①

Műszlab

Statisztika

Mai értelemben vett statisztika a 20. szd. elején indult.

Volt egy téma: agrometria - ez volt az egyik, ami motiválta a statisztika fejlesztését

50-es, 60-as, 70-es években → bizonyítékokon alapuló orvoslás

→ működik-e a terápia? (nem tudjuk, hogy miért, azt azonban tudni akarjuk, hogy egyáltalán működik-e)

pl: gyógyszeripar - hat-e a gyógyszer; van-e mellékhatása?

(gyógyszeripar költséges - 10 év, míg kifejlesztenek egy gyógyszert)

Még egy motiváció: élettani / kórélettani modellalkotás kérdései

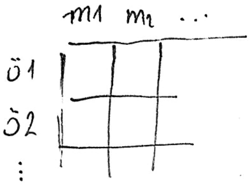
→ nem tudjuk, hogy miképpen kellene működnie az embernek

→ változók közötti összefüggések leírása → statisztika

Kísérlet tervezés

~~Pl.~~ 50-es években agrometriai problémák

két változónk van: pl több műtrágya és több öntözési stílus



m_i - i-edik műtrágya

$ö_i$ - i-edik öntözési stílus

Kérdés lehet például: lehet-e ~~[illegible]~~ választani legjobb műtrágyát az öntözési stílustól függetlenül

Teljes faktoriális kísérleti elrendezés: mindent párosítunk mindennel

Probléma: a paraméterek számának növekedésével kombinatorikusan nő ~~[illegible]~~ párosítások száma →

lehetetlen lesz mindegyiket megvalósítani

Kérdés: akkor melyiket valósítsuk meg?

→ Ez a kísérleti kérdés egyik témaköre

Másik kérdéskör:

A klinikai kísérletek kiindításakor egy csomó kérdés merült fel.

Kísérletek két típusa:

- <u>obszervacionális kísérletek</u> – a vizsgált alanyt nem tesszük ki semmilyen hatásnak, csak megfigyeljük
  epidemiológia tipikus vizsgálati módszere
  pl: dohányzás káros-e az egészségre?
  tipikus kérdés: egy adott környezeti hatásnak van-e valamilyen egészségügyi hatása
- <u>experimentális</u> – beavatkozásos kísérlet

Alapfogalmak:

kohorsz: a szó jelentése: egy nemzedék – tehát olyan emberek részcsoportja, akik valamilyen jellemzőjük szerint vannak csoportosítva (pl. egy évben születettek a közötti korcsoport)

<u>prospektív kohorsz vizsgálat</u>

pl. 2000 egészséges embert vizsgálok – két egyenlő csoportjuk van – az egyik ki van téve valamilyen hatásnak, a másik nincsen
– 10 vagy 20 év múlva megnézzük, hogy mi van

Nagy értékű vizsgálat, ha nagy az elemszám, és gondos volt a válogatás, hogy a két csoport között az egyetlen különbség a vizsgált hatás legyen

Hátránya : macerás
- gondosan kell kivitelezni
- rettentő sokat kell rá várni

Case control - eset kontroll

gyengébb értékű, mint a kohorsz, de rugalmasabb és olcsóbb

nyilvántartások vizsgálatával megoldható

gyors

Olyanokat vizsgálok meg, akiknél kialakult az effektus

eset (case) → akiknél kialakult az effektus, a vizsgált betegség

ezen kívül keresek olyanokat, akik a lehető legnagyobb mértékben hasonlítanak a megbetegedettekre, csak nem betegek

Ez egy sokkal gyengébb vizsgálati módszer

→ kevésbé lehet jól megcsinálni

de néha muszáj ezt alkalmazni

pl: ritka betegségnél kezelhetetlen a kohorszt vizsgálat

Retrospektív kohorsz

Ez is nyilvántartások vizsgálatával dolgozik

Gondolatban visszamegyek 10 évvel ezelőttre

→ ott keresek kereső embereket, akik ki voltak téve a hatásnak, és olyanokat, akik nem, de ~~ettől eltekintve~~ ezt kivéve a legnagyobb mértékig hasonlítanak

→ Megnézem, hogy mi van velük a mában ...

Experimentációs vizsgálatok

→ az expozíciót én alkalmazom (pl. gyógyszer beadása)

→ 1000 ember kap gyógyszert / 1000 ember nem kap

Az első ilyen vizsgálat: 18. szd. → skorbut a hajó fedélzetén mivel gyógyítható

100 skorbutos matróz – 10-es csoportokra lett osztva

→ mindegyik kapott egy tippet

– amelyik C vitamint kapott, az lett jobban

Kérdés: hogyan döntsük el azt, hogy ki kap kezelést és ki nem?

A legjobb módszer a számítógéppel teljesen véletlenszerű válogatás – randomizáció

Nem jó, ha csak kezelt csoport van – kell kontroll is, akit nem kezelünk

DE: figyelembe kell venni a placebo hatást is

Placebo kontrollálás: mindegyik csoport kap valamit, de a kontroll hatás nélküli szert

Fontos: az orvosnak sem szabad tudnia azt, hogy mit ad → kettős vak kísérlet

Hármas vak kísérlet → a biostatisztikus az elemzéshez csak azt tudja, A v. B – persze ez extrém megvalósítás – a kettős vak az általános

A randomizált, placebo kontrollált, kettős vak vizsgálat nem mindig kivitelezhető

pl: vakbélműtétet ellenőrzését nem álszeni így csinálni

<u>Confounding variables</u> - olyan tényezők, amik hatnak a végeredményre, mégsem kezeljük őket

Pl: tűzoltók - ha kimennek → minél több tűzoltó volt kint, annál nagyobb lett a kár

confounding változó: tűzet mérete

Bizonyos esetekben ez nem nyilvánvaló

Pl. ABC - abortusz - mellrák hipotézis

→ abortusz megnöveli a mellrák valószínűségét

ezt a következtetést case controll vizsgálatokkal hozták ki

→ a mellrákosok között több volt, akinek előtte abortusza volt

viszont kohorsz vizsgálattal ez nem jött ki

Kiderült, hogy a case controllal az abortuszról nem dokumentummal győződtek meg, hanem kérdezés alapján

→ a mellrákosok kevésbé titkolták el az abortuszt, vagy kevésbé feledkeztek meg róla

<u>Valószínűség számítás</u>

Véletlen tömegjelenségekkel foglalkozó tudomány

Véletlen kísérlet: pl: kockadobás

→ sokszor megismételhető, de nem tudjuk, hogy mi lesz a kimenetele

$\Omega = \{ w_1, w_2, w_3, \dots \}$

w_i → lehetséges kimenetelei a kísérletnek - <u>elemi esemény</u>

Kockadobás:

$\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$ — diszkrét, véges

Duna vízállása

$\Omega = [0, 1000]$ végtelen eseménytér

Ω – az <u>eseménytér</u>

A kísérlet végrehajtásakor meg tudjuk állapítani, hogy az eseménytér melyik eleme következett be.

<u>Esemény</u>: az eseménytér egy részhalmaza

pl: $A = \{$⚀, ⚂, ⚄$\}$ – páratlan számú jött ki

Akkor mondom, hogy az esemény bekövetkezett, ha olyan elemi esemény jött ki, ami része az A halmaznak.

$A = \Omega$ → <u>biztos esemény</u>

$A = \emptyset$ → <u>lehetetlen esemény</u>

$A \cup B$ – események <u>összege</u>

$A \cap B$ – <u>események szorzata</u>

<u>Valószínűség</u>: egy függvény

$\mathbb{P} : \Omega \to [0, 1]$

az egyes elemi eseményekhez hozzárendel egy 0 és 1 közé eső számot

Bővebben:

$P : \mathcal{F} \to [0, 1]$ ~~elemi~~

$\mathcal{F}$: megfigyelhető események halmaza

lehetnek olyan ~~elemi~~ események, amelyhez nem akarunk valószínűséget rendelni

Ki kell kötnünk, hogy:

a, $\mathbb{P}(\Omega) = 1$

b, ha $A \cdot B = \emptyset$ → kizáró események

$\Rightarrow$ $\mathbb{P}(A+B) = \mathbb{P}(A) + \mathbb{P}(B)$ – szigma additivitás

Ezt nemcsak két, hanem több tagra is meg kell követelni

$(\Omega, \mathcal{F}, \mathbb{P})$ → ezt együtt Kolmogorov-féle valószínűségi mezőnek nevezzük

Rendelünk számokat az egyes elemi eseményekhez

→ valószínűségi változó – ez egy függvény

$X : \Omega \to \mathbb{R}$ pl.: ⚂ → 3

pl.: „768 cm a Duna vízállása" → 768

mögötte van egy természetes leképezés

Pénzérme dobási kísérletek

$r_n = \frac{k_n}{n}$ n: összes dobás

k_n: n-ből mennyi volt fej

→ valószínűség frekvencionista értelmezése

→ az a szám, ami körül r_n ingadozik

$\mathbb{P}(A) = \frac{1}{2}$

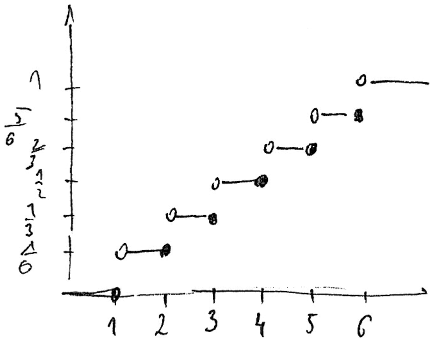
Eloszlás CDF - cumulative distribution function

$F_X(x)$ X valószínűségi változó eloszlásfüggvénye

$\mathbb{R} \to \mathbb{R}$ x az argumentuma

$F_X(x) = \mathbb{P}\{X < x\}$

pl. kockadobás. $F_X(1) = 0$; $F_X(2,5) = \frac{1}{3}$; $F_X(3,5) = \frac{1}{2}$; $F_X(12) = -1$



Az USA-beli értelmezés:

$$F_X(x) = \mathbb{P}\{ X < x)$$

Folytonos valószínűségi változó

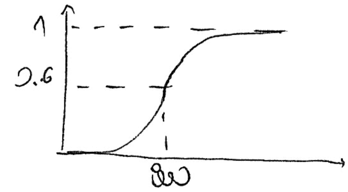
valószínűségi súlyfv. nem létezik

pl. annak valószínűsége, hogy a Duna vízállása egyetlen konkrét érték pontosan, az 0

viszont az eloszlás fv. létezik folytonos esetben is

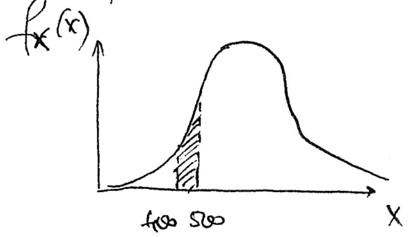
u. ezzel a definícióval

pl. Duna vízállása



Valószínűségi sűrűség függvény PDF - probability density function

pl. Duna vízállása



400 és 500 között a görbe alatti terület

$\to \mathbb{P}(400 \le X \le 500)$

$$\int_a^b f_X(x)\,dx = F_X(b) - F_X(a)$$

$$f_X(x) = F'_X(x)$$

a sűrűség függvény szemléletesebb, mint mint az eloszlás fv.

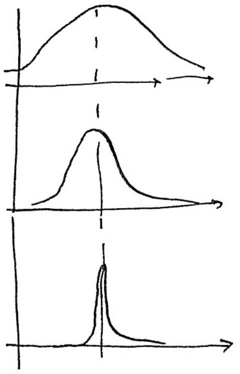
Várható érték

$\mathbb{E}X = \sum_{i_\omega} i\,p_i$ diszkrét esetben kockadobásnál $\mathbb{E}X = 3.5$

$\mathbb{E}X = \int_{-\infty} x\, f_X(x)\, dx$

Módusz - sűrűség fv. maximumának a helye

Ugyanakkora várható értéknél az eloszlás fv. különböző lehet:



Szórásnégyzet : $\mathbb{D}^2 X = \mathbb{E}\left((X-\mathbb{E}X)^2\right)$

Normális eloszlás - Gauss-görbe - haranggörbe

$N(\mu, \sigma^2)$

μ - várható érték, σ^2 - szórásnégyzet

Deskriptív statisztika

Van egy sokaságunk →
általában annak csak egy részhalmazáról van információ

→ minta
— mintavételes helyzet

Sokszor nehéz minden elemet megfigyelni
sőt a legtöbb esetben technikai okokból szinte lehetetlen

Teljeskörű megfigyelés nagyon ritka
→ esetleg ilyen a népszámlálás

A másik ok lehet, hogy a sokaság végtelen
ilyenek a fiktív sokaságok
pl. vérnyomáscsökkentő gyógyszer hatásossága – az, hogy kiket kell a gyógyszerrel kezelni végtelen...

Deskriptív statisztika
→ nem törődünk azzal, hogy a rendelkezésünkre álló adatok csak mintából

Egyváltozós / többváltozós vizsgálat

Egyváltozós vizsgálat – az alanyoknak csak egy adatát vizsgálom meg
amíg a változókat külön-külön vizsgálom egyváltozós vizsgálatról beszélünk

Többváltozós vizsgálat – a változók közötti kapcsolatot vizsgáljuk

Táblázatos formában:
változók oszlopokban
sorokban a megfigyelési egységek

Négy mérési skála
→ azt jelzi, hogy milyen műveleteknek van értelme az adott változókkal

1) nominális (pl. szemszín)
nincsen sorrend, csak különbözőség

2) ordinális (pl. iskolai végzettség szintje)
sorba lehet rakni az elemeket – kisebb-nagyobb relációt lehet kialakítani...

3) intervallum (pl. hőmérséklet °C-ban)

van értelme a különbségnek, de nincsen értelme
a hányadosnak

4) arány (pl. testmagasság)
van 0 pontja

1) 2) → kvalitatív adatok → nem valós skálán mennek
kategoriális adatok

3) 4) → mennyiségi skálán mért ismérvek

Egyváltozós nominális adatok:
egy dolgot lehet tenni:
gyakoriság / relatív gyakoriság

Ordinális adatoknál lehet u.azt, mint a nominálisoknál
+ még egy: kumulált gyakoriság

Az 1) → 2) → 3) → 4) sorrendben haladva az előző
típus műveleteit el lehet végezni, de jönnek
be újabbak

Ezek mindig ~~a~~ adattömörítések → a sok adatról
mondunk ~~egy v. több~~ kevés (egy v. több)
jellemző adatot
mindig információvesztéssel jár
(trade off → valamit valamiért)

Skálán mért adatok esetében
módszerek kétféle típusa:
- grafikus
- analitikus

Grafikus módszerek:

Hisztogram – gyakorisági sor

a változó által felvehető értékeket felosztjuk tartományokra

– oszlopokkal ábrázoljuk, hogy az egyes tartományokba mennyi elem tartozik

Probléma: le kell lőni az oszlopok szélességét

Általában equidisztáns intervallumokat veszek fel

Ezzel a módszerrel empirikusan tudjuk közelíteni az eloszlás függvényt

Empirikus sűrűségfüggvényt azonban definiálni így nem lehet..

Magfüggvény becslés → legyalulja a hisztogram lépcsőit

→ nagyon pici normális eloszlásoknak az összege

Itt is van paraméter – a kis Gauss-függvények szórását lehet állítani

Analitikus módszerek:

Átlag pl. $\overline{TM} = \frac{1}{N} \sum_{i=1}^{N} TM_i$

a várható értéket akarjuk vele becsülni

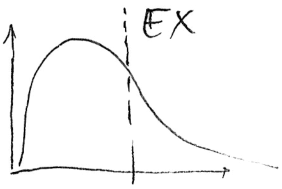
előnye: jól interpretálható

de: nem egy robusztus mutató

pl: a legépelésnél elütünk egy nullát → teljesen elviszi az átlag értékét

→ érzékeny az outlier-ekre (kiugró adatokra)

A másik probléma a ferde vagy hosszan elnyúló eloszlások esetében van:



Medián – a minta felező pontja

– a minta fele kisebb, a fele nagyobb nála

ha két középső van, akkor a medián ezek átlaga

Ez robusztus mutató

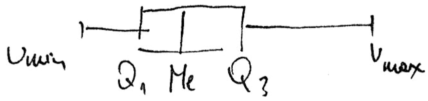
Hátrány: nehéz értelmezni

Quartilis : alsó és harmadik Q_1, Q_3

minta első vagy harmadik negyedelő pontja

Grafikus módszer:

Box plot



Szóródási mutatók:

Minimum / Maximum / Terjedelem (max − min)

↑

R: range

Empirikus szórásnégyzet

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$ – variancia

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$ – szórás

$$s^* = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$ – korrigált szórás

→ nem robusztus

Inter quartilis terjedelem

$$IQR = Q_3 - Q_1$$

Átlagos abszolút eltérés

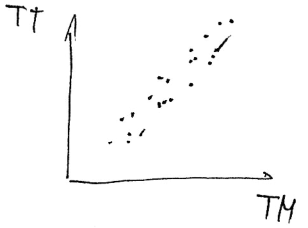
$$MAD = \frac{\sum |x_i - \bar{x}|}{n-1}$$

vagy

$$= \frac{\sum | x_i - m|}{n-1} \qquad m: \text{medián}$$

Többváltozós elemzések

Szóródási diagram – scattered dot

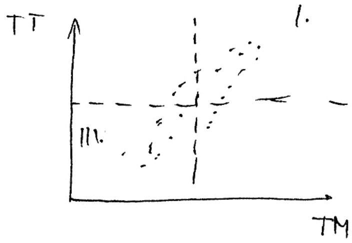


TT: testtömeg

TM: testmagasság

Sztochasztikus kapcsolat:

az egyik változó átlagnál nagyobb értékeihez tendenciaszerűen kapcsolódnak a másik változóból annak az átlagánál nagyobb értékei



Korreláció: $-1 \leq \text{corr}(TM, TT) \leq 1$

ha $= 0$, akkor nincs kapcsolat

az előjel iránya a kapcsolat irányát mondja meg

az abszolút értéke: az illeszkedés mértékére jellemző

(1-hez közeli érték – jó illeszkedés)

→ regressziós számításhoz vezet

Induktív statisztika

pl: $\overline{TM} = 172$ cm → ez az átlag mintáról mintára változik

A mintát ismerjük, tudunk-e mondani valamit a sokaságról

Biztosat mondani nem tudunk mondani, de valószínűségi jelentést tehetünk

Tudunk konstruálni egy tartományt, ahova a sokaság várható értéke egy bizonyos valószínűséggel esik

pl: a fenti $\overline{TM} = 172$ cm átlag után mondhatjuk, hogy a sokaság átlaga a (168 - 176 cm) tartományba 95%-os valószínűséggel esik

→ konfidencia intervallum

→ becsléselmélet

Hipotézis vizsgálat

van egy állításunk a sokaságról (pl. a várható értékről)

→ erről az állításról a minta alapján döntünk

Biostatisztika második előadás

Ismétlés

háttértudomány: valószínűségszámítás
(véletlen kísérlet, eseménytér, esemény, stb.)

Statisztika két része:
- deskriptív } statisztika
- induktív

Deskriptív statisztika lehet
- egyváltozós
- többváltozós (változók közötti kapcsolatokkal foglalkozik)

Adatok lehetnek:
nominálisak / ordinálisak / intervallum / [illegible] arány

Még egy osztás → feldolgozás lehet:
- grafikus
- analitikus

Átlag tömörebb, mint a hisztogram, de sokmindent elfed
a hisztogram komplexebb, de hűebb

A tömörítés mindig információvesztéssel jár

Nominális változóra lehet:
gyakorisági sor
grafikus oszlopdiagram
módusz

Ordinális változóknál az előzőeken kívül a mediánnak
lehet még értelme, de erőtlebb

Többváltozós elemzés

→ leszűkítjük a kétváltozósra

Egy grafikus módszer:

<u>Szóródási diagram</u> (scattered dot)

pl. testmagasság - testtömeg

kapcsolatot lehet vele kimutatni

→ nem determinisztikus kapcsolat van a két változó között, hanem sztochasztikus kapcsolat

Precíz definíció:

az egyik változó átlag feletti értékei tendenciaszerűen a másik változó átlag feletti értékeivel járnak együtt

→ ez a pozitív sztochasztikus kapcsolat

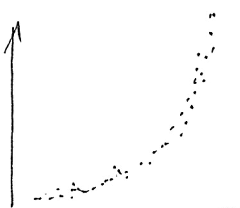
A szóródási diagramot a függőleges vagy vízszintes tengelyre vetítve a hisztogramot kapjuk

Egy kétváltozós vizsgálat többet mond, mint két egyváltozós vizsgálat

Analitikusan megadva:

<u>Pearson-féle (lineáris) korrelációs együttható</u> : R

Ez a szóródási diagram nagy R értéket ad



Ez a szóródási diagram nem ad nagy R értéket, mégis van kapcsolat a két érték között

Léteznek olyan korrelációs együtthatók, amik nem a lineáris kapcsolat erősségét, hanem a monoton kapcsolatot mutatják

pl.: <u>Spearmann-tau (τ)</u>

nem az adatokkal, hanem a rangjukkal számol

→ a rang megmondja, hogy az adott elem hányadik a rendezett mintában

Egy sima korrelációs együtthatót számolunk, de nem az eredeti adatokra, hanem a rangjukra

A monoton kapcsolatot mondja meg

→ mindegy neki, hogy milyen függvény szerinti a kapcsolat

Kategoriális adatokra (nominális és ordinális adatokra)

→ két változónál

<u>kontingencia tábla</u> v. <u>kereszt tábla</u>

a táblázatnak annyi sora és oszlopa van, amennyi kategóriája van az egyik és a másik változónak

Egy cellába az kerül, hogy hány elem volt a mintában, ami oda tartozott

→ Ennek egyváltozós redukciója a gyakorisági sor

→ úgy kaphatjuk, hogy a sorokat vagy az oszlopokat összeadjuk

Ezt <u>vetületi v. peremeloszlásnak</u> is hívják

Ha kettőnél több változó van, de egy változóban nincsen sok kategória, akkor is van értelme a kiterjesztésnek

Ha két változó van, akkor is csak akkor van értelme a kontingencia táblázatnak, ha nincsen 10-nél több kategóriája

→ ennél nagyobb táblázatok már átláthatatlanok

Főkomponens analízis v. faktor analízis

→ nagy korrelációs mátrixból kiszedi azokat, amelyek
kb. együtt mozognak
→ csoportok lesznek
kirajzolódnak az egymással együtt mozgó változók

Ha kategorikus ~~válto~~ adatokból van több →
erre is léteznek módszerek

pl: korrespondancia analízis CA

Ezek voltak a deskriptív statisztika módszerei

Ki nem mondott feltételezés
→ a kezünkben lévő adatokról mondunk valamit
a valóságban azonban a sokaságról szeretnénk mondani
valamit, amiből a mintát vettük

pl: első éves egyetemistákról szeretnék mondani valamit
→ az összes első éves egyetemistát nem tudom megvizsgálni
→ véletlenszerűen választok közülük → ezek nyertek mérést
→ azonban az összesről szeretnénk mondani valamit
→ induktív v. következtető statisztika

Van egy sokaságunk → erre irányul a kérdésfeltevésünk
→ ez a célsokaság
(pl. adott gyógyszer adott protokoll szerinti alkalmazása)
a sokaságot nem tudom egészében megfigyelni
→ amit meg tudok figyelni, azt nevezik mintának
→ Mintavételes helyzet
csak a mintáról van információm

Hogyan tudok ennek alapján a sokaságról mondani valamit? ⑳

A sokaságot a maga teljességében nehéz és drága megmérni.

Az is előfordul, hogy lehetetlen, mert a sokaság végtelen
(pl. egy adott típusú betegségben szenvedők halmaza
→ akárhány embert is vonok be a vizsgálatba
ez szükségképpen mindig csak minta lesz)

Ennek ellenére eléggé erős állításokat lehet megfogalmazni

Hibák:

Léteznek mintavételi és nem mintavételi hibák

Mintavételi hiba:
ami abból adódik, hogy miképpen vették a mintát
A valóságban nem csak ebből adódhat hiba
pl: válaszmegtagadás
nem jól kijelölt célsokaság
adat feldolgozási hiba

3000 főnél több mintát ritkán használnak
→ addig érdemes elmenni, amíg a mintavételi hiba kisebb, mint a nem mintavételi hiba

Bevezetés az induktív statisztikába:
deduktív statisztika:
ismert sokaságból ismeretlen mintára következtetünk

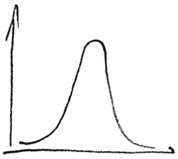
Egyelőre folytonos változóra szorítkozunk

Mi a jellemző eloszlása? → ezt ismerjük

feltételezés: a sokaság egyedileg adott pl $N = 100$
minta $n = 10$

Még a sokaságnak is véges az elemszáma

Az esetek többségében azonban a sokaság nem az elemeivel, hanem az eloszlásával adott



← pl. sűrűség függvényével

→ végtelen sokaságot az eloszlásával lehet megadni

Deduktív statisztika

← Ilyen az eloszlása → ebből veszek mintát

1) → egy elemet veszek ki

Arra lennék kíváncsi, hogy mennyi lesz a sokaság várható értéke · $EX = \mu$ várható érték

Ezt becslésnek hívják → az induktív statisztika egyik ága a becsléselmélet

A becslésre egy ellenvetés: ez lehetetlen → ui. ha csak egy elemet nem ismerünk a sokaságból → akkor már az átlag számításra lehet

pl: 100 elemű sokaságból ismerek 99-et, az utolsótól függően az átlag meg számítani is lehet

mégis ~~[illegible]~~ valószínűtlen az a helyzet, hogy a már meglévő 99 elem alapján számolt átlagtól nagyon eltérne a végleges átlag →

ehhez az utolsó elemnek extrém mértékben kellene különböznie a többitől

Feltéve persze, hogy a mintavételezés jól történt

Jó mintavételezés → reprezentatív minta

– ilyen a véletlen szám generálással kiválasztott minta

A mostani gondolatmenetben mindig véletlen mintavételezést tételezzük

Egy sokasági jellemzőt szeretnénk megbecsülni
(pl: várható érték, ferdeség...)

θ értéket becsüljük (egy ismeretlen sokasági jellemzőt)

$\hat{\theta}$ a becslés → θ becslése a mintaelemek alapján

$\hat{\theta} = f(y_1, y_2, \dots, y_n)$

f-et becslő függvénynek nevezzük

A becslő függvény függ attól, hogy mi a θ

Egy adott jellemző többféleképpen is becsülhető

A várható érték jó becslése az átlag

ha $\theta = \mu \Rightarrow \hat{\theta} = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$

Tfh. ismerjük a sokaságot

Legyen egy elemű mintánk

Az egy elemű minta eloszlása u.az lesz, mint az eredeti eloszlás → ezt háttéreloszlásnak is hívják

Két elemű minta: $n=2$

Átlag becslése: $\frac{y_1 + y_2}{2}$ → Mi lesz ennek az eloszlása

→ Ez nem egy triviális kérdés

Ha X is és Y követ valamilyen eloszlást, amit ismerünk
(jelölés: $X \sim$ és $Y \sim$)

Kérdés: Mi lesz $X+Y$ eloszlása? $(X+Y \sim ?)$

valószínűségi változóknál → konvolúció → egy db integrálással számítható

Mi esetünkben: egy eloszlásunk van, abból több mintát veszünk

Feltételezem, hogy a mintavétel független

→ így könnyen leírható a konvolúció

\+ u.abból a háttéreloszlásból vesszük a mintát

Ezt: FAE mintának hívjuk (i.i.d mintának is hívják angolul)

A gyakorlatban ez nem teljesül

Pl. feltételezem, hogy a sokaság normális eloszlású

$X \sim N(\mu, \sigma^2)$

FAE mintavételezéssel

normális eloszlás zárt a konvolúcióra

$X_1 + X_2 \sim N(2\mu, 2\sigma^2)$

\- Nem a szórások adódnak össze, hanem a varianciák

Két elemű minta átlaga:

$$\bar{X}_2 = \frac{X_1 + X_2}{2}$$

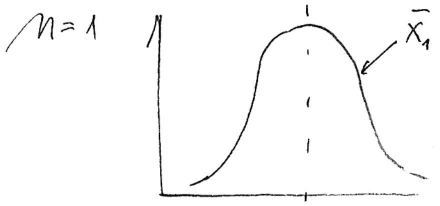
Normális eloszlás a lineáris transzformációra zárt

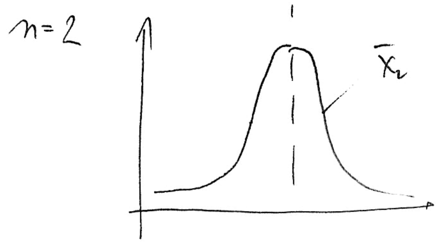
$$\bar{X}_2 \sim N\left(\mu, \frac{2\sigma^2}{2^2}\right) = N\left(\mu, \frac{\sigma^2}{2}\right)$$

$$\bar{X}_1 \sim N\left(\mu, \frac{n\sigma^2}{n^2}\right) = N\left(\mu, \frac{\sigma^2}{n}\right)$$

ui. $\mathbb{E}(ax) = a\,\mathbb{E}(X)$

$\mathbb{D}^2(ax) = a^2\,\mathbb{D}(X)$





Kicsit szűkebb

Egy elemű mintánál – elvileg bármilyen értéket felvehet a normális eloszlású véletlen minta

Akárhány elemű mintánál az átlag lehet kisebb is és nagyobb is mint a valós mennyiség
- elvileg akármennyivel

Azonban a minta nagyságának növelésével egyre jobban ráfókuszálunk a jó mennyiségre
- → $\bar{X}_n$ eloszlása egyre keskenyebb Gauss-görbe lesz n növelésével

→ Nem igaz, hogy semmit sem tudok mondani a minta alapján a sokaságról

Igen, biztosat nem tudok megfogalmazni

A becslő függvénytől elvárjuk, hogy a jó érték körül ingadozzon

→ Szeretnénk, ha $E(\hat{\Theta}) = \Theta$ lenne (a becslő függvény várható értéke a becsült mennyiség lenne)

→ Ezt hívjuk <u>torzítatlanságnak</u>

A minta átlag torzítatlan becslő minden mintaméret mellett

<u>Aszimptotikusan torzítatlan becslés</u>

$E(\hat{\Theta}) \neq \Theta$, de

$\lim_{n\to\infty} E(\hat{\Theta}) = \Theta$

Emellett $D^2(\hat{\Theta})$-tól azt szeretnénk, hogy minél kisebb legyen

Vegyünk egy alternatív becslő függvényt

$f(x_1, x_2, \dots, x_n) = x_1 \sim N(\mu, \sigma^2)$

Ez egy torzítatlan becslő függvény

DE: átlagszámításkor növelhetjük n-t → mindig egyre kisebb lesz a szórásnégyzete

A torzítatlan becslő függvények közül a legkisebb szórásút **hatásos becslő függvénynek** hívjuk

Bizonyítható, hogy az átlag hatásos becslője a várható értéknek

Konzisztens egy becslő függvény, ha két feltétel teljesül:

$$\lim_{n \to \infty} \mathbb{D}^2 \hat{\Theta} = 0 \quad \text{és}$$

$$\lim_{n \to \infty} \mathbb{E} \hat{\Theta} = \Theta$$

$f(x_1, x_2, \dots, x_n) = x_1$ nem konzisztens

Kérdés: Hogyan találunk jó becslő függvényt egy adott problémára?

pl. Mi a 170 cm-nél magasabbak mediánja

Az, hogy az átlag a várható érték egy ilyen jó becslés, és ennyire egyszerű – csak a szerencsén múlott

Léteznek elvek, amivel ~~megtalálhatók~~ találhatók becslések

pl. maximum likelihood elv

→ becslés készítési elv

A statisztika egy másik ága:

hipotézis vizsgálat

→ erről nem esett több szó...