem konzisztens becslése μ-nek! Mit mondhatunk a Cramér- Rao-egyenlőtlenség alapján?

(b) Számítsuk ki a minta Fisher-információját! Számítsuk ki $\mathrm{D}^2_\mu(\overline{X})$-ot is! Igazoljuk, hogy $\overline{X}$ hatásos becslése μ-nek!

(c) Torzítatlan becslése-e μ^2 -nek $X_1 X_2$? Mennyi a szórásnégyzete? Mondhatunk-e valamit a Cramér- Rao-egyenlőtlenség alapján?

(d) Torzítatlan becslése-e μ^2-nek $\overline{X}^2$? Ha nem, tegyük azzá, és számítsuk ki a szórásnégyzetét!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \sim \mathcal{N}(0, \vartheta)$ ($\vartheta = \sigma^2$) független minta.

(a) Adjuk maximum likelihood becslést ϑ-ra!

(b) Igazoljuk, hogy $S_1^2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ hatásos becslése σ^2-nek!

(c) Igazoljuk, hogy a korrigált empirikus szórásnégyzet nem hatásos becslése a σ^2 paraméternek!

Tipp:

(a) Alkalmazzuk a definíciót (l.)

(b) Számítsuk ki a minta ϑ-ra vonatkozó Fisher-információját (l.). és a $\hat{\vartheta}$ M-L becslés szórásnégyzetét

(c) Közvetlen számolás.

Válasz:

(a) $S_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

(b) $I_n(\vartheta) = \frac{1}{2\vartheta^2}$, $D^2(\hat{\vartheta}) = 2\vartheta^2$.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, λ paraméterű Poisson eloszlású minta.

(a) Vegyük λ maximum likelihood becslését! Minden realizáció mellett létezik-e maximum likelihood becslés?

(b) Igazoljuk, hogy a maximum likelihood módszerrel kapott becslés torzítatlan és számítsuk ki a szórásnégyzetét! Mit mondhatunk a Cramér- Rao-egyenlőtlenség alapján?

(c) Igazoljuk, hogy X_1 is torzítatlan becslése λ -nak! Az $\overline{X}$ elégséges statisztika segítségével blackwellizáljuk az X_1 becslést!

(d) Torzítatlan becslése-e λ-nak az empirikus szórásnégyzet? Ha nem, tegyük azzá! Hatásos becslést kapunk-e így?

(e) A fenti becslések közül melyik konzisztens?

Tipp:

(a) Közvetlen számolás.

(b) Közvetlen számolás; számítsuk ki a minta $I_n(\lambda)$ Fisher-információját.

(c) Közvetlen számolás. Alkalmazzuk tulajdonságait, és vegyük észre, hogy az $X_1, \ldots, X_n$ mintaelemek szerepe szimmetrikus!

(d) Vegyük észre, hogy empirikus szórásnégyzet mindig torzítatlan becslése a szórásnégyzetnek. Alkalmazzuk paragrafusban a szóránégyzet becslésére megfogalmazott állítást!

(e) Alkalmazzuk az előző részfeladatok eredményeit!

Válasz:

(a) Igen.

(b) Az információs határ eléretik, tehát a M-L becslés hatásos.

(c) A mintaátlag (azaz a M-L becslés) lesz a blackwellizált.

(d) Igen. A becslés nem lesz hatásos, bár ennek ellenőrzése az Útmutatás alapján hosszadalmas, a cáfolathoz elegendő λ egyetlen értékére elvégezni a számolást.

(e) (c) kivételével mindegyik.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \sim Bin(5, p)$.

(a) Vizsgáljuk meg a maximum likelihood és a momentumok módszerével kapott becslések torzítatlanságát és hatásosságát!

(b) Számítsuk ki a minta Fisher-információját!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Adjunk becslést a negatív binomiális eloszlás paramétereire momentumok módszerével!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Tekintsük az

$$f_{a,p}(x) = \begin{cases} \dfrac{pa^p}{x^{p+1}}, & \text{ha } x \geq a, \\ 0 & \text{különben} \end{cases}$$

sűrűségfüggvényű Pareto-eloszlást, ahol $a, p > 0$ paraméterek. Adjunk maximum likelihood becslést $\theta = (a, p)$-re! Tegyük fel, hogy $p > 2$. Adjunk becslést θ-ra a momentumok módszerével!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Tekintsünk egy kételemű független, $(\mu, 1)$ paraméterű Cauchy eloszlású mintát! A (μ, σ) paraméterű Cauchy eloszlás sűrűségfüggvénye:

$$f_{\mu,\sigma}(x) = \frac{\sigma}{\pi(\sigma^2 + (x-\mu)^2)}.$$

(a) Adjunk maximum likelihood becslést μ-re az x_1, x_2 realizáció segítségével!

(b) Tudunk-e becslést adni momentumok módszerével? Használjuk ki, hogy 1-nél kisebb momentumok is léteznek!

Tipp: Tegyük fel, hogy a két mintalem x_1 és x_2, ekkor a likelihood függvény konstans szorzótól eltekintve

$$L_\theta(x_1, x_2) = \frac{1}{(1 + (x_1 - \theta)^2)\,(1 + (x_2 - \theta)^2)}$$

A $\frac{\partial L_\theta(x_1,x_2)}{\partial \theta} = 0$ egyenlet θ-ban harmadfokú, de szimmetria meggondolásokból következik, hogy a $\theta = \frac{x_1+x_2}{2}$ megoldás. Osszuk el az egyenletet $\theta - \frac{x_1+x_2}{2}$-vel, így már egy másodfokú egyenletet kapunk.

Válasz: A 3 megoldás:

$$\theta_{1,2,3} = \begin{cases} \frac{x_1+x_2}{2} \\ \frac{x_1+x_2+\sqrt{(x_1-x_2)^2-4}}{2} \\ \frac{x_1+x_2-\sqrt{(x_1-x_2)^2-4}}{2} \end{cases}$$

Látható, hogy θ_2 és $\theta_3 |x_1 - x_2| \geq 4$ estén valósak. Tovább analizálva a megoldásokat kiderül az a meglepő tény, hogy a ezek valós megoldások a maximumhelyei a likelihood függvénynek, így éppen a távoli mintaelemek esetén nem egyértelmű a megoldás, és a megoldások egymástól távoliak; $|x_1 - x_2| > 2$ esetén a mintaátlag lokális minimum!

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, $[a, b]$ intervallumon egyenletes eloszlású minta.

(a) Adjunk becslést (a, b)-re a momentumok módszerével!

(b) Adjunk maximum likelihood becslést (a, b)-re!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független minta. Tudunk-e adni $1-\varepsilon$ megbízhatósági szintű konfidencia intervallumot σ-ra

(a) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$,

(b) $\frac{nS_n^2}{\sigma^2}$ ($S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$) segítségével?

Tipp:

(a) Vizsgáljuk meg milyen statisztika alapján kellene konfidencia intervallumot adni!

(b) Vizsgáljuk meg milyen statisztika alapján kellene konfidencia intervallumot adni!

Válasz:

(a) Nem, mert a $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ statisztika standard normális eloszlású, ebből egyik paraméterre sem vonhatunk le következtetést.

(b) Nem, mert a $\frac{nS_n^2}{\sigma^2}$ statisztika $\chi^2(n)$ eloszlású, ebből egyik paraméterre sem vonhatunk le következtetést.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Egy cukorgyárban kockacukrokat gyártanak. Tegyük fel, hogy a cukrok élhossza közelítőleg normális eloszlású. Megmérjük 16 cukor élhosszúságát. Az adatok átlaga 10,06 mm, tapasztalati szórása 0,46 mm. Adjunk 95% megbízhatósági szintű konfidencia intervallumot μ^3-re (azaz egy „átlagos" kockacukor térfogatára)!

Tipp: Alkalmazzuk a paragrafus példáját standard normális eloszlás helyett a $t(15)$ Student eloszlással a kocka élhosszára, majd használjuk fel azt a tényt, hogy az x^3 függvény monoton.

Válasz: Táblázatból ismert, hogy ha $X \sim t(15)$, akkor $\mathbb{P}(X > 2,12) = 0,975$ így a kocka élére a $10,06 \pm 2,12 \cdot 0,46/4$ intevallum 95 megbízhatósági szintű konfidencia intervallum. A térfogatra a $[945,87mm^3, 1093,94mm^3]$ nem szimmetrikus konfidencia intervallumot kapjuk.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyenek $X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ és $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$ független minták. Adjunk $1-\varepsilon$ szintű konfidencia intervallumot $\mu_1 - \mu_2$-re $\overline{X} - \overline{Y}$ segítségével ((n, m, σ) ismert!)

Tipp: várható éetékű valószínűségi változó határozzuk meg σ_e^2 szórásnegyzetét, majd alkalmazzuk paragrafusban kidolgozott példát $\mu = \mu_1 - \mu_2$-re .

Válasz: $\sigma_e^2 = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$

A konfidencia intervallum:

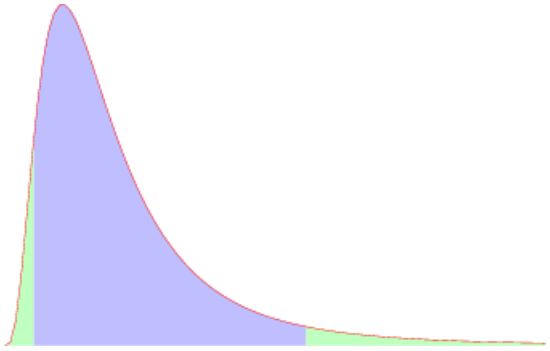
$$\overline{X} - \overline{Y} \pm \frac{\sigma_e \cdot \Phi^{-1}(1-\varepsilon/2)}{\sqrt{n}}$$

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyenek $X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ és $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$ független minták. Adjunk $1-\varepsilon$ szintű konfidencia intervallumot σ_1/σ_2-re!

Tipp: Tekintsük az

$$\eta = \frac{\frac{\sum_{j=1}^n (X_j-\mu_1)^2}{n}}{\frac{\sum_{j=1}^m (Y_j-\mu_2)^2}{n}}$$

statisztikát, vegyük észre, hogy $\frac{\sigma_2^2}{\sigma_1^2}\eta \sim F(n,m)$. Jelöljön ξ egy $F(n,m)$ eloszlású valószínűségi változót; keressük meg azt az F_1 (F_2) értéket amelyre a $P(\xi < F_1) = \varepsilon/2$ ($P(\xi > F_2) = \varepsilon/2$)



Válasz: A $P(\,)$ argumentumát alkalmas átrendezése a

$$P(\eta/F_2 < \frac{\sigma_1^2}{\sigma_2^2}) = \varepsilon/2 \quad \text{és} \quad \frac{\sigma_1^2}{\sigma_2^2} < \eta/F_1 = 1-\varepsilon/2$$

egyenlőségre vezet.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, a $[0,\theta]$ intervallumon egyenletes eloszlásból vett minta. Adjunk $1-\varepsilon$ megbízhatósági szintű konfidencia intervallumot θ-ra

(a) $X_1 + X_2$,

(b) X_n^* segítségével!

Tipp:

(a) Nyilvánvaló, hogy a minta töredékével $(X_1 + X_2)$ túlságosan tág konfidencia intervallumot kapunk.

(b) Alsó határnak az maga az X_n^* megfelel, hiszen θ nem lehet ennél kisebb. A θ_f felső határ meghatározásához vegyünk egy $0 < \delta < \theta$ számot és vizsgáljuk a $P(\delta < X_n^* < \theta) = P(\theta < X_n^* + \delta) = 1-\varepsilon$ valószínűséget. A jobb oldal valószínűsége $1 - \left(\frac{\theta-\delta}{\theta}\right)^n$, ami egyenlő $1-\varepsilon$-nal. Ebből δ-ra kapunk egy egyenletet. Oldjuk meg és rendezzük át a középső valószínűség argumentumát.

Válasz:

(a) Az $X_1 + X_2$ eset irreleváns.

(b) A javasolt számitásokat eredménye: $\theta_f = X_n^*/\varepsilon^{1/n}$.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, λ paraméterű Poisson eloszlású minta. Adjunk λ-ra $1-\varepsilon$ megbízhatósági szintű konfidencia intervallumot a centrális határeloszlás-tétellel!

Tipp: Lásd a pargrafusban az $\mathcal{N}(\mu, \sigma_0^2)$ re kidolgozott példát. Itt $\sigma_0^2 = \lambda$, ezért, ha $\overline{X} \pm r_\varepsilon$ alakban keressük a konfidencia intervallumot.

Válasz: $r_\varepsilon = \overline{X} - y$ lesz, ahol y az $(1-y)^2 - \frac{\Phi^{-1}(1-\varepsilon/2)y}{\sqrt{n}}$ másodfokú egyenletnek az a megoldása amelyre r_ε $1/\sqrt{n}$ nagyságrendű.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, λ paraméterű exponenciális eloszlású minta. Adjunk λ-ra $1-\varepsilon$ megbízhatósági szintű konfidencia intervallumot

(a) $\sum_{i=1}^n X_i$ segítségével!

(b) a Csebisev-egyenlőtlenség felhasználásával!

(c) a centrális határeloszlás-tétellel!

Tipp: Mindhárom $1/x$-etben a $\mu = \frac{1}{\lambda}$ paraméterre (várhatóan $1/b < \lambda < 1/a$) esszünk egy $a < \mu < b$ intervallumot, majd felhasznalva az függvény monoton voltát az intervallum lesz a keresett konfidencia intervallum.

(a) $\lambda \sum_{i=1}^{n} X_i \sim \Gamma(n, 1)$.

(b) A Csebisev-egyenlőtlenséget az $\bar{X} - \mu$ valószínűségi változóra írjuk fel:

$$\mathbb{P}\left(|\bar{X} - \mu| > a\mathbb{D}(\bar{X})\right) < \frac{1}{a^2},$$

ahol $\mathbb{D}(\bar{X}) = \frac{\mu}{\sqrt{n}}$, $a = 1/\sqrt{\varepsilon}$.

(c) A centrális határeloszlástétel alkalmazásánál hasonló módon járunk el: keressük azt az a értéket, amelyre

$$\mathbb{P}\left(\left|\bar{X} - \mu\right| > a\right) < \varepsilon.$$

Legyen a_ε a standard normális eloszlás $1 - \varepsilon/2$ kvantilise. Mivel elég nagy mintaelemszámra a $\frac{\sqrt{n}(\bar{X}-\mu)}{\mathbf{D}(X)}$ véletlen változó jó közelítéssel standard normális eloszlású és $\mathbf{D}(X) = \mu$.

Válasz:

(a)

(b)

$$\mathbb{P}\left(\bar{X} - \frac{\mu\sqrt{\varepsilon}}{\sqrt{n}} < \mu < \bar{X} + \frac{\mu\sqrt{\varepsilon}}{\sqrt{n}}\right) < \varepsilon, \qquad (1)$$

???

λ-ra szerkesztett intervallumhoz ennek az intervallum végpontjainak reciprokait vesszük. (Tanulságos kiszámolni a keresett $[a, b]$ intervallumot n és ε konkrét értékeire; legyen például $n = 100$ és $\varepsilon = 0,05$. Ekkor $a = 4,47$, $\mathbb{D}(\bar{X}) = 0,1\bar{X}$ és a (2.11) egyenlőtlenség-ben a jobb oldalon álló esemény a következő alakot ölti:

$$\bar{X} - 0,447\mu < \mu < \bar{X} + 0,447\mu.$$

A szükséges műveleteket elvégezve, μ-re a következő konfidencia intervallumot kapjuk:

$$0,68\bar{X} < \mu < 1,82\bar{X}.$$

)

(c)

$$\mathbb{P}\left(\left|\frac{\sqrt{n}(\bar{X} - \mu)}{\mu}\right| > a_\varepsilon\right) \mathbb{P}\left(\frac{\sqrt{n} - a_\varepsilon}{\sqrt{n}\bar{X}} < \frac{1}{\mu} < \frac{\sqrt{n} + a_\varepsilon}{\sqrt{n}\bar{X}}\right) \approx \varepsilon.$$

(Az (a) ponthoz hasonlóan legyen $n = 100$ és $\varepsilon = 0,05$. Ekkor $a_\varepsilon = 1,96$. Az (a) pont gondolatmenetével analóg módon adódik:

$$\bar{X} - 0,196\mu < \mu < \bar{X} + 0,196\mu.$$

A μ paraméterre a következő konfidencia intervallumot kapjuk:

$$0,836\bar{X} < \mu < 1,244\bar{X}.$$

Vegyük észre, hogy ugyanarra a mintaelemszámra és megbízhatósági szintre a Csebisev-egyenlőtlenség felhasználásával $1,14\bar{X}$ méretű, a centrális határeloszlástétel alkalmazásával $0,408\bar{X}$ méretű, konfidencia intervallumot szerkesztettünk. Természetesen az utóbbi csak közelítőleg pontos eredmény. A hiba becslése lehetséges, de meghaladja a Tananyag kereteit.)

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiii$K_n$iiiiiiiiiii) Végezzü $1-\varepsilon$ el n -szer egy kísérletet, legyen az A $p=\mathrm{P}(A)$ bek$n=10$zése$n=10000$a . Szerkesszünk rá megbízhatósági szintű konfidencia intervallumot -ra és esetén is!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1,\dots,X_n$ független, a $(\theta-1/2,\ \theta+1/2)$ intervallumon egyenletes eloszlású minta. Adjunk $1-\varepsilon$ megbízhatósági szintű konfidencia intervallumot θ -ra $T(\mathbf{X})=(X_1^*+X_n^*)/2$ segítségével!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen X egy egyelemű minta, sűrűségfüggvénye $e^{\theta-x}$, ha $x>\theta$. Szerkesszünk $1-\varepsilon$ megbízhatósági szintű konfidencia intervallumot a θ paraméterre X segítségével!

Tipp: Vegyük észre, hogy $e^{\theta-x}$ éppen annak a valószínűsége, hogy az X mintaelem nagyobb, mint x. Feltesszük, hogy olyan $[\theta_a,\theta_f]$ konfidenciaintervallumot kívánunk szerkeszteni, amely az alábbi értelemben szimmetrikus:

$$\mathbb{P}_{\theta_a}(X>\theta_a)=\varepsilon/2 \text{ és } \mathbb{P}_{\theta_f}(X<\theta_f)=\varepsilon/2.$$

Ehhez oldjuk meg az

$$\exp(\theta_a-x)=\varepsilon/2 \text{ és } \exp(\theta_f-x)=1-\varepsilon/2$$

egyenleteket.

Válasz: $\theta_a=\log\varepsilon-\log 2,\quad \theta_f=\log(1-\varepsilon/2)\approx 1-\varepsilon/2$.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1,\dots,X_n$ független, λ paraméterű exponenciális eloszlású minta.

(a) Konstruáljunk ε terjedelmű próbát $H_0:\lambda=\lambda_0$ versus $H_1:\lambda=\lambda_1<\lambda_0$ hipotézisek ellenőrzésére X_1^* segítségével!

(b) Konstruáljunk ε terjedelmű próbát ugyanezen hipozézisekre $\sum X_j$ alapján !

(c) A fenti próbák közül melyik konzisztens?

Tipp:

(a) Vegyük észre, hogy X_1^* eloszlásfüggvénye $1-e^{-n\lambda x}$ A kritikus tartomány $X_1^*>c(n)$ alakú. A konzisztencia ellenőrzéséhez a $\mathbb{P}_{\lambda_1}(X_1^*<c(n))$ valószínűség $(n\to\infty)$ határértékét kell megvizsgálni.

(b) A $\sum j=1^n X_j \mathcal{G}(n,\lambda)$ eloszlású, ennek alapján is szerkeszthetünk próbát de n nagy értékeire egyszerűbb a centrális határeloszlestételt alkalmazni $Y_n=\overline{X}$ -re, figyelembevéve, hogy $\mathbb{E}(Y_n)=\frac{1}{\lambda}$ és $\mathbb{D}(Y_n)=\frac{1}{\lambda\sqrt{n}}$ A kritikus tartományt most is $Y_n>c(n)$ alakban keressük. A konzisztencia ellenőrzése is hasonló módon történik.

(c) Az (a) és (b) pontokban megadtuk a módszert.

Válasz:

(a) A $\mathbb{P}_{\lambda_0}(X_1^*<c(n))=e^{-n\lambda_0 c(n)}=\varepsilon$ egyenletből adódik a $c(n)=\frac{-\log\varepsilon}{n\lambda_0}$. A konzisztencia ellenőrzéséhez a $\mathbb{P}_1=\mathbb{P}_{\lambda_1}(X_1^*>c(n))=1-(-n\lambda_1 c(n))$ valószínűséget kell meghatározni; helyettesítsük be $c(n)$ értékét $\mathbb{P}_1$ definíciójába:

$$\mathbb{P}_1=1-\exp\left(\log\varepsilon\frac{\lambda_1}{\lambda_0}\right).$$

Tehát $\mathbb{P}_1$ nem függ n-től.

(b) A centrális határeloszlás tétel szerint elég nagy n-re ($n > 20$) $(Y_n - \frac{1}{\lambda_0})\lambda_0\sqrt{n}$ standard normális eloszlású. Tehát $c(n) = \frac{1}{\lambda_0} + \frac{1}{\sqrt{n}\lambda_0}\Phi^{-1}(1-\varepsilon)$, ahol $\Phi(\cdot)$ a standard normális eloszlásfüggvény.

(c) Az (a) esetben már igazoltuk, hogy a másodfajú hiba független n-től, tehát a próba nem konzisztens. A (b) esetben a centrális határeloszlás tételt a $\lambda_1 < \lambda_0$ paraméterrel alklamazzuk. Ekkor az $(Y_n - \frac{1}{\lambda_1})\lambda_1\sqrt{n}$ valószínűségi változó lesz standard normális eloszlású. Mivel $1/\lambda_1 > 1/\lambda_0$ és $1/\sqrt{n}$ tart nullához belátható, hogy ha $n \to \infty$ akkor $\mathbb{P}_{\lambda_1}(Y_n < c(n))$ nullahoz tart, tehát a próba konzisztens.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Valódi (θ) selejtarányra szeretnénk minőségellenőrzést. Vegyünk egy $n = 25$ elemű független Bernoulli-mintát: $X_1, \ldots, X_n$. Konstruáljunk $\varepsilon = 0,05$ terjedelmű (randomizált) próbát a $H_0 : \theta = \theta_0 = 0,05$ és $H_1 : \theta = \theta_1 = 0,1$ választáshoz! Határozzuk meg a másodfajú hibát is. A $\mathcal{B}(25, 0, 05)$ (F_0) és a $\mathcal{B}(25, 0, 1)$ (F_1) binomiális eloszlásokról az alábbi adatok ismertek:

$F_0(2) = 0,873$
$F_0(3) = 0,9,66$
$P_0(3) = 0,093$
$F_1(2) = 0,873$
$F_1(3) = 0,9,66$
$P_1(3) = 0,093,$

ahol $P_0(3)$ ($P_1(3)$) annak a valószínűsége, hogy egy $\mathcal{B}(25, 0, 05)$ ($\mathcal{B}(25, 0, 1)$) eloszlású valószínűségi változó pontosan a 3 értéket veszi fel.

Tipp: Alkalmazzuk a -lemmát. Vegyük észre, hogy az így konstrált próba kritikus tartománya $x > c$ alakú, ahol x a mintában levő selejtes termékek x száma. Látható, hogy olyan kritikus tartomány nincs, amely pontosan 0,05 terjedelemű próbát adna, ($F_0(2) < 0,95$, $F_0(3) > 0,95$), ezért randomizálnunk kell. Keressük meg azt a $\delta > 0$ számot, amelyre $F_0(2) + \delta P_0(3) = 0,95$.

Döntésünk: ha $x > 3$ elvetjük a null-hipotézist, ha $x = 3$ akkor$1 - \delta$ valószínűséggel vetjük el a null-hipotézist. A másodfajú hiba kiszámításához határozzuk meg a $\mathcal{B}(25, 0, 1)$ binomiális eloszlás szerinti valószínűségét annak az eseménynek, hogy a null-hipotézst elfogadju, azaz $x \leq 2$ plusz $\delta \cdot P_1(3)$.

Válasz: A döntésben szereplő szorzó $\delta = 0,828$, a másodfajú hiba valószínűsége 0,725.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen X_1 egy egyelemű, p paraméterű geometriai eloszlású minta. A H_0: $p = 0,5$ versus H_1: $p = 0,9$ esetén a mekkora a terjedelme annak a véletlenített próbának, amelynek próbafüggvénye

$$\Psi(X_1) = \begin{cases} 0 & k \geq 3 \\ 0,5 & k = 2 \\ 1 & k = 1 \end{cases}$$

Adjuk meg a másodfajú hiba valószínűségét is!

Tipp: Az előző feladathoz hasonló módon járunk el, azzal a könnyebbséggel, hogy itt a próbafüggvény adott és a hibavalószínűségeket kell kiszámítani. (A geometriai

eloszlás megfelelő valószínűségeit l. .)

Válasz: Terjedelem: 0,375.

Másodfajú hiba 0,046.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n$ független, λ paraméterű exponenciális eloszlású minta. Konstruáljuk meg a $H_0 : \lambda = \lambda_0$ és $H_1 : \lambda = \lambda_1 > \lambda_0$ egyszerű alternatívához tartozó ε terjedelmű próbát a Neyman-Pearson alaplemma segítségével!

Tipp: Miveλ_0-le$\lambda_1$nában szereplő likelihood hán$\lambda_1 > \lambda_0$az $Y = \sum_{j=1}^{n} X_j$ monoton függvénye (a monotonitás iránya függ és viszonyátol) a próbafüggvény esetben

$$\Psi(Y) = \begin{cases} 0 & Y \geq c \\ 1 & Y < c \end{cases}$$

alakú lesz. Ha $F \sim \mathcal{G}(n, \lambda_0)$ akkor a $c = F^{-1}(\varepsilon)$ lesz az alkalmas konstans.

Válasz: Az Útmutató alapján c értéke konkrét n és λ_0 értékekre kiszámolható, l. ábra.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) $X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ független minta. Konstruáljuk meg a $H_0 : \sigma = \sigma_0$ és $H_1 : \sigma = \sigma_1$ egyszerű alternatívához tartozó ε terjedelmű próbát a Neyman-Pearson alaplemma segítségével!

Tipp: Mivel -lemmában szereplő likelihood hányados az $Y = \sum_{j=1}^{n} X_j^2$ monoton függvénye (a monotonitás iránya függ σ_0 és σ_1 viszonyától).

Válasz: A próbafüggvény $\sigma_1 > \sigma_0$ esetben

$$\Psi(Y) = \begin{cases} 0 & Y \leq c \\ 1 & Y > c \end{cases}$$

alakú lesz. Ha $F \sim \chi^2(n)$ akkor a $c = F^{-1}(1 - \varepsilon)$ lesz az alkalmas konstans.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Írjuk fel n elemű mintára a likelihood-hányados próba $\lambda_n(\mathbf{X})$ statisztikáját, ahol

(a) $\mathbf{X} \sim geom(p)$ és $H_0 : p = p_0$ vs $H_1 : p \neq p_0$.

(b) $\mathbf{X} \sim Poisson(\lambda)$ és $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$.

(c) $\mathbf{X} \sim exp(\lambda)$ és $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$.

(d) $\mathbf{X} \sim U(0, \theta)$ és $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

Tipp: Az (a), (b), (c) esetekben alkalmazzuk a A paragrafusban adott formulát. A számlálóban a likelihood függvénynek az az alakja szerepel, amelyben a paraméter az *egyszerű* null-hipotézishez tartozó érték; nevezőben pedig (ahol a szuprémum szerepel) a likelihood függvénynek az az alakja szerepel, amelyben a paraméter helyett annak M-L becslése áll. A (d) eset külön meggondolást igényel.

Válasz:

(a) Legyen $Y = \sum_{j=1}^{n} X_j$, és $\hat{p} = n/Y$

$$\lambda_n(\mathbf{X}) = \frac{p^n (1-p)^{Y-n}}{\hat{p}^n (1-\hat{p})^{Y-n}}$$

(b) Legyen $Y = \sum_{j=1}^{n} X_j$, és $\hat{\lambda} = Y/n$

$$\lambda_n(\mathbf{X}) = \frac{\lambda^Y e^{-\lambda}}{\hat{\lambda}^Y e^{-\hat{\lambda}}}$$

(c) Legyen $Y = \sum_{j=1}^{n} X_j$, és $\hat{\lambda} = n/Y$

$$\lambda_n(\mathbf{X}) = \frac{\lambda^n e^{-n\lambda Y}}{\hat{\lambda}^n e^{-n\hat{\lambda} Y}}$$

(d) Ha $X_n^* > \theta_0$ elvetjük a null-hipotézist, mert egy lehetetlen esemény következett be.

Ellenkező esetben $\lambda_n(\mathbf{X}) = \frac{X_n^*}{\theta_0}$

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $H_0 : \sigma = Y_n \sim H_1 : \sigma \neq 1$ független minta, mindkét paraméter ismeretlen (elegendően nagy). Legyen és . Konstruáljunk ezekhez 0,05 terjedelmű likelihood-hányados próbát!

Tipp: Az egyszerűség kedvéért legyen $S = \frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{n}$ Ismeretes, hogy $nS\chi^2(n-1)$ eloszlasú.

A próbastatisztika:

$\log\left(S^{n/2}e^{-nS/2+n/2}\right)$

Válasz: Na itt a csalás: logaritmálás es 2-vel való szorzás után a próbastatisztika: $n\log S + n - nS$

Felhasználva a $\log(1+\varepsilon) \approx \varepsilon$ közelítést és a centrális határeloszlás tételt, mely szerint $\sqrt{n}(S-1) \sim \mathcal{N}(0,1)$, $n\log S \approx n(S-1)$. Ugyanakkor a próbastatisztika második tagja pontosan $n(1-S)$.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyenek $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ és $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$ független minták.

(a) Írjuk fel a $H_0 : \sigma = \sigma_0$ és $H_1 : \sigma \neq \sigma_0$ hipotézisekhez konstruált likelihood-hányados próba statisztikáját!

(b) Írjuk fel a $H_0 : \mu_1 = \mu_2$ és $H_1 : \mu_1 \neq \mu_2$ hipotézisekhez konstruált likelihood-hányados próba statisztikáját, ha σ ismert!

(c) Írjuk fel a $H_0 : \mu_1 = \mu_2$ és $H_1 : \mu_1 \neq \mu_2$ hipotézisekhez konstruált likelihood-hányados próba statisztikáját, ha σ ismeretlen!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független minta. Tekintsük a $H_0 : \sigma = \sigma_0$ és $H_1 : \sigma > \sigma_0$ hipotéziseket, és azt a próbát, amelyre $\mathcal{X}_k = \{\mathbf{x} : nS_n^2/\sigma_0^2 > c\}$

($S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ az empirikus szórásnégyzet). Torzítatlan-e az adott próba?

Tipp: Keressük meg a ben a χ^2 eloszlás sűrűségfüggvényét, és alkalmazzuk az $y = \sigma x$ helyettesítést:

$$f_Y(x) = \frac{x^{n/2-1}e^{-x/(2/\sigma)}}{(\sigma 2)^{n/2}\Gamma(n/2)}, \quad y \geq 0.$$

Vizsgáljuk meg, hogy a próba ereje hova tart, ha $\sigma \to \infty$!

Válasz: Nem.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Igaz-e, hogy az ε terjedelmű (kétoldali) u-próba pontosan akkor fogadja el a nullhipotézist, ha μ_0 benne van az $\overline{X}$ segítségével μ-re szerkesztett $1-\varepsilon$ szintű konfidencia-intervallumban?

Tipp: Írjuk fel az elfogadási tartomány és alakítsuk át!

Válasz: Igaz.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$ független minta, (σ_0 ismert). Legyen $H_0 : \mu = \mu_0$ és $H_1 : \mu \neq \mu_0$. Konstruáljunk ezekhez 0,05 terjedelmű likelihood-hányados próbát! Vessük össze a kapott próbát az u-próbával (két- és egyoldali változatával is)!

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $(X_1, Y_1), \ldots (X_n, Y_n) \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$, ahol

$\mathbf{m}^\top = (\mu_1, \mu_2)$ és $\mathbf{C} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$.

Tegyük fel, hogy a szórások ismertek. Szerkesszünk $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ hipotézisekre

(a) kétmintás u-próbát!

(b) Alkalmazzunk önkontrollos vizsgálatot!

Tipp: Mindkét esetben az $\overline{X} - \overline{Y}$ valószínűségi változó σ_e^2 szórásnégyzetét kell meghatározni.

Válasz: A próbastatisztika a standard normális eloszlású $\overline{X} - \overline{Y}/\sigma_e$.

(a) A kétmintás u-próbánál $\sigma_e^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

(b) Az önkontrollos vizsgálatnál $\sigma_e^2 = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}$.

A két próbafüggvény azonos, mert $n = n_1 = n_2$.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független minta. Tekintsük a $H_0 \colon \mu = \mu_0$ versus $H_1 \colon \mu \neq \mu_0$ (a szórásnégyzet ismeretlen) t-próba statisztikáját:

$$t(\mathbf{X}) = \frac{\overline{X} - \mu_0}{S_n^*/\sqrt{n}}$$

(a) Igazoljuk, hogy a likelihood-hányados próbához tartozó statisztika

$$\lambda_n(\mathbf{X}) = \left(\frac{\sum_{j=1}^n (X_j - \overline{X})^2}{\sum_{j=1}^n (X_j - \mu_0)^2} \right)^{n/2}$$

alakú.

(b) Igazoljuk, hogy

$$\lambda_n(\mathbf{X}) = \left(\frac{1}{1 + \frac{t^2(\mathbf{X})}{n-1}} \right)^{n/2}.$$

(c) Mutassuk meg, hogy ez azt jelenti, hogy a fenti likelihood-hányados próba a t-próba kétoldali változatával ekvivalens!

Tipp: A paragrafusban keressük meg a likelihood-hányados próba szerkesztésének módját. Itt a paramétertér 2 dimenziós:

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\},$$

a 0-hipotézis által kijelölt 1-dimenziós részsokaság pedig

$$\Theta_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}.$$

Az $\mathbf{X} = (X_1, \ldots, X_n)$ független, azonos eloszlású minta alapján felírjuk az

$$L_{\mu,\sigma^2}(\mathbf{X}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right)$$

likelihood-függvényt, majd vesszük ennek szuprémumát a Θ illetve a Θ_0 halmazon:

$$\sup_{(\mu,\sigma^2)\in\Theta} L_{\mu,\sigma^2}(\mathbf{X}) = \frac{1}{\left(2\pi\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2\right)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{2\left(\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2\right)}\right) =$$
$$= \left(\frac{n}{2\pi\sum_{i=1}^n (X_i-\bar{X})^2}\right)^{n/2} e^{-\frac{n}{2}},$$

$$\sup_{(\mu,\sigma^2)\in\Theta_0} L_{\mu,\sigma^2}(\mathbf{X}) = \frac{1}{\left(2\pi\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2\right)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{2\left(\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)^2\right)}\right) =$$
$$= \left(\frac{n}{2\pi\sum_{i=1}^n (X_i-\mu_0)^2}\right)^{n/2} e^{-\frac{n}{2}},$$

A fenti számolásból nyilvanvalóan adódik (a), és egyszerű algebrai átalakításokkal (b).

Válasz: Az Útmutatóban (a) és (b) megoldása már szerepel, a (c) abból következik, hogy likelihood-hányados próba statisztika monoton függvénye a kétoldali t-próba statisztikájának.

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Határozzuk meg az egyoldali u -próba erőfüggvényét! Igazoljuk, hogy a próba torzítatlan és konzisztens is! Hogyan változik a próba ereje, ha

(a) ε,

(b) $\theta-\theta_0$,

(c) n nő?

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Tekintsük az $(X_1,Y_1),\dots,(X_n,Y_n)$ mintát és az r_{sp} Spearman-féle rangkorrelációs együtthatót.

(a) Igazoljuk, hogy $|r_{sp}|\le 1$ és egyenlőség pontosan akkor teljesül, ha minden $i\neq j$ párra $X_i\le X_j$ az $Y_i\le Y_j$, illetve $Y_i\ge Y_j$ relációt vonja maga után (r_{sp} előjelének megfelelően).

(b) Igazoljuk, hogy ha a háttérváltozók függetlenek, akkor $\mathrm{E}(r_{sp})=0$.

Tipp:

Válasz:

(iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii) Legyen $X_1,X_2,\dots\sim exp(\lambda)$ független azonos eloszlású minta. Adjunk a $H_0:\lambda=\lambda_0$ vs. $H_1:\lambda=\lambda_1$ egyszerű alternatíva eldöntésére szekvenciális eljárást (ε_1 elsőfajú és ε_2 másodfajú hibával)! Adjuk meg a várható lépésszámokat!

Tipp:

Válasz:

3. Tesztek

(i) Milyen eloszlású a λ paraméterű exponenciális eloszlásból vett n elemű rendezett minta első eleme?

(a) exp($n\lambda$)

(b) exp(λ/n)

(c) Gamma(n,λ)

(d) Béta($1,n$)

Válasz: (a)

(ii) Tekintsünk egy $N(m, \sigma^2)$ vett mintát, legyen $\overline{X}$ a mintaátlag. Igaz-e, hogy $\overline{X}$ elégséges statisztika (m, σ^2) paraméternek?

(a) igen, a Neyman-Fisher faktorizáció miatt

(b) igen, mivel torzítatlan becslése a várható értéknek

(c) nem, mert két paraméterre nem lehet megadni elégséges statisztikát

(d) nem, mert a mintának a mintaátlagra vett feltételes eloszlása μ-től független, de σ^2-től nem.

Válasz: (d)

(iii) Az alábbiak közül melyik az exponenciális eloszlás várható értékére elégséges statisztika?

(a) X_n^*

(b) $X_{\lfloor n/2 \rfloor} * + X_{\lceil n/2 \rceil} *$

(c) $X_1 \dots X_n$

(d) $X_1 + \ldots + X_n$

Válasz: (d)

(iiii) Tekintsünk egy n elemű $N(m, \sigma^2)$ eloszlásból vett mintát. Milyen becslése σ^2 -nek $(\sum_{i=1}^n X_i^2 - \overline{X}^2)/(n+1)$?

(a) Nem torzítatlan, de aszimptotikusan torzítatlan, erősen konzisztens.

(b) Nem torzítatlan, de aszimptotikusan torzítatlan, gyengén sem konzisztens.

(c) Torzítatlan, a Cramér-Rao egyenlőtlenség alapján hatásos, erősen konzisztens.

(d) Torzítatlan, de a Cramér-Rao egyenlőtlenség alapján nem hatásos, erősen konzisztens.

Válasz: (a)

(iiiii) Tekintsünk egy n elemű $N(0, \sigma^2)$ eloszlásból vett mintát. Milyen becslése σ^2-nek $(\sum_{i=1}^n X_i^2)/n$?

(a) Nem torzítatlan, de aszimptotikusan torzítatlan, erősen konzisztens.

(b) Nem torzítatlan, de aszimptotikusan torzítatlan, gyengén sem konzisztens.

(c) Torzítatlan, a Cramér-Rao egyenlőtlenség alapján hatásos, erősen konzisztens.

(d) Torzítatlan, de a Cramér-Rao egyenlőtlenség alapján nem hatásos, erősen konzisztens.

Válasz: (c)

(iiiiii) Tekintsünk egy n elemű $U(0, \theta)$ eloszlásból vett mintát. Milyen becslése θ-nak a maximum likelihood becslés?

(a) Nem torzítatlan, de aszimptotikusan torzítatlan, erősen konzisztens.

(b) Nem torzítatlan, de aszimptotikusan torzítatlan, gyengén sem konzisztens.

(c) Torzítatlan, hatásos, gyengén konzisztens.

(d) Torzítatlan, nem hatásos, gyengén konzisztens.

Válasz: (a)

(iiiiiii) Tekintsünk egy n elemű Poisson(λ) eloszlásból vett mintát. Milyen becslése λ-nak a momentumok módszerével vett becslés?

(a) Nem torzítatlan, de aszimptotikusan torzítatlan, erősen konzisztens.

(b) Nem torzítatlan, de aszimptotikusan torzítatlan, gyengén sem konzisztens.

(c) Torzítatlan, hatásos, erősen konzisztens.

(d) Torzítatlan, nem hatásos, erősen konzisztens.

Válasz: (c)

(iiiiiiii) Mi a kapcsolat a normális eloszlás várható értékére ismeretlen szórás esetén adott konfidenciaintervallumnak és a t-próba között?

(a) A t-próba elfogadja a nullhipotézist, ha tesztelt érték a konfidenciaintervallumba esik.

(b) A t-próba elfogadja a nullhipotézist, ha $\overline{X}$ a konfidenciaintervallumba esik.

(c) A t-próba elutasítja a nullhipotézist, ha tesztelt érték a konfidenciaintervallumba esik.

(d) A t-próba elutasítja a nullhipotézist, ha $\overline{X}$ a konfidenciaintervallumba esik.

Válasz: (a)

(iiiiiiiii) Létezik-e az exponenciális eloszlás paraméterére vonatkozó, $H_0 : \lambda = \lambda_0$ és $H_1 : \lambda = \lambda_1$ hipotéziseket tesztelő ε terjedelmű legerősebb próba ($\varepsilon > 0$ tetszőleges)?

(a) Nem, mert $1/\overline{X}$ nem torzítatlan becslése λ-nak.

(b) Igen, a likelihood-hányados próba ilyen.

(c) Igen, a Neyman-Pearson alaplemma alapján.

(d) Igen, a Wald-féle szekvenciális eljárás ilyet ad.

Válasz: (c)

(iiiiiiiii) Mennyi az ε terjedelmű egymintás, egyoldali u-próba másodfajú hibája?

(a) $1 - \varepsilon$

(b) $1/\varepsilon$

(c) $\beta_n(m\varepsilon) = 1 - \Phi(u_\varepsilon - (\mu - \mu_0)/(\sigma_0/\sqrt{n}))$

(d) $1 - \beta_n(m\varepsilon) = \Phi(u_\varepsilon - (\mu - \mu_0)/(\sigma_0/\sqrt{n}))$

Válasz: (d)

(iiiiiiiiii) Az egymintás egyoldali u-próba

(a) torzítatlan és konzisztens.

(b) nem torzítatlan de konzisztens.

(c) torzítatlan de nem konzisztens.

(d) nem torzítatlan és nem konzisztens.

Válasz: (a)

(iiiiiiiiiiii) Alkalmazható-e a t próba ismert szórás esetén?

(a) Igen.

(b) Csak normális eloszlású kis minta esetén.

(c) Csak normális eloszlású nagy minta esetén.

(d) Nem, mert az ismeretlen szórás feltétel, ismert szórás esetén csak az u próbát alkalmazhatjuk.

Válasz: (a)

(iiiiiiiiiiiii) Mikor használhatjuk a χ^2 próbákat?

(a) Mindig.

(b) Diszkrét háttérváltozó esetén mindig, folytonos háttérváltozó diszkretizálása esetén csak nagy mintaelemszám mellett.

(c) Az illeszkedévizsgálatra vonatkozó χ^2 próbát mindig, a többit csak nagy mintaelemszám esetén.

(d) Csak nagy mintaelemszám esetén (mindegyiket, minden háttérváltozó esetén).

Válasz: (a)

3. fejezet - A többdimenziós normális eloszlás, Wishart eloszlás

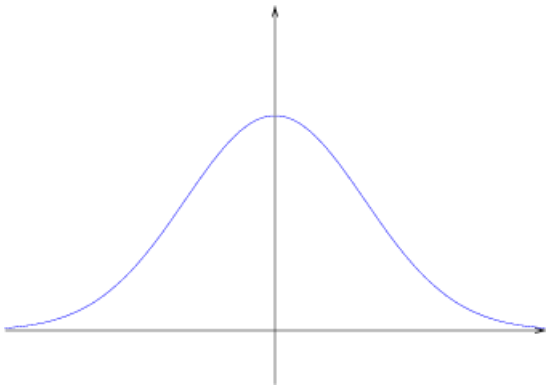
1. Elméleti háttér

1.1. Többdimenziós normális eloszlás

A p -dimenziós, nem-elfajult normális eloszlást az p -dimenziós standard normális eloszlás lineáris transzformáltjaként vezetjük be.

3.1.1.1. Definíció. *Azt mondjuk, hogy az* $\mathbf{Y}$ *véletlen vektor* p-dimenziós standard normális eloszlású, *ha komponensei 1-dimenziós standard normális eloszlásúak és függetlenek. Erre az* $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ *jelölést használjuk, utalva arra, hogy a* p *-dimenziós* $\mathbf{Y}$ *véletlen vektor várható érték vektora a* $\mathbf{0}$ *vektor, kovarianciamátrixa pedig* $\mathbf{I}_p$ *(ezek az eloszlás paraméterei).*

Az alábbi ábra mutatja a standard normális eloszláshoz, azaz $\mathcal{N}(0,1)$-hez tartozó sűrűségfüggvényt.



$\phi(x)$ $\mathbf{Y}$ sűrűségfüggvénye a függetlenség miatt a komponensek sűrűségfüggvényeinek szorzata, azaz

$$g(\mathbf{y}) = \prod_{i=1}^{p} \phi(y_i) = \frac{1}{\sqrt{2\pi}^p} e^{-(\sum_{i=1}^{p} y_i^2)/2} = \frac{1}{(2\pi)^{p/2}} e^{-\|\mathbf{y}\|^2/2},$$

ahol ϕ jelöli a standard normális sűrűségfüggvényt (Gauss-görbét), az $\mathbf{y} = (y_1, \ldots, y_p)^T$ vektor pedig az együttes sűrűségfüggvény argumentuma.

Alkalmazzuk most a fenti $\mathbf{Y}$-ra az

$$\mathbf{X} = \mathbf{AY} + \mathbf{m} \qquad (1)$$

???lineáris transzformációt, ahol $\mathbf{A}$ $p \times p$-s nem-szinguláris mátrix, $\mathbf{m}$ pedig p-dimenziós vektor. Könnyű látni, hogy $\mathbf{X}$ várható érték vektora $\mathbf{m}$, kovarianciamátrixa pedig:

$$\mathbf{C} = \mathbb{E}(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T = \mathbb{E}(\mathbf{AY})(\mathbf{AY})^T =$$
$$= \mathbb{E}(\mathbf{AYY}^T\mathbf{A}^T) = \mathbf{A}\mathbb{E}(\mathbf{YY}^T)\mathbf{A}^T = \mathbf{AI}_n\mathbf{A}^T = \mathbf{AA}^T,$$

ahol a vektorok oszlopvektorok, egy vektor várható értéke a komponensek várható értékeiből álló vektor, egy mátrix várható értéke pedig az elemeinek a várható értékeiből álló mátrix.

3.1.1.2. Definíció. *Az* $\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ *többdimenziós standard normális eloszlású véletlen vektor-ból a fenti (invertálható) lineáris transzformációval kapott* $\mathbf{X}$ *véletlen vektort* nem-elfajult többdimenziós normális eloszlás*ának nevezzük, és ennek kifejezésére röviden az* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ *formulát használjuk.*

A nem-elfajult p-dimenziós normális eloszlású $\mathbf{X}$ véletlen $p \times p$-ktor eloszlásának paraméterei tehát a p dimenzió, $c_{ij} = c_{ji}$ ráható X_i ték változó X_j tor és a kovarianciamátrix. A $(i \neq j)$-s c_{ii} zimmetrikus, pozitív definit mátrix elemei: az és $\mathbb{D}^2\mathbf{X}$ nponensek kovarianciája $\mathbf{A}$, pedig szórásnégyzete (varianciája). A kovarianciamátrix $\mathbf{C} = \mathbf{A}\mathbf{A}^T$ jelölést fogjuk használni. Ha -ról kikötjük, hogy négyzetes és nem-szinguláris mátrix, akkor a kovarianciamátrix pozitív definit. Megjegyezzük, hogy szinguláris $\mathbf{C}$ mátrixszal végrehajtva 3.1 transzformációt, szinguláris, pozitív szemidefinit -hez jutunk. Ilyen esetekben rangja is kisebb lesz, mint , ekkor *elfajult többdimenziós normális eloszlás*ról beszélünk.

A továbbiakban, hacsak külön nem mondjuk, akkor mindig a nem-elfajult esetre gondolunk.

3.1.1.3. Állítás. *Ha a $\mathbf{C}$ mátrix invertálható, akkor az $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ véletlen vektor sűrűségfüggvénye:*

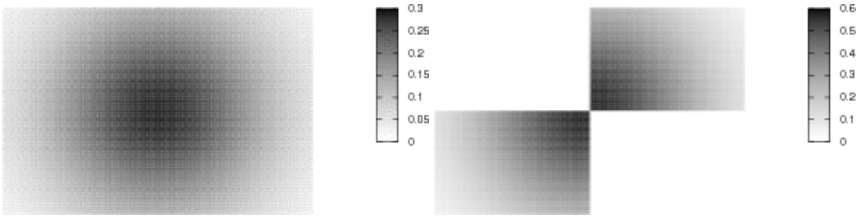
$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}, \quad \mathbf{x} \in \mathbb{R}^p. \qquad (1)$$

Megjegyezzük, hogy az elfajult többdimenziós normális eloszlás alacsonyabb dimenziós sűrűségfüggvénye például úgy kapható meg, hogy az (3.2) képletben $\mathbf{C}^{-1}$ helyett $\mathbf{C}^{+}$-t írunk (azaz a szinguláris $\mathbf{C}$ mátrix általánosított inverzét, l.) $|\mathbf{C}|$ helyett pedig $\mathbf{C}$ pozitív sajátértékeinek szorzatát.

3.1.1.4. Állítás. *Az $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ véletlen vektor komponensei pontosan akkor teljesen függetlenek, ha a $\mathbf{C}$ kovarianciamátrix diagonális.*

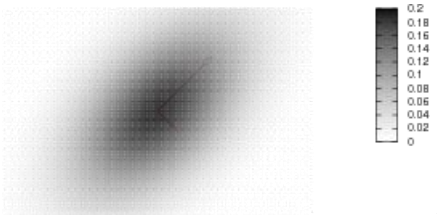
Megjegyezzük, hogy $p = 2$ esetén $\mathbf{Y}$ sűrűségfüggvénye körszimmetrikus és maximumhelye az origóban van.

Az alábbi ábrákon látható a kétdimenziós standard normális eloszlás sűrűsége és egy, a segítségével konstruált olyan együttesen nem normális eloszlás sűrűsége, amely marginálisai standard normálisok.



2 dimenziós standard normális és nem 2 dimenziós normális sűrűség???

$\mathbf{X} = \mathbf{A}\mathbf{Y} + \mathbf{m}$ sűrűségfüggvényének a maximumhelye viszont $\mathbf{m}$-ben van, nívóhalmazai pedig ellipszisek, melynek tengelyirányait a nem-szinguláris $\mathbf{C}$ kovarianciamátrix sajátvektorai jelölik ki, a tengelyek hossza pedig a megfelelő sajátértékek négyzetgyökével arányos.



2 dimenziós normális normális sűrűség egy szintvonal tengelyeivel???

Ez a legegyszerűbben az (1.2)-beli sűrűségfüggvény exponensében álló kvadratikus alak

$$(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) = (\mathbf{x}-\mathbf{m})^T\mathbf{U}\Lambda^{-1}\mathbf{U}^T(\mathbf{x}-\mathbf{m}) = \mathbf{z}^T\Lambda^{-1}\mathbf{z} =$$
$$= \sum_{i=1}^{2} \frac{1}{\lambda_i} z_i^2 = \frac{z_1^2}{\sqrt{\lambda_1}^2} + \frac{z_2^2}{\sqrt{\lambda_2}^2} \qquad (1)$$

főtengely-transzformációjából látható; a nívóhalmazokat úgy kapjuk, hogy a fenti kvadratikus alakot valamely nemnegatív konstanssal tesszük egyenlővé. (Gondoljuk meg, milyen értékhatárok közt mozoghat e konstans ahhoz, hogy valódi ellipsziseket kapjunk!) Az is látható, hogy a nívóhalmazok pontosan akkor körök, hogy ha a

sajátértékek egyenlőek, ez viszont ekvivalens azzal, hogy a komponensek függetlenek és azonos szórásúak. Ezt mindjárt általános p-re is belátjuk.

Egy $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ valószínűségi változó sűrűségében álló kvadratikus alak hasonló módon

$$(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) = \mathbf{z}^T\Lambda^{-1}\mathbf{z} = \sum_{i=1}^{p}\frac{1}{\lambda_i}z_i^2 = \sum_{i=1}^{p}\frac{z_i^2}{\sqrt{\lambda_i}^2}$$

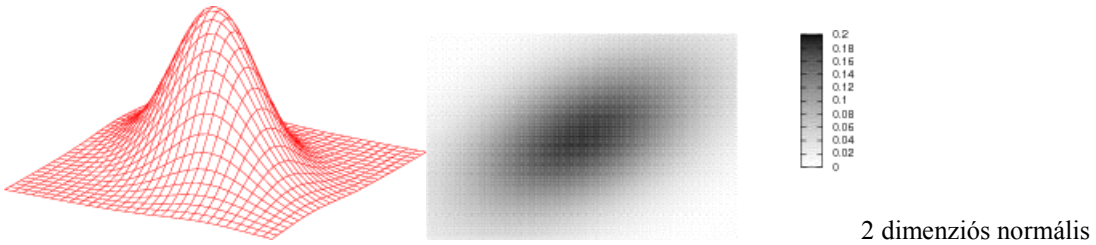
alakúvá transzformálható a $\mathbf{z} = \mathbf{U}^T(\mathbf{x}-\mathbf{m})$ koordinátatranszformációval (ami egy eltolást, majd egy forgatást jelent). Eredményképp egy olyan p-dimenziós ellipszoid egyenletét kapjuk, mely főtengelyeinek hossza a sajátértékek gyökével arányos, irányukat pedig a sajátvektorok jelölik ki. Az ellipszoid pontosan akkor lesz gömb, ha $\lambda_1 = \cdots = \lambda_p = \lambda$, ekkor a kovarianciamátrix

$$\mathbf{C} = \mathbf{U}(\lambda\mathbf{I}_p)\mathbf{U}^T = \lambda\mathbf{I}_p$$

alakú, ami ekvivalens azzal, hogy a komponensek függetlenek és azonos ($\sqrt{\lambda}$) szórásúak. Könnyű látni, hogy amennyiben a komponensek függetlenek, de nem azonos szórásúak, ellipszoidot kapunk, melynek tengelyirányai a koordinátatengelyekkel párhuzamosak. Minden más esetben olyan ellipszoidok adódnak nívófelületekként, melyek tengelyei (legalábbis egy részük) elfordulnak (2-dimenziós esetben az elfordulás szögéből következtethetünk a két komponens közti korreláció mértékére): az alábbi ábrákon a $\mathbf{0}$ várható érték vektorú,

$$\begin{pmatrix} 1 & 0.6 \\ 0.6 & 2 \end{pmatrix}$$

kovarianciamátrixú 2-dimenziós normális eloszlás sűrűségfüggvénye láthatók 3 dimenziós és szürkeárnyalatos ábrázolásban.



2 dimenziós normális sűrűségek

A későbbiekben használni fogjuk a következő tételt.

3.1.1.5. Tétel. *Ha $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ és a $\mathbf{C}$ kovarianciamátrix pozitív definit, akkor*

$$(\mathbf{X}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{X}-\mathbf{m}) \sim \chi^2(p).$$

Az érdekesség kedvéért megemlítjük a normális eloszlás egy Harald Cramértől származó karakterizációját.

3.1.1.6. Tétel. *Ha X és Y független valószínűségi változók és $X+Y$ normális eloszlású, akkor X és Y külön-külön is normális eloszlásúak.*

A statisztikai vizsgálatokban előforduló véletlen változók általában **nem** együttesen normális eloszlásúak, a normális eloszlásra kiszámolt statisztikai módszerek alkalmazásása indokolható az alábbi Tétellel. Emellett a skalár, sőt a diszkrét értékű valószínűségi változók statisztikai vizsgálatában olyan gyakran alkalmazott módszerek mint a χ^2-próba jogosságának indoklásában is szükségünk van a centrális határeloszlás tétel többdimenziós alakjára.

3.1.1.7. Tétel. *Legyenek $\mathbf{X}_1, \mathbf{X}_2, \ldots$ független, azonos eloszlású p-dimenziós véletlen vektorok, melyek $\mathbf{m}$ várható érték vektora és $\mathbf{C}$ kovarianciamátrixa létezik (utóbbi nem feltétlenül invertálható). Legyen $\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$, $n = 1, 2, \ldots$. Akkor a standardizált részletösszegek sorozata, azaz az $\frac{1}{\sqrt{n}}(\mathbf{S}_n - n\mathbf{m})$ véletlen vektor sorozat eloszlása konvergál az $\mathcal{N}_p(\mathbf{0}, \mathbf{C})$ eloszláshoz, ha $n \to \infty$.*

Itt jegyezzük meg, hogy n növelésével a többdimenziós normális eloszlás $n^{C/\varepsilon}$ színűségeinek numerikus integrálással történő kiszámításának a műveletigénye C megengedett hiba esetén nagyságrendű, még abban az esetben is, amikor egy -dimenziós téglatest kovarianciamátrixú normális eloszlás szerinti valószínűségét akarjuk meghatározni. Léteznek az Hermite-polinomok szerinti sorfejtésen alapuló módszerek, de ezek csak akkor működnek, ha közel van az -dimenziós egységmátrixhoz (növelésével a korrelációknak csökkenni $1/\varepsilon^2$. Nagy értékre a Monte Carlo módszert kell alkalmazni, ennek műveletigenye a dimenziótól függetlenül .

3.1.1.8. Állítás. *Az $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ véletlen vektor komponensei pontosan akkor teljesen függetlenek, ha a $\mathbf{C}$ kovarianciamátrix diagonális.*

A későbbiekben használni fogjuk a következő tételt.

3.1.1.9. Tétel. *Ha $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ és a $\mathbf{C}$ kovarianciamátrix pozitív definit, akkor*

$$(\mathbf{X}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{X}-\mathbf{m}) \sim \chi^2(p).$$

A korrelációs mátrixok gyakran képezik a statisztikai vizsgálatok alapját, geometriai struktúrájukat azonban csak az elmúlt 20 évben kezdték feltérképezni. A 2×2-es korrelációs mátrixok struktúrája a nyilvánvaló, a $\mathbf{C}$ főtlátlóban 1-ek állnak, így egyetlen szabad paraméterük van $(-1 \le \rho \le 1)$.

A 3×3-as korrelációs mátrixok struktúrája mmár eléggé összetett de még áttekinthető.

A következőkben elsősorban Grone és társszerzői [16] dolgozatára támaszkodunk. A $\mathbf{C}$ mátrixot most paraméterezzük a a következőképpen:

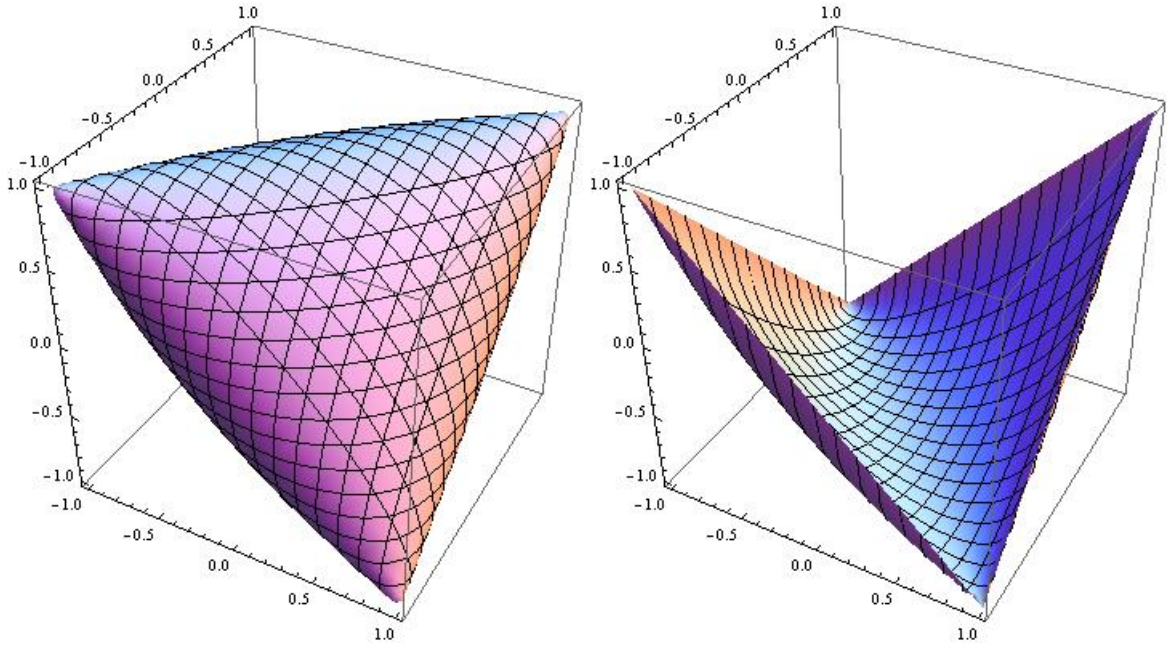
$$\mathbf{C} = \begin{pmatrix} 1 & x & z \\ x & 1 & y \\ z & y & 1 \end{pmatrix}$$

Az $(x, y, z)^T$ vektor nyilván mindig $[-1,1] \times [-1,1] \times [-1,1] \subset \mathbb{R}^3$ kockán van, és ezek a vektorok konvex halmazt alkotnak.

Most pontosan leírjuk ennek a konvex halmaznak a határát. Vegyük észre, hogy a kocka 8 csúcsából 4 NEM lehet a keresett konvex halmaz eleme; ezek: $(-1,-1,-1)(-1,1,1)(1,-1,1)(1,-1,1)$.

Viszont az $(1,-1,-1)(-1,1,-1)(-1,-1,1)(1,1,1)$ lehetséges

paramétervektorok. Konvex burkuk egy tetraéder, melynek 6 élea kocka 6 lapjának egy-egy átlója. A [16] dolgozat (implicit módon) bebizonyítja, hogy a keresett határ a ??? ábrákon látható „felfújt" tetraéder, aminek határát a kocka lap-párjával párhuzamos síkokban levő ellipszissereg alkotja. Innen az **elliptóp** elnevezés. Pl. A vízszintes - (x, y) - síkokban levő ellipszisek a kocka alaplapján $[(1,-1-1),(-1,1,-1)]$ szakasszá fedőlapján a $[(-1,-11),(1,1,1)]$ szakasszá fajulnak el, míg a $z = 0$ síkon egy, a tetraéder maradék 4 élét metsző körbe mennek át. Az ellipszisek minden $y = c(-1 \le c \le 1)$ síkban metszik a tetraéder maradék 4 élét. Ez a tulajdonság még nem határozza meg az ellipsziseket; figyelembe kell vennünk, hogy a lehetséges (x, y, z) paraméterek halmazának határán legfeljebb 2 rangú korrelációs mátrix paraméterei lehetnek, ami a $\det|\mathbf{C}| = 0$ feltétellel ekvivalens. így egy adott síkban fekvő ellipszisnek legalább 5 különböző pontját adhatjuk meg, ami egyértelműen meghatározza az ellipszist. Így készültek az ábrák, *animáció* és *interaktív animáció*.



elliptop

1.2. Wishart eloszlás

A többdimenziós normális eloszlás paramétereinek becsléséhez és a paraméterekre vonatkozó hipotézisek vizsgálatához. Ehhez szükségünk van a becslésekben fellépő többdimenziós statisztikák eloszlásának meghatározására.

3.1.2.1. Definíció. *A $p \times p$-s $\mathbf{W}$ véletlen mátrixot p-dimenziós, n* szabadságfokú, *$\mathbf{C}$ kovarianciájú (centrális)* Wishart-mátrix*nak nevezzük, ha előállítható $\mathbf{W} = \mathbf{X}\mathbf{X}^T$ alakban, ahol a $p \times n$-es $\mathbf{X}$ véletlen mátrix oszlopvektorai függetlenek és $\mathcal{N}_p(\mathbf{0}, \mathbf{C})$-eloszlásúak. Egy ilyen $\mathbf{W}$ véletlen mátrix elemeinek együttes eloszlását $p, n, \mathbf{C}$ paraméterű (centrális)* Wishart-eloszlás*nak nevezzük, és a következőképpen jelöljük: $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{C})$.*

$\mathbf{W}$ szimmetriája miatt valójában $p(p+1)/2$-dimenziós eloszlásról van szó. Megjegyezzük, hogy a nem-centrális Wishart-eloszlás definíciója ugyanígy kezdődik, csak ott $\mathbf{X}$ oszlopvektorai független $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ eloszlásúak lesznek. Ilyenekkel mi nem foglalkozunk, és a továbbiakban Wishart eloszláson mindig a centrálisat értjük. Az $\mathbf{X}$ mátrix oszlopvektorait $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$-nel jelölve vegyük észre, hogy $\mathbf{W} = \sum_{k=1}^{n} \mathbf{X}_k \mathbf{X}_k^T$. Az ilyen előállítást diádösszegnek hívjuk. Amennyiben az $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ vektorok független mintaelemek egy $\mathcal{N}_p(\mathbf{0}, \mathbf{C})$ eloszlású véletlen vektorra, az $\mathbf{X}^T$ mátrixot *adatmátrix*nak is szokták nevezni, amely tehát soronként tartalmazza a megfigyeléseket. A $\mathcal{W}_p(n, \mathbf{I})$ eloszlást *standard Wishart-eloszlás*nak nevezzük. Itt tehát az

$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ vektorok p-dimenziós standard normális eloszlásúak. Ha speciálisan $p = 1$, akkor $\mathbf{W} = \sum_{k=1}^{n} X_k^2$, ami definíció szerint $\chi^2(n)$-eloszlású.

3.1.2.2. Tétel. *Legyen a $p \times p$-s $\mathbf{C}$ kovarianciamátrix pozitív definit. $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{C})$ pontosan akkor teljesül, ha $\mathbf{C}^{-1/2}\mathbf{W}\mathbf{C}^{-1/2} \sim \mathcal{W}_p(n, \mathbf{I})$.*

A fenti tétel azt fejezi ki, hogy egy Wishart-mátrix standardizáltja standard Wishart-eloszlású.

Wishart-mátrixra példa az empirikus kovarianciamátrix konstansszorosa. Ezt fogalmazza meg pontosan a következő tétel.

3.1.2.3. Tétel. *Legyen $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ független elemű minta egy $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ eloszlású véletlen vektorra, továbbá legyen*

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k \quad \text{és} \quad \mathbf{S} = \sum_{k=1}^{n} (\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{X}_k - \bar{\mathbf{X}})^T.$$

Akkor

(1) $\bar{\mathbf{X}} \sim \mathcal{N}_p(\mathbf{m}, \frac{1}{n}\mathbf{C})$,

(2) $\mathbf{S} \sim \mathcal{W}_p(n-1, \mathbf{C})$,

(3) $\bar{\mathbf{X}}$ *és* $\mathbf{S}$ *függetlenek egymástól.*

3.1.2.4. Tétel. *Legyenek* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *független azonos eloszlású* $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ *változók* $(p < n)$, *és* $\mathbf{X} := (\mathbf{X}_1, \ldots, \mathbf{X}_n) p \times n$*-es mátrix. Akkor a* $\mathbf{W} = \mathbf{X}\mathbf{X}^T$ *standard Wishart-mátrix sűrűsége*

$$c_{np}|\mathbf{W}|^{\frac{n-p-1}{2}} e^{-\frac{1}{2}\mathrm{tr}\mathbf{W}} \qquad (1)$$

alakú, ahol c_{np} *csak* p*-től és* n*-től függő konstans.*

A bizonyításról (részleteket ld. [5] 217-219. o.) csak annyit jegyzünk meg, hogy az $\mathbf{X}$ véletlen mátrix sűrűségéből kell kindulni, ami nem más, mint az $\mathbf{X}_1, \ldots, \mathbf{X}_n$ független azonos eloszlású minta alapján felírt likelihood-függvény:

$$\frac{1}{(2\pi)^{np/2}} e^{-\frac{1}{2}\mathrm{tr}\mathbf{W}}.$$

Ebből $\mathbf{W}$ eleminek együttes eloszlása mértéktranszformációval határozható meg. Ecélból mátrixok lineáris transzformáltjainak Jacobi-determinánsait kell meghatároznunk (itt $|\mathbf{A}|$ az $\mathbf{A}$ matrix determinánsának abszolút értéke):

(1) $\mathbf{X} = \mathbf{A}\mathbf{Y}$, ahol $\mathbf{A}$ tetszőleges $p \times p$-s nemszinguláris mátrix, $\mathbf{X}$ a $p \times n$-es minta. Közvetlen számolással adódik a

$$\left|\frac{\partial \mathbf{X}}{\partial \mathbf{Y}}\right| = |A|^n.$$

(2) $\mathbf{A}$ mint (1)-ben, $\mathbf{W}$ a $p \times p$ Wishart mátrix, $\mathbf{W} = A\mathbf{V}A^T$. Ekkor az ún. Sverdrup-lemma [27] szerint

$$\left|\frac{\partial \mathbf{W}}{\partial \mathbf{V}}\right| = |A|^{p+1}.$$

A Wishart-mátrix volt az első véletlen mátrix, amit a matematikusok intenzíven tanulmányoztak (1937 óta).

Vegyük észre, hogy a (3.4) formula szerint a Wishart mátrix sűrűségfüggvénye a csak a sajatértékek osszegén és szorzatán (determináns, trace) keresztül függ a a mátrixelemektől, de ez nem a Wishart-mátrix spektrumának az eloszlása. A Wishart mátrix sajátértékeinek empirikus eloszlására vonatkozik a Marcsenko-Pasztur tétel (l. [????]). Tegyük fel, hogy mind n, mind pedig p végtelenbe tart olymódon, hogy $\frac{p}{n} \to c$, ekkor

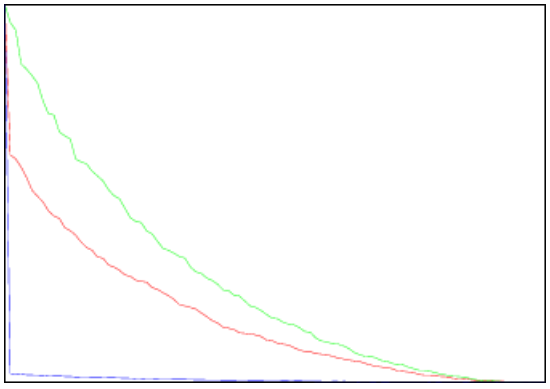
$$\frac{1}{p}\#\left\{\lambda_j^p \colon \lambda_j^p < x\right\} \to F(x), \qquad (1)$$

???

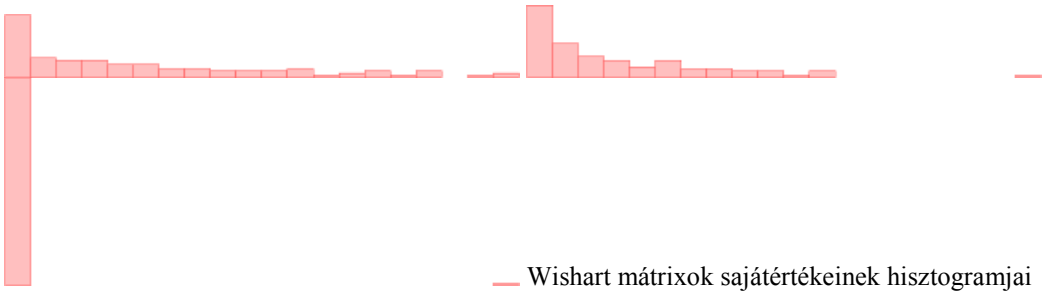
ahol λ_j^p a $W \sim \mathcal{W}_p(n, \mathbf{I})$ mátrix j-edik sajátértéke (monoton nemcsökkenő rendezés mellett) és

$$F'(x) = \frac{1}{2\pi x c}\sqrt{(b-x)(x-a)}, \quad a < x < b.$$

A (3.5) formulabeli konvergencia majdnem biztos, ha $0 < c \leq 1$. Az F eloszlás várható értéke 1, szorásnégyzete $1 + c$.


Wishart-mátrixok sajátértékei

A zöld grafikon standard Wishart mátrix sajátértékeit mutatja, a kék pedig egy olyanét, amelyhez tartozó C mártix minden eleme közel 1. Az előbbi ábra sajátértékei láthatóak hisztogramon is ábrázolva.


Wishart mátrixok sajátértékeinek hisztogramjai

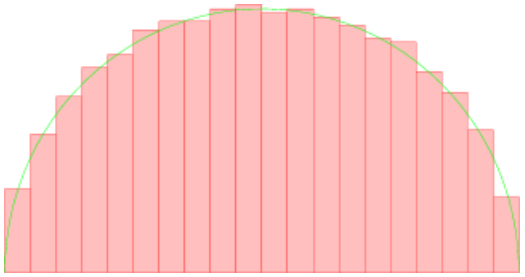
Meglepő módon a legegyszerűbb véletlen mátrix - a független $\mathcal{N}(0,1)$ eloszlású elemekből álló $n \times n$ szimmetrikus mátrix - empirikus spekrumának viselkedést csak az 1940-es években kezdte el tanulmányozni Wigner Jenő, a kaotikus kvantumrendszerek leírása céljából. Az ilyen mátrixok λ sajátértékeinek rendezett mintáját $\mathrm{const} \cdot \sqrt{n}$-nel normálva kapjuk a híres félkör-törvényt.

$$\frac{1}{p}\#\left\{\lambda_j^p \colon \lambda_j^p < x\right\} \to F(x), \qquad (1)$$

???ahol

$$F'(x) = \frac{2}{\pi}\sqrt{(1-x^2)}, \quad -1 < x < 1.$$

A (3.6) formulabeli konvergencia is majdnem biztos.


Wigner hisztogram

2. Feladatok

(i) Van-e olyan többdimenziós normális eloszlású vektorváltozó, amely komponensei nem függetlenek, de páronként korrelálatlanok?

Tipp:

Válasz: Nincs.

(ii) Igaz-e, hogy ha $Y_1, \dots, Y_m$ független normális eloszlásúak, akkor együttes eloszlásuk m-dimenziós normális?

Tipp:

Válasz: Igaz.

(iii) Adjunk olyan (legalább 3 dimenziós) véletlen vektorváltozót, amely komponensei 1-dimenziós normális eloszlásúak, ő maga nem többdimenziós (és nem is elfajult többdimenziós) normális eloszlású!

Tipp: Lásd a 3.2 ábrát!

Válasz:

$$f(x_1, \dots, x_n) = \begin{cases} c\phi(x_1)\dots\phi(x_n), & \text{ha} \quad x_1 \dots x_n > 0 \\ (1-c)\phi(x_1)\dots\phi(x_n), & \text{ha} \quad xy \leq 0, \end{cases}$$

ahol $0 < c \leq 1$ és $\phi(x)$ a standard normális eloszlás sűrűségfüggvénye.

(iiii)

Legyen $\mathbf{Y} \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$, ahol $\mathbf{C}$ pozitív definit, $\mathbf{B}$ pedig egy $d \times d$-s nemszinguláris mátrix. Milyen eloszlású $\mathbf{X} = \mathbf{BY}$?

Tipp: Az $\mathbf{X}$ véletlen vektor várható értéke $\mathbf{Bm}$, ennek ismerteben feltehető, hogy a szóban forgó véletlen vektorok várható értéke a $\mathbf{0}$vektor. $\mathbf{D}$kovarianciamátrixát pedig a $\mathbf{D} = \mathbb{E}(\mathbf{X}\mathbf{X}^\top) = \mathbb{E}(\mathbf{B}\mathbf{Y}\mathbf{B}\mathbf{Y}^\top)$ képlet alapján számíthatjuk ki.

Válasz: $\mathbf{X} \sim \mathcal{N}_d(\mathbf{Bm}, \mathbf{BCB}^\top)$.

(iiiii) Legyen $\mathbf{X} \sim \mathcal{N}_2(\mathbf{m}, \mathbf{C})$.

(a) Adjuk meg $\mathbf{X}$ komponenseinek tetszőleges $aX_1 + bX_2$ lineáris kombinációjának eloszlását!

(b) Adjuk meg $\mathbf{X}$ komponenseinek korrelációs mátrixát!

(c) Adjuk meg annak a lineáris transzformációnak a mátrixát, amely $\mathbf{X}$ véletlen vektort a 2-dimenziós standard normális eloszlásúba viszi át. Egyértelmű-e ez a mátrix?

Tipp: Jelölje c_{11}, c_{12}, c_{22} a $\mathbf{C}$ mátrix független elemeit.

(a) $\mathrm{D}^2(aX_1 + bX_2) = \mathrm{Cov}(aX_1 + bX_2 aX_1 + bX_2)$, használjuk a definíciót és a várható érték tulajdonságait!

(b) Normáljuk alklamasan a $\mathbf{C}$ mátrixot.

(c) Tetszőleges olyan $\mathbf{A}$ mátrix, amelyre $\mathbf{ACA}^\top = \mathbf{I}_2$.

Válasz:

(a) $\mathcal{N}(am_1 + bM_2, a^2c_{11} + 2abc_{12} + b^2c_{22}, a^2c_{11} + 2abc_{12} + b^2c_{22})$

(b) a korrelacios mátrix főatlójában 1-ek állnak, az r_{12} korrelációs együttható pedig $r_{1,2} = \frac{c_{12}}{\sqrt{c_{11}}\sqrt{c_{22}}}$

(c) Az $\mathbf{A} = \mathbf{C}^{-1/2}$ például jó választás, egy 2×2 pozitív definit mátrixnak általában 4 különböző négyzetgyöke van, és ezzel a lehetséges mátrixok köre még nem merült ki, mert ha $\mathbf{D}$ alkalmas mátrix, $\mathbf{V}$ pedig ortonormált, akkor $\mathbf{DV}$ is alkalmas mátrix.

(iiiiii) Legyenek $\mathbf{X}_i \sim \mathcal{N}_d(\mathbf{m}_i, \mathbf{C}_i),\ i = 1, \dots, n$ független véletlen vektorok. Adjuk meg $\sum_{i=1}^n \mathbf{X}_i$ eloszlását!

Tipp: Analóg a független skalár $\mathcal{N}(m_i, \sigma_i^2)$

k esetével.

Válasz:

$$\mathcal{N}_d(\sum_{i=1}^{n} \mathbf{m}_i, sum_{i=1}^{n}\mathbf{C}_i)$$

(iiiiiii) Legyen $\mathbf{X}$ egy d dimenziós ún. szimmetrikus normális eloszlású vektor, azaz komponensei azonos eloszlásúak és bármely két komponens kovarianciája ugyanakkora.

(a) Határozzuk meg a korrelációs mátrix spektrálfelbontását!

(b) Határozzuk meg $\mathbf{C}^{-1}$-et, ahol $\mathbf{C}$ a kovarianciamátrix!

(c) Adjuk meg annak a lineáris transzformációnak a mátrixát, amely $\mathbf{X}$ véletlen vektort a d-dimenziós standard normális eloszlásúba viszi át.

(d) Mutassuk meg, hogy bármely két komponens korrelációja nagyobb mint $(1-d)^{-1}$.

Tipp: Jelölje $\mathbf{R}$ a korrelációs mátrixot, ami

$$\mathbf{R}\begin{pmatrix} 1 & \rho\dots & \rho \\ \rho & 1\dots & \rho \\ \vdots & \vdots & \vdots \\ \rho & \rho\dots & 1 \end{pmatrix}$$

alakú, ahol $\rho \in [0,1]$.

Ezen speciális alak miatt $\mathbf{C} = \sigma^2\mathbf{R}$.

(a) Az

$$\mathbf{R} - (1-\rho)\mathbf{I}_d = \begin{pmatrix} \rho & \rho\dots & \rho \\ \rho & \rho\dots & \rho \\ \vdots & \vdots & \vdots \\ \rho & \rho\dots & \rho \end{pmatrix}$$

mátrix 1-rangú, és egyetlen nem 0 sajatértéke $d\rho$. Ismeretes, hogy ha egy $\mathbf{A}d \times d$-s mátrix sajátértékei $\lambda_1, \dots, \lambda_d$, akkor $\mathbf{A} + c\mathbf{I}_d$ sajátértékei $\lambda_1 + c, \dots, \lambda_d + c$ (spektrál-leképezés tétel). Ennek alapján $\mathbf{R}$, és igy $\mathbf{C}$ spektruma meghatározható Az utolsó $d-1(\lambda_2, \dots \lambda_d)$ sajátérték egyenlő, míg λ_1 különbözik tőlük. A λ_1-hez tartozó $\mathbf{u}_1$ sajátvektor koordinátái egyenlők, tehát normálva $\mathbf{u}_1 = (\frac{1}{\sqrt{d}}, \dots, \frac{1}{\sqrt{d}})^\top$ Az $\mathbf{R}$ többi sajátvektorai tetszőleges $\mathbf{u}_1$-re és egymásra ortogonális oszlopvektorok. Ilyen sokféle van, különösebb számolás nélkül meghatározhatók azok amelyeknek 1 eleme negatív, a fölötte levő elemek 1-ek, az alatta levők 0-k.

(b) $\mathbf{C}^{-1} = \sigma^{-2}\mathbf{R}^{-1}$.

Ha ismerjük azt az $\mathbf{U}$ ortonormált matrixot, amelynek oszlopai az $\mathbf{u}_1, \dots, \mathbf{u}_d$ sajátvektorok, és $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, akkor a tétel miatt $\mathbf{R} = \mathbf{U}\Lambda\mathbf{U}^\top$, ezért $\mathbf{C}^{-1} = \sigma^{-1}UU\Lambda^{-1}\mathcal{U}^\top$.

(c) A a(c) ponthoz hasonlóan $\mathbf{C}^{-1/2} = \sigma^{-1/2}UU\Lambda^{-1/2}\mathcal{U}^\top$.

(d) Vizsgáljuk meg az (a) pontban kapott sajatértékeket. Mivel $\mathbf{R}$ szükségképpen nemnegatív definit, és a $\lambda_2 = \lambda_3, \dots = \lambda_d = 1-\rho$ sajátértékek nemnegatívak, a $\lambda_1 > 0$ feltételnek kell teljesülnie.

Válasz:

(a) Az $\mathbf{R}$ korrelációs mátrix sajátértékei $\lambda_1 = 1 + (d-1)\rho, \lambda_2 = \lambda_3, \dots = \lambda_d = 1-\rho.$ Itt $d = 4$ -re megmutatjuk $\mathbf{u}2$, $\mathbf{u}_3$ és $\mathbf{u}_4$ konstrukcióját, amiből az általános eset már könnyen leolvasható.

$$\mathbf{U} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{6} & \frac{\sqrt{12}}{12} \\ \frac{1}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{6} & \frac{\sqrt{12}}{12} \\ \frac{1}{2} & 0 & -\frac{\sqrt{6}}{3} & \frac{\sqrt{12}}{12} \\ \frac{1}{2} & 0 & 0 & \frac{-\sqrt{12}}{4} \end{pmatrix}$$

89898

(b) Az Útmutató és (a) pont alapján nyilvánvaló.

(c) Az Útmutató és (a) pont alapján nyilvánvaló.

(d) Az Útmutató és λ_1 értéke alapján nyilvánvaló

(iiiiiiii) * Legyen A és B két $n \times n$-es pozitív definit mátrix. Mutassuk meg, hogy elemenkénti szorzatuk is pozitív definit!

Tipp: Jelölje $\mathbf{A} = \{a_{ij}\} \quad i = 1, \ldots, n\, j = 1, \ldots, n \qquad \mathbf{B} = \{b_{ij}\} \quad i = 1, \ldots, n\, j = 1, \ldots, n$ és $\mathbf{C} = \{c_{ij} = a_{ij}b_{ij}\} \quad i = 1, \ldots, n\, j = 1, \ldots, n$ A feladatban szereplo mátrixokat; $\mathbf{A}$ és $\mathbf{B}$ pozitiv definitása miatt léteznek $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, A)$ és $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, B)$ véletlen vektorok. Tegyük fel, hogy függetlenek. Ekkor a (NEM GAUSS) $\mathbf{Z} = (z_1 = x_1 y_1, \ldots, z_n = x_n y_n)^\top$ veletlen vektor kovarianciamátrixa éppen $\mathbf{C}$.

Válasz: Mivel minden kovarianciamátrix nem negatív definit, és $\mathbf{Z}$ koordinatái lineárisan függetlenek, $\mathbf{C}$ pozitív definit.

A feladtra van tisztán algebrai bizonyítás is: tekintsük az $\mathbf{A} \otimes \mathbf{B} n^2 \times n^2$-es tenzorszorzat mátrixot, ami szintén pozitív definit, és található olyan invariáns altere amiben éppen $\mathbf{C}$ által definiált operátor hat.

(iiiiiiiii) Igaz-e, hogy egy d-dimenziós normális eloszlású vektorváltozó komponensei közül $(d > k)$-t tetszőlegesen kiválasztva azok együttes eloszlása k-dimenziós normális?

Tipp: Próbáljuk felírni a definícióban szereplő $\mathbf{A}$ mátrixot. Feltehető, hogy a definícióban szereplő $\mathbf{A}$ alsó trianguláris, a szimmetria miatt feltehető, hogy az első k komponenst választottuk.

Válasz: Igaz.

(iiiiiiiiii) Igaz-e, hogy $(X_1, X_2) \sim \mathcal{N}_2(\mathbf{0}, \mathbf{C}_d)$ esetén $X_1^2/c_{1,1} + X_2^2/c_{2,2}$ pontosan akkor $\chi^2(2)$ eloszlású, ha X_1 és X_2 korrelálatlanok?

Tipp: Vegyük észre, hogy X_1 és X_2 együttesen Gauss-eloszlású valószínűségi változók pontosan akkor függetlenek, ha korrelálatlanok. Hasonlóan, vegyük észre, hogy két $\mathcal{N}(0,1)$ valószínűségi változó négyzeteinek összege pontosan akkor $\chi^2(2)$ eloszlású, ha függetlenek.

Válasz: Igaz.

(iiiiiiiiiii) Legyen $\mathbf{Y} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$, továbbá $\mathbf{A}$ egy $d \times d$-s szimmetrikus r rangú mátrix. Igaz-e, hogy $\mathbf{Y}^\top \mathbf{A} \mathbf{Y} \sim \chi^2(r)$ pontosan akkor teljesül, ha $\mathbf{A}\mathbf{A} = \mathbf{A}$?

Tipp: Az $\mathbf{A}\mathbf{A} = \mathbf{A}, \quad \mathbf{A} = \mathbf{A}^\top, \quad \mathrm{rang}(\mathbf{A}) = r$ feltétel éppen azt jelenti, hogy $\mathbf{A}$ egy r dimenziós altérre való vetítés mátrixa.

Válasz: Igaz, mivel $\mathbf{A}\mathbf{Y}$ kovariancimátrixa $\mathbf{I}_r$, ezért $\mathbf{Y}^\top \mathbf{A} \mathbf{Y} = \mathbf{Y}^\top \mathbf{A} \mathbf{A} \mathbf{Y} r$ darab független standard normális eloszlású valószínűségi változó négyzetének összege.

(iiiiiiiiiii) Tekintsük az $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ mátrixot, amely oszlopvektorai $\mathbf{X}_i \sim \mathcal{N}_d(\mathbf{0}, \mathbf{C}), i = 1, \ldots, n$ független azonos eloszlású változók, valamint a $\mathbf{W} = \mathbf{X}\mathbf{X}^\top$ Wishart-mátrixot!

(a) Milyen eloszlású $\mathbf{W}^\top$?

(b) Hogy változik meg $\mathbf{W}$, ha $\mathbf{X}$ két oszlopát felcseréljük?

(c) Hogy változik meg $\mathbf{W}$, ha $\mathbf{X}$ két sorát felcseréljük?

(d) Adjunk meg $\mathbf{W}$ várható értékét!

(e) Milyen eloszlású $\mathbf{W}k$-adik főminora?

Tipp: Vegyük észre, hogy $\mathbf{W}$ szimmetrikus. Figyeljük meg a $\mathbf{W}$ definícióját.

Válasz:

(a) $\mathbf{W} = \mathbf{W}^\top$ tehát $\mathbf{W}^\top \sim \mathcal{W}_d(n, \mathbf{C})$

(b) $\mathbf{W}$ nem változik.

(c) Tegyük fel hogy az i-edik és a j-edik sort cseréltük fel. Ekkor $\mathbf{W}$-ben a w_{ii}-t és a w_{jj}-t tartalmazó oszlopok es sorok felcserélődnek.

(d) Ha $n = 1 \mathbb{E}(\mathbf{W}) = \mathbf{C}$, tehát $\mathbb{E}(\mathbf{W}) = n\mathbf{C}$.

(e) $\mathcal{W}_k(n, \mathbf{C}')$, ahol $\mathbf{C}'$ a $\mathbf{C}$ mátrix k-adik főminora.

(iiiiiiiiiiii) Legyenek $\mathbf{W}_i \sim \mathcal{W}_d(n_i, \mathbf{C}),\ i = 1, \dots, k$ független Wishart-mártixok. Milyen eloszlású $\sum_{i=1}^k \mathbf{W}_i$?

Tipp: Emlékezzünk arra, hogy a Wishart-eloszlás a χ^2-eloszlás (l.) analogonja.

Válasz: Legyen $n = n_1 + \dots + n_k \sum_{i=1}^k \mathbf{W}_i \sim \mathcal{W}_d(n, \mathbf{C})$.

(iiiiiiiiiiiii) Legyen $\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{C})$ és $a \in \mathbb{R}^+$. Milyen eloszlású $a\mathbf{W}$?

Tipp: Emlékezzünk arra, hogy a Wishart-eloszlás a chi^2-eloszlás analogonja.

Válasz: $a\mathbf{W} \sim \mathcal{W}_d(n, a\mathbf{C})$

(iiiiiiiiiiiiii) Legyen $\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{C})$ és $\mathbf{B}$ egy $d \times d$-s nemszinguláris mátrix. Milyen eloszlású $\mathbf{BWB}^\top$?

Tipp: Számoljuk ki a $\mathbf{BX}$ kovarianciamátrixát, ahol $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{C})$. Ha $\mathbf{W} = \mathbf{XX}^\top$ mivel egyenlő a $\mathbf{BXBX}^\top$?

Válasz: $\mathbf{BWB}^\top \sim \mathcal{W}_d(n, \mathbf{BCB}^\top)$.

(iiiiiiiiiiiiiii) Legyen $\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{I})$.

(a) Milyen eloszlásúak $\mathbf{W}$ diagonális elemei?

(b) Milyen eloszlású $\mathrm{tr}\mathbf{W}$?

(c) Igazoljuk, hogy $\mathbf{W}$ nemdiagonális elemei előállnak két független $\chi^2(n)$ eloszlású változó különbségének konstansszorosaként!

Tipp:

(a) Alkalmazzuk a definíciót.

(b) Alkalmazzuk a definíciót, és keressük meg a χ^2 eloszlás definióját ben.

(c) Alkalmazzuk az $(a+b)(a-b) = a^2 - b^2,\ \ (a+b)^2 = a^2 + 2ab + b^2,\ \ (a-b)^2 = a^2 - 2ab + b^2$ azonosságokat.

Válasz:

(a) $\chi^2(n)$

(b) $\chi^2(nd)$

(c) Ha n $\mathcal{N}(0,2)$ Y független standard normális eolszlású valószínűségi változók, akkor $X+Y$ és $X-Y$ $(X+Y)^2/4-(X-Y)^2/$ lószínűségi változók (az előbb idézett azonosság mia χ^2 Továbbá k XY ggetlen valószínűségi változó különbsége melyeknek 2-s n resei el XY ásúak. Ugyanakkor ez a különbség . A standard Wishart mátrix diagonálison kívüli elemei független alakú valószínűségi változó összege.

3. Tesztek

(i) $X_1,\ldots,X_n$ egydimenziós normális eloszlásúak. Melyik állítás igaz?

(a) Együttes eloszlásuk csak akkor többdimenziós normális, ha függetlenek.

(b) Ha függetlenek, akkor együttes eloszlásuk többdimenziós normális.

(c) Együttes eloszlásuk csak akkor többdimenziós normális, ha nem függetlenek.

(d) Ha nem függetlenek, akkor együttes eloszlásuk többdimenziós normális.

Válasz: (b)

(ii) Egy többdimenziós normális eloszlású változó komponensei standard normális eloszlásúak. Igaz-e, hogy együttesen is standard normális eloszlású?

(a) Igen, mert ez a definíció.

(b) Igen, mert a többdimenziós standard normális eloszlású változó lineáris transzformációjaként kapjuk, az pedig egyértelmű.

(c) Igen, mert a függetlenségből következik a korrelálatlanság.

(d) Nem, csak ha a komponensek korrelálatlanok.

Válasz: (d)

(iii) Legyenek $\mathbf{X}_1,\ldots,\mathbf{X}_n \sim N_d(\mathbf{0},\mathbf{C})$ függetlenek. Milyen eloszlású $\sum_{i=1}^n \mathbf{X}_1+\ldots+\mathbf{X}_n$?

(a) $N_d(\mathbf{0},\mathbf{C})$

(b) $N_d(\mathbf{0},n\mathbf{C})$

(c) $N_d(\mathbf{0},n^2\mathbf{C})$

(d) $\mathcal{W}_d(n,\mathbf{C})$

Válasz: (b)

(iiii) Legyenek $\mathbf{X}_1,\ldots,\mathbf{X}_n \sim N_m(\mathbf{m},\mathbf{I})$ függetlenek. Milyen eloszlású $\sum_{k=1}^n(\mathbf{X}_k-\mathbf{m})(\mathbf{X}_k-\mathbf{m})^\top$?

(a) $\chi^2(n)$

(b) $\chi^2(nd)$

(c) $\mathcal{W}_m(n,I)$

(d) $\mathcal{W}_n(m,I)$

Válasz: (c)

(iiiii) Valójában hány dimenziós változó egy $\mathcal{W}_d(n,\mathbf{C})$ eloszlású Wishart-mátrix?

(a) d^2

(b) $d(d+1)/2$

(c) nd

(d) $(nd+1)/2$

Válasz: (b)

(iiiiii) Milyen eloszlásúak az n darab d dimenziós standard normális eloszlású változó segítségével kapott Wishart-mátrix főátlójának elemei?

(a) Standard normális

(b) $\chi^2(1)$

(c) $\chi^2(d)$

(d) $\chi^2(n)$

Válasz: (d)

4. fejezet - Paraméterbecslés és hiptézisvizsgálat többdimenziós normális modellben

1. Elméleti háttér

1.1. Paraméterbecslés többdimenziós normális modellben

Ebben a paragrafusban csak azokra a fogalmakra és tételekre térünk ki, amelyek természetüknél fogva lényegesen különböznek azok egydimenziós változataiktól. *Hatásosság:* A torzítatlan becslések között keressük a leghatásosabbat. Mivel a több paraméter esetén a becslésk szórásnégyzetei helyett azok kovarianciamátrixait kell összehasonlítanunk, a hatásosság mérésére egy erősebb fogalmat vezetünk be.

4.1.1.1. Definíció. *A* $\underline{\theta} \in \Theta$ *paraméter* $\mathbf{T}_1$ *becslése legalább olyan hatásos, mint* $\mathbf{T}_2$ *becslése, ha*

$$\mathbb{D}^2_{\underline{\theta}}(\mathbf{T}_1) \leq \mathbb{D}^2_{\underline{\theta}}(\mathbf{T}_2),$$

ahol a mátrixok közötti $\mathbf{A} \leq \mathbf{B}$ *rendezés úgy értendő, hogy* $\mathbf{B} - \mathbf{A}$ *pozitív szemidefinit.*

Ilyen értelemben alkalmazza a rendezést a Cramér- Rao egyenlőtlenség több paraméterre vonatkozó alakja:

4.1.1.2. Tétel. *A Cramér- Rao egyenlőtlenség többváltozós alakja (bizonyos - itt teljesülő - regularitási feltételek esetén) alsó korlátot ad a torzítatlan becslések szórásmátrixára:*

$$\mathbb{D}^2_{\underline{\theta}}(\mathbf{T}) \geq \frac{1}{n}\mathbf{I}_1^{-1}(\underline{\theta}) = I_n^{-1}(\underline{\theta}), \qquad \underline{\theta} \in \Theta$$

$\mathbf{I}_1(\underline{\theta})$ *jelöli az ún. Fisher-féle információs mátrixot, amit 1-elemű mintából számolhatunk:*

$$\mathbf{I}_1(\underline{\theta}) = \mathbb{E}_{\underline{\theta}}\left(\frac{\partial}{\partial \underline{\theta}} \ln f_{\underline{\theta}}(\mathbf{X}_1)\right)\left(\frac{\partial}{\partial \underline{\theta}} \ln f_{\underline{\theta}}(\mathbf{X}_1)\right)^T = \mathbb{D}^2_{\underline{\theta}}\left(\frac{\partial}{\partial \underline{\theta}} \ln f_{\underline{\theta}}(\mathbf{X}_1)\right),$$

Megjegyezzük, hogy többdimenziós normális eloszlásnál egyenlőség az $(\bar{\mathbf{X}}, mxS/(n-1))$ párra nem érhető el.

1.1.1. A többdimenziós normális eloszlás paramétereinek maximum-likelihood becslése.

Mielőtt hozzáfognánk ennek a feladatnak a megoldásához, felidézzük a Steiner-egyenlőséget többdimenziós változatát.

4.1.1.1.1. Lemma. *(Steiner-egyenlőség). Legyenek* $\mathbf{x}_1, \ldots \mathbf{x}_n \in \mathbb{R}^p$ *vektorok, , továbbá legyen* $\bar{\mathbf{x}}$ *az átlaguk és* $\mathbf{v} \in \mathbb{R}^p$ *egy tetszőleges vektor. Ekkor*

$$\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{v})(\mathbf{x}_k - \mathbf{v})^\top = \sum_{k=1}^{n}(\mathbf{X}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top + (\bar{\mathbf{x}} - \mathbf{v})(\bar{\mathbf{x}} - \mathbf{v})^\top. \quad (1)$$

Speciálisan, ha $\mathbf{v} = \mathbf{0}$

$$\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top = \sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^\top - n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top.$$

Legyen $\mathbf{X}_1, \ldots, \mathbf{X}_n$ független elemű minta az $\mathbf{X} \in \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ véletlen vektorra, tegyük fel, hogy $n > p$. $\mathbf{C}$ mintaelemek alapján szeretnénk becslést adni az ismeretlen $\mathbf{m}$ várható érték vektorra és a kovarianciamátrixra, melyről feltesszük, hogy pozitív definit. Ehhez a maximum likelihood módszert használjuk, azaz a mintaelemek együttes sűrűségfüggvényével definiált likelihood-függvényt maximalizáljuk a két ismeretlen paraméterben. A mintaelemek függetlensége következtében az együttes sűrűségfüggvény a külön-külön vett sűrűségfüggvények szorzata, melyek mindegyike (a mintaelemek azonos eloszlása miatt) az (3.2) alakban írható (csak az argumentumokba most a mintaelemeket írjuk):

$$L_{\mathbf{m},\mathbf{C}}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \frac{1}{(2\pi)^{np/2}|\mathbf{C}|^{n/2}} e^{-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{X}_k-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{X}_k-\mathbf{m})}. \qquad (1)$$

???

Vegyük észre exponensbeli

$$\sum_{k=1}^{n}(\mathbf{X}_k - \mathbf{m})^T\mathbf{C}^{-1}(\mathbf{X}_k - \mathbf{m})$$

kvadratikus alak tulajdonképpen egy 1×1-es mátrix nyoma (trace-e), ami a trace függvény ciklikus permutációkkal szembeni invarianciája miatt

$$\mathrm{tr}\mathbf{C}^{-1}(\mathbf{X}_k - \mathbf{m})(\mathbf{X}_k - \mathbf{m})^T \qquad (1)$$

???

alakban is írható (erről közvetlen számolással is meggyőződhetünk). A formulák kezelése szempontjából ez az alak gyakran előnyösebb, mint a kvadratikus forma írásmód.

Az előző rész jelöléseit használjuk:

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{X}_k$$

jelöli a mintaátlagot és

$$\mathbf{S} = \sum_{k=1}^{n}(\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{X}_k - \bar{\mathbf{X}})^T$$

az empirikus kovarianciamátrix n-szeresét. A likelihood-függvényt most a (4.3) formula és a (4.1) többdimenziós Steiner-egyenlőseg segítségével úgy alakítjuk át, hogy benne ezek a statisztikák jelenjenek meg:

$$L(\mathbf{X}_1, \ldots, \mathbf{X}_n; \mathbf{m}, \mathbf{C}) = \frac{1}{(2\pi)^{np/2}|\mathbf{C}|^{n/2}} e^{-\frac{1}{2}\mathrm{tr}\mathbf{C}^{-1}\mathbf{S}} \cdot e^{-\frac{1}{2}n(\bar{\mathbf{X}}-\mathbf{m})^T\mathbf{C}^{-1}(\bar{\mathbf{X}}-\mathbf{m})}. \quad (1)$$

???A fenti (4.4) függvényt $\mathbf{m}$-ben és $\mathbf{C}$-ben kell maximalizálnunk, hogy megkapjuk $\hat{\mathbf{m}}$ és $\hat{\mathbf{C}}$ becsléseket. A (4.4) függvény akkor lesz $\mathbf{m}$-ben maximális, ha a kitevőben lévő kvadratikus alak értéke 0, ezért

$$\hat{\mathbf{m}} = \overline{\mathbf{X}}.$$

Mivel ez a szélsőérték független a $\mathbf{C}$ paramétertől a (4.4) függvényt ugy maximalizálhatjuk $\mathbf{C}$ szerint (valojában $\mathbf{C}^{-1}$ szerint) $\hat{\mathbf{m}} = \overline{\mathbf{X}}$-szel helyettesítjük. A további számolás a fejezetben ismertetett $\frac{\partial|\mathbf{A}|}{\partial\mathbf{A}} = \mathrm{adj}(\mathbf{A}^\top)$ képlet alkalmazásával végezhető el, ezt nem részletezzük, csak a végeredményt közöljük:

$$\hat{\mathbf{C}} = \frac{\mathbf{S}}{n}.$$

1.2. Hipotézisvizsgálat többdimenziós normális modellben

Az egyváltozós esethez hasonlóan hipotéziseket is vizsgálhatunk a várható érték vektorra és a kovarianciamátrixra vonatkozóan. Ehhez megismételjük likelihood hányados próba, és bevezetjük a Hotelling T^2-eloszlás definícióját.

4.1.2.1. Definíció. *Legyen $\underline{\theta}$ a $\mathbf{X}_1, \ldots, \mathbf{X}_n$ iségfüggvényű eloszlás ismeretlen μ H_0 am H_1 rvektora, $\underline{\theta} \in \Theta$ ($\Theta \subset \mathbb{R}^k$ többdimenziós tartomány). Az minta alapján dönteni szeretnénk a és hipotézisek között:*

$$H_0 : \underline{\theta} \in \Theta_0 \quad \text{vers.} \quad H_1 : \underline{\theta} \in \Theta_1,$$

ahol $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$, és a $\dim(\Theta_0) = r$, $\dim(\Theta) = k$ jelöléssel $r < k$ teljesül. Az n-elemű minta alapján konstruálandó próbastatisztika:

$$\lambda_n(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \frac{L_0^*}{L_1^*} = \frac{\sup\limits_{\underline{\theta} \in \Theta_0} L_{\underline{\theta}}(\mathbf{X}_1, \ldots, \mathbf{X}_n)}{\sup\limits_{\underline{\theta} \in \Theta} L_{\underline{\theta}}(\mathbf{X}_1, \ldots, \mathbf{X}_n)}.$$

Amennyiben ismerjük a $\lambda_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ próbastatisztika eloszlását H_0 fennállása esetén, adott $1 - \varepsilon$ szignifikanciaszinthez (ε ``kicsi") megkonstruáljuk a mintatér részét képező

$$\mathcal{X}_k = \{(\mathbf{x}_1, \ldots, \mathbf{x}_n) : \lambda_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) \leq \lambda_\varepsilon\}$$

kritikus tartományt, ahol a λ_ε kritikus értéket úgy határozzuk meg, hogy a próba terjedelme ε legyen, azaz $\sup\limits_{\underline{\theta} \in \Theta_0} \mathbb{P}_{\underline{\theta}}((\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathcal{X}_k) = \varepsilon$. Ezután, ha mintánk a kritikus tartományba esik, elutasítjuk, különben pedig elfogadjuk a nullhipotézist.

4.1.2.2. Definíció. *Legyenek a $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{I})$ - $\mathbf{W}$ pozitív definit (ez 1 valószínűséggel teljesül, ha $n > p$) - és a $\mathbf{X} :\sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ valószínűségi változók függetlenek. Akkor a*

$$T^2 = n\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$$

összefüggéssel definiált T^2 valószínűségi változót Hotelling-féleT^2-eloszlás*ának nevezzük n, p paraméterekkel. A továbbiakban az n paraméterre, mint szabadságfokra hivatkozunk.*

Megjegyezzük, hogy a Hotelling-féle T^2-eloszlás a Student-féle t-eloszlás többdimenziós általánosítása: a $p = 1$, $\mathbf{C} = 1$ esetben $T^2 \equiv t^2/n$.

4.1.2.3. Állítás. *A $\mathbf{W} \sim \mathcal{W}_p(n, \mathbf{C})$ és $\mathbf{X} :\sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$ esetben*

$$T^2 = n(\mathbf{X} - \mathbf{m})\mathbf{W}^{-1}(\mathbf{X} - \mathbf{m})^\top$$

valószínűségi változó szintén T^2-eloszlású n és p paraméterekkel.

4.1.2.4. Tétel. *Ha a T^2 statisztika Hotelling elosszlású n és p paraméterekkel, akkor*

$$\frac{n - p + 1}{p} \cdot T^2 \sim \mathcal{F}(p, n - p + 1),$$

azaz T^2 megfelelő konstansszorosa Fisher-féle F-eloszlású a zárójelben felsorolt paraméterekkel.

Részleteket ld. [5] 228-230. o. (6.2. Tétel).

2. Feladatok

(i) Igazoljuk a Steiner-egyenlőség következő többdimenziós változatát:

ha $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{v} \in \mathbb{R}^d$, akkor

$$\sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{v})(\mathbf{x}_k - \mathbf{v})^\top = \sum_{k=1}^{n} (\mathbf{x}_k - \overline{\mathbf{x}})(\mathbf{x}_k - \overline{\mathbf{x}})^\top + n(\overline{\mathbf{x}} - \mathbf{v})(\overline{\mathbf{x}} - \mathbf{v})^\top.$$

Tipp:

Válasz:

(ii) Legyen $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$ független minta. Igazoljuk, hogy

$\mathrm{Cov}(\overline{\mathbf{X}}, \mathbf{X}_i - \overline{\mathbf{X}}) = \mathbf{0}$.

Tipp:

Válasz:

(iii) Legyen $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ minta. Adjuk meg az $\mathbf{I}_1$ Fisher-féle információs mátrixot!

Tipp: Alkalmazzuk a többdimenziós Fisher- Cochran-tételbeli definíciót.

Válasz:

$\mathbf{I}_1 = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^4} \end{pmatrix}$

(iiii) Legyen $X_1, \dots, X_n \sim U(a, b)$ független minta. Adjuk meg az $\mathbf{I}_1$ és $\mathbf{I}_n$ Fisher-féle információs mátrixokat!

Tipp: Alkalmazzuk a többdimenziós Fisher- Cochran-tételbeli definíciót.

Válasz:

$\mathbf{I}_1 = \begin{pmatrix} \frac{1}{(b-a)^2} & \frac{1}{(b-a)^2} \\ \frac{1}{(b-a)^2} & \frac{1}{(b-a)^2} \end{pmatrix}$,

$\mathbf{I}_1 = \begin{pmatrix} \frac{n^2}{(b-a)^2} & \frac{n^2}{(b-a)^2} \\ \frac{n^2}{(b-a)^2} & \frac{n^2}{(b-a)^2} \end{pmatrix}$.

(iiiii) $\mathbf{X}_1, \dots, \mathbf{X}_n$ egy $\mathbf{a}$ középpontú b sugarú d-dimenzós gömbben egyenletes eloszlásból vett független minta.

(a) Adjuk meg az $\mathbf{I}_1$ Fisher-féle információs mátrixot!

(b) Adjunk maximum likelihood becslést $\mathbf{a}$-ra $b = 1$ esetben!

(c) Adjunk maximum likelihood becslést $(\mathbf{a}, b)$-re!

Tipp:

(a) Vegyük észre, hogy a sűrűségfüggvény értéke nem függ az $\mathbf{a}$ vektortól abban a tartományban, ahol ez az érték nem 0. Ugyanezt az elvet alkalmaztuk pl. $[0, \theta]$ intervallumon egyenletes minta Fisher-féle információjának kiszámításákor, és az előző feladatban is. Az előző feladat azért is érdekes, mert $d = 1$-re alkalmasan átparaméterezve $\left(c = \frac{a+b}{2} \text{ és } r = \frac{b-a}{2}\right)$ ugyanez a helyzet.

(b) Minden olyan $\mathbf{a}$ vektor M-L becslés lesz, amely körüli 1 sugarú gömb tartalmazza a mintát.

(c) $\mathbf{a}$ M-L becslése az a vektor lesz, amely körüli - a teljes mintát tartalmazó - körlap sugara minimális, míg b M-L becslése ez a minimális sugár

Válasz:

(a) Figyelembevéve, hogy a d-dimenziós gömb térfogata $C_d b^d$, ahol C_d egy a dimenziótól függő kosntans - ami a számolás során kiesik:

$\begin{pmatrix} \frac{d^2}{b^2} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$.

(b) Az Útmutató alapján pl. a síkon viszonylag egyszerű algoritmussal a mintát egy olyan négyzettel burkoljuk, amely egyik élének iránya tetszőleges, ennek középpontja alkalmas becslés.

(c) Nem tudok rá gyors algoritmust.

(iiiiii) 49 idős embert az orvos két csoportba sorolt aszerint, hogy van-e szenilis faktor a viselkedésükben (I. csoport) vagy sem (II. csoport). Ezután elvégeztettek velük 4 pszichológiai tesztet (1. információ, 2. hasonlóság, 3. aritmetika, 4. képfelismerés), melyekre kapott átlagpontszámok az alábbi táblázatban láthatók:

	I. (n=37)	II. (m=12)
1.	12,57	8,75
2.	9,57	5,33
3.	11,49	8,50
4.	7,97	4,75

Vizsgálja meg, 95%-os szignifikanciaszinten elfogadható-e az a nullhipotézis, hogy a két csoport várhatóan nem különbözik szignifikánsan a teszteredmények alapján. Feltesszük, hogy az egyes emberek teszteredményei 4-dimenziós normális eloszlást követnek ismeretlen (közös) kovarianciamátrixszal. Az egyesített (49) elemő mintából számolt $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$ mátrix inverze:

$$\mathbf{S}^{-1} = \begin{pmatrix} 0{,}0052 & -0{,}0028 & -0{,}0012 & -0{,}0012 \\ -0{,}0028 & 0{,}0038 & -0{,}0008 & -0{,}0002 \\ -0{,}0012 & -0{,}0008 & 0{,}0030 & -0{,}0004 \\ -0{,}0012 & -0{,}0002 & -0{,}0004 & 0{,}0042 \end{pmatrix}.$$

Tipp:

Válasz:

(iiiiiii) Legyen $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$ független minta, ahol $\mathbf{C}$ ismert.

(a) Adjuk meg az $\mathbf{I}_1$ Fisher-féle információs mátrixot!

(b) Igazoljuk, hogy $\overline{\mathbf{X}}$ hatásos becslése $\mathbf{m}$-nek! (Használjuk a Cramér-Rao egyenlőtlenség többdimenziós változatát!)

(c) Igazoljuk, hogy a $H_0 : \mathbf{m} = \mathbf{m}_0,\ H_1 : \mathbf{m} \neq \mathbf{m}_0$ hipotézisek vizsgálatára konstruált próba likelihood-hányados teszt!

(d) Igazoljuk, hogy az előző pontbeli teszt az u-próba általánosítása!

Tipp:

Válasz:

(iiiiiiii) 20 fiatal emberre az A, B, C stimuláló szerek hatását vizsgálták a reakcióidő szempontjából (századmásodpercben).

$$\overline{X}_A = 21{,}05 \quad \overline{X}_B = 21{,}65 \quad \overline{X}_C = 28{,}95,$$

$$\mathbf{S} = \begin{pmatrix} 45{,}2 & 43{,}6 & 32{,}6 \\ 43{,}6 & 53{,}2 & 36{,}4 \\ 32{,}6 & 36{,}4 & 49{,}4 \end{pmatrix}.$$

95%-os szignifikanciaszinten vizsgálja meg az egyenlő hatás elvét a $B - A, C - B$ különbségekre! (Feltesszük, hogy a hatások többdimenziós normális eloszlást követnek, és azt teszteljük, hogy a B és A hatás különbsége, valamint a C és B hatás különbsége mint 2-dimenziós normális eloszlású véletlen vektor $\mathbf{0}$ várható érték vektorúnak tekinthető-e.) Megjegyezzük, hogy valójában a három stimulálószer hatása várható értékének egyenlősége itt a nullhipotézis, azonban megfigyeléseink nem független mintákra, hanem ugyanarra a 20 emberre vonatkoznak. Így a javasolt vizsgálat a t-próbánál bevezetett önkontrollos vizsgálat többdimenziós általánosításának tekinthető.

Tipp:

Válasz:

(iiiiiiiii) Legyen $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$ független minta. Vegyük az $(\mathbf{m}, \mathbf{C})$ paraméter $(\hat{\mathbf{m}}, \hat{\mathbf{C}}) = (\overline{\mathbf{X}}, \mathbf{S}/n)$ (maximum likelihood) becsléseit!

(a) Igazoljuk, hogy $(\overline{\mathbf{X}}, \mathbf{S})$ elégséges statisztika $(\mathbf{m}, \mathbf{C})$-re!

(b) Torzítatlan becslése-e $(\overline{\mathbf{X}}, \mathbf{S}/n)$ az $(\mathbf{m}, \mathbf{C})$ paraméternek? Ha nem, korrigáljuk!

(c) Mutassuk meg, hogy a (Hotelling-féle) T^2-próba a t-próba (kétoldali változatának) általánosítása (de az egyoldalinak nem)!

(d) Konstruáljunk likelihood-hányados próbát a $H_0 : \mathbf{C} = \mathbf{C}_0$ hipotézis tesztelésére!

(e) Konstruáljunk ε terjedelmű egyenletesen legerősebb próbát a Neyman-Pearson alaplemma segítségével a $H_0 : (\mathbf{m}, \mathbf{C}) = (\mathbf{m}_0, \mathbf{C}_0)$ vs. $H_1 : (\mathbf{m}, \mathbf{C}) = (\mathbf{m}_1, \mathbf{C}_0)$ egyszerű alternatíva vizsgálatára!

Tipp:

Válasz:

(iiiiiiiiii) Igazoljuk, hogy a (Hotelling-féle) kétmintás T^2-próba likelihood-hányados próba! Igazoljuk, hogy ez a teszt a kétmintás t-próba általánosítása!

Tipp:

Válasz:

(iiiiiiiiii) Legyen $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1} \sim \mathcal{N}_d(\mathbf{m}_1, \mathbf{C}_1)$ és $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2} \sim \mathcal{N}_d(\mathbf{m}_2, \mathbf{C}_2)$ független minták. Konstruáljunk likelihood-hányados próbát a $H_0 : \mathbf{C}_1 = \mathbf{C}_2$, $H_1 : \mathbf{C}_1 \neq \mathbf{C}_2$ hipotézisek vizsgálatára (kétmintás T^2 próba feltételének ellenőrzése)!

Tipp:

Válasz:

(iiiiiiiiiii) Legyen $\mathbf{X}_1, \mathbf{X}_2, \ldots \sim \mathcal{N}_d(\mathbf{m}, \mathbf{C})$ fae. Adjunk a $H_0 : (\mathbf{m}, \mathbf{C}) = (\mathbf{m}_0, \mathbf{C}_0)$ vs. $H_1 : (\mathbf{m}, \mathbf{C}) = (\mathbf{m}_1, \mathbf{C}_0)$ egyszerű alternatíva eldöntésére szekvenciális eljárást (ε_1 elsőfajú és ε_2 másodfajú hibával)! Adjuk meg a várható lépésszámokat!

Tipp:

Válasz:

(iiiiiiiiiiii) Legyen $A_1, \ldots, A_k$ teljes eseményrendszer, $\mathrm{P}(A_i) = p_i$. Legyen $\mathbf{X}$ az eseményrendszer k-dimenziós indikátorváltozója, valamint $\mathbf{p} = (p_1, \ldots, p_k)^\top$. Legyenek $\mathbf{X}_1, \mathbf{X_2} \ldots$ független vektorok, amelyek eloszlása megegyezik $\mathbf{X}$ eloszlásával.

(a) Mutassuk meg, hogy $\sum_{i=1}^{n} \mathbf{X}_i \sim Poly_n(p_1, \ldots, p_k)$.

(b) Adjunk maximum likelihood becslést az első n mintaelem alapján $\mathbf{p}$-re a Lagrange-multiplikátor módszerével!

(c) Adjunk maximum likelihood becslést az első n mintaelem alapján $\mathbf{p}$-re $p_k = 1 - p_1 - \ldots - p_{k-1}$ felhasználásával is!

(d) Adjunk a $H_0 : \mathbf{p} = \mathbf{p}_0$ vs. $H_1 : \mathbf{p} = \mathbf{p}_1$ egyszerű alternatíva eldöntésére szekvenciális eljárást (ε_1 elsőfajú és ε_2 másodfajú hibával)! Adjuk meg a várható lépésszámokat!

Tipp:

Válasz:

3. Tesztek

(i) Tekintsünk egy n elemű $\mathcal{N}_d(\mathbf{m}, \mathbf{C})$ eloszlásból vett mintát (feltesszük, hogy $\mathbf{C}$ invertálható, a több dimenziós Fisher $\mathbf{I}_1$ mátrix a $\mathbf{C}$ mátrix inverze). Milyen becslése a $\mathbf{m}$-nek a maximum likelihood becslés?

(a) Nem torzítatlan, de aszimptotikusan torzítatlan, erősen konzisztens.

(b) Nem torzítatlan, de aszimptotikusan torzítatlan, gyengén sem konzisztens.

(c) Torzítatlan, hatásos, erősen konzisztens.

(d) Torzítatlan, nem hatásos, gyengén sem konzisztens.

Válasz: (c)

(ii) Tekintsünk egy n elemű $N_d(\mathbf{m}, \mathbf{C})$ eloszlásból vett mintát. Milyen becslése a $\mathbf{C}$-nek a maximum likelihood becslés?

(a) Nem torzítatlan, de aszimptotikusan torzítatlan, erősen konzisztens.

(b) Nem torzítatlan, de aszimptotikusan torzítatlan, gyengén sem konzisztens.

(c) Torzítatlan, hatásos, erősen konzisztens.

(d) Torzítatlan, nem hatásos, gyengén sem konzisztens.

Válasz: (a)

(iii) Melyik teszt általánosítása a Hotelling-féle T^2 próba (azaz egy dimenziós esetben melyiket kapjuk)?

(a) u próba

(b) t próba

(c) F próba

(d) χ^2 próba

Válasz: (b)

(iiii) Hogy lehet két (egy- vagy többdimenziós) standard normális eloszlás (amelyek együttesen is normális eloszlásúak) függetlenségének tesztelésére alkalmazni a normális eloszlás kovarianciamátrixára vonatkozó próbát?

(a) Sehogy, mert az a többdimenziós normális eloszlás kovarianciamátrixára vonatkozik, nem függetlenségre.

(b) Ha azonos a dimenziószam, a különbségváltozó kovarianciamátrixát teszteljük, hogy 0-e.

(c) Összefűzött változót teszteljük, kovarianciamátrixa egységmátrix-e.

(d) Külön-külön teszteljük a két változót, kovarianciamátrixa egységmárix-e és megnézzük, a két teszt ugyanazt adta-e eredményül.

Válasz: (c)

5. fejezet - Lineáris módszerek 1.: főkomponensanalízis, faktoranalízis

1. Elméleti háttér

1.1. Főkomponensanalízis

Legyen $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$, és tegyük fel, hogy a $\mathbf{C}$ kovarianciamátrix pozitív definit. A modell a következő: keressük $\mathbf{X}$ előállítását

$$\mathbf{X} = \mathbf{V}\mathbf{Y} + \mathbf{m} \qquad (1)$$

???

alakban, ahol $\mathbf{m} = \mathbb{E}\mathbf{X}$, $\mathbf{V}$ $p \times p$-s ortogonális mátrix (azaz $\mathbf{V}^{-1} = \mathbf{V}^T$), $\mathbf{Y}$ pedig független komponensű, p-dimenziós normális eloszlású véletlen vektor

Vegyük észre, hogy az (5.1) előállítás hasonló a 3. fejezetben tárgyalt (3.1)-beli $\mathbf{X} = \mathbf{A}\mathbf{Y} + \mathbf{m}$

felbontáshoz, de ott $\mathbf{Y}$ p-dimenziós standard normális eloszlású volt, a $p \times p$-s $\mathbf{A}$ mátrix pedig az $\mathbf{A}\mathbf{A}^T = \mathbf{C}$ (nem egyértelmű) felbontásból adódott. Ott $\mathbf{Y}$ komponensei függetlenek és 1 szórásúak voltak, míg a fenti (1.1) előállításban $\mathbf{Y}$ komponenseitől csak a függetlenséget követeljük meg, míg a transzformációs mátrixtól ortogonalitást várunk el. Ez az előállítás már egyértelmű, ha $\mathbf{Y}$ komponenseit varianciáik (szórásnégyzeteik) csökkenő sorredjében rendezzük. (Ha a varianciák között adódnak egyenlőek, akkor nincs egyértelműség, ennek feltételét az alábbi eljárásból olvashatjuk ki.)

Most megadjuk (5.1) a előállítást. Mivel $\mathbf{V}$ invertálható, ezért (5.1) ekvivalens az

$$\mathbf{Y} = \mathbf{V}^{-1}(\mathbf{X} - \mathbf{m}) = \mathbf{V}^T(\mathbf{X} - \mathbf{m})$$

felbontással. Jelölje $\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^T$ az $\mathbf{X}$ véletlen vektor kovarianciamátrixának spektrálfelbontását. Ezzel $\mathbf{Y}$ kovarianciamátrixának diagonálisnak kell lennie. A spektrálfelbontás egyértelműsége értelmében

$$\begin{aligned}\mathbb{E}\mathbf{Y}\mathbf{Y}^T &= \mathbb{E}\left[\mathbf{V}^{-1}(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T\mathbf{V}\right] = \mathbf{V}^{-1}\mathbb{E}\left[(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T\right]\mathbf{V} = \\ &= \mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{V}^{-1}\mathbf{U}\Lambda\mathbf{U}^T\mathbf{V} = (\mathbf{V}^{-1}\mathbf{U})\Lambda(\mathbf{V}^{-1}\mathbf{U})^T\end{aligned}$$

diagonális mátrix fődiagonálisában csökkenő elemekkel akkor és csak akkor, ha $\mathbf{V}^{-1}\mathbf{U} = \mathbf{I}_p$, azaz $\mathbf{V} = \mathbf{U}$. (Itt kihasználtuk, hogy $\mathbf{V}$, $\mathbf{U}$, következésképpen $\mathbf{V}^{-1}\mathbf{U}$ is ortogonális mátrix.) Megjegyezzük, hogy többszörös multiplicitású sajátértékek esetén az $\mathbf{U}$ mátrix megfelelő oszlopai sem egyértelműek (l.). Így

$$\mathbf{X} = \mathbf{U}\mathbf{Z} + \mathbf{m}$$

lesz a kívánt felbontás, ahol $\mathbf{Z}$ jelöli a $\mathbf{V} = \mathbf{U}$ választás melletti $\mathbf{Y}$-t, azaz

$$\mathbf{Z} = \mathbf{U}^{-1}(\mathbf{X} - \mathbf{m}) = \mathbf{U}^T(\mathbf{X} - \mathbf{m}).$$

Ezt a $\mathbf{Z}$-t az $\mathbf{X}$ véletlen vektor *főkomponensvektor*ának, komponenseit pedig *főkomponensek*nek nevezzük. Vegyük észre, hogy a k-adik főkomponens az $\mathbf{X} - \mathbf{m}$ változó komponenseinek az $\mathbf{u}_k$ vektor koordinátáival vett lineáris kombinációja:

$$Z_k = \mathbf{u}_k^T(\mathbf{X} - \mathbf{m}) \qquad (k = 1, \ldots, p),$$

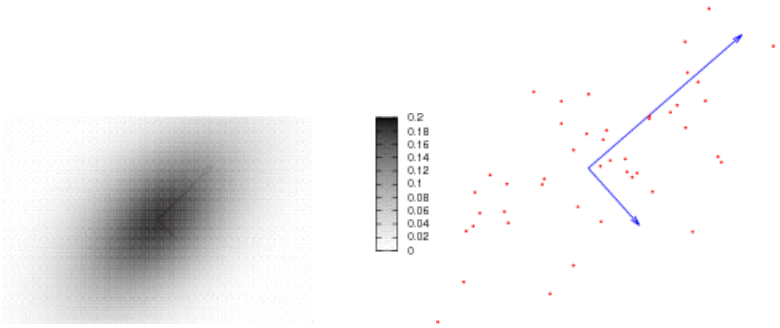
ahol $\mathbf{u}_k$ a $\mathbf{C}$ mátrix λ_k sajátértékéhez tartozó normált sajátvektora ($\mathbf{U}k$-adik oszlopa), $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

Az $\mathbf{X}$ véletlen vektor fenti felbontása eleget tesz az alább ismertetendő optimalitási kritériumnak (a főkomponenseket ezzel is be lehetne vezetni).

5.1.1.1. Tétel. *Az első főkomponens, Z_1 szórása maximális az $\mathbf{X}-\mathbf{m}$ véletlZ_2 vektor komponenseinek összes lehetséges ıZ_1rmált (egységvektorral képzett) lineáris kombinációéi között;k szórása maximálisZ_kz összes lehetséges, -től független norm$Z_1,\dots,Z_{k-1}$ kombinációéi közt; s.i.t. a -adik főkomponens, szórása $k=3,\dots,p$z összes lehetséges, -től független normált lineáris kombináció szórása közt ( ).*

Bizonyítást ld. [5] 240-241. o. 1.1. Tétel.

Tehát a $\mathbf{Z}$ p-dimenziós normális eloszlású véletlen vektor komponensei függetlenek és varianciáik a $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ számokkal egyeznek meg. Ezt szemlélteti az alábbi ábra.



Főkomponensek elméleti és empirikus szórásnégyzetei

A $\sum_{i=1}^{p}\lambda_i$ összeg a főkomponensek varianciáinak az összege (a továbbiakban teljes varianciának nevezzük), eredeti változóink teljes varianciája pedig a $\mathbf{C}$ kovarianciamátrix fődiagonálisbeli elemeinek összege, azaz $\mathrm{tr}\mathbf{C}$. Mivel a λ_i számok $\mathbf{C}$ sajátértékei, ezért $\sum_{i=1}^{p}\lambda_i=\mathrm{tr}\mathbf{C}$, ami a varianciák nyelvén azt jelenti, hogy főkomponensek teljes varianciája megegyezik az eredeti változók teljes varianciájával, és ebből a főkomponensek csökkenő sorrendben részesülnek. A főkomponensek szórásai az ún. *kanonikus szórások* (ezek a $\sqrt{\lambda_i}$ számok, $i=1,\dots,p$). Mivel a várható érték vektor hozzáadása csak egy eltolást jelent, a továbbaiakban ezt már levontnak képzeljük el, és eleve $\mathbf{0}$ várható érték vektorú $\mathbf{X}$ véletlen vektor-ból indulunk ki. Ezekután a $\mathbf{Z}=\mathbf{U}^T\mathbf{X}$ főkomponenstranszformáció (a sajátvektorok alkalmas előjelezésével) egy p-dimenziós forgatás, hiszen az $\mathbf{U}^T$ mátrix ortogonális. A fentiek alapján a főkomponens transzformáció egyben azt is jelenti, hogy ha az $\mathbf{u}_1,\dots,\mathbf{u}_p$ sajátvektorok alkotta bázisra térünk át, akkor ezekben az irányokban a transzformált változó varianciája maximális.

A következő állítás mondanivalója az, hogy a főkomponens tranzformáció *forgatásinvariáns*.

5.1.1.2. Állítás. *Legyen az $\mathbf{X}$ p-dimenziós véletlen vektor várható érték vektora $\mathbf{0}$, kovarianciamátrixa pedig $\mathbf{C}$. Tetszőleges $\mathcal{O}$ $p\times p$-s ortogonális mátrix választása esetén az $\mathbf{X}$ és $\mathcal{O}\mathbf{X}$ véletlen vektork főkomponensvektora megegyezik.*

Az állításban megfogalmazott tulajdonság úgy is mondható, hogy a főkomponensvektor *forgatásinvariáns*, azaz ha $\mathbf{X}$ koordinátáit másik ortogonális koordinátarendszerben adjuk meg ($\mathcal{O}$-val elforgatjuk), attól még főkomponensvektora ugyanaz marad. Viszont a főkomponensanalízis *nem skálainvariáns*. Ha például $\mathbf{X}$ komponensei mérések, és mértékegységeinket megváltoztatjuk, akkor ez az $\mathbf{X}\to\mathbf{S}\mathbf{X}$ transzformációt jelenti, ahol az $\mathbf{S}$ diagonális mátrix tartalmazza a váltószámokat. Az $\mathbf{S}\mathbf{X}$ véletlen vektor kovarianciamátrixa $\mathbf{S}\mathbf{C}\mathbf{S}^T=\mathbf{S}CC\mathbf{S}$ lesz. A skálainvariancia azt jelentené, hogy a transzformált $\mathbf{S}\mathbf{X}$ változó főkomponensvektora ugyancsak az $\mathbf{U}$ ortogonális transzformációval adódna, azaz az (1.2) mintájára

$$(\mathbf{S}\mathbf{U})^T\mathbf{S}\mathbf{X}=\mathbf{U}^T\mathbf{S}^2\mathbf{X}=\mathbf{U}^T\mathbf{X}$$

teljesülne. Az $\mathbf{S}\mathbf{U}$ mátrix oszlopvektorai viszont általában nem ortogonálisak, és semmi közük az $\mathbf{S}\mathbf{C}\mathbf{S}$ mátrix sajátvektoraihoz.

A főkomponensanalízis másik fontos optimumtulajdonságát fogalmazza meg a következő tétel: nevezetesen, hogy az első k főkomponens változónk legjobb k-dimenziós közelítését adja az alábbi értelemben. Az $\mathbf{X}$ p-dimenziós véletlen vektor k-dimenziós $(k<p)$ közelítése alatt egy olyan véletlen vektort értünk, amely $\mathbf{A}\mathbf{X}$ alakban áll elő valamely $p\times p$-s, k-rangú $\mathbf{A}$ mátrixszal. Ugyanis $\mathbf{A}\mathbf{X}$ értékeit 1 valószínűséggel az $\mathbf{A}$ oszlopvektorai által kifeszített (k-dimenziós) altérben veszi fel.

5.1.1.3. Tétel. *Legyen* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{C})$ *véletlen vektor Rögzített* $k < p$*-re az*

$$\mathbb{E}||\mathbf{X} - \mathbf{A}\mathbf{X}||^2$$

legkisebb négyzetes eltérést minimalizáló k*-rangú közelítés annak a projekciónak a mátrixával adható meg, amely a* $\mathbf{C}$ *kovarianciamátrix* k *legnagyobb sajátértékéhez tartozó sajátvektora által kifeszített altérre vetít. (A* $\lambda_k = \lambda_{k+1}$ *esetben ez az altér nem egyértelmű.)*

Így a főkomponensanalízis a kovarianciamátrixnak nemcsak a 1.1. Tételbeli optimális felbontását adja, hanem a a kovarianciamátrixnak és így az eredeti változónak is alacsonyabb dimenziós közelítésére ad lehetőséget a 109 Tétel alapján (az első egynéhány főkomponens megtartásával). A fenti tétel alkalmazásakor felmerül k választásának kérdése. Ehhez a

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \ldots\ldots + \lambda_p}$$

hányadost használjuk, amely azt mutatja, hogy az első k főkomponens a teljes variancia hányad részét magyarázza (általában olyan k-t célszerű választani, melyre ``nagy'' az ugrás λ_k és λ_{k+1} közt).

A gyakorlatban az empirikus kovarianciamátrixból indulunk, amely - többdimenziós normális eloszlást feltételezve - az elméleti kovarianciamátrix maximum likelihood becslése. Mivel a sajátértékek és sajátvektorok a kovarianciamátrix folytonos függvényei, az empirikus kovarianciamátrix sajátértékei és sajátvektorai az elméletiek maximum likelihood becslései lesznek (amennyiben a kovarianciamátrix sajátértékei mind különbözőek). A főkomponensanalízisnek akkor van értelme, ha kovarianciamátrixunknak vannak kiugró sajátértékei. k kiugró sajátérték megléte a

$$H_0 : \lambda_{k+1} = \cdots = \lambda_{p-1} = \lambda_p$$

hipotézis elfogadásával ekvivalens, hiszen H_0 fennállása azt jelenti, hogy a legkisebb $p-k$ sajátérték egyenlő. A hipotézisvizsgálatot a $k = 0, 1, \ldots, p-1$ egészekre ilyen sorrendben addig végezzük, amíg adott szinten el nem fogadjuk a null-hipotézist. Ezzel a k-val megegyező számú főkomponenst fogunk beválasztani.

Likelihood hányados próbával adódik, hogy a

$$-2 \ln \lambda_n = n(p-k) \ln \frac{a}{g}$$

statisztika (l. [26]) H_0 fennállása esetén (amennyiben a mintaelemszám elég nagy) közel χ^2_f eloszlást követ, ahol a és g a $\hat{\mathbf{C}}$ empirikus kovarianciamátrix sajátértékeinek számtani- és mértani közepét jelöli:

$$a = \frac{\hat{\lambda}_{k+1} + \cdots + \hat{\lambda}_p}{p-k} \quad \text{és} \quad g = (\hat{\lambda}_{k+1} \ldots \hat{\lambda}_p)^{\frac{1}{p-k}},$$

a χ^2 eloszlás szabadságfoka pedig

$$f = \frac{1}{2}(p-k+2)(p-k-1).$$

Ez az f nem más, mint a sajátértékek egyenlőségére tett feltételek mellett a paraméterek számának a csökkenése. H_0 fenállása esetén a sajátértékek (p) száma csökken $(p-k-1)$-gyel, a sajátvektorokat tartalmazó $p \times p$-s ortogonális mátrixban levő szabad paraméterek száma ($(p-1)p/2$) pedig $(p-k-1)(p-k)/2$-vel, a $(p-k) \times (p-k)$-as forgatások szabad paramétereinek számával (hiszen az azonos sajátértékhez tartozó sajátvektorok egy $(p-k)$-dimenziós altérben tetszőlegesen elforgathatók).

1.2. Faktoranalízis

A főkomponensanalízisnél láttuk, hogy a módszer alkalmas a változók számának csökkentésére. A faktoranalízis célja eleve ez: nagyszámú korrelált változó magyarázata kevesebb korrelálatlannal (többdimenziós normális eloszlás esetén a korrelálatlan helyett független mondható). Ezek a *közös faktorok* azonban nem magyaráznak meg mindent a változókból, csak azoknak az ún. ``közös részét''. Ezen kívül van a változóknak egy ``egyedi része'' is, amelynek leválasztása szintén a modell feladata. A közös faktorokra itt nem úgy kell gondolni, mintha közvetlenül megfigyelhető változók lennének. A k-faktor modell tehát a következő.

Adott a p-dimenziós $\mathbf{X}$ véletlen vektor $\mathbf{m}$ várható érték vektorral és $\mathbf{C}$ kovarianciamátrixszal, többdimenziós normalitás esetén $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}, \mathbf{C})$. Adott $k (1 \leq k < p)$ egészre keressük az

$$\mathbf{X} = \mathbf{A}\mathbf{f} + \mathbf{e} + \mathbf{m} \qquad (1)$$

???felbontást, ahol $\mathbf{A} p \times k$-as mátrix, az $\mathbf{f}$ *közös faktor*$\mathbf{0}$ várható érték vektorú, korrelálatlan komponensű, k-dimenziós véletlen vektor, komponensei 1 szórásúak, az $\mathbf{e}$ *egyedi faktor* p-dimenziós korrelálatlan komponensű véletlen vektor, ráadásul komponensei még $\mathbf{f}$ komponenseivel is korrelálatlanok. A modell feltevései formálisan:

$$\mathbb{E}\mathbf{f} = \mathbf{0}, \quad \mathbb{E}\mathbf{f}\mathbf{f}^T = \mathbf{I}_k,$$
$$\mathbb{E}\varepsilon = \mathbf{0}, \quad \mathbb{E}\mathbf{e}\mathbf{e}^T = \mathbf{f},$$
$$\mathbb{E}\mathbf{e}\mathbf{e}^T = \mathbf{0} \quad \text{a} \quad k \times p\text{-es azonosan 0 mátrix.}$$

Koordinátákra lebontva ez a következőt jelenti:

$$X_i = \sum_{j=1}^{k} a_{ij} f_j + e_i + \mu_i, \quad i = 1, \ldots, p.$$

Mivel e_i és f_j korrelálatlanok, X_i varianciája

$$c_{ii} = \sum_{j=1}^{k} a_{ij}^2 + d_{ii},$$

ahol d_{ii} a $\mathbf{D}$ diagonális mátrix i-edik diagonális eleme nem más, mint az e_i változó (i-edik egyedi faktor) varianciája. Tehát X_i varianciájából a $\sum_{j=1}^{k} a_{ij}^2$ részt magyarázzák a közös faktorok - ezt nevezzük az X_i változó *kommunalitás*ának - , d_{ii} pedig az *egyedi variancia*. A modell paraméterei az $\mathbf{A}$ és $\mathbf{D}$ mátrixok. Az $\mathbf{A}$ mátrixot *faktorsúly-mátrix*nak (más terminológiával átviteli mátrixnak) nevezzük. Ezekkel a modell mátrixalakja a következő:

$$\mathbf{C} = \mathbf{A}\mathbf{A}^T + \mathbf{D}. \qquad (1)$$

???Látható, hogy $\mathbf{X}$ tetszőleges átskálázás után is leírható a k-faktor modellel, ugyanis

$$\mathbf{S}\mathbf{X} = (\mathbf{S}\mathbf{A})\mathbf{f} + \mathbf{e} + \mathbf{S}\mathbf{m}$$

teljesíti a (5.2) modell feltételeit. Az is látható, hogy az $\mathbf{A}$ faktorsúly-mátrix sorainak tetszőleges elforgatása után (azaz az $\mathbf{A}\mathbf{O}$ transzformáció után is, ahol $\mathbf{O} k \times k$-as ortogonális mátrix) faktorsúly-mátrix marad a (5.2) modellben. Még adott k esetén is nehéz megtalálni a (5.3) felbontást. Az egyértelműség kedvéért szokás ezen kívül még további kényszerfeltételeket tenni az $\mathbf{A}$ mátrixra. Például többdimenziós normális eloszlású $\mathbf{X}$, $\mathbf{e}$, $\mathbf{e}$ esetén a k-faktor modell paramétereinek maximum likelihood becslését keresve fel szokták tenni, hogy a $\mathbf{C}$ kovarianciamátrix nem-szinguláris, az

$$\mathbf{A}^T\mathbf{D}^{-1}\mathbf{A} \qquad (1)$$

???mátrix pedig diagonális, diagonális elemei különbözőek, és nem-csökkenő sorrendbe vannak rendezve. Ez a feltétel bizonyos egyértelműséget biztosít a faktorok maximum likelihood becsléséhez, és a számolásokat is egyszerűbbé teszi. A faktorok számát, k-t ``kicsire'' célszerű választani. Kérdés azonban, hogy milyen $k < p$ természetes számokra írható le az n-dimenziós $\mathbf{X}$ véletlen vektor a k-faktor modellel. Ehhez számoljuk össze a (5.3) modell paramétereit: $\mathbf{A}$-ban és $\mathcal{D}$-ben összesen $pk + p$ ismeretlen paraméter van, a (5.4) kényszerfeltétel azonban a diagonálison kívüli elemek 0 voltára vonatkozón $(1/2)(k^2 - k) = (1/2)k(k-1)$ egyenletet jelent (ez megegyezik a $k \times k$-as forgatások szabad paramétereinek számával). Alapvetően pedig van $(1/2)p(p+1)$ egyenletünk (a $\mathbf{C}$ kovarianciamátrix különböző elemei a szimmetria miatt). A felírható egyenletek és a szabad paraméterek számának különbsége:

$$s = (1/2)p(p+1) + (1/2)k(k-1) - (pk + p) = (1/2)(p-k)^2 - (p+k).$$

Általánosságban $s \leq 0$ esetén várható az egyenlet algebrai megoldásának létezése. Ekkor

$$k \geq (2p + 1 - \sqrt{8p+1})/2. \qquad (1)$$

A?A faktormodell identifikálhatóságán azt értjük, hogy rögzített k esetén egyértelműen meg tudjuk adni $\mathbf{D}$-t és -t.

5.1.2.1. Tétel. *Adott $k < p$ természetes szám esetén a (5.3) egyenlet pontosan akkor oldható meg, ha van olyan $p \times p$ -s diagonális $\mathbf{D}$ mátrix (fődiagonálisában nemnegatív elemekkel), hogy a $\mathbf{C} - \mathbf{D}$ mátrix pozitív szemidefinit és rangja nem nagyobb k-nál.*

A tétel valójában a $\mathbf{C} - \mathbf{D}$ mátrix spektrálfelbontásából következik. A faktorok (5.4) melletti maximum likelihood becsléséhez legyen $\mathbf{X} \in \mathcal{N}_p(\mathbf{m}, \mathbf{C})$, $\mathbf{e} \in \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k)$ és $\mathbf{e} \in \mathcal{N}_p(\mathbf{0}, \mathbf{D})$. Jelölje $\hat{\mathbf{C}}$ az $\mathbf{X}$-re vett n-elemű mintából számolt empirikus kovarianciamátrixot. Ezekkel a likelihood függvény logaritmusa

$$-\frac{1}{2}n \log |\mathbf{C}| - \frac{1}{2}n \mathrm{tr}\mathbf{C}^{-1}\hat{\mathbf{C}} + c$$

lesz, ahol c konstans (l. hyperref több dim. gauss parmeter ML becslése, csak ott az $\mathbf{S}$ jelölést használtuk az empirikus kovarianciamátrix n-szeresére: $\mathbf{S} = n\hat{\mathbf{C}}$). Ezekkel a likelihood függvény logaritmusa a (5.3)-beli $\mathbf{C} = \mathbf{A}\mathbf{A}^T + \mathbf{D}$ modell-egyenlet miatt $\mathbf{A}$ és $\mathbf{D}$ függvényének tekinthető, és ezekben kell maximalizálni. Könnyen látható, hogy a feladat ekvivalens az

$$F(\mathbf{A}, \mathbf{D}) = \log |\mathbf{A}\mathbf{A}^T + \mathbf{D}| + \mathrm{tr}(\mathbf{A}\mathbf{A}^T + \mathbf{D})^{-1}\hat{\mathbf{C}}$$

függvény minimalizálásával. Ecélból az $\mathbf{D}$ és $\mathbf{D}$ elemei szerinti parciális deriváltakat 0-val egyenlővé tesszük, következő mátrix-egyenletrendszert megoldjuk:

$$\mathbf{C}^{-1}(\mathbf{C} - \hat{\mathbf{C}})\mathbf{C}^{-1}\mathbf{A} = \mathbf{0}$$
$$\mathrm{diag}[\mathbf{C}^{-1}(\mathbf{C} - \hat{\mathbf{C}})\mathbf{C}^{-1}] = \mathbf{0},$$

ahol $\mathbf{C} = \mathbf{A}\mathbf{A}^T + \mathbf{D}$, $\mathrm{diag}(\cdot)$ pedig olyan diagonális mátrixot jelent, melynek fődiagonálisa megegyezik a zárójelben álló mátrix fődiagonálisával.

A fenti egyenletrendszert numerikus közelítéssel szokták megoldani.

2. Feladatok

(i) Legyen $\mathbf{X}$ egy d-dimenziós vektorváltozó és $\mathbf{Y}$ a hozzá tartozó főkomponensvektor. Adjuk meg X_i és Y_j kovarianciáját!

Tipp: Az általánosság megszorítása nélkül feltehető, hogy $\mathbb{E}(\mathbf{X}) = \mathbf{0}$, a továbbiakban, amikor ennek értelme van ezt mindig feltesszük. Ismeretes hogy $\mathbf{Y} = \mathbf{U}^\top\mathbf{X}$, ahol $\mathbf{U}\{u_{ij}|i = 1, j = 1^n\}$ az $\mathbf{X}$ véletlen vektor $\mathbf{C} = \{c_{ij}|_{i=1,j=1}^n\}$ kovarianciamátrixának $\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^\top$ spektrálelőállításában szereplő ortonormált mátrix. Eszerint

$$Y_j = \sum_{k=1}^{n} u_{kj}X_k \text{ és így } \mathbb{E}(X_i \cdot Y_j) = \sum_{k=1}^{n} u_{kj}\mathbb{E}(X_iX_k)$$

Válasz:

$$\mathbb{E}(X_i \cdot Y_j) = \sum_{k=1}^{n} u_{kj}c_{ik}$$

(ii) Legyen $\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{C})$, ahol $\mathbf{C} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, ahol $0 < \rho < 1$. Adjuk meg a főkomponenseket és a főkomponensvektor kovarianciamátrixát!

Tipp: Az előző feladat Útmutatásában szereplő definíciók alapján meg kell keresni a $\mathbf{C}$ matrix 2 sajátértéket, és a hozzájuk tartozó 1 normáju sajátvektorokat, melyekből összeáll az U mátrix.

Válasz:

$\lambda_1 = 1+\rho,\ \lambda_2 = 1-\rho\ \mathbf{U} = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$

$\mathbf{Y} = \mathbf{U}^\top \mathbf{X}, \mathrm{Cov}(\mathbf{Y}) \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$

Megjegyezzük, hogy $\rho > 0$ esetén a fenti mátrixok a kanonikus (a sajátértékek csökkenő sorrendjnek megfelelő) mátrixok.

(iii) Legyen $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{C})$, ahol $\mathbf{C}$ diagonális mátrix főátlójában különböző (pozitív) értékekkel. Adjuk meg a főkomponensvektort!

Tipp: Ha a $\mathbf{C}$ mátrix diagonális, akkor a főkomponensanalízis feladata a főkomponensek sorrendjétől eltekintve megoldott.

Válasz: $Y_i = X_{\pi(i)}$, ahol π az a permutáció, amely a $\mathbf{C}$ matrix sajátértékeit nemnövekvő sorrendbe rendezi.

(iiii) Legyen $\mathbf{X} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{C})$, ahol $\mathbf{C}$ fődiagonálisának minden eleme 1, minden más eleme r valamely $0 < r < 1$ számra.

(a) Adjuk meg $\mathbf{X}$ első főkomponensét!

(b) Adjuk meg a főkomponensek szórásnégyzeteit!

Tipp: Ez a feladat a 2. feladat általánosítása, a $\mathbf{C}$ sajátértékei: $\mathbf{1} + (d-1)r, 1-r, \ldots, 1-r$, Az $1+r$ (maximális sajátértékhez tartozó) sajátvektor: $\frac{\sqrt{d}}{d}(1,\ldots,1)^\top$, és mivel a maradék $d-1$ sajátérték egyenlő a többi sajátvektor nincs (így az $\mathbf{U}$ mátrix és Y_1-en kívül a többi főkomponens sincs) egyértelműen meghatározva.

Válasz: $Y_1 = \frac{\sqrt{d}}{d}\sum_{j=1}^n X_j$. A főkomponensek szórásnégyzetei a Tippben megadott sajátértékek.

(iiiii) Legyen $\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{C})$, ahol $\mathbf{C} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. Adjunk maximum likelihood becslést $\mathbf{C}$ sajátértékeire!

Tipp: Az $\mathbf{X}$ vektor két komponense (X_1, X_2) két fuüggetlen normális eloszlású 0 várható értékű valószínűségi változó ezért λ_1 és λ_2 M-L becslése a komponensek alapján meghatározhatók, a skalár valószínűségi változók esetében szokásos módon.

Válasz: $\hat{\lambda}_j = \frac{1}{n}\sum_{k=1}^n X_{jk}^2 (j = 1,2)$

Itt n a mintaelemeszám.

(iiiiii) A főkomponensanalízis egy módosított változatában az $\mathbf{R} = r_{ij}|_{ij=1}^n d \times d$-s korrelációs mátrixból indulunk ki.

(a) Mutassuk meg, hogy ezzel a módszerrel más megoldást kapunk, mint a kovarianciamátrixot használó modellben!

(b) A Kaiser-kritérium azon sajátvektorokkal konstruált főkomponenseket választja, amelyekhez tartozó sajátérték legalább a sajátértékek átlaga. Igazoljuk, hogy tetszőleges nemszinguláris korrelációs mátrix sajátértékeinek átlaga 1!

(c) Tegyük fel, hogy a korrelációs mátrix minden eleme nagyobb mint $1-\varepsilon$. Adjunk ε-tol olyan alsó becslést a legnagyobb sajátértékre, amely tart d-hez, midőn $\varepsilon \to 0$ (egy nagy és sok kis szórású főkomponens van)!

(d) Tegyük fel, hogy a korrelációs mátrix sajátértékei a legnagyobb kivételével kisebbek mint ε. Adjunk ε-tol olyan alsó becslést korrelációk minimumára, amely tart 1-hez, midőn $\varepsilon \to 0$.

Tipp:

(a) Elegendő észrevenni azt, hogy a korrelációs mátrix független az $\mathbf{X}$ komponenseinek X_j átskálázásától, míg a kovariancia mátrix függ ettől, megváltoztathatja a sajátértékek sorrendjét, az Y_i valószínűségi változók együtthatóit az főkomponensekben.

(b) Ismeretes, hogy a mátrix nyoma független attól, hogy a mátrix által definiált operátort milyen koordináta rendszerben felírt mátrixszal adjuk meg, így $\mathbf{R}$ sajátértékeinek összege d, átlaga 1.

(c) Legyen $\rho = \min r_{ij}$, és írjuk fel a korrelációs mátrixot $\mathbf{R} = \mathbf{R}_1 + \mathbf{R}_2$ alakban, ahol

$$\mathbf{R}_1 = \begin{pmatrix} 1 & \rho \dots & \rho \\ \rho & 1 \dots & \rho \\ \vdots & \vdots & \vdots \\ \rho & \rho \dots & 1 \end{pmatrix}$$

alakú, míg mátrixot $\mathbf{R}_2$ főátlójában 0-k, állnak, a többi eleme pedig nem nagyobb, mint ε. Alkalmazzuk $\mathbf{R}_2$-ra a t, az összegre pedig a t.

(d) Tegyük fel, hogy $\mathbf{R}$ első sora $(r) = (1, r_2, \dots r_d)$ a legnagyobb sajátértékhez tartozó sajatvektor pedig $\mathbf{e}(1, e_2, \dots e_d)^\top$ (az általanosság korlátozása nélkül feltehetjük, hogy $\mathbf{e}$ első koordinatája 1). Ekkor $\mathbf{Re}$ első koordinátája: $1 + \sum_{j=2}^{d} r_j e_j$ A Schwartz-egyenlőtlenség miatt ez az összeg akkor maximális, ha $\forall j\, e_j = r_j$, azaz a fenti összeg maximuma: $1 + \sum_{j=2}^{d} r_j^2$, ami a feltétel miatt angyobb, mint $1 - d\varepsilon$.

Válasz:

(a) Az Útmutató alapján nyilvánvaló.

(b) Az Útmutató alapján nyilvánvaló.

(c) $d(1 - 2\varepsilon)$ becslést kapunk.

(d) Mivel $\forall j\, |r_j| \leq 1$, a Tippből következik, hogy nincs olyan j, amire $r_j^2 < 1 - d\varepsilon$. Ugyanezt a meggondolás $\mathbf{R}$ minden sorára működik.

(iiiiiii) Tekintsük az $\mathbf{X} = \mathbf{Af} + \mathbf{e} + \mathbf{m}k$-faktor modellt ($\mathbf{X}$ egy d-dimenziós vektorváltozó, $\mathbf{A}$ a $d \times k$-as faktorsúlymátrix, $\mathbf{f}$ a k-dimenziós közös faktor $\mathbf{I}_k$ kovarianciamátrixszal, $\mathbf{e}$ d-dimenziós egyedi faktor D diagonális kovarianciamátrixszal, amelyre $\mathrm{E}(\mathbf{fe}^\top) = \mathbf{0}$).

(a) Mutassuk meg, hogy ha $i \neq j$, akkor X_i és e_j korrelálatlanok!

(b) Adjuk meg X_i változó és e_i egyedi faktorkomponens kovarianciáját!

(c) Adjuk meg X_i változó és f_j közös faktorkomponens kovarianciáját!

Tipp:

(a) Az $\mathbf{X}$ vektorváltozó $i$edik koordinátája: $X_i = \sum_{\ell=1}^{k} a_{i\ell} f_\ell + e_i$ Vegyük figyelemeb, hogy $\mathbb{E}\mathbf{fe}$ a $k \times p$-s azonosan 0 mátrix.

(b) A (a) pont alapján $\mathbb{E}e_i e_j$

(c) Alkalmazzuk X_i (a) pontbeli felírásáat.

Válasz:

(a) Vegyük észre, hogy $\mathbf{e}_j$ az X_i komponens Tippben kifejtett alakjában szereplő minden taggal korrelálatlan, ha $i \neq j$.

(b) A definíciója alapján d_{ij}

(c) A definíciója alapján d_{ij} és a Tipp (a) pontja alapján a_{ij}.

$\mathbf{AA}^\top = \mathbf{BB}^\top$aktoranalízis modelljében legyen $\mathbf{A}$ ($\mathbf{G}k \times$ 't $p \times k$-s ($p > k$) faktorsúly-mát$\mathbf{B} = \mathbf{AG}$yekre . Mutassuk meg, hogy ekkor van olyan méretű ortogonális mátrix, amelyre .

Tipp: Vegyük észre, hogy a $p \times p$-s $\mathbf{AA}^\top$ és $\mathbf{BB}^\top$. mátrixok teljesen leírják a $\mathbf{A}$ és $\mathbf{B}$ mátrixok p darab k dimenziós sora által alkotott $\mathbb{R}^k$ térbeli alakzat geometriai struktúráját: a vektorok hosszait, és bármely két vektor által bezárt szöget. Tehát a két alakzat egybevágó.

Válasz: Bármely két $\mathbb{R}^k$-beli egybevágó alakzat átvihető egymásba egy k-dimenziós forgatással, és esetleg még egy tükrözés alkalmazasával. Ez eppen egy $\mathbf{G}$ ortonormált mátrixszal való szorzás; ha $|\mathbf{G}| = -1$, akkor tükrözni is kell.

(iiiiiiiii) A faktoranalízis modelljének mátrixalakja $\mathbf{C} = \mathbf{AA}^\top + \mathbf{D}$, ahol $\mathbf{A}$ egy $d \times k$-s mátrix, $\mathbf{D}$ pedig egy $d \times d$-s diagonális mátrix nemnegatív elemekkel. Tekintsük a $d = 2$ és $k = 1$ esetet!

(a) Mikor van megoldása a fenti modellnek?

(b) Adjunk maximum likelihood becslést $\mathbf{A}$-ra és $\mathbf{D}$-re!

Tipp:

(a) A modellben 4 paraméter van: a_1, a_2, d_1, d_2 és 3 egyenlet:

$$\begin{aligned} C_{11} &= a_1^2 + d_1 \\ C_{12} &= a_1 a_2 \\ C_{22} &= a_2^2 + d_2, \end{aligned} \qquad (1)$$

???ezért ha van megoldás az általában nem egyértelmű.

Honnan vesszük észre, hogy egy mátrix $\mathbf{AA}^\top$ alakú? A rangja 1, es nemnegatív definit, azaz bevezetve az $a > 0$ és az x paramétereket fennáll az

$$\begin{aligned} a_1^2 &= a \\ a_1 a_2 &= xa \\ a_2^2 &= x^2 a \end{aligned} \qquad (1)$$

???egyenletrendszer. Írjuk be a (5.6) egyenletrendszerbe a (5.7) egyenletrendszert, és oldjuk meg, feltéve, hogy $d_1 = 0$

(b) Írjuk be az (a) pont megoldását a $\mathbf{C}$ mátrix M-L becslésébe.

Válasz:

(a) A megoldás a-ra és x-re: $a = c_{11} x = c_{12}/c_{11}$, ezért $a_1 = \sqrt{c_{11}} a_2 = (a_1 c_{12})/c_{11}$. Mivel a fentiekből következik, hogy $d_2 = c_{22} - \frac{c_{12}^2}{c_{11}}$, azaz amegoldhatosaág feltétele $c_{22} > \frac{c_{12}^2}{c_{11}}$.

(b) A $\mathbf{C}$ mátrix M-L becslése $\frac{1}{n}\mathbf{S}$, ahol n a mintaelemszám.

3. Tesztek

6. fejezet - Lineáris módszerek 2.: regresszióanalízis, a legkisebb négyzetek módszere

1. Elméleti háttér

1.1. Regresszióanalízis

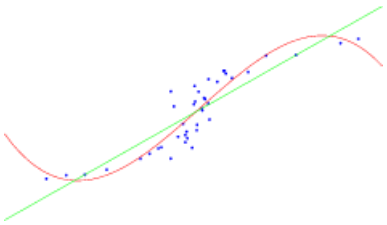
A többváltozós regressziós problémában az Y valószínűségi változót (*függő változó*) szeretnénk az $X_1, \dots, X_p$ valószínűségi változók (*független változók*) függvényével közelíteni legkisebb négyzetes értelemben. Amennyiben ismerjük az $Y, X_1, \dots, X_p$ véletlen vektor együttes eloszlását (tegyük fel, hogy ez abszolút folytonos, az együttes sűrűségfüggvényt jelölje $f(y, x_1, \dots, x_p)$), akkor

$$\mathbb{E}(Y - g(X_1, \dots, X_p))^2$$

minimumát a p-változós g függvények körében Y-nak az $X_1, \dots, X_p$ változók adott értéke mellett vett feltételes várható értéke szolgáltatja:

$$g_{opt}(x_1, \dots, x_p) = \mathbb{E}(Y \,|\, X_1 = x_1, \dots, X_p = x_p) = \frac{\int_{-\infty}^{\infty} y f(y, x_1, \dots, x_p) dy}{\int_{-\infty}^{\infty} f(y, x_1, \dots, x_p) dy},$$

ezt nevezzük regressziós függvénynek.



Regressziós görbe becslése

Adott f sűrűségfüggvény mellett sem mindig triviális a fenti integrál kiszámolása, általában azonban f nem adott, csak egy statisztikai mintánk van a függő és független változókra az $(Y^{(m)}, X_1^{(m)}, \dots, X_p^{(m)})$, $(m = 1, \dots, n)$ független, $(p+1)$-dimenziós megfigyelések formájában. A legegyszerűbb ilyenkor a fenti minimumot a lineáris függvények körében keresni, ezt nevezzük *lineáris regresszió*nak. Erre az esetre vezethető vissza olyan függvényekkel való közelítése Y-nak, amely az X_i változók lineáris függvényének monoton (például exponenciális, logaritmikus) transzformációja. Ilyenkor az inverz transzformációt alkalmazva Y-ra, az így kapott új függő változón hajtunk végre lineáris regressziót az eredeti független változók alapján.

A másik érv a lineáris regresszió mellett az, hogy amennyiben $Y, X_1, \dots, X_p$ együttes eloszlása $(p+1)$-dimenziós normális, akkor a feltétele várható érték képzés valóban lineáris függvényt ad megoldásul (l. 17 Állítást, es (6.1???) Feladatot).

Térjünk rá a lineáris regresszióra. A legjobb

$$Y \sim l(\mathbf{X}) = a_1 X_1 + \dots + a_p X_p + b$$

lineáris közelítést keressük legkisebb négyzetes értelemben, azaz minimalizálni akarjuk az

$$\mathbb{E}(Y - (a_1 X_1 + \dots + a_p X_p + b))^2$$

kifejezést az $a_1, \dots, a_p$ és b együtthatókban. A megoldáshoz először is szabaduljunk meg a várható értékektől, azok csak zavarnak a számolásban, a változók szórása, kovarianciája, mint látni fogjuk, nem változik meg ezáltal. Tehát legyen

$$Y' = Y - \mathbb{E}Y, \quad X_i' = X_i - \mathbb{E}X_i, \quad (i = 1, \dots, p),$$

ezeknek az ún. centrált változóknak a várható értéke már 0 lesz. Így célfüggvényünkön az

$$\begin{aligned} &\mathbb{E}(Y - a_1X_1 - \cdots - a_pX_p - b)^2 = \\ =&\mathbb{E}(\{Y' - a_1X_1' - \cdots - a_pX_p'\}+ \\ +&[\mathbb{E}Y - a_1\mathbb{E}X_1 - \cdots - a_p\mathbb{E}X_p - b])^2 = \\ =&\mathbb{E}(Y' - a_1X_1' - \cdots - a_pX_p')^2 \end{aligned} \qquad (1)$$

???

átalakítás végezhető el, mivel

$$\mathbb{E}Y - a_1\mathbb{E}X_1 - \cdots - a_p\mathbb{E}X_p - b = 0.$$

Ebből a b együtthatóra (ha a_i-k már ismertek lennének) rögtön adódik, hogy

$$b = \mathbb{E}Y - a_1\mathbb{E}X_1 - \cdots - a_p\mathbb{E}X_p,$$

így b-vel a továbbiakban már nem foglalkozunk.

Ezek után az

$$Y' \sim l(\mathbf{X}') = a_1X_1' + \cdots + a_pX_p'$$

lineáris közelítést keressük legkisebb négyzetes értelemben, azaz minimalizálni akarjuk az

$$\mathbb{E}(Y' - (a_1X_1' + \cdots + a_pX_p'))^2 \qquad (1)$$

???kifejezést az $a_1, \dots, a_p$ együtthatókban, feltéve, hogy $\mathbb{E}(Y') = \mathbb{E}(X_1') = \cdots = \mathbb{E}(X_p') = 0$.

Ecélból a

$$\mathbf{Ca} = \mathbf{d}$$

egyenletrendszert kell megoldani, ahol $\mathbf{a} = (a_1, \dots, a_p)^T$, $\mathbf{C}$ jelöli az $\mathbf{X}$ változó $p \times p$-s kovarianciamátrixát, a $\mathbf{d} \in \mathbb{R}^p$ vektor pedig az Y változónak $\mathbf{X}$ komponenseivel vett (kereszt)kovarianciáit tartalmazza. Ennek az egyenletrendszernek létezik egyértelmű megoldása, ha a $\mathbf{C}$ kovarianciamátrix invertálható, tehát $\mathbf{a} = \mathbf{C}^{-1}\mathbf{d}$.

A fenti közelítés maximalizálja korrelációt a következő értelemben.

Jelöljük $\ell(\mathbf{X})$ a fenti lineáris regressziós feladat megoldását, es vezessük be a többszörös korrelációs együttható fogalmát.

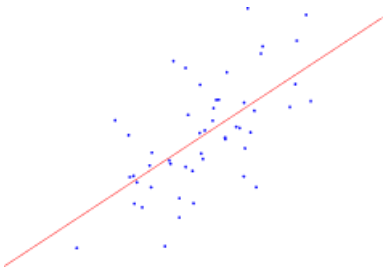
6.1.1.1. Definíció. *Az Y független- és az $X_1, \dots, X_p$ függő változók közötti* többszörös korrelációs együtthatón *Y és $l(\mathbf{X})$ korrelációját értjük és $r_{Y(X_1,\dots,X_p)}$-vel jelöljük.*

A $p = 1$ esetben a többszörös korrelációs együttható a függő- és az egyetlen független változó közötti valódi korrelációs együttható.

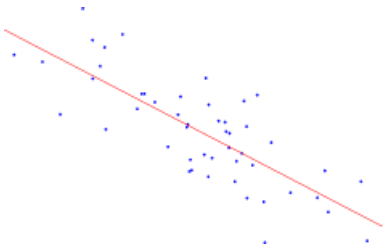
6.1.1.2. Állítás. *Az $X_1, \dots, X_p$ valószínűségi változók tetszőleges $h(\mathbf{X})$ lineáris kombinációjára*

$$|r_{Y(X_1,\dots,X_p)}| = |\mathbf{Corr}(Y, \ell(\mathbf{X}))| \geq |\mathbf{Corr}(Y, h(\mathbf{X}))|.$$

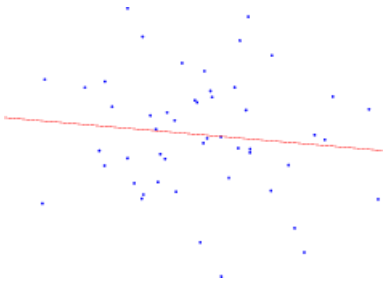
Az alábbi ábrák egyváltozós esetben mutatják a becsléseket.



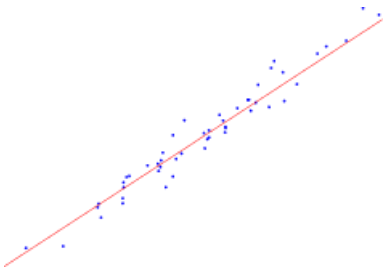
Regressziós egyenes pozitív korreláció esetén



Regressziós egyenes negatív korreláció esetén



Regressziós egyenes függetlenn minta esetén



Regressziós egyenes nagy korreláció esetén

1.2. Legkisebb négyzetek módszere

Legyenek $x_1, \ldots, x_p$ mérési pontok, melyek beállíthatók (tehát nem valószínűségi változók), méréseink pedig ezek valamely ismeretlen $a_1, \ldots, a_p$ paraméterekkel való lineáris kombinációira vonatkoznak, és mérési hibával terheltek. Jelölje ε a mérési hibát, Y a mért értéket, ezek valószínűségi változók. Feltehető, hogy $\mathbb{E}(\varepsilon) = 0$. Modellünk tehát a következő:

$$Y = a_1 x_1 + \cdots + a_p x_p + \varepsilon,$$

ami hasonlít a többváltozós regresszióéhoz, csak ott X_i-k valószínűségi változók. Itt $\mathbb{E}(Y) = \sum_{j=1}^{p} a_j x_j$.

Célunk az ismeretlen $\mathbf{a} = (a_1, \ldots, a_p)^T$ paramétervektor (oszlopvektor) legkisebb négyzetes becslése n mérés alapján ($n \geq p$, általában n sokkal nagyobb, mint p). Az i-edik mérés az $(x_{i1}, \ldots, x_{ip})$ p-dimenziós pontban történik, a mért értéket jelölje Y_i, a mérési hibát pedig ε_i, $(i = 1, \ldots, n)$. Vezessük be még a következő jelöléseket is:

$$\mathbf{Y} := (Y_1, \ldots, Y_n)^T, \quad \underline{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)^T$$

n-dimenziós oszlopvektorok, az $x_{ij} (i = 1, \ldots, n; j = 1, \ldots, p)$ mérési pontokat pedig az $n \times p$-s $\mathbf{X}$ mátrixban gyűjtjük össze. $\mathbf{X}$ oszlopvektorait jelölje $\mathbf{x}_1, \ldots, \mathbf{x}_p$! Ezekkel a jelölésekkel a (4.1) rendszeregyenlet

$$\mathbf{Y} = \mathbf{X}\mathbf{a} + \underline{\varepsilon}$$

alakban írható, ahol tehát $\mathbb{E}(\underline{\varepsilon}) = \mathbf{0}$, továbbá tegyük fel, hogy a mérési hibák korrelálatlanok (normális eloszlás esetén függetlenek) és azonos szórásúak, azaz $\underline{\varepsilon}$ kovarianciamátrixa $\sigma^2\mathbf{I}_n$ alakú. Ekkor persze a mérések is korrelálatlanok, és ugyanaz a kovarianciamátrixszuk, mint $\underline{\varepsilon}$-é:

$$\mathbb{E}(\mathbf{Y} - \mathbf{X}\mathbf{a})(\mathbf{Y} - \mathbf{X}\mathbf{a})^T = \mathbb{E}\underline{\varepsilon}\underline{\varepsilon}^T = \sigma^2\mathbf{I}_n,$$

ahol σ szintén ismeretlen paraméter, melyet majd a végén becsülni fogunk. Az $\mathbf{a}$ ismeretlen paraméter legkisebb négyzetes becslésén azt az $\mathbf{a}$ vektort értjük, amelyre a mérési hibák négyzetösszege,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{a}\|^2 = (\mathbf{Y} - \mathbf{X}\mathbf{a})^T(\mathbf{Y} - \mathbf{X}\mathbf{a}) = (\mathbf{Y}^T - \mathbf{a}^T\mathbf{X}^T)(\mathbf{Y} - \mathbf{X}\mathbf{a}) =$$
$$= \mathbf{Y}^T\mathbf{Y} - \mathbf{a}^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{a} + \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a}$$

minimális. Az $\mathbf{a}$ vektor szerint deriválva és a deriváltat $\mathbf{0}$-val egyenlővé téve a

$$-\mathbf{X}^T\mathbf{Y} - (\mathbf{Y}^T\mathbf{X})^T + 2\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{0}$$

vektoregyenlet adódik, amelyből átrendezéssel és 2-vel osztva

$$\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}^T\mathbf{Y} \qquad (1)$$

???következik. Ezt nevezzük Gauss-féle *normálegyenlet*nek.

A normálegyenleteket a geometriai szemlélet alapján is megkaphatjuk következőképpen. $\|\mathbf{Y} - \mathbf{X}\mathbf{a}\|^2$ nyilván akkor minimális $\mathbf{a}$-ban, ha $\mathbf{X}\mathbf{a}$ az $\mathbf{Y}$ vektornak az F altérre való merőleges vetülete, ahol az $F \subset \mathbb{R}^n$ alteret $\mathbf{X}$ oszlopvektorai (az $\mathbf{x}_1, \ldots, \mathbf{x}_p$ vektorok) feszítik ki, $\dim(F) = r \leq p$ (tipikusan p-vel egyenlő, ha az $\mathbf{x}_i$ vektorok lineárisan függetlenek). Jelölje $\mathbf{P}$ ennek az r-rangú ortogonális projekciónak az $n \times n$-es mátrixát! Ezzel az optimális $\mathbf{a}$-ra $\mathbf{X}\mathbf{a} = \mathbf{P}\mathbf{Y}$ és

$$\mathbf{Y} = \mathbf{P}\mathbf{Y} + (\mathbf{I} - \mathbf{P})\mathbf{Y}, \quad \text{azaz} \quad \mathbf{Y} = \mathbf{X}\mathbf{a} + (\mathbf{Y} - \mathbf{X}\mathbf{a}), \qquad (1)$$

???ugyanis az $\mathbf{X}\mathbf{a}$ vektor az $\mathbf{x}_1, \ldots \mathbf{x}_p$ vektorok lineáris kombinációja. Mivel $\mathbf{X}\mathbf{a} \in F$, $\mathbf{Y} - \mathbf{X}\mathbf{a}$ pedig merőleges F-re, ezért $\mathbf{Y} - \mathbf{X}$ merőleges F tetszőleges vektorára, ami $\mathbf{X}\mathbf{b}$ alakú lesz valamely $\mathbf{b} \in \mathbb{R}^p$ vektorral. Így

$$(\mathbf{X}\mathbf{b})^T \cdot (\mathbf{Y} - \mathbf{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p.$$

Ebből

$$\mathbf{b}^T\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{a}) = 0, \quad \forall \mathbf{b} \in \mathbb{R}^p.$$

Ez csak úgy lehetséges, ha

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\mathbf{a}) = \mathbf{0},$$

azaz

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\mathbf{a}$$

adódik, ami nem más, mint a (6.3) normálegyenlet. A normálegyenlet mindig konzisztens, hiszen az $\mathbf{X}^T\mathbf{Y}$ vektor benne van az $\mathbf{X}^T$ mátrix oszlopvektorai által kifeszített altérben, és ugyanezt az alteret feszítik ki az $\mathbf{X}^T\mathbf{X}$ mátrix oszlopai is. A megoldás pontosan akkor egyértelmű, ha az $\mathbf{X}^T\mathbf{X}$ mátrix rangja $r = p(\leq n)$, ilyenkor a megoldás

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

alakban írható. A gyakorlatban általában az $\mathbf{X}^T\mathbf{X}$ mátrix invertálható. Az $\mathbf{a}$ vektornak a normálegyenlet megoldásaként kapott becslése torzítatlan, igaz a következő állítás:

6.1.2.1. Állítás. *Ha* $r = p$ *és* $\underline{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, *akkor* $\hat{\mathbf{a}} \sim \mathcal{N}_p(\mathbf{a}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

A Gauss- Markov-tétel szerint $\hat{\mathbf{a}}$ minimális kovarianciamátrixú az $\mathbf{a}$-ra vonatkozó lineáris, torzítatlan becslések között.

6.1.2.2. Tétel. *Legyen* $r = p$ *és* $\tilde{\mathbf{a}}$ *az* $\mathbf{a}$ *paramétervektor tetszőleges lineáris torzítatlan becslése. Ekkor*

$$\mathbb{D}^2(\hat{\mathbf{a}}) \leq \mathbb{D}^2(\tilde{\mathbf{a}}),$$

azaz a $\mathbb{D}^2(\tilde{\mathbf{a}}) - \mathbb{D}^2(\hat{\mathbf{a}})$ *mátrix pozitív szemidefinit.*

A σ^2 közös szórásnégyzet becsléséhez vezessük be a következő jelölést:

$$S^2_{\underline{\varepsilon}} := ||\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}}||^2 = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}}),$$

ezt a mennyiséget *reziduális varianciának* nevezzük.

A geometriai szemlélet (projekciók) alapján $S^2_{\underline{\varepsilon}}$ a következő alakban is írható:

$$S^2_{\underline{\varepsilon}} = (\mathbf{Y} - \mathbf{P}\mathbf{Y})^T(\mathbf{Y} - \mathbf{P}\mathbf{Y}) = ((\mathbf{I} - \mathbf{P})\mathbf{Y})^T((\mathbf{I} - \mathbf{P})\mathbf{Y}) =$$
$$= \mathbf{Y}^T(\mathbf{I} - \mathbf{P})^2\mathbf{Y} = \mathbf{Y}^T(\mathbf{I} - \mathbf{P})\mathbf{Y},$$

mivel $\mathbf{I} - \mathbf{P}$ is egy projekció mátrixa, melynek rangja $n - p$. Ezért $S^2_{\underline{\varepsilon}}$ az I.3.6. Állítás a. része alapján előállítható $n - p$ db. független, σ^2 varianciájú, normális eloszlású valószínűségi változó négyzetösszegeként, így $S^2_{\underline{\varepsilon}} \sim \sigma^2\chi^2_{n-p}$, továbbá $\mathbb{E}(S^2_{\underline{\varepsilon}}) = \sigma^2(n - p)$. Ebből az is következik, hogy

$$\hat{\sigma}^2 = \frac{S^2_{\underline{\varepsilon}}}{n - p}$$

torzítatlan becslés σ^2-re. Megjegyezzük, hogy amennyiben az $\mathbf{X}$ mátrix rangja $r < p$, a $\mathbf{P}$ projekció rangja is r, következésképpen

$$\hat{\sigma}^2 = \frac{S^2_{\underline{\varepsilon}}}{n - r}$$

a σ^2 paraméter torzítatlan becslése. Megjegyezzük, hogy ha a konstans tagot is becsüljük, akkor a nevezőben $n - r - 1$ áll.

Az alábbi *animáció* szemlélteti, hogy nagy szórás esetén egy pont mennyire változtatja meg a becslést.

A

$$H_0\colon\ a_1 = \cdots = a_n = 0$$

Nullhipotézis tesztelésére a likelihood-hányados próbát használjuk, ebben a szerencsés esetben a λ_n próbafüggvény az ismert $(\mathcal{F}(p, n - p))$ eloszlású

$$F = \frac{\mathbf{Y}^\top\mathbf{P}\mathbf{Y}}{\mathbf{Y}^\top(\mathbf{I} - \mathbf{P})\mathbf{Y}} \cdot \frac{n - p}{p}$$

statisztikának szigorúan monoton függvénye.

2. Feladatok

(i) Legyen $(Y, X_1, \ldots, X_m) \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, ahol $c_{ii} = 1$ és $c_{1i} = c_{i1} = 1/m$, $\mathbf{C}$ minden más eleme 0. Adjuk meg az $\mathrm{E}((Y - g(X_1, \ldots, X_m))^2)$-et mininalizáló regressziós függvényt!

Tipp: $\mathbf{a}$ meghatározásához ld.

Válasz: $l(\mathbf{X}) = (X_1 + \ldots + X_m)/m$.

(ii) Igaz-e, hogy ha X, Y véges szórású valószínűségi változók, valamint $Y \sim aX + b$ a legjobb lineáris közelítés négyzetes értelemben, akkor

(a) $\mathrm{r}(X,Y) = a \cdot \dfrac{\mathbb{D}(X)}{\mathbb{D}(Y)}$?

(b) Tetszőleges valós számokra $\mathrm{E}((Y-(aX+b))^2) \geq (1-\mathrm{r}(X,Y))\mathbb{D}^2(Y)$?

Tipp: Centráljuk az Y és X valószínűségi változókat:

$X' = \mathbb{E}(X)\, Y' = Y - \mathbb{E}(Y).$

Ebből a modell alapján azonnal leolvasható, hogy ha a ismert, akkor

$b = \mathbb{E}(Y) - a\mathbb{E}(X).$

Válasz: Mindkettő igaz.

(iii) Legyen $(Y, X_1, \ldots, X_m) \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$. Adjuk meg az $\mathrm{E}((Y - g(X_1, \ldots, X_m))^2)$ -et mininalizáló regressziós függvényt!

Tipp: Jelölje $\ell(x_1, \ldots, x_m)$ azt a lineáris függvényt amely a lineáris függvények körében minimalizálja a $\mathrm{E}((Y - \ell(X_1, \ldots, X_m))^2)$ négyzetes eltérést. $\mathrm{E}((Y - \ell(X_1, \ldots, X_m))X_j) = 0$ minden $j = 1, \ldots, m$-re. A 90 Állítás miatt ebből következik, hogy $Y - \ell(X_1, \ldots, X_m)$ független az X_j valószínűségi változóktól.

Válasz: Alkalmazzuk a 16 és 17 Állításokat

(iiii) Igazoljuk, hogy ha X, Y véges szórású valószínűségi változók, valamint $Y \sim aX + b$ a legjobb lineáris közelítés négyzetes értelemben, akkor

(a) $\mathrm{r}(X,Y) = a \cdot \dfrac{\mathbb{D}(X)}{\mathbb{D}(Y)}$,

(b) Tetszőleges valós számokra $\mathrm{E}((Y-(aX+b))^2) \geq (1-\mathrm{r}(X,Y))\mathbb{D}^2(Y)$.

Tipp: Centráljuk az Y és X valószínűségi változókat:

$X' = \mathbb{E}(X)\, Y' = Y - \mathbb{E}(Y).$

(a) Ebből a modell alapján azonnal leolvasható, hogy ha a ismert, akkor

$b = \mathbb{E}(Y) - a\mathbb{E}(X).$

(b) Ezek után az a paramétert becsülhetjükaz $Y' \sim aX'$ modell alapján.

Válasz:

(a) Az Útmutató (b) pontja alapján nyilvánvaló.

(b) Ha a és b a becslés alapján kapott számok, akkor a kérdés (b) pontjában egyenlőség áll, egybként pedig a Schwartz-egyenlőtlenség következménye.

(iiiii) Tekintsük az (X, Y) véletlen vektort, az $l_1(X) = aX + b$ (amelyre $\mathrm{E}((Y - l_1(X))^2)$ minimális) és az $l_2(Y) = cY + d$ (amelyre $\mathrm{E}(X - l_2(Y))^2$ minimális) regressziós egyeneseket. Mikor teljesül, hogy $c = 1/a$?

Tipp: Oldjuk meg a

$\mathbb{E}(Y) = a + \mathbb{E}(X)b\mathbb{E}(XY) \quad = \mathbb{E}(X)a + [\mathbb{E}(X)]^2 b$

normálegyenletet, és ugyanezt az $X \leftrightarrow Y$ szerepcserével.

Válasz: Ha $\mathrm{Cov}(X,Y) = \pm 1$.

$Y_i = ax_i + b + \epsilon_i, \quad i = 1, \ldots, n$ mérési pontok, továbbá $Y_1, \ldots, Y_n$ valt$\epsilon_1, \ldots, \epsilon_n \sim \mathcal{N}(0, \sigma^2)$légítik a regressziós modellt, ahol a mérési hibák független valószínűségi változók.

(a) Adjunk maximum likelihood becslést az (a, b, σ^2) paraméterre a $\mathbf{Y}$ minta segítségével! (Mi köze a kapott becslésnek a legkisebb négyzetek módszeréhez?)

(b) Igazoljuk, hogy a és b fenti becslései pontosan akkor korrelálatlanok, ha $\overline{x} = 0$.

(c) Adjunk konfidencia-intervallumot a-ra, ha $b = 0$ és σ ismert.

(d) Konstruáljunk a $H_0 : a = a_0$ és $H_1 : a \neq a_0$ hipotézisekhez ε terjedelmű próbát, feltéve, hogy b és σ^2 ismert!

(e) Konstruáljunk likelihood-hányados próbát $H_0 : a = a_0$ és $H_1 : a \neq a_0$ hipotézisekhez, ha $b = 0$ és σ^2 ismeretlen!

(f) Konstruáljunk likelihood-hányados próbát $H_0 : a = a_0$ és $H_1 : a \neq a_0$ hipotézisekhez, ha b és σ^2 ismeretlen!

(g) Hogyan ellenőrizhetjük a modell alkalmazhatóságát, azaz a mérési hibákra vonatkozó feltételek teljesülését?

Tipp: Az egyszerűbb írásmód kedvéért bevezetjük a következő jelöléseket:

$\mathbf{X} = (x_1, \ldots, x_n)^\top$
$\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$.

Továbbá írjuk fel a minta sűrűségfüggvényét ismert a, b és σ^2 paraméterek mellett (Nota Bene: x_i-k NEM valószínűségi változók):

$$f(y_1, \ldots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{\sum_{i=1}^n (y_i - ax_i - b)^2}{2\sigma^2}\} \qquad (1)$$

???

(a) Írjuk fel a modell alapján az $Y_1, \ldots, Y_n$ valószínűségi változók likelihood függvényének logaritmusát. Az a és b paraméterek becslése éppen a (legkisebb négyzetek módszere) alapfeladatának megoldása. Ezután alkalmazzuk a paragrafusban tárgyalt módszert.

(b) Írjuk fel a normálegyenletet, ami ekkor két független egyenlet lesz a-ra és b-re, $\hat{b} = \bar{\mathbf{Y}}$. Megforditva: oldjuk meg a normálegyenletet.

(c) Az egyszerűség kedvéért tegyük fel, hogy $\sigma^2 = 1$. A normálegyenlet megoldása: $\hat{a} = \frac{\mathbf{X}^\top \mathbf{Y}}{\mathbf{X}^\top \mathbf{X}}$ Ekkor $\hat{a} \sim \mathcal{N}(a, (\mathbf{X}^\top \mathbf{X})^{-1})$.

(d) Alkalmazzuk az u-próbát a (c) pont felhasználásával.

(e)

A próbafüggvényt két sűrűségfüggvény hányadosaként kapjuk meg: a számlálóban a minta sűrűségfüggvényében $a = a_0$, $b = 0$ és σ^2 ugyanezen feltevések melletti $S(\varepsilon, a_0, 0) = \sum_{i=1}^n (y_i - a_0 x_i)^2/n$ beslése áll, míg a nevezőbeli sűrűségfüggvényben $a = \hat{a}$, $b = 0$ és σ^2 ugyanezen feltevések melletti $S(\varepsilon, \hat{a}, 0) = \sum_{i=1}^n (y_i - \hat{a}x_i)^2/n$ becslése áll. Vegyük észre, hogy az exponenciális faktor mindkét esetben $e^{-n/2}$-vé egyszerűsödik.

(f) Hasonló a (d) ponthoz, csak σ^2 becsleésében $b = 0$ helyett mind a számlálóban mind a nevezőben $b = \hat{b}$ áll.

(g) Vizsgáljuk meg likelihood-hányados próbaval, hogy teljesül-e a $H_0 : \mathbf{a} = \mathbf{0}$ hipotézis! Ha igen, akkor nincs értelme legkisebb négyzetek módszere alkalmazásának. A próbastatisztikát az alábbi hányadossal definiáljuk.

$$\lambda_n = \frac{\sup\limits_{\mathbf{a}=\mathbf{0},\sigma^2} L_{\mathbf{a},\sigma^2}(\mathbf{Y})}{\sup\limits_{\mathbf{a},\sigma^2} L_{\mathbf{a},\sigma^2}(\mathbf{Y})}.$$

Ezen kívül azt kell ellenőrizni, hogy az egyes reziduális epszilonok független azonos eloszlásúk-e.

Válasz:

(a) Az Útmutató alapján csak a σ^2 becslésére kell kitérni: Jelölje $\hat{a}$, illetve $\hat{b}$ az a, illetve b paraméterek M-L becsléseit továbbá legyen $S(\varepsilon) = \sum_{i=1}^{n}(Y_i - \hat{a}x_i - \hat{b})^2$ reziduális szórásnégyzet. A σ^2 M-L becslése $S(\varepsilon)/n$.

(b) Az egyik irány várható érték képzéssel adódik a Tippből. A másik irány abból következik, hogy a normálegyenlet megoldásaként (l. (c) pont) számított $\mathrm{Cov}(\hat{a}, \hat{b}) = c\bar{x}$, ahol $c \neq 0$.

(c) $\hat{a} \pm \frac{1}{\sqrt{\mathbf{X}^\top \mathbf{X}}} \Phi^{-1}(1 - \varepsilon/2)$.

(d) Ha

$$\hat{a} \notin \left[a_0 - \frac{1}{\sqrt{\mathbf{X}^\top \mathbf{X}}} \Phi^{-1}(1 - \varepsilon/2),\, a_0 + \frac{1}{\sqrt{\mathbf{X}^\top \mathbf{X}}} \Phi^{-1}(1 - \varepsilon/2) \right].$$

elvetjük a H_0 hipotézist.

(e) Az Útmutató alapján a $\lambda(y_1, \ldots, y_n)$ próbafüggvény az exponenciális tényezők elött álló tényezők hányadosa lesz:

$$\lambda(y_1, \ldots, y_n) = \left(\frac{\sum_{i=1}^{n}(y_i - \hat{a}x_i)^2}{\sum_{i=1}^{n}(y_i - a_0 x_i)^2} \right)^{n/2}$$

(f)

$$\lambda(y_1, \ldots, y_n) = \left(\frac{\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2}{\sum_{i=1}^{n}(y_i - a_0 x_i - \hat{b})^2} \right)^{n/2}$$

(g) Vizsgáljuk meg likelihood-hányados próbaval (hyperref!!), hogy teljesül-e a $H_0 : \mathbf{a} = \mathbf{0}$ hipotézis! A próbastatisztika a

$$\frac{(\mathbf{Y}^T \mathbf{P} \mathbf{Y})/p}{\mathbf{Y}^T(\mathbf{I} - \mathbf{P})\mathbf{Y}/(n-p)}, \qquad (1)$$

(ahol $\mathbf{P}$ korábban definiált mátrixa) $\mathcal{F}(p, (n-p)$ eloszlású, ezért a kritikus tartomány konstrukciója nyilvánvaló.

Ezután autokovarianciát alkalmazunk, ami itt azt jelenti, hogy a rezidualis szórások indexeit 1-gyel eltoljuk és az eredeti valamint az eltolt vektor kovarianciáját számojuk.

(iiiiiii) Tekintsük az $Y = \mathbf{a}^\top \mathbf{x} + \epsilon$ regressziós modellt, ahol $\epsilon \sim \mathcal{N}(0, \sigma^2)$, σ^2 ismert értékre. Konstruáljuk meg a Neyman-Pearson alaplemma segítségével a $H_0 : \mathbf{a} = \mathbf{a}_0$ vs. $H_1 : \mathbf{a} = \mathbf{a}_1$ egyszerű alternatívához tartozó ε terjedelmű próbát!

Tipp: Írjuk fel a feladatban szereplő modellt koordinátánként.

$$Y_i = \sum_{j=1}^{d} a_j x_{ij}$$

Írjuk fel a minta sűrűségfüggvényeit ismert $\mathbf{a}_0$, ($\mathbf{a}_1$) és σ^2 paraméterek mellett:

$$f_0(y_1,\ldots,y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{\sum_{i=1}^n (y_i - \sum_{j=1}^d a_{0,j}x_{ij})^2}{2\sigma^2}\}$$

$$f_1(y_1,\ldots,y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{\sum_{i=1}^n (y_i - \sum_{j=1}^d a_{1,j}x_{ij})^2}{2\sigma^2}\}$$

Tegyünk két észrevételt.

(a) f_1/f_0 hányados kitevőjében csak a $\sum_{i=1}^n y_i(\sum_{j=1}^d a_{1,j}x_{ij} - \sum_{j=1}^d a_{0,j}x_{ij})$ tag konstansszorosa szerepel.

(b) Mivel az elsőfajú hiba rögzitett a feladat valójában nem más mint u-próba szerkesztése $\sum_{j=1}^d a_{0,j}x_{ij}$ várható értékű σ^2 szórásnégyzetű normális eloszlásra n minta alapján.

Válasz: Ha $\sum_{j=1}^d a_{1,j}x_{ij} > \sum_{j=1}^d a_{0,j}x_{ij}$ akkor a kritikus tartomány

$$\left\{\sqrt{n}\frac{\bar{Y} - \sum_{j=1}^d a_{0,j}x_{ij}}{\sigma} > \Phi^{-1}(1-\varepsilon)\right\}$$

(iiiiiiii) Tekintsük az $Y = a_1x_1 + \ldots + a_dx_d + b + \epsilon$ regressziós modellt és a $H_0 : a_1 = \ldots = a_d = 0$ hipotézist tesztelő regresszióanalízist.

(a) Legyen $Q = \sum_{i=1}^n (Y_i - \overline{Y})^2$, $Q_r = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ és $Q_e = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$, ahol $\hat{Y}_i = \hat{a}_1x_{i,1} + \ldots + \hat{a}_dx_{i,d} + \hat{b}$. Igazoljuk, hogy $Q = Q_r + Q_e$.

(b) Jelölje R_n a többszörös korrelációs együttható becslését. Mutassuk meg, hogy $R_n^2 = \frac{Q_r}{Q}$.

(c) Igazoljuk, hogy a próbastatisztika $F = \frac{(n-d-1)Q_r}{dQ_e} = \frac{(n-d-1)R_n^2}{d(1-R_n^2)}$ alakokban is felírható!

(d) Vessük össze a regresszióanalízist a korrelációs együtthatókra vonatkozó tesztekkel! Indokolt-e a regresszióanalízist függetlenség tesztelésére használni?

Tipp:

Válasz:

(iiiiiiiii) Vessük össze a lineáris regresszió megoldását ($\mathbf{a} = \mathbf{C}^{-1}\mathbf{d}$, ha a várható értékek 0-k) a determinisztikus változók esetén kapott megoldással ($\hat{\mathbf{a}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$)!

Tipp:

Válasz: Vegyük észre, hogy $(\mathbf{X}\mathbf{X}^\top)$ éppen $\mathbf{C}$ M-L becslése.

(iiiiiiiii) Igazoljuk, hogy $\mathbf{X}^\top\mathbf{X}$ pontosan akkor nemszinguláris, ha $\mathbf{X}$ oszlopvektorai lineárisan függetlenek.

Tipp: Lehetne hivatkozni lineáris algebrai tételekre, de a legkisebb négyzetek módszerének témaköréhez tartozó egyszerű meggondolás is célravezető.

Válasz: A legkisebb négyzetek módszerének geometriai interpretációja következő: Keressük az $\mathbf{Y}$ vektornak az $\mathbf{X}$ mátrix oszlopvektorai által kifeszített térre való merőleges vetületét. Ez a vetület pontosan akkor fejezhető ki egyértelműen ezen vektorok lineáris kombinációjával, ha lineárisan függetlenek. A normálegyenlet egyértelmű megoldhatóságanak pedig éppen az a szükséges és elegséges felétele, hogy az $\mathbf{X}\mathbf{X}^{top}$ mátrix nemszinguláris.

(iiiiiiiiii) Tekintsük a következő multiplikatív modellt: $Y = bX_1^{a_1} \cdot \ldots \cdot X_k^{a_k}$. Vezessük vissza a lineáris modellre, és adjunk becslést a paraméterekre a módosított modellben a legkisebb négyzetek módszerével! Más becslést kapnánk-e, ha a legkisebb négyzetek módszerét közvetlenül az eredeti modellre alkalmaznánk?

Tipp: Az eredeti modell helyett tekintsük az alábbi „logaritmikus" modellt: $\log Y = \log b + a_1 \log X_1 + \ldots + a_k \log X_k$.

Válasz: A feladat eloő részb = 0, $a_2 = 0, \ldots, a_k = 0$za a Tipp, a második részre a válasz, IGEN, mas becslést kapnánk, ez ellenőrizhető a modellen két mintaelem esetén.

(iiiiiiiiiii) Polinomiális regresszió esetén a modell $Y = b + a_1 X + \ldots + a_k X^k$ alakú. A megoldást úgy keresik, hogy az $X^i = X_i$ valószínűségi változókat formálisan függetleneknek tekintik és megoldják a rájuk vonatkozó többváltozós lineáris regresszió feladatát. Viszont X^i és X^j általában nem független változók. Okoz-e ez problémát a megoldás egyértelműsége tekintetében? Miért?

Tipp: Írjuk fel a modellhez tartozó normálegyenlet mátrixát a várható érték képzés előtt, pl $k = 2$-re:

$$\mathbf{R}\begin{pmatrix} 1 & X & X^2 \end{pmatrix}$$

Ez a mátrix a egy valószínűséggel 1-rangú, amiből nem következik, hogy a várható érték vétel után is 1-rangú marad.

Válasz: valójában nem okoz problémát, mert Y-t az X Hermite-polinomjaival is közelíthetjük (ezek éppen a Gauss-sűrűségre nézve ortogonális polinomok, amelyekből az X hatványai egyértelműen visszaszámolhatók) és ebben a sémában a normálegyenlet mátrixa diagonalis lesz! Mármost ez túl megy a zaróvizsga tételeken!!!!!

3. Tesztek

7. fejezet - Lineáris módszerek 3.: Egy- és többszempontos varianciaanalízis

1. Elméleti háttér

A varianciaanalízis speciális lineáris modelleket vizsgál, kísérlettervezésben és minőségellenőrzésben felmerülő hipotézisek tesztelésére. A tekintett modellek specifikuma az, hogy a legkisebb négyzetek módszerénel alkalmazott modellben a beállítható mérési pontok mátrixa helyett 0-1 elemekből álló ún. *struktúramátrix*szal dolgozunk, amelyet úgy állítunk össze, hogy bizonyos megfigyelések csak bizonyos paraméterektől függjenek. A hipotézisek vizsgálata is a likelihood hányados próba analógiájára történik.

Gyakorlati alkalmazásokban olyan mintákat vizsgálunk, melyeket különböző körülmények közt figyeltünk meg, és célunk éppen annak a megállapítása, vajon ezek a körülmények jelentősen befolyásolják-e a mért értékeket. Tehát mintánkat eleve csoportokba osztottan kapjuk, feltesszük azonban, hogy a különböző csoportokban felvett minták egymástól függetlenek, normális eloszlásúak és azonos szórásúak.

A Tananyagban csak az egyszempontos varianciaanalízissel és a kétszempontos varianciaanalízis interakciót tesztelő változatával foglalkozunk, ugyanis az interakció nélküli kétszempontos varianciaanalízis csak formálisan bonyolultabb az egyszempontosnál, de új jelenséget nem vizsgál.

1.1. Egyszempontos varianciaanalízis

Valamilyen szempont alapján (például különböző kezelések) k csoportban külön végzünk megfigyeléseket. Az egyes csoportokban a mintaelemek száma általában nem egyenlő: jelölje n_i az i. csoportbeli mintaelemek számát, $n = \sum_{i=1}^{k} n_i$ pedig az összminta elemszámát. Az i. csoportban az $X_i \sim \mathcal{N}(b_i, \sigma^2)$ valószínűségi változóra vett mintaelemeket

$$X_{ij} \sim \mathcal{N}(b_i, \sigma^2), \quad (j = 1, \ldots, n_i)$$

jelöli. Ezek egymás közt és különböző i-kre is függetlenek, azonos szórásúak. A várható értékekre a $b_i = m + a_i$ felbontást alkalmazzuk, ahol m a várható értékek súlyozott átlaga, a_i pedig az i. csoport hatása:

$$m = \frac{1}{n} \sum_{i=1}^{k} n_i b_i, \qquad a_i = b_i - m \quad (i = 1, \ldots, k).$$

Könnyen látható, hogy

$$\sum_{i=1}^{k} n_i a_i = 0. \qquad (1)$$

???Ezekkel a jelölésekkel az egyszempontos modell

$$X_{ij} = m + a_i + \varepsilon_{ij} \qquad (j = 1, \ldots, n_i; i = 1, \ldots, k)$$

alakban írható, ahol az $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ független valószínűségi változók véletlen hibák.

Lineáris modellről van szó, hiszen ha megfigyeléseinket az

$$\mathbf{Y} := (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}, \ldots, X_{k1}, \ldots, X_{kn_k})^T$$
$$\underline{\varepsilon} := (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \varepsilon_{21}, \ldots, \varepsilon_{2n_2}, \ldots, \varepsilon_{k1}, \ldots, \varepsilon_{kn_k})^T$$

$\sum_{i=1}^{k} n_i = n$-dimenziós vektorban, a_i paramétereinket pedig az $\mathbf{a} = (a_1, \ldots, a_k)^T$ vektorban helyezzük el, akkor az (5.2) modell az

$\mathbf{Y} = \mathbf{B} \cdot \mathbf{a} + \mathbf{1} \cdot m + \underline{\varepsilon}$

alakban írható, ahol $\mathbf{1} \in \mathbb{R}^n$ az azonosan 1 koordinátájú vektor, $\mathbf{B}$ pedig az alábbi (7.2) alakú struktúramátrix:

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad (1)$$

(Ebben a példában $k = 3$, $n_1 = 3$,$n_2 = 4$ és $n_3 = 5$.)

Látható, hogy $\mathrm{rang}\mathbf{B} = k$, az oszlopok által kifeszített k-dimenziós alteret jelölje F; nyilván $\mathbf{1} \in F$. A paramétereket közvetlenül a legkisebb négyzetek módszerével becsüljük, azaz keressük a

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - m - a_i)^2 \qquad (1)$$

???kifejezés minimumát az $m, a_1, \ldots, a_k$ paraméterekben az (7.1) kényszerfeltétel mellett. Vezessük be a csoportátlagokra ill. a teljes mintaátlagra az

$$\bar{X}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \quad (i = 1, \ldots, k) \quad \text{ill.} \quad \bar{X}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$$

jelöléseket! Könnyen látható, hogy a paraméterek legkisebb négyzetes becslései

$$\hat{m} = \bar{X}_{..} \quad \text{és} \quad \hat{a}_i = \bar{X}_{i.} - \bar{X}_{..} \quad (i = 1, \ldots, k)$$

lesznek. Ugyanis m helyébe a nyilvánvaló $\bar{X}_{..}$-ot í rva az (7.3) kifejezés minimuma kereshető az egyes a_i-kben külön-külön - csak a külső szumma i-edik tagjában álló négyzetösszeg minimalizálásával - , hiszen a_i becslése csak az $X_{ij}, \quad j = 1, \ldots, n_i$ mintaelemektől függ $(i = 1, \ldots, k)$, és a Steiner-tétel alapján a fenti lesz. (A szélshőérték számítás módszereivel ellenőrízhető a fenti heurisztikus számolás helyessége.)

A minimum értéke

$$Q_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \hat{m} - \hat{a}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2$$

lesz. A *Legkisebb négyzetek módszere* paragrafus jelöléseivel Q_e az $S_{\underline{\varepsilon}}^2$ reziduális variancia. Az alább taglalandó vetítéssel Q_e a merőleges komponens hosszának a négyzete, míg a vetület hosszának négyzete:

$$Q_a = \|\mathbf{B}\hat{\mathbf{a}}\|^2 = \sum_{i=1}^{n_i} n_i \hat{\mathbf{a}}_i^2 = \sum_{i=1}^{k} n_i (\bar{X}_{i.} - \bar{X}_{..})^2.$$

Ebben az egyszerű esetben minden projekciót pontosan leírunk. A Q_e kvadratikus alakot definiáló projekció $\mathbf{A}$ mátrixa, amellyel

$$Q_e = \mathbf{Y}^T \mathbf{A} \mathbf{Y},$$

a következő szimmetrikus, idempotens mátrix:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{A}_k \end{pmatrix},$$

ahol az $\mathbf{A}_i$ diagonális blokkok:

$$\mathbf{A}_i = \begin{pmatrix} 1-\frac{1}{n_i} & -\frac{1}{n_i} & \dots & -\frac{1}{n_i} \\ -\frac{1}{n_i} & 1-\frac{1}{n_i} & \dots & -\frac{1}{n_i} \\ \dots & \dots & \dots & \dots \\ -\frac{1}{n_i} & -\frac{1}{n_i} & \dots & 1-\frac{1}{n_i} \end{pmatrix} \quad (i = 1, \dots, k)$$

alakúak, és az $\mathbf{A}$ projekció az F altér $\mathbb{R}^n$-beli ortogonális kiegészítő alterére vetít. Rangja $n-k$.

A Q_a kvadratikus alakot definiáló

$$Q_a = \mathbf{Y}^T\mathbf{P}\mathbf{Y}$$

$\mathbf{P}$ projekció az $\mathbf{1} \in \mathbb{R}^n$ vektornak az F altérbeli ortogonális kiegészítő alterére vetít, rangja $k-1$. A $Q = Q_a + Q_e$ kvadratikus alaknak megfelelő projekció itt most nem $\mathbf{I}_n$, hanem

$$\mathbf{A} + \mathbf{P} = \mathbf{I}_n - \mathbf{1}\mathbf{1}^T,$$

amely az $\mathbf{1}$ vektor $\mathbb{R}^n$-beli ortogonális kiegészítő alterére vetít.

A gyakorlati alkalmazók terminológiájával élve: a fenti kvadratikus alakok segítségével a mintaelemek teljes mintaátlagtól vett eltéréseinek négyzetösszege (Q) felbomlik csoportok közötti (between, Q_a) ill. csoportokon belüli (within, Q_e) részre a következőképpen:

$$\begin{aligned} Q &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})]^2 = \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})^2 = \\ &= \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 = Q_a + Q_e, \end{aligned}$$

és ezt a felbontást a projekciók ismerete nélkül, viszonylag egyszerű számolással is megkaphattuk volna, miután a $[\dots]^2$ négyzetreemelésnél kihasználható, hogy a kétszeres szorzatok összege 0.

A fenti felbontásokat az alábbi ún. **ANOVA** (ANalysis Of VAriances) táblázatban foglaljuk össze.

A szóródás oka	Négyzetösszeg	Szabadsági fok	Empirikus szórásnégyzet
Csoportok között	$Q_a = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$	$k-1$	$s_a^2 = \frac{Q_a}{k-1}$
Csoportokon belül	$Q_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$	$n-k$	$s_e^2 = \frac{Q_e}{n-k}$
Teljes	$Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2$	$n-1$	-

A fenti modellben először az $m = 0$ hipotézist teszteljük. Ha ezt elutasítjuk (az összes várható érték nem 0, azaz van ún. főhatás), akkor a

$$H_0 : a_1 = \cdots = a_k = 0, \quad \text{tömören} \quad \mathbf{a} = \mathbf{0}$$

hipotézist vizsgáljuk. A *A legkisebb négyzetek módszere* paragrafusban leírtakhoz hasonlóan látható, hogy a likelihood-hányados statisztika a Q_a/Q_e hányados monoton fogyó függvénye (ez a hányados annál ``nagyobb'', minél ``nagyobb'' a csoportok közötti variancia a csoportokon belülihez képest, ami ellentmond H_0-nak).

A Q_e-ben szereplő lineáris kifejezések mindegyikének várható értéke 0, ugyanis a csoportokon belül a várható értékek egyenlőek a mintaátlag várható értékével:

$$\mathbb{E}(X_{ij} - \bar{X}_{i.}) = \mathbb{E}(X_{ij}) - \mathbb{E}(\bar{X}_{i.}) = a_i - a_i = 0, \quad (i = 1, \ldots, k)$$

akár igaz H_0, akár nem. Tehát az I.3.6. Állítás a. része értelmében $Q_e \sim \sigma^2\chi^2(n-k)$.

A Q_a-ben szereplő lineáris kifejezések várható értéke:

$$\mathbb{E}(\bar{X}_{i.} - \bar{X}_{..}) = \mathbb{E}(\bar{X}_{i.}) - \mathbb{E}(\bar{X}_{..}) = a_i - \frac{1}{n}\sum_{j=1}^{k} n_j a_j \qquad (i = 1, \ldots, k),$$

amely csak akkor lehet minden i-re 0, ha H_0 fennáll. Ezesetben szintén az I.3.6. Állítás a. része miatt $Q_a \sim \sigma^2\chi^2(k-1)$, és az előbbi állítás b. része alapján Q_e és Q_a függetlenek (megjegyezzük, hogy csak a null-hipotézis fennállása esetén lesz Q_a centrális χ^2-eloszlású).

Így bevezetve az

$$s_a^2 = \frac{Q_a}{k-1} \qquad \text{ill.} \qquad s_e^2 = \frac{Q_e}{n-k}$$

kifejezéseket, ezek azonos (σ^2) szórásúak, függetlenek, hányadosuk pedig H_0 fenállása esetén F-eloszlást követ $k-1$ ill. $n-k$ szabadsági fokkal:

$$F = \frac{s_a^2}{s_e^2} = \frac{Q_a}{Q_e} \cdot \frac{n-k}{k-1} \sim \mathcal{F}(k-1, n-k),$$

és ez az F is szigorúan monoton csökkenő függvénye a likelihood hányados statisztikának.

Megjegyezzük, hogy a a fenti F statisztika levezethető a likelihood hányados próba alkalmazása és a vetítések felírása nélkül is.

1.2. Többszempontos varianciaanalízis interakcióval

Itt is két különböző szempont alapján kialakított $k \cdot p$ csoportban végzünk megfigyeléseket, de cellánként több (mondjuk minden cellában n) megfigyelést. Az előző rész példájával élve: k féle technológiával p féle gépen gyártanak alkatrészeket és mérik azok szakítószilárdságát. Itt azonban feltételezzük, hogy a kétféle szempont hatása nem független, (nem mindegy, hogy melyik gépen melyik gyártási technológiát alkalmazzuk).

Jelölje X_{ijl} az első szempont alapján i-edik, a második szempont alapján pedig j-edik csoportban végzett l-edik megfigyelést, példánkban az i-edik technológiával a j-edik gépen gyártott l-edik termék szakítószilárdságát $(i = 1, \ldots, k; j = 1, \ldots, p; l = 1, \ldots, n)$.

Tehát összmintánk elemszáma kpn. A mintaelemek függetlenek és

$X_{ijl} \sim \mathcal{N}(m + a_i + b_j + c_{ij}, \sigma^2)$, azaz lineáris modellünk most a következő:

$$X_{ijl} = m + a_i + b_j + c_{ij} + \varepsilon_{ijl}, \qquad (i = 1, \ldots, k; j = 1, \ldots, p) \qquad (1)$$

???ahol az $\varepsilon_{ijl} \sim \mathcal{N}(0, \sigma^2)$ független valószí nűségi változók véletlen hibák. Itt a_i-k jelölik az egyik, b_j-k a másik tényező hatásait, c_{ij}-k pedig az interakciókat. Feltesszük (m-be való beolvasztással elérhető), hogy

$$\sum_{i=1}^{k} a_i = 0, \qquad \sum_{j=1}^{p} b_j = 0,$$

$$\sum_{i=1}^{k} c_{ij} = 0 \qquad (j = 1, \ldots, p) \quad \text{és}$$

$$\sum_{j=1}^{p} c_{ij} = 0 \qquad (i = 1, \ldots, k).$$

??? A $\mathbf{B}$ struktúramátrix alakja most:

$$\begin{pmatrix}
10 & 100 & 100000 \\
10 & 010 & 010000 \\
10 & 001 & 001000 \\
01 & 100 & 000100 \\
01 & 010 & 010010 \\
01 & 001 & 001001 \\
10 & 100 & 100000 \\
10 & 010 & 010000 \\
10 & 001 & 001000 \\
01 & 100 & 000100 \\
01 & 010 & 010010 \\
01 & 001 & 001001
\end{pmatrix} \qquad (1)$$

(Ebben a példában $k = 2$, $p = 3$, és $n = 2$.)

Az (7.4) modell az

$$\mathbf{Y} := (X_{111}, \ldots, X_{11n}, X_{121}, \ldots, X_{12n}, \ldots, X_{kp1}, \ldots, X_{kpn})^T$$
$$\underline{\varepsilon} := (\varepsilon_{111}, \ldots, \varepsilon_{11n}, \varepsilon_{121}, \ldots, \varepsilon_{12n}, \ldots, \varepsilon_{kp1}, \ldots, \varepsilon_{kpn})^T$$

és az

$$\underline{abc} = (a_1, \ldots, a_k, b_1, \ldots, b_p, c_{11}, \ldots, c_{kp})^T$$

jelölések, továbbá a segítségével az

$$\mathbf{Y} = \mathbf{B} \cdot \underline{abc} + \mathbf{1} \cdot m + \underline{\varepsilon}$$

lineáris modell alakját ölti, ahol $\mathbf{1} \in \mathbb{R}^{kpn}$ az azonosan 1 komponensű vektor, l. (7.5).

Jelölje F a $\mathbf{B}$ mátrix oszlopvektorai által kifeszített alteret, míg F_a, F_b, és F_c jelölje rendre az ekső k a következő p oszlop és az utolsó $k \cdot p$ oszlop által kifeszített alteret.

Jelölje $F\mathbf{B}$ mátrix oszlopvektorai által kifeszített alteret, míg F_a, F_b, és F_c jelölje rendre az eső k a következő p oszlop és az utolsó $k \cdot p$ oszlop által kifeszített alteret.

Vegyük észre, hogy $\mathbf{1} \in F_a$, $\mathbf{1} \in F_b$ és $\mathbf{1} \in F_c$. Jelölje F_{a1} illetve F_{b1} az $\mathbf{1}$ vektor ortogonális kiegészítőjét F_a-ban illetve F_b-ben, továbbá F_{cab} az F_a és F_b által generált altér ortogonális kiegészítőjét F_c-ben, valamint F_e az F ortogonális kiegészítőjét $\mathbb{R}^n$ -ben. Mivel az $\mathbf{1}$ vektort F_a F_b és F_c is tartalmazza: $\dim(F_{a1}) = k - 1$, $\dim(F_{b1}) = p - 1$, $\dim(F_{cab}) = kp - (k-1) - (p-1) + 1 = (k-1)(p-1)$, és $\dim(F_e) = kp(n-1)$. Jelölje az F_{a1}-ra, F_{b1}-re, F_{cab}-re és F_e-re vetítő projekciókat rendre $\mathbf{P}_a$, $\mathbf{P}_b$, $\mathbf{P}_c$ és $\mathbf{P}_e$. A fentiek miatt

$$\mathbf{I}_n = \mathbf{1}\mathbf{1}^T + \mathbf{P}_a + \mathbf{P}_b + \mathbf{P}_c + \mathbf{P}_e.$$

Először a legkisebb négyzetek módszerével megbecsüljük a paramétereket. Ehhez keressük a

$$\sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n} \varepsilon_{ijl}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{l=1}^{n} (X_{ijl} - m - a_i - b_j - c_{ij})^2 \qquad (1)$$

???kifejezés minimumát az $m, a_1, \ldots, a_k, b_1, \ldots, b_p$ paraméterekben az (7.1.2) kényszerfeltételek mellett.

Vezessünk be néhány jelölést:

$$\bar{X}_{i..} = \frac{1}{pn} \sum_{j=1}^{p} \sum_{l=1}^{n} X_{ijl} \quad (i = 1, \ldots, k)$$

$$\bar{X}_{.j.} = \frac{1}{kn} \sum_{i=1}^{k} \sum_{l=1}^{n} X_{ijl} \quad (j = 1, \ldots, p)$$

$$\bar{X}_{ij.} = \frac{1}{n} \sum_{l=1}^{n} X_{ijl} \qquad (i = 1, \ldots, k; j = 1, \ldots, p)$$

$$\bar{X}_{...} = \frac{1}{kpn} \sum_{i=1}^{k} \sum_{j=1}^{p} \sum_{l=1}^{n} X_{ijl}.$$

Ezekkel a paraméterek legkisebb négyzetes becslései:

$$\hat{m} = \bar{X}_{...},$$
$$\hat{a}_i = \bar{X}_{i..} - \bar{X}_{...} \qquad (i = 1, \ldots, k),$$
$$\hat{b}_j = \bar{X}_{.j.} - \bar{X}_{...} \qquad (j = 1, \ldots, p),$$
$$\hat{c}_{ij} = \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...} \qquad (i = 1, \ldots, k; \quad j = 1, \ldots, p),$$

az (7.6) kifejezés minimuma pedig

$$Q_e = \sum_{i=1}^{k} \sum_{j=1}^{p} \sum_{l=1}^{n} (X_{ijl} - \hat{m} - \hat{a}_i - \hat{b}_j - \hat{c}_{ij})^2$$

lesz.

Ha a $\mathbf{P}_a$, $\mathbf{P}_b$, $\mathbf{P}_c$ és $\mathbf{P}_e$ projekcóknak rendre az $\mathbf{Y}$ vektorral képzett Q_a, Q_b, Q_c és Q_e kvadratikus formák felelnek meg, akkor igaz a

$$Q = Q_a + Q_b + Q_c + Q_e \qquad (1)$$

???varianciafelbontás, ahol a mintaelemek teljes mintaátlagtól vett eltéréseinek négyzetösszegét (Q) felbontjuk a következő ANOVA-táblázat szerint:

A szóródás oka	Négyzetösszeg	Szabadsági fok	Ei szór
a-hatások	$Q_a = pn \sum_{i=1}^{k} (\bar{X}_{i..} - \bar{X}_{...})^2$	$k-1$	s
b-hatások	$Q_b = kn \sum_{j=1}^{p} (\bar{X}_{.j.} - \bar{X}_{...})^2$	$p-1$	s
ab-interakció	$Q_c = n \sum_{i=1}^{k} \sum_{j=1}^{p} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2$	$(k-1)(p-1)$	s_c^2 =
Véletlen hiba	$Q_e = \sum_{i=1}^{k} \sum_{j=1}^{p} \sum_{l=1}^{n} (X_{ijl} - \bar{X}_{ij.})^2$	$kp(n-1)$	s_e^2
Teljes	$Q = \sum_{i=1}^{k} \sum_{j=1}^{p} \sum_{l=1}^{n} (X_{ijl} - \bar{X}_{...})^2$	$kpn-1$	

Miután az $m = 0$ hipotézist elutasítottuk, a fenti modellben háromféle null-hipotézist akarunk vizsgálni, az egyik és a másik szempont szerint megnézni, hogy a csoporthatások azonosak-e, továbbá, hogy interakciók léteznek-e. Az első tényező hatására vonatkozóan tehát vizsgáljuk a

$$H_{0a} : a_1 = a_2 = \cdots = a_k = 0$$

hipotézist (példánkban azt, hogy a gyártási technológia nincs hatással az alkatrész szakítószilárdságára). Ezzel párhuzamosan a második tényező hatására vonatkozóan vizsgáljuk a

$$H_{0b}: b_1 = b_2 = \cdots = b_p = 0$$

hipotézist (példánkban azt, hogy a gyártó gép megválasztása nincs hatással az alkatrész szakítószilárdságára). Továbbá az interakciókra vonatkozóan vizsgáljuk a

$$H_{0ab}: c_{ij} = 0, \quad (i = 1, \ldots, k; j = 1, \ldots, p)$$

hipotézist (példánkban azt, hogy a gyártó gép nem hat a gyártási technológiára).

A Q_e-ben szereplő lineáris kifejezések mindegyikének várható értéke 0. A Q_a-ban szereplő lineáris kifejezések várható értéke csak akkor lehet minden i-re 0, ha H_{0a} fennáll. Hasonlóan, a Q_b-ben szereplő lineáris kifejezések várható értéke csak akkor lehet minden j-re 0, ha H_{0b} fennáll. A Q_c-ben szereplő lineáris kifejezések várható értéke pedig csak akkor lehet minden (i, j)-re 0, ha H_{0ab} fennáll.

Az (7.7) felbontásban a kvadratikus alakok rangja itt is összeadódik:

$$kpn - 1 = (k - 1) + (p - 1) + (k - 1)(p - 1) + kp(n - 1).$$

Így igazak az alábbi állítások:

* e. $Q_e/\sigma^2 \sim \chi^2(kp(n-1))$, akár fennállnak a nullhipotézisek, akár nem.

* a. H_{0a} fennállása esetén $Q_a/\sigma^2 \sim \chi^2(k-1)$ és független Q_e-től.

* b. H_{0b} fennállása esetén $Q_b/\sigma^2 \sim \chi^2(p-1)$ és független Q_e-től.

* c. H_{0ab} fennállása esetén $Q_c/\sigma^2 \sim \chi^2((k-1)(p-1))$ és független Q_e-től.

Ezért nullhipotéziseink vizsgálatára a következő statisztikákat használhatjuk. Először a kölcsönhatást, vagyis a H_{0ab} hipotézist vizsgáljuk. Ennek fennálása esetén

$$F_{ab} = \frac{s_c^2}{s_e^2} \sim \mathcal{F}((k-1)(p-1), kp(n-1)),$$

azaz, ha a fenti F_{ab} statisztika értéke nagyobb vagy egyenlő, mint az $\mathcal{F}((k-1)(p-1), kp(n-1))$-eloszlás $(1-\alpha)$-kvantilise, akkor H_{0ab}-t $1-\alpha$ szinten elutasítjuk, vagyis elfogadjuk, hogy van kölcsönhatás a két szempont között, legalábbis bizonyos (i, j) indexpárokra. Ebben az esetben a H_{0a}, H_{0b} hipotéziseket nincs értelme vizsgálni.

Amennyiben H_{0ab}-t elfogadjuk, akkor a H_{0a} és H_{0b} hipotézisektől függetlenül $Q_c \sim \chi^2((k-1)(p-1))$ és független Q_e -től. Így ezeket összeadhatjuk, és a σ^2 szórásnégyzetre most már a $(k-1)(p-1) + kp(n-1) = kpn - k - p + 1$ szabadságfokú

$$\tilde{s}_e^2 = \frac{Q_c + Q_e}{kpn - k - p + 1}$$

becslést kapjuk.

Ezekután a H_{0a} hipotézis vizsgálatára az

$$F_a = \frac{s_a^2}{\tilde{s}_e^2}$$

statisztikát használjuk, amely H_{0a} fennállása esetén $\mathcal{F}(p-1, kpn-k-p+1)$-eloszlást követ. Hasonlóan, a H_{0b} hipotézis vizsgálatára az

$$F_b = \frac{s_b^2}{\tilde{s}_e^2}$$

statisztikát használjuk, amely H_{0b} fennállása esetén $\mathcal{F}(k-1, kpn-k-p+1)$-eloszlású. Ha a H_{0a} vagy/és H_{0b} hipotézist elutasítjuk, akkor az előző pontokéhoz hasonlóan vizsgálhatjuk az a - vagy/és b -hatásokat ill. azok különbségét.

2. Feladatok

(i) Tekintsük az egyszempontos varianciaanalízis modelljében a paraméterek legkisebb négyzetek módszerével kapott becsléseit.

(a) Mutassuk meg, hogy ezek maximum likelihood becslések!

(b) * Számoljuk ki ezeket a becsléseket Lagrange-multiplikátor módszerrel!

Tipp: Lásd 4. feladat (a) pontját.

Válasz: Az Útmutató alapján nyilvánvaló.

(ii) Tekintsük az egyszempontos varianciaanalízis csoporthatás-vizsgálatát, ahol $Q_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2$ és $Q_a = \sum_{i=1}^{k} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$.

(a) Mutassuk meg, hogy $Q_e/\sigma^2 \sim \chi^2(n-k)$!

(b) Igazoljuk, hogy H_0 teljesülése mellett $Q_a/\sigma^2 \sim \chi^2(k-1)$, de ha H_0 nem teljesül, Q_a nem χ^2 eloszlású!

(c) Adjuk meg H_0 mellett Q_a és Q_e várható értékét és szórásnégyzetét!

Tipp:

(a) A Fisher- Cochran-tételhez fűzött megjegyzeés a szabadságfokok heurisztikus számolásáról alapján itt a szabadságfok $n-k$, mert az n valószínűségi változót tartalmazo kvadratikus alakban k becsult paraméter van.

(b) Ha H_0 fennáll akkor (a) az pontbeli eredmény és Fisher- Cochran tétel közvetlen következménye, míg ha nem teljesül, akkor Q_a nem 0 várható értékű valószínűségi változók négyzetének összege.

(c) A alapján számolunk

Válasz:

(a) Az Útmutató alapjaán nyilvanvaló.

(b) Az Útmutató alapjaán nyilvanvaló.

(c) $\mathbb{E}Q_a = (k-1)/\sigma^2\mathbb{D}^2Q_a = 2(k-1)/\sigma^2\mathbb{E}Q_e = (n-k)/\sigma^2\mathbb{D}^2Q_a = 2(n-k)/\sigma^2$

(iii) Adjunk maximum likelihood becslést σ^2-re az egyszempontos varianciaanalízis modelljében! Torzítatlan lesz-e becslésünk?

Tipp: Az előző feladatban szereplő Q_a és Q_e független kvadratikus alakok alpajan számoljunk.

Válasz: $\hat{\sigma}^2 = (Q_a + Q_e)/n$, ami torzított becslés.

(iiii) Mutassuk meg, hogy az egyszempontos varianciaanalízis csoporthatás-vizsgálata

(a) likelihood-hányados próba!

(b) a kétmintás t-próba általánosítása több mintára!

Tipp: Valójában F-próba.

Válasz:

(iiiii) Tekintsük az (X, Y) vektorváltozót, ahol Y normális eloszlású, Y pedig véges sok értéket felvevő diszkrét változó. Csoportosítsuk a mintát az X értékei szerint. Alkalmazhatjuk-e az egyszempontos varianciaanalízist és Y függetlenségének tesztelésére?

Tipp: Vizsgáljuk meg milyen hipotézist tesztel a varianciaanalízis!

Válasz: Csak a várható értékek azonos voltát teszteli, nem a függetlenséget.

(iiiiii) Tekintsük a kovarianciaanalízis modelljét és ebben egy n elemű mintát egy előre tervezett hatás és egy kísérő változó esetén: $Y_i = b_i a + d_i c + \varepsilon_i$, ahol a, c paraméterek, b_i-k tervezett hatások, d_i-k kísérő változók, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$ független hibák.

(a) Adjunk becslést a paraméterekre a legkisebb négyzetek módszerével!

(b) Konstruáljunk likelihood-hányados próbát a $H_0 : c = 0$ hipotézis tesztelésére!

Tipp: Vegyük észre, hogy a feladat független a kovarianciaanalízis modelljétől, egyszerű kétváltozós lineáris modellről van szó.

(a) A

$$\sum_{i=1}^{n} Y_i b_i = a \sum_{i=1}^{n} b_i^2 + c \sum_{i=1}^{n} b_i d_i \quad \sum_{i=1}^{n} Y_i d_i = a \sum_{i=1}^{n} b_i d_i + c \sum_{i=1}^{n} d_i^2$$

normálegyenletet kell megoldani.

(b) A $\lambda(y_1, \ldots, y_n)$ próbafüggvény

$$\lambda(y_1, \ldots, y_n) = \left(\frac{\sum_{i=1}^{n} (y_i - \hat{a} b_i - \hat{c} d_i)^2}{\sum_{i=1}^{n} (y_i - \hat{a} b_i)^2} \right)^{n/2}$$

alakú lesz (l. 6.4 feladat (e) pontját)

Válasz: Az Útmutatók alapján nyilvánvaló.

(iiiiiii) Tekintsünk egy mintát, amely teljesíti az alábbi modellt:

$Y_{i,j} = a x_{i,j} + c_i + \varepsilon_{i,j}$,

$i = 1, \ldots, r$, $j = 1, \ldots, n_i$, ahol $c_1, \ldots, c_r$ és a paraméterek, $x_{i,j}$-k (determinisztikus) kísérő változók, $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ független hibák.

(a) Adjunk becslést a paraméterekre a legkisebb négyzetek módszerével!

(b) Mutassuk meg, hogy a fenti modell a kovarianciaanalízis egy modellje.

Tipp:

Válasz:

3. Tesztek

8. fejezet - Kontingenciatáblák elemzése: diszkriminanciaanalízis, korrespondenciaanalízis, információelmélet

1. Elméleti háttér

1.1. Diszkriminanciaanalízis

Jelen feladatban objektumokat szeretnénk a rajtuk végrehajtott többdimenziós megfigyelések alapján előre adott osztályokba besorolni. Például pácienseket klinikai- vagy pszichiátriai teszteredményeik alapján szeretnénk beteg- ill. kontrollcsoportba, vagy többféle betegcsoportba besorolni; vagy egy új egyedet mért értékei alapján valamely ismert fajba akarunk besorolni. A módszert úgy kell elképzelni, hogy első lépésben egy ún. tanuló-algoritmust hajtunk végre. Az objektumoknak kezdetben létezik egy osztálybasorolása. Ezt úgy adjuk meg, hogy a megfigyelt többdimenziós, folytonos eloszlású valószínűségi változó komponensein kívül bevezetünk egy, az osztálybatartozásra jellemző diszkrét valószínűségi változót, mely annyiféle értéket vesz fel, ahány osztály van; ez utóbbit egy szakértő a mérésektől függetlenül állapítja meg. Az egyes osztályok adatai alapján diszkrimináló algoritmust készítünk, és megnézzük, hogy az algoritmus szerint melyik osztályba kerülnének eredeti objektumaink. Amennyiben a téves osztálybasorolások száma nem túl nagy, úgy tekintjük, hogy az algoritmus által adott diszkrimináló függvény a továbbiakban is használható az adott csoportok elkülönítésére. A tényleges osztályozás figyelembevételével bevezetjük a következőket. Jelölje k az osztályok számát, továbbá

a. jelölje az egyes osztályokhoz tartozó p-dimenziós mintaelemek sűrűségfüggvényét $f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})$ (abszolút folytonos eloszlásokat feltételezünk);

b. jelölje $\pi_1, \ldots, \pi_k$ az egyes osztályok a priori valószínűségeit;

Az a.-beli sűrűségeket osztályonként becsüljük a mintatákból, a b.-beli a priori valószínűségek pedig lehetnek az egyes osztályok relatív gyakoriságai. Így visszük bele ``tudásunkat'' az alábbi algoritmusba. Ha már adva lenne a p-dimenziós mintatér egy $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_k$ partíciója, akkor a $\mathbf{x} \in \mathcal{X}$ mintaelemet akkor soroljuk a j-edik osztályba, ha $\mathbf{x} \in \mathcal{X}_j$. A cél az, hogy a legkisebb veszteséggel járó partíciót megkeressük. Ehhez jelölje $r_{ij} \geq 0$ $(i, j = 1, \ldots k)$ azt a veszteséget, ami akkor keletkezik, ha egy i-edik osztálybelit a j-edik osztályba sorolunk (a veszteségek nem feltétlenül szimmetrikusak, de feltesszük, hogy $r_{ii} = 0$), és legyen L_i az i-edik osztálybeliek besorolásának átlagos vesztesége (rizikója):

$$L_i = \int_{\mathcal{X}_1} r_{i1} f_i(\mathbf{x})\, d\mathbf{x} + \cdots + \int_{\mathcal{X}_k} r_{ik} f_i(\mathbf{x})\, d\mathbf{x}, \quad (i = 1, \ldots, k),$$

ahol összegeztük a veszteségeket azokra az esetekre, mikor az i-edik osztálybelit az $1., \ldots, k.$ osztályba soroltuk. Most nem az egyes L_i veszteségeket, hanem az

$$L = \sum_{i=1}^{k} \pi_i L_i$$

átlagos Bayes-féle veszteséget (rizikót) minimalizáljuk.

$$L = \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} \int_{\mathcal{X}_j} r_{ij} f_i(\mathbf{x})\, d\mathbf{x} = \sum_{j=1}^{k} \int_{\mathcal{X}_j} \sum_{i=1}^{k} \pi_i r_{ij} f_i(\mathbf{x})\, d\mathbf{x} = -\sum_{j=1}^{k} \int_{\mathcal{X}_j} S_j(\mathbf{x})\, d\mathbf{x},$$

ahol az

$$S_j(\mathbf{x}) = -[\pi_1 r_{1j} f_1(\mathbf{x}) + \cdots + \pi_k r_{kj} f_k(\mathbf{x})]$$

$(j = 1, \ldots, k)$edik *diszkrimináló informánsnak* nevezzük, és argumentumában az $\mathbf{x}$ mintaelem szerepel S_j. A negatív előjel miatt -k növekedése az átlagos veszteség csökkenését eredményezi, azaz a

$$\sum_{j=1}^{k} \int_{\mathcal{X}_j} S_j(\mathbf{x})\, d\mathbf{x}$$

kifejezést szeretnénk maximalizálni a mintatér összes lehetséges mérhető partícióján.

Célszerűnek tűnik tehát egy $\mathbf{x}$ mért értékekkel rendelkező objektumot abba az osztályba sorolni, melyre diszkrimináló informánsa a legnagyobb értéket veszi fel. Ennek az eljárásnak a jogosságát a következő tétel biztosítja.

8.1.1.1. Tétel. *Legyen az $\mathcal{X}$ mintatér $\mathcal{X}_1^* \cup \cdots \cup \mathcal{X}_k^*$ partíciója olyan, hogy $\mathbf{x} \in \mathcal{X}_j^*$-ból $S_j(\mathbf{x}) \geq S_i(\mathbf{x})$ következik az összes $i \neq j$ indexekre $(j = 1, \ldots, k)$. Akkor az $\mathcal{X}_1^*, \ldots, \mathcal{X}_k^*$ osztályozással az L átlagos veszteség minimális lesz.*

A tétel állítása az alábbi lemma közvetlen következménye.

8.1.1.2. Lemma. *Legyenek $g_1, \ldots, g_k$ $\mathbb{R}^p$-n értelmezett valós függvények. Legyen $\mathbb{R}^p = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_k$ a p-dimenziós euklideszi tér egy partíciója. Tegyük fel, hogy az $\mathbb{R}^p = \mathcal{X}_1^* \cup \cdots \cup \mathcal{X}_n^*$ partícióra teljesülnek a*

$$g_i(\mathbf{x}) \geq g_j(\mathbf{x}), \quad \text{ha} \quad \mathbf{x} \in \mathcal{X}_i^* \quad \forall j \neq i; \qquad i = 1, \ldots, k$$

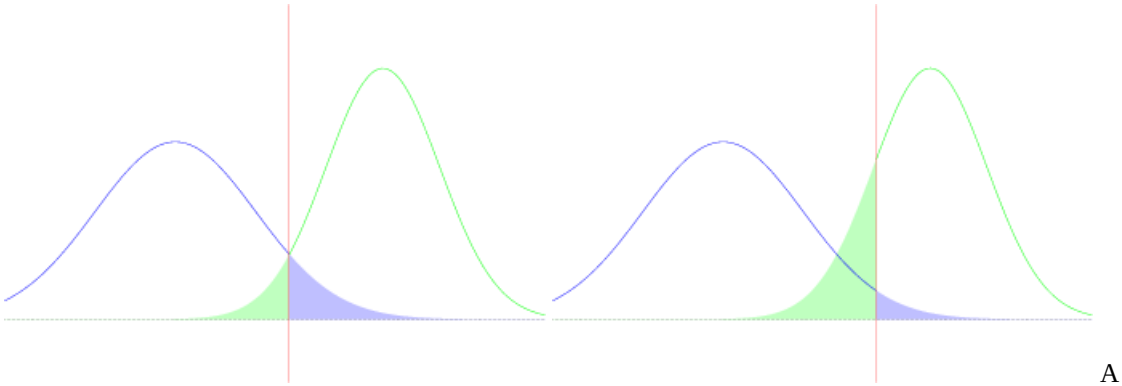
egyenlőtlenségek. Ekkor

$$\sum_{i=1}^{k} \int_{\mathcal{X}_i^*} g_i(\mathbf{x}) \geq \sum_{i=1}^{k} \int_{\mathcal{X}_i} g_i(\mathbf{x}). \qquad (1)$$

A Lemma bizonyítását egy ábra szemlélteti (ld. [5] 306. o. 2.1. ábra). Jelölje $I_A(\mathbf{x})$ az $A \subset \mathbb{R}^p$ halmaz indikátorfüggvényét! A (8.1)-beli egyenlőtlenségek miatt

$$\sum_{i=1}^{k} I_{\mathcal{X}_i^*}(\mathbf{x}) g_i(\mathbf{x}) = \max_{i \in \{1,\ldots,k\}} g_i(\mathbf{x}) \geq \sum_{i=1}^{k} I_{\mathcal{X}_i}(\mathbf{x}) g_i(\mathbf{x}). \qquad (1)$$

???



A mintatér felosztása diszkrimináló informánsokkal

A (8.1) egyenlőtlenség (8.2) integrálásával adódik.

Megjegyezzük, hogy az alkalmazásokban az optimális partíciót a (2.4) egyenlőtlenségek segítségével definiáljuk. A partíció nem egyértelmű, ha van olyan $i \neq j$ indexpár, hogy $g_i(\mathbf{x}) = g_j(\mathbf{x})$ egy nem-0 mértékű halmazon. Ilyenkor ezt a halmazt tetszőlegesen oszthatjuk fel $\mathcal{X}_i^*$ és $\mathcal{X}_j^*$ között.

A $g_i(\mathbf{x}) = S_i(\mathbf{x})$ helyettesítéssel adódik a tétel állítása.

Most néhány egyszerűsítő feltevést vezetünk be. Ha az r_{ij} veszteségekre nincsenek adataink, és az összes téves besorolást egyformán akarjuk büntetni, akkor jobb híján az $r_{ij} = 1(i \neq j)$ és $r_{ii} = 0$ választással élünk. Ezzel

$$S_j(\mathbf{x}) = -\sum_{i=1}^{k} \pi_i r_{ij} f_i(\mathbf{x}) = -\sum_{i \neq j} \pi_i f_i(\mathbf{x}) = -\sum_{i=1}^{k} \pi_i f_i(\mathbf{x}) + \pi_j f_j(\mathbf{x}) = \pi_j f_j(\mathbf{x}) + c,$$

ahol a c konstans nem függ j-től. Valójában tehát az $\mathbf{x}$ mért értékekkel rendelkező objektumot az l. osztályba soroljuk, ha

$$\pi_l f_l(\mathbf{x}) = \max_{j \in \{1,\dots,k\}} \pi_j f_j(\mathbf{x}).$$

Tegyük fel, hogy az egyes osztályoknak különböző paraméterű, p-dimenziós normális eloszlások felelnek meg. Azaz, ha $\mathbf{X} \in \mathcal{N}_p(\mathbf{m}_j, \mathbf{C}_j)$, akkor

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{C}_j|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_j)^T \mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{m}_j)}.$$

Tekintsük az osztálybasorolás alapját képező $\pi_j f_j(\mathbf{x})$ mennyiségek természetes alapú logaritmusát, a logaritmus monoton transzformáció lévén ez ugyanarra a j-re lesz maximális, mint az eredeti kifejezés, sőt az összes j-re közös $\ln \frac{1}{(2\pi)^{p/2}}$-től is eltekinthetünk. Az így kapott módosított j-edik diszkrimináló informánst S'_j-vel jelöljük, és alakja miatt *kvadratikus diszkriminancia szkór*nak is szokás nevezni:

$$S'_j(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{C}_j| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_j)^T \mathbf{C}_j^{-1}(\mathbf{x}-\mathbf{m}_j) + \ln \pi_j.$$

Ha a kovarianciamátrixok azonosak: $\mathbf{C}_1 = \cdots = \mathbf{C}_k = \mathbf{C}$, akkor $S'_j(\mathbf{x})$-ből a j-től független $-\frac{1}{2}\ln|\mathbf{C}|$ és a kvadratikus alak kifejtésében fellépő, j-től ugyancsak független $-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}$ rész elhagyható, a maradék pedig $\mathbf{x}$ lineáris függvényeként írható. Ezt nevezzük *lineáris informáns*nak:

$$S''_j(\mathbf{x}) = \mathbf{m}_j^T \mathbf{C}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{m}_j^T \mathbf{C}^{-1}\mathbf{m}_j + \ln \pi_j. \qquad (1)$$

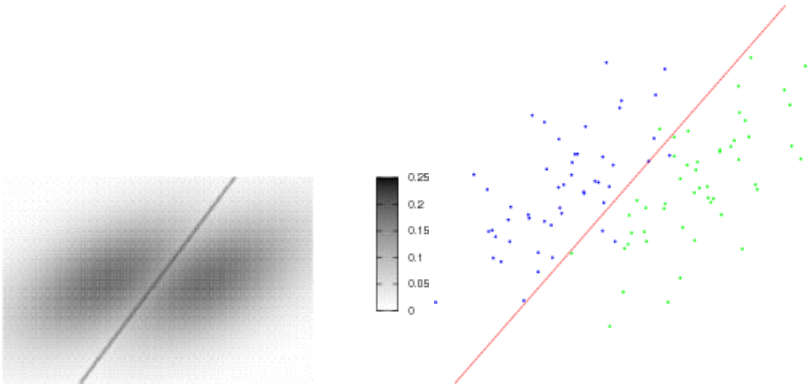
???

Eljárásunk tehát a következő: minden osztályra kiszámoljuk az $S''_j(\mathbf{x})$ értékét $(j = 1, \dots k)$, és objektumunkat abba az osztályba soroljuk, amelyikre az $S''_j(\mathbf{x})$ lineáris informáns értéke a legnagyobb. A 115 Tétel garantálja, hogy ekkor átlagos veszteségünk minimális lesz. Amennyiben csak két osztályunk van, objektumunkat az $\mathbf{x}$ megfigyelés alapján az első osztályba soroljuk, ha $S''_1(\mathbf{x}) \geq S''_2(\mathbf{x})$, különben a másodikba. Azaz az $S''_1(\mathbf{x}) - S''_2(\mathbf{x})$ különbség előjele fogja eldönteni az osztálybatartozást. De

$$S''_1(\mathbf{x}) - S''_2(\mathbf{x}) = L(\mathbf{x}) - c,$$

ahol (8.3) alapján

$$L(\mathbf{x}) = (\mathbf{m}_1^T - \mathbf{m}_2^T)\mathbf{C}^{-1}\mathbf{x} \quad \text{és}$$
$$c = \frac{1}{2}(\mathbf{m}_1^T \mathbf{C}^{-1}\mathbf{m}_1 - \mathbf{m}_2^T \mathbf{C}^{-1}\mathbf{m}_2) - \ln \pi_1 + \ln \pi_2.$$



diszkriminanciafüggvény 2 dimenzióban

Elméleti és empirikus

A fenti $L(\mathbf{x})$-et *Fisher-féle diszkriminancia függvény*nek is szokták nevezni, és ennek alapján döntjük el az osztálybatartozást: ha $L(\mathbf{x}) \geq c$, akkor objektumunkat az első, ha pedig $L(\mathbf{x}) < c$, akkor a második osztályba soroljuk. Az $L(\mathbf{x})$ lineáris kifejezésben az egyes x_i változók együtthatói egyfajta súlyokként is szolgálnak, azok a változók fejtik ki a legerősebb hatást a két csoport diszkriminálásában, amely a legnagyobb súllyal szerepelnek. Ha az átlagos veszteséget akarjuk minimalizálni, normális eloszlású minták esetén a fenti eljárás keresztülvihető az egyes osztályokban számolt empirikus kovarianciamátrixokkal és az osztályok relatív gyakoriságaival becsült apriori valószínűségek segítségével. Létezhetnek azonban ún. látens osztályok (pl. egy újfajta betegség, újfajta faj), ami ronthat a módszer alkalmazhatóságán. Szükség van ezért különféle hipotézisvizsgálatokra. Pl. két osztály esetén, az első osztályba való besorolhatóság a

$$T_1 = \frac{[(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{X} - \mathbf{m}_1)]^2}{(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)} \sim \chi^2(1) \qquad (1)$$

???statisztikával, míg a második osztályba való besorolhatóság a

$$T_2 = \frac{[(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{X} - \mathbf{m}_2)]^2}{(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)} \sim \chi^2(1) \qquad (1)$$

???statisztikával tesztelhető, ugyanis ha $\mathbf{X} \sim \mathcal{N}_p(\mathbf{m}_j, \mathbf{C})$, akkor $\mathbf{C}^{-1}(\mathbf{X} - \mathbf{m}_j) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{C}^{-1})$, $(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{X} - \mathbf{m}_j) \sim \mathcal{N}_p(\mathbf{0}, (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{C}^{-1}(\mathbf{m}_2 - \mathbf{m}_1))$, s utóbbinak standardizáltja lesz a (8.4)- ill. (8.5)-beli T_j statisztika $(j = 1, 2)$.

Ha mind T_1, mind T_2 szignifikánsan nagyobb az 1-paraméterű χ^2-eloszlás adott (pl. 95%-os) kvantilisénél, akkor egy látens harmadik osztály jelenlétére gyanakodhatunk.

Számítsuk most ki két p-dimenziós normális eloszlású, azonos $\mathbf{C}$ kovarianciamátrixú minta esetén a helytelen osztálybasorolások valószínűségeit! Az egyszerűség kedvéért legyen most két egyforma népességű mintánk, azaz az apriori valószínűségekre a $\pi_1 = \pi_2 = 1/2$ feltételezéssel élünk. A számolást nem részletezzük, ebben az esetben a veégeredmény meglepően egyszerű:

Legyen

$$\sigma^2 = (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \qquad (1)$$

???

Ekkor mindkét típusú hibás osztálybasorolás valószínűsége:

$$\mathbb{P} = 1 - \Phi\left(\frac{1}{2}\sigma\right).$$

Ez nem meglepő, hiszen a (8.6) szerint σ annál nagyobb, minél távolabb vannak egymástól a két csoport „standardizált" várható értékei. A diszkrimináló információkban szereplő paramétereket a mintából becsüljük, minél több a paraméter, annál pontatlanabb az egyes paraméterek becslése; azt is mondhatjuk, hogy a paraméterek a konkrét mintához vannak adaptálva. Ezért, ha az eljárás rizikóját a nem megfelelő osztályba

sorolt egyedek száma alapján az alább ismertetendő módon becsüljük, a valódi veszteségfüggvénynél kisebb torzított becslést kapunk. E torzítás kivédésére alkalmazzák az ún. *cross-validation (kereszt-kiértékelés)* módszert: a paramétereket a minta egy része (60% a szokásos hányad) alapján becsüljük, míg az osztályozás minőségét a paraméterbecslésben fel nem használt mintaelemekkel teszteljük (40%). A torzítás csökkentésére Tukey [] javasolt egy szellemes - általa jackknife-nak (bicskának) nevezett, nagy számolásigényű módszert. Ezt a módszert az **algoritmikus modellek fejezetben ismertetjuk**

1.2. Korrespondanciaanalízis

Ebben és a következő paragrafusban minden eloszlás diszkrét és véges, ezt a továbbiakban külön nem említjük.

A korrespondanciaanalízis kategórikus változók közti kapcsolatok elemzésére szolgál a változó-kategóriák metrikus megjelenítése alapján. Kategórikus, más néven kvalitatív változó alatt olyan diszkrét eloszlású valószínűségi változót értünk, amely véges sok értéket vesz fel, és az értékek általában nem nagyságrendet tükröznek, hanem csak a változó lehetséges értékeit kódolják (pl. a hajszín változó szőke, barna, fekete, vörös értékei az 1,2,3,4 számokkal kódolhatók). A Tananyagban csak két kategórikus változót vizsgálunk, az adatok kontingenciatábla (gyakoriság- vagy rekatív gyakoriságtábla) formájában vannak megadva. A probléma a következő: az X és Y diszkrét valószínűségi változók n ill. m különböző kategóriát tartalmaznak, az egyszerűség kedvéért jelölje értékkészletüket az $\{1,2,\ldots,n\}$ ill. az $\{1,2,\ldots,m\}$ halmaz. X és Y nem függetlenek, értékeiket nem specifikáljuk, célunk éppen az értékek alkalmas megválasztása lesz. Egy közös megfigyelésükre vonatkozó minta alapján adva van egy $n \times m$ -es *kontingenciatábla* az f_{ij} ún. *cellagyakoriságok*kal (f_{ij} az X változó i-edik és az Y változó j-edik kategóriájába eső megfigyelések számát jelenti). Legyen $N = \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij}$ a megfigyelések száma, ezzel callánként leosztva az

$$r_{ij} = \frac{f_{ij}}{N} \qquad (i = 1,\ldots,n;\ j = 1,\ldots,m)$$

relatív gyakoriságokhoz jutunk. Ezeket tekinthetjük a két diszkrét eloszlású valószínűségi változó (az egyik n, a másik m különböző értéket vesz fel) együttes eloszlásának, és R-rel jelöljük. Ugyancsak $\mathbb{R}$ jelöli az r_{ij} számok alkotta $n \times m$-es mátrixot. Jelölje

$$p_i = r_{i.} \quad (i = 1,\ldots,n) \quad \text{ill.} \quad q_j = r_{.j} \quad (j = 1,\ldots,m)$$

a peremeloszlásokat (azaz az egyes kategóriák valószínűségeit), ezeket röviden P-nek ill. Q-nak fogjuk nevezni, az elemeiket fődiagonálisként tartalmazó $n \times n$-es ill. $m \times m$-es diagonális mátrixokat pedig $\mathbf{P}$ ill. $\mathbf{Q}$ jelöli. Célunk a kontingenciatáblának valamilyen alacsonyabb rangú táblával való közelíése. Ehhez a kanonikus korrelációanalízisnél leírtakhoz hasonlóan keresünk olyan, értékeiket a P- ill. Q-eloszlás valószínűségei szerint felvevő, egységszórású, páronként korrelálatlan valószínűségi változókat, ún. faktorokat úgy, hogy a megegyező indexű faktorok korrelációja maximális legyen. Ilyen módon a kontingenciatábla előáll a faktor valószínűségi változók értékei (szkórok) diádszorzatainak súlyozott összegeként. A legnagyobb súlyok közül bizonyos számút megtartva a kontingenciatábla egy alacsonyabb rangú közelítését kapjuk.

Mi csak a 2 rangú közelítéssel foglalkozunk, ami visszavezethető a *Rényi-féle maximálkorreláció* feladatára: adott két kategórikus változó együttes eloszlása (együttes relatív gyakorisága, azaz egy $n \times m$ gyakoriságtábal). Keressük azokat az $\underline{\alpha}$ és $\underline{\beta}$ valós számértékű véletlen vektorokat, amelyek marginális eloszlásai megegyeznek az adott kontingencia táblából számolt marginális eloszlásokkal, és az együttes eloszlás alapján számított korrelációjuk maximális. A marginális eloszlás általános és egzakt definícióját l. a következő paragrafusban (117).

Látni fogjuk, hogy ezen véletlen vektorok együttes eloszlása az eredeti kontingenciatábla 2 rangú közelítése. Ha az itt tárgyalt módszerrel magasabb rangú közelítéseket is számulunk, akkor ezek "együttes eloszlásában" NEGETÍV valószínűségek is előfordulhatnak. A feladat pontos leírásához jelölje α_l ill. β_l a sor- ill. oszlop-faktorokat ($l = 1,2\ldots,\min\{n,m\}$). A faktorok szórására és korrelálatlanságára tett feltevések azt jelentik, hogy

$$\mathbb{E}_P \alpha_l \alpha_{l'} = \sum_{i=1}^{n} \alpha_l(i)\alpha_{l'}(i)p_i = \delta_{ll'} \qquad (l, l' = 2 \ldots, \min\{n, m\}),$$

$$\mathbb{E}_Q \beta_l \beta_{l'} = \sum_{j=1}^{m} \beta_l(j)\beta_{l'}(j)q_j = \delta_{ll'} \qquad (l, l' = 2 \ldots, \min\{n, m\}),$$

ahol $\delta_{ll'}$ a Kronecker-delta, $\alpha_l(i)$ ill. $\beta_l(j)$ pedig az α_l ill. β_l valószínűségi változók p_i ill. q_j valószínűséggel felvett értékei. A cél az α_l, β_l párok egymásutáni meghatározása oly módon, hogy az előzőekkel való korrelálatlansági feltételek mellett

$$\mathbb{E}_R \alpha_l \beta_l = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_l(i)\beta_l(j)r_{ij} \qquad (l = 1 \ldots, \min\{n, m\}) \qquad (1)$$

???maximális legyen. A korrespondanciafaktorok $l > 1$ esetén egységszórásúak, később pedig látni fogjuk, hogy várható értékük 0, ezért (8.7) egyben az azonos indexű faktorpárok közti korrelációt is jelenti. Az $l = 1$ esetben adódó faktorpár tagjaitól nem követeljük meg, hogy 0 várható értékűek és 1 szórásúak legyenek, de (8.7) ezesetben is maximális. A megoldáshoz egy α, β változópárt a következő transzformációnak vetünk alá:

$$x(i) := \sqrt{p_i}\alpha(i), \quad (i = 1, \ldots, n),$$
$$y(j) := \sqrt{q_j}\beta(j), \quad (j = 1, \ldots, m).$$

Jelölje $\mathbf{x} = (x(1), \ldots, x(n))^T$ ill. $\mathbf{y} = (y(1), \ldots, y(m))^T$ a fenti komponensekből álló vektort. Amennyiben $\underline{\alpha}$ ill. $\underline{\beta}$ jelöli az α ill. β valószínűségi változók felvett értékeiből álló n- ill. m-dimenziós vektort,

$$\underline{\alpha} = \mathbf{P}^{-1/2}\mathbf{x} \quad \text{ill.} \quad \underline{\beta} = \mathbf{Q}^{-1/2}\mathbf{y}.$$

$$\underline{\alpha} = \mathbf{P}^{-1/2}\mathbf{x} \quad \text{ill.} \quad \underline{\beta} = \mathbf{Q}^{-1/2}\mathbf{y}.$$

Az α, β valószínűségi változókra tett (1.1) feltételek miatt $||\mathbf{x}||$=1 és $||\mathbf{y}||$=1. A maximalizálandó (8.7) kifejezés pedig:

$$\mathbb{E}_R \alpha\beta = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha(i)\beta(j)r_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} x(i)y(j)\frac{r_{ij}}{\sqrt{p_i}\sqrt{q_j}} = \mathbf{x}^T\mathbf{B}\mathbf{y}$$

alakban írható, ahol az $n \times m$-es $\mathbb{B}$ mátrix a következő:

$$\mathbf{B} = \mathbf{P}^{-1/2}\mathbb{R}\mathbf{Q}^{-1/2}.$$

Keresendő

$$\max_{\mathbb{E}_P\alpha^2=1, \mathbb{E}_Q\beta^2=1} \mathbb{E}_R \alpha\beta = \max_{||\mathbf{x}||=1, ||\mathbf{y}||=1} \mathbf{x}^T\mathbf{B}\mathbf{y}.$$

Az 158. Tétel alapján az utóbbi kifejezés maximuma a $\mathbf{B}$ mátrix legnagyobb szinguláris értéke, és felvétetik az ehhez tartozó saját bázispáron, jelölje ezeket $\mathbf{u}_1$ ill. $\mathbf{v}_1$. Így

$$\underline{\alpha}_1 = \mathcal{P}^{-1/2}\mathbf{u}_1 \quad \text{ill.} \quad \underline{\beta}_1 = \mathbf{Q}^{-1/2}\mathbf{v}_1$$

lesz az első összetartozó faktorpár. Könnyű látni, hogy $\alpha_1 \equiv \mathbf{1}$, $\beta_1 \equiv \mathbf{1}$ és $s_1 = 1$, ui. a Cauchy- Schwarz egyenlőtlenség miatt $\mathbb{E}_R\alpha\beta \leq 1$, ugyanakkor az azonosan 1 értéket felvevő α , β párokkal $\mathbb{E}_R\alpha\beta = \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} = 1$ teljesül. Az α_1, β_1 faktorokat *triviális faktorok*nak is szokták nevezni, várható értékük 1, szórásuk 0, kovarianciájuk is 0. A többi faktor korrelálatlansága velük éppen azt jelenti, hogy azok várható értéke 0. Tekintsünk egy ilyen α, β párt. Ezekre tehát

$$\mathbb{E}_P\alpha = 0, \quad \mathbb{D}^2_P\alpha = \mathbb{E}_P\alpha^2 = 1, \quad \mathbb{E}_Q\beta = 0, \quad \mathbb{D}^2_Q\beta = \mathbb{E}_Q\beta^2 = 1.$$

Tekintsük most velük egy szekvenciális feltételes szélsőértékkeresési feladatot, melyből csak az első lépést adjuk meg: keresendő

$$\max_{\mathbb{E}_P\alpha=0,\mathbb{E}_Q\beta=0\mathbb{D}_P\alpha=1,\mathbb{D}_Q\beta=1} \mathbb{E}_R\alpha\beta = \max_{\|\mathbf{x}\|=1,\|\mathbf{y}\|=1\mathbf{x}^T\mathbf{u}_1=0,\mathbf{y}^T\mathbf{v}_1=0} \mathbf{x}^T\mathbf{B}\mathbf{y}.$$

Ismét a 158. Tételre hivatkozva adódik, hogy a maximum a $\mathbf{B}$ mátrix második legnagyobb szinguláris értéke, s_2, és az $\mathbf{u}_2$, $\mathbf{v}_2$ saját bázispáron vétetik fel. Ezek transzformáltjai lesznek az

$$\underline{\alpha}_2 = \mathcal{P}^{-1/2}\mathbf{u}_2 \quad \text{ill.} \quad \underline{\beta}_2 = \mathbf{Q}^{-1/2}\mathbf{v}_2$$

ún. korrespondancia-faktorok. Az s_2 szám éppen a Rényi-féle maximál korreláció.

1.3. Információelméleti módszerek

Mivel itt is diszkrét eloszlásokkal foglalkozunk, az $(\Omega, \mathcal{A}, \mathbb{P})$ valószínűségi mező definíciójában szereplő Ω halmaz mindig véges. Az Ω-án definiálható összes eloszlások családját $\mathcal{D}(\Omega)$-val jelöljük.

A vizsgált eloszlások tipikus példája, a d-szempontos osztályozás, amikor a valószínűségek a d-dimenziós tömbbe vannak rendezve. Az i-edik szempont kategóriáinak számát jelölje r_i, ekkor az Ω elemei

$$\omega = (j_1, \ldots, j_d), \quad 1 \le j_1 \le r_1, 1 \le j_2 \le r_2, \ldots, 1 \le j_d \le r_d$$

alakúak; ezeket szokták celláknak nevezni. Az $\mathbf{X}(\omega) = \mathbf{X}(j_1, \ldots j_d)$ cellagyakoriságokból állló mitát d-dimenziós kontingenciatáblának, pontosabban $r_1 \times r_2, \times \cdots \times r_d$ méretű táblának nevezzük.

8.1.3.1. Definíció. *(Marginális eloszlás.) Megjegyezzük, hogy az elnevezés a latin ,,margo" (genitivus: marginis) szóból származik. Tetszőleges $\gamma \in \{1, \ldots, d\}$ az $\mathbf{X}$ kontingenciatábla, illetve egy $p \in \mathcal{D}(\Omega)$ eloszlás γ-marginálisán azt a $\prod_{i\in\gamma} r_i$-dimenziós $\mathbf{X}^\gamma$ vektort, illetve p^γ vektort értjük, amelynek $\mathbf{X}^\gamma(i_1, \ldots, i_{|\gamma|})$, illetve $p^\gamma(i_1, \ldots, i_{|\gamma|})$ komponensei mindazon $\mathbf{X}(\omega)$, illetve $p(\omega)$ elemek összegével egyenlők, melyekre $\omega = (j_1, \ldots, j_d)$ -nek γ-beli indexű koordinátái rendre $i_1, \ldots, i_{|\gamma|}$. Ha $|\gamma| = k$, akkor k-dimenziós marginálisról beszélünk.*

A fenti formális definíció nehezen érthető, de az alábbi, a $d = 2$, $r_1 = 3r_2 = 3$ esetet illusztráló táblazatokból kitűnik, hogy valójában csak egy jólismert fogalom általánosításának kissé nehézkes, de elkerülhetetlen formalizálásáról van szó.

A könnyebb olvashatóság kedvéért a valószínűségeket százalékban adjuk meg. A 8.1 Táblázat egy háromdimenziós eloszlás táblázata, a szemléletesség kedvéért gondoljuk az i és j koordináták által meghatározott 3×3 (i-vel és j-vel indexelt táblázatokat 3 vízszintes rétegnek, míg a k index az egyes rétegek magasságat jelzi).

	k_1	k_1	k_1	k_2	k_2	k_2	k_3	k_3	k_3
j_1	2	5	2	1	3	4	6	15	6
j_2	1	3	4	2	5	2	3	9	12
j_3	1	1	1	1	1	1	3	3	3
	i_1	i_2	i_3	i_1	i_2	i_3	i_1	i_2	i_3

Háromdimenziós eloszlás???

A 8.2 Táblázat az eredeti háromdimenziós eloszlás (i, k) kétdimenziós marginálisát illusztrálja: a j indexre össszegzünk 9 rögzített (i, k) párra.

	i_1	i_2	i_3
k_1	4	9	7
k_2	4	9	7
k_3	12	27	21

Marginálisok???

Végül a fenti kétdimenziós marginális eloszlás elemeit a k index szerint összegezzük (ami ekvivalens azzal, hogy az eredeti eloszlás elemeit a j és a k indexekre összegezzük minden rögzített i értékre).

i_1	i_2	i_3
20	45	35

Összegzett marginálisok

Ennek a paragrafusnak az a célja, hogy a többdimenzós gyakorisagtáblázatok mögötti eloszlást minél kevesebb paraméterrel írja le információelméleti módszerek segítségével. A becslési feladatoknak két típusát különböztetik meg .

Külső feltételekkel meghatározott feladatok. Ebben az esetben feltételezzük, hogy az $\mathbf{X}$ minta p valódi eloszlása egy $\mathcal{F}$ eloszláscsaládhoz tartozik. A $p \in \mathcal{F}$ eloszlás meghatározásának általánosan elfogadott módja, hogy megkeressük azt a $p^* \in \mathcal{F}$ eloszlást amely az alább ismertetett eltérések valamelyikének értelmében legközelebb van a $p_{\mathbf{X}}$ empirikus eloszláshoz. Ugyanez a módszer a 121 Lemma alapján alkalmazható annak a hipotézisnek a vizsgálatára, hogy az $\mathbf{X}$ minta származhat-e egy $\mathcal{F}$-beli eloszlásból.

Belső feltételekkel meghatározott (modellalkotási) feladatok. Itt az $\mathbf{X}$ mintában foglalt információt kevesebb adattal, általában bizonyos $S_1, \ldots, S_r$ statisztikák mintabeli átlagaival kívánjuk reprezentálni. Ha „ismereteink mintavétel előtti állapotát" $q \in \mathcal{D}(\Omega)$ eloszlás jellemzi (ennek legtöbbször az Ω-án értelmezett egyenletes eloszlást vesszük), akkor az

$$\mathcal{F} = \left\{ p : \sum_{\omega \in \Omega} p(\omega) S_i(\omega) = \sum_{\omega \in \Omega} p_{\mathbf{X}}(\omega) S_i(\omega), \quad i = 1, \ldots, r \right\} \qquad (1)$$

???

eloszláshalmazhoz legközelebbi p^* eloszlást tekintjük a modellalkotási feladat megoldásának.

1.3.1. Eloszlások eltérése

Az eloszlások egymástól való eltérésére számos, az információelméletben használatos mérőszám ismeretes, ezek általánosítását az ún. f-eltérést Csiszár Imre vezette be (l. [9]) 1967-ben.

Mielőtt rátérnénk az információs geometria tárgyalására itt közöljük az ehhez kapcsolódó feladatokban szükséges Jensen-egyenlőtlenséget.

8.1.3.1.1. Tétel (Jensen-egyenlőtlenség.). *Legyen $f(x)$ ($x \in \mathbb{R}$) valós értékű konvex függvény, X pedig egy valószínűségi változó. Tegyük fel, hogy $\mathbb{E}(X)$ és $\mathbb{E}(f(X))$ léteznek. Ekkor*

$$\mathbb{E}\left(f(X)\right) \geq f\left(\mathbb{E}(X)\right). \qquad (1)$$

Legyen $f(u)$ a pozitív félegyenesen értélmezett konvex függvény, amelyre $f(1) = 0$, és legyen megállapodás szerint

$$f(0) = \lim_{u \to 0} f(u), \qquad 0 f(\frac{a}{0}) = a \cdot \lim_{u \to \infty} \frac{f(u)}{u}.$$

8.1.3.1.2. Definíció (f-eltérés). *Tetszőleges $p \in \mathcal{D}(\Omega)$ és $q \in \mathcal{D}(\Omega)$ eloszlások f-eltérésén a*

$$D_f(p||q) = \sum_{\omega \in \Omega} q(\omega) f\left(\frac{p(\omega)}{q(\omega)}\right) \qquad (1)$$

mennyiséget értjük.

A tananyagban $f(u)$-t háromféleképpen választjuk meg:

* (i) $f(u) = |u - 1|$

* (ii) $f(u) = (u-1)^2$

* (iii) $f(u) = u \log u$

Az (i), (ii) és (iii) függvényeknek rendre a $\sum_\omega |p(\omega) - q(\omega)|$ variációs távolság, a $\sum_\omega \frac{1}{q(\omega)}(p(\omega) - q(\omega))^2$ Pearson-féle χ^2-eltérés, illetve a

$$D_f(p||q) = \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)} \qquad (1)$$

???Kullback- Leibler-féle „diszkrimináló információ" (ezt a rövidség kedvéért a továbbiakban egyszerűen **divergenciának** nevezzük) felel meg.

8.1.3.1.3. Lemma. *$D_f(p||q) \geq 0$, ha $f(u)$ az $u = 1$ pontban szigorúan konvex, akkor az egyenlőség csak $p = q$ esetén teljesül.*

Bizonyítás Lásd ???? Feladat.

A fenti Lemma állításából nem következik, hogy az f-eltérés távolság, mert általában sem a szimmetria, sem a háromszög egyenlőtlenség nem teljesül. A felsorolt 3 eltérés közül csak az (i) variációs távolság valódi távolság.

Jelölje $T(p)$ a p eloszlás tartóját:

$$T(p) := \{\omega : p(\omega) > 0\}.$$

Nyilvánvaló, hogy $D(p||q)$ akkor és csak akkor véges, ha $T(p) \subseteq T(q)$.

A következő Lemma lehetőséget teremt az f-eltérések statisztikai próbákban történő felhasználására.

8.1.3.1.4. Lemma. *(Az f-eltérés és a χ^2-eloszlás kapcsolata) Ha az eltérést definiáló $f(u)$ függvény az $u = 1$ pontban szigorúan konvex, az $u = 1$ pont egy környezetében kétszer folytonosan differenciálható, és $f''(1) > 0$, akkor az egymáshoz közeli p és q eloszlások f-eltérése a χ^2-eltérésük egy konstansszorosával közelíthető, pontosabban bármely $\varepsilon > 0$-hoz van olyan $\delta > 0$, hogy*

$$\left(\frac{f''(1)}{2} - \varepsilon\right) \sum_{\omega \in \Omega} \frac{(p(\omega) - q(\omega))^2}{q(\omega)} \leq D(p||q) \leq \\ \leq \left(\frac{f''(1)}{2} + \varepsilon\right) \sum_{\omega \in \Omega} \frac{(p(\omega) - q(\omega))^2}{q(\omega)}, \qquad (1)$$

ha $|p(\omega) - q(\omega)| \leq \delta q(\omega)$ minden $\omega \in \Omega$-ra.

A Lemma feltétele teljesül a divergenciára.

A kontingenciatáblázatok elemzésekor alapfeladat az, hogy egy megkeressük egy $\mathcal{F} \subseteq \mathcal{D}(\Omega)$ eloszláscsaládnak adott p eloszlástól legkevésbé eltérő elemét. Ezt kétféleképpen tehetjük meg.

8.1.3.1.5. Definíció (Vetületek).

I-vetület *Egy $q \in \mathcal{D}(\Omega)$ eloszlásnak $\mathcal{F} \in \mathcal{D}(\Omega)$ eloszláshalmazra vonatkozó I-vetülete az a $p^* \in \mathcal{F}$ eloszlás, amelyre*

$$D(p^*||q) = \min_{p \in \mathcal{F}} D(p||q) < \infty. \qquad (1)$$

L-vetület *Egy $p \in \mathcal{D}(\Omega)$ eloszlásnak $\mathcal{F} \in \mathcal{D}(\Omega)$ eloszláshalmazra vonatkozó L-vetülete az a $q^* \in \mathcal{F}$ eloszlás, amelyre*

$$D(p||q^*) = \min_{q \in \mathcal{F}} D(p||q) < \infty. \qquad (1)$$

Az ??? feladatban fogalmaztuk meg a következő lemma egyik állítását.

Mielőtt a lemmát kimondanánk vezessük be a $p_A(\omega) := \frac{p(\omega)}{P(A)}$ ha $\omega \in A, p_A(\omega) := 0,$ ha $\omega \notin A$ jelölést, és analóg módon a $q_A(\omega)$ jelölést is.

8.1.3.1.6. Lemma.

Legyenek, $A_1, \ldots, A_r$ az Ω valószínűségi tér páronként diszjunkt részhamazai melyekre $\cup_{i=1}^r A_i = \Omega$ (teljes eseményrendszer). Ekkor tetszőleges p és q eloszlásokra:

$$D_f(p||q) \geq \sum_{i=1}^{r} q(A_i) f\left(\frac{p(A_i)}{q(A_i)}\right). \qquad (1)$$

Egyenlőség akkor érvényes ha $p_{A_i} = q_{A_i}$ minden olyan i-re, amelyre $p(A_i)q_(A_i) > 0.$ Ha f szigorúan konvex, akkor az egyenlőségnek ez elégséges feltétele.

A fenti Lemma lehetővé teszi, hogy egy q eloszlásnak meghatározzuk az I-vetületét egy speciális eloszláshalmazra; nevezetesen azon eloszlások halmazára, amelyek szerint egy $A_1, \ldots, A_r$ teljes eseményrendszer elemeinek valószínűségei adottak:

$$\mathcal{F}\{p : p(A_i) = \pi_i\}. \qquad (1)$$

???

8.1.3.1.7. Tétel. *(Jeffrey-szabály.) Ha $q(A_i) > 0$ minden i-re, amelyre $\pi_i \neq 0$*

$$\min_{p \in \mathcal{F}} D(p||q) = D(p^*||q) = \sum_{i=1}^{r} q(A_i) f\left(\frac{\pi_i}{q(A_i)}\right),$$

ahol

$$p^*(\omega) = \frac{\pi_i}{q(A_i)} q(\omega)$$

minden $\omega \in \Omega$-ra.

Vegyük észre, hogy ebben az esetben az I-vetület nem függ az eltérést meghatározó függvénytől; ez általában nincs így.

A Jeffrey-szabállyal egy speciális külső feltételekkel megadott feladatot oldunk meg, ugyanis ha $q = p_{\mathbf{X}}$, akkor p^* az (8.16) $\mathcal{F}$ eloszláscsalád $p_{\mathbf{X}}$-hez legközelebbi eleme lesz a becslés eredménye. Ugyanakkor a Jeffrey szabállyal kapott p^* becslés teljesíti a belső feltételekkel megadott feladat (8.8) egyenlőségét is.

Minimális diszkrimináló információ módszernek (MDI) nevezzük azt az eljárást, amikor a becslés az $\mathcal{F}$ eloszláscsaládnak a q eloszláshoz Kullback- Leibler értelemben legközelebbi p eleme

Most megmutatjuk, hogy a polinomiális eloszlás e az empirikus eloszlás divergencia szerinti L-vetülete a polinomiális eloszlások halmazára. Minden $\omega \in \Omega$-ra az ω kategóriába eső elemek száma legyen $\mathbf{X}(\omega)$, az $\mathbf{X}(\omega)$ komponenseiből alkotott vektor az $\mathbf{X}$ minta, a mintaelemszám $N := \sum_{\omega \in \Omega} \mathbf{X}(\omega), p_{\mathbf{X}} = \frac{1}{N}\mathbf{X}$

Ezekkel a jelölésekkel az $\mathbf{X}$ minta függvénye:

$$
\begin{aligned}
L(p_{\mathbf{X}}) &= \log\left[\frac{N!}{\prod_{\omega\in\Omega}\mathbf{X}(\omega)!}\prod_{\omega\in\Omega}p(\omega)^{\mathbf{X}(\omega)}\right] \\
&= a(\mathbf{X}) + \sum_{\omega\in\Omega}\mathbf{X}(\omega)\log p(\omega) \\
&= a(\mathbf{X}) - N\sum_{\omega\in\Omega}\mathbf{X}(\omega)p_{\mathbf{X}}\log\frac{1}{p(\omega)} \quad (1) \\
&= a(\mathbf{X}) - N\sum_{\omega\in\Omega}\mathbf{X}(\omega)p_{\mathbf{X}}\log\frac{p_{\mathbf{X}}(\omega)}{p(\omega)} - \sum_{\omega\in\Omega}\mathbf{X}(\omega)p_{\mathbf{X}}(\omega)\log p_{\mathbf{X}}(\omega) \\
&= b(\mathbf{X}) - N\sum_{\omega\in\Omega}\mathbf{X}(\omega)p_{\mathbf{X}}\log\frac{p_{\mathbf{X}}(\omega)}{p(\omega)} = b(\mathbf{X}) - ND(p_{\mathbf{X}}||p)
\end{aligned}
$$

???

Ahol $a(\mathbf{X})$ és $b(\mathbf{X})$ csak a mintától (a becsülendő p paramétervektortól nem) függő - így a maximumot nem befolyásoló függvényeket jelölnek.

A fenti egyenlőségből adódik

$$ND(p_{\mathbf{X}}||p) = L(p_{\mathbf{X}}) - b(\mathbf{X}),$$

tehát $L(p_{\mathbf{X}})$ ugyanarra a p vektorra veszi fel a maximumát, amelyre $ND(p_{\mathbf{X}}||p)$ a minimumát.

Ez a becslési módszer a külső feltételekkel megadott feladat megoldását adja abban a speciális esetben, amikor az $\mathcal{F}$ eloszláshalmaz az Ω véges halmazon értelmezett összes lehetséges eloszlást tartalmazza.

Ha q az Ω-án egyenletes eloszlás, akkor a divergencia definíciójából következik

$$D(p||q) = \sum_{\omega\in\Omega}p(\omega)\log p(\omega) + \log|\Omega|,$$

tehát az I-vetület most éppen az a $p\in\mathcal{F}$ eloszlás, amelynek a

$$H(p) = -\sum_{\omega\in\Omega}p(\omega)\log p(\omega)$$

Shannon-entrópiája maximális.

Ezért a rendkívül népszerű maximális-entrópia becslési módszer speciális esetként tartalmazza az MDI-módszert.

Az f-eltérés nem távolság, ennek ellenére bizonyos geometriai állítások az f-eltérésre is igazak. Az információelmélet geometriai megközelítése az elemi matematikai példatáráról jól ismert N. N. Csencov [8] orosz matematikustól származik. Most megmutatjuk, hogy speciális „duális" eloszláscsaládok esetén az f-eltérésre teljesül a Pitagorasz-tétel.

Legyenek $S_1,\dots,S_r$ az Ω halmazon értelmezett tetszőleges valós függvények, és legyen S_0 az azonosan 1 függvény. Jelölje $\mathbf{S}$ azt az $(r+1)\times|\Omega|$ típusú mátrixot, amelynek i-edik sora $S_i(\omega)$, $i=0,\dots,r$

Az $\mathbf{S}$ mátrix segítségével két eloszláscsaládot definiálunk.

8.1.3.1.8. Definíció. *(Lineáris és exponenciális eloszláscsalád.) Legyenek $p_0\in\mathcal{D}(\Omega)$ és $q_0\in\mathcal{D}(\Omega)$ tetszőleges eloszlások. Az*

$$\mathcal{L} = \mathcal{L}(\mathbf{S},p_0) := \{p:\ \mathbf{S}p = \mathbf{S}p_0\} \qquad (1)$$

*eloszláscsaládot az $\mathbf{S}$ mátrixhoz és p_0 eloszláshoz tartozó **lineáris eloszláscsalád**nak nevezzük. Az*

$$\mathbb{E} = \mathbb{E}(\mathbf{S},q_0) := \{q:\ q = q_0\exp(\mathbf{S}^\top\tau)\}, \qquad (1)$$

$r+1$ $q_0 \exp(\mathbf{S}^\top \tau)$ *a* $q_0(\omega)\exp(\sum_{i=0}^{r} \mathbf{S}_i$ $q_0 \exp(\mathbf{S}^\top \tau) \in \mathcal{D}(\Omega)$*ből álló vektort jelenti, és* τ *befutja mindazokat az* r*-dimenziós* $\tau_1, \ldots, \tau_r$*rokat amelyekre* $\in \mathcal{D}(\Omega)$*, exponenciális eloszláscsaládnak nevezzük. Ez azt jelenti, hogy* $\mathbf{S}$ *tetszőlegesek és*

$$\tau_0 = -\log \sum_{\omega \in \Omega} q_0(\omega) \exp(\sum_{i=1}^{r} S_i(\omega)\tau_i).$$

Vegyük észre, hogy a definiáló q_0 eloszlás eleme $\mathcal{E}$-nak a $\tau = (\tau_0 = 0, \tau_1 = 0, \ldots, \tau_r = 0)$ választással.

A (8.18) definícióból következik, hogy a lineáris eloszláscsalád az $\mathbb{R}^{|\Omega|}$ euklideszi tér zárt halmaza, míg az (8.19) definíció alapján látható, hogy az exponenciális eloszláscsalád nem zárt (egy valószínűség tetszőlegesen közel lehet 0-hoz, de nem lehet egyenlő vele). A család lezártját $\mathrm{cl}\mathbb{E}(\mathbf{S}, q_0)$-lal, vagy egyszerűen $\mathrm{cl}\mathbb{E}$-vel jelöljük.

Jegyezzük meg, hogy minden $q \in \mathcal{E}$ eloszlásra $T(q) = T(q_0)$ és minden $q \in \mathrm{cl}\mathcal{E}$ eloszlásra $T(q) \subset T(q_0)$

Továbbá vegyük észre, hogy ha a (8.18) és a (8.19) definíciókban szereplő $\mathbf{S}$ mátrixot újabb sorok hozzávételével egy $\bar{\mathbf{S}}$ mátrixszá egészítjük ki, akkor

$$\tilde{\mathcal{L}}(\bar{\mathbf{S}}, p_0) \subset \mathcal{L}(\mathbf{S}, p_0) \text{ és } \tilde{\mathcal{E}}(\bar{\mathbf{S}}, p_0) \supset \mathcal{E}(\mathbf{S}, p_0).$$

8.1.3.1.9. Tétel. *(Az információs geometria Pitagorasz-tétele.) Tegyük fel, hogy a (8.18) és a (8.19) eloszláshalmazok metszete nem üres:* $\mathcal{L} \cap \mathrm{cl}\mathcal{E} \neq \emptyset$*. Ekkor* $\mathcal{L}$*-nek és* $\mathcal{E}$*-nek pontosan egy* p^* *közös eleme van, és erre*

$$D(p||q) = D(p||p^*) + D(p^*||q), \quad \text{ha} \quad p \in \mathcal{L}, q \in \mathrm{cl}\mathcal{E}, \qquad (1)$$

továbbá

$$T(p^*) = \bigcup_{\substack{p \in \mathcal{L} \\ T(p) \subset T(q_0)}} T(p) \qquad (1)$$

és

$$p^*(\omega) = q_0(\omega) \exp\left(\sum_{i=0}^{r} S_i(\omega)\tau_i^*\right), \text{ ha } \omega \in T(p^*) \qquad (1)$$

ahol $\tau^* = (\tau_1^*, \ldots \tau_r^*)^\top$ *alkalmas vektor.*

Bizonyítás A Tananyagban általában nem közlünk bizonyításokat, de a 126 Tétel érdekessége miatt a (8.20) azonosságot bebizonyítjuk.

A divergencia definíciójából következik , hogy tetszőleges $q = q_0 \exp(\mathbf{S}^\top \tau) \in \mathbb{E}$ és $q' = q_0 \exp(\mathbf{S}^\top \tau') \in \mathbb{E}$ eloszlásokra és a $t(p) \subset T(q_0)$ feltételt kielégítő $p \in \mathcal{D}(\Omega)$ eloszlásra

$$D(p||q) - D(p||q') = \sum_{\omega \in \Omega} p(\omega) \log \frac{q'(\omega)}{q(\omega)} = p^\top \mathbf{S}^\top (\tau' - \tau).$$

Ezért $p \in \mathcal{L}, T(p) \subset T(q_0)$ esetén

$$D(p||q) - D(p||q') = f(q, q') \qquad (1)$$

???(azaz nem függ p-től, ha $q \in \mathcal{E}, q' \in \mathcal{E}$). Határátmenettel adódik, hogy (8.23) akkor is igaz marad ha q és q' a bővebb $\mathrm{cl}\mathcal{E}$-nek eleme, kizárva azokat a p-ket amelyekre $D(p||q) = \infty$. Most $p^* \in \mathcal{L} \cap \mathcal{E}$ esetén q' szerepét p^*-nak adva a (8.23) egyenlőségből adódik, hogy

$$D(p||q) - D(p||p^*) = D(p^*||q) - D(p^*||p^*)$$

Mivel $D(p^*||p^*) = 0$ a (8.20) egyenlőséget bebizonyítottuk.

Kiegészítés. A $\mathcal{L} \cap \mathrm{cl}\mathcal{E} \neq \emptyset$ feltétel pontosan akkor teljesül, ha $T(p) \subseteq T(q_0)$.

8.1.3.1.10. Megjegyzés. *A divergencia nemnegatív voltából következik, hogy a $\{p^*\} = \mathcal{L} \cap \mathcal{E}$ halmaz egyetlen eleme egyidejűleg a q eloszlás $\mathcal{L}$-re vett I-vetülete és a p-eloszlás $\mathcal{E}$-re vett L-vetülete.*

1.3.2. A belső és külső feltételekkel meghatározott feladatok részletesebb elemzése

1. Belső feltételekkel meghatározott feladatok. Legyen $p_{\mathbf{X}}$ az $\mathbf{X}$ minta empirikus eloszlása, q_0 a mintavétel előtti ismereteinket jellemző eloszlás, és legyenek $S_1, \ldots, S_r$ azok a statisztikák, amelyeknek mintabeli átlagait a már vázolt modellalkotási feladathoz fel kívánjuk használni. Ekkor a modellalkotási feladat MDI-megoldásán a q_0-nak az

$$\underline{\mathcal{L}} = \mathcal{L}(\mathbf{S}, p_{\mathbf{X}}) = \{p : \mathbf{S}p = \mathbf{S}p_{\mathbf{X}}\} \qquad (1)$$

???lineáris eloszláscsaládra vonatkozó p^* I-vetületét értjük. A továbbiakban feltesszük, hogy $T(q) = \Omega$. A 126 Tétel kiegészítése szerint a p^* I-vetület létezik és egyértelmű.

Struktúrális 0-nak nevezzük a (8.24) eloszláscsaládra nézve azokat az $\omega \in \Omega$ elemeket, amlyekre minden $p \in \underline{\mathcal{L}}$ eloszlásra $p(\omega) = 0$. Feltesszük, hogy az $\mathbf{X}$ mintában nincsenek struktúrális 0-k. Ez a helyzet, ha minden $\omega \in \Omega$-ra az $\mathbf{X}(\omega) \neq 0$. Ekkor a már említett kiegészítés szerint a p^* I-vetület az $\mathcal{L} \cap \mathcal{E}$ metszet egyetlen eleme, (éppen a struktúrális 0-k hiánya miatt nem kell $\mathcal{E}$ lezárását tekinteni), és p^* megegyezik a $p_{\mathbf{X}}\mathcal{E}$-ra vonatkozó L-vetületével, azaz az ismeretlen eloszlás maximum-likelihood becslésével [l. (8.17)].

H az $\mathbf{X}$ kontingenciatáblában van struktúrális 0 akkor a modellalkotási feladat p^* megoldasa csak a $\mathrm{cl}\mathcal{E}$-ben és

$$p^*(\omega) = \begin{cases} q_0(\omega) \exp\left(\sum_{\gamma \in \Gamma} \tau_\omega^\gamma\right), & \text{ha} \quad \omega \in T(p^*) \\ 0, & \text{ha} \quad \omega \notin T(p^*). \end{cases}$$

Az MDI-megoldásként kapott p^* eloszlás akkor tekinthető a $p_{\mathbf{X}}$ empirikus eloszlás adekvát modelljének, ha a $D(p_{\mathbf{X}}||p^*)$ divergencia kicsi, ennek kvantitatív mérésére az 121 Lemma nyújt lehetőséget.

Ha az $\mathbf{X}$ egy $q \in \mathcal{D}(\Omega)$ eloszlásból vett N elemű minta , akkor a (8.12) képlet alapján:

$$2ND(p_{\mathbf{X}}||q) \sim \sum_{\omega \in \Omega} \frac{(X(\omega) - Nq(\omega))^2}{Nq(\omega)}, \quad \text{ha} \quad N \to \infty. \qquad (1)$$

Itt a $\sim$ jel azt jelenti, hogy a két oldal hányadosa tart 1-hez. A jobb oldali tört aszimptotikusan $|\Omega| - 1$ szabadságfokú χ^2 eloszlású.

1. Külső feltételekkel meghatározott feladatok. Ezekben a feladatokban az MDI-módszer akkor célszerű, ha az ott szereplő $\mathcal{F}$ eloszláscsalád egy $\mathcal{L}(\mathbf{S}, p_0)$ lineáris eloszláscsalád. Ha feltesszük, hoy az $\mathbf{X}$ minta valamelyik (ismeretlen) $p \in \mathcal{L}$ eloszlásból származik, ennek az eloszlásnek az MDI-becslésén a $p_{\mathbf{X}}$ empirikus eloszlás $\mathcal{L}$-re vonatkozó p^* I-vetületét értjük, feltéve, hogy erre teljesül $T(p^*) = T(p_{\mathbf{X}})$. (Az I-vetület (8.13) definíciójából következik, hogy $T(p^*) \subseteq T(p_{\mathbf{X}})$, azonban a valódi tartalmazás kizárható, mert ekkor az $\mathbf{X}$ minta biztosan nem származhatna a p^* eloszlásból.)

A 126 Tétel szerint a p^* MDI-becslés $p_{\mathbf{X}}$ helyett bármely $q \in \mathcal{E}(\mathbf{S}, p_{\mathbf{X}})\mathcal{L}$-re vonatkozó I-vetületeként is megkapható. Ez azt jelenti, hogy az adott MDI-becslési feladat eredménye nem változik, ha a $p_{\mathbf{X}}$ empirikus eloszlást egy korábbi MDI-becsléssel helyettesítjük, feltéve, hogy abban a becslésben alkalmazott az $\mathcal{L}'$ családot definiáló $\mathbf{S}'$ mátrix sorai benne vannak az $\mathbf{S}$ sorai által kifeszített altérben. (l. ??? Feladat).

Az MDI-becslés most is felhasználható a $p \in \mathcal{L}$ hipotézis tesztelésére, ugyanis a (8.25) formulához hasonlóan adódik, hogy ha a valódi eloszlás p, akkor

$$2ND_f(p||p_{\mathbf{X}}) \sim \sum_{\omega \in \Omega} \frac{(X(\omega) - Np(\omega))^2}{Np(\omega)}, \quad \text{ha} \quad N \to \infty. \qquad (1)$$

???Itt a D_f eltéri$|T_p| - 1 f(u) = -\log u$ vχ^2asztással kell számolni. A (8.26) képlettel definiált statisztika aszimptotikusan szabadságfokú eloszlást követ. A 126 Tétel (8.20) képlétét alkalmazva a k'ek[varianciaanalíizis]ből ismert szórásnégyzet felbontást is kaphatunk:

$$2ND_f(p||p_{\mathbf{X}}) = 2ND_f(p||p^*) + 2ND_f(p^*||p_{\mathbf{X}}),$$

ahol az összeadandók aszimptotikusan függetlenek, az első tag szabadságfoka $|T_p| - 1 - r$, míg a második tag szabadságfoka r azaz az $\mathcal{L}$ lineáris családot definiáló mátrix nem konstans sorainak száma.

1.4. Az I-vetület numerikus meghatározása

Ebben a pontban egyetlen módszert ismertetünk nevezetesen azt amelyik akkor alkalmazható, ha az $\mathcal{L}$ lineáris család olyan $\mathcal{L}_1, \ldots, \mathcal{L}_r$ lineáris családok metszete amelyekre való egyes I-vetületek explicite meghatározhatók. Ez a helyzet, amikor az eloszláscsalád bizonyos γ-marginálisok előírásával van megadva:

$$\mathcal{L} = \{p\colon\ p^\gamma = p_0^\gamma, \gamma \in \Gamma\}.$$

8.1.4.1. Tétel. *Legyenek $\mathcal{L}_1, \ldots, \mathcal{L}_r$ lineáris eloszláscsaládok, $\mathcal{L} \cap_{i=1}^r \mathcal{L}_i$ és legyen q_0 tetszőleges olyan eloszlás, amelyhez található a $T(p) \subseteq T(q_0)$ feltételt kielégítő $p \in \mathcal{L}$. Értelmezzük a $p_1^*, p_2^*, \cdots$ eloszlásokat a következő iterációval: $p_0^* = q_0$, és $n = 1, 2, \ldots$ esetén*

p_n^* a $p_{n-1}^*\mathcal{L}$-re vonatkozó I-vetülete,

ahol $\mathcal{L}_n = \mathcal{L}_i$ ha $n = kr + i$.

Ekkor q_0-nak $\mathcal{L}$-re vonatkozó I-vetülete:

$$p^* = \lim_{n\to\infty} p_n^*.$$

2. Feladatok

(i) Bizonyítsuk 120 Lemmát, azaz azt az állítást,

hogy ha az t definiáló $f(u)$ függvény az $u = 1$ pontban szigorúan konvex, akkor $D_f(p||q) \geq 0$, és egyenlőség csak akkor áll fenn, ha $p = q$.

Tipp: Alkalmazzuk a et az $f(u)$ fügvényre, az $X = \frac{p(\omega)}{q(\omega)}$ valószínűségi változóra és a q eloszlás szerinti várható értékre. Vegyük észre, hogy ebben a szereposztásban

$$f(\mathrm{E}[X]) = f\left(\sum_{\omega\in\Omega} q(\omega)\cdot\frac{p(\omega)}{q(\omega)}\right) = f(1) = 0.$$

Ha $f(u)$ az $u = 1$ pontban szigorúan konvex, és $p \neq q$ akkor $f(p/q) > 0$ így $\mathrm{E}[f(X)] > 0$.

Válasz:

(ii) Bizonyítsuk be a következő állítást.

Legyenek, $A_1, \ldots, A_r$ az Ω halmaz páronként diszjunkt részhamazai melyekre $\cup_{i=1}^r A_i = \Omega$. Ekkor tetszőleges p és q eloszlásokra:

$$D_f(p||q) \geq \sum_{i=1}^r q(A_i) f\left(\frac{p(A_i)}{q(A_i)}\right).$$

Az állítás szemléletes tartalma az, hogy a durvított eloszlások f-eltérése nem nagyobb, mint az eredeti eloszlásoké.

Tipp: Vezessük be a $p_A(\omega) := \frac{p(\omega)}{P(A)}$ ha $\omega \in A, p_A(\omega) := 0$, ha $\omega \notin A$ jelölést, és analóg módon a $q_A(\omega)$ jelölést.

A fenti jelölésekkel

$$D_f(p||q) = \sum_{\omega\in A_i}^{r} q(\omega)\left(\frac{p(\omega)}{q(\omega)}\right).$$

Alkalmazzuk a et az f fügvényre, a $\frac{p(\omega)}{q(\omega)}$ valószínűségi változóra a $q_{A_i}(\omega)$ feltételes eloszlás szerinti várható értékkel.

Válasz:

(iii) Legyen Ω tetszőleges véges halmaz. Keressük meg azt az Ω-n értelmezett $p(\omega)$ eloszlást amelyre a

$$H(p) = -\sum_{\omega\in\Omega} p(\omega)\log p(\omega)$$

entrópia maximális. Mennyi a maximális érték?

Tipp: Alkalmazzuk a a szélsőérték-számítás Lagrange-multiplikátor módszerét! (Aki nem ismeri ezt a módszert, oldja meg a feladatot az $|\Omega| = 2$ esetben.)

Válasz: $p(\omega) = \frac{1}{|\Omega|}$, $H = \log|\Omega|$.

(iiii) Legyen $\Omega = \{0, 1, \ldots, n\}$,$r = 1$,$S_1(\omega) = \omega$. Legyen továbbá $p_0 \in \mathcal{D}(\Omega)$ tetszőleges q_0 pedig az $(n, \frac{1}{2})$ paraméterű binomiális eloszlás.

(a) Bizonyítsuk be, hogy a fenti jelölésekkel az $\mathcal{L}(\mathbf{S}, p_0)$ lineáris eloszlászcsalád mindazon $p = (p(0), p(1), \ldots p(n))$ eloszlások összessége, amelyek várható E_0 értéke megegyezik p_0-éval, azaz

$$\sum_{i=0}^{n} p(i)i = \sum_{i=0}^{n} p_0(i)i,$$

az $\mathcal{E}(\mathbf{S}, q_0)$ exponenciális eloszláscsalád az n, π paraméterű binomiális eloszlások összessége, ahol $n\pi = \mathrm{E}_0$.

(b) Adjuk meg az exponenciális család $q = q_0\exp(\mathbf{S}^\top\tau)$ előállításában szereplő $\tau = (\tau_0, \tau_1)^\top$ vektort a binomiális eloszlás π paraméterével.

Tipp: Idézzük fel a k'ek[lineáris és exponenciális eloszláscsalád definícióját]

Válasz:

$$\tau_1 = \log\frac{\pi}{1-\pi}, \quad \tau_0 = n\log(2-2\pi).$$

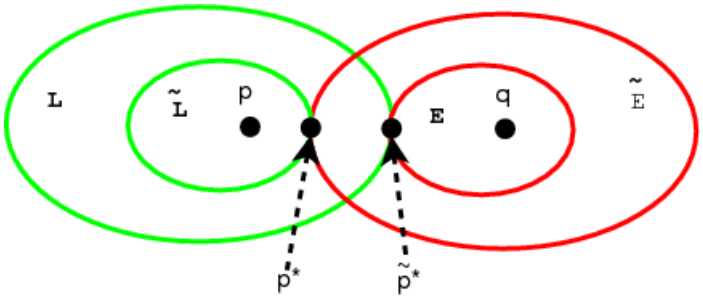
(iiiii) Legyen $\bar{\mathbf{S}}$ olyan mátrix, amely az $\mathbf{S}$ mátrixból további sorok hozzáadásával nyertünk. Jelölje az $\bar{\mathbf{S}}$ mátrix az eredeti p_0, és q_0 által definiált eloszláscsaládokat $\bar{\mathcal{L}}(\bar{\mathbf{S}}, p_0)$ és $\bar{\mathcal{E}}(\bar{\mathbf{S}}q_0)$. (Az $\bar{\mathcal{E}}(\bar{\mathbf{S}}q_0)$ definíciójában szereplő τ vektorok lehetséges halmaza is kibővül.)

Tegyük fel, hogy $\mathcal{L}\cap\mathrm{cl}\mathcal{E} \neq \emptyset$ és $\bar{\mathcal{L}}\cap\mathrm{cl}\bar{\mathcal{E}} \neq \emptyset$.

Ekkor minden $p \in \bar{\mathcal{L}}$ és $q \in \mathrm{cl}\mathcal{E}$ eloszlásra

$$\begin{aligned} D(p||q) &= D(p||p^*) + D(p^*||q) \\ D(p||p^*) &= D(p||\bar{p}^*) + D(\bar{p}^*||p^*), \end{aligned} \qquad (1)$$

ahol $p^* \in \mathcal{L}\cap\mathrm{cl}\mathcal{E}$ és $\bar{p}^* \in \bar{\mathcal{L}}\cap\mathrm{cl}\bar{\mathcal{E}}$.



Eloszláscsaládok

Tipp: Idézzük fel a definícióját és az . A különböző eloszláscsaládok viszonyait, és az ebben elhelyezkedő eloszlásokat az alábbi ábra szemlélteti.

Válasz:

3. Tesztek

(i) Az alábbi f fügvények közül jelöljük meg azokat amelyekhez tartozó f-eltérés távolság.

(a) $f(u) = (u-1)^2$

(b) $f(u) = (1-\sqrt{u})$

(c) $f(u) = |u-1|$

(d) $f(u) = u - \log u$

Válasz: c

(ii) Az X és Y véletlen változók 4-4 értéket vehetnek fel,

együttes eloszlásukat az alábbi mátrix tartalmazza.

$$\begin{pmatrix} 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 \end{pmatrix}$$

Az alábbi sorok melyikében állnak az X illetve az Y valószínűségi változóhoz tartozó marginális eloszlás valószínűségei?

(a) (1, 2, 3, 4)

(b) (1, 2, 3, 4)

(c) (1, 2, 3, 4)

(d) (1, 2, 3, 4)

Válasz: válasz: itt a számoktól függ,

(iii) Az alábbi állítások közül melyik igaz Jeffrey-szabályra?

(a) A Jeffrey-szabállyal csak I-vetületet számolunk.

(b) A Jeffrey-szabállyal csak L-vetületet számolunk.

(c) A Jeffrey-szabállyal I- és L-vetületet számolunk.

(d) A Jeffrey-szabállyal nem vetületet számolunk.

Válasz: c

(iiii) Az alábbi állítások közül melyik igaz Jeffrey-szabályra?

(a) A Jeffrey-szabály a lineáris eloszláscsaládra érvényes.

(b) A Jeffrey-szabály az exponenciális eloszláscsaládra érvényes.

(c) A Jeffrey-szabály eredménye függ az eltérést definiáló függvénytől.

(d) A fentiek közül egyik sem igaz.

Válasz: d

(iiiii) A lineáris (exponenciális) eloszláscsaládot egy $\mathbf{S}(r+1) \times |\Omega|$ típusú mátrix definiálja.

Az alábbi állítások közül melyek igazak?

(a) Ha az $\mathbf{S}$ mátrixot további sorokkal bővítjük, az általa definiált lineáris eloszláscsalád bővül, valamint az általa definiált exponenciális eloszláscsalád bővül.

(b) Ha az $\mathbf{S}$ mátrixot további sorokkal bővítjük, az általa definiált lineáris eloszláscsalád szűkül, valamint az általa definiált exponenciális eloszláscsalád bővül.

(c) Ha az $\mathbf{S}$ mátrixot további sorokkal bővítjük, az általa definiált lineáris eloszláscsalád bővül, valamint az általa definiált exponenciális eloszláscsalád szűkül.

(d) Ha az $\mathbf{S}$ mátrixot további sorokkal bővítjük, az általa definiált lineáris eloszláscsalád szűkül, valamint az általa definiált exponenciális eloszláscsalád szűkül.

Válasz: b

9. fejezet - Klaszteranalízis, többdimenziós skálázás

1. Elméleti háttér

1.1. Klaszteranalízis

A diszkriminanciaanalízistől eltérően itt nem adott osztályokkal dolgozunk, hanem magukat az osztályokat (*klaszter*eket) keressük, azaz objektumokat szeretnénk osztályozni a rajtuk végrehajtott többdimenziós megfigyelések alapján (ugyanez megtehető a változókkal is az objektumok alapján).

A minimalizálandó veszteségfüggvény, aminek segítségével az osztályozást végrehajtjuk - egyelőre csak vázlatosan - a következő. Az n db objektum a p-dimenziós mintatér pontjainak tekinthető $(p < n)$, és euklideszi metrikában dolgozunk. Tekintsük minden egyes osztályra az adott osztálybeli objektumok súlypontját, és vegyük az objektumok négyzetes eltérését (távolság-négyzetét) a súlyponttól. Az így kapott mennyiségeket utána összegezzük az osztályokra és keressük azt az osztályszámot, hozzá pedig az osztályokat, melyekre ez a veszteség minimális. Arra vonatkozóan, hogy hogyan alakult ki ez a veszteségfüggvény, röviden utalunk a varianciaanalízisre, ahol a

$$T = W + B$$

szórásnégyzet-felbontás alapvető. A minta teljes (Total) varianciáját a csoportokon belüli (Within) és a csoportok közötti (Between) varianciákra bontjuk fel. Az objektumok minden egyes partíciójához létezik ilyen felbontás, és a *klaszterezés* (osztálybasorolás) annál homogénebb, minél kisebb W a B-hez képest, azaz a

$$\frac{W}{B} = \frac{W}{T - W}$$

kifejezést szeretnénk minimalizálni, ami (T fix lévén) W minimalizálásával ekvivalens.

Legyenek $\mathcal{C}_1, \ldots, \mathcal{C}_k$ a klaszterek (ezek a mintateret alkotó objektumok partícióját jelentik diszjunkt, nem-üres részhalmazokra). A j. klaszter súlypontja

$$\mathbf{s}_j = \frac{1}{|\mathcal{C}_j|} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i.$$

A $\mathcal{C}_j$-beliek négyzetes eltéréseinek összege $\mathbf{s}_j$-től:

$$W_j = \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{s}_j\|^2 = \frac{1}{|\mathcal{C}_j|} \sum {}^{i,\mathbf{x}_{i'} \in \mathcal{C}_j}_{i<i'} \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2.$$

(Az utolsó egyenlőség egyszerű geometriai meggondolásból adódik, így még a súlypont kiszámolása sem szükséges.) Megjegyezzük, hogy a fenti euklideszi távolságok az eredeti adatok ortogonális transzformációira invariánsak, a célfüggvény csak a pontok kölcsönös helyzetétől függ. Ezekután keresendő a

$$W = \sum_{j=1}^{k} W_j \to min.$$

veszteség-minimum, amelynek fizikai jelentése a k db. súlypontra vonatkozó tehetetlenségi (inercia) nyomatékok összege. Itt az euklideszi távolságnégyzetek helyett más metrikával is dolgozhatunk, pl. vehetjük az $f(\|\mathbf{x}_i\|)$ függvényeket, ahol f folytonos, monoton növő A minimalizálás természetesen az összes lehetséges k-ra $(1 \le k \le n)$, és emelett az összes lehetséges klaszterbesorolásra vonatkozik. Ismert tény, hogy az összes partíciók száma az ún. Bell-szám:

$$\omega(n) = \sum_{k=1}^{n} \left\{ {n \atop k} \right\},$$

ahol az $\left\{ {n \atop k} \right\}$-val jelölt ún. másodfajú Stirling-féle $(k = 1, \ldots, n)$ elemű halmaz k nem-üres, diszjunkt részhalmazra való összes lehetséges partícióinak számát jelöli . Ezek és függvényében meghatározhatók az

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{r=0}^{k-1} (-1)^r \binom{k}{r} (k-r)^n$$

egzakt formulával $(n = 1, 2, \ldots; k = 1, 2, \ldots, n)$.

A W veszteségfüggvény kiértékelése a kombinatorikusan lehetséges véges számú esetre elvileg keresztülvihető, a gyakorlatban azonban nagyon időigényes lenne, ui. be lehet látni (l. [20]), hogy $\left\{ {n \atop n-k} \right\}$ az n-nek $2k$-fokú polinomja (8 objektum, 4 klaszter esetén is $\left\{ {8 \atop 4} \right\} = 1701$ lehetőséget kellene végigszámolnunk). Nézzünk helyette inkább egy jól bevált algoritmust:

k-közép (MacQueen) módszer: a minimalizálandó veszteségfüggvény

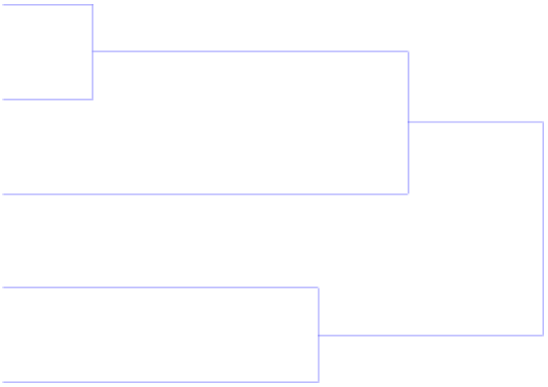
$$W = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{s}_j\|^2.$$

Itt k adott (geometriai vagy előzetes meggondolásokból adódik), és induljunk ki egy kezdeti $\mathcal{C}_1^{(0)}, \ldots, \mathcal{C}_k^{(0)}$ klaszterbesorolásból (pl. kiszemelünk k távoli objektumot, és mindegyikhez a hozzájuk közelieket soroljuk, egyelőre csak durva megközelítésben). Egy iterációt hajtunk végre, a lépéseket jelölje $m = 1, 2, \ldots$. Tegyük fel, hogy az $(m-1)$-edik lépésben az objektomoknak már létezik egy k klaszterbe sorolása: $\mathcal{C}_1^{(m-1)}, \ldots, \mathcal{C}_k^{(m-1)}$, a klaszterek súlypontját pedig jelölje $\mathbf{s}_1^{(m-1)}, \ldots, \mathbf{s}_k^{(m-1)}$ (a 0. lépésbeli besorolásnak a kezdő klaszterezés felel meg). Az m-edik lépésben átsoroljuk az objektumokat a klaszterek között a következőképpen: egy objektumot abba a klaszterbe sorolunk, melynek súlypontjához a legközelebb van. Pl. $\mathbf{x}_i$-t az l. klaszterbe rakjuk, ha

$$\|\mathbf{x}_i - \mathbf{s}_l^{(m-1)}\| = \min_{j \in \{1,\ldots,k\}} \|\mathbf{x}_i - \mathbf{s}_j^{(m-1)}\|$$

(ha a minimum több klaszterre is eléretik, akkor a legkisebb indexű ilyenbe soroljuk be), azaz $\mathbf{x}_i \in \mathcal{C}_l^{(m)}$ lesz. Kétféle módon is el lehet végezni az objektumok átsorolását: vagy az összes objektumot átsoroljuk az $(m-1)$-edik lépésben kialakult klaszter-súlypontokkal számolva, majd a régi súlypontok körül kialakult új klaszteknek módosítjuk a súlypontját, vagy pedig az objektumokat $\mathbf{x}_1, \ldots, \mathbf{x}_n$ szerint sorravéve, mihelyt egy objektum átkerül egy új klaszterbe, módosítjuk annak súlypontját. Így a végén nem kell már újra súlypontokat számolnunk, és az iterációszám is csökkenhet, ui. célratörőbb (``mohó'') az algoritmus. Miután az összes objektumot átsoroltuk, az új $\mathcal{C}_1^{(m)}, \ldots, \mathcal{C}_k^{(m)}$ klaszterezásből és az új $\mathbf{s}_1^{(m)}, \ldots, \mathbf{s}_k^{(m)}$ súlypontokból kiindulva ismét teszünk egy lépést. Meddig? Választhatunk többféle leállási kritériumot is, pl. azt, hogy az objektumok már stabilizálódnak a klaszterekben, és a klaszterek nem változnak az iteráció során. Az eljárást *animáció* szemlélteti.

Az *agglomeratív ill. divizív* módszerek a klaszterszámot fokozatosan csökkentik ill. növelik. Ezek közül is az ún. *hierarchikus* eljárások terjedtek el, ahol úgy csökkentjük ill. növeljük a klaszterszámot, hogy minden lépésben bizonyos klasztereket összevonunk ill. szétvágunk. Például nézzünk egy agglomeratív, hierarchikus eljárást. A kezdeti klaszterszám $k^{(0)} = n$, tehát kezdetben minden objektum egy külön klasztert alkot. Az iteráció a következő: tegyük fel, hogy az m. lépésben már csak $k^{(m)}$ db. klaszterünk van. Számítsuk ki a klaszter-középpontokat (súlypontokat). Ezek euklideszi távolságai egy $k^{(m)} \times k^{(m)}$-es, szimmetrikus ún. távolság-mátrixot alkotnak (főátlója 0). Azokat a klasztereket, melyek távolsága egy adott korlátnál kisebb, egy klaszterbe vonjuk össze, ilyen módon egy lépésben persze kettőnél több klaszter is összevonódhat. Végül, legfeljebb n lépésben már minden összeolvad, és csak egy klaszterünk lesz.



dendrogram???

A mellékelt ún. *dendrogram* (l. 9.1 ábra) egy agglomeratív eljárást szemléltet (5 objektummal). Az eljárás megtekinthető *animáción* is. Nem szükséges persze végigcsinálni az összes lépést. Agglomeratív eljárások esetén a W veszteségfüggvény általában monoton nő, azt kell megfigyelni, hol ugrik meg drasztikusan. Ha végigcsináljuk az összes lépést, a dendrogramot szemlélve próbálunk meg egy ésszerű klaszterszámot találni (a mellékelt példában lehetne ez 2). Ilyen agglomeratív, hierarchikus eljárás a *legközelebbi szomszéd* módszer is, amely akkor is összevon két klasztert, ha létezik közöttük egy lánc, amelyben az egymás utáni elemek már közelebb vannak egymáshoz egy adott korlátnál. Ezt az algoritmust Kruskal dolgozta ki (l. [18]).

1.2. Többdimenziós skálázás

Tegyük fel, hogy n db. objektum mindegyikén végeztünk p számú megfigyelést (n és p viszonya most tetszőleges). Célunk az objektumok vagy/és változók megjelenítése valamely (lehetőleg alacsony dimenziós) euklideszi tér pontjaiként. Amenynyiben megfigyeléseink egy $n \times p$-es adatmátrix formájában vannak megadva, ennek sorai tekinthetők az objektumokat reprezentáló p-, oszlopai pedig a változókat reprezentáló n-dimenziós pontoknak. A probléma az, hogy n és p általában ``nagy'', mi pedig inkább 1-,2-, esetleg 3-dimenziós ábrákon szeretnénk tájékozódni. Előfordulhat az is, hogy nincsen szabályos adatmátrixunk, hanem csak az objektumok vagy/és változók közti ún. hasonlósági vagy különbözőségi mérőszámok adottak, és csupán ezek alapján szeretnénk reprezentálni adatainkat. A következőkben az objektumok alacsony dimenziós reprezentálásával (skálázásával) fogunk foglalkozni. A leírtak értelemszerűen alkalmazhatók a változókra is. A precíz tárgyaláshoz bevezetünk néhány definíciót és jelölést.

9.1.2.1. Definíció. *A* $\mathbf{D} = (d_{ij})_{i,j=1}^{n}$ *mátrixot* távolság-mátrix*nak nevezzük, ha*

* *(i)* $d_{ii} = 0, \ i = 1, \ldots, n;$

* *(ii)* $d_{ij} = d_{ji} \geq 0, 1 \leq i < j \leq n; d_{ik} \leq d_{ij} + d_{jk}, \ i, j, k \in \{1, \ldots, n\}.$

9.1.2.2. Definíció. *Az* $n \times n$*-es* $\mathbf{D}$ *távolságmátrixot* euklideszi*nek nevezzük, ha valamely* p *pozitív egész mellett vannak olyan* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ *vektorok, hogy*

$$d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j|| \qquad (i, j = 1, \ldots n).$$

Legyen $\mathbf{H}_n := \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ az ún. *centráló mátrix*. Miután n-et rögzítettük, a $\mathbf{H}$ mátrix alsó indexét elhagyjuk. A következő tétel szükséges és elégséges feltételt ad arra, hogy egy távolságmátrix euklideszi legyen.

9.1.2.3. Tétel. *Az* $n \times n$*-es* mxD *távolság-mátrix akkor és csak akkor euklideszi, ha a* $\mathbf{B} := \mathbf{HAH}$ *mátrix pozitív szemidefinit, ahol az* $\mathbf{A}$ *mátrix elemei:* $a_{ij} = -\frac{1}{2}d_{ij}^2$.

A Tételt nem bizonyítjuk, de megmutatjuk, hogy ha a $\mathbf{B}$ mátrix pozitív szemidefinit, akkor hogyan találjuk meg egy alkalmas $\mathbb{R}^p$ euklideszi térben a pontoknak megfelelő vektorokat. Mivel $\mathbf{B}$ Gram-mátrix előáll $\mathbf{B} = \mathbf{X}\mathbf{X}^{top}$ alakban, ahol $\mathbf{X}$ egy $n \times p$,átrix, melynek sorai az $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ vektorok. Ekkor igaz a $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$ összefüggés.

Általában semmi garancia nincs arra, hogy a $\mathbf{D}$ távolságmátrix euklideszi. Ha $\mathbf{D}$ nem euklideszi, akkor 131 Tételben szereplő $\mathbf{B}$ mátrix indefinit. Tegyük fel, hogy az $n \times n$-es $\mathbf{B}$-nek p darab pozitív sajátértéke van

$(\lambda_1(\mathbf{B}) \geq \cdots \geq \lambda_p(\mathbf{B}))$ és a $\mathbf{B} = \mathbf{U}\Lambda\mathbf{U}^\top$ spektrálfelbontásbeli Λ-ban a sajátértékek nem-növekvő sorrendben vannak rendezve. Az 153 Tétel (Weyl perturbációs tétel) szerint tetszőleges $\mathbf{B}_p$ szimmetrikus mátrixra

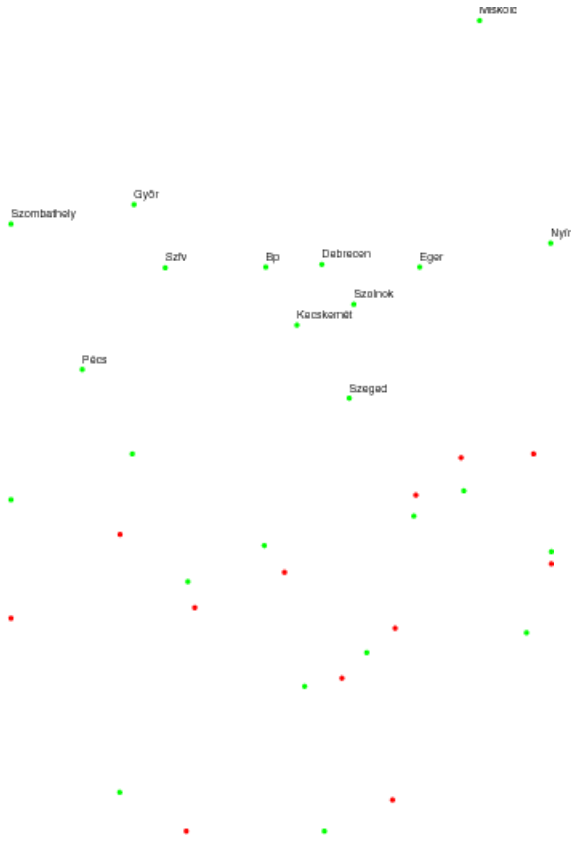
$$\max_j |\lambda_j(\mathbf{B}) - \lambda_j(\mathbf{B}_p)| \leq \|\mathbf{B} - \mathbf{B}_p\|.$$

A fenti egyenlőtlenség bal oldalának minimuma a p rangú, pozitív szemidefinit $\mathbf{B}_p$ mátrixok körében a $\mathbf{B}$ mátrix legnagyobb abszolút értékű negatív sajátértéke. A $\widehat{\mathbf{B}}_p = \sum_{i=1}^{p} \lambda_i(\mathbf{B})\mathbf{u}_i\mathbf{u}_i^T$ mátrixon ez a minimum eléretik. Ily módon $\widehat{\mathbf{B}}_p$-ből a fenti módon konstruált $\widehat{\mathbf{D}}$ távolságmátrixot a $\mathbf{D}$ mátrix euklideszi távolságmátrixszal való optimális közelítésének tekinthetjük.

Városok eredeti pozíciójukban

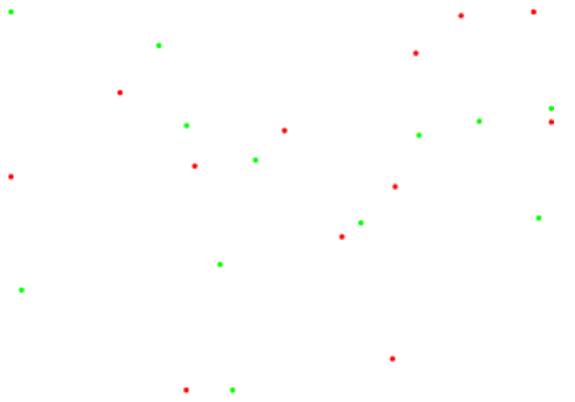
Városok közelítése légvonalbeli távolságmátrix alapján

Városok közelítése közúton mért távolságmátrix alapján

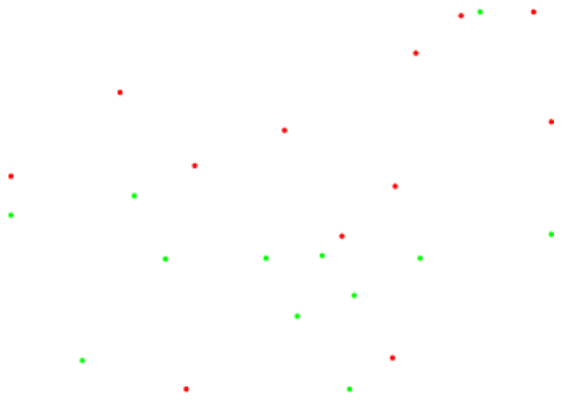


Városok közelítése Manhattan távolságmátrix alapján

Eredeti és légvonalban mért távolságmátrix alapján kapott térkép



Eredeti és közúton mért távolságmátrix alapján kapott térkép



Eredeti és Manhattan távolságmátrix alapján kapott térkép

10. fejezet - Randomizált módszerek nagyméretű problémákra

1. Elméleti háttér

A többváltozós statisztikai módszerek jelentős része (faktor-, klaszter és korrespondenciaanalízis) valamely mátrix spektrális vagy szinguláris felbontásán alapul, s mivel a statisztika egyik célja nagy adattömeg leírása minél kevesebb adattal ezen módszerekben csak néhány kiugró saját- vagy szinguláris értéket és a hozzájuk tartozó sajátvektorokat, illetve sajátvektor párokat kell meghatároznunk. A napjainkban egyre elterjedtebb ún. adatbányászatnak is a szinguláris érték felbontás az alapja. Itt mátrixok mérete $(m \times n)$ milliószor milliós lehet, ugyanakkor a hagyományos szinguláris érték felbontási algoritmusok számításigénye $\mathcal{O}(\min mn^2, m^2n)$. Több kezdeti kísérlet után Frieze, A., Kannan, R., és Vempala, S. [13] javasoltak véletlen kiválasztáson alapuló hatékony módszert egy nagyméretű $\mathbf{A}$ mátrix k-nál kisebb rangú $\mathbf{D}^*$ mátrixszal való közelítésére. Az általuk alkalmazott véletlen kiválasztásnál a sorok kiválasztásának valószínűsége arányos a *sor euklideszi norma négyzete / $\mathbf{A}$ hyperref[?]Frobenius-norma négyzete* mennyiséggel, a soron belül az elemek kiválasztásának valószínűsége (feltéve, hogy az adott sort kiválasztottuk) arányos *az adott elem négyzete / $\mathbf{A}$ Frobenius-norma négyzete* mennyiséggel. Alaptételük a következőt állítja.

10.1.1. Tétel. *Legyen $\mathbf{A}$ egy $m \times n$ mátrix, legyen rögzítve $k \in \mathbb{Z}^+ \varepsilon > 0$ és $\delta > 0$. Ekkor van olyan véletlenített algoritmus, amely leírja azt a legfeljebb k-rangú $\mathbf{D}^*$ mátrixot amelyre lagalább $1 - \delta$ valószínűséggel teljesül a*

$$||A - \mathbf{D}^*||_F^2 \leq \min_{\mathbf{D}, \mathrm{rk}\mathbf{D} \leq k} ||\mathbf{A} - D||_F^2 + \varepsilon ||\mathbf{A}||_F^2.$$

Az algoritmus csak k-ban, $\frac{1}{\varepsilon}$-ban és $\log \frac{1}{\delta}$-ban polinomidejű, m-től és n-től független. Az így kapott leírás alapján $\mathbf{D}^*$ explicit módon kiszámítható $\mathcal{O}(kmn)$ lépésben.

A következő tétel Achlioptas-tól és McSherrytől származik [1]. Mielőtt kimondanánk bevezetjük egy $m \times n$-es $\mathbf{A}$ mátrixszal azonos méretű mátrixban meglevő minimális lineáris struktúrát mérő Ψ mennyiséget. Legyen $b = \max_{i,j} |a_{ij}|$ és legyen $\mathcal{Q}$ egy olyan $m \times n$-es $\mathbf{Q}$ mátrixok halmaza, amelyek elemei b-vel vagy $-b$-vel egyenlők.

$$\Psi(\mathbf{A}) = \min_{\mathbf{Q} \in \mathcal{Q}} ||\mathbf{Q}||$$

10.1.2. Tétel. *Legyen $\mathbf{A}$ tetszőleges $m \times n$-es mátrix és $s > 1$ tetszőleges valós szám. Legyen továbbá $\hat{\mathbf{A}}$ olyan $m \times n$-es véletlen mátrix, melynek elemi függetlenek és tetszőleges i, j indexpárra*

$$\hat{a}_{ij} = \begin{cases} 0, & 1 - \frac{1}{s} \text{ valószínűséggel} \\ sa_{ij}, & \frac{1}{s} \text{ valószínűséggel.} \end{cases}$$

(A függetlenség visszatevéses mintavétellel mindig elérhető)

Ha még

$$s \leq \frac{m+n}{11^6} \log^6(m+n)$$

is teljesül, akkor

$$\mathbb{P}\left(||\mathbf{A} - \hat{\mathbf{A}}_k|| \leq ||\mathbf{A} - \mathbf{A}_k|| + 7\sqrt{s}\Psi(\mathbf{A}) \right) \geq 1 - \frac{1}{m+n},$$

ahol $\mathbf{A}_k$, illetve $\hat{\mathbf{A}}_k$ jelöli az $\mathbf{A}$, illetve $\hat{\mathbf{A}}$ mátrixot legjobban közelítő k-rangú mátrixot.

A tétel bizonyítása azon alapszik, hogy az $\mathbf{A} - \hat{\mathbf{A}}$ mátrix alkalmas elrendezéssel Wigner-típusú mátrixszá alakítható. A Wigner-mátrixok maximális sajátértéke eloszlásának felső farkára jó becslések ismertek.

11. fejezet - Algoritmikus modellek

1. Elméleti háttér

1.1. ACE-algoritmus (általánosított regresszióra)

A Breiman és Friedman ([[6]]) által kifejlesztett algoritmus az alábbiakban vázolt általános regressziós feladat numerikus megoldására szolgál igen tág keretek között (kategorikus adatokra, idősorokra ugyanúgy alkalmazható, mint olyan többváltozós adatokra, ahol a változók egy része abszolút folytonos, más része diszkrét).

Az Y függő és az $X_1, \ldots, X_p$ független változóknak keresendők olyan $\Psi, \Phi_1, \ldots, \Phi_p$ mérhető, nem-konstans valós értékű függvényei (szkórjai), amelyekkel

$$e^2(\Psi, \Phi_1, \ldots, \Phi_p) = \mathbb{E}\left[\Psi(Y) - \sum_{j=1}^{p} \Phi_j(X_j)\right]^2 / \mathbb{D}^2(\Psi(Y)) \qquad (1)$$

???

minimális adott $\{(y_k, x_{k1}, \ldots, x_{kp} : k = 1, \ldots, n)\}$ adatrendszer alapján. Valójában feltételes minimumot keresünk a $\mathbb{D}^2(\Psi(Y)) = 1$ feltétel mellett.

Lineáris transzformációkkal elérhető, hogy $\mathbb{E}(\Psi(Y)) = \mathbb{E}(\Phi_1(X_1)) = \cdots = \mathbb{E}(\Phi_p(X_p)) = 0$ és $\mathbb{D}^2(\Psi(Y)) = 1$ legyen.

Amennyiben a változók együttes $(p+1)$-dimenziós eloszlása ismert, az algoritmus a következő. Legyenek $\Psi^{(0)}(Y), \Phi_1^{(0)}(X_1), \ldots, \Phi_p^{(0)}(X_p)$ a feltételeknek eleget tevő kezdeti függvények. Az iteráció $(m+1)$-edik lépése a következő (mindig csak egyik függvényt változtatjuk).

1. Rögzített $\Phi_1^{(m)}(X_1), \ldots, \Phi_p^{(m)}(X_p)$ esetén

$$\Psi^{(m+1)}(Y) := \frac{\mathbb{E}(\sum_{j=1}^{p} \Phi_j^{(m)}(X_j)|Y)}{\mathbb{D}(\sum_{j=1}^{p} \Phi_j^{(m)}(X_j)|Y)}.$$

2. Rögzített $\Psi^{(m+1)}, \Phi_1^{(m+1)}(X_1), \ldots, \Phi_{i-1}^{(m+1)}(X_{i-1}), \Phi_{i+1}^{(m)}(X_{i+1}), \ldots, \Phi_p^{(m)}(X_p)$ esetén

$$\Phi_i^{(m+1)}(X_i) := \mathbb{E}\left([\Psi^{(m+1)}(Y) - \sum_{j=1}^{i-1} \Phi_j^{(m+1)}(X_j) - \sum_{j=i+1}^{p} \Phi_j^{(m)}(X_j)]|X_i\right)$$

$i = 1, \ldots, p.$

Az iterációt akkor hagyjuk abba, ha a (11.1)-beli célfüggvény értéke már ``keveset'' változik.

Az algoritmust részletesebben leírjuk abban az esetben, amikor a valószínűségi változók ismeretlen folytonos eloszlásúak, és a feltételes várható érték vételt a simítás helyettesíti.

Nyilván világos az algoritmus elnevezése: ACE=Alternating Conditional Expectation (alternáló feltételes várható érték).

Ha az együttes eloszlást nem ismerjük, az n mintaelemet tartalmazó adatrendszer alapján minimalizálandó célfüggvényt akkor is felírhatjuk

$$\frac{1}{n}\sum_{k=1}^{n}\left[\Psi(y_k) - \sum_{j=1}^{p} \Phi_j(x_{kj})\right]^2$$

alakban, melyet azzal a kényszerfeltétellel minimalizálunk, hogy $\Psi(Y)$ empirikus szórásnégyzete 1. Az iterációs lépések a fentie$p = 1$al a különbséggel, hogy a feltételes várható értéket is a minta alapján képezzük. Például 2 változó esetén ( ) ennek becslése a következő:

$$\hat{\mathbb{E}}(\Phi(X)|Y = y) = \sum_{k:x_k=x} \Phi(x_k) / \sum_{k:y_k=y} 1,$$

vagyis átlagoljuk az azonos Y értéket felvevő mintaelemekhez rendelt $\Phi(x_k)$-kat Y összes megfigyelt értékére. Pl. ha Y a szemszín és $\Phi(X)$ a hajszín szkórja, akkor átlagoljuk az azonos szemszínűek hajszín-szkórjait, majd átlagoljuk az azonos hajszínűek az $\Psi(y)$ szemszín-szkórjait, és normálunk. Az algoritmus lényege éppen abban áll, hogy ezt felváltva hajtjuk végre, miközben a másik változót rögzítjük.

A fenti algoritmus ismeretlen mintaeloszlások esetén csak akkor működik, ha a tapasztalati feltételes várható értékek kiszámíthatók, azaz a minta együttes eloszlása diszkrét. Breiman és Friedman a minták simításának módszerét ajánlották folytonos valószínűségi változók esetére. A jelölésekben - melyek kissé eltérnek a szokásostól - az idézett dolgozatot követjük.

Jelölje $\mathbf{X}$ az adathalmazt (mintát), azaz az $\mathbb{R}^p$ euklideszi tér N pontjából álló $\{\mathbf{x}_1, \ldots \mathbf{x}_N\}$, azaz

$$\begin{pmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & & \vdots \\ x_{N1} & \ldots & x_{Np} \end{pmatrix}$$

adatmátrixot. Rögzített $\mathbf{X}$-re legyen $F(\mathbf{X})$ az összes $\mathbf{X}$-en értelmezett valósértékű Φ fügvények tere, azaz egy $\Phi \in F(\mathbf{X})$ függvényt N valós szám $(\{\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_N\}))$ definiál. Legyen továbbá $F(x_j)(j = 1, \ldots, p)$ az összes $\{x_{1j}, \ldots, x_{Nj}\}$ halmazon értelmezett valósértékű függvények tere.

11.1.1.1. Definíció. *Az $\mathbf{X}$ mintára értelmezett $S\colon F(\mathbf{X}) \mapsto F(x_j) S_j$ függvényt az $\mathbf{X}$ minta x_j szerinti simításának nevezzük. Ha $\Phi \in F(\mathbf{X})$, jelöljük az $F(x_j)$ térben S_j képét $S_j(\Phi|x_j)$-vel, a függvény értékét a k-adik adaton pedig $S_j(\Phi|x_{kj})$-vel*

Feltesszük, hogy az alábbi tulajdonságok teljesülnek.

(i) Linearitás: minden $\Phi_1, \Phi_2 \in F(\mathbf{X})$, valamint minden valós α és β számra

$$S(\alpha\Phi_1 + \beta\Phi_2) = \alpha S\Phi_1 + \beta S\Phi_2.$$

(ii) Konstans megőrzés: ha $\Phi \in D$ azonosan konstans ($\Phi \equiv c$), akkor $S\Phi = \Phi$.

(iii) Korlátosság: Az S simítás korlátja M, ha minden $\Phi \in F(\mathbf{X})$-re

$$||S\Phi||_N \leq M||\Phi||_N,$$

ahol $||\cdot||_N$ az Np dimenziós euklideszi norma. (Egy $\mathbf{X}$ minta N darab p dimenziós vektorból áll!)

Példák.

1. Legközelebbi szomszéd módszer: Rögzitsünk egy $M < \frac{N}{2}$ természetes számot. Rendezzük a mintát a j-edik koordinatája szerint. Az itt alkalmazott jelölésekben ez azt jelenti, hogy $x_{1j} < x_{2j} < \cdots < x_{Nj}$; feltesszük, hogy nincsenek egyenlő elemek. Legyen

$$S(\Phi|x_{kj}) = \frac{1}{2M} \sum_{m=-M, m\neq 0}^{N} \Phi(\mathbf{x}_{k+m}).$$

Ha valamelyik oldalon (pl. a végén) már nincs M pont, egészítsük ki az összegzést a másik oldalról (pl. az elejéről) vett pontokkal.

2. Magfüggvény módszer: Legyen $K(x)$ olyan valós nemnegatív értékű függvény, amely maximumát a 0 pontban veszi fel.

Legyen

$$S(\Phi|x_{kj}) = \frac{\sum_{m=1}^{N} \Phi(\mathbf{x}_m) K(x_{mj} - x_{k,j})}{\sum_{m=1}^{N} K(x_{mj} - x_{k,j})}$$

Vegyük észre, hogy ha a j-edik változó szerint simítunk, akkor lényegében a $\Phi(\mathbf{x})$ függvényt átlagoljuk a j-edik változó mentén, ez felel meg a megfelelő feltételes várható érték vételnek.

Most egy kettős ciklussal definiáljuk a Breiman- Friedman numerikus algoritmust. Az algoritmus külső ciklusában θ-t, belső ciklusában Φ_j-ket $j = 1, \ldots, p$ változtatjuk. A külső ciklus n-edik lépése után e szerzők két lehetőséget javasolnak:

(a) Megtartjuk a belső ciklusban kapott Φ-k értékeit (restart),

(b) Kinullázzuk a korábbi Φ értékeket (friss start).

Kettős ciklus.

0. Inicializálás:

$$\theta^{(0)}(y_k) = y_k \quad \Phi_j^{(0)}(y_{kj}) = 0.$$

1. Külső ciklus ($n = 1, 2, \ldots$-re): legyen $\theta^{(n)} = S_y(\sum_j^p \Phi_j)/||S_y(\sum_j^p \Phi_j)||_N$. Térjünk vissza a belső ciklushoz minden j-re $\Phi_j^{(0)} = \Phi_j$-vel (restart) vagy minden j-re $\Phi_j^{(0)} = 0$-val (friss start).

2. Belső ciklus ($m = 0, 2, \ldots$-re): a külső ciklus n-edik szintjén $\theta^{(n)}$-nel és $\Phi_j^{(0)}$-vel ($j = 1, \ldots, p$) kezdünk.

Futtasuk a legbelső ciklust m-et növelve.

3. Legbelső ciklus (j-re, m fix): $j = 1, 2, \ldots, p$. Legyen

$$\Phi_j^{(m+1)} = S_j \left(\theta^{(n)} - \sum_{i<j} \Phi_i^{(m+1)} - \sum_{i>j} \Phi_i^{(m)} \right) \qquad (1)$$

???

3' Legbelső ciklus vége.

2' A belső ciklus megáll ha $\sum_{j=1}^{p} ||\Phi_j^{(m+1)} - \Phi_j^{(m)}||_m$ növelésével alig változik.

1' A külső ciklus megáll, ha $||\theta^{(n)} - \sum_{j=1}^{p} \Phi_j||_n$ növelésével alig változik.

Kettős ciklus vége.

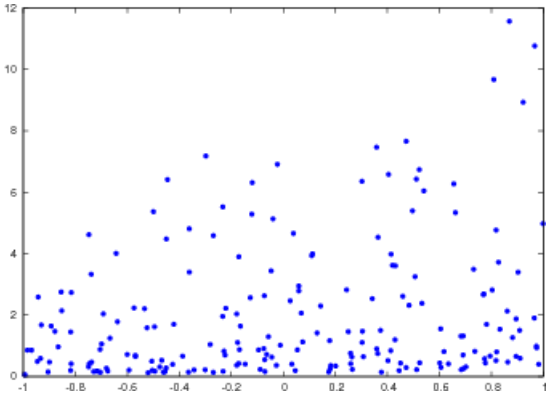
11.1.1.2. Megjegyzés. *Vegyük észre, hogy*

1. A belső ciklusban, amikor a j-edik változó szerint simítunk, (a (11.2) formula) akkor $\theta^{(m)} - \sum_{i<j} \Phi_i^{(m+1)}$-nek a j-edik változó szerinti feltételes várható értékét vesszük.

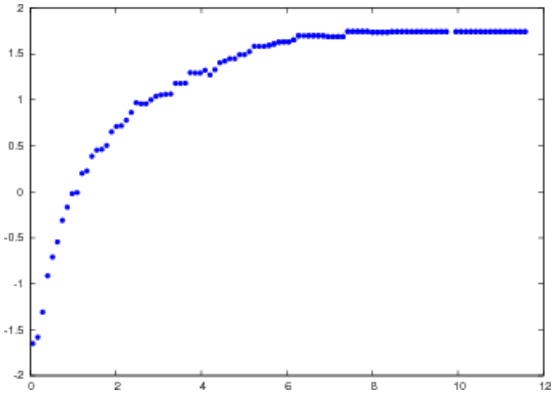
2. A külső ciklusban az y változó szerint simítunk, ezt formálisan nem definiáltuk, de belevehettük volna az $\mathbf{X}$ mintába, $p + 1$-edik változóként.

A fenti algoritmus konvergenciáját A Breiman és Friedman ([6]) speciális, nehezen ellenőrizhető feltételek mellett igazolták. A gyakorlat azt mutatja, hogy a módszer a feladatok széles körére jól alkalmazható.

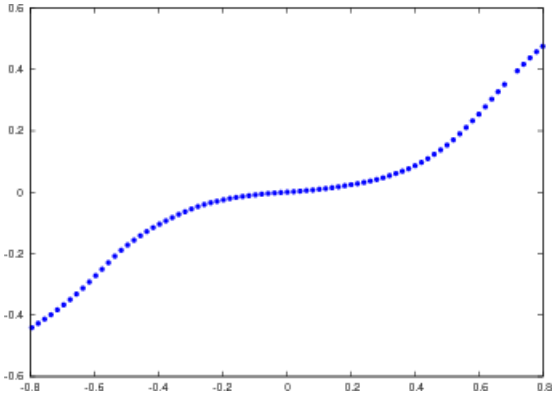
Az alábbiakban az ACE algoritmus működését illusztráljuk egy 200 elemből álló adathalmazon; $p = 1$, $y_k = \exp\{|x_k^3 + \varepsilon_k|\}$, ahol x_k és ε_k független azonos standard normális eloszlású valószínűségi változók. A simítás legközelebbi szomszéd módszerrel történt. Az 1. ábrán a nyers adatokat, a 2. illetve a 3. ábrán a $\theta(y)$, illetve a $\Phi(x)$ függvények becslése látható.



Kettős ciklus - adatok



Kettős ciklus - $\theta(y)$ becslése



Kettős ciklus - $\Phi(x)$ becslése

1.2. Jackknife eljárás

Az M. H. Quenouille [28] által 1954-ben által javasolt, a becslés torzítását csökkentő módszernek J. W. Tukey [32] adta a jackknife (zsebkés) elnevezést. Az elnevezés azt fejezi ki, hogy maga az eljárás - elsősorban kis minták esetén - számos más célra is alkalmazható, mert a normális eloszlásra kidolgozott módszereket jól imitálja olyan esetekben is, amikor a normalitás sérül. A jackknife azonban nem mindenre jó gyógyszer, egy egyszerű ellenpéldán megmutatjuk korlátjait.

A jackknife az adatok jól megválasztott csoportosításán alapszik, a csoportok kombinációi alapján becsléseket konstruálunk, amelyek átlaga lesz a jackknife becslés. Itt csak az egyelemű csoportokat használó eljárást ismertetjük.

A jackknife módszer alábbi vázlatos ismertetésében Rupert Miller [23] és [24] dolgozataira támaszkodunk.

Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta egy $\mathbb{P}_\theta$ eloszlásból, ahol $\theta \in \Theta$ ismeretlen paraméter. Jelölje $\hat{\theta} := \hat{\theta}(\mathbf{X})$ a θ paraméter valamilyen becslését a teljes minta alapján; a továbbiakban a becslések argumentumába nem írjuk be a mintaelemeket. Jelölje $\hat{\theta}_{-i} (i = 1, \ldots, n)$ azt a becslést, amelyet az i-edik mintaelem elhagyásával kapunk. Képezzük az ún. *pszeudoértékeket* (az elnevezés Tukey-től származik):

$$\widetilde{\theta}_i := n\hat{\theta} - (n-1)\hat{\theta}_{-i} \qquad (1)$$

???

11.1.2.1. Definíció. *A θ paraméter* jackknife becslése *a $\widetilde{\theta}_i$ pszeudoértékek átlaga:*

$$\widetilde{\theta}_\bullet = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-\bullet}, \qquad (1)$$

ahol $\hat{\theta}_{-\bullet} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}$.

11.1.2.2. Állítás. *A jacknife becslés pontosan eliminálja a torzítás $\frac{1}{n}$ rendű tagját.*

Mivel ez az állítás éppen a jackknife-becslés alapvető tulajdonságát jellemzi (tulajdonképpen ezt a célt valósítja meg az eljárás) közöljük a rövid és tanulságos bizonyítást.

Bizonyítás Ha $\mathbb{E}(\hat{\theta}) = \theta + \frac{a}{n} + \frac{b}{n^2} + \cdots$, akkor

$$\mathbb{E}(\widetilde{\theta}_\bullet) = n(\theta + \frac{a}{n} + \frac{b}{n^2} + \ldots) - (n-1)(\theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + \cdots = \theta - \frac{b}{n(n-1)} + \ldots)$$

QED

Tukey szerint a $\widetilde{\theta}_i$ pszeudoértékek közelítőleg függetlenek; ha ez a feltevés igaz, akkor $\mathbb{D}^2(\widetilde{\theta}_\bullet)$ becslése az

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\widetilde{\theta}_i - \widetilde{\theta}_\bullet)^2 \qquad (1)$$

???statisztika lehet, és a

$$t = (\widetilde{\theta}_\bullet - \theta)\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(\widetilde{\theta}_i - \widetilde{\theta}_\bullet)^2\right]^{-1/2} \qquad (1)$$

???

statisztika közelítőleg $t(n-1)$ eloszlású, így alkalmas hipotézisvizsgálatra és konfidenciaintervallum szerkesztésre. Ezt illusztráljuk a következő példán.

Legyen $X_1, \ldots, X_n$ független, azonos $F((x-\mu)/\sigma)$ eloszlású minta, ahol F ismeretlen eloszlásfüggvény μ és σ ismeretlen lokációs és skálaparaméterekkel ($\mu = \mathbb{E}(X_1)$, $\sigma^2 = \mathbb{D}^2(X_1)$). Tegyük fel, hogy F-nek létezik a negyedik momentuma. A σ^2 paraméter torzítatlan becslése

$$S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Alkalmazzuk a jackknife eljárást!

$$\begin{aligned}\widetilde{\theta}_i &= S_n^{*2} + \frac{n}{n-2}\left((X_i - \bar{X})^2 - n^{-1}\cdot\sum_{j=1}^{n}(X_j - \bar{X})^2\right),\\ \widetilde{\theta}_\bullet &= S_n^{*2} \quad \text{és} \qquad (1)\\ \sum_{i=1}^{n}(\widetilde{\theta}_i - \widetilde{\theta}_\bullet)^2 &= \frac{n^2}{(n-2)^2}\sum_{i=1}^{n}\left((X_i - \bar{X})^2 - n^{-1}\sum_{j=1}^{n}(X_j - \bar{X})^2\right)^2.\end{aligned}$$

???Ahogyan az (11.5) becslés alapján megkonstruáltuk az (5.4) statisztikát, az (11.7) statisztikák alapján σ^2 paraméter $\widetilde{\theta}_\bullet$ jackknife becslésére (ami itt azonos a hagyományos S_n^{*2} torzítatlan becsléssel!) közelítő t-statisztikát konstruálhatunk:

$$t = (\widetilde{\theta}_\bullet - \sigma^2) \left[\frac{1}{n(n-1)} \sum_{i=1}^{n} (\widetilde{\theta}_i - \widetilde{\theta}_\bullet)^2 \right]^{-1/2}.$$

Egy - kissé mesterkélt - ellenpéldán megmutatható, hogy az (11.6) statisztika eloszlása erősen eltérhet az $n-1$ szabadsági fokú Student-eloszlástól. A példát nem ismertetjük.

jackknife módszer a diszkriminanciaanalízis kereszt-kiértékelésére. Tegyük fel, hogy N elemű $(X_1, \ldots, X_N)$ mintára alkalmazunk egy tetszőleges diszkrimináló eljárást. A következőt kell tennünk: az eljárást N-szer végrehajtjuk úgy, hogy kihagyjuk az $X_i, i = 1, \ldots, N$ mintaelemet, majd megnézzük, hogy a kihagyott (X_i) elemet melyik osztályba sorolta az így szerkesztett eljárás. A kapott eredményeket átlagolva megkapjuk a hibás (és természetesen a helyes) besorolások relatív gyakoriságát.

1.3. Bootstrap eljárás

A paragrafusnak ebben a részében elsősorban A. B. Efron 1997-ben megjelent alapvető [10] dolgozatára, valamint G. J. Babunak és C. Radhakrishna Rao-nak a Handbook of Statistics [2] 9. kötetében megjelent összefoglaló ismertetésére, és az abban idézett irodalomra támaszkodunk. A paragrafus elején ismertetett jackknife algoritmus elsősorban arra alkalmas, hogy valamely eloszlás ismeretlen paraméterének a torzítását csökkentse, és számos esteben jó közelítést adjon a becslés szórásnégyzetére. Az Efron által javasolt *bootstrap* (szó szerint csizmahúzó); a statisztikán kívül pl. az informatikában is használatos elnevezés a bonyolult problémákat kezelő általános receptekre) módszerrel a becslő statisztikák eloszlása is jól kezelhető.

A bootstrap statisztika definíciója és eloszlásának meghatározása. Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ független minta egy tetszőleges F eloszlásból, és legyen $T(\mathbf{X}, F)$ az $\mathbf{X}$ mintától függő statisztika.

A korábbi - a paraméteres statisztikával foglalkozó fejezetekben - F-ről általában feltettük, hogy normális eloszlású, és ekkor a gyakran alkalmazott $T(\mathbf{X}, F)$ statisztikák eloszlását analitikusan is meg tudtuk határozni. Más esetben - ha statisztika független azonos eloszlású valószínűségi változók normált összege volt - a centrális határeloszlás-tételre hivatkoztunk.

Kis mintaelemszám és ismeretlen F esetén a $T(\mathbf{X}, F)$ statisztika eloszlását közelíthetjük a mintából becsült $\hat{F}_n$ empirikus eloszlás alapján számított eloszlással. Megjegyezzük, hogy pl. az $\bar{\mathbf{X}}$ átlag eloszlásának kiszámításához az $\hat{F}_n$ n-szeres konvolúcióra van szükség, amelynek műveletigénye $\mathcal{O}((\log n)n^2)$, ami elfogadható, ennek ellenére a bonyolultabb statisztikák eloszlásának az $\hat{F}_n$ empirikus eloszlás alapján történő közvetlen meghatározása körülményes. Erre is alkalmas az Efron [10] által javasolt bootstrap eljárás.

A bootstrap statisztika eloszlása meghatározásának laggyakrabban használt módszere a „nyers rő", azaz a Mont Carlo módszer. Rögzített $\hat{F}$-hez vegyünk egy függatlen azonos $(\hat{F}_n)$ eloszlású $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ ún. bootstrap mintát. Ez a gyakorlatban azt jelenti, hogy az eredeti $\mathbf{X}$ mintából visszatevéssel kiválasztunk n elemet.

Ennél szofisztikáltabb módszer a centrális határeloszlás-tétel élesítésének alkalmazása a bootstrap mintára. Ha az $F(x)$ folytonos eloszlás harmadik abszolút momentuma véges, akkor a klasszikus Berry- Esseen-tétel (l. pl [15] szerint

$$\sup_x \left| \mathbb{P}\left(\bar{\mathbf{X}} - \mu \leq x\sigma\right) - \Phi(x) \right| = \mathcal{O}(n^{-1/2}) \qquad (1)$$

???

Ez az egyenlőtlenség nem javítható, de ha az F eloszlásnak létezik a k-adik $(k > 3)$ abszolút momentuma, akkor a (11.8) képletben szereplő explicit módon megadható, és a különbség rendje $\mathcal{O}\left(1/\sqrt{n^{s-2}}\right)$ lesz (Ljapunov tétele l. [15]). Mivel az $\hat{F}_n$ eloszlás momentumai megegyeznek a tapasztalati momentumokkal, az idézett tétel alkalmazható az $\hat{F}_n$ eloszlás analitikus alakban történő közelítésére ($\mathbf{X}$ helyett $\tilde{\mathbf{X}}$, $\mu = overline\mathbf{X}$ szereposztással).

Most megfogalmazunk egy tételt, amely az $\overline{\mathbf{X}}$ és bootsrap minta átlaga közötti eltérésére állít a (11.8) egyenlőtlenségnél pontosabb becslést. Mielőtt ezt kimondanánk, emlékeztetünk a rácsos eloszlás fogalmára: egy F eloszlás rácsos, ha növekedési pontjainak halmaza $\mathbb{R}$ ekvidisztáns pontjaiból áll. Az $\hat{F}$ eloszlás szerinti mértéket $\tilde{\mathbb{P}}$-vel jelöljük. K. Singh (l. [31]) tétele:

11.1.3.1. Tétel *Tegyük fel, hogy* $\mathbf{X} = (X_1, \ldots, X_n)$ *független minta egy* F *n*$\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)$*ból,*$\hat{F}_n$ *melynek várható értéke szórása és a harmadik abszolú*$(X_1, \ldots, X_n, \ldots)$*es. Legyen az alapján kisorsolt bootstrap minta. Ekkor majdnem minden realizációra*

$$\sup_x \left| \mathbb{P}\left((\bar{\mathbf{X}} - \mu) \le \sigma\right) - \tilde{\mathbb{P}}\left((\bar{\tilde{\mathbf{X}}} - \bar{\mathbf{X}}) \le x \sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{\mathbf{X}})^2}\right) \right| = o(n^{-1/2})$$

A következő - Babutól származó - példa (l. [2]) illusztrálja, hogy nem lehet vakon bízni a bootstrap módszerben. Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ standard normális eloszlásból származó független minta. Mivel $\sqrt{n}\bar{\mathbf{X}}$ standard normális eloszlású, $\mu = 0$, $n(\bar{\mathbf{X}})^2 - \mu^2 \sim \chi^2(1)$. Legyen $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ a bootstrap minta. Megmutatható, hogy az $(\bar{\tilde{\mathbf{X}}}^2 - \bar{\tilde{\mathbf{X}}}^2)$ majdnem minden végtelen $(X_1, \ldots, \mathbf{X}_n, \ldots)$ realizációra divergál!

ebből feladat gyártható: miért mond ez látszólag ellent a Steiner egyenlőtlenségnek?

Második példánk a diszkriminanciaanalízis hibabecslése. Az egyszerűseg kedvéért tegyük fel, hogy csak két mintánk van:

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim F = \mathcal{N}(\mathbf{m}_1, \mathbf{C})$$

és

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_m \sim G = \mathcal{N}(\mathbf{m}_2, \mathbf{C}),$$

ahol az $\mathbf{X}_i$ és $\mathbf{Y}_j p$-dimenziós véletlen vektorok teljesen függetlenek. A megfigyelt értékek: $\mathbf{x}_1, \ldots \mathbf{x}_n$, illetve $\mathbf{y}_1, \ldots, \mathbf{y}_m$. A minta alapján megbecsüljük az $\mathbf{m}_1$ és $\mathbf{m}_2$ várhatóérték vektort, valamint a $\mathbf{C}$ kovariancaiamátrixot, legyenek a becslések: $\widehat{\mathbf{m}}_1$, $\widehat{\mathbf{m}}_2$ és $\widehat{\mathbf{C}}$. Ezeket a becsléseket a **A diszkrdec25.tex-beli szovegben most szamozatlan a regi konyvben 311. o. 2.9 en itt nem tudom beirni...** formulába beírva eljárást kapunk arra, hogy eldöntsük: egy új $\mathbf{x}$ megfigyelést az F vagy a G eloszlást követi-e. Ha

$$\mathbf{x} \in B := \{\mathbf{x} : (\widehat{\mathbf{m}}_2^T - \widehat{\mathbf{m}}_1^T)\widehat{\mathbf{C}}^{-1}\mathbf{x} > c\}$$

akkor az $\mathbf{x}$ megfigyelést a G eloszlást követők csoportjába soroljuk. Az osztályozás várható hibáját még az új megfigyelések beérkezése előtt szeretnénk megbecsülni. Az

$$\widehat{\text{error}} := \frac{|\{i : \mathbf{x}_i \in B\}|}{m} \qquad (1)$$

???nyilván alulbecsüli a hibát, mert az osztályozó eljárást a minta alapján szerkesztettük, az mintegy adaptálódott a mintához. A valódi várható hiba

$$\text{error} := \mathbb{P}_F\{i : \mathbf{x}_i \in B\}$$

lenne.

$$R((\mathbf{X}, \mathbf{Y}), (F, G)) := \text{error} - \widehat{\text{error}}.$$

Az $\widetilde{R}$ bootstrap veszteség momentumait ``nyers erő"-vel (Monte Carlo módszerrel) határozhatjuk meg. Az $\hat{F}$ és $\hat{G}$ eloszlásból generálunk n, illetve $m\widetilde{\mathbf{x}}_i$, illetve $\widetilde{\mathbf{y}}_j$ bootstrap mintaelemet, ezek alapján kiszámítjuk az $\hat{F}$ és $\hat{G}$ eloszlások paramétereit, meghatározzuk a $\widetilde{B}$ bootstrap kritikus tartományt. Így az $\widetilde{R}$ bootstrap veszteség egy realizációja:

$$\widetilde{R} = R((\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}), (\hat{F}, \hat{G})) = \frac{|\{i : \mathbf{x}_i \in \widetilde{B}\}|}{m} - \frac{|\{i : \widetilde{\mathbf{x}}_i \in \widetilde{B}\}|}{m}.$$

Ezen eljárás elegendően sok független ismétlése után a keresett momentumok átlagolással nyerhetők. Ilymódon becslést kapunk az R veszteségfüggvény várható értékére, amivel az osztályozás hibájának (11.9) becslését korrigálhatjuk.

Megjegyezzük, hogy a programcsomagok kiszámítják a hibavalószínűség jackknife becslését is olymódon, hogy minden egyes mintaelem kihagyásával megszerkesztik a kritikus tartományt, majd megvizsgálják, hogy a

kihagyott elem melyik tartományhoz tartozik. Az így tapasztalt hibás döntések relatív gyakorisága a hibavalószínűség becslése. Efron idézett dolgozatában egy 10 és egy 20 elemű mintára ismerteti mindkét eljárás eredményét; nincs lényeges különbség.

2. Feladatok

(i) Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ standard normális eloszlásból származó független minta. Mivel $\sqrt{n}\bar{\mathbf{X}}$ standard normális eloszlású, $\mu = 0$, $n(\bar{\mathbf{X}})^2 - \mu^2 \sim \chi^2(1)$. Legyen $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{\mathbf{X}}_n)$ a bootstrap minta. Megmutatható, hogy az $(\bar{\tilde{\mathbf{X}}}^2 - \bar{\mathbf{X}}^2)$ majdnem minden végtelen $(X_1, \ldots, X_n, \ldots)$. Mutassuk meg, hogy ez az állítás látszólag ellentmond a Steiner-egynlőségnek.

Tipp: Az $\frac{1}{n}[\sum_{j=1}^{n}(\tilde{X}_j - \bar{\mathbf{X}})]^2$ valószínűségi változók aszimptotikusan valóban $\chi^2(1)$ eloszlásúak, Irjuk fel rájuk a Steiner-egyenlőséget, felhasználva, hogy $\mathbb{E}(\tilde{X}_j) = \bar{\mathbf{X}}$.

Válasz:

$$\frac{1}{n}\left[sum_{j=1}^{n}(\tilde{X}_j - \bar{\mathbf{X}})\right]^2 - (n\bar{\tilde{\mathbf{X}}}^2 - \bar{\mathbf{X}}^2) = 2\bar{\mathbf{X}}^2 - 2\bar{\mathbf{X}}\bar{\tilde{\mathbf{X}}}.$$

A fenti egyenlőség jobb oldala a nagy számok törvénye miatt 0-hoz tart, de finomabb meggondolások alapján kiderül, hogy ez nem elegendő az $(n\bar{\tilde{\mathbf{X}}}^2 - \bar{\mathbf{X}}^2)$ bootstrap statisztika eloszlás szerinti konvergenciájához.

12. fejezet - Függelék

1. Függelék 1: Lineáris algebrai emlékeztető

Jelölje $\mathbb{R}^n$ az n-dimenziós valós euklideszi teret (elemei n-dimenziós valós komponensűvektorok, melyek összeadása és valós számmal való szorzása értelmezve van a szokásos műveleti tulajdonságokkal, továbbá a vektortér a $\langle \cdot, \cdot \rangle$ skaláris szorzás műveletével is el van látva). Az $\mathbb{R}^n$ térben tekintsük a standard $\varepsilon_1, \ldots, \varepsilon_n$ bázist (az ε_i vektor i-edik koordinátája 1, többi koordinátája pedig 0). Ha a skaláris szorzást nem definiáljuk konkrét formulával, akkor fel kell tennünk, hogy az $\varepsilon_1, \ldots, \varepsilon_n$ bázis *ortonormált*:

$$\langle \varepsilon_i, \varepsilon_j \rangle = \delta_{ij} = \begin{cases} 0, & \text{ha } i \neq j \\ 1, & \text{ha } i = j. \end{cases} \qquad (1)$$

???Az $\mathbb{R}^n$ vektorait $\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$-vel jelöljük, ezeket oszlopvektoroknak tekintjük; ha sorvektorokként szeretnénk tekinteni, akkor az $\mathbf{x}^\top, \mathbf{y}^\top, \mathbf{z}^\top, \ldots$ jelölést használjuk. Az $\mathbf{x}$ vektor kooordinátái ebben a bázisban $x_1, \ldots, x_n$, azaz $\mathbf{x} = \sum_{i=1}^n x_i \varepsilon_i$. Az (12.1) megállapodás miatt $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$, az

$\mathbf{x}$ vektor euklideszi normája pedig $||\mathbf{x}|| = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{\sum_{i=1}^n x_i^2}$.

Az $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ lineáris transzformációt azonosítjuk azzal az $n \times n$-es $\mathbf{A} := (a_{ij})_{i,j=1}^n$ mátrixszal, melynek j-edik oszlopában az $\mathcal{A}\varepsilon_j$ vektor koordinátái állnak. Ha egy $\mathbf{x}$ vektor $\mathcal{A}$-val való transzformáltja $\mathbf{y}$, azt az $\mathcal{A}\mathbf{x} = \mathbf{y}$, vagy mátrixalakban az $\mathbf{A}\mathbf{x} = \mathbf{y}$ ($y_i = \sum_{j=1}^n a_{ij} x_j$) jelöléssel fejezzük ki. Az $\mathbf{A} := (a_{ij})$ és $\mathbf{B} := (b_{ij}) n \times n$-es mátrixok szorzata definí ció szerint $\mathbf{AB} := (c_{ik}) = (\sum_{j=1}^n a_{ij} b_{jk})$.

Az $\mathbf{I} := (\delta_{ij})_{i,j=1}^n$ mátrixot n -dimenziós egységmátrixnak (*identitás*nak) nevezzük. Az elnevezést az $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$ öszefüggés indokolja. Az $n \times n$-es $\mathbf{A}$ mátrix $\mathbf{A}^{-1}$ inverzét az $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ összefüggés definiálja (ez pontosan akkor létezik, ha az $|\mathbf{A}|$ mátrix alább definiált determinánsa nem 0). Közvetlen számolással meggyőződhetünk arról, hogy, ha az $\mathbf{A}$ és $\mathbf{B}$ mátrixok invertálhatók, akkor az $\mathbf{AB}$ mátrix is invertálható, és $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Az $\mathbf{A}$ mátrix $|\mathbf{A}|$*determináns*a a mátrix oszlopavektorai által definiált n-dimenziós parallelepipedon előjeles térfogata, ami az alábbi képlettel számítható ki:

$$|\mathbf{A}| = \sum_{\substack{\pi \in \text{az } (1,\ldots,n) \\ \text{permutációinak halmaza}}} (-1)^{\pi \text{ [inverzióinak száma]}} a_{1\pi(1)} \cdot \cdots \cdot a_{n\pi(n)}. \quad (1)$$

???Jelöljük A_{ij}-vel annak az $(n-1) \times (n-1)$-es mátrixnak a determinánsát, amelyet úgy kapunk $\mathbf{A}$-ból, hogy elhagyjuk az i-edik sorát és a j-edik oszlopát. Az $\text{adj}(\mathbf{A}) := ((-1)^{i+j} A_{ji})_{j,i=1}^n$ mátrixot $\mathbf{A}$*adjungált mátrix*ának nevezik, l. [30]. Az $\mathbf{A}^{-1}$ mátrix pontosan akkor létezik, ha $|\mathbf{A}| \neq 0$, és ekkor

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \text{adj}(\mathbf{A}).$$

Vegyük észre, hogy a determináns egy n^2 változós függvény (polinom), így van értelme a mátrixelemek szerinti deriválásnak. A (12.2)-beli definíciót felhasználva kapjuk, hogy

$$\frac{\partial |\mathbf{A}|}{\partial a_{ij}} = (-1)^{i+j} A_{ij}. \qquad (1)$$

???Egy $f(\mathbf{A}) (f : \mathbb{R}^{n^2} \to \mathbb{R})$ mátrixfüggvény mátrixelemek szerinti deriváltjaiból álló mátrixot szokás $\frac{\partial f}{\partial \mathbf{A}}$-val is jelölni, ezzel a jelöléssel (12.3) a

$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = \text{adj}\,(\mathbf{A}^q top)$$

tömör alakba írható át.

Ha az $\mathbf{A}^{-1}$ mátrix nem létezik, akkor azt mondjuk, hogy az $\mathbf{A}$ által definiált $\mathcal{A}$ transzformáció szinguláris.

A mátrix-jelölést alkalmazva $\mathrm{Im}(\mathbf{A})$ az $\mathbf{A}$ mátrix $\mathbf{ab}_1, \ldots, \mathbf{ab}_n$ oszlopvektorai által kifeszített $\mathrm{Span}(\mathbf{ab}_1, \ldots, \mathbf{ab}_n)$ altér (ezt onnan is látni, hogy $\mathbf{Ax} = \sum_{i=1}^{n} x_i \mathbf{ab}_i$), a $\mathrm{Ker}(\mathbf{A})$ altér pedig azon $\mathbf{x}$ vektorokból áll, amelyek ortogonálisak az $\mathbf{A}$ mátrix soraira, azaz az $\mathbf{A}^\top$ ($\mathbf{A}$ transzponáltja) oszlopaira, vagyis az $\mathrm{Im}(\mathbf{A}^\top)$ altérre. Ezzel igazoltuk a következőt.

12.1.1. Állítás. *A $\mathrm{Ker}(\mathbf{A})$ és $\mathrm{Im}(\mathbf{A}^\top)$ alterek egymás ortogonális komplementerei $\mathbb{R}^n$-ben, tehát $\dim(\mathrm{Ker}(\mathbf{A})) + \dim(\mathrm{Im}(\mathbf{A}^\top)) = n$.*

12.1.2. Definíció. *Az $\mathcal{U}$ transzformáció* ortogonális*, ha definiáló mátrixára igaz az $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ összefüggés.*

Ez azt jelenti, hogy $\mathbf{U}$ oszlopai ortonormáltak. Belátható, hogy ekkor $\mathbf{U}$ sorai is ortonormátak, ezért igaz az $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$ összefüggés is. Az ilyen $\mathbf{U}$ mátrixot *ortonormált mátrix*nak is szokták nevezni.

12.1.3. Definíció (szimmetrikus mátrix). *Az $\mathbf{A}$ $n \times n$-es valós mátrix szimmetrikus, ha $\mathbf{A}^\top = \mathbf{A}$, vagy, ami ugyanaz: $a_{ij} = a_{ji}$ minden $(i, j)(i = 1, \ldots, n;\ j = 1, \ldots, n)$ indexpárra.*

12.1.4. Definíció (projekció). *$\mathcal{P}$ transzformáció* ortogonális projekció*, ha $\mathbb{P}$ szimmetrikus és idempotens, azaz $\mathbb{P}\mathbb{P} = \mathbb{P}$.*

A $\mathcal{P}$ operátor az $\mathrm{Im}(\mathbb{P})$ altérre vetít. Mivel $\mathbb{P}$ szimmetrikus, 139. állítás miatt a $\mathrm{Ker}(\mathbb{P})$ és a $\mathrm{Im}(\mathbb{P})$ egymás ortogonális komplementerei, tehát minden $\mathbf{x} \in \mathbb{R}^n$ vektor előáll $\mathbf{x} = \mathbf{y} + \mathbf{z}$ alakban, ahol $\mathbf{y} \in \mathrm{Im}(\mathbb{P})$, $\mathbf{z} \in \mathrm{Ker}(\mathbb{P})$. Ezért $\mathbb{P}\mathbf{x} = \mathbf{y}$, innen az elnevezés. Ha $H \subset \mathbb{R}^n$ egy altér, $\mathbb{P}_H$ jelöli a H-ra való vetítést.

12.1.5. Állítás. *Ha $\mathbf{A}$ és $\mathbf{B}$ teszőleges $n \times n$-es mátrixok és $\mathbf{x} \in \mathbb{R}^n$ tetszőleges vektor, akkor $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^t op$ és*

$$(\mathbf{A}^\top\mathbf{x})^\top\mathbf{Bx} = \mathbf{x}^\top\mathbf{WBx}.$$

12.1.6. Definíció (kvadratikus alak, definitás). *Legyen $\mathbf{A}$ egy $n \times n$-es, szimmetrikus mátrix. Az*

$$\mathbf{x}^\top\mathbf{Ax} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j$$

számot az $\mathbf{A}$ által definiált kvadratikus alak*nak nevezzük. Az a_{ij} illetve x_i számok az $\mathbf{A}$ mátrix elemei illetve az $\mathbf{x}$ vektor koordinátái. Az $\mathbf{A}$ mátrixot* pozitív definit (szemidefinit)*nek nevezzük, ha az $\mathbf{x}^\top\mathbf{Ax}$ kvadratikus alak pozitív (nem-negatí v) minden, nem azonosan 0 komponensű $\mathbf{x}$ vektorra. Hasonlóan, az $\mathbf{A}$ mátrix* negatív definit (szemidefinit)*, ha az $\mathbf{x}^\top\mathbf{Ax}$ kvadratikus alak negatív (nem-pozitív) minden, nem azonosan 0 komponensű $\mathbf{x}$ vektorra. Ha pedig az $\mathbf{x}^\top\mathbf{Ax}$ kvadratikus alak mind pozitív, mind negatív értékeket felvehet (természetesen más-más $\mathbf{x}$ vektorokra), akkor az $\mathbf{A}$ mátrixot* indefinit*nek nevezzük. Szinguláris (nem invertálható) mátrixok a szemidefinitek és az indefinitek egy része.*

12.1.7. Definíció. *Legyenek $\mathbf{A}$ és $\mathbf{B}$ szimmetrikus mátrixok. Azt mondjuk, hogy $\mathbf{A} > \mathbf{B}$, ha $\mathbf{A} - \mathbf{B}$ szigorúan pozitív definit. Azt mondjuk, hogy $\mathbf{A} \geq \mathbf{B}$, ha $\mathbf{A} - \mathbf{B}$ pozitív szemidefinit.*

12.1.8. Tétel. *Az $\mathbf{A}$ mátrix akkor és csak akkor szimmetrikus, ha minden $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ vektorpárra*

$$\mathbf{x}^\top\mathbf{Ay} = \mathbf{y}^\top\mathbf{Ax}.$$

Megjegyezzük, hogy egy $\mathbf{B}$ mátrix pontosan akkor pozitív szemidefinit, ha ún. Gram-mátrix, azaz van olyan $\mathbf{A}$ mátrix, hogy $\mathbf{B} = \mathbf{A}^\top\mathbf{A}$.

Az alábbi tétel (l. [19] 149. o.) kovarianciamátrixok összehasonlításánál hasznos lehet.

12.1.9. Tétel. *Legyenek $\mathbf{A}$ és $\mathbf{B}$ invertálható szimmetrikus mátrixok. Ha $\mathbf{A} \leq \mathbf{B}$, akkor $\mathbf{B}^{-1} \leq \mathbf{A}^{-1}$*

12.1.10. Definíció (sajátérték, sajátvektor). *Az $\mathbf{u} \in \mathbb{R}^n$ nem azonosan 0 komponensűvektort az $n \times n$-es $\mathbf{A}$ mátrix sajátvektorának nevezzük, ha van olyan λ valós szám* (sajátérték)*, amellyel $\mathbf{Au} = \lambda\mathbf{u}$ teljesül.*

Ezzel ekvivalens a következő állítás: $\dim(\mathrm{Ker}(\mathbf{A}-\lambda\mathbf{I}))>0$, illetve $\dim(\mathrm{Im}(\mathbf{A}-\lambda\mathbf{I}))<n$, azaz az $\mathbf{A}-\lambda\mathbf{I}$ mátrix nem invertálható.

A sajátértékek geometriájáról a Gersgorin-tétel segítségével nyerhetünk hasznos információt.

12.1.11. Tétel (Gersgorin). *Legyen $\mathbf{A}$ egy tetszőleges (komplex elemű) $n\times n$-es mátrix. Legyen C_i az a_{ii} körüli $r_i:=\sum_{k=1k\neq i}^{n}|a_{ik}|$ sugarú nyílt körlemez a komplex számsíkon. Ekkor az $\mathbf{A}$ mátrix valamennyi sajátértéke a*

$$D:=\cup_{i=1}^{n}C_i$$

tartományban helyezkedik el.

12.1.12. Megjegyzés. *Az alábbi egyszerűészrevétel is rendkívül hasznos lehet a sajátértékek geometriájának vizsgálatánál.*

12.1.13. Tétel (spektrál-leképezés tétel). *Ha $P(\cdot)$ tetszőleges polinom, és λ az $\mathbf{A}$ mátrix sajátértéke, akkor $P(\lambda)$ a $P(\mathbf{A})$ mátrix sajátértéke.*

12.1.14. Tétel (spektrálfelbontási tétel). *Az $n\times n$-es szimmetrikus, valós elemű$\mathbf{A}$ mátrixnak van pontosan n valós sajátértéke (nagyság szerint csökkenő sorrendben jelölje őket $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n$), és az ezekhez tartozó $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_n$ sajátvektorok megválaszthatók úgy, hogy ortonormáltak legyenek (egy ilyen $\mathbf{u}_1,\ldots,\mathbf{u}_n$ rendszert ortonormált sajátvektor rendszernek nevezünk). Mátrixalakban ez az*

$$\mathbf{A}=\mathbf{U}\Lambda\mathbf{U}^T=\sum_{i=1}^{n}\lambda_i\mathbf{u}_i\mathbf{u}_i^T \qquad (1)$$

felbontást jelenti, ahol az $n\times n$ -es Λ diagonális mátrix a $\lambda_1,\ldots,\lambda_n$ sajátértékeket tartalmazza fődiagonálisában, az $\mathbf{U}$ ortogonális mátrix pedig a hozzájuk tartozó sajátvektorokat tartalmazza oszlopaiban, a sajátértékek sorrendjének megfelelően. Az (12.4) felbontást az $\mathbf{A}$ mátrix spektrálfelbontás*ának nevezzük.*

Szimmetrikus mátrixok sajátértékeinek becslésének hasznos eszköze a Weyl perturbációs tétel

12.1.15. Tétel.

$$\max_j|\lambda_j(\mathbf{A})-\lambda_j(\mathbb{B})|\leq||\mathbf{A}-\mathbb{B}||. \qquad (1)$$

Vegyük észre, hogy ha a $\mathbb{B}$ mátrix k-rangú, akkor (12.5) baloldala nem kisebb, mint $\lambda_{k+1}^*(\mathbf{A})$, viszont a $\widehat{\mathbb{B}}:=\sum_{i=1}^{k}\lambda_i^*\mathbf{u}_i^*\mathbf{u}_i^{*T}$ mátrixra teljesül

$$||\mathbf{A}-\widehat{\mathbb{B}}||=\lambda_{k+1}^*(\mathbf{A}).$$

Ezzel bebizonyítottuk, hogy a k-rangú szimmetrikus mátrixok körében $\mathbf{A}$ legjobb közelítése $\widehat{\mathbb{B}}$.

Ez az észrevétel képezi a főkomponensanalízis alapját.

A Weyl perturbációs tétel tetszőleges mátrixokra is általánosítható.

12.1.16. Tétel. *Legyen $\mathbf{A}$ tetszőleges $m\times n$-es valós elemű mátrix. Akkor*

$$\min_{\mathbb{B}\,k\text{-rangú}}||\mathbf{A}-\mathbb{B}||=s_{k+1},$$

és a minimum a $\widehat{\mathbb{B}}=\mathbf{V}\mathbb{S}_k\mathbf{U}$ mátrixon éretik el, ahol $\mathbb{S}_k$ az első k szinguláris értéket, valamint 0-kat tartalmazó (esetleg téglalap alakú) diagonális mátrix, $\mathbf{U}$ és $\mathbf{V}$ pedig az $\mathbf{A}$ mátrix szinguláris felbontásában szereplő ortogonális márixok.

12.1.17. Megjegyzés. *Az (12.4) formula azt jelenti, hogy az $\mathbf{A}$ mátrix egydimenziós alterekre való merőleges vetítések valós lineáris kombinációjaként áll elő.*

Tetszőleges valós $n \times n$-es $|\mathbf{A} - \lambda\mathbf{I}| = 0$em lehet ortogonális bázisban diagonalizálni, sőt egyáltalán nem lehet diagonalizálni, mert pl. a karakterisztikus egyenletnek komplex gyökei vannak, ilyen pl. a sík szöggel való elforgatását megadó

$$\begin{pmatrix} \sin\alpha & \cos\alpha \\ -\cos\alpha & \sin\alpha \end{pmatrix}$$

mátrix. Ilyenkor a mátrix komplex euklideszi térbeli ortogonális bázisban diagonalizálható, de ha a karakterisztikus egyenletnek többszörös (valós vagy komplex) gyöke van, akkor előfordulhat (nem szükségképpen!), hogy a mátrixnak még a komplex térben is n-nél kevesebb sajátvektora van, így ``ferde'' bázisban sem diagonalizálható, pl.

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Más módszert kell találni a mátrixok egyszerűbb alakban való felírására. Erre szolgál a poláris felbontás tétele, amely a komplex számok $z = re^{i\varphi}$ alakú felírásának messzemenő általánosítása.

12.1.18. Tétel (a poláris felbontás tétele). *Tetszőleges $\mathbf{A}$ négyzetes mátrix felírható $\mathbf{WB}$ alakban, ahol $\mathbf{B}$ pozitív szemidefinit (szimmetrikus), $\mathbf{W}$ pedig ortogonális. A $\mathbf{B}$ mátrix mindig egyértelműen meghatározott, míg $\mathbf{W}$ csak abban az esetben, ha $\mathbf{A}$ invertálható.*

A tétel közvetlen következménye a négyzetes mátrixokra vonatkozó

12.1.19. Tétel (szinguláris felbontási tétel). *Tetszőleges $\mathbf{A}$ négyzetes mátrixhoz van olyan $\mathbf{S} = \operatorname{diag}(s_1, \ldots, s_n)$ diagonális, valamint $\mathbf{U}$ és $\mathbf{V}$ unitér mátrix, hogy*

$$\mathbf{A} = \mathbf{VSU}^T = \sum_{i=1}^{n} s_i \mathbf{v}_i \mathbf{u}_i^T. \qquad (1)$$

* 1. A poláris (és a szinguláris) felbontásban szereplő $\mathbf{U}$ mátrix $\mathbf{u}_1, \ldots, \mathbf{u}_n$ oszlopvektorai rendelkeznek a következő tulajdonsággal:

$$(\mathbf{A}\mathbf{u}_i)^T(\mathbf{A}\mathbf{u}_j) = \delta_{ij}s_i^2$$

* 2. A $\mathbf{V}$ mátrix $\mathbf{v}_1, \ldots, \mathbf{v}_n$ oszlopvektoraira igaz az $s_i \cdot \mathbf{v}_i = A\mathbf{u}_i$ összefüggés.

* 3. Az $\mathbf{u}_1, \ldots, \mathbf{u}_n$ vektorrendszer az $\mathbf{A}^T\mathbf{A}$, míg a $\mathbf{v}_1, \ldots, \mathbf{v}_n$ vektorrendszer az $\mathbf{A}\mathbf{A}^T$ sajátvektorrendszere. (Az első állítás a konstrukció következménye, a második pedig az $\mathbf{A}\mathbf{A}^T = \mathbf{VSU}^T\mathbf{USV}^T = \mathbf{VS}^2\mathbf{V}^T$ egyenlőségsorozatból adódik.)

* 4. Egy szimmetrikus mátrix szinguláris értékei a sajátértékek abszolút értékei. Egyik oldali szinguláris vektoroknak megfelel a sajátvektorok bármely rendszere, legyen ez az $\mathbf{u}_i$ rendszer, a másik oldali szinguláris vektorok pedig a $\mathbf{v}_i = \pm\mathbf{u}_i$ vektorok lesznek, ahol az előjel a megfelelő λ_i sajátérték előjele.

* 5. $||\mathbf{A}|| = s_1$.

12.1.20. Tétel. *Legyen $\mathbf{A}$ tetszőleges $m \times n$-es valós eleműmátrix. Akkor*

$$\min_{\mathbf{B}\ k\text{-rangú}} ||\mathbf{A} - \mathbf{B}|| = s_{k+1},$$

és a minimum a $\widehat{\mathbf{B}} = \mathbf{VS}_k\mathbf{U}$ mátrixon éretik el, ahol $\mathbf{S}_k$ az első k szinguláris értéket, valamint 0-kat tartalmazó (esetleg téglalap alakú) diagonális mátrix, $\mathbf{U}$ és $\mathbf{V}$ pedig az $\mathbf{A}$ mátrix szinguláris felbontásában szereplő ortogonális márixok.

12.1.21. Definíció (mátrix nyoma). *A $\operatorname{tr}\mathbf{A} = \sum_{i=1}^{n} a_{ii}$ mennyiséget az $\mathbf{A}n \times n$-es mátrix* nyom*ának nevezzük.*

általában nem igaz, hogy az $1, \ldots, k$ számok tetszőleges $\pi(\cdot)$ permutációjára

$\mathrm{tr}\,(\mathbf{A}_1 \dots \mathbf{A}_k) = \mathrm{tr}\,(\mathbf{A}_{\pi(1)} \dots \mathbf{A}_{\pi(k)})$,

de ha $\pi(\cdot)$ ciklikus, akkor a tr$(\cdot)$ függvény `` kommutatív'':

$\mathrm{tr}\,(\mathbf{A}_1 \dots \mathbf{A}_k) = \mathrm{tr}\,(\mathbf{A}_2 \dots \mathbf{A}_k\mathbf{A}_1) = \mathrm{tr}\,(\mathbf{A}_3 \dots \mathbf{A}_k\mathbf{A}_1\mathbf{A}_2)$,

s.í.t.. Szükségünk lesz még a $p \times n$-es $\mathbf{A}$ és a $q \times m$-es $\mathbf{B}$ mátrixok *Kronecker- vagy tenzor-szorzatára*. Ez alatt azt a $pq \times nm$-es, $\mathbf{A} \otimes \mathbf{B}$-vel jelölt hipermátrixot értjük, melynek pn darab $q \times m$ méretűblokkja van: az (i, j) blokk az $a_{ij}\mathbf{B}$ mátrix $(i = 1, \dots p; j = 1, \dots, n)$. A Kronecker-szorzás asszociatív, a mátrixösszeadásra nézve disztributív, viszont általában nem kommutatív. Igaz azonban, hogy

$(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T$.

Amennyiben $\mathbf{A}$ és $\mathbf{B}$ négyzetes mátrixok - például $\mathbf{A}n \times n$-es, $\mathbf{B}$ pedig $m \times m$-es, akkor

$|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^m \cdot |\mathbf{B}|^n$,

továbbá, ha mindkettő invertálható, akkor Kronecker-szorzatuk is az, és

$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.

2. Függelék 2: Valószínűségelméleti képletgyűjtemény

2.1. Kolmogorov axiómái:

* (i) Adva van egy nem üres Ω halmaz (az eseménytér), Ω elemeit elemi eseményeknek nevezzük, és ω-val jelöljük.

* (ii) Ki van tüntetve az Ω részhalmazainak egy $\mathcal{A}$ algebrája ($\Omega \in \mathcal{A}$, $A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$, $A \in \mathcal{A} \& B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$).

* (iii) $\mathcal{A}\sigma$-algebra, azaz $A_k \in \mathcal{A}\,(k = 1, 2, \dots) \Rightarrow \cup_{k=1}^{\infty} A_k \in \mathcal{A}$.

* (iv) Minden $A \in \mathcal{A}$ eseményhez hozzá van rendelve egy $P(A)$ nemnegatív szám, az A esemény valószínűsége.

* (v) $P(\Omega) = 1$.

* (vi) Ha $A_k \in \mathcal{A}\,(k = 1, 2, \dots)$ páronként egymást kizáró események, akkor $P(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$.

2.2. Szitaformula:

$n = 3$ esetben:

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$$

Tetszőleges n-re:

$$P(A_1 \cup \dots \cup A_n) = \sum_{k=1}^{n} (-1)^k S_k^{(n)},$$

ahol

$$S_k^{(n)} := \sum_{1 \le i_1 < \dots < i_k \le n} P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}).$$

2.3. Események függetlensége, feltételes valószínűség

$1 \le j < k \le n$) **függetlenek**, ha $P(A_j \cap A_k) = P(A_j) \cdot P(A_k)$ páronként ($1 \le i_1 < \dots < i_k \le n$-ek, ha minden $P(A_{i_1} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot \dots \cdot P(A_{i_k})$ (ill. minden indexsorozatra). A teljes függetlenség implikálja a páronkénti függetlenséget.

Feltételes valószínűség: $P(A|B) := \frac{P(A \cap B)}{P(B)}$, ha $P(B) > 0$.

Teljes eseményrendszer:

$A_1, \dots, A_n \in \mathcal{A}, \quad P(A_i \cap A_j) = 0$ ha $i \neq j$ és $P(\cup_{i=1}^{n} A_k) = 1$.

Bayes tétele: Ha $A_1, \dots, A_n$ teljes eseményrendszer és $P(B) > 0$:

$$P(A_1|B) := \frac{P(B|A_1) \cdot P(A_1)}{\sum_{k=1}^{n} P(B|A_k) \cdot P(A_k)}.$$

2.4. Valószínűségi változó

Valószínűségi változó: Az Ω halmazon értelmezett olyan $\xi(\omega)$ valós értékűfüggvény, amelyre $\{\xi(\omega) < x\} \in \mathcal{A}$ minden valós x-re. Ha ξ értékkészlete a természetes számok halmaza, akkor diszkrét valószínűségi változóról beszélünk.

Függetlenség: A $\xi_1, \dots, \xi_n$ valószínűségi változók páronként (ill. teljesen) függetlenek, ha a $\{\xi_1(\omega) < x_1\}, \dots, \{\xi_n(\omega) < x_n\}$ események páronként (ill. teljesen) függetlenek $x_1, \dots, x_n$ minden értékére.

Eloszlás (általános eset):

A ξ valószínűségi változó $F(x)$ eloszlásfüggvénye:

$F_\xi(x) := P\{\xi < x\}$

$F_\xi(x)$ monoton nemcsökkenő balról folytonos függvény, $F_\xi(-\infty) = 0$, $F_\xi(\infty) = 1$.

Diszkrét eset:

A ξ valószínűségi változó $\{p_j\}$ eloszlása:

$p_j := P\{\xi = j\}$ $j = 0, 1, \dots$

Abszolút folytonos eset:

Ha $F_\xi(t) = \int_{-\infty}^{t} F'_\xi(x)dx$, akkor az

$f_\xi(x) := F'_\xi(x)$ függvény a ξ valószínűségi változó sűrűségfüggvénye.

Eloszlások konvolúciója:

A diszkrét eset: ha $\{p_i\}$ a ξ és $\{q_j\}$ az η független valószínűségi változók eloszlásai akkor a $\zeta = \xi + \eta$ valószínűségi változó eloszlása $\{r_k\}$:

$$r_k = \sum_{i=0}^{k} p_i \cdot q_{k-i} = \sum_{j=0}^{k} p_{k-j} \cdot q_j.$$

Az abszolút folytonos eset: ha ξ és η független valószínűségi változók, akkor

$$f_{\xi+\eta}(z) = \int_{-\infty}^{\infty} f_\xi(z-y) \cdot f_\eta(y)dy = \int_{-\infty}^{\infty} f_\xi(x) \cdot f_\eta(z-x)dx.$$

Valószínűségi változó függvényének eloszlása: (Csak az abszolút folytonos esetetet vizsgáljuk.) Legyen $\psi(x)$ invertálható függvény. Ha $f_\xi(x)$ a ξ valószínűségi változó sűrűségfüggvénye, akkor az $\psi(\xi)$ sűrűségfüggvénye:

$$f_\psi(y) = \begin{cases} \frac{f_\xi(\psi^{-1}(y))}{|\partial\psi(x_i)/\partial y_j|}, & \text{ha} \quad \inf\psi(x_i) < y_j < \sup\psi(x_i) \\ 0, & \text{különben} \end{cases}$$

2.5. Valószínűségi változó momentumai:

A diszkrét eset: ha $\{p_k\}$ a ξ valószínűségi változó eloszlása, az

$$M_{n,\xi} := \sum_{k=1}^{\infty} k^n \cdot p_k$$

összeget (amennyiben konvergens) a ξn-edik momentumának nevezzük, míg a

$$M_{n,\xi}^{(c)} := \sum_{k=1}^{\infty} (k - M_1)^n \cdot p_k$$

összeget a ξn-edik centrált momentumának nevezzük.

Az abszolút folytonos eset: ha $f(x)$ a ξ valószínűségi változó sűrűségfüggvénye, az

$$M_{n,\xi} := \int_{-\infty}^{\infty} x^n \cdot f(x)dx$$

integrált (amennyiben létezik) a ξn-edik momentumának nevezzük, míg a

$$M_{n,\xi}^{(c)} := \int_{-\infty}^{\infty} (x - M_1)^n \cdot f(x)dx$$

integrált a ξn-edik centrált momentumának nevezzük.

Ha ξ és η független valószínűségi változók, akkor

$$M_{n,\xi\cdot\eta} = M_{n,\xi} \cdot M_{n,\eta}.$$

Ha $k < n$ és $M_{n,\xi}$ létezik, akkor $M_{k,\xi}$ is létezik.

Várható érték, szórásnégyzet:

A ξ valószínűségi változó várható értéke: $E(\xi) := M_{1,\xi}$ szórásnégyzete: $D^2(\xi) := M_{2,\xi}^{(c)}$.

Legyen $\psi(x)$ egy tetszőleges valós értékűfüggvény.

$$E(\psi(\xi)) = \begin{cases} \sum_{k=0}^{\infty} \psi(k) \cdot p_k, & \text{ha} \quad \xi \ \text{diszkrét}, \\ \int_{-\infty}^{\infty} \psi(x) \cdot f(x)dx, & \text{ha} \quad \xi \ \text{abszolút folytonos}, \end{cases}$$

amennyiben a jobboldalon álló összeg (integrál) létezik.

Ha ξ és η tetszőleges valószínűségi változók, amelyeknek létezik a várható értékük, akkor $E(\xi + \eta) = E(\xi) + E(\eta)$.

Ha $\xi_1, \ldots, \xi_n$ páronként független valószínűségi változók, akkor $D^2(\xi_1 + \cdots + \xi_n) = D^2(\xi_1) + \cdots + D^2(\xi_n)$, ha a jobboldal létezik.

A Steiner-képlet:

$$D^2(\xi) := M_{2,\xi} - (E(\xi))^2$$

2.6. A generátorfüggvény:

A $\{p_j\}$ eloszlású ξ diszkrét valószínűségi változó $G_\xi(s)$ generátorfüggvénye:

$$G_\xi(s) := E(s^\xi) = \sum_{k=0}^{\infty} s^k \cdot p_k$$

$G_\xi(s)$ analitikus az egyeségkörben, $G_\xi(1) = 1$, $G'_\xi(1) = E(\xi)$.

Ha a $\xi_1, \ldots, \xi_n$ valószínűségi változók teljesen függetlenek, akkor

$$G_{\xi_1+\cdots+\xi_n}(s) = G_{\xi_1}(s) \cdot \cdots \cdot G_{\xi_n}(s).$$

Ha $\xi_1, \xi_2, \cdots$ azonos eloszlású teljesen független valószínűségi változók, és ν tőlük független diszkrét valószínűségi változó, akkor

$$G_{\xi_1+\cdots+\xi_\nu}(s) = G_\nu(G_\xi(s)).$$

A generátorfüggvény egyértelműen meghatározza az eloszlást:

$$p_n = \frac{1}{n!}\frac{d^n}{ds^n}G_\xi(s)|_{s=0},\ n = 1, 2, \ldots$$

A generátorfüggvény $s = 1$ pontbeli deriváltjai meghatározzák az ún. faktoriális momentumokat:

$$E[\xi(\xi-1)\ldots(\xi-k)] = \frac{d^k}{ds^k}G_\xi(s)/Vert_{s=1}$$

2.7. A karakterisztikus függvény:

ξ valószínűségi változó $\varphi_\xi(t)$ karakterisztikus függvénye:

$$\varphi_\xi(t) := E(e^{\mathbf{i}\cdot\xi\cdot t}) = \begin{cases} \sum_{k=0}^{\infty} e^{\mathbf{i}\cdot k\cdot t}\cdot p_k, & \text{ha} \quad \xi \text{ diszkrét}, \\ \int_{-\infty}^{\infty} e^{\mathbf{i}\cdot x\cdot t}\cdot f_\xi(x)dx, & \text{ha} \quad \xi \text{ abszolút folytonos}, \end{cases}$$

ahol $\mathbf{i} = \sqrt{-1}$.

Ha ξ diszkrét, akkor $\varphi_\xi(t) = G_\xi(e^{\mathbf{i}\cdot t})$.

A $\varphi_\xi(t)$ a t-nek a $(-\infty < t < \infty)$ intervallumon egyenletesen folytonos függvénye, $\varphi_\xi(0) = 1$, $|\varphi_\xi(t)| \le 1$ minden t-re, $\varphi_{a+b\xi}(t) = e^{\mathbf{i}\cdot a\cdot t}\varphi_\xi(b\cdot t)$.

$$M_{n,\xi} = (-\mathbf{i})^n \frac{d^n}{dt^n}\varphi_\xi(t)|_{t=0}.$$

Ha a $\xi_1, \ldots, \xi_n$ valószínűségi változók teljesen függetlenek, akkor

$$\varphi_{\xi_1+\cdots+\xi_n}(t) = \varphi_{\xi_1}(t) \cdot \cdots \cdot \varphi_{\xi_n}(t).$$

A karakterisztikus függvény egyértelműen meghatározza az eloszlást; abszolút folytonos eloszlás esetén, ha $|\varphi_{\xi_n}(t)|$ integrálható:

$$f_\xi(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\mathbf{i}\cdot x\cdot t}\varphi_\xi(t)dt.$$

A karakterisztikus függvény $t = 0$ pontbeli deriváltjai alapján kiszámíthatók a momentumok:

$$E(\xi^k) = \mathbf{i}^{-k}\frac{d^k}{dt^k}\varphi(t)$$

2.8. Nevezetes diszkrét eloszlások:

Bernoulli-eloszlás (egyszerű alternatíva):

$P\{\xi = 1\} = p,\ P\{\xi = 0\} = q,\ p + q = 1.$

$E(\xi) = p,\ D^2(\xi) = p \cdot q,\ G_\xi(s) = q + p \cdot s.$

Binomiális eloszlás (n független Bernoulli összege):

$P\{\xi = k\} = \binom{n}{k} p^k q^{n-k},\ p + q = 1,\ k = 0, 1, \ldots, n.$

$E(\xi) = n \cdot p,\ D^2(\xi) = n \cdot p \cdot q,\ G_\xi(s) = (q + p \cdot s)^n.$

Poisson-eloszlás (binomiális eloszlás limesze, ha $n \to \infty$ és $p \cdot n = \lambda$):

$P\{\xi = k\} = \frac{1}{k!}\lambda^k \cdot e^{-\lambda},\ \lambda > 0,\ k = 0, 1, \ldots$

$E(\xi) = \lambda,\ D^2(\xi) = \lambda,\ G_\xi(s) = e^{\lambda \cdot (s-1)}.$

Geometriai eloszlás (az egyszerűalternatíva független ismétléseinek száma az első 1-es megjelenéséig):

$P\{\xi = k\} = p \cdot q^{k-1},\ p + q = 1,\ k = 1, 2, \ldots$

$E(\xi) = \frac{1}{p},\ D^2(\xi) = \frac{q}{p^2},\ G_\xi(s) = \frac{p \cdot s}{1 - q \cdot s}.$

Negatív binomiális eloszlás (r darab geometriai összege):

$P\{\xi = r + k\} = \binom{k+r-1}{r-1} p^r q^k,\ p + q = 1,\ k = 0, 1, \ldots$

$E(\xi) = \frac{r}{p},\ D^2(\xi) = \frac{r \cdot q}{p^2},\ G_\xi(s) = \left(\frac{p \cdot s}{1 - q \cdot s}\right)^r.$

Hipergeometrikus eloszlás (visszatevés nélküli mintavétel):

$P\{\xi = k\} = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}} M < N,\ n \leq N,\ k = 0, 1, \ldots, n.$

$E(\xi) = n \cdot \frac{M}{N},\ D^2(\xi) = n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot (1 - \frac{n-1}{N-1}).$

2.9. Nevezetes abszolút folytonos eloszlások:

Normális (Gauss-) eloszlás:

$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}},\ -\infty < x < \infty,\ -\infty < m < \infty,\ 0 < \sigma < \infty.$

$E(\xi) = m,\ D^2(\xi) = \sigma^2,$

továbbá, ha $m = 0,\ k = 1, 2, \ldots E(\xi^{2k-1}) = 0$ és

$E(\xi^{2k}) = 1 \cdot 3 \cdot \cdots \cdot (2k-1)\sigma^{2k}.$

$\psi_\xi(t) = e^{\mathrm{i} \cdot m \cdot t - \frac{\sigma^2}{2} t^2}.$

Lognormális eloszlás (e^ξ eloszlása, ahol ξ Gauss):

$f_\xi(x) = \frac{1}{x \cdot \sqrt{2\pi}\sigma} e^{-\frac{(\ln x - m)^2}{2\sigma^2}},\ 0 < x < \infty,\ -\infty < m < \infty,\ 0 < \sigma < \infty.$

$E(\xi) = e^{m + \sigma^2/2},\ D^2(\xi) = e^{2m + \sigma^2} \cdot (e^{\sigma^2} - 1).$

Exponenciális eloszlás:

$f_\xi(x) = \lambda \cdot e^{-\lambda \cdot x},,\ 0 < x < \infty,\ 0 < \lambda < \infty.$

$E(\xi) = \frac{1}{\lambda}$, $D^2(\xi) = \frac{1}{\lambda^2}$ $\psi_\xi(t) = \frac{1}{1-\frac{\mathbf{i}t}{\lambda}}$.

Az exponeciális eloszlást karakterizálja az ún. örökifjú tulajdonság:

$$\mathrm{P}(\xi > x + y | \xi > x) = \mathrm{P}(\xi > y)$$

Gamma-eloszlás $(\mathcal{G}(\lambda, \alpha))$:

$f_\xi(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x \geq 0$

($\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$)

$E(\xi) = \frac{\alpha}{\lambda}$ $D^2(\xi) = \frac{\alpha}{\lambda^2}$ $\psi_\xi(t) = \left(1 - \mathbf{i}\frac{\mathbf{t}}{\lambda}\right)^{-\alpha}$.

χ^2 eloszlás n szabadságfokkal:

$f_\xi(x) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)}$, $x \geq 0$

$E(\xi) = n$ $D^2(\xi) = 2n$ $\psi_\xi(t) = \left(1 - \mathbf{i}\frac{\mathbf{t}}{\mathbf{2}}\right)^{-n/2}$.

t (Student-) eloszlás n szabadságfokkal: A ξ/η eloszlása, ahol ξ és η függetlenek, $\xi \sim \mathcal{N}(0,1)$ $\eta \sim \chi^2(n)$

$f_\xi(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$,

$E(\xi) = 0$ ha $n > 1$ $D^2(\xi) = \frac{n}{n-2}$ ha $n > 2$.

Béta-eloszlás a, b paraméterrel $(\mathcal{B}(a,b))$:

$f_\xi(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$ $x \in [0,1]$

$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

$E(\xi) = \frac{a}{a+b}$ $D^2(\xi) = \frac{ab}{(a+b)^2(a+b+1)}$

Másodfajú Béta-eloszlás a, b paraméterrel:

$f_\xi(x) = \frac{x^{a-1}(1+x)^{-a-b}}{B(a,b)}$ $x \in [0, \infty)$

$E(\xi) = \frac{a}{b-1}$ ha $b > 1$ $D^2(\xi) = \frac{a(a+b-1)}{(b-1)^2(b-2)}$ ha $b > 2$

Fisher-féle F-eloszlás n és m paraméterekkel $(\mathcal{F}(n,m))$, A ξ/η eloszlása, ahol ξ és η függetlenek:

$f_\xi(x) = \frac{n(\frac{n}{m}x)^{\frac{n}{2}-1}(1+\frac{n}{m}x)^{-\frac{n+m}{2}}}{mB(\frac{n}{2},\frac{m}{2})}$ Az $\eta = \frac{n}{m}\xi$ valószínűségi változó Másodfajú Béta-eloszlás $\frac{n}{2}$, $\frac{m}{2}$ paraméterrel!

Egyenletes eloszlás (az (a,b) intervallumon):

$f_\xi(x) = \frac{1}{b-a}$, ha $a < x < b$, 0 különben.

$E(\xi) = \frac{a+b}{2}$, $D^2(\xi) = \frac{1}{12}(b-a)^2$ ha $a = -b$: $\psi_\xi(t) = \frac{\sin bt}{b \cdot t}$.

2.10. Sztochasztikus konvergencia, majdnem biztos konvergencia:

A ξ_n valószínűségi változó sorozat sztochasztikusan konvergál a ξ valószínűségi változóhoz, ($\xi_n \overset{szt}{\to} \xi$) ha

bármely ε-hoz van olyan N, hogy minden $n > N$-re

$$P\{|\xi_n - \xi| > \varepsilon\} < \varepsilon.$$

A ξ_n valószínűségi változó sorozat majdnem biztosan (1 valószínűséggel) konvergál a ξ valószínűségi változóhoz, ($\xi_n \overset{mb}{\to} \xi$) ha

$$P\{\lim_{n\to\infty}\xi_n = \xi\} = 1.$$

A majdnem biztos konvergencia implikálja a sztochasztikus konvergenciát.

2.11. Nevezetes összefüggések

12.2.11.1. Tétel (Markov-egyenlőtlenség). *Ha a $E(\xi)$ létezik, akkor minden pozitiv a számra:*

$$P\{|\xi| \geq a\} \leq \frac{E(|\xi|)}{a}.$$

Csebisev-egyenlőtlenség:

Ha a $D^2(\xi)$ létezik, akkor minden pozitiv a számra:

$$P\{|\xi - E(\xi)| \geq a\} \leq \frac{D^2(\xi)}{a^2}.$$

12.2.11.2. Tétel (Nagy számok gyenge törvénye). *Ha $\xi_1, \xi_2, \cdots$ páronként független azonos eloszlású valószínűségi változók sorozata, és léteznek a $D^2(\xi_k)$ szórásnégyzetek, akkor*

$$\frac{1}{n}(\xi_1 + \cdots + \xi_n) \overset{szt}{\to} E(\xi).$$

12.2.11.3. Tétel (Nagy számok erős törvénye). *Legyen $\xi_1, \xi_2, \cdots$ teljesen független azonos eloszlású valószínűségi változók sorozata. Annak szükséges és elégséges feltétele, hogy az $\frac{1}{n}(\xi_1 + \cdots + \xi_n)$ sorozat majdnem biztosan konvergáljon egy m számhoz az, hogy létezzen az $E(\xi)$ várható érték. Ekkor $m = E(\xi)$.*

12.2.11.4. Tétel (Centrális határeloszlás tétel). *Ha $\xi, \xi_1, \xi_2, \cdots$ teljesen független azonos eloszlású valószínűségi változók sorozata, és létezik a $D^2(\xi)$ szórásnégyzet, akkor*

$$\lim_{n\to\infty} P\left\{\frac{\xi_1 + \cdots + \xi_n - n \cdot E(\xi)}{\sqrt{D^2(\xi) \cdot n}} < x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds.$$

Hivatkozások

[1] Achlioptas, D., McSherry, F., Fast Computation of Low Rank mátrix approximations J. ACM 54 2 (2007) Art. 9 (elektronikus) 19 o.

[2] Babu, Bootstrapping Statistics with Linear Combination of Chi-squares as a Weak Limit, The Indian Statist. J. 46 (1984) 85-93.

[3] Borovkov, A. A., Matematikai statisztika, Typotex, Bp., 1999

[4] Bevezetés a matematikai statisztikába, KLTE jegyzet, Szerk. Fazekas István, Kossuth Egyetemi Kiadó, 2005

[5] Bolla Marianna, Krámli András, Statisztikai következtetések elmélete (Második, javított kiadás), Typotex, 2012

[6] Breiman, L., Friedman, J. H., Estimating Optimal Transformation for multiple Regression and Correlation, J. Amer. Stat. Assoc. 80 391 (1985) 580- 598.

[6] Breiman, L., Friedman, J. H., Estimating Optimal Transformation for multiple Regression and Correlation, J. Amer. Stat. Assoc. 80 391 (1985) 580- 598.

[8] Csencov, N. N., Statisztikai Döntési Szabályok és Optimális Következtetések (oroszul), NAUKA, Moszkva, 1972

[9] Csiszár Imre, Eloszlások eltérésének információ típusú mértékszámai. MTA III. Oszt. Közleményei 17, 123- 149, 1967

[10] Efron, B., Bootstrap methods: another look at the jackknife Ann. Statist. 7 (1979), 1-45

[11] Fisher, R. A. Theorz of statistical estimations, Proc. Cambridge Phylosoph. Soc. 22 (1925), 700.

[12] Flury, A first course in multivariate statistics, Sringer, 1997

[13] Frieze, A., Kannan, R., Vempala, S., Fast Monte Carlo Algorithms for Finding Low-Rank Approximation, J. ACM 51 6 (2004) 1025- 1041.

[14] Giri, Multivariate statistical analysis, Marcel Dekker, 2004

[15] Gnyegyenko, B. V., Kolmogorov, A. N., Független valószínűségi változók öszegeinek határeloszlásai, Akadémiai Kiadó, Budapest, 1951

[16] Grone, R., Pierce, S., Watkins W., Extremal correlation matrices, Lin. Alg. Appl. 134 (1990), 63- 70.

[17] Hofmann, T., Schölkopf, B., Smola, J., Kernel methods in machine learning, Ann. Statist. 36 3 (2008) 1171- 1220.

[18] Kruskal, J. B., On the shortest spanning subtree of a grapf and the travelling salesman problem. Problem. Amer. Math. Soc. 7 (1956), 48- 50

[19]

[20] Lovász, L., Kombinatiorikai problémák és feladatok Typotex, Bp., 1999

[21] Lukacs, E., The stochastic independence of symmetric and homogeneous linear and quadratic statistics, Ann. Math. Statist. 23 (1952), 442- 449.

[22] Mika, S., Schölkopf, B., Smola, A. J. Müller, K. R., Kernel PCA and de-noising in feature spaces, Advances in neural information processing systems 11 (1), 536-542

[23] Miller, Rupert, G., Jr., A trustworthy jackknife, Ann. Math. Statist. 35 (1964), 1594-1605

[24] Miller, Rupert, G., Jr.,Jackknifing variances, Ann. Math. Statist. 39 (1968), 567-582

[25] Móri, Szeidl, Zempléni: Matematikai statisztika példatár, ELTE Eötvös Kiadó, 1997

[26] Móri Tamás, Székely J. Gábor (szerk.), Többváltozós Statisztikai Analizis, Műszaki Könyvkiadó, Budapest, 1972

[27] Olkin, I., Pierce, S. The 70th anniversary of random matrices, Lin. Alg. Appl. 354 (2002), 231-243.

[28] Quenouille, M., H., Notes on bias in estimation, Biometrika, 43 (1956) 353-360

[29] R., ed. Handbook of Statistics, V. 9. 627-659 Elsevier Science Pulisher, 1993

[30] Rózsa, P., Lineáris algebra és alkalmazásai, Műszaki Könyvkiadó, Bp., 1974

[31] Singh, K., On the asymptotoic accuracy of Efron's bootstrap, Ann. Statist. 9 (1981) 1187- 1195.

[32] Tukey, J., W., Abstract, Ann. Math. Statist. 29 (1958), 612