Statisztika Jegyzet
az üzelti informatika szakirány számára

(kézirat gyanánt)

Telcs András

December 16, 2005

CONTENTS

0.1 Elõszó

E jegyzet informatikus hallgatók számára íródott. Ezen belül az Üzleti informatika szakirány keretében kerül felhasználásra. Szerkesztésekor e két szempont alapján arra törekedtünk, hogy az informatikus képzés során szerzett ismeretekre, informatikus és mérnöki szemléletre alapozzunk. Ez egyben azt is jelenti, hogy lehetõség szerint a gyakorlati igénnyel fellépõ olvasó számára íródott. Ezt erõsíti a szakirány szabta feladat. A keretek szabta korlátok között az üzleti életben felmerülõ problémákon keresztül kerülnek elõ egyes statisztikai kérdések. Bizonyos mértékig az üzleti életben szokásos zsargont is becsempésztük a szövegbe, hogy a szerzett ismeretek késõbi alkalmazását ezzel is könnyítsük.

A kurzus több szinten sajátítható el. Mindenki maga dönti el mit céloz meg.

A jegyzet anyaga lehetõvé teszis, hogy minimális üzleti, statisztikai szókincsre tegyen szert az olvasó. Képes legyen egy üzleti, gazdasági poblémában felismerni a statisztikai feladatot és azonosítani a feladat megoldásához szükséges módszert. Végül megértse a mások által megoldott statisztikai elemzés fõbb üzenetét, interpretációját.

Az aki ennél többre törekszik, az alaposabb elsajátítás révén képessé válhat a statisztikusokkal együttmûködni a feladat korrekt specifikálásában, mediálni a vállalat és a statisztikusok között biztosítva, hogy a válasz valóban arra a kérdésre adatik amit az üzleti élet felvetett.

A kitúzhetõ maximális cél, hogy a tárgyhoz kapcsolódó laborra is támaszkodva képessé válik a feladat meghatározásától a teljes megvalósításig minden lépést megoldani. Azaz végighaladnia következõ lépéseken:

1. probléma specifikálás
2. munkafeltevések megfogalmazása
3. módszer kiválasztás
4. adatigény felmérése
5. adatgyûjtés megtervezése, kivitelezése, adatfeldolgozás
6. adatjavítás, szûrés, tesztelés
7. elõzetes statisztikai feldolgozás
8. a statisztikai módszer alkalmazása
9. egybevetés a munkafeltevésekkel, egyéb szempontokkal

10. esetleg a 2-8 ciklus részleges vagy teljes ismétlése

11. következtetések levonása, interpretálás, az eredmények visszafordítása az üzleti probléma nyelvére

Aki sikerrel megbirkózott a kurzus ismereteinek ilyen magas szintû elsajátításával hasznos és erõs fegyvertár birtokába jutott. Természetesen nem rendelkezik a profi statisztikus teljes arzenáljával, de egye-egy irányban már képes lehet önállóan is fejleszteni ezirányú ismereteit illetve, ha kedvet érez haladó statisztikai stúdiumokra is vállalkozhat mint például a napjainkban oly fontos nemparaméteres eljárások vagy adatbányászati módszerek.

Annak érdekében, hogy a kitûzött feladatoknak megfeleljen, jegyzet felépítése a következõ. Az elsõ fejezetben a valószínûségszámítás alapogalmai kerülnek röviden összefoglalásra majd a másodikban statisztika alapfogalmai kerülnek ismertetésre. A következõ fejezetek egyre öszetettebb módszereket ismertetnek. Minden fejezet végén külön papíron illetve számítógép segítségével megoldható feladatok találhatóak. Ezek megoldása biztosítja az anyag elsajátításának második illetve harmadik szintjét.

Köszönetnyilvánítás

A szerző köszönetét szeretné kifejezni Maricza Istvánnak, azért, hogy jegyzetének idősor fejezete átemeléséhez hozzájárult. Ugyanőt illeti még a közönet a baráti és szakmai beszélgetésekért, amellyel a szerzőt messzemenően segítette.

0.2 Bevezetés

A statisztika a rendszerezett számbavétel igényébõl fejlõdõtt ki az évszázadok során. Az elsõ számbavételi feladatok az idõszámításunk elõtti 4000 évre nyúlnak vissza. Kínában már ekkor összeírták a lakosságot, házakat, birtokokat. Ezen adatok a hatalom két nagyon fontos célját szolgálták az adókivetést és a katonai szolgálatot. Hasonló számbavételi igénye volt az egyiptomi uralkodóknak is a termény és a munkáerõ felmérése kapcsán. Mózes is számbaveszi nemzettségét (Lásd Mózes IV. könyve) és nem kevesebb mint 603.550 felnõtt férfit tud magáénak. A gyermekek a magas halandóság miatt, illetve mert még sem munkára sem hadra nem foghatóak, nem számítottak, ahogy a nõk sem. Másutt több nemzettségfõ részletes vagyonfelsorolását találhatjuk a Bibliában mennyi nény illetve kétlábú jószágga rendelkeztek.

Az elsõ hivatalos összeírásról is a Biblia számol be (Lukács 2.) ”Augusztus császár rendeletet adott ki, hogy az egész földkerekséget összeírják össze. Ezt az elsõ összeírást Cirinus, Szíria helytartója, bonyoltította le. Mindenki elment a maga városába, hogy összeírják.” Adat Rómáról maradt fennt, fénykorában mintegy 1 millió lakosa volt. Ezt megelõzõ idõbõl származó adatok szerint Athénban és Khorinthoszban 400 illetve 460 ezer rabszolga volt, amely feltehetõen nagyobb a valóságnál.

A következõ nagyszabású vállalkozásra ezer évet kell várni. A XI. századba születik az un. Doomsday Book, ami az akkori anglia fölbirtok és hûbéri viszonyainka felmérését szolgálta, megint csak adó és katonai szolgálat céljából . E hatalmas mû egyben számos hely és kortörténeti leírással volt kiegészítve.

Az elsõ modernnek tekinthetõ statisztika John Graunt és William Petty nevéhez fûzõdik 1650-bõl. Õk készítették az elsõ születési és halálozási statisztikát.

S leíró statisztika, ezen belül a grafikus ábrázlás úttörõje volt Florence Nightingale (1820-1910) aki adatsorokra támaszkodó ábrákkal gyözte meg Nagybritannia katonai vezetését, hogy a Krími háború sebesültjei közül többen pusztultak el a hadikórhzak rossz rökülményeinek következtében mint a sérülésekben a csatatéren. Õ az elsõ statisztikus nõ, egyben az ápolási szakma megteremtõje is.

E régi idõkre visszanyúló példák is mutatják a statisztika gyakorlati jelentõségét. Napjaink üzleti döntései pedig végképp elképzelhetetlenek azok nélkül. Hogy csak egy példát említsünk, az egyes termékek szenozális fogyasztási szokásai ismeretében gyártanak a termelõk, rendelnek és készleteznek a kereskedõk, például sört. A késõbbiekben számos ilyen jellegû példát founk még ismertetni, konkrét módszerek kapcsán.

Chapter 1
A LEíRÓ STATISZTIKA ELEMEI

Ebben a fejezetben a statszitika legrégibb ágával a leíró statisztikával ismerkedünk meg.

A leíró statisztika a vizsgálandó kérdés kapcsán szóba jöhetõ összes objektum megfigyelésén alapul. Vegyünk egy példát. Csoportkirándulást szervezzünk. Az étkezések megrendeléséhez néhány kérdést kell tisztázni. Van-e a résztvevõk között vegetáriánus, cukorbeteg, tej- vagy lisztérzékeny. Ki milyen italt fogyaszt reggelire, kávét, teát, tejet, kakaót. Természesen ezeket az információkat egyszerûen összegyûjthetjük a (tegyük fel) 36 résztvevõtõl. Ennek eredményeképpen 36 válaszlap lesz a kezünkben. Természetesen az étkezést nem személyenként fogjuk megrendelni. A megrendelésen az kell, hogy szerepeljen, hogy a fenti kérdésekre hány igen válasz érkezett. Elkezdünk tehát strigulázni. Az eredmény már a leíró statisztika egyik alapeleme.

Definíció 1 *Alapsokaság a vizsgált objektumok, egyedek összesége.*

Esetünkben a kiránduló csoport tagjai.

Definíció 2 *Oszály vagy kategória a megfigyelt objektumok valamely ismérv szerinti felosztása.*

Esetünkben például a reggeli ital mint az objektumok egy attribútuma négy lehetséges dolog lehet: kávé, tea, tej, kakaó. Ennek következtében a megfigyelt egyedek halmazának egy partícióját kapjuk, aszerint, hogy ki mit választ (kizárva a többszörös választást és a nem választ).

Definíció 3 *A partíció osztályainak elemszámát abszolút gyakoriságnak nevezzük.*

Fenti kérdéseink kissé egyoldalúak voltak, a kapott válaszok mind kategoriális ismérveket tartalaztak. Természetesen numerikus adatokra is szükségünk lehet, mint például, egy vagy két kávét iszik reggel. Általában a megfigyelt objektumok attribítumai, hasonlóan egy adatbázi egy rekordjáának mezõihez különbözõ jellegûek lehetnek. Az attribítumok két nagy osztályát különböztetjük meg.

Definíció 4 *Beszélhetünk **kvantitatív** és **kvalitatív** ismérvekrõl. Ezen belül különbözõ skálákról szokás beszélni. Kvalitatív skála lehet nominális, ha az attribítum csak egy név, cimke az objektumok valamely osztályba sorolását jelöli. Ilyen például a hajszín vagy a személyi azonosító szám. Lehet a skála ordinális, avagy rendezett amikor az osztályok között valamilyen értelem szerû sorrend van, de továbbra is csak osztálycimkék, nevek az adataink. Ilyen például a fizetési kategóriák, A7, B4, J11 vagy a labdarúgó csapatok liga besorolása.*

Vigyázat a nominális cimke is lehet szám, mint a személyi azonosító szám, mégsem érdemes például átlagot számítani belõle. Ennél is óvatosabban kezelendõ, hogy gyakran a számítógépes feldolgozás céljára az osztályok neve helyett számokat használnak, pédául a hajszín kódok fekete=1, barna=2 és így tovább, de e számok szintén csak cimkék, mûveleteket végezni nincs értelme velük. Ez ugyanakkor nem jelenti azt, hogy az elõfordulások számával ne lehetne mûveletekret végezni és a sagítségükkel statisztikai következtetésekre jutni.

Definíció 5 *A kvantitatív adatok a következõ csoportokba sorolhatóak.* ***Intervallum skáláról*** *lehet beszélni, ha az ismérv számszerû és elemei egy intervallumból kerülnek ki. Ilyen például az õszi félév oktatási napjainka dátuma.*
Hányados skálával *van dolgunk, ha a megfigyelt objektumok kérdéses numerikus ismérveinek hányadosa ismert. Például nem tudjuk, ki mennyi cukrot tesz az italába, de azt igen, hogy X másfélszer annyit mint Y. Y pedig 0.9-szer annyit mint Z.*

A leíró statisztika célja, hogy a rendelkezésre álló adatok alapján általános képet kapjunk a vizsgált objektumok összeségérõl és egyben az adatok minõségérõl. Ezt gyakran grafikus ábrázolással lehet elérni. Kiindulásul a megfigyelt gyakoriságok szolgálnak.

Egy ideig figyelmünket olyan vizsgálatokra korlátozzuk, amelyekbne az objektumok egyetlen paraméterével foglalkozunk. Természeseten késõbb több paraméter vizsgálatára sor kerül, hiszen gyakran igazán azok az izgalmas kérdések, hogy az egyik jól megfigyelhetõ paraméter viselkedésébõl következtessünk a másik esetleg kevésbé megfigyelhetõre.

Definíció 6 ***Gyakorisági tábla. Legyen*** *a megfigyelt összes objektum egy adott ismérv szerint osztályokba van sorolva. Az egyes osztályokba esõ elemek száma a gyakoriság, az osztálycimkék és e gyakoriságok alkotják a gyakorisági táblát. Például egy csoportban a hajszín gyakorisági táblája*

fekete	4
barna	12
vörös	1
szõke	3
összesen	20

Jelölje f_i az i-edik osztály gyakoriságát, frekvenciáját.

Definíció 7 *A* ***relatív gyakoriság***

$$r_i = \frac{f_i}{N}$$

ahol N az összes megfigyelt elemek száma, $i = 1, 2...K$ pedig az osztályok sorszáma, r_i az osztályok részirányát fejezi ki 0 és 1 között.

Definíció 8 *A gyakorisági diagramm avagy hisztogramm a gyakorisági táblát ábrázolja grafikusan.*

```
1  ****
2  **********
3  *
4  **
```

Abban az esetben is lehet gyakoriságról, relatív gyakoriságról beszélni, illetve ábrát készíteni. Ilyenkor jól megválasztott (általában egyenlõ) szélességû intervallumok alkotják az osztályokat és a gyakoriság az abba esõ elemek száma. Az intervallum szélességének megválasztásakor úgy célszerû eljárni, hogy a kapott gyakoriságok jól tükrözzék a vizsgált kérdést, egy-egy intervallumba az összes elemek közül se túl sok se túl kevés ne essen.

Szokás még a relatív gyakoriságokat kördiagrammon is ábrázolni. **[ide abra]**

A gyakoriságok elemzését néhány egyszerû statisztika segíti.

Definíció 9 ***Módusz*** *az az osztály vagy osztályok, amelyek a maximális elemszámot tartalmazzák.*

Fenti példánkban a barna a modális osztály.

A továbbiakban numerikus adatok, kvalitatív ismérvekre szorítkozunk.

Definíció 10 ***Medián*** *a vizsgált mennyiség skáláján egy olyan érték, amely két egyenlõ számú részre vágja a megfigyelt elemeket halmazát. Páratlan elemszám esetén ez a sorba rendezett értékek közul a középsõ, páros számú elem esetén a középsõ kettõ átlaga.*

Vegyünk egy példát. Az alábbi adatok

7.57, 9.55, 8.82, 8.72, 6.96, 6.83, 11.42, 16.08, 6.83, 13.05, 11.36, 2.72, 8.55, 10.79, 9.97, 9.85, 7.86, 7.72, 12.90, 18.17, 7.72, 14.75, 12.84, 3.08, 9.67, 12.19, 11.26, 11.13, 8.89, 2.41

30 háztartás papírhulladék "termelésének" értékei. A sorbarendezés után a középsõ két elem, azaz a 15 és 16. átlaga 9.61. Azaz ez a medián.

Definíció 11 *Hasonlóan definiáljuk a percentilist. a p-percentilis az a skálaérték, amely alatt a p százaléka található az elemeknek.*

Definíció 12 *Nevezetes percentilis az alsó és felsõ kvartilis, Q_1illteve Q_3, amelyek a 25 illetve 75 százalékos percentilisnek felelnek meg.*

Példánkban a 30% percentilis 8.21, $Q_1 = 7.72, Q_3 = 11.81$.

Definíció 13 *A fenti statisztikákat szokás* ***helyzeti paraméter****eknek is nevezni, mert az adatok elhelyezkedésérõl adnak felvilágosítást. Talán a legfontosabb helyzeti paraméter az* ***átlag****. azaz a mért $x_i, i = 1, 2...N$ értékek számtani közepe.*

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Az adatok szóródásának jellemzésére is több statisztikát lehet használni.

Definíció 14 *A* ***terjedelem*** *nem más mint a maximális és minimális mért érték különbsége.*

$$range = \max_{1 \le i \le N} x_i - \min_{1 \le i \le N} x_i.$$

Definíció 15 *Az **interkvartilis terjedelem***

$$IQR = Q_3 - Q_1.$$

Definíció 16 *A leghasznosabb szóródási mérõszám a **variancia***

$$VAR = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

*illetve az eredeti skálára visszatérve a **szórás***

$$\overline{\sigma} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Gyakorlat 1 *Számoljuk ki példánkban az átlagot és a szórást.*

Definíció 17 *A **korrigált tapasztalati szórás** kis elemszámok esetén játszik szerepet, jelentõségérõl még késõbb lesz szó.*

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Szokás még az adatok lapultsgára illetve jobb vagy baloldalra hízására vonatkozó mérõszámokta is bevezetni. Ezekre itt nem térünk ki.

Chapter 2
VALÓSZíNÛSÉGSZÁMíTÁSI ALAPFOGALMAK

2.1 A valószínûségi mezõ

Korábbi tanulmányaink során megismerkedtünk a Kolmogorov féle valószínüségi mezõ és azon értelmezett valószínûségi változó fogalmával. Az alábbiakban röviden áttekintjük az ide vonatkozó és a statisztikában nélkülözhetetlen fogalmakat és összefüggéseket. Bõvebb bevezetõért, példákért, gyakorlatokért lásd [**?**].

Definíció 18 *Legyen Ω tetszõleges halmaz. Ez az összes események halmaza. Szoktuk ezt **eseménytérnek** is nevezni. Abban az esetben, ha Ω véges vagy megszámláhatóan végtelen, akkor **atomos eseménytér**rõl beszélünk. Az $\omega \in \Omega$ eseményeket szokás elemi eseményeknek nevezni. Ezek már nem bonthatóak további eseményekre.*

Véges eseménytér esetén könnyû elképzelni az eseményeket. Ilyen például, hogy a tanulócsoportból véletlenül kiválasztott diák hajszíne fekete.

Definíció 19 *Összetett esemény minden nem egyelemû $A \subset \Omega$ részhalmaza $\Omega-$nak.*

Definíció 20 *Bevezetünk mûveleteket az események illetve az azokkal azonosított halmazok között.*

$$A + B = \{\omega : \omega \in A \text{ vagy } \omega \in B\}.$$
$$AB = \{\omega : \omega \in A \text{ és } \omega \in B\}. \tag{2.1}$$

E definíciók az eseményalgebrában szokásos jelöléseket használják, természetesen ugyanakkor egybeesnek a halmazok közötti mûveletekkel. $A + B = A \cup B,\ AB = A \cap B$. Ha adott $A_1, ...A_n..$ megszámlálhaóan végtelen esemény ezek összegét jelölje

$$\sum_{i=1}^{\infty} A_i = A_1 + A_2 + ...A_n + ... \tag{2.2}$$

Definíció 21 *Az $A \subset \Omega$ esemény komplementere*

$$\overline{A} = \{\omega \in \Omega : \omega \notin A\}. \tag{2.3}$$

Definíció 22 *A lehetetlen esemény az üres halmaz, jele $\emptyset$.*

Definíció 23 *$\mathcal{F}$ az Ω részhalmazainak egy családja szigmaalgebra, ha nem vezet ki belõle a (2.2)összeadás, (2.1) szorzás és (2.3) komplementer képzés.*

Definíció 24 *Az $(\Omega, \mathcal{F}, \mathbb{P})$ hármast valószínüségi mezõ, ha Ω valamely alaphalmaz, $\mathcal{F}$ e fölött egy szigmaalgebra, $\mathbb{P}$ pedig valószínüségi mérték, ami azt jelenti, hogy az alábbi axiómákat teljesíti. Legyen $A, B \subset \Omega$.*

1. $P(A) \geq 0$.

2. $P(\Omega) = 1$

3. ha $AB = \emptyset$ *akkor*

$$P(A + B) = P(A) + P(B)$$

Definíció 25 *Feltételes valószínüséget definiálhatunk tetszõleges $B \subset \Omega, P(B) \neq 0$ eseményre*

$$P(A|B) = \frac{P(A|B)}{P(B)}$$

segítségével.

Gyakorlat 2 *Lássuk be, hogy a feltételes valószínûség is kielégíti az 1.-3. axiomákat.*

Definíció 26 *A $\mathbb{B}$ valós halmazok családja Borel szigmaalgebrát alkot, ha tartalmazza az összes $[x, y)$ intervallumot és szigmaalgebra. A $B \in \mathbb{B}$ halmazokat Borel hamlazoknak nevezzük.*

Definíció 27 *Az $(\Omega, \mathcal{F}, \mathbb{P})$-n egy X leképezés, $X : \Omega \to \mathbb{R}$, valószínüségi változó, ha minden B Borel hamlamzra annak õsképe, azaz*

$$X^{-1}B = \{\omega : X\omega \in B\} \in \mathcal{F}$$

azaz $\mathcal{F}$ beli. Az ilyen halmazokat szoktuk még $\mathcal{F}$-mérhetõnek is nevezni.

Ha Ω véges vagy megszámlálhatóan végtelen, azaz

$$\Omega = \{\omega_1, ...\omega_n...\}$$

akkor diszkrét valószínüségi mezõrõl beszélünk, egyébként folytonosról. A várható érték és szórás fogalmát elõször diszkrét valószínüségi mezõn definiáljuk.

A diszkrét valószínüségi változó valószínûség eloszlását meghatározzák a

$$p_i = \mathbb{P}(\omega_i)$$

valószínüségek.

Definíció 28 *Legyen a diszkrét $(\Omega, \mathcal{F}, \mathbb{P})$-n egy X valószínüségi változó akkor ennek várható értéke, amennyiben az alábbi összeg létezik*

$$E(X) = \sum_{i=1}^{\infty} X(\omega_i)\,\mathbb{P}(\omega_i) =: \sum_{i=1}^{\infty} x_i p_i.$$

Jelölje mint általában ezt röviden $\mu = E(X)$.

Definíció 29 *Legyen a diszkrét $(\Omega, \mathcal{F}, \mathbb{P})$-n egy X valószínüségi változó akkor ennek szórásnégyzete, avagy varianciája*

$$V(X) = \sum_{i=1}^{\infty} (x_i - \mu)^2 p_i$$

amennyiben az összeg létezik. Szórása ez esetben

$$\sigma(X) = \sqrt{\sum_{i=1}^{\infty} (x_i - \mu)^2 p_i}.$$

Folytonos valószínûségi mezõ esetén az általánosság enyhe szûkítésével definiáljuk a várható értéket és szórást.

Definíció 30 *Legyen az X valószínûségi változó az $(\Omega, \mathcal{F}, \mathbb{P})$-n. X eloszlásfüggvénye $F(x)$, ha minden $x \in \operatorname{Im} X = \{y \in \mathbb{R} : y = X(\omega)\}$-ra*

$$0 \leq F(x) = \mathbb{P}(X < x) \leq 1$$

valószínüség sûrûségfüggvénye pedig $f(x) \geq 0$, ha $F(x)$ abszolút folytonos és minden $x - re$, ahol F értelmezett.

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$

Definíció 31 *Adott $(\Omega, \mathcal{F}, \mathbb{P})$-n az $A, B \subset \Omega$, eseményekre a feltétele valószínüséget a*

$$P(A|B) = \frac{P(AB)}{P(B)}$$

összefüggés definiálja, ahol feltesszük, hogy $P(B) > 0$.

Definíció 32 *Adott $(\Omega, \mathcal{F}, \mathbb{P})$-n az $A, B \subset \Omega$, események függetlenek, ha*

$$P(A|B) = P(A).$$

Ez ekvivalens azzal, hogy

$$P(AB) = P(A) P(B).$$

Definíció 33 *Adott $(\Omega, \mathcal{F}, \mathbb{P})$-n, az $X_i, i = 1, 2...$ valószínûségi változók páronként függetlenek, ha*

$$\mathbb{P}(\{X_i < x\}\{X_j < y\}) = \mathbb{P}(X_i < x)\, \mathbb{P}(X_j < y)$$

teljesen függetlenek, ha minden $k - ra$ és minden $i_1..i_k$ index k-asra

$$\begin{aligned} & \mathbb{P}(\{X_{i_1} < x_{i_1}\}\{X_{i_2} < x_{i_2}\}..\{X_{i_k} < x_{i_k}\}) \\ = \; & \mathbb{P}(X_{i_1} < x_{i_1})\, \mathbb{P}(X_{i_2} < x_{i_2}) ... \mathbb{P}(X_{i_k} < x_{i_k}). \end{aligned}$$

Definíció 34 *Adott* $(\Omega, \mathcal{F}, \mathbb{P})$*-n,* $A \subset \Omega, \mathbb{P}(A) > 0$ *eseményre értelmeztük a feltételes valószínüség fogalmát. Ez egy újabb valószínûségi mezõt is definiál* $(\Omega, \mathcal{F}, \mathbb{P}(.|A))$*-t, amin az* A*-ra vonatkozó feltételes várható érték valamely* X*-re a fenti definicióból adódik, jele* $E(X|A)$*, diszkrét esetben*

$$E(X|A) = \sum_{i=0}^{\infty} x_i P(X = x_i|A) \mathbb{I}(A)$$

folytonos esetben, tegyük fel, hogy létezik a $(\Omega, \mathcal{F}, \mathbb{P}(.|A))$*-n az* X *feltételes sûrûség függvénye* $f(x|A)$ *ekkor*

$$E(X|A) = \int x f(x|A)\, dx.$$

A feltételes várható érték fogalmát kissé szûkített értelmezéssel definiáljuk, elkerülendõ bizonyos fogalmi nehézségeket. Diszkrét esetben a definíció viszonylag egyszerû.

Definíció 35 *Legyen,* $(\Omega, \mathcal{F}, \mathbb{P})$ *valószínüségi mezõ,* X, Y *két valószínüségi változó. legyenek értékkészleteik rendre* x_i*,*y_i*,azt az eseményt pedig, hogy* $X = x_i, A_i$ *illetve* $Y = y_j$ *jelöje* B_j*. Ekkor*

$$\begin{aligned} E(X|Y) &= \sum_{j=0}^{\infty} E(X|B_i) \mathbb{I}(B_i) \\ &= \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} x_i \mathbb{P}(A_i|B_j) \mathbb{I}(B_i). \end{aligned}$$

Definíció 36 *Folytonos* $(\Omega, \mathcal{F}, \mathbb{P})$ *valószínûségi mezõ és* X, Y *változók esetén az* $f(x|Y)$ *sûrûség függvényt a* $h(x, y)$ *együttes sûrûségfüggvényen keresztül definiáljuk, feltéve annak létezését, valamint, hogy* Y *sûrûségfüggvénye* $f_Y(y) > 0$*:*

$$f(x|y) = \frac{h(x, y)}{f_Y(y)}.$$

Definíció 37 *Folytonos* $(\Omega, \mathcal{F}, \mathbb{P})$ *valószínûségi mezõ és* X, Y *változók esetén tegyük fel, hogy létezik az* $f(x|Y)$ *sûrûség függvény, ekkor a feltételes várható érték*

$$E(X|Y) = \int x f(x|Y)\, dx.$$

Chapter 3
STATISZTIKAI ALAPFOGALMAK

3.1 Bevezetõ, sokaság, minta

A statisztika mint azt a bevezetõben említettük két fõ ágra bomlik, leíró statisztikára és matematikai statisztikára. Utóbbi módszerei matematikai alapon nyugszanak, lényegüket tekintve viszont olyan praktikusan alkalmazható módszereket foglal össze, amelyek segítségével korlátozott ismeretek alapán, kövtekeztetéseket lehet levonni, ezek alapján például üzleti döntéseket lehet hozni mégpedig úgy, hogy eközben a következtetés, döntés megbízhatóságára vonatkozóan is vannak ismereteink. Az élet számos területén találkozunk ilyen feladatokkal, az üzleti élet különösen sok ilyet állít elénk. Állandóan döntéseket kell hoznunk, kockázatválalást, esélylatolgatást kell végeznünk. Álljon itt csak egyetlen példa. Új terméket szereténk piacra dobni. A termék fogadtatása nagyban függ a bevezetõ reklám sikerétõl. A kampány kialakítására több alternatívát is kidolgozunk. Melyik lesz igazán eredményes? Melyiket válasszuk? A szubjekív vezetõi döntés, a vállalati tapasztalatok felhasználása éppúgy elképzelhetõ, mint a megcélzott fogyasztói csoportonból kiválasztott kis csoporton végzett kísérlet, mérés. Azaz ún peer csoportot hívunk meg mintát választunk ki. Ezt több alcsoportra bontjuk és az egyes csoportokon mefelelõ módszerekkel lemérjük a reklám hatékonyságát. Mit jegyeznek meg, mivel társítják s.t.b. . Az így kapott eredményeket mintegy kivetítjük a teljes fogyasztói körre, célközönségre, feltételezzük, hogy (kellõen nagy és gondosan választott minta és gondosan kivitelezett mérés eredményeképpen) az alternatív kampányok közül az lesz a leghatékonyab a piacon, amelyik a kísérleti körülmények között az volt.

Általánosan tehát egy sokaság, azaz egyedek objektumok összeségének bizonyos, teljes egészében nem megfigyelhetõ tulajdonságairól szeretnénk tudomást szerezni. Ehhez mintát veszünk (megfelelõ módon) a sokaságból. A minta minden elemének kérdéses tulajdonságait megvizsgáljuk, majd a kapott eredmény segítségével visszakövetkeztetünk a teljes sokaság ugyanezen tulajdonságaira. Szokás ezt statisztikai következtetésnek nevezni. (lásd **??** ábrát).

Ismrekedésünket a statisztikai módszerekkel az egyváltozós statisztika körében kezdjük, azaz amikor a sokaság egyedeinek egyetlen tulajdonságát vizsgáljuk. Ilyen például az az egyszerû adat, hogy ki hány percet beszél telefonon, hányas cipõt visel, hány éves. Ha egy értékpapírt vizsgálunk, a sokaság lehet a papír kereskedésének napjai, a tulajdonság pedig a napi árváltozás.

Általánosságban egy $(\Omega, \mathcal{F}, \mathbb{P})$ valószínüségi mezõt vizsgálunk, ahol $\mathbb{P}$ nem ismert. A $\mathbb{P}$ valószínüségi mérték gyakran egy ismert mértékcsalád eleme, amelyet (egy vagy több) paraméter jellemez. Ilyen lehet például a normális eloszlás vagy a Poisson eloszlás. Ez esetben paraméteres problémát vizsgálunk.

Késõbb nem paraméteres problémák vizsgálatára is szép módszerekkel ismerkedünk majd meg, de elõször a paraméteres problémákkal foglalkozunk, mivel történetileg is ezek alakultak ki elõbb és talán az elsõ ismerkedés is velül a könyebb.

Fenti példáink is számszerüsíthetõ tulajdonságra vonatkoztak, tehát alapvetõen egy $(\Omega, \mathcal{F}, \mathbb{P})$ valószínüségi mezõn értelmezett X valószínüségi változót vizsálunk. Mirõl is van szó? Miért ne a sokaság összes eleme a vizsgálódás tárgya. Általában nincs mód megmérni a sokaság minden elemére vonatkozóan a kérdéses mennyiséget, ez eleve lehetetlen, vagy igen költséges, esetleg más okból kerülendõ. Például egy új termék bevezetése elõtt természetesen nincs információnk a fogadtatásró, ha meg már bevezettünk a kérdés már eldõlt. Más esetekben a teljes körû mérés bár lehetséges lenne, de a konkurensek elõtt titkolni szeretnénk milyen döntésre készülünk, ezért ezt el kell kerülni.

Igy tehát modellt alkutunk. A sokaság bármely eleme lehetne a megfigyelésünk tárgya, ezért azt tesszük fel, hogy a mefigyelt objektum véletlen, annak tulajdonságát méri X, ez pedig egy valószínüségi változó. Érdeklõdésünket erre az X-re összpontosítjuk. Ahhoz hogy az X eloszlásáról képet alkussunk mintát vezünk a sokaságból, megnérjük a minta elemein a kérdéses mennyiséget, a kapott értékeket jelölje $X_1, X_2 ... X_n$.

A statisztika készítés gyakorlatának egyik sarkalatos lépése a minta kiválasztása. Késõbb röviden majd kitérünk majd a mintavételi eljárások alapjaira, de ezek alapvetõen meghaladják e jegyzet kereteit.

Mindvégig feltesszük, hogy az X_i valószínüségi változók teljesen függetlenek és azonos eloszlásúak, ha ettõl a konvenciótól eltérünk, azt expliciten ott megjegyezzük.

3.2 Alapstatisztikák

Legyenek $X_1, X_2 ... X_n$ független azonos eloszlású valószínüségi változók. Azaz az X valószínüségi változó független replikátumai. Azt a tényt, hogy az X, Y változó azonos eloszlású az $X \sim Y$-al fogjuk röviden jelölni. A fenti feltevés tehát azt jelenti, hogy $X_i \sim X$ minden $i = 1, 2, ...n$-re.

Definíció 38 *Az*

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

értéket mintaátlagnak nevezzük. Ha konkrét relaizációról beszélünk akkor ezt néha a

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

jelöléssel hangsúlyozzuk.

Állítás 1 *Ha létezik a $\mu = E(X)$ várható érték, akkor*

$$E\left(\overline{X}\right) = \mu.$$

A bizonyítást az olvasóra bízzuk.

Definíció 39 *A minta tapasztalati szórásnégyzete, avagy varianciája ($n > 1$ esetén)*

$$V = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

és a tapasztalati szórás

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

Ha a konkrét realizációt akarjuk hangsúlyozni, akkor a

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

jelölést fogjuk használni.

Tétel 3 *(Steiner) Tetszõleges x_i és c valósakra*

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

Bizonyítás.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - c)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i - \overline{x})(\overline{x} - c) + (\overline{x} - c)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 + (\overline{x} - c)^2\end{aligned}$$

mert a $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. ■

Következmény 1

$$\min_c \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Állítás 2

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2 \tag{3.1}$$

Bizonyítás. Következik a Steiner Tételbõl $c = 0-$val. ■

A tapasztalait szorásnégyzet és szórás melett bevezetjük a korrigált tapasztalati szórásnégyzetet szórást.

Definíció 40 *A korrigált tapasztalati szórásnégyzet*

$$V^* = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

és korrigált tapasztalati szórás

$$s^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

3.3 Határeloszlástételek avagy a valóság megismerhetõsége

A cím esetleg talányosnak vagy fellengzõsnek tûnik, de mindjárt látni fogjuk, hogy a határeloszlás tételek valóban a valóság megismerésének egyik kulcsa. A formális állítások elõtt hagy utaljunk vissza a legegyszerûbb, mondhatni ma már közhely számba menõ összfüggésre. Ha egy pézdarabbal sok dobást végzünk, a fejel részaránya egyre pontosabban közelíti az 1/2 értéket. Ez kicsit általánosabban is így van, ha a pénz nem szabályos, vagy más függelenül ismételhetõ azonos módon lezajló kísérletet ismételgetünk, amelynek több véletlen kimenetele van, akkor egy adott kimenetel relatív gyakorisága egyre pontosabban közelíti azt az értéket, amit axiomakent mint valószíûség ahhoz rendelünk. Példa lehet akár a kocka dobás esetén a 6-s kimenetele. De lehet az a megfigyelt jelenség, hogy milyen valószínüséggel választanak a vevõk a jobb illetve a bal kezük ügyébe esõ ugyanazon termékbõl.

Az elõzõ részben láttuk, hogy a tapasztalati átlag átlaga maga a sokaság átlaga (lásd 1 Állítás), azaz a mintaátlag jól céloz. Ennél jóval több is igaz. Elõször is a mintaátlag szórására vonatkozó alapvetõ észrevétel következik.

Állítás 3 *Ha X-nek létezik várható értéke és szórása σ,akkor az $X-$bõl vett n elemû minta átlagára igaz, hogy*

$$\sigma\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}. \tag{3.2}$$

Bizonyítás. A függetlenség alapján

$$\sigma^2\left(\overline{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\left(X_i\right) = \frac{n\sigma}{n^2}$$

amibõl az állítás már adódik. ■

Definíció 41 *Ha egy A esemény bekövetkezésére vonatkozóan ismételt független kísérleteket végzünk, kiszámolhatjuk a tapasztalati részarányt a bekövetkezések k_A száma és a megfigyelések n számának hányadosaként.*

$$\overline{p} = \frac{k_A}{n}.$$

Termeszetesen azt várjuk, hogy $\overline{p}$ jól közelíti az ismeretlen $p = \overline{P}(A)$ értéket. Valóban, igaz a következõ.

Tétel 4 *(A nagy számok Bernoulli-féle törvénye)*

Tekintsünk egy A eseményt, amelynek valószínüsége $p = \mathbb{P}(A)$ $(0 < p < 1)$. Legyen n független megfigyelésbõl a relatív gyakoriság $\overline{p} = \overline{p}_n(A)$, ekkor minden $\varepsilon, \delta > 0$ -ra létezik $N = N(\varepsilon, \delta)$,hogy minden $n > N$-re

$$\mathbb{P}(|\overline{p}_n - p| < \varepsilon) > 1 - \delta.$$

Definíció 42 *Legyen adott egy F eloszlásfüggvényû valószínüségi változó X. Az ennõl vett n elemõ minta egy tapasztalati eloszlásfüggvényt határoz meg. Legyen $k_n(x)$ az n elemû mintában az x érték alá esõ x_i megfigyelések száma. Ekkor az $F_n(x)$ tapasztalati eloszlásfüggvény a következõ*

$$F_n(x) = \frac{k_n(x)}{n}.$$

Következmény 2 *Ha speciálisan $A = \{X < x\}$ valamely valószínüségi változóra, akkor*

$$\mathbb{P}(|F_n(x) - F(x)| < \varepsilon) > 1 - \delta$$

ha $n > N$.

Tétel 5 *(Nagy számok erõs törvénye)*

Legyen egy esemény $A, p = \mathbb{P}(A)$ $(0 < p < 1)$ valószínûséggel. Legyen n független megfigyeléspõl a relatív gyakorisága $\overline{p} = \overline{p}_n(A)$, ekkor minden $\varepsilon > 0$ -ra létezik $N = N(\varepsilon)$,hogy minden $n > N$-re

$$\mathbb{P}(|F_n(x) - F(x)| < \varepsilon) = 1.$$

Tétel 6 *(Centrális határeloszlás tétel) Ha $X_1, X_2 ... X_n$ független azonos eloszlású valószínüségi változók μ várható értékkel és σ szórással, akkor*

$$\mathbb{P}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < y\right) \underset{n\to\infty}{\to} \Phi(y)$$

ahol Φ a standard normális eloszlás eloszlásfüggvénye.

Megjegyzés 1 *Az állítás kissé pongyolán azt jelenti, hogy elég nagy n esetén (n>30 használható mint ökölszabály) az $\overline{X}$ mint valószínüségi változó közelítõleg normális eloszlású, pontosabban $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ eloszlású.*

Tétel 7 *Ha X normális eloszlású valószínüségi változó, akkor*

$$\mathbb{P}\left(\frac{\overline{X} - \mu}{s/\sqrt{n}} < y\right) = t^{(n-1)}(y)$$

ahol $t^{(n-1)}(y)$ az $n-1$ szabadságfokú Student eloszlás eloszlásfüggvénye.

Megjegyzés 2 *Ha $n > 30$ a Student eloszlás igen jól közelíti a standard normális eloszlást. Vegyük észre, hogy a második tétel sokkal erõsebb feltevésen alapszik, miszerint az alapsokaság normális eloszlású, cserébe viszont nem csak közelítõ állítás adható, hanem az eloszlás, a Student eloszlás akkor is adódik, ha a szórás értékének ismerete hiányában $s/\sqrt{n}$-el normáljuk a tapasztalati várható értéket.*

Megjegyzés 3 *Megjegyezzük, hogy normális eloszlásíú valószínüségi változók összege is normális, azaz a tétel feltételei melett $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$maga standard normális eloszlású. A gyakorlatban az okoz nehézséget, hogy a sokaság szórása, sõt esetleg várható értéke sem ismert. Többek között a küzdelem ezen paraméterek meghatározásáért folyik és mindkét tétel fõ mondanivalója éppen az, hogy a minta elemszámának növelésével a minta és a sokaság átlaga kis, kontrolált valószínüséggel tér csak el egymástól.*

A centrális határeloszlás tétel speciális esete a bevezetõben említett feladat egzakt megoldása.

Tétel 8

$$\mathbb{P}\left(\frac{\overline{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < y\right) \underset{n\to\infty}{\to} \Phi(y).$$

Megjegyzés 4 *Azaz megint csak kissé egyszerüsítve, nagy n-re $\overline{p}$ közelítõleg $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ eloszlású. Igaz továbbá az is, hogy a nem ismert p helyett annak közelítését helyettesítve a képletbe az továbbra is fennáll, azaz:*

$$\mathbb{P}\left(\frac{\overline{p}-p}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} < y\right) \underset{n\to\infty}{\to} \Phi(y).$$

*Az 6,7,7 Tételeket nem igazoljuk (lásd [**?**]).*

Ezek után a megismerhetõség még teljesebb voltát igazoló tételt, a statisztika alaptételét mondjuk ki. Ez lényegében azt állítja, hogy az F eloszlásfüggvényû valószínüségi változóból vett egyre nagyobb mintából az F tetszõlegesen pontosan meghatározható.

Tétel 9 *(A statisztika alaptétele)*

$$\sup_x |F_n(x) - F(x)| \underset{n\to\infty}{\to} 0$$

1 valószínüséggel.

A bizonyítástól eltekintünk. Az egyetlen nehézséget a szuprémum kezelése jelenti, hiszen minden fix x-re az $A = \{X < x\}$ eseményre alkalmazható a fenti tétel, hiszen $F(x) = \mathbb{P}(A)$, $F_n(x) = \overline{p}(A)$.

Az alábbiakban sokkal erõsebb, úgynevezett próbákat is biztosító élesebb eredményeket ismertetünk.

Tétel 10 *(Szmirnov)*

$$\lim_{n\to\infty} \mathbb{P}\left[\sqrt{n}\left(F_n\left(x\right)-F\left(x\right)\right)<y\right]=S\left(y\right)$$

ahol

$$S\left(y\right)=\left\{\begin{array}{cll} 0 & \textit{ha} & y\leq 0 \\ 1-e^{-2y^2} & \textit{ha} & y>0 \end{array}\right. .$$

Tétel 11 *(Kolmogorov)*

$$\lim_{n\to\infty} \mathbb{P}\left[\sqrt{n}\left|F_n\left(x\right)-F\left(x\right)\right|<y\right]=K\left(y\right)$$

ahol

$$K\left(y\right)=\left\{\begin{array}{cll} 0 & \textit{ha} & y\leq 0 \\ \sum_{i=-\infty}^{\infty}\left(-1\right)^i e^{-2i^2y^2} & \textit{ha} & y>0 \end{array}\right. .$$

Tétel 12 *(Gnyegyenko) Legyen* $c=\left\lceil y\sqrt{2n}\right\rceil$ *, ha* F_n *és* G_n *két tapasztalati eloszlásfüggvény, valamely* F,G *eloszlásfüggvény* n *elemû mintáira vonatkozóan, akkor* $F=G$ *-bõl következkzik, hogy*

$$\mathbb{P}\left(\sqrt{\frac{n}{2}}\sup_x\left(F_n\left(x\right)-G_n\left(x\right)\right)<y\right)=\left\{\begin{array}{cll} 0 & \textit{ha} & y\leq 0 \\ 1-\frac{\binom{2n}{n+c}}{\binom{2n}{n}} & \textit{ha} & 0<y<\sqrt{\frac{n}{2}} \\ 1 & & \textit{egyébként} \end{array}\right.$$

A tétel igazlásához bevezetõül egy klasszikus kombinatorikus gondolatmenetre van szükségünk.

Probléma 13 *Az alább ismertetendõ balott-tétel a szavazatszámlás lefolyására vonatkozó egy érdekes összefüggést mutat be. Ha egy polgármesteri posztért két jelölt* A *és* B *küzdött, a lehetséges* $n=a+b$ *szavazatból* $a-t$ *szerzett meg* A*,* $b<a-t$ *a pedig* B*, felfet[dik, hogy mi annak a valószínüsége, hogy a szavazatok összeszámlálása során mindvégig* A *vezet, azaz az összes részeredmény is az õ gyõzelmét jelenti? Az összeszámlálás folyamatát jól lehet ábrázolni. Tekintsük a szokásos síkbeli koordináta rendszert. Induljunk ki az origóból és rajzoljunk egy* $(1,1)$ *vektort, ha az elsõ szavazatot* A *kapta, ellenkezõ esetben az* $(1,-1)$ *vektort rajzoljuk. Ezt az eljárást folytatjuk mindíg az elõbbi két vektor egyikének lerajzolásával "megtoldva" az addig lerajzolt törött vonalat. (lásd a* **??** *ábrát.) Világos, hogy a törött vonal* n *hosszúságú és az* $y=a-b$ *magasságban ér véget, azaz a* $(0,0)$ *és* $(n,a-b)$ *pontokat köti össze. Hívjuk röviden az ílyen szabállyal rajzolt törött vonalakat utaknak.*

Tétel 14 *(ballot-tétel) Legyen* $n\geq a>0, b=n-a$*. Azon utak hossza amelyek azorigóból indulnak és* $(n,a-b)$*-ben úgy végzõdnek, hogy közben végig a felsõ félsíkban haladnak az* x *tengely érintése nélkül egyenlõ*

$$\frac{a-b}{n}\binom{n}{a}.$$

A bizonyítás az igen sok helyen alkalmazható tükrözési elven alapul. Jelölje egy síkbeli C pont x tengelyre vonatkozó tükörképét C'.

Tétel 15 *(tükrözési elv) Az x tengelyt érintõ vagy metszõ C pontból D-be vezetõ utak száma megegyezik az összes C' és D közötti utak számával.*

Bizonyítás. A bizonyítás a leszámlálások köréből ismert módon történik, úgy, hogy a kérdéses utak között egy egy-egy értelmû megfeleltetést adunk, amibõl persze következik, hogy az utak száma is egyenlõ. Vegyüknk szemre egy adott olyan utat ami $C-$bõl D-be halad és érinti vagy metszi az x tengelyt. Ezen metszések közül van elsõ, azaz balról jobbra (a lerajzolás sorrendje szerint, azaz idõben) az elsõ pont ahol az út érinti vagy metszi az x tengelyt. Legyen ez a pont $(k,0)$, azaz a metszés a k-adik szakasz lerajzolásakor. Tükrözzük az út $(0,0)$ és $(k,0)$ közötti részét az x tengelyre (lásd **??** Ábra). Ezzel egy C'-bõl D-be tartró úthoz jutunk. Az is világos, hogy minden ilyen C'-bõl D-be tartró út valahol metszi az x tengelyt és mivel az elsõ érintést illetve metszést választottuk a C, D úton, ezért a $(k,0)$ pont lesz az elsõ metszés a C', D úton is. Ez az elsõ (érintés illetve) metszéspont az utak mindkét szóban forgó halmazát diszjunkt részhalmazokra osztja. Az egy-egy értelmû megfeleltetést ezen részhalmazokon hozzuk létre. A tükrözés kölcsönösen egyértelmû megfeleltetés, ezért az utak között így kapott megfeleltetés is egy-egy értelmû. Ebbõl viszont következik, hogy a megfelelõ utak száma is egyenlõ. ■

Következmény 3 *Annak a valószínüsége, hogy a fenti szavazatszámlálási problémában végig A vezet $\frac{a-b}{a+b}$.*

Bizonyítás. Világos, hogyha az összes szavazatsorrendek száma $\binom{a+b}{a}$, ha ezek mind egyenlõen valószínûek, akkor egy adott sorrend valószínüsége $\frac{1}{\binom{a+b}{a}}$. A 14 tételbõl ezért következik, hogy azon utak száma amik végig az x tengely felett haladnak az összes lehetséges utak számából kivéve azokat, amelyek érintenek vagy metszenek. Természetesen az elsõ lépés felfelé kell, hogy történjen, ezért a vizsgált összes út $(1,1)$-bõl halad $(n, a-b)$-be, ahol $n = a+b$, vagyis egyel balra és letolva $(0,0)$ $-$bõl halad $(n-1, a-b-1)$-be, ezek száma $\binom{n-1}{a-1}$,ugyanakkor a ballot tétel szerint az érintõ vagy metszõ utak száma egyenlõ a $(0,-1)$, (n,a) utak számával, azaz $\binom{n-1}{a}$-val, a keresett nem érintõ utak száma ezért

$$\begin{aligned}\binom{n-1}{a-1} - \binom{n-1}{a} &= \frac{(n-1)!}{(a-1)!b!} - \frac{(n-1)!}{a!\,(b-1)!} \\ &= \frac{a}{n}\frac{n!}{a!b!} - \frac{b}{n}\frac{n!}{a!b!} = \frac{a-b}{a+b}\binom{n}{a}.\end{aligned}$$

Mivel az egyes utak (szavazócédula sorrendek) valószínüsége azonos, ezek összes száma pedig $\binom{n}{a}$,ezzel igazoltuk, hogy annak a valószínûsége, hogy a számlálás során végig A vezet $\frac{a-b}{a+b}$. ■

A ballot tétel gondolatmenete segítségével igazolhatjuk Gnyegyenko tételét.

Elõször egy újabb egyszerõ kombinatorikus gondolatra van szükségünk. Legyen $X_1, X_2 ... X_n$ az F illetve $Y_1, Y_2 ... Y_n$ a G-ból választott n elemû minta. Keverjük össze és

rendezzük nagyság szerint õket. Így a $Z_1....Z_{2n}$ sorozathoz jutunk. Feltesszük, hogy ezen értékek mind különbözõek. Legyen

$$\varepsilon_i = \begin{cases} 1 & \text{ha } Z_i \text{ az } X_k \text{ sorozat eleme} \\ -1 & \text{egyébként} \end{cases}.$$

Ezzel megint egy 2n hosszú töröttvonalat is kapunk, ha az $(1, \varepsilon_i)$ vektorokat összefûzzük.

Lemma 1 *A fenti jelölések mellett*

$$\sup_x \left(F_n(x) - G_n(x)\right) = \frac{1}{n} \max_{0 \le i \le 2n} S_i.$$

Bizonyítás. Az $n\left(F_n(x) - G_n(x)\right)$ kifejezés az x-nél kisebb X_j beli és Y_k beli elemek számának különbsége. Az x növekedtével ez pontosan akkor változik, mégpedig ε_i-vel, ha X_i illetve Y_i éppen x. ■

A 12 Tétel bizonyítása. Mivel a feltevés szerint $F = G$ ezért az $X_i$és Y_i sorozatok elemei azonos eloszlásúak és teljesen függetlenek, amibõl következik, hogy azok összes sorrendje azonos valószínûségû. Azaz bármelyik sorozat valószínûsége $\frac{1}{\binom{2n}{n}}$. Azon sorozatok száma pedig amelyekre $\max_{0 \le i \le 2n} S_i < z\sqrt{2n}$ azon utak száma, amelyek a $c = \left\lceil z\sqrt{2n} \right\rceil$ egyenes alatt maradnak. Alkalmazzuk a tükrözési elvet most az $y = c$ egyenesre (lásd **??** Ábra). Világos, hogy $(0,0)$ −ból a $(2n,0)$ −ba haladó $c - t$ nem érintõ utak száma $\binom{2n}{n} - \binom{2n}{n-c}$ lesz. Osztva az összes utak számával adódik az állítás. ■

Megjegyzés 5 *Határátmenet képzésével Gnyegyenkó 12 tételébõl követezik Szmirnov 10 tétele. A tükrözési elv ismételt alkalmazásával és határátmenet képzésével lehet igazolni Kolmogorov 11 tételét is.*

Chapter 4
BECSLÉSELMÉLET

A vizsgált $(\Omega, \mathcal{F}, \mathbb{P})$ valószínüségi mezõt gyakran jellemzi egy a $\mathbb{P}$ valószínüségi métrékhez kapcsolódó ϑ paraméter. Ezt néha a $\mathbb{P}_\vartheta$ jelöléssel is hangsúlyozzák. Ilyen paraméter mondjuk az exponenciális eloszlás $\lambda-$ja, vagy a normális eloszlás várható értéke, μ,vagy szorása σ. De ha például egy pénz feldobásánál a fej falószínûségét vizsgáljuk, lehet a ϑ paraméter az $1/2-$tõl való eltérés is. Igen gyakori feladat, hogy egy $X_1, X_2...X_n$ mintából megbecsüljük $\vartheta - t$.

Definíció 43 *Az $X_1, X_2...X_n$ mintából készített statisztika*

$$\overline{\vartheta} = f(X_1, X_2, ..., X_n)$$

az aminek segítségével vissza kívánunk következtetni a sokaságot jellemzõ ϑ-ra.

A feladat megfogalmazása már magában rejti azt, hogy valamilyen elõfeltevéssel élünk (általában) a sokaságot jellemzõ $\mathbb{P}$ valószínûségi mértéket illetõen. Várható értékrõl sok esetben lehet például beszélni, de a Cauchy eloszlásnak nincs várható értéke, helyette helyzeti paraméterrõl szokás beszélni. Hasonlóan λ-ról beszélünk az exponenciális eloszlás esetében, tehát például, ha egymást követõ telefonhívások között eltelt idõt tekintünk, ilyenkor jó okkal feltesszük, hogy valamilyen ismeretlen λ paraméterû, de exponenciális eloszlással van dolgunk. Ha viszont a jelenségrõl tudható, hogy norális eloszlást követ, semmi értelme λ becslsérõl beszélni.

Megközelítésünk teháte elõzetes feltevésre, a jelenség valamilyen szintû ismeretére alapoz. Ez általában jellemzi a paraméteres statisztikai vizsgálatokat. A következõ fejezetekben errõl lesz szó.

Elkészítve egy $\overline{\vartheta} = f(X_1, X_2, ..., X_n)$ statisztikát, illetve becslést felmerül a kérdés, meyyire "jó" ez a becslés. Például megéri-e n elemen elvégezni a mérést. Ez a kérdés különösen akkor fontos, ha a vizsgálat tönkreteszi annak tárgyát, például a sorozatgyártásból kiemelt villanyégõket tartóssági tesztnek vetik alá. Több ezer órán át égnek, mérik a teljes élettaratmukat. Ezt pedig csak akkor állapíthatják meg, ha a lámpa végül kiég. Természetesen ez az eljárás hosszadalmas és költséges, ha a teljes legyártott mennyiségen hajtanánk végre, akkor remek becslésünk lenne az átlagos élettartamra, csak nem lenne egy eladható égõnk sem. Így tisztázni kell, hogy hány elemû mintát érdemes használni. Ehhez egyrészt azt kell tisztázni, milyen pontos becslésre, milyen megbízható becslésre van szükségünk, másrészt azt kell valahogy megállapítanunk, hogy maga a mérési módszer, esetünkben a $\overline{\vartheta} = f(X_1, X_2, ..., X_n)$ statisztika milyen hibát "hordoz" magában. Az elõbbi példára visszatérve, az égõk élettartamára elég ± 10 óra pontos becslést adni. Jó

lenne ugyanakkor, ha az adott becslésünk megbízható lenne, azaz mondjuk, ha 1000 szállítmányra alkalmazzuk, akkor lehetõleg csak egy-két esetben forduljon elõ, hogy a valódi élettartam és az általunk ígért (mérés alapján becsült) érték jobban eltér mint ± 10 óra. Szintén jogos elvárás lehet, hogy a költség, azaz a minta elemszámának növekedtével javuljon a becslés valamelyik fenti értelemben.

A becslés elmélet ezen kérdések ekzakt megközelítésére szolgál, melynek alapai kerülnek ebben a fejezetben bemutatásra.

Az alábbiakban tehát egy ismeretlen ϑ paraméter $\overline{\vartheta}$ becslésének jóságát leíró fogalmakat vezetünk be, majd néhány egyszerû statisztikára alkalmazzuk e fogalmakat.

Definíció 44 *Azt mondjuk, hogy* $\overline{\vartheta} = f(X_1, X_2, ..., X_n)$ *torzítatlan becslése* ϑ*-nek, ha minden* ϑ *esetén*

$$\mathbb{E}_\vartheta\left(\overline{\vartheta}\right) = \vartheta.$$

Itt a várható érték a $\mathbb{P}_\vartheta$ valószínûségi mértékbõl vett független X_i mintaelemekre vonatkozik.

Állítás 4 *Legyen*

$$T(X_1, X_2 ... X_n) = \sum_{i=1}^{n} c_i X_i$$

Tegyük fel, hogy $X-$*nek létezik* μ *várható értéke. Ekkor* T *akkor és csak akkor torzítatlan, ha*

$$\sum_{i=1}^{n} c_i = 1.$$

Bizonyítás. A várható érték linearitása miatt

$$\mathbb{E}_\mu(T) = \sum_{i=1}^{n} c_i E(X_i) = \mu \sum_{i=1}^{n} c_i$$

amibõl következik az állítás. ■

Következmény 4 *Ebbõl következik* $c_i = \frac{1}{n}-$*et tekintve az is, hogy a mintaátlag a sokaság átlagának torzítatlan becslése, azaz*

$$\mathbb{E}\left(\overline{X}\right) = \mu.$$

Definíció 45 *Azt mondjuk, hogy egy* $\overline{\vartheta}$ *becslés hatásosabb mint egy* $\overline{\vartheta}'$*, ha torzítatlan becslések és*

$$\sigma\left(\overline{\vartheta}\right) \leq \sigma\left(\overline{\vartheta}'\right),$$

feltéve persze a szórások létezését.

Definíció 46 *Azt mondjuk, hogy* $\overline{\vartheta}$ *hatásos, ha minden más* $\overline{\vartheta}'$*-nél hatásosabb.*

Definíció 47 *Egy $\overline{\vartheta}'$ becslés hatásfokáról akkor beszélhetünk, ha létezik egy hatásos $\overline{\vartheta}$ becslés, ekkor a hatásfok (efficiencia)*

$$e\left(\overline{\vartheta}'\right) = \frac{\sigma\left(\overline{\vartheta}'\right)}{\sigma\left(\overline{\vartheta}\right)}.$$

Állítás 5 *Ha a vizsgált eloszlás normális akkor $\overline{X}$ hatásos.*

Ezt az állítást nem igazoljuk, de a következõt viszont igen.

Állítás 6 *A várható érték lineáris becslései közül a mintaátlag a leghatékonyabb, ha létezik második momentum..*

Bizonyítás. Tekintsünk egy

$$T = \sum_{i=1}^{n} c_i X_i$$

lineáris becslést. Ekkor a függetlenség miatt

$$\sigma^2(T) = \sigma^2\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \sigma^2.$$

De a számtani és mértani közép közötti összefüggésbõl tudjuk hogy

$$\sum_{i=1}^{n} c_i^2 \geq \frac{1}{n}\left(\sum_{i=1}^{n} c_i\right)^2 = \frac{1}{n}$$

és egyenlõség akkor és csak akkor áll fenn ha minden c_i azonos, egyenlõ $\frac{1}{n}$-el. Így tehát

$$\sigma^2(T) \geq \sigma^2\left(\overline{X}\right).$$

■

Definíció 48 *Egy $\overline{\vartheta}_n = f(X_1, X_2, ..., X_n)$ becslés asszimptotikusan torzítatlan, ha*

$$\mathbb{E}\left(\overline{\vartheta}_n\right) \underset{n\to\infty}{\to} \vartheta.$$

Definíció 49 *Egy $\overline{\vartheta}$ becslés gyengén konzisztns, ha minden $\varepsilon > 0$-ra létezik $\delta > 0$ és $N > 0$, hogy minden $n > N$-re*

$$\mathbb{P}_\vartheta\left(\left|\overline{\vartheta} - \vartheta\right| < \varepsilon\right) \geq 1 - \delta. \tag{4.1}$$

Erõsen konzisztens, ha

$$\mathbb{P}_\vartheta\left(\lim_{n\to\infty} \overline{\vartheta} = \vartheta\right) = 1 \tag{4.2}$$

valamint n;gyzetes k0z;pben, ha

$$\mathbb{E}_\vartheta\left(\left|\overline{\vartheta} - \vartheta\right|^2\right) \underset{n\to\infty}{\to} 0. \tag{4.3}$$

Mint azt a valószínûségszámítás alapjainál megismertük az erõs (4.2)és négyzetes középben (4.3) vett konvergencia (esetünkben konzisztencia) egyaránt maga után vonja a gyenge értelemben vett konvergenciát (4.1)(konzisztenciát).

Tétel 16 *Ha létezik második momentum, akkor $\overline{X}$ erõsen konzisztens becslés.*

Az állítás következi a nagy számok erõs törvényébõl.

Tétel 17 *A tapasztalati szórásnégyzet V és a korrigált tapasztalati szórásnégyzet V^* egyaránt erõsen konzisztens.*

Bizonyítás.

$$V = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \to \mathbb{E}\left(\overline{X}^2\right) - \mathbb{E}\left(\overline{X}\right)^2$$

igaz egy valószínûséggel a nagy számok erõs törvénye szerint. Hasonlóan a korrigált tapasztalati szórásnégyzetre is $\frac{n-1}{n} \to 1$ miatt. ■

Tétel 18 *A V tapasztalati szórás nem torzítatlan becslése σ^2-nek, S^* viszont az.*

Bizonyítás.

$$\begin{aligned}
\mathbb{E}(V) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu+\mu-\overline{x})^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{(x_i-\mu)^2 - 2(x_i-\mu)(\overline{x}-\mu)+(\overline{x}-\mu)^2\right\}\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - 2n(\overline{x}-\mu)^2 + n(\overline{x}-\mu)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - n(\overline{x}-\mu)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right] - \mathbb{E}\left[(\overline{x}-\mu)^2\right]
\end{aligned}$$

De mint láttuk $\sigma^2\left(\overline{X}\right) = \frac{\sigma}{n}$ (lásd 3.2) ezért az utolsó kifejezés egyenlõ

$$\sigma^2 - \mathbb{E}\left[(\overline{x}-\mu)^2\right] = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Természetesen a korrigálás éppen kioltja az $\frac{n-1}{n}$ tényezõt, így a korrigált tapasztalati szórásnégyzet már torzítatlan. ■

Definíció 50 *Egy $\overline{\vartheta}$ statisztikát elégségesnek nevezünk, ha a becsülendõ ϑ paraméterre vonatkozó minden információt tartalmaz, azaz ha $F_{\alpha,n}$ jelöli az $X_1, X_2, ..., X_n$ együttes eloszlásfüggvényét, akkor*

$$F_{\alpha,n}\left(x|\overline{\vartheta}=t\right)=F_n\left(x|\overline{\vartheta}=t\right),$$

azaz a $\overline{\vartheta}=t$ feltevés melett a jobboldal mint formula sem tartalmazza a α paramétert.

E fogalom megértéséhez az alábbi példa a dhat segítséget.

Állítás 7 *Egy Poisson eloszlású sokaság paraméterének becslésére a mintaközép elégséges becslés.*

Bizonyítás. Tegyük fel, hogy a paraméter λ. Mint tudjuk, Poisson eloszlású változók összszege is ilyen és a paramétereik összegzõdnek, ezért speciálisan $n\overline{X}$ is Poisson eloszlású $n\lambda$ paraméterrel. Tekintsük a minta együttes eloszlásfüggvényét és használjuk ki a függetlenséget. Jelölje $\overline{\lambda}=\frac{N}{n}$,ahol $N=\sum_{i=1}^{n}x_i$

$$\begin{aligned}
& P\left(X_1=x_1,X_2=x_2,...,X_n=x_n|\overline{X}=\overline{\lambda}\right)\\
= \; & P\left(X_1=x_1|\overline{X}=\overline{\lambda}\right)P\left(X_2=x_2|\overline{X}=\overline{\lambda}\right)...P\left(X_n=x_n|\overline{X}=\overline{\lambda}\right)P^{-1}\left(n\overline{X}=N\right)\\
= \; & \frac{\lambda^{x_1}}{x_1!}e^{-x_1\lambda}\frac{\lambda^{x_2}}{x_1!}e^{-x_2\lambda}...\frac{\lambda^{x_n}}{x_1!}e^{-x_n\lambda}\frac{N!}{(n\lambda)^N}e^{\lambda N}=\frac{1}{n^N}\binom{N}{x_1,x_{2,...,x_n}}.
\end{aligned}$$

Azaz a kapott képlet valóben nem tartalmazza a λ paramétert, tehát $\overline{X}$ elégséges statisztika.
■

Chapter 5
A LEGNAGYOBB VALÓSZíNÛSÉG ELVE

Ha egy doboz franciadrazsé között zöld, piros és kék színûek tatálhatóak, számuk pedig 20,50,33, akkor ha arra a kérdésre kell választ adnunk, hogy milyen színû lesz a véletlenül híhúzott drazsé, mindenki azt feleli, hogy piros, hiszen ezekbõl van a legtöbb. Ugy is mondhatjuk,, hogy az adott esetben ennek a valószíûsége $\frac{50}{103}$ a legnagyobb, mindenki azt várja, hogy a legnagyobb valószínûségû esemény következik be. Nos ezt a természetes gondolatmenetet sok esetben alkalmazza a valószínûség számítsá és a statisztka is. Ebben a fejezetben a legnagyobb valószínûség elvén alapuló módszerrel idegen szóval a maximum likelihppd módszerrel ismerkedünk meg, újra az egyváltozós egyparaméteres problémák körére szorítkozva.

Példa 19 *Legyen egy legyártott szériában a hibás gyártmányok részaránya ismeretlen p. Azaz ha véletlen szerûen egyet kihízunk a szériából, annak a valószínûsége, hogy selejtest válatszunk p. Szeretnénk mintavétel segítségével meghatározni p-t. Tegyük fel, hogy egy n elemû mintát vettünk ki és abból k bizonyult selejtesnek. Számoljunk formálisan, mi ennek a valószínûsége.*

$$\mathbb{P}_p\left(X=k\right)=\binom{n}{k}p^k\left(1-p\right)^{n-k}$$

Keressük azt a p-t amire ez az érték maximális. Tegyük fel, hogy $0<k<n$.

$$\frac{d}{dp}\mathbb{P}_p\left(X=k\right)=\binom{n}{k}kp^{k-1}\left(1-p\right)^{n-k}-\binom{n}{k}p^k\left(n-k\right)\left(1-p\right)^{n-k-1}$$

A jobboldalt nullával egyenlõvé téve

$$0=k\left(1-p\right)-p\left(n-k\right)$$

adódik, amibõl átrendezéssel a nem túl meglepõ

$$p=\frac{k}{n}$$

adódik. Természetesen meg kell gyõzõdnünk arról, hogy a kapott érték valóban szélsõérték és maximum hely-e, de ez ebben az esetben jól látható.

A fenti példa a modellje a maximum likelihood módszernek.

Definíció 51 *Egy ismeretlen ϑ paraméterû $\mathbb{P}_\vartheta$ valószínûségi mértékbõl vett elemû minta maximum likelihood függvénye diszkrét valószínüségi változó esetében*

$$L(x_1, x_2, ..., x_n|\vartheta) = \mathbb{P}_\vartheta(X_1 = x_1)\,\mathbb{P}_\vartheta(X_2 = x_2)\,...\mathbb{P}_\vartheta(X_n = x_n)$$

illetve folytonos valószínüségi változó esetében

$$L(x_1, x_2, ..., x_n|\vartheta) = f_\vartheta(x_1)\,f_\vartheta(x_2)\,...f_\vartheta(x_n)$$

ahol $f_\vartheta(x)$ a ϑ paraméterû valószínüségi változó sûrûségfüggvénye.

Példa 20 *Határozzük meg a maximum likelihood módszer segítségével a normális eloszlásu sokaság várható értékének és szórásának becslését. Folytonos esetben felhasználva a logaritmus függvény monotonicitását célszerû a likelihood függvény logaritmusát tekinteni. Mivel a normális eloszlás sûrûségfüggvénye*

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ezért

$$\ln L(x_1, x_2, ...x_n|\vartheta) = \sum_{i=1}^{n}\left[-\ln(\sigma) - \frac{1}{2}\ln 2\pi - \frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

Véve ennek μ illetve σ szerinti deriváltját

$$\frac{\partial}{\partial\mu}\ln L(x_1, x_2, ...x_n|\vartheta) = \sum_{i=1}^{n}\frac{2(x_i-\mu)}{2\sigma^2}, \tag{5.1}$$

$$\frac{\partial}{\partial\sigma}\ln L(x_1, x_2, ...x_n|\vartheta) = \sum_{i=1}^{n}\left[-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}\right]. \tag{5.2}$$

Keresve a gyököket

$$\sum_{i=1}^{n}\frac{2(x_i-\mu)}{2\sigma^2} = 0$$

egyenletbõl

$$\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.3}$$

adódik, illetve (5.2)*-bõl abba behelyettesítve* $(5.3) - t$

$$\begin{aligned}\sum_{i=1}^{n}\left[-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3}\right] &= 0\\ \sum_{i=1}^{n}(x_i-\overline{\mu})^2 &= n\overline{\sigma}^2\\ \overline{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{\mu})^2\end{aligned}$$

adódik.

5.1 További példák, feladatok

Gyakorlat 21 *Legyen most a sokaság ismeretlen λ paraméterû Poisson eloszlásu. ekkor a maximum likelihood függvény logaritmusa az*

$$f_\lambda(x) = \frac{\lambda^k}{k!} e^{-\lambda k}$$

diszkrét eloszlásfüggvény alapján

$$\ln L(x_1, x_2, ..., x_n|\lambda) = \log \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!) - n\lambda$$

amibõl

$$\frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$$

alapján a maximum likelihood becslés $\lambda-$ra $\overline{\lambda}_n = \overline{X}$. Ehhez természetesen még szükséges annak ellenõrzése, hogy ez maximum hely, ami következik a második derivált negativitásából:

$$\frac{\partial^2}{\partial \lambda^2} \ln L(x_1, x_2, ..., x_n|\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} x_i < 0$$

ha $x_i \neq 0$.

Gyakorlat 22 *Ha a sokaság λ paraméterû exponenciális eloszlásu, az okoskodás igen hasonló. A sûrûség függvény*

$$f_\lambda(x) = \lambda e^{-x\lambda},$$

ezért a maximum likelihood függvény logaritmusa

$$\ln L(x_1, x_2, ..., x_n|\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Innen deriválásall a

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

egyenlethez jutunk. Tehát $\overline{\lambda} = \overline{X}$, ha a maximum globális, ami megint következik a

$$\frac{\partial^2}{\partial \lambda^2} \ln L(x_1, x_2, ..., x_n|\lambda) = -\frac{n}{\lambda^2} < 0$$

összefüggésbõl.

Gyakorlat 23 *Legyen most egy ismeretlen λ paraméterû Poisson eloszlásunk. Igazoljuk, hogy a mintaátlag elégséges statisztika λ-ra nézve.*

A fenti példákban szinte automatikusan lehet alkalmazni a maximum likelihood módszert, az alábbiakban néhány trükkösebb példa következik.

Gyakorlat 24 *Legyen most az* $X \in [\alpha, 2\alpha]$ *valószínüségi változó, amelynek sûrûségfüggvénye*

$$f(x) = \frac{2x}{3\alpha^2},$$

$x \in [\alpha, 2\alpha]$*-n értelmezve. Adjunk maximum likelihood becslést* α*-ra.*

Gyakorlat 25 *Az visszatevéses mintavétel módszere. Szeretnénk megszámolni, hány kékbálna él egy tengerszakaszon. Legyen a keresett szám* N*. Most ez az ismeretlen paraméter. A következõképpen járunk el. hétig "vadászva" sárga festéklövedékkel jelöljuk meg a bálnákat. Legyen a megjelölt bálnák száma* M*. Ezek után a következõ héten* n *darabot láttunk, ezek közül pedig* s *darab volt sárga festékkel megjelölve. Mi az* N *értékének maximum likelihood becslése? Határozzuk meg elõször az* $L(s) = L(s|N)$ *maximum likelihood függvényt.*

$$L(s|N) = \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{}}$$

n.

Vizsgáljuk a

$$\frac{L(s+1)}{L(s)}$$

hányadost, mely értékekre $>$ *illetve* $<$ *mint* 1*.*

$$\begin{aligned}
\frac{L(s+1)}{L(s)} &= \frac{\frac{M!}{s!(M-s)!}\frac{(N-M)!}{(n-s)!(N-M-n+s)!}\frac{n!(N-n)!}{N!}}{\frac{M!}{(s-1)!(M-s+1)!}\frac{(N-M)!}{(n-s+1)!(N-M-n+s-1)!}\frac{n!(N-n)!}{N!}} \\
&= \frac{\frac{1}{s!(M-s)!}\frac{1}{(n-s)!(N-M-n+s)!}}{\frac{1}{(s-1)!(M-s+1)!}\frac{1}{(n-s+1)!(N-M-n+s-1)!}} \\
&= \frac{(M-s+1)(n-s+1)}{s(N-M-n+s)} \geq 1
\end{aligned}$$

ha

$$N \leq n\frac{M}{s} - 1$$

azaz L *ilyen értékekre nõ, nagyobbakra csökken. ellenõrizni!*

Gyakorlat 26 *Legyen most az* $X \in [\alpha, 2\alpha]$ *egyenletes eloszlású valószínüségi változó. Adjunk maximum likelihood becslést* α*-ra.*

*A problémához egy paradoxon is tartozik (lásd még [***?***]). Ennek ismertetése elõtt oldjuk meg az eredeti feladatot. Nyilván a sûrûség függvény*

$$f_\alpha(x) = \frac{x}{\alpha} : x \in [\alpha, 2\alpha].$$

A maximum likelihood függvény logaritmusa

$$\ln L\left(x_1, x_2, ..., x_n|\alpha\right) = \sum_{i=1}^{n} \ln \frac{x_i}{\alpha}$$

és ennek

$$\frac{\partial}{\partial\alpha} \ln L\left(x_1, x_2, ..., x_n|\alpha\right) = -\frac{n}{\alpha},$$

ahol $X_i \in [\alpha, 2\alpha]$ *miatt* -$\frac{n}{\alpha}$ *a maximumát*

$$\overline{\alpha} = X_n^* = \max_{1\leq i\leq n} X_i$$

helyen veszi fel. Ez tehát amaximum likelihood becslés. Tekintsük ezek után a

$$\beta = \frac{1}{2}\frac{n+1}{n+2}\overline{\alpha}$$

statisztikát és igazoljk, hogy ez torzítatlan becslése α*-nak. Igazoljuk továbbá, hogy*

$$\sigma^2\left(\beta\right) = \frac{1}{4n^2}.$$

Ezután készítsük el a

$$\gamma = \frac{n+1}{5n+4}\left(\min_{1\leq i\leq n} X_i + 2\max_{1\leq i\leq n} X_i\right)$$

statisztikát és lássuk be, hogy ez hatásosabb mint β*. Igaz ugyanis, hogy*

$$\sigma^2\left(\gamma\right) \simeq \frac{1}{5n^2}.$$

Hogyan lehetséges ez? A válasz összefügg az elégségesség fogalmával. Szemléletesen is érthetõ, hogy γ*több információt hordoz* α*- ról mint* β*. Sõt az is igaz, hogy* γ *elégséges statisztika* α*-ra nézve, ezért aztán nem meglepõ, hogy az kisebb hibával közelíti. Ezzel elárultuk a választ, de természetesen az olvasónak még maratd feladata, igazolni kell, hogy a szóban forgó becslések torzítatlanok, fennállnak a szórásokra vonatkozó állítások, és érdemes megpróbálkozni a* γ *elégségess égének igazolásával is.*

Gyakorlat 27 *A következõ szintén sokaságméretre vonatkozó példa igazán történelmi. A II. Világháború alatt az angol katonai hírszerzés meg kívánta becsülni a német ipar tankgyártási kapacitását. A hagyományos hírszerzési módszerek alapján a havi termelésre vonatkozóan 1550 darabos becslést adtak. Statisztikusok a következõ módszert javasolták. Írják össze a kilõtt tankok gyártási sorozatszámait. Ennek alapán adnak majd becslést. Mint kiderült a precíz németek, minden hónapban más betüjellel kezdõdõ sorozatszámmal látták el a legyártott tankokat. Az 1941 juniusában gyártott tankok közül a kilõtt tankok közül legnagyobb sorozatszám a 244 volt. Adjunk maximum likelihood becslést ha feltesszük, hogy kilõtt tankok sorszáma egyenletesen oszlik el az összes sorszám között. Használjuk az elõzõ gyakorlat becsléseit. (A legkisebb leolvasott sorozatszám 31 volt).*

Megjegyzés 6 *Itt nem részletezendõ szintén a populáció méretére vonatkozó becsléssel állapították meg angol statisztikusok (Bradley és, Efron [**?**]), hogy Shakespeare aktív szókincse 31534 szóból állt, paszívan további 35.000-t ismert. Érdemes ezt összevetni azzal, hogy az átlagember 2000 szót használ akítvan, 3-5000 szó a passzív szókincse, míg az igen választékos irodalmi közlés hozzávetõleg 8000-10.000 szó aktív ismeretét tételezi fel.*

Megjegyzés 7 *Az átlag és a módusz (illetve a legvalószínûbb osztály) használata némi óvatosságot igéényel. Nem igaz például, hogy az átlagos a leggyakoribb. Ez egyszerûen azért igaz, mert ferde eloszlások esetén az átlag és a módusz nem esik egybe.* ***[ábra!]*** *Például egy társadalomban sok szegény ebre él és igen kevés gazdag. Átlagos vagyoni helyzetú igen kevés van.*

J. Reynolds vélte úgy, hogy az átlagot látjuk „szépnek" Nem világos, hogy e kijelentést hogyan kell értelmezni. Legyen ugyanis h az átlagmagasság, w az átlagsúly. Ha valaki arányos testfelépítésû akkor X magassághoz

$$Y = c_w X^3 \tag{5.4}$$

súly tartozik valamilyen c_w arányossági tényezõvel. Mi következik akkor az átlagokra?

$$w = \mathbb{E}(Y)$$

ugyanakkor

$$h = \mathbb{E}(X).$$

Véve (5.4) *mindkét oldalán a várható értéket*

$$w = c_w E\left(X^3\right).$$

Viszont nyilvánvaló, hogy általában $E(X^3) \neq E(X)^3 = h^3$, azaz az átlag magassághoz nem átlagos súly tartozik, vagyis, Átlag Polgár, akinek h magasságot tulajdonítunk és w súlyt nem lesz arányos, nem lesz "szp".

Chapter 6
HIPOTÉZIS VIZSGÁLAT

Ebben a fejezetben igen hasznos, jól alkalmazható módszereket ismertetünk amelyek egy-egy üzleti döntéshez adhatnak megbízható támpontot. Mintapéldánk lehet egy új termék, szolgáltatás bevezetése. Egy új szappanopera sikeréhez mondjuk az a minimum feltétel, hogy az egyidoben sugárzott mûsorokkal szemben szerezzen 25% nézettséget, azaz egy-két hét után a piaci részesedése legyen 25%. A sorozat sorsa felõl megint minta alapján döntünk. Zártkörû vetítést tartunk nézõk egy csoportjának, akik mint otthon szabadon választhatnak a csatornák között. (Esetleg nem is tudják a vizsgálat valódi tárgyát, mondjuk chips-et és üdítõt kapnak, választhatnak a TV nézéshez mit fogyasztanak és ezekrõl kérjük a véleményüket.)

6.1 Intervallum becslés

A p piaci részarány megállapítása nem jelent mást mint annak a $p = \mathbb{P}(A)$ valószínûségnek a becslése, hogy egy véletlenül kiválasztott nézõ sorozatot választja-e. Mint má láttuk, a $\overline{p}$ relatív gyakoriság remek pontbecslése p-nek. Ugyanakkor mint az az átlagmagasság kapcsán megjegyeztük, sem egyetlen ember esetében, sem egy $n > 1$ elemû minta esetében sem fog pontosan fennállni, hogy $p = \overline{p}_n$. Mit mondhatunk akkor e helyett?

1. p és $\overline{p}_n$ közel van egymáshoz. Mennyire?
2. p és $\overline{p}_n$ távolsága kisebb mint δ. Ez biztos, vagy csak valamilyen valószínûséggel igaz;

A centrális határeloszlás tétel segítségéével pontos válasz adható ezekre a kérdésekre.

Vegyünk egy tipikus példát. Egy gyártmány számos paraméterrel rendelkezik, hogy csak a llegegyszerûbb esetet vegyük egy üdítõs palackra az van rá írva, hogy 1 litert tartalmaz. A vevõ ezt el is várja. De valóban annyi van az üvegekben? Vegyünk mintát az üvegek közül és mérjuk meg a tartalmukat. Legyen az n elemû minta alapán számolt tapasztalati átlag

$$\overline{X} = \overline{X}_n.$$

A centrális határeloszlás tétel alapján $\overline{X}$ igen jó közelítéssel normális eloszlású, ha a vizsgált X valószínüségi változónak létezik várható értéke és szórása, a mintaelemek függetlenek és $n > 30$. Még pontosabban várható értéke μ lesz és szórása $\frac{\sigma}{\sqrt{n}}$, azaz

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right),$$

Megjegyzés 8 *Vegyük észre, hogy a szórás σ-rõl $\frac{\sigma}{\sqrt{n}}-$re változott. Mint mindjárt látni fogjuk, ez biztosítja azt, hogy a becslésünk egyre jobb lesz a mintaelemszám növekedtével.*

Ebbõl következik, hogy

$$\mathbb{P}\left(\overline{X} < x\right) = \Phi\left(\frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

vagy másképpen

$$\mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < x\right) = \Phi\left(x\right).$$

A normális eloszlás szimmetriája miatt, akkor az is igaz, hogy

$$\mathbb{P}\left(\left|\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right| < x\right) = 2\Phi\left(x\right) - 1. \tag{6.1}$$

A kényelem kedvéért szokás a jobb oldalon található valószínûséget valamilyen „nevezetes" közmegegyezéssel elfogadott értéknek választani, pl. 0.95, 0.975, 0.99. Azaz a standard normális eloszlás inverzét használjuk. Tekintsük elõször a

$$\mathbb{P}\left(X < x\right) = \Phi\left(x\right) = 1-\alpha$$

adott $\alpha-$hoz tartozó $x = z_\alpha$ értéket. (lásd ábra) Ez a z_α érték vág ki a standard normális eloszlás „farkából"α valószínûséget. Természetesen akkor

$$\mathbb{P}\left(-z_\alpha < X < z_\alpha\right) = 1-2\alpha.$$

Ha tehát a kívánalom mondjuk, hogy (6.1) jobb oldalán .95 valószínûség álljon, akkor a kivágott össz valószínûséget legyen $1-\alpha = .95$, akkor a z kritiuks értékek az $\alpha/2$ helyekhez tartozóak. Tehát a standard normális eloszlás esetén

$$\mathbb{P}\left(-z_{\alpha/2} < X < z_{\alpha/2}\right) = 1-\alpha.$$

amibõl feladatunkban

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) = 1-\alpha,$$

ezt átrendezve

$$\mathbb{P}\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

illetve

$$\mathbb{P}\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

A kapott összefüggés azt jelenti, hogy az ismeretlen μ sokaság átlag, az

$$\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

intervallumba esik $1-\alpha$ valószínûséggel. Ezzel megszerkesztettük a μ-re vonatkozó konfidencia intervallumot.

Definíció 52 *Másképp azt szokták mondani, hogy az* $\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\sigma\frac{\sigma}{\sqrt{n}}\right)$ *intervallum* $1-\alpha$ *megbízhatósággal tartalmazza* $\mu - t$*, vagyis ez a* μ*-re vonatkozó* $1-\alpha$ *szintû* ***megbízhatósági (vagy konfidencia) intervallum****. Ezzel úgynevezett* ***intervallum becslést*** *adtunk* μ*-re .Szokás* $\alpha - t$ *a szignifikancia szintjének nevezni.*

Általában kétoldali konfidencia intervallumokat szokás használni, de speciális esetekben (pl. az üvegbe eleve nem fér 1 liternél több) lehet egyoldalú konfidencia intervallumot is szerkeszteni, ilyenkor persze α nem felezõdik meg:

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu\right) = 1 - \alpha$$

vagy

$$\mathbb{P}\left(\mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

6.1.1 $t-$ eloszlásra épített konfidencia intervallum

Az elõzõ szakaszban láttuk, hogy amennyiben a tapasztalati átlag normális illetve közel normális eloszlású, akkor ennek a tudásunknak a birtokában konfidencia intervallumot lehet szerkeszteni az ismeretlen sokaság, azaz populáció átlagra. Abban az esetben, amikor a vizsgály valószínüségi változó X nem normális eloszlású, továbbá a populáció σ szórása ismeretlen, akkor viszonylag kis minták esetén a tapasztalati szórás segítségével normált statisztika

$$\frac{\overline{X} - \mu}{s}$$

nem nem standard normális eloszlású.

Állítás 8 *Ha* $X_1, X_2 ... X_n$ *független azonos eloszlású valószínüségi változók, akkor a mintaátlag alábbi standardizáltja*

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t^{(n-1)}$$

$n-1$ *szabadságfokú Student, avagy t-eloszlás követ. A jelölés némi keverésével jelölje* $t^{(n-1)}(x)$ *ennek az eloszlásnak a sûrûségfüggvényét.*

A $t-$ eloszlás értékeit célszerû táblázatból kikeresni vagy számítógép segítségével meghatározni. A 8 Állítás segítségével a megint lehetõség van konfidencia intervallumot szerkeszteni.

Állítás 9 *Legyen* $t_{\alpha/2}$ *az a kritikus érték, amely a* t*-eloszlás jobboldali „farkából"* $\alpha/2$ *területet vág ki, azaz, ha* Y $t-$*eloszlású, akkor*

$$\mathbb{P}\left(Y > t_{\alpha/2}\right) = \frac{\alpha}{2}$$

ekkor az eloszlás szimmetriája miatt

$$\mathbb{P}\left(t_{\alpha/2} < Y < t_{\alpha/2}\right) = 1 - \alpha. \tag{6.2}$$

Következmény 5

$$\mathbb{P}\left(t_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} < t_{\alpha/2}\right) = 1-\alpha$$

és

$$\mathbb{P}\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha.$$

Az álltís nyilván következik a 8 Állításból és (6.2)-bõl.

Gyakorlat 28 *Készítsünk „döntési diagrammot mikor kell normális és mikor kell t eloszláson alapuló intervallumot készíteni.*

Állítás 10 *Részarány konfidencia intervalluma mindíg normális eloszlásra épül, azaz*

$$\mathbb{P}\left(\overline{p} - z_{\alpha/2}\sqrt{\frac{\overline{p}\,(1-\overline{p})}{n}} < p < \overline{p} + z_{\alpha/2}\sqrt{\frac{\overline{p}\,(1-\overline{p})}{n}}\right) = 1-\alpha. \tag{6.3}$$

Bizonyítás. *A centrális határeloszlás tételbõl tudjuk, hogy*

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

ezért a részarányra vonatkozóan is igaz, hogy

$$\frac{\overline{p}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1). \tag{6.4}$$

Az (6.3) összefüggés ezután abból következik, hogy a binomiális eloszlás szórása $\sigma = \sqrt{p\,(1-p)}$, ezért $\overline{p}$ szórása $\sqrt{\frac{p(1-p)}{n}}$, amibe ha $\overline{p}$ helyettesítünk az elkövetett hiba „másodrendben" $kicsis(6.$ továbbra is igaz marad. ■

Gyakorlat 29 *Készítsünk „döntési diagrammot"$mikorkellnormlissmikorkellteloszlsonalapulintervall$*

6.1.2 Konfidencia intervallum az ismeretlen szórásra

Az elõzõ rész sémája a következõ módon foglalható össze. Tegyók fel, hogy Y valószínüségi változó, statisztika szolgál egy η paraméter becslésére. Ha tudjuk, hogy $Y-\eta$ vagy $\frac{Y}{\eta}$ milyen eloszlást követ, akkor ebbõl konfidencia intervallum szerkeszthetõ η-ra.

Definíció 53 *A χ^2 eloszlást iplicit módon definiáljuk. Legyen*

$$Y = \sum_{i=1}^{n} X_i^2$$

ahol $X_1, X_2 ... X_n$ független standard normális eloszlás valószínüségi változók. Ekkor azt monjuk, hogy Y eloszlása $(n-1)$ szabadságfokú χ^2 (olvasd khi-négyzet) eloszlás.

Állítás 11 *Ha egy normális populációból vettünk $X_1, X_2...X_n$ független mintát, akkor*

$$\frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2.$$

Bizonyítás. *Az állítás közvetlenkövetkezménye a* (53) *Definíciónak.* ■

Állítás 12 *Az ismeretlen σ^2 szórásnégyzet besclésére normális eloszlású populáció esetén a*

$$\mathbb{P}\left(\frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)\,s^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha$$

intervallum szerkeszthetõ. Azaz σ^2-t $1-\alpha$ valószínûséggel tartalmazza a $\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right)$ intervallum.

6.1.3 A mintaméret megválasztása

A gyakorlati életben gyakran az a feladat, hogy a populáció átlagot bizonyos hibahatáron belül tartsuk, Mint az üvegtöltés példájában a térfogat $1 \pm 0,01$ liter kell, hogy legyen. Ehhez elõszor a populációátlagot kell a fenti módon jól becsülnünk, ezután lehet az egyes üvegek töltésére következtetni az átalg becsése melett a szórás becslését is felhasználva. Foglalkozzunk a legegyszerûbb konfidencia intervallummal. A

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

egyenõtlenség $1-\alpha$ valószínûséggel fenáll. Szeretnénk elérni, hogy μ -re $\overline{X} \pm B$ intervallumot kapjunk ugyanezen $1-\alpha$ megbízhatósággal.adott, rögzített B melett. Világos, hogy ehhez a

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = B$$

összefüggésnek kell teljesülnie. Ha α rögzített, ez meghatározza $z_{\alpha/2}$-t, ezért az n mintaméret az egyetlen választható paraméter. A legkisebb megfelelõ n

$$n = \left\lceil \frac{z^2_{\alpha/2}\sigma^2}{B^2} \right\rceil.$$

Gyakorlat 30 *Adjuk meg az analóg mintaméretre vonatkozó összefüggéseket $t-$eloszlás esetére és részaránybecslés esetére is.*

6.2 Hipotézis vizsgálat

A mindennapi életben, döntési szituációban gyakran nem az érdekel minket, hogy konfidencia intervallumot szerkesszünk, inkább az a kérdés bent van-e az ismeretlen átlag egy adott intervallumban vagy sem. Például a palackok töltési átlaga 1 liter. Ha igen, minden rendben, de ha nem akkor állítani kell a töltõberendezésen, le kell állítani a gépsort. Ez persze termelés, profit kiesést jelent, ezért ha lehet, csak akkor döntünk így, ha eléggé biztosak vagyunk abban, hogy a töltési átlag lényegesen eltér az elõírttól. A mintavétel,

a mérés és statisztikai vizsgálat alapján egy dichotóm döntést hozunk, egy liter az töltési átlag, vagy nem az. Ennek a példának a végigvitelével mutatjuk be a hipotési vizsgálat tipikus menetét.

A döntési folyamat elsõ lépése a munkahipotézis felállítása. Ez az egyetlen „kényes" lépés, a többi rutin feladat.

Legyen $\mu_0 = 1$. Elvárásunk, hogy a populáció várható értéke μ ezzel egybe essék. Ezt fogalmazzuk meg mint **null hipotézist**.

$$H_0 : \mu = \mu_0$$

Ennek logikai ellenéte az **alternatív hipotézis**

$$H_1 : \mu \neq \mu_0.$$

Mielõtt mintát veszünk, számításokat végzünk el kell, döntenünk, hogy mennyire „fontos", hogy jól döntsünk.A döntést minta alapján fogjuk meghozni. A minta véletlentõl függõ, ezért több módon is felléphet hiba. Ezeket mutatja az alábbi táblázat.

		a valóság	
		H_0 igaz	H_0 hamis
hogyan döntünk	elfogadjuk H_0-t	jó a döntés	másodfajú hibát vétünk
	elvetjük H_0-t	elsõfajú hibát vétünk	jó a döntés

Mielõtt a hiba valószínûségének kiszámításához látunk, két jelölést vezetünk be.

$$\alpha = \mathbb{P}_{H_0} \left(\text{elvetjük } H_0\text{-t}\right)$$

ahol a $\mathbb{P}_{H_0}$ azon valószínûségi mérték amely H_0 fennállása esetén jellemzi a sokaságot.

$$\beta = \mathbb{P}_{H_1} \left(\text{elfogadjuk } H_0\text{-t}\right)$$

itt pedig $\mathbb{P}_{H_1}$ azon valószínûségi mérték amely H_1 fennállása esetén jellemzi a sokaságot. Ez utóbbi nem feltétlenül jól definiált, erre késõbb visszatérünk.

Ha a feltételek biztosítják, hogy a mintaátlag jó közelítéssel normális akkor ez igaz a

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

úgynevezett **próbafüggvényre** is, pontosabban z közel standard normális eloszlású lesz. Ebbõl következik, hogy

$$\mathbb{P}\left(|z| < z_{\alpha/2}\right) = 1 - a$$

azaz

$$\mathbb{P}\left(\left|\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| < z_{\alpha/2}\right) = 1 - a$$

illetve annak a valószínûsége, hogy a próbafüggvény abszolút értéke nagyobb legyen mint $z_{\alpha/2}$ egyenlõ α. Megint azt az okoskodást, használjuk, hogy egy kísérletben a nagy valószínûségû esemény bekövetkeztét tételezzük fel. Ha tehát feltesszük, hogy

$$\mu = \mu_0$$

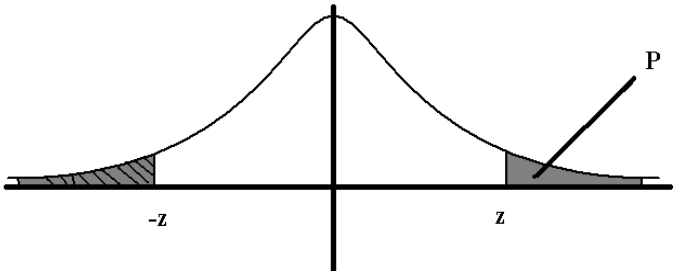


Figure 1 p az eloszlás farkában

akkor

$$\mathbb{P}\left(\frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} < -z_{\alpha/2} \text{ vagy } \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{\alpha/2}\right) = a.$$

Az (**??**) ábrára tekintve ez azt jelenti, hogy a z próbafüggvény a kis, $\alpha/2-\alpha/2$ valószínüségû alsó vagy felsõ farokba mutat. Ez nincs összhangban azzal a feltevésünkkel, hogy az $1-\alpha >> \alpha$ valószínûségû esemény következik be, ezért a $\mu = \mu_0$ feltevést vetjük el. Természetesen valamilyen más μ lehet igaz, aminek megfelelõen netrálva $\overline{X} - t$ a próbafüggvény már a $\left(-z_{\alpha/2}, z_{\alpha/2}\right)$ intervallumba mutat, (lásd (1)Ábrát.).

Definíció 54 *A $\left(-z_{\alpha/2}, z_{\alpha/2}\right)$ intervallumot szokás **elfogadási tartománynak** nevezni, a $\left(-\infty, -z_{\alpha/2},\right) \cup \left(z_{\alpha/2}, \infty\right)$ −t pedig **elutasítási tartománynak**. A **próba szintje**, vagy **szignifikancia szintje** α, a **kritikus érték**(ek) α illetve $\alpha/2$.*

Azt mondjuk, hogy α szinten elutasítjuk H_0-t ha z az elutasítási tartományba esik, illetve, hogy nem utasítjuk el, ha az ellenkezõje igaz. Kicsit pongyolán lehet azt is mondani, hogy elfogadjuk H_0-t, de mint látni fogjuk ez nem igazán szerencsés, kissé félrevezetõ.

Megjegyzés 9 *Vegyük észre, hogy α-t tetszés szerint megválaszthatjuk a döntési eljárás elején. Ebbõl pedig az is következik, hogy az elsõfajú hiba valószínûségét meg tudjuk szabni, azaz annak a valószínûségét, hogy elutasítjuk H_0-t pedig igaz. Hiszen ez a szituáció akkor áll elõ, amikor $\mu = \mu_0$ és a próbafüggvény mégis a $\left(-\infty, -z_{\alpha/2},\right) \cup \left(z_{\alpha/2}, \infty\right)$ elutasítási tartományba esik, aminek a valószínûsége éppen α.*

6.2.1 A hipotézis vizsgálat menete

A hipotézis vizsgálat lépései a következõkben foglalhatóak össze.

1. A null hipotézis H_0 megfogalmazása.
2. A próbafüggvény kiválasztása a feltételek alapján.
3. A α szignifikancia szint meghatározása a probléma természetétõl függõen.
4. A próbafüggvény eloszlásának alapján az elutasítási tartomány meghatározása.
5. Mitavételezés, a próbafüggvény kiszámítása.

6. Döntés attól függõen, hogy a próbafüggvény hova esik.

Mint említettük az egyetlen kényes feladat a nullhipotézis felállítása (és talán mint minden statisztikai feladatnál a jó mintavételezés). A gyakorlati életben a $\mu = \mu_0$ null hipotézis melett igen gyakori a $\mu \geq \mu_0$ illetve a $\mu \leq \mu_0$ alakú feltevés. Mint a palacktöltõ sor mûködtetõje a $\mu = \mu_0$ feltevés volt fontos számunkra. A vevõ szempontja más. Õt az érdekli, hogy elég sokat vagy elég jót kapjon a pénzéért. A hirdetések is igen gyakran azzal kínálnak egy portékát, hogy az azt jellemzõ valamilyen mennyiség kisebb vagy nagyobb mint egy adott éreték. Erre lehet példa a gyümölcslé, üdítõ. Nagy betûkkel a csomagolás oldalán hirdeti, hogy több mint 50% gyümölcstartalommal, vagy „kevesebb mint 1% cukor tartalom". A fogyasztó vagy a konkurencia esetleg gyanút fog, hogy a gyártó nem teljesíti azt, amit a hirdetésben állít. kritikus esetben ez pert vonhat maga után, a bíróságnak kell döntenie a felek között. Vegyük ebbõl a nézõponból szemügyre a problámát.

A gyártó azt állítja, hogy

$$\mu \geq \mu_0$$

ahol mondjuk $\mu_0 = 50\%$. A konkurencia szerint ez nem igaz. Az ártatlanság vélelmének jogelvét követve azt mondjuk, hogy *a gyártó ártatlan mindaddig amíg bûnössége minden kétséget kizáróan bebizonyosodik.* Abban az esetben amikor nem lehet minden egyes terméket, vitás objektumot megvizsgálni, természetesen a bíróság is csak statisztikai módszerek alapján dönthet. A minta véletlen természetébõl fakadóan a döntés sem lehet 100% bizonyosságú, ezért az ártatlanság vélelme azt kívánja, hogy jól kontrolálni tudjuk annak a valószínûségét, hogy bár a gyrátó ártatlan, a szerencsétlen véletlen mintaválasztás azt eredményezi, hogy statisztikai vizsgálat bûnösnek találja a gyártót. Az elõbb vázolt hipotézis vizsgálat keretei között ez azt jelenti, hogy az α elsõfajú hiba valószínûségét tudjuk kontrolálni, a vitázó feleknek meg kell állapodnia ebben az $\alpha-$ban, ezt közlik a bírósággal (természetesen ez a minta nagyságát az eljárás költségét is befolyásolja). A null hipotézist ennek a logikának az értelmében úgy kell fölállítani, hogy az elsõ fajú hiba a leírt legyen, azaz a null hipotési a gyártó ártatlanságát, a hirdetésben közölt állítást kell, hogy tartalmazza, azaz

$$H_0 : \mu \geq \mu_0$$

a helyesen választott null hipotézis. Természetesen nem minden vizsgálandó probléma fordítható le a jog nyelvére, általában célszerû a nullhipotézis irányát úgy megválasztani, hogy az elkövethetõ hibák közul az legyen az elsõfajú aminek bekövetkezése a nagyobb kárt okozza, mivel ennek a valószínûségét tudjuk α-val szabályozni, míg a másodfaúét nem.

Vegyünk erre még egy példát. Fémlemezeket vonunk be rozsdamentesítõ festékkel. Ha túl sok festéket viszünk fel, akkor fölösleges költséget hozunk létre, ha viszont nem elég vastag a védõréteg, a lemezt hamar kikezdi a korrózió, tönkremegy a lemez, esetleg ennek következtében a lemez értékénél is nagyobb kár keletkezik. Ezért tehát a minõségellenõrzés során a μ átlagos festékvastagságra vonatkozóan célszerû a

$$H_0 : \mu \leq \mu_0$$

nullhipotésit alkalmazni, hiszen akkor lesz igaz az hogy a két lehetséges hiba közul a nagyobb kárt okozó lesz az elsõfajú hiba. Konkrétan azt a valószínûséget kontrolálja α,hogy a

festékréteg vékonyabb mint szükséges, de ezt a próba nem mutatta ki és ezért elvetjük a null hipotézist.

6.2.2 Paraméteres próbák

A hipotézis vizsgálat 5.6. lépése az úgynevezett próba. A körülményektõl függõen különbözõ próbák ismeretesek, azoknak pedig további variációi. Az alábbiakban a legfontosabbakat ismertetjük. Elõször egy sokaság valamely paraméterére vonatkozó próbákat mutatjuk be, ezek az egymintás próbák, majd két sokaság valamilyen paraméterének az összehasonlítására vonatkozó kétmintás próbák következnek.

A kétoldali z próba (néha nevezik u próbának is) került a fejezet elején ismertetésre.

1. Ekkor a null hipotézis
$$H_0 : \mu = \mu_0$$
az alternatív hipotézis
$$H_1 : \mu \neq \mu_0.$$

2. Ha ismert a sokaság szórása σ vagy a minta elemszáma elég nagy ($n > 30$), akkor próbafüggvény
$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
ami közel standard normális eloszlású.

3. A hétköznapi, üzleti életben általában a $\alpha = 0.05$ megfelelõ választás.

4. Az elutasítási tartomány $\left(-\infty, -z_{\alpha/2},\right) \cup \left(z_{\alpha/2}, \infty\right)$,ahol $z_{\alpha/2} = 1.96$

5. Legyen a példa kedvéért $\overline{X} = 1.002\%$ liter a mintában a palackokba töltött folyadék átlagos mennyisége. Legyen $\sigma = 0.01\%$ és $n = 100$. Ekkor
$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{0.002}{0.01} 10 = 2.$$

6. Viszont $z_{\alpha/2} = 1.96 < z = 2$,ezÉrt a null hipotézist el kell vetnünk. Le kell állítani a gépsort és újraszabályozni a töltést.

6.2.3 Az egymintás próbák további esetei

Ebben a szakaszban a hipotézis vizsgálat variánsait mutatjuk be amelyet a feltételek változása hoz létre.

Abban az esetben, ha a populáció szórása ismeretlen azt a tapasztalati szórással helyettesítkük amikor a próbafüggvényt kiszámoljuik, azaz a mintaátlagot standardizáljuk:
$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Ilyen esetben a standardizált kifejezés normális eloszlást követ, ha a populáció maga is normális eloszlású volt vagy ha $n > 30$.

Állítás 13 *Ellenkezõ esetben azaz ha a populáció nem normális, illetve ismeretlen szórást követ, létezik a várható érték és szórás (de ismeretlenek) akkor a*

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

próbafüggvénny Student eloszlású és ennek megfelelõen alakul a 3. lépés, illetve az 5. lépésben a kritikus értékeket $t_{\alpha/2}$ illetve egyoldali hipot;zis eset;ben t_α adja.

Állítás 14 *Ha egy p valószínúség, illetve részarány becslése a feladat akkor mindíg a*

$$z = \frac{\overline{p} - \mu}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}}$$

próbafüggvény alkalmazható, ami standard normális eloszlású.

6.2.4 Kétmintás próbák

Gyakori és természetes kérdések merülnek fel, két populácoó összehasonlítása során. Igaze, hogy a villányi szõlõ cukorfoka magasabb mint a sporonié? Az ilyen típus kérdések vizsgálatát is két példán mutatjuk elõször be, majd tömören ismertetjük a technikai részleteket.

Legyen μ_1 a Villányi szõlõ átlagos cukorfoka, μ_2 a Sopronié. vegyünk $X_1, X_2 ... X_n$ független azonos eloszlású valószínüségi változókat, mintát a Villányon termett mustokbó és $Y_1, Y_2 ... Y_m$ a Soproni.

Az X_i és Y_j mintaelemekrõl feltesszük, hogy teljesen függetlenek.

Tegyük fel, hogy $n, m > 30$, azaz a minták „nagyok". Mi legyen a nullhipotézis? Min korábban hangsúlyoztuk a null hipotésis megválasztása függ attól, hogyhibás döntés esetén milyen kár keletkezik. Most egy olyan szituációt vázolunk, amiben a korábbival éppen ellentétes módon célcerû a null hipotézistválasztani. Ha egy hirdetés állítaná, hogy a villányi szõlõ cukorfoka magasabb mint a sporonié akkor a független döntést a $H_0 : \mu_1 \geq \mu_2$ hipotézisbõl kiindulva kell hozni. Ha viszont az a helyzet, hogy a pincészetünknek régóta sporoni mustot veszünk, akkor természetesen mindaddig kitartunk e melett míg ennek az ellenkezõje alaposan be nem igazolódik, azaz a statisztika nyelvén, ki nem derül, hogy a villányi cukorfoka **szignifikénsan** magasabb. Ekkor tehát a

$$H_0 : \mu_1 \leq \mu_2$$

null hipotézisbõl indulunk ki. Érdemes a null hipotézist kissé átalakítani,

$$H_0 : \mu_1 - \mu_2 \leq D_0 = 0.$$

Esetünkben az állítás annyiról szól, hogy az egyik átlag nagyobb mint a másik, ilyenkor $(\mu_1 - \mu_2)_0 = D_0 = 0$,de lehetne a kezdeti állítás pl az, hogy a villányi szõlõ átlagos cukorfoka 2 fokkal magasabb mint a sporoni, ekkor $D_0 = 2$ lenne.

Világos, hogy ezt akkor fogjuk elvetni, ha

$$z = \frac{\left(\overline{X} - \overline{Y}\right) - D_0}{s_{\mu_1 - \mu_2}} > z_\alpha \tag{6.5}$$

ahol

$$s_{\mu_1-\mu_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

A (6.5) egyenõtlenséget átírva

$$\mu_1 > \mu_2 + z_a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

összefüggést kapjuk, amibõl azonnal világos, hogy mit értünk szignifikánsan nagyobb alatt. μ_1 nem egyszerûen nagyobb kell, hogy legyen mint μ_2 hanem $\mu_2 + z_a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$-nél is azaz a véletlen szóródás $z_\alpha-$szorosánál is nagyobb az eltérés. A proba kialakitásakor fontos, hogy α a szignifikancia szintje, $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ pedig a mintaátlagok különbségének szórása. A proba további menete már megegyezik az egymintás próbáéval.

Abban az esetben, ha a hipotézis kétoldali, azaz a null hipotézis

$$H_0 : \mu_1 = \mu_2$$

illetve

$$H_0 : \mu_1 - \mu_2 = D_0 = 0$$

akkor is (6.5) a próbafüggvény, de a kritikus tartomány kétoldali, azaz $-z_{\alpha/2}, z_{\alpha/2}$ a két kritikus érték.

Kis mintás próba két átlag különbségére vonatkozóan

A fenti példa kis minták esetén a

$$t = \frac{\left(\overline{X} - \overline{Y}\right) - D_0}{s_{\mu_1-\mu_2}} \tag{6.6}$$

Student eloszlású próbafüggvény segítségével oldható meg. A **szabadságfokot** a

$$df = \frac{\frac{s_1}{n_1} + \frac{s_2}{n_2}}{\frac{\left(\frac{s_1}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s2}{n_2}\right)^2}{n_2-1}}$$

képlet határozza meg.

Kombinált szórásbecslés

Kis minták esetében javítható a szórásra vonatkozó becslés, ha egyéb körülmények alapján tudható, hogy a két populáció szórása ugyanaz, azaz $\sigma_1 = \sigma_2$. Ekkor s_1, s_2 ugyanazt a közös σ értéket becsli, ezért célszerû a kevert mintaszórással számolni, ennek négyzete:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}.$$

Ekkor próbafüggvényünk a

$$t = \frac{\left(\overline{X} - \overline{Y}\right) - D_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

lesz. E két képletet, azaz a kevert szórás becslést érdemes nagy minta esetén is használni, ha $\sigma_1 = \sigma_2$ feltehetõ.

Részarányok összehasonlítása

Igaz-e hogy a diplomások körében az A párt népszerûbb mint az alacsonyabb végzettségûek körében? Erre a kétmintás részarányokra vonatkozó próba adhat választ. Legye a null hipotézis

$$H_0 : p_1 - p_2 \geq D_0 = 0$$

ahol p_1 a diplomások között az A-t preferálók részaránya, p_2 a másik csoportban. Nyilván egyoldali hipotézisrõl van szó, a próbafüggvény pedig

$$z = \frac{(p_1 - p_2) - D_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}.$$

Kétoldali

$$H_0 : p_1 - p_2 = 0$$

hipotozis esetén lehet a kozös p becslését, $\widehat{p}$-t használni, ezért a próbafüggvény

$$z = \frac{p_1 - p_2 - 0}{\sqrt{\widehat{p}\,(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

ahol

$$\widehat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.$$

Próba párosított mintával

Az összes eddigi esetben feltettük, hogy a két minta elemei teljesen függetlenek. Bizonyos esetekben ez bár nem áll fenn, mégis igen jó teszt készíthetõ. Tegyük fel, hogy egy készülõ felhasználói szoftverhez két kezelõfelület terv készült. El kívánjuk dönteni, hogy melyik tevnek kedvezõbbek az ergonómiai tulajdonságai, melyiken hatékonyabb a munkavégzés. Ezért a következõ „kísérletet" végezzük. 50 rutinos opertátort teljesen azonos feltételek közöt megtanítunk minkét kezelõfelület használatára. Ezek után ugyanazt a feladatot elvégzik az egyik illetve a másik felületen, a szükséges idõt mérjük. Legyen X_i az 1. Y_i a 2. kezelõfeluleten mért idõ $i = 1, ..., 50$. Az esetlegesfáradási tényezõt is kiküszöböljük, 25-en elõszõr az 1. kezelõfelületet használják utána a 2-at, a másik huszonöt operátor fordítva.

Számítsuk ki a mért idõk d_i különbségeit

$$d_i = X_i - Y_i.$$

Ha nincs egyéb elõfeltevésünk, akkor a

$$H_0 : \mu_1 = \mu_2$$

null hipotézist használjuk. Ez természetesen ekvivalens a

$$H_0 : \mu_1 - \mu_2 = 0$$

illetve ha $\mu_d = E(X - Y) = \mu_1 - \mu_2$ jelöli a várható értékek különbségét, akkor ez szintén ekvivalens a

$$H_0 : \mu_d = 0$$

hipotézissel. Ezután a d_i mintadifferenciákra mint egymintás próba járhatunk el. A módszer kis és nagymintás variánsai, részarányra vonatkozó variánsa egyaránt egyszerûen értelmezhetõ.

6.3 Próbák a szórásra vonatkozóan

6.3.1 Egymintás próba

Eddig kizárólag az átlag illetve a részarányra vonatkozó próbákról volt szó. De ahogy a szórásnégyzetre lehet konfidencia intervallumot szerkeszteni, természetesen hipotézis vizsgálat is végezhetõ. Legyen egy normális eloszlású sokaságunk, aminek szórása ismeretlen. Korábbi példánkban, a palacktöltõsor esetén vizsgálani érdemes a szóródást. Nem kedvezõ, ha túl nagy (ha kicsi általában nem baj.) Ha egy kihaló félben lévõ faj genetikus változatosságáról van szó, akkor a nagy szórás éppen kívánatos is lehet. Az elõbbi esetnél maradva legyen a kívánt szórás σ_0, ekkor a null hipotézis, a gyártó nézõpontjából

$$H_0 : \sigma^2 \leq \sigma_0^2$$

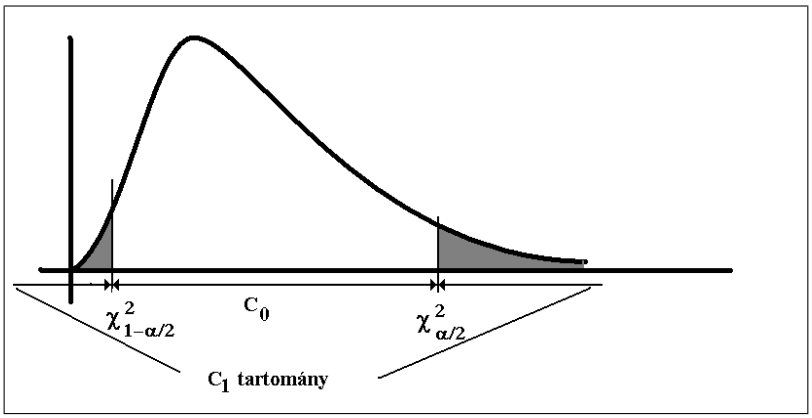
módon választandó. Az (11) Állításból tudjuk, hogy

$$\frac{(n-1)\, s^2}{\sigma^2} \sim \chi^2$$

ahol χ^2 szabads$'gfoka$n-1.$Ezrta\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$próbafüggvényt alkalmazva elutasítjuk a null hipotézist, ha

$$\chi^2 > \chi_\alpha^2$$

ahol a χ_α^2 az $n-1$ szabadságfokú χ^2 eloszlás $\alpha-$hoz tartozó kritikus értéke.



6.3.2 Kétmintás próba

Az átlagokra vonatkozó kétmintás próbák között már felmerült a kérdés, egyenlõ-e a két populáció szórása. Ha igen akkor a szórásra vonatkozóan igen jó becslés adható a kombonált szórással (Lásd 6.2.4 alfejezetet.) . Ehhez elõször ellenõriznóünk kell, hogy a két szórás valóban egyenlõ -e, azaz igaz-e a

$$H_0 : \sigma_1^2 = \sigma_2^2$$

A próba kivitelezése a következõ állításon alapszik.

Állítás 15 *Ha két normális populációból származó n illetve m elemû teljesn független mintát veszünk, akkor*

$$\frac{s_1^2}{s_2^2} \sim F^{(n-1,m-1)}$$

hányados $(n-1, m-1)$ szabadságfokú F eloszlást követ.

Az F eloszlást nem definiáljuk egzakt módon, lényegében két khinégyzet eloszlású független valószínüségi változó hányadosának eloszlását írja le.

Állítás 16

$$F^{(n-1,m-1)}(1-x) = \frac{1}{F^{(m-1,n-1)}(x)}.$$

Az állítást nem bizonyítjuk, de intuitíve látható a definicióból

Ennek alapján a próbafüggvény

$$F = \frac{s_1^2}{s_2^2}$$

F eloszlású ha igaz a null hipotézis. A kritikus érték $F_{\alpha/2}$ illetve $F_{1-\alpha/2}$,de ha az általánosság megszorítása nélkül, úgy számoztuk meg a mintákat, hogy $s_1 :> s_2$,akkor elég a felsõ kritikus tartománnyal foglalkozni. Abban az esetben, ha

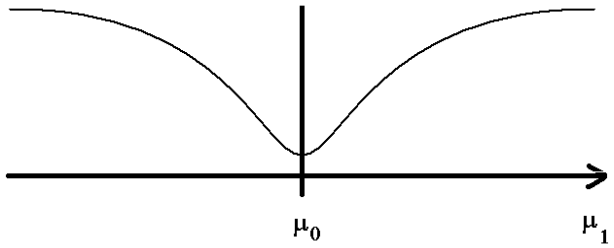
$$F > F_{\alpha/2}$$

elvetjük a null hipotézist.

6.3.3 A másodfajú hiba

Emlékezzünk vissza arra, hogy másodjajú hibát akkor vétünk a hipotézis vizsgálat során, ha a null hipotézis nem áll fenn, mégis a melett döntünk. Ezzen az eseménynek a valószínûsége (a véletlen mintavétel következtében) β. Az elsõfajú hibával ellentétben β általában nem számítható ki a valódi paraméter érték ismerete nélkül.

Definíció 55 *A **próba erejének** szokás nevezni* $1-\beta$*, vagyis azt a valószínûséget, hogy helyesen ismeri fel, hogy* H_0 *nem áll fenn. Szokás próba **erõfüggvényét** definiálni a*



A prba erőfggvnye a μ *vltozsval*

paraméter függvényében $f(\vartheta) = 1-\beta$.

Például egy jobboldali, várható értékre vonatkozó rpóba erõfüggvényét mutatja a (**??**)Ábra. **[Ide abrat]**

A másodfajú hiba kiszámítása

Képzeljük el a következõ szituációt. A hõ ellenállók selejtes hadianyagot gyártanak a munkaszolgálat során. Természetesen ügyelni kell, hogy a szabotázs ne derüljön ki. A gyártott fegyver például ûrmérete $\mu_0 = 9.00$ kell, hogy legyen. A gyártás során ennél lényegesen kisebb nem is lehet, viszont a furat készítésekor kis trükkel ennél nagyobbra lehet készíteni, ami selejtté teszi a fegyvert. A minõségellenõrzés a

$$H_0 : \mu \leq \mu_0$$

feltevést ellenõrzni. Mi a valószínûsége, hogy a szabotázs során gyártott $\mu = 9.07$ átlaggal nem buknak le az ellenállók, azaz H_0 nem igaz, de ezt a pr=ba nem mutatja ki. Tegy-k fel, hogy a szórás $\sigma = 0.02$, a szignifikancia szintje $\alpha = 0.05$ és a szokásos mintaméret $n = 30$? Készítsük el a próbafügvényt.

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Ekkor a próba alapján nem vetik el a null hipotézist, ha $z < z_\alpha$,azaz

$$\overline{x} < \mu_0 + z_a \frac{\sigma}{\sqrt{n}} =: C. \tag{6.7}$$

Ha tudjuk, hogy $\mu = 9.02$, akkor

$$\overline{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

alapján adódik, hogy a másodfajú hiba β valószínûsége (azaz, hogy elkerülik a lebukást)

$$\mathbb{P}\left(\overline{x} < C\right) = \mathbb{P}\left(\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{C-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi\left(\frac{C-\mu}{\frac{\sigma}{\sqrt{n}}}\right).$$

Érdemes C-be behelyettesíteni annak értékét (6.7)-bõl. Az átlag eredeti skáláján

$$\frac{C-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}} + z_\alpha$$

lesz a kritikus érték. Jól látható, hogy a kritkus tartomány nõ, ha μ nõ. Másképpen a $\frac{\sigma}{\sqrt{n}}$ szórású mintaátlag alapján jól el lehet különíteni $\mu > \mu_0$ átlagot, ha az a szórás többszörösével haladja me μ_0-t, egyébként nem.

Chapter 7
NEM PARAMÉTERES PRÓBÁK

7.0.4 Khinégyzet próbák

Ebben a fejezetben a **nem paraméteres statisztika** egy kis részét a **nem paraméteres próbák**at érintjük, ezen belül is elõször az úgynevezett χ^2 **próbákról** lesz szó. Közös vonásuk, mint nevük mutatja, hogy χ^2 eloszlásra vezetõ szellemes konstrukciókkal ragadnak meg igen összetett problémákat. A módszer mintapéldája a **multinomiális eloszlás** tesztelése.

Példa 31 *Képzeljük el, hogy a kóla termékek piaci részesedését követjük figyelemmel. Korábban azt tapasztaltuk, hogy a $p_1, p_2, ..p_k$ volt az $1, 2, ...k$ termék piaci részesedése az elõzõ idõszakban. Minket elsõ nekifutásra az érdekel, történt-e elmozdulás, vagy sem. Ezért null hipotézisként a*

$$H_0 : p_i' = p_i, i = 1, ...k$$

feltevéssel élünk, ahol p_i' az i-edik termék ismeretlen jelenlegi piaci részaránya. Legyen egy független n elmeû mintánk, amiben azt találtuk, hogy x_i fogyasztó választotta az i-edik terméket. Világos, hogy ekkor x_i binomiális eloszlású valószínüségi változó (n, p_i) paraméterekel)feltéve, hogy a null hipotézis fennáll. De akkor az is igaz, hogy igen jó közelítéssel ($x_i > 5$ feltevése már elegendõ) igaz, hogy

$$\frac{x_i - np_i}{\sqrt{np_i\left(1 - p_i\right)}} \sim N\left(0, 1\right).$$

Tekintettel arra, hogy az osztályok száma általában elég nagy $(1 - p_i) \approx 1$ ezt el szokták hagyni. Jelölje e_i a hipotézis alapján a várt értéket,

$$e_i = np_i$$

ezzel a jelöléssel

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(x_i - e_i\right)^2}{e_i}$$

próbafüggvény χ^2 eloszlást követ, ha H_0 igaz. Ezért a hipotézis ellenõrzése innen a szokásos módon történhet.

Homogenitás vizsgálat

Igen hasonló az úgynevezett **homogenitás vizsgálat**, amely két populáció rétegezettségének azonos voltát hivatott ellenõrizni. Adott A, B populációk és ezek azonos ismérv szerinti

felosztása. Az egyes osztályok részaránya legyenm p_i, q_i. Példaként állhat itt mondjuk a következõ A a városi lakosság, B a kistelepülési lakosság. az egyes osztályokat a szokásos legmagasabb iskolai végzettség szerint alakítjuk ki. A null hipotézis, hogy a legmagasabb végzettség szerint a városi és kistelepulési megoszlás azonos, azaz

$$H_0 : p_i = q_i, i = 1, ...k$$

Megint feltesszük, hogy $1 - p_i \approx 1, 1 - q_i \approx 1$. Legyen a két vizsgált minta elemszáma n, m a talált osztályok mérete pedig x_i,y_i. Tulajdon képpen az

$$\frac{x_i}{n} - \frac{y_i}{m} \approx 0$$

összefüggést ellenõrizzük. Ha igaz H_0,akkor

$$\frac{x_i}{n} \approx N\left(p_i, \sqrt{\frac{p_i}{n}}\right)$$

ezért $\frac{x_i}{n} - \frac{y_i}{m}$ szórása jó közelítéssel

$$\sqrt{p_i \frac{m+n}{mn}},$$

a p_i közösbecslése pedig

$$\frac{x_i + y_i}{n+m}$$

ezért

$$z = \frac{\frac{x_i}{n} - \frac{y_i}{m}}{\sqrt{\frac{x_i+y_i}{n+m}}} \sim N(0,1)$$

amibõl

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(\frac{x_i}{n} - \frac{y_i}{m}\right)^2}{\frac{x_i+y_i}{nm}}$$

χ^2 eloszlású próbafüggvény kapható.

Függetlenség vizsgálat

Igen gyakori kérdés, hogy két tulajdonság között van-e kapcsolat. Például a végzettség és az egyes napilapok kedveltsége között van-e összefüggés. Azaz egy sokaságot két módon is osztályba sorolunk, példánk szerint végzettség illetve a vásárolt napilap szerint. Tegyük fel, hogy m különbözõ választ adhatnak a megkérdezettek az újág kérdésre (nem olvas is ide tartozhat) és n féle végzettséget különböztetünk meg. Ekkor ha N volt a megkérdezettek száma készítsünk el egy úgynevezett kontingencia táblát, amelynek i-edik sorának j-edik eleme (azaz oszlopa) azon válaszadók $k_{i,j}$ számát tartalmazza, akik az

i-edik újságot olvassák és végzettségük alapján a j-edik osztályba tartoznak.

	1	2		i		m
1						
2						
j				$k_{i,j}$		
n						

A függetlenség azt jelentené, hogy ha $A_i$annak a valószínûsége, hogy egy véletlenül választott ember az $i-$edik ujságot választja, illetve B_j,hogy a $j-$edik végzettségi kategriába esik, akkor

$$\mathbb{P}\left(A_i B_j\right) = \mathbb{P}\left(A_i\right)\mathbb{P}\left(B_j\right).$$

Ugyanakkor a bal és jobb oldalon található valószínûségekre a kontingencia táblázat becséseket tartalmaz. Világos, hogy ha

$$r_i = \sum_{j=1}^{m} k_{i_j}$$

akkor

$$\frac{r_i}{N}$$

becsli $\mathbb{P}\left(A_i\right)$-t, illetve ha

$$c_i = \sum_{i=1}^{n} k_{i_j}$$

akkor

$$\frac{c_j}{N}$$

becsli $\mathbb{P}\left(B_j\right)$,végül pedig

$$\frac{k_{i,j}}{N}$$

becsli $\mathbb{P}\left(A_i B_j\right)$-t. Ennek alapján a következõ próba függvény konstruálható.

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\left(k_{i,j} - e_{i,j}\right)^2}{e_{i,j}} \tag{7.1}$$

ahol

$$e_{i,j} = \frac{c_i o_j}{N}.$$

Állítás 17 *A* (7.1)*beli próbafüggvény jó közelítéssel* χ^2 *eloszlású* $(n-1)(m-1)$ *szabadságfokkal, ha az egyes cellákban kapott elvárt* $e_{i,j}$ *értékek mindegyike nagyobb mint* 5.

Az állítást nem biszonyítjuk, helyessége a korábbi gondolatmenetbbõl intuitíven látható.

Megjegyzés 10 *A leírt függetlenség vizsgálat valószínûségi változók függetlenségének eldöntésére is alkalmazható oly módon, hogy elkészítjük a kétdimenziós tapasztalati együttes eloszlás alapján megfelelõ intervallumok választásával a kontingencia táblázatot.*

Gyakorlat 32 *Készítsük el a megjegyzés szerinti függetlenségvizsgálat teljes menetét.*

7.0.5 Illeszkedés, normalitás vizsgálat

A multinomiális eloszlásnál látott módon tetszõleges eloszláshoz való illeszkedést is lehet tesztelni.

Legyen Ψ egy tetszleges eloszlás függvény. Tegyük fel, hogy adott $x_i x_{i+1}\ i = 1..k$ értékekhez tartoznak a

$$p_i = \Psi\left(x_{i+1}\right) - \Psi\left(x_i\right)$$

értékek. Ekkor az $X_1, X_2 ... X_n$ független azonos eloszlású valószínüségi változók akkor származnak a Ψ eloszlásból (α szinten), ha a

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(k_i - Np_i\right)^2}{Np_i}$$

χ^2 eloszlású próbafüggvény az $1 - \alpha/2, \alpha/2$ kritkus értékek köz; esik.

Az eljárással ellenõrizhetõ az a feltevés, hogy X normális eloszlású-e. Pontosabban célszerû a $z = \frac{X - \overline{X}}{s}$ standardizáltat vizsgálni, hogy az illeszkedik-e a standard normális eloszláshoz.

Gyakorlat 33 *Készítsük el a standard normális eloszlás esetén azt az x_i sorozatot , amelyre $p_i = 0.15$.*

Megjegyzés 11 *Az eljárás nem jól alkalmazható olyan paraméteres eloszlások esetén, ahol nem áll rendelkezésre a standardizáláshoz hasonló az ismeretlen paramétert egyszerû transzformációval kiküszöbölõ módszer. Ilyenkor ugyanis bár lehet, hogy az eloszlás $\Psi_a - t$követi, de ha azt egy $b \neq a$ paraméterû Ψ_b-hez illesztjük, akkor a próbafüggvény esetleg igen nagy lehet. Erre a legegyszerûbb példa, ha a μ várható értékû normális eloszlást egy μ'-vel centráljuk természetesen nagyon rossz illeszkedést kapunk (lásd (**??**)Ábra).*

7.0.6 Próbák helyzeti paraméterek vizsgálatára

Az elõzõ (11) Megjegyzés is muitatja, hogy a helyzeti paraméterek próbái milyen fontosak. Az alábbiakban olyan teszteket ismertetünk amelyek többek között a helyzeti paraméterekrõl szolgálnak információval.

Elõjel próba

Ha egy eloszlásnak nem ismert az m mediánja, azaz az a m érték, melyre $\mathbb{P}\left(X < m\right) = 1/2$, az alábbi állítás segítségével lehet a

$$H_0 : m = m_0 \tag{7.2}$$

hipotézist ellenõrizni.

Állítás 18 *Legyen $X_1, X_2 ... X_n$ független azonos eloszlású valószínüségi változók*

$$\begin{aligned} Y_i &= \begin{cases} 1 & \textit{ha } X_i > m \\ 0 & \textit{egyébként} \end{cases} . \\ Y &= \sum_{i=1}^{n} Y_i \end{aligned}$$

Ekkor, ha $n \geq 20$

$$Y \sim N\left(\frac{1}{2}n, \sqrt{\frac{1}{4}n}\right)$$

Következmény 6

$$z = \frac{Y - \frac{1}{n}n}{\sqrt{\frac{1}{4}n}}$$

próbafüggvény standard normális eloszlású, alkalmas (7.2) *ellenõrzésére.*

Minkét állítás evidens a binomiális eloszlásra vonatkozó centrális határeloszlás tételbõl.

Wilcoxon féle elõjeles rang test

A most ismertetésre kerülõ módszer párosított (azaz nem független minta esetén) az X ill. Y változók F, G eloszlásának azonosságát ellenõrzni. Legyen $X_1, X_2 ... X_n$ illetve $Y_1, Y_2 ... Y_n$ független azonos eloszlású valószínüségi változók (X illetve Y példányai).

X	Y	d	$\lvert d\rvert$	rang $\lvert d\rvert$	elõjeles rang
x_1	y_1	d_1	$\lvert d_1\rvert$	r_1	$sign\,(d_1)\, r_1$
x_2	y_2	d_2	$\lvert d_2\rvert$	r_2	$sign\,(d_2)\, r_2$
x_n	y_n	d_n	$\lvert d_n\rvert$	r_n	$sign\,(d_n)\, r_n$
					$T = \sum_{i=1}^{n} sign\,(d_i)\, r_i$

Állítás 19 *Ha $F = G$ akkor a fenti T próbafüggvény jó közelítéssel normális eloszlású, és*

$$\begin{aligned} \mathbb{E}(T) &= 0 \\ \sigma_T &= \sqrt{\frac{n\,(n + 1\,(2n+1))}{6}} \end{aligned}$$

$$z = \frac{T}{\sigma_T}$$

alkalmas próbafüggvény.

Az állítás a centrális hatéreloszlás tétel egy élesítéséböl következik, nem bizonyítjuk.

7.0.7 Man-Whitney próba

Az alábbi próba újra a két eloszlás azonossága, az

$$H_0 : F = G$$

feltevés ellenõrzésére szolgál.

Legyen mint elõbb $X_1, X_2...X_n$ független F eloszlású valószínüségi változók illetve $Y_1, Y_2...Y_m$ független G eloszlású valószínüségi változók . Keverjük össze a két mintát és rendezzük nagyság szerint. Legyenek az X_i elemek rangjai $r_{\pi(i)}$ezek összege pedig $T = sr_X$.

Állítás 20 *Ga $F = G$, akkor a fenti T statisztikár igazak az alábbiak.*

$$\begin{aligned} E(T) &= \frac{1}{2}n(n+m+1) \\ \sigma_T &= \sqrt{\frac{1}{2}nm(n+m+1)} \end{aligned}$$

és

$$z = \frac{T - \frac{1}{2}n(n+m+1)}{\sigma_t}$$

standard normális eloszlású valószínüségi változó.

A Kruskal-Wallis teszt

A következõ teszt az összetett

$$H_0 : F_1 = F_2 = ... = F_k$$

hipotézis ellenõrzésére szolgál., azaz k minta alaápán k eloszlás azonosságát teszteli. Legyen $X_{1,j}, X_{2,j}...X_{n,j}$ független azonos eloszlású valószínüségi változók $j = 1, 2...k$-ra. Hasonlóan mint a Man-Whitney próba esetén keverjük össze a mintákat minden elem kapja meg a neki megfelelõ rangot. Jelölje R_j a $j-$edik minta elemeinek rangösszegét, n_T pedig az összes elemek számát.

Állítás 21 *Ha igaz a*

$$H_0 : F_1 = F_2 = ... = F_k$$

feltevés és minden $j - re\ n_j \geq 5$, akkor

$$W = \frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n_T+1)$$

próbafüggvény $k-1$ szabadságfokú χ^2 eloszlást követ.

A Spreman féle rangkorreláció

Párosított minták esetén gyakori kérdés, hogy mi a kapcsolat, mi a korreláció a két változó között. Idõnként célszerú ezt a kérdést is nem paraméteres módon megközelíteni. Erre szolgál a következõ fogalom.

Definíció 56 *A Spreman féle rangkorreláció. Legyenek a két párosított minta rangjai* r_i, s_i. *Legye* $d_i = r_i - s_i$. *Ekkor*

$$r_S = 1 - \frac{\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}$$

a tapasztalati rangkorreláció,

$$\rho_S = \mathbb{E}\left(r_S\right)$$

pedig a Spreman féle rangkorreláció.

Állítás 22 *Ha*

$$H_0 : \rho_S = 0$$

akkor persze

$$\begin{aligned} \mathbb{E}\left(r_S\right) &= 0 \\ \sigma\left(r_S\right) &= \sqrt{\frac{1}{n-1}} \end{aligned}$$

és

$$z = \frac{r_s - 0}{\sigma\left(r_S\right)}$$

közel standard normális eloszlás eloszlású, ha $n \geq 30$.

Chapter 8
SZÓRÁSANALíZIS

A szórás analízis (analysis of variance ANOVA) a paraméteres próbák egy érdekes családja, amely egy közös modellre épít. A legegyszerûbb az egyváltozós, egy faktoros eset. A módszer megismerését egy ilyen példával kezdjük.

Példa 34 *Egy étteremlánc ugyanazt a hamburgerét sok étteremben kínálja. A hamburger népszerüsítésére akciót kíván indítani. Elõször azt vizsgálja országonként 30-30 éttermeben, hogy mennyit nõtt a nyeresége ezen a terméken. Azaz rendelkezésre állnak az* $X_{i,j}$ *eredmények, ahol* i *az ország felsorolás,* $j = 1..30$ *pedig az éttermek sorszáma. Ez azt jelenti, hogy* $X_{i,j}$ $j = 1..30$ *egy független* 30 *elemû minta egy* Y_i *változóból. Elsõ feltevésük, hogy az akció eredménye független az országtól, azaz*

$$\mu_i = E(Y_i) = \mu$$

minden $i = 1..k-$*ra, azaz*

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

összetett hipotézist kell ellenõrizni.

A statisztika hagyományos kifejezésével, szokás az i-edik kezelésrõl beszélni, mert a módszert a növénytermesztésben, az egyes földterületek eltérõ kezelésének összevetésére alkalmazták elõször. Az azonos kezelésnek alávetett egyedek alkotnak egy csoportot. Szokás szerint olyan statisztikát keresünk, ami a H_0 feltevés mellett jól viselkedik. Az egye mintaelemekre a modell szerinti feltevés a következõ:

$$X_i(j) = \mu_i + \varepsilon_i(j)$$

ahol

$$\varepsilon_i(j) \sim N(0, \sigma)$$

fügetlen valószínüségi változók ismeretlen közös $\sigma-$el.

Szokás a modelt a

$$X_i(j) = \mu + \alpha_i + \varepsilon_i(j)$$

alakban megfogalmazni, ahol α_i az i-edik kezelés egyedi hatása. Ha H_0 igaz, akkor a sokaságban tapasztalt szórásranégyzetre a teljes minta alapján becslés is adható. Ugyanakkor az egyes csoportokon belül is adható variancia-becslés, ami a feltevés szerint független becsléseket ad, ezek átlaga szintén becsli a teljes sokaság szórásnégyzetét. Így kétféle becslést lehet készíteni a szórásnégyzetre. Ehhez elõször a tapasztalati átlagokra vezessünk be jelölést.

Legyen az n_i elemû i-edik csoport tapasztalati átlaga

$$\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_i(j),$$

a teljes átlag

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} X_i(j)$$

ahol $n = \sum_{i=1}^{p} n_i$ az összes mintaelemek száma. Bevezetjük a

$$SST = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_i(j) - \overline{X} \right)^2$$

teljes négyzetösszeget valamint a

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(X_i(j) - \overline{X}_i \right)^2$$

csoportokon beküli négyzetösszegek összegét, végül pedig a

$$SSTR = \sum_{i=1}^{p} n_i \left(\overline{X}_i - \overline{X} \right)^2$$

a csoportosításból fakadó négyzetösszeget.

Lemma 2

$$SST = SSE + SSTR$$

Bizonyítás. Az állítás egyszerû aritmetikai átalakítással igazolható. A kidolgozást az olvasóra bízzuk. ■

A próba megalapozását a következõ már nehezebb állítás biztosítja.

Állítás 23 *Jelölje*

$$MSR = \frac{1}{p-1} SSTR$$

a csoportosításból fakadó átlagos tapasztalati négyzetösszeget. Ekkor

$$E(MSTR) = \sigma^2 + \frac{1}{p-1} \sum_{i=1}^{p} n_i (\mu_i - \mu)^2.$$

Bizonyítás. Kezdjük a definíció szerinit zárójel bõvítésével és felbontásával.

$$\begin{aligned} \left(\overline{x}_i - \overline{\overline{x}} \right)^2 &= \left(\overline{x}_i - \mu_i + \mu_i - \mu + \mu - \overline{\overline{x}} \right)^2 \\ &= (\overline{x}_i - \mu_i)^2 + (\mu_i - \mu)^2 + \left(\mu - \overline{\overline{x}} \right)^2 \\ &\quad + 2(\overline{x}_i - \mu_i)(\mu_i - \mu) + (\overline{x}_i - \mu_i)\left(\mu - \overline{\overline{x}} \right) + 2(\mu_i - \mu)\left(\mu - \overline{\overline{x}} \right) \end{aligned}$$

Vegyük észre, hogy a cetrálás miatt, illetve mert a szorzat másik tagja konstans

$$E\left[2\left(\overline{x}_i-\mu_i\right)\left(\mu_i-\mu\right)\right]=E\left[2\left(\mu_i-\mu\right)\left(\mu-\overline{\overline{x}}\right)\right]=0.$$

Az alábbi két tagban mintaátlagok szórásai szerepelnek, ezért

$$\sum_{i=1}^{r} n_i E\left[\left(\overline{x}_i-\mu_i\right)^2+\left(\mu-\overline{\overline{x}}\right)^2\right]=\sum_{i=1}^{r} n_i\left(\frac{\sigma^2}{n_i}+\frac{\sigma^2}{n}\right)=r+1. \tag{8.1}$$

A harmadik négyzetes tag a jobboldalon kívánt összege a μ_i-k és a μ négyzetes eltérésösszegének:

$$\frac{1}{r-1}\sum_{i=1}^{r} n_i\left(\mu_i-\mu\right)^2 \tag{8.2}$$

Foglalkozzunk a megmaradt egyetlen keresztszorzattal.

$$\begin{aligned}
& E\left[\sum_{i=1}^{r} 2n_i\left(\overline{x}_i-\mu_i\right)\left(\mu-\overline{\overline{x}}\right)\right] \\
= \; & E\left[\sum_{i=1}^{r} \frac{2n_i}{n}\left(\overline{x}_i-\mu_i\right)\left(n\mu-n\overline{\overline{x}}\right)\right] \\
= \; & E\left[\sum_{i=1}^{r} \frac{2n_i}{n}\left(\overline{x}_i-\mu_i\right)\left(n\mu-n_i\mu_i+n_i\mu_i-\sum_{j=1}^{n_i} x_{i,j}+\sum_{j=1}^{n_i} x_{i,j}-n\overline{\overline{x}}\right)\right] \\
= \; & E\left[\sum_{i=1}^{r} \frac{2n_i}{n}\left(\overline{x}_i-\mu_i\right)\left(\left(n_i\mu_i-\sum_{j=1}^{n_i} x_{i,j}\right)+\left(n\mu-n_i\mu_i-\sum_{k\neq i}\sum_{j=1}^{n_i} x_{k,j}\right)\right)\right] \\
= \; & E\left[\sum_{i=1}^{r} \frac{2n_i^2}{n}\left(\overline{x}_i-\mu_i\right)\left(\mu_i-\overline{x}_i\right)+\sum_{i=1}^{r} \frac{n_i}{n}\left(\overline{x}_i-\mu_i\right)\left(n\mu-n_i\mu_i-\sum_{k\neq i}\sum_{j=1}^{n_i} x_{k,j}\right)\right] \\
= \; & E\left[-2\sum_{i=1}^{r} \frac{n_i^2}{n}\left(\overline{x}_i-\mu_i\right)^2+2\sum_{i=1}^{r} \frac{n_i}{n}\left(\overline{x}_i-\mu_i\right)\left(n\mu-n_i\mu_i-\sum_{k\neq i}\sum_{j=1}^{n_i} x_{k,j}\right)\right]
\end{aligned}$$

Vegyük észre, hogy a második tagban a szorzat két tényezõje független, hiszen az utoóbbiban pont az i-edik csoport elemei nem kerülnek összegzésre. A várható érték így a várható értékek szorzata és mindkettõ centrált, mindkettõ nulla. Az elsõ tag pedig megint a mintaátlagok szórása, így

$$E\left[-2\sum_{i=1}^{r} \frac{n_i^2}{n}\left(\overline{x}_i-\mu_i\right)^2\right]=-2\sum_{i=1}^{r} \frac{n_i^2}{n}\frac{\sigma^2}{n_i}=-2\sigma^2. \tag{8.3}$$

Összevetve (8.1), (**??**) és (8.3)-t kapjuk az állítást. ■

Ez tehát azt jelent, hogy MSR pontosan akkor torzítatlan becslése σ^2-nek ha az összes márható érték egyenlõ, azaz a null hipotézis teljesül. Ellenkezõ esetben σ^2-tõl felfelé tér el.

Állítás 24 *A model feltevései melett igaz, hogy*

$$E\left(\frac{1}{n-p}SSE\right) = \sigma^2$$

továbbá

$$E\left(\frac{1}{n-1}SST\right) = \sigma^2$$

és $MSE = \frac{1}{n-p}SSE$ *illetve* $MSRT = \frac{1}{p-1}SSTR$ *függetlenek. Igaz továbbá, hogy*

$$\frac{SSE}{\sigma^2}$$

$n-p$ szabadságfokú χ^2 eloszlású valószínüségi változó továbbá, ha igaz a null hipotézis akkor

$$\frac{SSRT}{\sigma^2}$$

$p-1$ szabadságfokú χ^2 valószínüségi változó.

Az állítások evidensek, kivéve a függetlenséget, ezt nem bizonyítjuk.
Mindezek alapján a kövtekezõ próba végezhetõ.

Tétel 35 *Ha a model feltevései fennállnak és igaz a null hipotézis, akkor a*

$$F = \frac{MSTR}{MSE}$$

próbafüggvény $(p-1, n-p)$ *szabadságfokú* F *eloszlást követ. Azaz* F *értéke közel van* $1-hez$. *Ellenkezõ esetben* F *értéke nagyobb.*

8.0.8 Kétrészes osztályozás

Chapter 9
LINEÁRIS REGRESSZIÓ

A modell:

$$Y = \alpha X + \beta + \varepsilon$$

ahol $\varepsilon \tilde{} N(0, \sigma)$ független az X-től. Keressük az

$$F(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

minimumát a, b-ben. Keressük a $\frac{\partial}{\partial b} F(a, b) = 0$, $\frac{\partial}{\partial a} F(a, b) = 0$ megoldásokat.

$$\frac{\partial}{\partial b} F(a, b) = -2 \sum_{i=1}^{n} (y_i - ax_i) - nb$$

amiből

$$\begin{aligned} nb &= \sum_{i=1}^{n} (y_i - ax_i) \\ b &= \overline{y} - a\overline{x}. \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial b} F(a, b) &= 2 \sum_{i=1}^{n} (y_i - ax_i - b) x_i \\ &= 2 \sum_{i=1}^{n} (y_i - ax_i - \overline{y} - a\overline{x}) x_i \\ &= 2 \sum_{i=1}^{n} \left[y_i x_i - a x_i^2 \right] - n\overline{xy} - an (\overline{x})^2 . \end{aligned}$$

Ebből az $s_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\overline{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$ jelöléssel

$$\begin{aligned} ans_x^2 &= a \left(\sum_{i=1}^{n} x_i^2 - n (\overline{x})^2 \right) \\ &= \sum_{i=1}^{n} y_i x_i - n\overline{xy} \end{aligned}$$

azaz

$$a = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{n s_x^2},$$

vagy másképpen bevezetve az $s_{x,y} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$ jelölést

$$\begin{aligned} a &= \frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{n s_x^2} = \frac{n\left(\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{xy}\right)}{n s_x^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{xy}}{s_x^2}, \end{aligned}$$

azt kapjuk, hogy

$$a = \frac{s_{x,y} - \overline{xy}}{s_x^2}.$$

Összefoglalva:

$$\begin{aligned} a &= \frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{n s_x^2}, \\ b &= \overline{y} - a\overline{x}. \end{aligned}$$

Ebből, ha a tapasztalati covarianciát illetbe korrelációs együtthatót $Cov_{x,y}$ illetve $r_{x,y}$ jelöli, azt nyerjük, hogy

$$a = \frac{Cov_{x,y}}{s_x^2} = \frac{r_{x,y} s_x s_y}{s_x^2} = r_{x,y}\frac{s_y}{s_x}.$$

Definíció 57 *Jelölje* $\widehat{y}_i = ax_i + b$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

a lineáris illeszkedés négyzetes hibájának átlagát, vagy másképpen az egyenes körüli)korrigált tapasztalati szórásnégyzetet.

Definíció 58 *Vezessük be a következő rövidítéseket*

$$\begin{array}{ll} SX = \sum_{i=1}^{n} x_i & SX^2 = \sum_{i=1}^{n} x_i^2 \\ SY = \sum_{i=1}^{n} y_i & SY^2 = \sum_{i=1}^{n} y_i^2 \\ SXY = \sum_{i=1}^{n} x_i y_i & \end{array}$$

Megjegyzés 12 *Természetesen*

$$\begin{aligned} \overline{x} &= nSX. \\ s_x^2 &= \frac{1}{n}SX^2 - (\overline{x})^2 \end{aligned}$$

Tétel 36 *Ha* $\varepsilon_i \tilde{} N(0,\sigma)$ *és korrelálatlanok, akkor*

$$\begin{aligned} E(a) &= \alpha, \\ E(b) &= \beta, \\ \sigma^2(a) &= \frac{s^2}{s_x}, \\ \sigma^2(b) &= s^2 \frac{SX^2}{n s_x^2}. \end{aligned}$$

Az állítást nem bizonyítjuk.

Definíció 59

$$\begin{aligned} SST &= \sum_{i=1}^{n} (y_i - \overline{y})^2, \\ SSR &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2, \\ SSE &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2. \end{aligned}$$

SST az y ingadozását méri, SSE az egyenes és a méréspontok közötti hiba négyzetes összegét.

Megjegyzés 13

$$\begin{aligned} SSR &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 \\ &= \sum_{i=1}^{n} (a x_i + b - a\overline{x} - b)^2 \\ &= \sum_{i=1}^{n} (a x_i - a\overline{x})^2 \\ &= a^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= a^2 n s_x^2 = \\ &\quad a^2 \left[SX^2 - \frac{1}{n} (SX)^2 \right]. \end{aligned}$$

Ha figyelembe vesszük, hogy

$$a = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{n s_x^2} = \frac{SXY - \frac{1}{n} SX SY}{n s_x^2}$$

akkor egyrészt azt kapjuk, hogy

$$SSR = a^2 n s_x^2 = \frac{\left[SXY - \frac{1}{n} SXSY\right]^2}{n s_x^2} \tag{9.1}$$

azaz

$$SSR = \frac{\left[SXY - \frac{1}{n} SXSY\right]^2}{SX^2 - \frac{1}{n}\left(SX\right)^2}.$$

Ugyanakkor a korrelációs együtthatóval kifejezve:

$$\begin{aligned} SSR &= a^2 n s_x^2 \\ &= \left[r\left(x,y\right)\frac{s_y}{s_x}\right]^2 n s_x^2 \\ &= r\left(x,y\right)^2 n s_y s_x = nCov\left(x,y\right). \end{aligned}$$

Tehát az SSR a lineáris kapcsolat erősségét fejezi ki.

Lemma 3

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} y_i x_i - n\overline{xy}$$

Bizonyítás.

$$\begin{aligned} & \sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right) \\ = & \sum_{i=1}^{n}\left[y_i x_i - x_i\overline{y} - \overline{x}y_i + \overline{xy}\right] \\ = & \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i\overline{y} - \overline{x}\sum_{i=1}^{n} y_i + n\overline{xy} \\ = & \sum_{i=1}^{n} y_i x_i - n\overline{xy} - n\overline{xy} + n\overline{xy} \\ = & \sum_{i=1}^{n} y_i x_i - n\overline{xy}. \end{aligned}$$

■

Tétel 37

$$SST = SSE + SSR.$$

Bizonyítás.

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - ax_i - (\overline{y} - a\overline{x}))^2 \\ &= \sum_{i=1}^{n} [(y_i - \overline{y}) - a(x_i - \overline{x})]^2 \\ &= \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2a \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) + a^2 \sum_{i=1}^{n} (x_i - \overline{x}) \\ &= SST - 2a \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) + a^2 n s_x^2. \end{aligned}$$

Megint helyettesítsünk a-ba.

$$\begin{aligned} a^2 n s_x^2 &= \left[\frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{n s_x^2} \right]^2 n s_x^2 \\ &= \frac{\left[\sum_{i=1}^{n} y_i x_i - n\overline{xy}\right]^2}{n s_x^2} \end{aligned}$$

amiből $\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) = \sum_{i=1}^{n} y_i x_i - n\overline{xy}$ miatt

$$\begin{aligned} SSE &= SST - 2\frac{\left[\sum_{i=1}^{n} y_i x_i - n\overline{xy}\right]^2}{n s_x^2} \\ &\quad + \frac{\left[\sum_{i=1}^{n} y_i x_i - n\overline{xy}\right]^2}{n s_x^2} \\ &= SST - \frac{\left[\sum_{i=1}^{n} y_i x_i - n\overline{xy}\right]^2}{n s_x^2} \\ &= SST - a^2 n s_x^2. \end{aligned}$$

Ugyanakkor (9.1)-ban láttuk, hogy

$$SSR = a^2 n s_x^2,$$

azaz

$$SSE = SST - SSR.$$

■

Definíció 60

$$r^2 = \frac{SSR}{SST} = \frac{SXY - \frac{1}{n} SX SY}{SY^2 - \frac{1}{n} (SY)^2}.$$

Megjegyzés 14 *Tudjuk, hogy*

$$SSR = a n s_x^2,$$

$$SST = n s_y^2$$

és

$$\begin{aligned} a &= r_{x,y}\frac{s_y}{s_x} \\ a^2 &= .r_{x,y}^2\left(\frac{s_y}{s_x}\right)^2, \end{aligned}$$

ezért

$$\begin{aligned} r_{x,y}^2 &= \left(\frac{s_x}{s_y}\right)^2 a^2 = \frac{ns_x^2}{ns_y^2}\frac{SSR}{ns_x^2} \\ &= \frac{SSR}{ns_y^2} = \frac{SSR}{SST} = r^2. \end{aligned}$$

Az $r_{x,y}$ előjelét pedig az egyenes a meredekségének előjele határozza meg, ezért

$$r_{x,y} = sign\,(a)\,\sqrt{r^2}.$$

Megjegyzés 15 *Vegyük észre, hogy*

$$\begin{aligned} r^2 &= \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \\ |r| &= \sqrt{1 - \frac{SSE}{SST}}. \end{aligned}$$

Chapter 10
FŐKOMPONENS ANALíZIS

10.1 A lineáris algebra néhány eleme

Definíció 61 *Ha $a, b \in \mathbb{R}^p$ akkor skaláris szorzatuk*

$$a^T b = (a, b) = \sum_{i=1}^{p} a_i b_i.$$

Definíció 62 *Ha $a \in \mathbb{R}^n, b \in \mathbb{R}^p$ akkor diadikus szorzatuk*

$$ab^T = (a_i b_j)_{n \times p}$$

$n \times p$-s mátrix.

Definíció 63 *Ha $a \in \mathbb{R}^p$ akkor az a vektor l_2 normája, avagy hossza:*

$$\|a\|_2 = \|a\| = a^T a = (a, a) = \sum_{i=1}^{p} a_i^2.$$

Definíció 64 *Egy u vektort normáltnak nevezünk, ha $\|u\| = 1$, azaz hossza 1.*

Definíció 65 *Ha A $p \times p$-s mátrix, akkor azt mondjuk, hogy az $u \in \mathbb{R}^p$ vektor az A mátriz sajátvektora a λ sajátértékkel, ha*

$$Au = \lambda u.$$

Állítás 25 *Ha A szimmetrikus mátrix akkor minden sajátértéke valós, ezek száma p. Legyenek $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ a sajátértékek ϕ_i a hozzájuk tartozó (normált) sajátvektorok.*

$$\lambda_1 = \max_{\|u\|=1} (u, Au),$$

$$\lambda_i = \max_{\|u\|=1, (u,\phi_i)=0: i=1...i-1} (u, Au). \tag{10.1}$$

Azaz az i-edik sajátvektor az első $i-1$ feszítette altérre merőleges.

Állítás 26 *Legyen $U = \left(\phi_1, ...\phi_p\right)$ a sajátvektorokból alkotott mátrix. Ekkor*

$$A = U^T \Lambda U$$

ahol

$$\Lambda = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ & 0 & & \\ 0 & & & 0 \quad \lambda_p \end{pmatrix}$$

diagonális mátrix. Ez az A mátrix spektrál felbontása.

Definíció 66 *Egy U mátrix ortonormált, ha*

$$U^T U = I,$$

ami egyben azt is jelenti, hogy

$$U^{-1} = U^T.$$

Definíció 67 *Egy A mátrix nyoma (trace-e)*

$$\mathrm{tr}\,(A) = \sum_{i=1}^{p} a_{i,i}$$

a diagonális elemeinek összege.

Lemma 4 *Tetszőleges A, B $k \times l$-s mátrixokra (ahol $k, l \geq 1$)*

$$\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$$

Bizonyítás.

$$\begin{aligned} \mathrm{tr}\,(AB) &= \sum_{i=1}^{p} (AB)_{i,i} = \sum_{i=1}^{p} \left(\sum_{k=1}^{p} a_{i,k} b_{k,i} \right) \\ &= \sum_{i=1}^{p} \left(\sum_{k=1}^{p} a_{i,k} b_{k,i} \right) = \sum_{i=1}^{p} \left(\sum_{k=1}^{p} b_{k,i} a_{i,k} \right) \\ &= \sum_{k=1}^{p} \left(\sum_{i=1}^{p} b_{k,i} a_{i,k} \right) = \sum_{k=1}^{p} (BA)_{k,k} = \mathrm{tr}\,(BA)\,. \end{aligned}$$

■

Lemma 5

$$u^T A v = \mathrm{tr}\left(A v u^T\right).$$

Bizonyítás. Mivel $u^T A v \in \mathbb{R}$ skalár, ezért $u^T A v = \mathrm{tr}\left(u^T A v\right)$. Viszont a 4 Lemma miatt $\mathrm{tr}\left(u^T A v\right) = \mathrm{tr}\left(A v u^T\right)$. ■

Lemma 6 *Ha U ortonormált akkor*

$$\mathrm{tr}\left(U^T A U\right) = \mathrm{tr}\,(A)\,.$$

Bizonyítás. Alkalmazzuk a 4 Lemmát $A = U^T, B = AU$ szereposztással.

$$\mathrm{tr}\left(U^T AU\right) = \mathrm{tr}\left(AUU^T\right).$$

Viszont $U^T U = I$, amiből következik az állítás. ■

Következmény 7 *Ha A szimmetrikus mátrix, akkor*

$$\mathrm{tr}\left(A\right) = \sum_{i=1}^{p} \lambda_i.$$

Bizonyítás. A feltétel miatt alkalmazható a 26 Állítás, azaz

$$A = U^T \Lambda U$$

és alkalmazva a 6 Lemmát $V = U^T$-re

$$\mathrm{tr}\left(V^T AV\right) = \mathrm{tr}\left(UU^T \Lambda UU^T\right) = \mathrm{tr}\left(\Lambda\right)$$

amiből következik az állítás. ■

10.2 Véletlen vektorok elforgatása

Definíció 68 *Legyen* $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \in \mathbb{R}^p$ *p-dimenziós valószínűségi vektorváltozó. Ennek várható értéke* $\mu \in \mathbb{R}^p$:

$$\mu = E\left(X\right) = \begin{pmatrix} E\left(X_1\right) \\ E\left(X_2\right) \\ \vdots \\ E\left(X_p\right) \end{pmatrix}.$$

Az X kovariancia mátrixa:

$$\Sigma = Cov\left(X\right) = \left(Cov\left(X_i, X_j\right)\right)_{p\times p}.$$

Vegyük észre, hogy Σ főátlójában $Cov\left(X_i, X_j\right) = \sigma^2\left(X_i\right)$-k állnak.

Lemma 7 *Σ szimmetrikus, pozitív szemidefinit mátrix.*

Bizonyítás. A szimmetria nyilvánvaló. Ugyanakkor $\Sigma = E\left(\left(X-\mu\right)\left(X-\mu\right)^T\right)$ ezért minden $a \in \mathbb{R}^p$-re

$$a^T \Sigma a = E\left(a^T\left(X-\mu\right)\left(X-\mu\right)^T a\right) = E\left(Y^T Y\right),$$

ahol $Y = a^T\left(X-\mu\right) \in \mathbb{R}$. Azaz

$$a^T \Sigma a = E\left(Y^T Y\right) = E\left(Y^2\right) \geq 0.$$

■

Következmény 8 *A pozitív szemidefinitség egyben azt is jelenti, hogy a Σ sajátértékei mind nem negatívak, hiszen*

$$0 \leq \phi_i^T \Sigma \phi_i = \lambda_i \phi_i^T \phi_i = \lambda_i.$$

Következmény 9

$$\mathrm{tr}\left(\Sigma\right) = \sum_{i=1}^{p} \sigma^2\left(X_i\right)$$

ugyanakkor mivel Σ szimmetrikus a 7 Következmény miatt

$$\mathrm{tr}\left(\Sigma\right) = \sum_{i=1}^{p} \sigma^2\left(X_i\right) = \sum_{i=1}^{p} \lambda_i,$$

azaz a sajétértékek összege azonos az X komponenseinek teljes szórásnégyzetével. Ezért a sajátértékek (amelyek mint tudjuk nemnegatívak) a teljes szórás egy másik felbontását adják.

Legyen $e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ egységvektor. Világos, hogy

$$e_i^T \Sigma e_i = \sigma^2\left(X_i\right)$$

azaz az X e_i irányú szórásnégyzete $\sigma^2\left(X_i\right)$. Hasonlóan bármely u normált vektorra $u^T \Sigma u$-t tekinthetjük az X u irányú szórásának. Ez egyben a sajátértékek (10.1) előállítása alapján azt is jelenti, hogy ϕ_1 az az irány amelyre az X legnagyobb szóródása "esik" ez pedig λ_1. A következő sajátvektor ϕ_2 a ϕ_1-re merőleges altérben az az irány amely a maximális szórást adja, ez λ_2, sorra így tovább. Ezzel az X teljes szórásátnégyzetét nemcsak felbontottuk a λ_i-k összegére, de magtaláltuk azokat az egymásra merőleges irányokat, amelyekre eső szórásnégyzetek összege kiadja a teljes szórásnégyzet összeget is.

10.3 A vektrováltozó elemi statisztikai viselkedése

Legyen $X_1, ...X_n$ az X eloszlásából származó n elemű minta. Jelölje

$$M = \left(X_1, ...X_n\right)$$

a $p \times n$ dimenziós úgynevezett minta mátrixot. Vigyázat, itt most mindegyik X_i maga egy p-dimenziós vektor, nem pedig az X vektor i-edik komponense. Ha valahol ez félreértésre ad okot, ott $\left(X_i\right)_m$ fogja jelölni az X_i vektor m-edik komponensét. Jelölje

$$\overline{X} = \frac{1}{n} \sum X_i,$$

tapasztalati vagy mintaátlag,

$$S_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T$$

a tapasztalati kovariancia mátrix,

$$S_n^* = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T$$

pedig a korrigált tapasztalati kovariancia mátrix.

Az egydimenziós Steiner tétel alábbi általánosítása is igaz.

Tétel 38 *(Steiner tétele) Minden $a \in \mathbb{R}^p$-re*

$$nS_n = \sum_{i=1}^{n}\left(X_i - a\right)\left(X_i - a\right)^T - n\left(a - \overline{X}\right)\left(a - \overline{X}\right)^T.$$

Bizonyítás. A bizonyítás lényegében azonos az egydimenziós eset igazolásával. Kivitelezését az olvasóra bízzuk mint gyakorlat. ■

Tétel 39 *1. $\overline{X}$ torzítatlan, konzisztens becslése μ-nek,*

2. S_{n-1} torzítatlan, konzisztens becslése Σ-nak,

Ha igaz továbbá, hogy $X_i \in N_p\left(\mu, \Sigma\right)$ $i = 1, 2...n$-re, akkor

3. $\overline{X} \in N_p\left(\mu, \frac{1}{n}\Sigma\right)$,

4. $\overline{X}$ és S_n független.

5. Ha Σ pozitív definit, azaz nem elfajuló és $n > p$, akkor S_n is pozitív definit 1 valószínűséggel, azaz a tapasztalati kovariancia mátrix sem elfajuló.

Tétel 40 *Ha a kovariancia mátrix pizitív definit, és*

$$\Sigma = U^T \Lambda U$$

a spektrális felbontása, akkor

$$Y = U\left(X - \mu\right)$$

valószínűségi változóra

$$\begin{aligned} E\left(Y\right) &= 0 \\ Cov\left(Y\right) &= \Lambda, \end{aligned}$$

ha továbbá $X \in N_p\left(\mu, \Sigma\right)$, akkor

$$Y \in N\left(0, \Lambda\right).$$

Bizonyítás.

$$\begin{aligned} E\left(YY^T\right) &= E\left(U\left(X-\mu\right)\left(X-\mu\right)^T U^T\right) \\ &= UU^T\Lambda UU^T = \Lambda. \end{aligned}$$

■

Tétel 41 *Az X L_2 normában (négyzetes távolságban) mért legjobb k dimenziós becslését az X-nek az X első k főkomponense által kifeszített altérre vonatkozó vetülete biztosítja.*

Nem bizonyítjuk.

Tétel 42 *Legyen W ortonormált mátrix. Ekkor az Z és WZ főkomponensei azonosak.*

Bizonyítás. Legyen most $X = WZ$. A Z kovariancia mátrixának $U^T\Lambda U$ felbontásával a Z főkomponense UZ-ne adódik. Ugyanakkor X-re

$$\begin{aligned} E\left(XX^T\right) &= E\left(WZZ^TW^T\right) = WE\left(ZZ^T\right)W^T \\ &= W\Sigma W^T = WU^T\Lambda UW^T, \end{aligned}$$

ezért az X főkomponense $Y = \left(UW^T\right)X = \left(UW^T\right)WZ$. Ebből viszont az következik, hogy $Y = UW^TWZ = UZ$ amit igazolni akartunk. ■

Megjegyzés 16 *Ezzel beláttuk, hogy az ortonormált transzformáció, azaz az elforgatás nem változtatja a főkomponenst. Az egyes koordináták átskálázása viszont igen, azaz nyújtás, affinitásra a főkomponens nem invariáns.*

10.4 A tapszatalati főkomponens

A fentik alapján mostmár azt vizsgáljuk, hogy, ha az X véletlen vektor eloszlásából egy n elemű mintával rendelkezünk, akkor ebből, hogyan lehet jól közelíteni az X főkomponenseit tapasztalati főkomponensekkel.

Keressük azt az a irányt amire az X vetületének szórása, azaz a^TX szórása maximális. Ezt a maximumot úgy keressük, hogy az a irányában a minta szórását maximalizáljuk. Ezzel a feladatot visszavezettük a fenti gondolatmenetre, úgy hogy a tapasztalati kovariancia mátrix spektrális felbontását használjuk. Legyenek a tapasztalati kovariancia mátrix sajátrétékei $\widehat{\lambda}_1 \geq \widehat{\lambda}_2, ...\widehat{\lambda}_p$, sajátvektoraik pedig $\widehat{a}_1, \widehat{a}_2, ...\widehat{a}_p$. Világos, hogy tapasztalati főkomponensek

$$\widehat{Y}_1 = \widehat{a}_1^T X,$$

$\widehat{Y}_2 = \widehat{a}_2^T X, ...\widehat{Y}_p = \widehat{a}_p^T X$ korrelálatlanok és rendre maximalizálják az a maradék szórásból egy-egy vektorra vetíthető szórást.

Definíció 69 *Tekintsük a sajátértékek következő hányadosát.*

$$\Psi = \frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_k + ... + \lambda_p}$$

és ennek tapasztaliti megfelelőjét.

$$\widehat{\Psi} = \frac{\widehat{\lambda}_1 + \widehat{\lambda}_2 + ... + \widehat{\lambda}_k}{\widehat{\lambda}_1 + \widehat{\lambda}_2 + ... + \widehat{\lambda}_k + ... + \widehat{\lambda}_p}$$

Tétel 43 *Ha* $X_i \in N_p(\mu, \Sigma)$ *akkor*

$$\widehat{\Psi} \in N_p(\Psi, .).$$

A szórásnégyzetet szándékosan nem adtuk meg. Az állítást nem bizonyítjuk.

Következmény 10 *A fenti tétel alapján a*

$$H_0 : \Psi = \Psi_0$$

hipotézist a $\widehat{\Psi}$ *próbafüggvény segítségével lehet ellenőrizni.*

Chapter 11
OSZTÁLYOZÁS, KLASZTEREZÉS

11.1 Osztályozás

Az osztályozás feladata igen sok gyakorlati problema során felmerül. Az üzleti élet mindennapi feladata a vevők, ügyfelek osztályozása. Jó példa erre a hitelbírálat. A banktisztviselő az ügyfél adatai alapján kell, hogy döntsön arról, hogy kaphat-e hitelt, vagy differenciáltabb esetben, milyen hitelkonstrukciókat érdemes ajánlani a számára aszerint, hogy melyik ügyfélosztályba esik. Ezen ügyfélosztályok előzetesen kerülnek kialakításra, ennek módjáról később esik szó. Az informatika is számos osztályozási feladattal foglalkozik. Klasszikus példa a karakterolvasó program. Ez egy pixelhalmazról kell, hogy eldöntse, melyik betűhöz hasonlít leginkább, melyik gépi jellel azonosítsa.

Feladatunk a következő képpen fogalmazható meg. Adott M osztály, $D = \{1, 2, ...M\}$. Az objektumok N attribútummal rendelkeznek, így az objektumok azonosíthatóak az $X \subset \mathbb{R}^N$ halmaza elemeivel, ezek az objektumokat leíró vektorok. Feltesszük, hogy adott egy $\{X, \mathcal{F}, P\}$ Kolmogorov féle valószínűségi mező továbbá egy veszteségfüggvény $w_{i,j}$, ami az objektumok hibás besorolásáakor felmerülő veszteséget jelenti.

$$w_{i,j} = \begin{cases} \geq 0 & \text{ha} \quad i \neq j \\ = 0 & \text{ha} \quad i = j \end{cases}.$$

Jelöljön $\xi \in X$ egy véletlen objektumot, amelynek eloszlása P. Legyen d diszkriminancia függvény, azaz osztályba sorolás és

$$\widehat{\eta} = d(\xi)$$

a d által javasolt osztály, legyen továbbá η a valódi osztálya ξ-nek. A rizikófüggvényt az

$$R(w, d) = E(w(\widehat{\eta}, \eta)) = E(w(\widehat{\eta}(\xi), \eta(\xi))).$$

Azt az osztálybasorolást amely R-t minimalizálja Bayes féle döntésnek, osztálybasorolásnak nevezzük. Jelölje ezt $d^*, \eta^* = d^*(\xi)$. Ha d ismert, akkor $X = X_1 \cup X_2 ... X_M$ osztályozás is ismert. Jelölje C_i^* a Bayes féle döntés osztályait és R^* a Bayes féle döntés rizikófüggvényét..

Állítás 27 *Legyen R egy tetszőleges*

$$R \geq R^*$$

Bizonyítás.

$$R = E(w(\widehat{\eta}, \eta)).$$

Alkalmazzuk a feltételes várható érték alaptulajdonságát, miszerint átlaga az eredeti átlag. Tekintsük először a diszkrét eloszlású $\xi ük$ esetét.

$$\begin{aligned} E\left(w\left(\widehat{\eta},\eta\right)\right) &= E\left[E\left(w\left(\widehat{\eta},\eta\right)|\xi\right)\right] \\ &= E\left[\sum w_{i,j}P\left(\eta=i,\widehat{\eta}=j|\xi\right)\right] \\ &= E\left[\sum w_{i,j}I\left(\widehat{\eta}\left(\xi\right)=j\right)P\left(\eta=i|\xi\right)\right] \\ &= E\left[\sum w_{i,j}I\left(\widehat{\eta}\left(\xi\right)=j\right)q_i\right], \end{aligned}$$

ahol

$$q_i = P\left(\eta\left(\xi\right)=i\right) = P\left(\xi\in C_i\right),$$

annak a valószínűsége, hogy a véletlen ξ objektum az i-edik osztályba tartozik. Így

$$\begin{aligned} R\left(w,d\right) &= E\left[\sum_{i,j} w_{i,j}I\left(\widehat{\eta}\left(\xi\right)=j\right)q_i\right] \\ &= E\left[\sum_{j} I\left(\widehat{\eta}\left(\xi\right)=j\right)\sum_{i} w_{i,j}q_i\right] \\ &\geq E\left[\min_j \sum_i w_{i,j}q_i\right]. \end{aligned}$$

Hasonlóan, abszolút folytonos eloszlású ξ esetén, ha $f_i\left(x\right)$ az i-edik osztályban a ξ sűrűségfüggvénye, akkor $p_i = P\left(\eta=1\right)$ melett

$$f\left(x\right) = \sum_{i=1}^{M} p_i f_i\left(x\right)$$

és Bayes tétele alapján

$$q_i\left(x\right) = P\left(\eta=i|\xi=x\right) = p_i\frac{f_i\left(x\right)}{f\left(x\right)}$$

ezért, ha d^* a Bayes döntés, akkor

$$\sum_{i,j} w_{i,j}f_i\left(x\right) \geq \sum_i w_{i,d^*(x)}p_i f_i\left(x\right)$$

és

$$R^* = \int \min_j \sum_{i=1}^{M} w_{i,j}p_i\left(x\right)f_i\left(x\right)dx.$$

■

11.1.1 A legközelebbi társ módszer

A legközelebbi társ módszer (Nearest Neighbor) az X téren definiált valamilyen metrikára épít. Az objektumok természetéről megszerzett ismereteink ebben a metrikában öltenek testet. Annál jobb a módszer, minnél pontosabb a metrika szétválasztó képessége.

Legyen $T = \{(\xi_1, \eta_1), (\xi_2, \eta_2) .... (\xi_n, \eta_n)\}$ n elemű "tananyag". Elemei jó besorolási döntéseket tartalmaznak. azaz ξ_i valóban a C_{η_i} osztályhoz tartozik. A legközelebbbi társ módszer az új ξ elemet abba a j_0 osztályba sorolja, amelyre igaz, hogy

$$d\left(\xi, \xi_{j_0}\right) \leq d\left(\xi, \xi_i\right) : \forall i = 1...n.$$

A legközelebbi társ ξ_{j_0}.

Legyen most két osztályunk, $M = 2$ és tegyük fel, hogy $W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Legyen

$$R^* = \int \min\left\{p_1 f_1\left(x\right), p_2 f_2\left(x\right)\right\} dx,$$

legyen továbbá

$$R = \lim R_n$$

ahol R_n az n elemű tananyag alapján adódó rizikófüggvény. Világos, hogyha $n \to \infty$ és az elemek valamilyen véletlen eloszlás szerint kerülnek a tananyagba akkor a végtelen sok mintaelem tökéletesen letapogatja az objektumok terét és ezért az erre épített döntés optimális lesz.

Tétel 44 *(Cover&Hart)*

$$R^* \leq R \leq 2R^*\left(1 - R^*\right)$$

továbbá

$$R_n \to R^* = R,$$

azaz a Bayes féle besorolás asszimptotikusan optimális.

11.2 Klaszter analízis

A klaszteranalízis feladata hasonló az osztályozás feladatához. A klaszterezési feladat során ismertek az osztályok, amikbe az objektumokat be kell sorolni, hanem éppen az a kérdés, hogy hogyan alakítsunk ki úgy klasztereket (csoportokat) adott objektumokból, hogy azok

1. viszonylag homogén klasztereket alkossanak
2. a klaszterek jól elkülönüljenek
3. az egyes klasztereket a feladat szellemeben jól tudjuk jellemezni.

Ilyen feladattal találkozhatunk éppen akkor, amikor egy gyártó vagy forgalmazó a fogyasztókat csoportokba kívánja sorolni, piacszegmentációt kíván végezni, annak érdekében, hogy az egyes szegmenseket egyedi módon célozza meg termékkel, marketinggel. Gyakori

feladat a biológia vagy az orvostudomány területén is, hogy nagyszámú megfigyelt objektumainkat csoportokba soroljuk, aminek alapján a lényegi eltérésekre lehet koncentrálni.

Legyenek $X(1), X(2) ... X(n)$ az adott objektumok, jelölje halmazukat X. Mindegyik azonosítható a saját leíró vektorával: $(X_1(i), ... X_N(i))^* \in \mathbb{R}^*$. Keressük az

$$X = \cup_{i=1}^{M} C_i$$

diszjunk osztályozást, ahol M sem ismert.

11.2.1 K-közép módszer

Ez a módszer előre választott M számú klaszter kialakítására szolgál. A módszer iteratív. Tegyük fel, hogy már hogy adottak a $C_1^k, ... C_M^k$ klaszterek. A $k+1$ generáció a következő képpen kapható. Képezzük a klaszterek Z_i^k középpontjait.

$$Z_i^k = \frac{1}{\left|C_i^k\right|} \sum_{X_j \in C_i} X_j.$$

Definiáljuk egy objektum és egy klaszter távolságát mint az objektum és a klaszter középpontjának távolságát:

$$d(X_j, C_i) = d\left(X_j, Z_i^k\right).$$

Sorra vesszük az X_j objektumokat. Tegyük fel, hogy

$$d(X_j, C_l) \leq d(X_j, C_i) : \forall i = 1...M.$$

Ekkor az X_j elemet a C_l osztályba helyezzük és ennek megfelelően módosítjuk az osztályközepeket.

$$Z_i^{k+1} = \frac{1}{Z_i^k + 1} \left(\left|C_i^k\right| Z_i^k + X_j\right).$$

Megállunk az eljárás iterálásával, ha elfogynak a besorolandó elemek vagy, ha a klaszterek nem nagyon változnak már.

Legyen a veszteségfüggvény

$$w = w\left(C_1^k, ... C_M^k\right) = \sum_i \sum_j d(X_j, C_i).$$

Állítás 28 *Igen általános feltételek melett w-t minimalizálja a módszer, a konvergencia kielégítő és független a kezdeti C^0 klaszterek megválasztásától.*

11.2.2 Hierarchikus eljárások

Ez a módszer az N objektum minden $M = 1, 2 ... N$ klaszterbe sorolását elvégzi és az eredmény alapján választhajuk meg az osztályok kívánt számát. Megint feltesszük, hogy az objektumok leíróvektorai között egy d távolság definiált, (nem feltétlenül Euklideszi ez a távolság). A klaszterek távolságát most a középpontjaik távolságával definiáljuk:

$$d(C_i, C_j) = d(Z_i, Z_j).$$

Az eljárás $M^0 = N$ klaszterrel indul, minden objektum külön klasztert alkot. Ezután összevonjuk azt a két osztályt amelyek távolsága minimális. Meghatározzuk az új klaszter

középpontját. Innen az előző lépés folytatható. Alternatív módszer, amely gyorsabb algoritmust eredményez, ha minden osztályt összevonunk, amelyek távolsága egy adott küszöb alatt van. Az eredményt szokás a klaszterek egymást követő generációjának fájával, úgynevezett dendogrammal ábrázolni. Világos, hogy az első módszer esetén a fa N levéllel rendelkezik, az alattuk lévő szinteken sorra $N-1, N-2..$ csúcspont helyezkedik el, a fa magassága pedig N. A második eljárásnál a szintek elemszáma gyorsabban is csökkenhet, így a fa magassága kisebb egyenlő mint N.

Chapter 12
IDŐSOROK

*

12.1 Alapfogalmak, definíciók

A továbbiakban olyan folyamatokat vizsgálunk, amelyeknél $X_1, \ldots, X_n$ nem független, azonos eloszlású változók.

Véges dimenziós eloszlások $X_t : t \in T$
$t_1, \ldots, t_n \in T : (X_{t_1}, \ldots, X_{t_n})$
$\mathrm{P}\left(X_{t_1} < x_1, \ldots, X_{t_n} < x_n\right)$
$\mathrm{P}\left((X_{t_1}, \ldots, X_{t_n}) \in B\right), B \in \Re^n$

[Kolmogorov] Ha adott véges dimenziós eloszlásoknak egy kompatibilis rendszere, akkor létezik egy valószínűségi mező és azon egy sztochasztikus folyamat, amelynek pont ezek a véges dimenziós eloszlásai (vagyis véges dimenziós eloszlásai meghatározzák a folyamatot).

Gauss folyamat Akkor nevezünk egy folyamatot *Gauss folyamat*nak, ha minden véges dimenziós eloszlása normális.

12.1.1 Összefüggőségi struktúrák

1. véges rendű Markov-i összefüggés
2. *martingál tulajdonság*: $\mathrm{E}\left(X_{n+1} | X_1, \ldots, X_n\right) = X_n$
3. *stacionaritás*

Erős stacionaritás Egy sztochasztikus folyamatot akkor nevezünk *erősen stacionárius*nak ha

$\forall t_1, \ldots, t_n, s : (X_{t_1}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_1+s}, \ldots, X_{t_n+s})$

Gyenge stacionaritás Egy sztochasztikus folyamatot akkor nevezünk *gyengén stacionárius*nak ha

1. $\forall t : \mathrm{E}\left(X_t\right) = \mathrm{E}\left(X_1\right)$
2. $\mathrm{cov}(X_t, X_s) = \gamma(t - s)$

A γ függvényt a folyamat *autokovariancia* függvényének nevezzük.

*Marizca István jegyzete alapján. A szerző engedélyével.

12.1.2 Az autokovariancia függvény tulajdonságai

1. $\gamma(0) = D^2[X_t] \geq 0$ (ha létezik)

2. $|\gamma(h)| \leq \gamma(0)$

   Bizonyítás: Cauchy-Schwartz: $\langle a, b\rangle^2 \leq ||a||^2 \, ||b||^2$.

$$\gamma(h) = \operatorname{cov}(X_t, X_{t+h})$$

$$|\operatorname{cov}(X_t, X_{t+h})|^2 \leq \; D^2[X_t] D^2[X_{t+h}]$$

3. $\gamma(h) = \gamma(-h), \forall h \in \mathbb{Z}$

4. pozitív szemidefinit

[Pozitív Szemidefinit] Az $M \in \Re^{n\times n}$ mátrix *pozitív szemidefinit* ha $\forall a \in \Re^n$ esetén $a^T M a \geq 0$.

Ha egy egész számokon értelmezett u függvényre igaz, az, hogy $\forall t_1, \ldots, t_n \in \mathbb{Z}$ esetén a $[u(t_i - t_j)]_{1\leq i\leq n, 1\leq j\leq n}$ mátrix pozitív szemidefinit akkor a függvényt *pozitív szemidefinit*nek nevezzük.

Gyengén stacionárius folyamat autokovariancia függvénye pozitív szemidefinit.

Bizonyítás. Legyen $Z := (X_{t_1} - \mathrm{E}(X), X_{t_2} - \mathrm{E}(X), \ldots, X_{t_n} - \mathrm{E}(X))^T$. Tudjuk, hogy $\mathrm{E}(Z) = 0$, ezért $\forall a \in \Re^n$: $\mathrm{E}\left(a^T Z\right) = 0$. Ezért:

$$D^2[a^T Z] = \mathrm{E}\left(a^T Z a^T Z\right).$$

Mivel $a^T Z$ egy skalár, ezért egyenlő önmaga transzponáltjával, így:

$$D^2[a^T Z] = \mathrm{E}\left(a^T Z Z^T a\right) = a^T \mathrm{E}\left(Z Z^T\right) a.$$

A ZZ^T mátrix egy olyan (diadikus) mátrix melynek i,j-edik eleme $(X_{t_i} - \mathrm{E}(X))(X_{t_j} - \mathrm{E}(X))$, így az $\mathrm{E}\left(ZZ^T\right)$ mátrix i,j-edik eleme $\mathrm{E}(X_{t_i} - \mathrm{E}(X))(X_{t_j} - \mathrm{E}(X)) = \gamma(t_i - t_j)$. Tudjuk azonban, hogy mivel $D^2[a^T Z]$ egy valószínűségi változó szórását jelöli ezért nem lehet negatív, így $\forall a \in \mathbb{Z}^n$:

$$0 \leq D^2[a^T Z] = a^T \mathrm{E}\left(ZZ^T\right) a.$$

Vagyis az $\mathrm{E}(ZZ^t)$ mátrix pozitív szemidefinit és így a γ függvény is. ■

[Herglotz] Legyen $X_1, \ldots, X_n, \ldots$ komplex értékű stacioner folyamat $\gamma(h) = \operatorname{cov}(X_t, \bar{X}_{t+h})$ autokovariancia-függvénnyel. Ekkor

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\nu} dF(\nu),$$

ahol $F(\nu)$ a spektrálfüggvény, amelyre igaz, hogy $F(-\pi) = 0$, jobbról folytonos, korlátos.

12.2 Idősorok transzformációja

Klasszikus dekompozíció:

$$X_t = m_t + \Delta_t + Y_t,$$

ahol

m_t**:** trend, lassan változó determinisztikus,

Δ_t**:** szezonalitás, periodikus függvény,

Y_t**:** stacionárius folyamat.

12.2.1 Nincs periodikus komponens

Kiindulunk egy ismert trendfüggvényből: $m_t = a + bt$. A minták alapján a-t, és b-t úgy határozzuk meg, hogy a $\sum_{t=1}^{n}[X_t - (a+bt)]^2$ négyzetes eltérés minimális legyen.

Mozgó átlagos simítás (moving average smoothing)

$W_t = \frac{1}{2q+1}\sum_{j=-q}^{q} X_{t+j}$ a simított folyamat.

$$\begin{aligned} X_t &\rightarrow \frac{1}{2q+1}\sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1}\sum_{j=-q}^{q} Y_{t+j} \\ \frac{1}{2q+1}\sum_{j=-q}^{q} Y_{t+j} &\approx 0 \end{aligned}$$

$W_t = \hat{m}_t$ a trend becslése, feltéve, ha az *lineáris*! $\hat{Y}_t = X_t - \hat{m}_t$.

Exponenciális simítás (exponential smoothing)

Legyen $a \in (0,1)$.

$$\begin{aligned} \hat{m}_1 &:= X_1 \\ \hat{m}_t &:= aX_t + (1-a)\hat{m}_{t-1} \\ \hat{m}_t &= \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^t X_1 \end{aligned}$$

Különbségképzés (differencing)

Definiáljuk a különbségképző (differencing, ∇), illetve backward shift (B) operátorokat a következőképpen:

$$B : \Re^{\mathbb{Z}} \mapsto \Re^{\mathbb{Z}}, BX_t = X_{t-1},$$

illetve

$$\nabla : \Re^{\mathbb{Z}} \mapsto \Re^{\mathbb{Z}}, \nabla = 1 - B$$
$$\nabla X_t = (1-B)X_t = X_t - X_{t-1}$$

Ezek felhasználásával: $\nabla(at+b) = a$, illetve lineáris $m_t(= at+b)$ trend estén $\nabla(m_t + Y_t) = a + \nabla Y_t$.

Hasonlóképp definiálhatjuk a ∇^k operátort is:

$$\nabla^k = (1-B)^k = 1 - \binom{k}{1}B + \binom{k}{2}B^2 + \cdots + \binom{k}{k}(-B)^k.$$

Például : $\nabla^2 X_h = X_h - 2X_{h-1} + X_{h-2}$, illetve: $\nabla^k(a_k t^k + \cdots + a_0) = k! a_k$.

12.2.2 Trend és szezonalitás

Lassú trend

Legyen

$$X_{j,k} = Y_{k+12(j-1)},$$

ahol j jelentheti pl. az évet és k a hónapot. Feltesszük, hogy egy éven belül a trend konstans: m_j.

$$\hat{m}_j := \frac{1}{12}\sum_{k=1}^{12} X_{j,k}.$$

Szezonalitás becslése:

$$\hat{s}_k := \frac{1}{20}\sum_{j=1}^{20}(X_{j,k} - \hat{m}_j).$$

$$\sum_{k=1}^{12} \hat{s}_k = 0$$

A szezonalitás periódusát ismernünk kell! Ennek meghatározásához használhatjuk például a periodogram módszert.

Mozgó átlagos simítás

Trend ($\hat{m}_t$) becslése:

$$\hat{m}_t = \begin{cases} \frac{1}{2q+1}\sum_{i=-q}^{q} X_{t+i} & : d = 2q+1 \\ \frac{1}{2q}(0.5(X_{t-q} + X_{t+q}) + \sum_{i=-q+1}^{q-1} X_{t+i}) & : d = 2q \end{cases}$$

$$\nabla_d X_t := X_t - X_{t-d} = X_t - B^d X_t = (1 - B^d)X_t$$

12.3 Tapasztalati autokovariancia és autokorreláció

Adott egy n elemű minta. Ekkor a *tapasztalati autokovariancia függvény*t a következőképpen definiáljuk :

$$\hat{\gamma}(h) := \frac{1}{n}\sum_{j=1}^{n-h}(X_{j+h} - \bar{X})(X_j - \bar{X}), \tag{12.1}$$

ahol $\bar{X}$ jelöli a mintaátlagot.

Bizonyítható, hogy az így definiált empirikus autokovariancia tagokból képzett $\Sigma = [\hat{\gamma}(i-j)]_{1\leq i,j\leq n}$ mátrix pozitív szemidefinit. Amennyiben a 12.1 egyenletben n helyett $(n-r)$-rel normálnánk, úgy a kapott empirikus autokovariancia mátrixra ez nem teljesülne.

Hasonlóképp értelmezhetjük az *emprikus autokorrelációs függvény*t is:

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

Box és Jenkins ökölszabálya szerint $n \geq 50$ és $h \leq n/4$ esetén van értelme az idősor analízisével foglalkozni.

Fontos megjegyezni, hogy az empirikus autokovariancia ill. autokorreláció függvényeket nem-stacionárius esetben is ki tudjuk számítani, ezért a kapott eredmények értelmezésénél ezt figyelembe kell venni.

Ha például $\hat{\rho}(h)$ függvény lecsengése lassú, hatványfüggvény jellegű, az az erős függés helyett jelentheti lassú determinisztikus trend jelenlétét is.

Amennyiben a folyamatban szezonalitás van jelen ez a $\hat{\rho}(h)$ periodicitását eredményezi.

12.4 Parciális autokovariancia függvény

$$\begin{aligned}\alpha(1) &= \operatorname{cov}(X_1, X_2)\\ \alpha(h) &= \operatorname{cov}(X_1 - \mathrm{E}\left(X_1|X_2,\ldots,X_h\right), X_{h+1} - \mathrm{E}\left(X_{h+1}|X_2,\ldots,X_h\right))\end{aligned}$$

$$\alpha(k) = \frac{\det R_k^*}{\det R_k},$$

ahol R_k az autokorreláció mátrix, $R_k = [\rho(i-j)]_{1\leq i,j\leq k}$, R_k^*-t pedig úgy kapjuk R_k-ból, hogy annak utolsó sorát a $[\rho_1,\ldots,\rho_k]$ vektorra cseréjük.

12.5 Fehér zaj

Fehér zaj *Fehér zaj*nak hívjuk, és $WN(0,\sigma^2)$-tel jelöljük azokat a folyamatokat, melyre

$$\gamma(k) = \begin{cases} \sigma^2, & k = 0\\ 0, & k \neq 0\end{cases}$$

A fehér zaj spektrálfüggvénye konstans.

$$f(\lambda) = \frac{\sigma^2}{2\pi}, \lambda \in [-\pi,\pi]$$

12.6 Mozgóátlag (MA) folyamatok

Mozgó átlag folyamat Akkor nevezünk egy X_t folyamatot *mozgó átlag folyamat*nak ha felírható

$$X_t = \Theta_0 e_t + \Theta_1 e_{t-1} + \cdots + \Theta_q e_{t-q}$$

alakban, ahol $e_t \sim WN(0,1)$ fehér zaj, $\Theta_i, 0 \leq i \leq q$ pedig konstansok.

X_t és e_t közötti összefüggést felírhatjuk a B backshift operátor segítséégével:

$$\begin{aligned} X_t &= \Theta_0 e_t + \Theta_1 B e_t + \cdots + \Theta_q B^q e_t \\ &= (\Theta_0 + \Theta_1 B + \cdots + \Theta_q B^q) e_t \end{aligned}$$

Formálisan definiálhatjuk a $\Theta(z)$ polinomot a következőképp: $\Theta(z) := \Theta_0 + \Theta_1 z + \cdots + \Theta_q z^q$. Ezzel a jelöléssel:

$$X_t = \Theta(B) e_t$$

Az algebra alaptétele szerint egy n változós polinomnak pontosan n darab nem feltétlenül különböző gyöke van. Ennek alapján felírhatjuk $\Theta(z)$ gyöktényezős alakját:

$$\Theta(z) = \Theta_q \prod_{i=1}^{q} (z - z_i).$$

$$\mathrm{E}\left(X_r e_s\right) = \begin{cases} \Theta_{r-s} & : r - q \leq s \leq r \\ 0 & : \text{különben} \end{cases}$$

$$\mathrm{E}\left(X_{t+k} X_t\right) = \mathrm{E}\left(X_{t+k} \sum_{i=0}^{q} \Theta_i e_{t-i}\right) = \sum_{i=0}^{q} \Theta_i E[X_{t+k} e_{t-i}] = \sum_{i=0}^{q-k} \Theta_i \Theta_{k+i}$$

Mivel $\mathrm{E}\left(X_{t+k} X_t\right)$ t-től nem, csak k-tól függ, ezért X_t gyengén stacionárius, és az autokovariancia függvénye:

$$\gamma(k) = \begin{cases} \sum_{i=0}^{q-k} \Theta_i \Theta_{k+i} & : k \leq q \\ 0 & : k > q \end{cases} \tag{12.2}$$

Az előző egyenletben kifejeztük $\gamma(k)$ értékét a Θ_i értékek segítségével. Felvetődik a kérdés, hogy lehet-e ugyanezt visszefelé, illetve mi annak a feltétele, hogy adott $\gamma(k)$, $k = 0{,}1{,}2 \ldots, q$ autokovariancia függvényhez létezzenek a 12.2 egyenletet kielégítő Θ_i értékek.

Vagyis a kérdés: adott $\gamma(k)$, $k = 0{,}1{,}2, \ldots, q$ esetén megoldható-e $\gamma(k) = b_0 b_k + b_1 b_{k+1} + \ldots + b_{g-k} b_q$, $k = 0{,}1{,}2, \ldots, q$ egyenletrendszer.

A válasz pedig, hogy a megoldhatóság feltétele, hogy a

$$\Gamma(s) := \gamma(0) + \sum_{k=1}^{q} \gamma(k)(s^k + s^{-k})$$

függvénynek az az $|s| = 1$ egységkörön csak páros multiplicitású gyökei legyenek.

12.7 Autoregresszív (AR) folyamatok

Autoregresszív folyamatok Akkor nevezünk egy (X_t) folyamatot *autoregresszív*nek, ha felírható

$$X_t = \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + e_t$$

alakban, ahol e_t fehér zaj.

A mozgó átlag folyamatoknál definiált Θ függvényhez hasonlóan definiálhatjuk a $\Phi(z) := \Phi_0 + \Phi_1 z + \cdots + \Phi_p z^p$ polinomot. Így

$$\Phi(B) X_t = e_t.$$

12.7.1 Példa, AR(1) folyamat

$$X_t := \Lambda X_{t-1} + e_t = e_t + \Lambda(\Lambda X_{t-2} + e_{t-1}) = \sum_{j=0}^{k} \Lambda^j e_{t-j} + \Lambda^{k+1} X_{t-k-1}$$

$$\left|\left| X_t - \sum_{j=0}^{k} \Lambda^j e_{t-j} \right|\right|^2 = \Lambda^{2(k+1)} \, ||X_{t-k-1}||^2 \to 0$$

ha $|\Lambda| < 1$. Ebben az esetben azt mondjuk, hogy az AR folyamat *kauzális*. Ebben az esetben felírhatjuk az X_t-t

$$X_t = \sum_{j=0}^{\infty} \Lambda^j e_{t-j}$$

alakban is, ami egy MA(∞) folyamatnak felel meg. Ezt nevezzük az AR(1) folyamat kauzális MA(∞) előállításának.

Észrevehetjük, hogy amennyiben az AR(1) folyamat kauzális, azaz létezik MA(∞) előállítása, akkor a

$$\Phi(z) = 1 - \Lambda z$$

polinomnak az egyedüli gyöke az egységkörön kívül helyezkedik el. Általánosságban is igaz a következő

Egy AR(p) folyamatnak akkor és csak akkor létezik kauzális MA(∞) előállítása ha a $\Phi(z) = 0$ egyenletnek nincsen a $|z| \leq 1$ egységkörön belül gyöke.

Tekintsük most azt az esetet amikor $|\Lambda| > 1$!

$$\begin{aligned} X_{t+1} &= \Lambda X_t + e_{t+1} \\ X_t &= \frac{1}{\lambda} X_{t+1} - \frac{1}{\Lambda} e_{t+1} = \\ &= -\sum_{j=1}^{k} \frac{1}{\Lambda^j} e_{t+j} + \frac{1}{\Lambda} x_{t+k} \end{aligned}$$

$|\Lambda| > 1$ esetén

$$X_t = -\sum_{j=1}^{\infty} \frac{1}{\Lambda^j} e_{t+j}$$

Vagyis $|\Lambda| > 1$ esetén is létezik MA(∞) előállítás, ez azonban nem kauzális.

AR(1) folyamatok autokovariancia - függvényét könnyen kifejezhetjük az MA(∞) előállításuk segítségével.

$$\operatorname{cov}(X_{t+h}, X_t) = \lim_{n\to\infty} \mathrm{E}\left(\sum_{j=0}^{n} \Lambda^j e_{t+k-j} \sum_{k=0}^{n} \Lambda^k e_{t-k} \right) = \Lambda^h \sum_{j=0}^{\infty} \Lambda^{2j} = \frac{\Lambda^h}{1-\Lambda^2}$$

12.7.2 Yule-Walker egyenletek

Ezeket az egyenleteket az $AR\,(p)$ folyamatokra tekintjük.

$$\begin{aligned} X_t &= \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + e_t \\ X_{t-k} X_t &= \Phi_1 X_{t-k} X_{t-1} + \cdots + \Phi_p X_{t-k} X_{t-p} + X_{t-k} e_t \\ \gamma(k) = \mathrm{E}\,(X_{t-k} X_t) &= \Phi_1 \gamma(k-1) + \cdots + \Phi_p \gamma(k-p) \\ \rho(k) &= \Phi_1 \rho(k-1) + \cdots + \Phi_p \rho(k-p) \end{aligned}$$

Ezen utóbbi egyenletet $k = 1,2,\ldots,p$ értékekre felírva és mátrix alakba rendezve kapjuk a *Yule-Walker* egyenletrendszert.

$$\begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \cdots & \rho_1 & \rho_0 \end{pmatrix} \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_p \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix}$$

Ennek segítségével találhatunk adott $\rho(1),\ldots,\rho(p)$ értékekhez olyan $\Phi_1,\ldots,\Phi_p$ együtthatókat, hogy a kapott AR(p) folyamat autokorreláció - függvényének első p eleme az adott $\rho(1),\ldots,\rho(p)$-val egyezzen. Ugyanez az egyenlet használható a $\hat{\Phi}_1,\ldots,\hat{\Phi}_p$ értékek becslésére is a $\hat{\rho}_1,\ldots,\hat{\rho}_p$ segítségével.

12.8 Autoregresszív - mozgóátlag (ARMA) folyamatok

Az X_t folyamatot *ARMA(p,q)* folyamatnak nevezzük, ha

$$\Phi(B) X_t = \Theta(B) e_t, \tag{12.3}$$

ahol $\Phi(B)$ p-ed és $\Theta(B)$ q-ad fokú polinomok.

Akkor nevezzük a folyamatot kauzálisnak, ha létezik MA(∞) előállítása. Egy kauzális ARMA folyamat autokovariancia - függvényét az MA(∞) előállításában szereplő együtthatókkal a következő tétel segítségével tudjuk kifejezni:

Amennyiben X_t stacionárius, és $\sum_{j=-\infty}^{\infty} |\Psi_j| < \infty$ akkor az $\eta_t := \sum_{j=-\infty}^{\infty} \Psi_j X_{t-j}$ függvény is stacionárius, és autokovariancia - függvénye:

$$\gamma_\eta(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \Psi_j \Psi_k \gamma_\xi(h-j+k)$$

12.8.1 A kauzalitás szükséges és elégséges feltétele

Tegyük fel, hogy a $\Phi(z)$ és $\Theta(z)$ polinomoknak nincs közös gyöke. Ezt nyugodtan feltehetjük, mivel ellenkező esetben a (12.3) egyenletben ezzel a közös gyökkel egyszerűsíthetünk.

Az X_t ARMA(p,q) folyamat kauzalitásának szükséges és elégséges feltétele, hogy a $\Phi(z) = 0$ egyenletnek ne legyen a $|z| \leq 1$ egységkörön belül gyöke.

$$\sum_{j=0}^{\infty} \Psi_j z^j = \frac{\Theta(z)}{\Phi(z)}, |z| \leq 1$$

Invertálhatóság Az X_t folyamatot *invertálható*nak nevezzük, ha $\exists \Pi_j : \sum_{j=0}^{\infty} |\Pi_j| < \infty$ és $e_t = \sum_{j=0}^{\infty} \Pi_j X_{t-j}$.

Az X_t ARMA(p,q) folyamat invertálhatóságának szükséges és elégséges feltétele, hogy a $\Theta(z) = 0$ egyenletnek ne legyen a $|z| \leq 1$ egységkörön belül gyöke.

$$\sum_{j=0}^{\infty} \Pi_j z^j = \frac{\Phi(z)}{\Theta(z)}, |z| \leq 1$$

12.9 Az átlag és az autokovariancia becslései

$$\mathrm{E}(X_t) = \mu, \mathrm{Cov}(X_t, X_{t+n}) = \gamma(n)$$

Az átlag természetes becslése

$$\bar{X_n} = \frac{X_1 + \ldots X_n}{n} \text{ és erre } \mathrm{E}\left(\bar{X_n}\right) = \mu$$

Az átlag szórásnégyzete

$$nD^2(\bar{X_n}) = \frac{1}{n} \sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j) = \sum_{i,j=1}^{n} \frac{1}{n} \gamma(i-j) =$$

$$= \sum_{h=-n+1}^{n-1} \sum_{j=1}^{n-|h|} \frac{1}{n} \gamma(h) = \sum_{|h|<n} \frac{n-|h|}{n} \gamma(h) \leq \sum_{|h|<n} |\gamma(h)|$$

Ha $\gamma(n) \to 0$, akkor

$$D^2(\bar{X_n}) \leq \sum_{|h|<n} |\gamma(h)| \to 0.$$

Ha pedig még az is igaz, hogy a $\gamma(n)$ sorozat abszolút konvergens, akkor az átlag szórásnégyzetére az alábbi aszimptotikát adhatjuk

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty \Rightarrow nD^2(\bar{X_n}) \to \sum_{h=-\infty}^{\infty} \gamma(h)$$

12.9.1 A spektrálfüggvény és az autokovariancia kapcsolata

[Inverz Fourier transzformált]

Igaz, hogy

$$\sum_{n=-\infty}^{\infty} |K(n)| < \infty \Longrightarrow K(h) = \int_{-\pi}^{\pi} e^{ih\nu} f(\nu) d\nu,$$

ahol

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{in\lambda} K(n)$$

ugyanis

$$\int_{-\pi}^{\pi} e^{ih\nu} f(\nu) d\nu = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} K(n) \int_{-\pi}^{\pi} e^{i(h-n)\nu} = K(h).$$

Ennek egyszerű következménye az alábbi állítás. Egy $\gamma(h)$ sorozatra

$$\begin{array}{c} \sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty \\ \text{és} \\ \gamma \text{ autokovariancia függvény} \end{array} \quad \Leftrightarrow \quad f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda} \gamma(n) \geq 0.$$

Ennek alkalmazásaként számítsuk ki, mikor lesz az alábbi alakú K függvény autokovariancia függvény!

$$K(h) = \begin{cases} 1 \text{ ha } h = 0 \\ \rho \text{ ha } h = \pm 1 \\ 0 \text{ egyébként} \end{cases}$$

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda} K(n) = \frac{1}{2\pi} (1 + 2\rho \cos \lambda) \geq 0 \Rightarrow |\rho| \leq \frac{1}{2}$$

12.9.2 Aszimptotikus normalitás

Legyen például

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j},$$

ahol Z_t független, azonos eloszlású 0 várható értékkel és σ^2 szórásnégyzettel, továbbá

$$\sum_{j=-\infty}^{\infty} |\Psi_j| < \infty \text{ de } \sum_{j=-\infty}^{\infty} \Psi_j \neq 0.$$

Ekkor

$$\bar{X_n} \xrightarrow{d} N\left(\mu, \frac{1}{n} \sum_{j=-\infty}^{\infty} \Psi(n)\right)$$

12.9.3 $\gamma(n)$ becslése

$$\widehat{\gamma(n)} = \frac{1}{n} \sum_{t=1}^{n-h} \left(X_t - \bar{X_n}\right) \left(X_{t+h} - \bar{X_n}\right),$$

ahol $0 \leq h \leq n-1$.

Ez általánosságban torzított becslés (bár bizonyos további feltételek mellett aszimptotikusan torzítatlan), de viszont a $\widehat{\Gamma_n} = \left[\widehat{\gamma(i-j)}\right]_{1 \leq i,j \leq n}$ mátrixa pozitív szemidefinit.

Ennek bizonyításához elég, hogy $\widehat{\Gamma_n} = \frac{1}{n} M M^T$ a következő az $Y_i = X_i - \bar{X_n}$ jelöléssel kifejezett M mátrixszal

$$M = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & Y_1 & Y_2 & \dots & Y_{n-1} & Y_n \\ 0 & 0 & \dots & 0 & Y_1 & Y_2 & Y_3 & \dots & Y_n & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & Y_1 & \dots & Y_{n-2} & Y_{n-1} & Y_n & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Általános ökölszabályként elmondhatjuk, hogy $\rho(h)$ becslése $\left(\widehat{\rho(h)} = \frac{\widehat{\rho(h)}}{\widehat{\rho(0)}}\right)$ akkor jó, ha $n \geq 50$ és $h \leq \frac{n}{4}$.

12.9.4 Az autokorrelációk mikor különböznek szignifikánsan 0*-tól?*

Ha a következő alakú szűrt független zajra

$$X_t - \mu = \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j}$$

és Z_t független, azonos eloszlású 0 várható értékkel és σ^2 szórásnégyzettel, továbbá

$$\sum_{j=-\infty}^{\infty} |\Psi_j| < \infty \text{ és } \mathrm{E}\left(Z_i^4\right) < \infty$$

akkor

$$\left[\widehat{\rho(1)}, \dots \widehat{\rho(h)}\right] \xrightarrow{d} N\left(\left[\rho(1), \dots \rho(h)\right], \frac{1}{n}W\right),$$

ahol W az ún. Bartlett mátrix.

$$W_{i,j} = \sum_{k=1}^{\infty} \left[\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\right]\left[\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\right].$$

Például a független fehér zaj folyamatra $\rho(l) \neq 0$, ha $l \neq 0$, azaz

$$W_{i,j} = \begin{cases} 1 \text{ ha } i = j \\ 0 \text{ha } i \neq j \end{cases}$$

azaz

$$\widehat{\rho(1)}, \dots, \widehat{\rho(h)} \approx \text{iid } N\left(0, \frac{1}{n}\right),$$

ennek konfidenciaintervalluma $\pm 1.96\frac{1}{\sqrt{n}}$, amely értéket a normális eloszlás táblázatából olvashatunk ki.

12.10 $ARMA$ modellek becslései

Amikor egy folyamatot $ARMA$ modellel közelítünk, a következő lépések szerint járunk el:

1. megbecsüljük p-t és q-t, az $ARMA$ folyamathoz tartozó két polinom fokszámát

2. megbecsüljük a polinomok együtthatóit
3. megbecsüljük a szórásnégyzetet

12.10.1 Ismert p és q

Tiszta autoregresszív esetben felírhatjuk a Yule-Walker egyenleteket:

$$X_t - \Phi_1 X_{t-1} - \cdots - \Phi_p X_{t-p} = e_t \text{ ahol } e_t \sim WN(0, \sigma^2)$$

$$\Gamma_p \Phi = \gamma_p,$$

ahol

$$\Gamma_p = [\gamma(i-j)]_{i,j=1}^{p}$$
$$\gamma_p^T = [\gamma(1), \ldots, \gamma(p)]$$
$$\Phi^T = [\Phi(1), \ldots, \Phi(p)].$$

Továbbá

$$\sigma^2 = D^2 e_t = D^2(X_t - \Phi_1 X_{t-1} - \cdots - \Phi_p X_{t-p}) = \gamma(0) - \Phi^T \gamma_p.$$

Így a Yule-Walker becslések a következő alakúak lesznek:

$$\widehat{\Gamma_p}\widehat{\Phi} = \widehat{\gamma_p}$$

és

$$\widehat{\sigma}^2 = \widehat{\gamma(0)} - \widehat{\Phi}^T \widehat{\gamma_p}.$$

12.10.2 Ismeretlen p

Ha p nem ismert és $AR(m)$-et próbálunk illeszteni, akkor azt várjuk, hogy $\widehat{\Phi_{m,m}}$ kicsi lesz.

$$\sqrt{n}\left(\widehat{\Phi_m} - \Phi_m\right) \xrightarrow{d} N_m(0, \sigma^2 \Gamma_m^{-1}),$$

ahol Φ_m az X_{m+1} legjobb lineáris közelítésének együtthatóvektora:

$$\left\| X_{m+1} - \Phi_m^T (X_1, \ldots, X_m) \right\| \to min.$$

12.10.3 A Durbin-Levinson algoritmus

A mátrixinvertálás kikerülésére a Durbin-Levinson algoritmust használjuk. Legyenek az m-ed rendű illesztés együtthatói

$$\widehat{\Phi}_m = \left(\widehat{\Phi}_{m,1}, \widehat{\Phi}_{m,2}, \ldots \widehat{\Phi}_{m,m}\right) = \widehat{R}_m^{-1} \widehat{\rho}_m$$

és

$$\widehat{v}_m = \widehat{\gamma}(0)\left[1 - \widehat{\rho}_m^T \widehat{R}_m^{-1} \widehat{\rho}_m\right].$$

Ekkor $\widehat{\Phi}_{1,1} = \widehat{\rho}(1)$ és $\widehat{v}_1 = \widehat{\gamma}(0)\left[1 - \widehat{\rho}^2(1)\right]$. Továbbá a becsült parciális autokovariancia függvény

$$\widehat{\Phi}_{m,m} = \left[\widehat{\gamma}(m) - \sum_{j=1}^{m-1} \widehat{\Phi}_{m-1,j}\widehat{\gamma}(m-j)\right] / \widehat{v}_{m-1}$$

$$\begin{bmatrix} \widehat{\Phi}_{m,1} \\ \vdots \\ \widehat{\Phi}_{m,m-1} \end{bmatrix} = \widehat{\Phi}_{m-1} - \widehat{\Phi}_{m,m} \begin{bmatrix} \widehat{\Phi}_{m-1,m-1} \\ \vdots \\ \widehat{\Phi}_{m-1,1} \end{bmatrix}$$

és

$$\widehat{v}_m = \widehat{v}_{m-1}\left(1 - \widehat{\Phi}^2_{m,m}\right).$$

Elméletileg $\alpha(m) = \Phi_{m,m} = 0$, ha $m > p$, gyakorlatilag $\sqrt{n}\widehat{\Phi}_{m,m} \to N(0,1)$, azaz $P\left(-1.96\frac{1}{\sqrt{n}} < \widehat{\Phi}_{m,m} < 1.96\frac{1}{\sqrt{n}}\right) = 0.95$. A rendre ezzel előzetes becslést adhatunk: $\widehat{p} = \min\left\{r : \forall m > r \;\; |\widehat{\Phi}_{m,m}| < 1.96\frac{1}{\sqrt{n}}\right\}$.

12.10.4 Az innovációs algoritmus

A Gram-Schmidt ortogonalizációs eljárással független vektorokból ortogonális rendszert készíthetünk vetítésekkel. Az eljárást idősorokra is alkalmazhatjuk a következő módon:

$$\mathrm{E}(X)_t = 0 \text{ és } \kappa(i,j) := \mathrm{E}(X_i X_j)$$

$$H_n := \langle X_1, \ldots, X_n\rangle = \langle X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n\rangle \text{ ahol } \hat{X}_{n+1} := pr_{H_n} X_{n+1}.$$

$$\hat{X}_{n+1} = \begin{cases} 0, & n = 0 \\ \sum_{j=0}^{n} \Theta_{n,j}\left(X_{n-j+1} - \hat{X}_{n-j+1}\right), & n \neq 0 \end{cases}.$$

A Θ együtthatók rekurzív kiszámítását a v segédváltozóval (szórásnégyzet) a következő rendszer adja meg:

$$v_n := \left\| X_{n+1} - \hat{X}_{n+1} \right\|^2$$

így

$$v_0 = \kappa(1,1)$$

$$\Theta_{n,n-k} = \frac{1}{v_k}\left[\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \Theta_{k,k-j}\Theta_{n,n-j}v_j\right] \text{ ahol } k = 0..n-1$$

$$v_n = \kappa(n+1,n+1) - \sum_{j=0}^{n-1} \Theta^2_{n,n-j}v_j$$

Ennek bizonyítását úgy kezdjük, hogy $0 \leq k \leq n$ esetén $X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n$ ortogonális, hiszen $X_i - \hat{X}_i \in H_{j-1}$, ha $i < j$, és $X_j - \hat{X}_j \perp H_{j-1}$. Tehát $\hat{X}_{n+1}$ definícióját használva

$$\langle \hat{X}_{n+1}, X_{k+1} - \hat{X}_{k+1} \rangle = \Theta_{n,n-k} v_k,$$

amely egyenlethez a $X_{n+1} - \hat{X}_{n+1} \perp X_{k+1} - \hat{X}_{k+1}$ azonosságot adva kapjuk, hogy

$$\langle X_{n+1}, X_{k+1} - \hat{X}_{k+1} \rangle = \Theta_{n,n-k} v_k.$$

Ebbe $\hat{X}_{k+1}$ defnícióját írva

$$\Theta_{n,n-k} = \frac{1}{v_k} \langle X_{n+1}, X_{k+1} - \sum_{j=0}^{k-1} \Theta_{k,k-j} \left(X_{j+1} - \hat{X}_j + 1 \right) \rangle =$$

$$\frac{1}{v_k} \left[\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \Theta_{k,k-j} \Theta_{n,n-j} v_j \right]$$

továbbá

$$v_n = \left| X_{n+1} - \hat{X}_{n+1} \right|^2 = |X_{n+1}|^2 + \left| \hat{X}_{n+1} \right|^2 =$$

$$\kappa(n+1, k+1) - \sum_{j=0}^{n-1} \Theta_{n,n-j}^2 v_j$$

Például a MA(1) folyamat predikcióját így adhatjuk meg:

$$X_t = Z_t + \Theta Z_{t-1} \text{ ahol} Z_t \sim WN(0, \sigma^2)$$

$$\kappa(i,j) = \begin{cases} \sigma^2(1+\Theta^2) & i = j \\ \Theta\sigma^2 & i = j+1 \\ 0 & \text{egyébként} \end{cases}$$

$$v_0 = \kappa(1,1) = \sigma^2(1+\Theta^2),$$

$$\Theta_{n,j} = \begin{cases} \frac{1}{v_{n-1}} \Theta\sigma^2 & j = 1 \\ 0 & 2 \le j \le n \end{cases}$$

$$v_n = (1+\Theta^2)\sigma^2 - \frac{1}{v_{n-1}^2} \Theta^2 \sigma^4,$$

$$r_n := \frac{v_n}{\sigma^2} = (1+\Theta^2) - \frac{1}{v_{n-1}^2} \Theta^2 \sigma^2.$$

Tehát a predikció:

$$\hat{X}_{n+1} = \frac{\Theta}{r_{n-1}} (X_n - \hat{X}_n)$$

12.10.5 Mozgóátlag folyamatok becslései

Az $X_1, \ldots, X_n$ adatokra a következő előfeltevést tesszük annak analógiájára, hogy az $X\hat{X}$ mennyiségek voltak a hibák:

$$X_t = Z_t + \hat{\Theta}_{m,1} Z_{t-1} + \cdots + \hat{\Theta}_{m,m} Z_{t-m} \text{ ahol } Z_t \sim WN(0, \hat{v}_m^2).$$

Ha $\hat{\gamma}(0) > 0$, akkor vezessük be a becsült együtthatók vektorára a $\hat{\Theta}_m = \left(\hat{\Theta}_{m,1}, \ldots, \hat{\Theta}_{m,m}\right)$ jelölést! Ezekre a következő rekurzív becslés érvényes:

$$\hat{v}_0 = \hat{\gamma}(0)$$

és

$$\hat{\Theta}_{m,m-k} = \hat{v}_k^{-1} \left[\hat{\gamma}(m-k) - \sum_{j=0}^{k-1} \hat{\Theta}_{m,m-j} \hat{\Theta}_{k,k-j} \hat{v}_j \right]$$

$$\hat{v}_m = \hat{\gamma}(0) - \sum_{j=0}^{k-1} \hat{\Theta}_{m,m-j}^2 \hat{v}_j \text{ ahol } k = 0, \ldots, m-1$$

12.10.6 Aszimptotikus viselkedés ARMA folyamatok esetén

A jelölések rövid leírása a következő:

$$\Phi(B) X_t = \Theta(B) Z_t \text{ ahol } Z_t \sim IID(0, \sigma^2) \text{ és } \mathrm{E}\left(Z_t^4\right) < \infty$$

$$\Psi(z) = \sum_{j=0}^{\infty} \frac{\Theta(z)}{\Phi(z)}, |z| \leq 1, \Psi_0 = 1$$

Ekkor minden k-ra

$$\sqrt{n} \left[\hat{\Theta}_{m,1} - \Psi_1, \ldots, \hat{\Theta}_{m,k} - \Psi_k \right] \xrightarrow{d} N(0, A)$$

ahol

$$A_{i,j} = \sum_{r=1}^{min(i,j)} \Psi_{i-r} \Psi_{j-r}$$

továbbá

$$m(n) \to \infty$$

úgy hogy

$$m(n) = o(\sqrt[3]{n}) \text{ és } \hat{v}_m \xrightarrow{p} \sigma^2.$$

Itt jegyezzük meg, hogy $AR(p)$ esetben a Durbin-Levinson algoritmus által a Φ_p-re adott $\hat{\Phi}_p = \left(\hat{\Phi}_{p,1}, \ldots, \hat{\Phi}_{p,p}\right)$ becslés konzisztens, ha $n \to \infty$. Viszont $MA(k)$ esetben az innovációs algoritmus által adott $\hat{\Theta}_q = \left(\hat{\Theta}_{q,1}, \ldots, \hat{\Theta}_{q,q}\right)$ becslés nem konzisztens, viszont a $\left(\hat{\Theta}_{m,1}, \ldots, \hat{\Theta}_{m,q}\right)$ már az.

A gyakorlatban $MA(q)$ esetben tudjuk, hogy $\rho(m) = 0$, ha $m > q$, és Bartlett tétele miatt

$$\hat{\rho}(m) \xrightarrow{d} N\left(0, \frac{1}{n}\sum_{i=-q}^{n-q}\rho(i)\right).$$

12.10.7 Maximum likelihood becslések

$E(X_t) = 0$ tulajdonságú Gauss folyamat esetén a $\Gamma_n = E\left(\underline{X}_n\underline{X}_n^T\right)$ jelöléssel a likelihood függvény a következő:

$$L(\Gamma_n) = \frac{1}{(2\pi)^{n/2}}\frac{1}{(\det\Gamma_n)^{1/2}}\exp\left(-\frac{1}{2}\underline{X}_n^T\Gamma_n^{-1}\underline{X}_n\right).$$

A mátrixinvertálás és a determinánsszámítás kikerülésére a következő algoritmus javasolt: $k = 0, \ldots, n-1$ esetén

$$\hat{X}_{k+1} = \sum_{j=0}^{k-1}\Theta_{k,k-j}\left(X_{j+1} - \hat{X}_{j+1}\right)$$

azaz

$$\begin{pmatrix}\hat{X}_1\\ \vdots\\ \hat{X}_n\end{pmatrix} = \begin{pmatrix}0 & 0 & 0 & \ldots & 0\\ \Theta_{1,1} & 0 & 0 & \ldots & 0\\ \Theta_{2,2} & \Theta_{2,1} & 0 & \ldots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \Theta_{n-1,n-1} & \Theta_{n-1,n-2} & \Theta_{n-1,n-3} & \ldots & 0\end{pmatrix}\begin{pmatrix}X_1 - \hat{X}_1\\ \vdots\\ X_n - \hat{X}_n\end{pmatrix}.$$

A képletben szereplő mátrixot jelöljük a következőképpen

$$\bar{C} = [\Theta_{i,i-j}]_{i,j=0}^{n-1} \text{ ahol } j \le 0 \text{ esetén } \Theta_{i,j} = 0.$$

Ezt a mátrisot módosítsuk úgy, hogy a főátlóba 1-eket írunk: $C := \bar{C} + Id$. Ekkor

$$C\left(\underline{X}_n - \underline{\hat{X}}_n\right) = (\bar{C} + Id)\left(\underline{X}_n - \underline{\hat{X}}_n\right) = \underline{\hat{X}}_n + \underline{X}_n - \underline{\hat{X}}_n = \underline{X}_n$$

azaz

$$\Gamma_n = \mathrm{E}\left(X_nX_n^T\right) = C\mathrm{E}\left(\left(\underline{X}_n - \underline{\hat{X}}_n\right)\left(\underline{X}_n - \underline{\hat{X}}_n\right)^T\right)C^T = CDC^T$$

ahol

$$D = Diag(v_0, v_1, \ldots, v_{n-1})$$

ezért a determináns egyszerűen számítható.

$$\det\Gamma_n = (\det C)^2\det D = v_0v_1\ldots v_{n-1}.$$

A kitevő is egyszerűbb alakra hozható:

$$\underline{X}_n^T \Gamma_n^{-1} \underline{X}_n = \left(\underline{X}_n - \underline{\hat{X}}_n\right)^T C^T \Gamma^{-1} C \left(\underline{X}_n - \underline{\hat{X}}_n\right) =$$

$$\left(\underline{X}_n - \underline{\hat{X}}_n\right)^T C^T C^{T-1} D^{-1} C^{-1} C \left(\underline{X}_n - \underline{\hat{X}}_n\right) =$$

$$\left(\underline{X}_n - \underline{\hat{X}}_n\right)^T D^{-1} \left(\underline{X}_n - \underline{\hat{X}}_n\right) = \sum_{j=1}^{n} \frac{\left(X_j - \hat{X}_j\right)^2}{v_{j-1}}$$

Tehát végeredményképpen a likelihood függvény legegyszerűbb alakja:

$$L(\Gamma_n) = (2\pi)^{-n/2} (v_0 v_1 \dots v_{n-1})^{-1/2} \exp\left[-\frac{1}{2} \sum_{j=1}^{n} \frac{\left(X_j - \hat{X}_j\right)^2}{v_{j-1}}\right].$$