

Valószínűségszámítás és statisztika előadás Info. BSC B-C szakosoknak

20018/2019 1. félév
Zempléni András
zempleni@caesar.elte.hu
zempleni.elte.hu

1. előadás: Bevezetés

- Irodalom, követelmények
- A félév célja
- Valószínűségszámítás tárgya
- Történet
- Alapfogalmak
- Valószínűségek kiszámítása

Irodalom

- Jegyzet
 - Baróti-Bognárné-Fejes Tóth-Mogyoródi: Valószínűségszámítás jegyzet programozó szakos hallatóknak
 - Matematikai statisztika jegyzet programozó matematikus hallatóknak
 - Internetes jegyzetek (BME, Debrecen, stb)
- Tankönyvek:
 - Prékopa: Valószínűségelmélet
 - Solt: Valószínűségszámítás
 - Bolla - Krámli: Statisztikai következetések elmélete
 - Pál: A valószínűségszámítás és a statisztika alapjai I-II
- Példatárák
 - Bognárné-Mogyoródi-Prékopa-Rényi-Szász: Valószínűségszámítási feladatgyűjtemény
 - Arató Miklós, Prokaj Vilmos és Zempléni András: Bevezetés a valószínűségszámításba és alkalmazásaiiba: példákkal, szimulációkkal (eletronikus jegyzet)
 - Mori-Szeidl-Zempléni: Matematikai statisztika példatár

Számonkérés

- Gyakorlatok
 - gyakorlati jegy: csoportonkénti zh-k alapján
- Vizsga: írásbeli, később egyeztetendő időpontban
- Az előadáson is kötelező a jelenlét (3 hiányzás lehetséges)! Papíros ellenőrzés lesz
- Az óraütközést az oktató igazolhatja
- Előadások anyaga: zempleni.elte.hu/okt.html

Cél

- Valószínűségszámítás és statisztika alapjainak ismertetése
- Feladatmegoldási készség kialakítása (elsősorban gyakorlaton)
- Alkalmazási lehetőségek bemutatása
- Modern módszerek (big data, R)

Véletlen tömegjelenségek

- Ismételhető/nagy számban végbemenő események (például: X éves férfi/nő mekkora valószínűsséggel köt 2 hónapon belül házasságot)
- Véletlen: az ismert feltételrendszer nem határozza meg egyértelműen az eredményt (pl. kockadobás). Nem is érdemes determinisztikus modellel kísérletezni, mert túl bonyolult lenne.

Valószínűségszámítás helye a tudományok között

- Matematikai tudomány, mert precízen megfogalmazott axiómákra épül.
- Gyakorlati alkalmazásai: statisztikai következtetések levonása (pl.: ha egy érmével 1000 dobásból 550 fej jött ki, akkor 99.9% valószínűséggel állítható, hogy az érme nem szabályos).

Történeti áttekintés 1.

- Első ismert feladat 1494-ből: játék idő előtti abbahagyása esetén hogyan osztozzanak? Helyes megoldás több, mint 100 ével későbbi: Pascal (1623 – 1662), Fermat (1601 – 1665)
- Könnyen adható szimulációs megoldás (precíz számítás a gyakorlaton)
- Cardano (1540 körül) könyvet írt a kockajátékokhoz kapcsolódó valószínűségszámítási kérdésekről

Történeti áttekintés 2.

- de Witt, Halley (1671): életjáradék-számítás valószínűségi alapon
- Jacob Bernoulli (1713): Ars Conjectandi (nagy számok törvénye)
- XVIII-XIX. sz: Moivre, Bayes, Gauss, Poisson
- Buffon: geometriai valószínűség bevezetése – paradoxonok
- XIX.sz: Csebisev, Markov, Ljapunov

Történeti áttekintés 3.

- Axiomatizálás: Kolmogorov (1933)
- Modern alkalmazások:
 - Információelmélet (Shannon)
 - Játékelmélet (Neumann)
 - Matematikai statisztika (Fisher)
 - Sztochasztikus folyamatok
- Magyar tudósok:
 - Jordán Károly (1871-1959)
 - Rényi Alfréd (1921-1970)

Alapfogalmak

- Eseménytér
 - Kísérlet egy lehetséges kimenetele: elemi esemény, jelölése ω .
 - Elemi események összessége: eseménytér, Ω .
 - Ω részhalmazai: események (A, B, C, \dots).
 - Esemény akkor következik be, ha az őt alkotó elemi események valamelyike bekövetkezik.

Példák

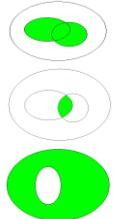
- Kockadobás: $\Omega = \{1, 2, \dots, 6\}$. Ha az A esemény: páros számot dobtunk, akkor $A = \{2, 4, 6\}$.
- Érmét kétszer feldobva: $\Omega = \{II, IF, FI, FF\}$ $A = \{II, IF\}$ az az esemény, hogy az első dobás írás.
- Érmét addig dobunk, míg fejet nem kapunk. $\Omega = \{F, IF, IIF, \dots, \omega_\infty\}$ ahol $\omega_\infty = III\dots$ (azaz minden dobás írás)

Események

- Speciális események:
 - Ω (biztos esemény)
 - \emptyset (lehetetlen esemény)
- Az események összessége: \mathcal{A} (halmazrendszer Ω részhalmazaiból)
- Műveletek eseményekkel: szokásos logikai műveletek = halmazműveletek

Műveletek eseményekkel

- $A \cup B$, vagy A vagy B bekövetkezik (az is lehet, hogy mindkettő)
- $A \cap B$: A és B is bekövetkezik
- A esemény ellentettje: \bar{A}



Tulajdonságok

$$A \setminus B = A \cap \bar{B}$$



$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$



(De Morgan)

$$\overline{\bar{A}} = A \quad \overline{\Omega} = \emptyset$$

Példák

- Kockadobás:
 - $A=\{\text{páros számot dobunk}\}$
 - $B=\{\text{legalább 3-ast dobunk}\}$
 - $A \cap B = \{4,6\}$
 - $A \cup B = \{2,3,4,5,6\}$
 - $\overline{A \cap B} = \{2\}$
 - $\overline{A} = \{1,3,5\}$

Valószínűség

- Szemléletes megfelelője: *relatív gyakoriság*. Ha n egymástól függetlenül, azonos körülmények között végrehajtott kísérletből az adott A esemény k -szor következett be, akkor a relatív gyakoriság k/n .
- Nagy n -re a relatív gyakoriság egy fix szám körül ingadozik: ezt nevezük az A valószínűségének.
- Szimulációk (appletek):
<http://www.randomservices.org/random/>
- Kocka-kísérlet

A valószínűség

- Jele: $P(A)$
- A relatív gyakoriság tulajdonságaiból:
 - Nemnegatív: $P(A) \geq 0$ minden A -ra
 - Egymást kizáró eseményekre, azaz, ha $A \cap B = \emptyset$: $P(A \cup B) = P(A) + P(B)$ (additivitás)
 - $P(\Omega) = 1$
- (Ω, \mathcal{A}, P) : valószínűségi mező

Tulajdonságok 1.

- Additivitás n eseményre: ha A_1, A_2, \dots, A_n páronként kizártó események, akkor

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Bizonyítás: indukcióval.

- $P(\emptyset) = 0$.

Bizonyítás: $\Omega = \Omega \cup \emptyset$ felbontásból és az additivitásból

Tulajdonságok 2.

$$P(A \setminus B) = P(A) - P(A \cap B)$$

Bizonyítás: $A = (A \cap B) \cup (A \setminus B)$ felbontásból és az additivitásból

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Bizonyítás: $A \cup B = B \cup (A \setminus B)$ felbontásból, az additivitásból és az előző tulajdonságból.

Kolmogorov-féle valószínűségi mező

- (Ω, \mathcal{A}, P) : Kolmogorov-féle valószínűségi mező, ha
 - Ω nemüres halmaz
 - \mathcal{A} az Ω részhalmazainak σ -algebrája
 - $P: \mathcal{A} \rightarrow [0, 1]$ halmazfüggvény (valószínűség), melyre
 - $P(\Omega) = 1$
- σ -additivitás: ha A_1, A_2, \dots , páronként kizártó események, akkor

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$

A valószínűség további tulajdonságai

A (Kolmogorov-féle) valószínűség végesen is additív: ha A_1, A_2, \dots, A_n páronként kizártó események, akkor

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Bizonyítás: $A_{n+1} = A_{n+2} = \dots = \emptyset$ választással alkalmazzuk a σ -additivitást.

Tehát a korábban belátott tulajdonságok a Kolmogorov-féle valószínűségi mezőre is érvényesek.

Véges valószínűségi mező

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}, \mathcal{A} = \mathcal{P}(\Omega).$$

Jelölés: $p_i = P(\omega_i)$.

$$\sum_{i=1}^n p_i = \sum_{i=1}^n P(\omega_i) = P(\Omega) = 1$$

az additivitásból.

$$P(A) = P(\bigcup_{\omega_i \in A} \omega_i) = \sum_{\omega_i \in A} p_i$$

Azaz a p_i nemnegatív, 1 összegű számok meghatározzák a valószínűséget.

Klasszikus valószínűségi mező 1

$p_i = 1/n$ minden i re (azonos valószínűségük az elemi események).

Ekkor $P(A) = \frac{k}{n}$ ahol k az A elemszáma,

n pedig az összes esetszám.

Másképpen: $P(A) =$ kedvező esetek száma/összes esetszám.

Visszatevéses mintavétel

- N termék, melyből M selejtes
- n elemű minta visszatevéssel
- A : pontosan k selejtes van a mintában

$$(k=0, \dots, n) \quad P(A) = \binom{n}{k} \left(\frac{M}{N} \right)^k \left(1 - \frac{M}{N} \right)^{n-k}$$

azaz a valószínűség kifejezhető a $p = M/N$ selejtarány segítségével:

$$P(A) = \binom{n}{k} p^k (1-p)^{n-k}$$

[Mintavétel](#)

Visszatevés nélküli mintavétel

- N termék, melyből M selejtes
- n elemű minta visszatevés nélkül
- A : pontosan k selejtes van a mintában

$$(k=0, \dots, n)$$

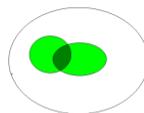
$$P(A) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$$

[Mintavétel](#)

Események uniójának valószínűsége

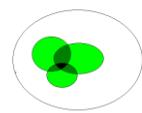
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Példa: Magyar kártyacsomagból kétszer húzunk visszatevessel. Mi a valószínűsége, hogy húzunk pirosat?



A : első piros, B : második piros
 $P(A)=P(B)=1/4$, $P(A \cap B)=1/16$
Tehát $P(A \cup B)=7/16$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$



Szita (Poincaré) formula

Képlet az általános esetre:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n (-1)^{i+1} S_i^{(n)}$$

ahol

$$S_i^{(n)} = \sum_{1 \leq j_1 < j_2 < \dots < j_i \leq n} P(A_{j_1} \cap A_{j_2} \cap \dots \cap A_{j_i})$$

az i tényezős metszetek valószínűségeinek összege.

Alkalmazások

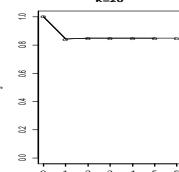
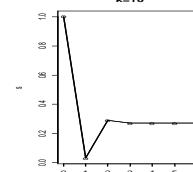
- Ha az egyes események és metszeteik is egyformán valószínűek, akkor
- $$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} P(A_1 \cap A_2 \cap \dots \cap A_i)$$
- Átfogalmazás metszetekre:
- $$P(\bar{A}_1 \cap \bar{A}_2 \cap \dots \cap \bar{A}_n) = 1 - P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=0}^n (-1)^i S_i^{(n)}$$
- (Megállapodás: $S_0=1$)
- Példa: Mi a valószínűsége, hogy adott (k) számú kockadobásból minden számot legalább egyszer megkaptunk?

Megoldás

- A_i : az i -számot nem dobtuk

$$P(\bar{A}_1 \cap \bar{A}_2 \cap \dots \cap \bar{A}_6) =$$

$$= \sum_{i=0}^6 (-1)^i \binom{6}{i} P(A_1 \cap A_2 \cap \dots \cap A_i) = \sum_{i=0}^6 (-1)^i \binom{6}{i} \left(\frac{6-i}{6} \right)^k$$



Feltételes valószínűség 1.

- Az A esemény valószínűségét keressük.
- Tudjuk, hogy B esemény bekövetkezett.
- A relatív gyakoriságokkal: csak azokat a kísérleteket nézzük, amelyekben B bekövetkezett. Ezen részsorozatban az A relatív gyakorisága:

$$r_{A \cap B} / r_B$$

Feltételes valószínűség 2.

- Megfelelője a valószínűségekre:

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

az A esemény B -re vonatkozó feltételes valószínűsége (feltétel: $P(B) > 0$).

- Példa: kockadobás. $A=\{\text{páros számot dobunk}\}$
 $B=\{\text{3-nál nagyobbat dobtunk}\}$
 $P(A|B)=2/3$.

Teljes eseményrendszer

- *Definíció.* Események A_1, A_2, \dots , sorozata *teljes eseményrendszer*, ha egymást páronként kizárták és egyesítésük Ω .
- *Tulajdonság:* $P(A_1) + P(A_2) + \dots = 1$
- Legtöbbször véges sok elemből álló teljes eseményrendszeret vizsgálunk.

Teljes valószínűség tétele

- Legyen B_1, B_2, \dots pozitív valószínűségű eseményekből álló teljes eseményrendszer, $A \in \mathcal{A}$ tetszőleges. Ekkor
$$P(A) = P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \dots$$
- *Bizonyítás.* $A = (A \cap B_1) \cup (A \cap B_2) \cup \dots$ diszjunkt tagokra bontás, tehát
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots$$
 és $P(A \cap B_i) = P(A | B_i)P(B_i)$ adja a tételelt.

Példa

- Összetett modellek (pl. nemtől függő valószínűségek): a színvakság valószínűsége a férfiaknál 0.01, a nőknél 0.001 (Tfh. ugyanannyi a férfi, mint a nő.) Mi a valószínűsége, hogy egy találomra válaszott ember színvak?
- A teljes eseményrendszer: $\{\text{férfi}\} \cup \{\text{nő}\}$.
 $p=0.01/2+0.001/2=0.0055$
- Ugyanígy tudunk számolni nem azonos valószínűségű eseményekre is

Bayes tétele

Legyen B_1, B_2, \dots pozitív valószínűségű eseményekből álló teljes eseményrendszer és $A \in \mathcal{A}$ pozitív valószínűségű. Ekkor

$$P(B_k | A) = \frac{P(A | B_k)P(B_k)}{\sum P(A | B_i)P(B_i)}$$

Bizonyítás. A nevező éppen $P(A)$ a teljes valószínűség tétele miatt.

- A számláló pedig $P(A \cap B)$, definíció szerint.
■ Spec.: Két elemű teljes eseményrendszerre:

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

Valószínűségszámítás és statisztika előadás Info. BSC B-C szakosoknak

20018/2019 1. félév
Zempléni András
2.előadás

Bayes tétele

Legyen B_1, B_2, \dots pozitív valószínűségű eseményekből álló teljes eseményrendszer és $A \in \mathcal{A}$ pozitív valószínűségű. Ekkor

$$P(B_k | A) = \frac{P(A | B_k)P(B_k)}{\sum P(A | B_i)P(B_i)}$$

Bizonyítás. A nevező éppen $P(A)$ a teljes valószínűség tétele miatt.

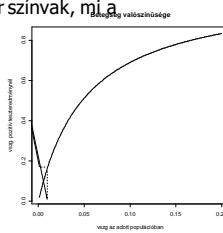
A számláló pedig $P(A \cap B)$, definíció szerint.

- Spec.: Két elemű teljes eseményrendszerre:

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

Példák

- Ha egy találomra válaszott ember színvak, mi a valószínűsége, hogy férfi?
 $p=0.005/(0.005+0.0005)=10/11.$



$$P(B|poz)=P(poziB)P(B)/[P(poziB)P(B)+P(poziE)P(E)]=p/(p+0.05(1-p))$$

Események függetlensége

- Ha a B esemény bekövetkezése nem befolyásolja az A valószínűségét, azaz $P(A|B)=P(A)$, akkor azt mondjuk, hogy az A és B függetlenek. Ez így nem ideális definíció (nem szimmetrikus, $P(B)>0$ kell hozzá), ezért
- Definíció. Az A és B események függetlenek, ha $P(A \cap B)=P(A)P(B)$.

Példák

- Húzunk egy lapot egy magyarkártya-csomagból. A : piros B : ász.
 $P(A)=1/4$, $P(B)=1/8$, $P(A \cap B)=1/32$, tehát függetlenek.
- A függetlenség nagyon ritka azonos kísérletből meghatározott eseményeknél!
- Tipikus eset függetlensége: A az első, B a második kísérlet eredménye.

Tulajdonságok

- Ha A és B diszjunktak, akkor csak triviális ($P(A)=0$ vagy $P(B)=0$) esetben függetlenek.
- Ha A és B függetlenek, akkor komplementereik is függetlenek.
- Önmaguktól csak a triviális események függetlenek.
- $A \subset B$ esetén csak akkor függetlenek, ha legalább az egyik triviális.

Általánosítás

- n esemény független, ha $P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_k})$ teljesül tetszőleges $1 \leq i_1 < i_2 < \dots < i_k \leq n$ indexsorozatra és minden $2 \leq k \leq n$ számra.
- Nem elég a fenti szorzat-tulajdonságot $k=2$ -re megkövetelni. Ha csak ez teljesül: páronkénti függetlenségről beszélünk.

Megjegyzések

- n független kísérlet esetén az egyes kísérletekhez tartozó események függetlenek. A gyakorlatban ez a tipikus, fontos előfordulása ennek a függetlenségnak.
- Klasszikus valószínűségi mező esetén független kísérleteket végezve, a kedvező és az összes események száma is összeszorozódik.
- Példa: szabályos kockával dobva: $P(\text{első dobás páros és a második hatos}) = 3/36$.

Valószínűségi változók

- A legtöbbször nem maga a kísérlet kimenetele (a realizálódott elemi esemény) hanem egy számszerűsíthető eredmény az érdekes.
- Példa: ipari termelés – minőségellenőrzés: a kérdés az esetleges selejtesek száma, nem pedig az, hogy pontosan melyik elemeket is választottak.
- Sok gyakorlati esetben nem is adódik természetesen az Ω halmaz (pl. időjárás megfigyelés).

Valószínűségi változók 2.

- Mintavételei példa (folyt). N termék, n elemű minta. Ω elemszáma: $\binom{N}{n}$
- Selejtesek száma (X): 0 és n közötti szám.
- Matematikailag: $X: \Omega \rightarrow \mathbb{R}$ függvény
- Feltétel: legyen értelme pl. annak a valószínűségről beszélni, hogy $X=a$. Hasonlóképpen más „termesztes” feltételek is legyen valószínűsége. Formalisan: megköveteljük, hogy $\{\omega: X(\omega) \in B\} \in \mathcal{A}$ teljesüljön minden, az intervallumokból megszámíthatóan sok halmazművelettel előállítható B -re.
- A gyakorlatban általában nem jelent problémát.

Példák

- Kockadobás:
 - X a dobott szám. $\Omega = \{1, 2, \dots, 6\}$, $X(i) = i$. Értékkészlete: $\{1, 2, \dots, 6\}$.
 - X az első olyan dobás sorszáma, amikor 6 jön ki. $\Omega = \{1, 2, \dots, 6\} \times \{1, 2, \dots, 6\} \times \dots$
 - X értékkészlete: $\{1, 2, \dots\}$
- Gyakorlati példák:
 - X az első selejt gyártásának időpontja. X értékkészlete: \mathbb{R}_+ .
 - X egy adott termék hossza. X értékkészlete: \mathbb{R}_+ részhalmaza (nem szükséges előzetesen korlátozni).

Diszkrét valószínűségi változók

- Definíció: az X diszkrét valószínűségi változó, ha értékkészlete (x_1, \dots, x_n) legfeljebb megszámítható.
- A valószínűségi változó definíciójából adódóan $\{\omega: X(\omega) = x\} = \{\omega: X(\omega) \in \mathcal{A}\}$ azaz $p_x := P(X=x)$ értelmes. Ezek meg is határozzák X eloszlását.
- Véges vagy megszámítható valószínűségi mezőn minden valószínűségi változó diszkrét.
- Nem célszerű a természetesen folytonos értékkészletű X diszkretizálása (egyszerűbbek a folytonos modellek – pl. esemény bekövetkezési ideje, file mérete, éves jövedelem).

Példák diszkrét valószínűségi változókra

- $X(\omega) = c$ minden ω -ra.
- Elnevezés: elfajult eloszlás.
- $$P(X=c)=1.$$
- X akkor 1, ha egy adott, p valószínűségű A esemény bekövetkezik és 0 különben (elnevezés: az A esemény indikátora).
- $$P(X=0)=1-p$$
- $$P(X=1)=p$$

Példák 2.

- Mintavételel legyen X a mintában levő selejtesek száma.
 - Visszatevéses esetben (binomiális eloszlás):

$$P(X=k) = \binom{n}{k} \left(\frac{M}{N}\right)^k \left(1 - \frac{M}{N}\right)^{n-k} \quad (k = 0, \dots, n)$$
 - Visszatevés nélküli esetben:

$$(hipergeometriai eloszlás) \quad P(X=k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \quad (k = 0, \dots, n)$$

Binomiális eloszlás alkalmazása

- Visszatevéses mintavétel más realizációja: független kísérletek azonos körülmények között. $P(A)=p$ esemény, végezzünk n (rögzített számú) független kísérletet.
- X : az A bekövetkezésének gyakorisága (pontosan hányszor jött ki az A). X eloszlása binomiális (n, p).
- $X = X_1 + X_2 + \dots + X_n$, ahol X_i az i -edik kísérletnél az A esemény indikátora. Ezek az indikátorok függetlenek is!
- Példák: 5 dobásból hány fej jön ki?
 - Egy ingyenes játék letöltői közül átlagosan minden 10. meg is veszi a haladó változatot. 100 letöltő közül hányan fognak vásárolni?

Geometriai (Pascal) eloszlás

- Független kísérletek azonos körülmények között. $P(A)=p$ esemény, addig kísérletezünk, míg A be nem következik.
 - X : az első sikeres kísérlet sorszáma.
- $$p_k = P(X=k) = p(1-p)^{k-1} \quad (k=1,2,\dots)$$
- Valóban valószínűségeseloszlás ($p_1+p_2+\dots=1$)
- geometriai eloszlás**
- Példák: hányadikra dobjuk az első fejet?
 - Hány hétag kell lottónunk az első nyerésig?
 - Hányadik honlapon találja meg a kereső az adott kifejezést?

Poisson eloszlás

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (k=0,1,2,\dots; \lambda > 0)$$

paraméter). Valóban eloszlás. [Grafikusan](#)

Állítás. Ha a binomiális eloszlás paramétereire $n \rightarrow \infty$ úgy, hogy $np \rightarrow \lambda$, akkor a határérték éppen a λ paraméterű Poisson eloszlás.

Bizonyítás. $\binom{n}{k} p^k (1-p)^{n-k} \approx \binom{n}{k} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \rightarrow \frac{\lambda^k e^{-\lambda}}{k!}$

Gyakorlati alkalmazások

- Első példa: lórugás áldozatainak száma a porosz hadseregben.
- Ritka események száma adott időszakban:
 - Balesetek száma
 - Viharok száma
 - Rendszer meghibásodásainak száma

Összefoglalás (diszkrét eloszlások)

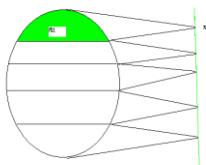
- Binomiális eloszlás
 - Rögzített számú kísérletnél adott esemény gyakorisága (pl. 10 kockadobásból a hatosok száma)
 - Nagy mintaelemszámról, kicsi valószínűségnél a Poisson eloszlással közelíthető
- Pascal (geometriai) eloszlás
 - Addig kísérletezünk, míg egy adott esemény be nem következik, az első sikeres sorszáma (pl. az első hatost hárnyadik kockadobásnál kapjuk meg)
- Hipergeometriai eloszlás
 - Visszatévezés nélküli mintavételnél adott típusú mintaelemek száma (pl. lottóhúzásnál az 5 találat valószínűsége)

Tulajdonságok

- Ha X diszkrét valószínűségi változó,
- $f: \mathbf{R} \rightarrow \mathbf{R}$ tetszőleges függvény, akkor $f(X)$ is diszkrét valószínűségi változó.
- Példa: X a gyártott termék hossza mm-ben. Tegyük fel, hogy $P(X=18)=\dots=$
 $=P(X=22)=1/5$. T.f.h. az ideális a 20 mm. Ekkor a $d=|X-20|$ eltérés eloszlása:
 $P(d=0)=1/5$, $P(d=1)=P(d=2)=2/5$.

Teljes eseményrendszer

- Ha X diszkrét valószínűségi változó, akkor az $A=\{\omega: X(\omega)=x\}$ események teljes eseményrendszer alkotnak.



Az eloszlásfüggvény

- Legyen $F_X(z)=P(X<z)$, az $F_X(z): \mathbf{R} \rightarrow \mathbf{R}$ függvény az X valószínűségi változó eloszlásfüggvénye.
- Tulajdonságai:
 - $0 \leq F_X(z) \leq 1$
 - $F_X(z)$ monoton növő
 - $\lim_{z \rightarrow -\infty} F_X(z)=1$, $\lim_{z \rightarrow +\infty} F_X(z)=0$
 - $F_X(z)$ balról folytonos.
- Bizonyítás: Az első kettő trivialis, az utolsó kettőhöz a valószínűség folytonossága kell:
 Ha $A_1 \supseteq A_2 \supseteq \dots$ akkor $\lim_{n \rightarrow \infty} P(A_n)=P(A)$
 ahol $A = \bigcap_{i=1}^{\infty} A_i$

Példák

- Tetszőleges 1-4 tulajdonságú F -hez létezik X , aminek F az eloszlásfüggvénye (pl. $\Omega = \mathbf{R}$, $P([a,b))=F(b)-F(a)$, X az idendifitásifüggvény)
- A c pontban elfajult eloszlás $F(z)=\begin{cases} 0, & \text{ha } z \leq c \\ 1, & \text{ha } z > c \end{cases}$
- Az indikátorváltozó eloszlásfüggvénye $F(z)=\begin{cases} 0, & \text{ha } z \leq 0 \\ 1-p, & \text{ha } 0 < z \leq 1 \\ 1, & \text{ha } z > 1 \end{cases}$

Valószínűségek kiszámítása

- $P(a \leq X < b) = F(b) - F(a)$
- $P(X=a) = F(a+0) - F(a)$, azaz ha F folytonos, minden egyes pont 0 valószínűségű.
- $P(a < X < b) = F(b) - F(a+0)$
- $P(a \leq X \leq b) = F(b+0) - F(a)$

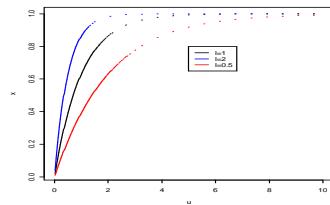
Folytonos eloszlások

- Definíció. X folytonos eloszlású, ha eloszlásfüggvénye folytonos.
- Példa: egyenletes eloszlás $[a,b]$ intervallumon:

$$F(z) = \begin{cases} 0, & \text{ha } z \leq a \\ \frac{z-a}{b-a}, & \text{ha } a < z \leq b \\ 1, & \text{ha } z > b \end{cases}$$

Exponenciális eloszlás

$$F(z) = \begin{cases} 0, & \text{ha } z \leq 0 \\ 1 - e^{-\lambda z}, & \text{ha } 0 < z \end{cases} \quad \text{ahol } \lambda > 0 \text{ paraméter}$$



Abszolút folytonos eloszlások

- Ha létezik f , hogy F előáll f integrálfüggvényeként:
$$F(z) = \int_{-\infty}^z f(t) dt$$

akkor azt mondjuk, hogy F abszolút folytonos, f **sűrűségfüggvénye**.

- Az események valószínűsége: $P(X \in A) = \int_A f(t) dt$
- f tulajdonságai: $f \geq 0$,
$$\int_{-\infty}^{\infty} f(t) dt = 1$$
- Ez elég is: minden ilyen f integrálfüggvénye eloszlásfüggvény.
- Kiszámítása a gyakorlatban: $f=F'$

A sűrűségfüggvény tulajdonságai

- Létezéséhez szükséges, hogy F folytonos legyen.
- Ha F abszolút folytonos, akkor $F'=f$, ahol F deriválható.
- f nem egyértelmű (pl. véges sok pontban tetszőleges értéket adhatunk neki), ezért a legegyszerűbb, szakaszonként folytonos változatot választjuk.
- Szemléletes jelentése:
$$P(a < X < b) = \int_a^b f(t) dt \approx f(a)(b-a)$$

azaz rövid intervallumokra a valószínűség közelíthető a sűrűségfüggvény értékének és az intervallum hosszának a szorzatával.

Szemléletes bevezetés

- Ha úgy közelítjük az abszolút folytonos eloszlást (pl. az év egy adott napján 12 órakor Bp-en a hőmérséklet), hogy egyre pontosabb eszközökkel mérjük meg, akkor $P(z < X < z + \delta) / \delta \approx f(z)$, azaz a valószínűségekből határátmenettel adódik a sűrűségfüggvény.

Példák

- Egyenletes eloszlás $[a,b]$ intervallumon

$$f(z) = \begin{cases} 0, & \text{ha } z \leq a \\ \frac{1}{b-a}, & \text{ha } a < z \leq b \\ 0, & \text{ha } z > b \end{cases}$$

- Exponenciális eloszlás

$$f(t) = \begin{cases} 0, & \text{ha } t \leq 0 \\ \lambda e^{-\lambda t}, & \text{ha } 0 < t \end{cases}$$

Valószínűségszámítás és statisztika előadás Info. BSC B-C szakosoknak

20018/2019 1. félév

Zempléni András
3.előadás

Példák

- Tetszőleges 1-4 tulajdonságú F -hez létezik X , aminek F az eloszlásfüggvénye (pl. $\Omega = \mathbf{R}$, $P([a,b])=F(b)-F(a)$, X az idenditásfüggvény)
- A c pontban elfajult eloszlás $F(z) = \begin{cases} 0, \text{ha } z \leq c \\ 1, \text{ha } z > c \end{cases}$
- Az indikátorváltozó eloszlásfüggvénye $F(z) = \begin{cases} 0, \text{ha } z \leq 0 \\ 1-p, \text{ha } 0 < z \leq 1 \\ 1, \text{ha } z > 1 \end{cases}$

Valószínűségek kiszámítása

- $P(a \leq X < b) = F(b) - F(a)$
- $P(X=a) = F(a+0) - F(a)$, azaz ha F folytonos, minden egyes pont 0 valószínűségű.
- $P(a < X < b) = F(b) - F(a+0)$
- $P(a \leq X \leq b) = F(b+0) - F(a)$

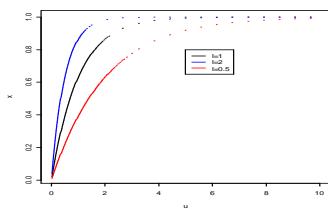
Folytonos eloszlások

- Definíció. X folytonos eloszlású, ha eloszlásfüggvénye folytonos.
- Példa: egyenletes eloszlás $[a,b]$ intervallumon:

$$F(z) = \begin{cases} 0, \text{ha } z \leq a \\ \frac{z-a}{b-a}, \text{ha } a < z \leq b \\ 1, \text{ha } z > b \end{cases}$$

Exponenciális eloszlás

$$F(z) = \begin{cases} 0, \text{ha } z \leq 0 \\ 1 - e^{-\lambda z}, \text{ha } 0 < z \end{cases} \quad \text{ahol } \lambda > 0 \text{ paraméter}$$



Abszolút folytonos eloszlások

- Ha létezik f , hogy F előáll f integrálfüggvényeként:
- $$F(z) = \int_{-\infty}^z f(t) dt$$
- akkor azt mondjuk, hogy F abszolút folytonos, f **sűrűségfüggvényel**.
- Az események valószínűsége: $P(X \in A) = \int_A f(t) dt$
 - f tulajdonsága: $f \geq 0$,
- $$\int_{-\infty}^{\infty} f(t) dt = 1$$
- Ez elég is: minden ilyen f integrálfüggvénye eloszlásfüggvény.
 - Kiszámítása a gyakorlatban: $f=F'$

A sűrűségfüggvény tulajdonságai

- Létezéséhez szükséges, hogy F folytonos legyen.
- Ha F abszolút folytonos, akkor $F'=f$, ahol F deriválható.
- f nem egyértelmű (pl. véges sok pontban tetszőleges értéket adhatunk neki), ezért a legegyszerűbb, szakaszonként folytonos változatot választjuk.
- Szemléletes jelentése:

$$P(a < X < b) = \int_a^b f(t) dt \approx f(a)(b-a)$$
azaz rövid intervallumokra á valószínűség közelíthető a sűrűségfüggvény értékének és az intervallum hosszának a szorzatával.

Szemléletes bevezetés

- Ha úgy közelítjük az abszolút folytonos eloszlást (pl. az év egy adott napján 12 órakor B_p -en a hőmérséklet), hogy egyre pontosabba eszközökkel mérjük meg, akkor $P(z < X < z+\delta)/\delta \approx f(z)$, azaz a valószínűségekből határatmenettel adódik a sűrűségfüggvény.

Példák

- Egyenletes eloszlás $[a,b]$ intervallumon

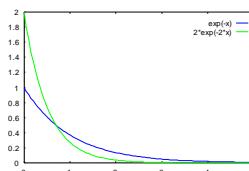
$$f(z) = \begin{cases} 0, & \text{ha } z \leq a \\ \frac{1}{b-a}, & \text{ha } a < z \leq b \\ 0, & \text{ha } z > b \end{cases}$$

- Exponenciális eloszlás

$$f(t) = \begin{cases} 0, & \text{ha } t \leq 0 \\ \lambda e^{-\lambda t}, & \text{ha } 0 < t \end{cases}$$

Exponenciális eloszlás

Exponenciális eloszlás
A sűrűségfüggvény $\lambda=1$ és $\lambda=2$ esetén



$g(X)$ eloszlása

- Legyen $g: \mathbf{R} \rightarrow \mathbf{R}$ (mérhető) függvény. Ekkor $g(X)$ is valószínűségi változó.
- Abból, hogy X eloszlása abszolút folytonos, nem következik még $g(X)$ eloszlásának folytonossága sem: pl. $g(x)=c$ esetén $g(X)$ elfajult eloszlású.

Példák

- $F_{aX+b}(z) = F_X((z-b)/a)$, ha $a > 0$ és
 $F_{aX+b}(z) = 1 - F_X((z-b)/a)$, ha $a < 0$.
 Ebből adódik, hogy ha X abszolút folytonos, és
 $g(z) = az + b$, akkor $g(X)$ sűrűségfüggvénye
 $f_{aX+b}(z) = f_X((z-b)/a) / |a|$.
 Általános eredmény: ha g szigorúan monoton,
 folytonosan deriválható, $g' \neq 0$, akkor

$$f_{g(X)}(z) = \frac{f_X(g^{-1}(z))}{|g'(g^{-1}(z))|}$$

Normális eloszlás

- #### ■ A standard normális eloszlás sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- Legyen m tetszőleges, σ pedig pozitív valós szám. Ha X standard normális eloszlású, akkor az $Y = \sigma X + m$ valószínűségi változó (m, σ) paraméterű normális eloszlású. Ennek sűrűségszűgvénye az

$f_{ax+b}(z) = f_x((z-b)/a)/|a|$ képletből

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

Diszkrét valószínűségi változók várható értéke

- Szerencsejátékban a pontos nyeremény nem látható előre. De: az átlagos nyereményről szeretnénk tudni. (Kedvező-e a játék? Fair játék: az ár éppen a várható érték.)
 - Példa: Dobókocka: annyi a nyereményünk, amennyit dobunk. Ennek átlagos értéke
$$\frac{1}{6}(1+2+\dots+6)=\frac{21}{6}=3.5$$
 - De ha nem szabályos a kocka, például az egyes helyett is 6 van, akkor az átlagos nyeremény
$$\frac{1}{6}(2+\dots+5)+6=\frac{3}{6}=13/3.$$
 - Definíció. A $p_i = P(X=x_i)$ eloszlással megadott szanszíűsík változó várható értéke $E(X) := p_1x_1 + p_2x_2 + \dots$, ha a sor abszolút konvergens.

Példák 2.

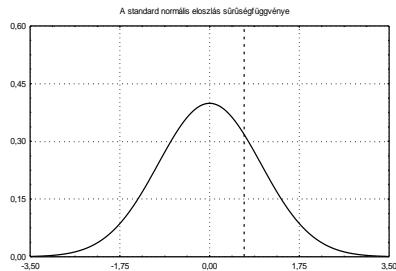
A hipergeometriai eloszlás várható értéke

$$E(X) = \sum_{k=1}^n k \binom{M}{k} \binom{N-M}{n-k} = \sum_{k=1}^n n \frac{M}{N} \binom{M-1}{k-1} \binom{N-1-(M-1)}{n-1-(k-1)} = n \frac{M}{N} \binom{N-1}{n-1}$$

A Poisson eloszlás várható értéke

$$E(X) = \sum_{k=1}^{\infty} k \lambda^k \frac{e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} \lambda^k \frac{e^{-\lambda}}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \lambda^{k-1} \frac{e^{-\lambda}}{(k-1)!} = \lambda$$

A standard normális sűrűségfüggvény



Példák

- Az elfajult eloszlás várható értéke:
 $E(X) = cP(X=c) = c,$
 - A p valószínűségű A esemény indikátorának várható értéke: $E(X)=1P(X=1)=p$
 - Az x_1, x_2, \dots, x_n számokon egyenletes eloszlás (mindegyik valószínűsége $1/n$) várható értéke a számok számtani középe.
 - Az (n,p) paraméterű binomiális eloszlás várható értéke:

$$E(X) = \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n np \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np$$
 - Amerikai roulette. Ha k számra teszünk, a nyeréményünk $36/k$. A várható nettó nyerémény $(36/k) \cdot (k/38) - 1 = -2/38$.

Tulajdonságok

- Nem minden valószínűségi változónak van véges várható értéke:
 $P(X=2^k) = (1/2)^k \quad k=1,2,\dots$
esetén $E(X) = 1+1+1+\dots = \infty$.
 - Azaz annak a játéknak az „ára”, ahol 2^k Ft-ot kapunk, ha szabályos érmével k -adikra dobunk először fejet: végtelen. Ez a Szt. Pétervári paradoxon; gyakorlatban persze nem realis így ez a játék, hiszen nincs az a bank, amely korlátlan pénzt tudna fizetni.
 - Ha $E(X)$ véges, akkor az abszolút konvergencia miatt egyértelmű is.
 - Ha EX véges, akkor $E(aX+b) = aEX+b$
 - Ha FX és FY véges, akkor $F(X+Y) = FX+FY$

Abszolút folytonos eloszlású valószínűségi változók várható értéke

- Az előbb látott határátmenet segítségével (egyre finomabb felosztással közelítjük a folytonos eloszlást) $E(X)=\sum zP(z < X < z+\delta) \approx \sum z\delta f(z) \approx \int zf(z)dz$
- Ebből a definíció: az abszolút folytonos eloszlású X várható értéke: $E(X)=\int_{-\infty}^{\infty} yf_X(y)dy$ ha az integrál létezik.

Tulajdonságok, példák

- Mivel a diszkrét esetből határátmennettel kaptuk a fogalmat, a tulajdonságok (pl. $E(aX+b)=aE(X)+b$, $E(X+Y)=E(X)+E(Y)$ stb.) most is érvényben maradnak.
- Ha X egyenletes eloszlású az $[a,b]$ -ben, akkor

$$E(X)=\int_a^b \frac{y}{b-a} dy = \left[\frac{y^2}{2(b-a)} \right]_a^b = \frac{a+b}{2}$$

További példák

- Ha X exponenciális eloszlású λ paraméterrel, akkor
- $$E(X)=\int_0^{\infty} \lambda y e^{-\lambda y} dy = \left[-ye^{-\lambda y} \right]_0^{\infty} + \int_0^{\infty} e^{-\lambda y} dy = \frac{1}{\lambda}$$
- Ha X standard normális eloszlású, akkor¹⁰
- $$E(X)=\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \left[-\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right]_{-\infty}^{\infty} = 0$$
- Ebből: ha $Y \sim N(\mu, \sigma)$ eloszlású, akkor $E(Y)=E(\sigma X + \mu) = \sigma E(X) + \mu = \mu$
 - Ha a Z változó Q_Z eloszlása keverék-eloszlás (azaz pl. p valószínűsséggel X -et, 1-p valószínűsséggel Y -t figyeljük meg), akkor $E(Z)=pE(X)+(1-p)E(Y)$.

Függvény várható értéke

- Ha X diszkrét: $p_i=P(X=x_i)$, akkor $E(g(X))=p_1g(x_1)+p_2g(x_2)+\dots$, ha a sor abszolút konvergens.
 - Legyen X sűrűségfüggvénye f és $Y=g(X)$ (g Borel mérhető). Ekkor
- $$E(Y)=\int_{-\infty}^{\infty} g(y) f_X(y) dy$$
- Bizonyítás a diszkrét valószínűségi változókra vonatkozó állításból határátmennettel.

Valószínűségszámítás és statisztika előadás Info. BSC B-C szakosoknak

20018/2019 1. félév
Zempléni András
4. előadás

Valószínűségi változók szórásnégyzete

- Nem minden, hogy mekkora a vizsgált véletlen mennyiségek ingadozása.
 - Jobb, ha a buszok pontosan 10 percenként jönnek, mintha időnként 3 jön egymás után, és aztán 30 percet kell várni.
 - Az ingadozás számszerűsítése: a várható értéktől vett átlagos négyzetes eltérés, elnevezése: szórásnégyzet. Formálisan:
- $$D^2(X) := E[(X - E(X))^2].$$
- Kiszámítása: $D^2(X) = E[X^2 - 2XE(X) + E^2(X)] = E(X^2) - 2E(X)E(X) + E^2(X)$
a várható érték linearitása miatt. Azaz
- $$D^2(X) = E(X^2) - E^2(X).$$

Tulajdonságok

- $D^2(X) \geq 0$, mert nemnegatív valószínűségi változó várható értéke.
- $D^2(aX+b) = a^2 D^2(X)$, mert $D^2(aX+b) = E[(aX+b-E(aX+b))^2] = E[(aX+b-aE(X)-b)^2] = E[(aX-aE(X))^2] = a^2 E[(X-E(X))^2]$.
- Abból, hogy $E(X)$ véges, még nem következik $D^2(X)$ végesére, hiszen ha $P(X=k) = c/k^3$ (egyértelműen megadható olyan c , amire ez eloszlás lesz) akkor $E(X)$ véges, de $E(X^2) = c(1+1/2^2+\dots+1/k^2+\dots)$, ami végtelen.

Példák

- Az elfajult eloszlás szórásnégyzete:
 $D^2(X) = E(X^2) - E^2(X) = c^2 - c^2 = 0$.
- Megfordítás: ha $D^2(X) = 0$, akkor X 1 valószínűséggel konstans.
Biz.: $(X-E(X))^2 \geq 0$, várható értéke 0, tehát ő maga is 1 valószínűséggel 0, azaz $X = E(X)$ 1 valószínűséggel.
- A p valószínűségű A esemény indikátorának szórásnégyzete:
 $D^2(X) = E(X^2) - E^2(X) = p \cdot p^2 = p(1-p)$. Azaz $p=0.5$ esetén a legnagyobb a szórásnégyzet.
- A kockadobás szórásnégyzete:
 $D^2(X) = E(X^2) - E^2(X) = (1+4+\dots+36)/6 - 49/4 = 91/6 - 49/4 = (182-147)/12 = 35/12$.

Példák 2

A Poisson eloszlás szórásnégyzete:

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2 \lambda^k \frac{e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k \lambda^k \frac{e^{-\lambda}}{(k-1)!} = \lambda \sum_{k=1}^{\infty} (k-1+1) \lambda^{k-1} \frac{e^{-\lambda}}{(k-1)!} \\ &= \lambda^2 \sum_{k=2}^{\infty} \lambda^{k-2} \frac{e^{-\lambda}}{(k-2)!} + \lambda = \lambda^2 + \lambda. \end{aligned}$$

Ebből

$$D^2(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Azaz a Poisson eloszlás várható értéke és szórásnégyzete megegyezik.

Példák 3

- Ha X egyenletes eloszlású az $[a,b]$ intervallumon, akkor
- $$E(X^2) = \int_a^b y^2 dy = \left[\frac{y^3}{3(b-a)} \right]_a^b = \frac{a^2 + ab + b^2}{3}$$
- $$D^2(X) = E(X^2) - E^2(X) = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{a^2 - 2ab + b^2}{12} = \frac{(a-b)^2}{12}$$
- Ha X exponenciális eloszlású, akkor
- $$E(X^2) = \int_0^{\infty} y^2 \lambda e^{-\lambda y} dy = \left[-y^2 e^{-\lambda y} \right]_0^{\infty} + \int_0^{\infty} 2y e^{-\lambda y} dy = \frac{2}{\lambda^2}$$
- $$D^2(X) = E(X^2) - E^2(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Összeg szórásnégyzete

- $D^2(X+Y) = E[(X+Y-E(X+Y))^2] = E[(X-E(X)+Y-E(Y))^2] = E[(X-E(X))^2] + E[(Y-E(Y))^2] + 2E[(X-E(X))(Y-E(Y))] = D^2(X) + D^2(Y) + 2E[(X-E(X))(Y-E(Y))]$
 - Példák:
 - $X=Y$ esetén $D^2(X+Y) = D^2(2X) = 4D^2(X)$
 - $X=-Y$ esetén $D^2(X+Y) = D^2(0) = 0$
- azaz nem csak X és Y egydimenziós eloszlásától, hanem az együttes viselkedésüktől, azaz az együttes eloszlásuktól is függ az összegük szórásnégyzete.

Valószínűségi változók függetlensége

- X és Y valószínűségi változók függetlenek, ha $P(X \in A \cap Y \in B) = P(X \in A)P(Y \in B)$ minden $A, B \subset \mathbb{R}$ -beli Borel halmazra
- diszkrét esetben ehhez elég, ha $P(\{X=x_i\} \cap \{Y=y_j\}) = P(X=x_i)P(Y=y_j)$ teljesül minden i, j értékre. Azaz az X -hez és az Y -hoz tartozó teljes eseményrendszer függetlenek.
- Megjegyzés:
 - az elfajult eloszlású valószínűségi változót minden valószínűségi változótól független.
 - Önmagától csak az elfajult eloszlású valószínűségi változót független.

A független val. változók esete

- Állítás. Ha X, Y függetlenek és $E(X), E(Y)$ végesek, akkor $E(XY) = E(X)E(Y)$.
 - Bizonyítás. $E(XY) = \sum_{k,m} x_k y_m P(X = x_k, Y = y_m)$
- ami a függetlenség miatt így írható:
- $$= \sum_k x_k P(X = x_k) \sum_m y_m P(Y = y_m) = E(X)E(Y).$$
- Ebből: $D^2(X+Y) = D^2(X) + D^2(Y)$, ha X és Y függetlenek.
 Indukcióval (páronként független val. változókra):
- $$D^2(X_1 + \dots + X_n) = \sum_{i=1}^n D^2(X_i)$$
- És így a binomiális eloszlás szórásnégyzete: $np(1-p)$.

A szórás

- Szórásnégyzet mértékegysége az eredeti X mértékegységének a négyzete (azaz pl. a buszok követési időközénél négyzetperc). Ez nem teszi egyszerűvé interpretációját.
- Szórás: $D(X)$ a szórásnégyzet pozitív négyzetgyöke. Ez már a megfelelő mértékegységű, $D(aX) = |a|D(X)$.

A normális eloszlás szórásnégyzete

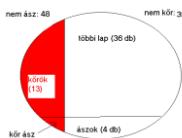
- Ha X standard normális eloszlású, akkor
- $$E(X^2) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{\infty} x(x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}) dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1$$
- ebből $D^2(X) = 1$, hiszen $E(X) = 0$.
- Tetszőleges (m, σ) paraméterű normális eloszlásra $D^2(Y) = \sigma^2$, hiszen $D^2(\sigma X + m) = \sigma^2 D^2(X)$.

Valószínűségi vektorváltozók

- $X = (X_1, X_2, \dots, X_d) : \Omega \rightarrow \mathbb{R}^d$ függvény valószínűségi vektorváltozó, ha $\{\omega : X(\omega) \in B\} \in \mathcal{A}$ minden B d -dimenziós Borel halmazra. (Pontosan akkor teljesül, ha X_i valószínűségi változó minden $1 \leq i \leq d$ -re.)
- $Q_x(B) := P\{\omega : X(\omega) \in B\}$ az X eloszlása \mathbb{R}^d Borel-halmazain.
- Ennek megadásához elegendő a $F_{X_i}(z) := P(X_i < z)$ valószínűségeket megadni ($z \in \mathbb{R}^d$), a $<$ reláció koordinátánként értendő, azaz $X_i < z$ pontosan akkor teljesül, ha $X_i < z$, minden $1 \leq i \leq d$ -re. Ezek meghatározzák $Q_x(B)$ értékét tetszőleges B -re (nem bizonyítjuk).

Diszkrét kétdimenziós eloszlás megadása táblázattal

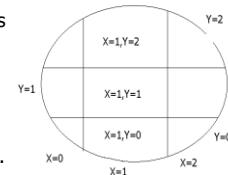
Példa: kétszer húzunk visszatevéssel egy francia kártyacsomagból



kör\ ász	0	1	2
0	$(36/52)^2$	$2 \cdot 3 \cdot 36 / (52)^2$	$(3/52)^2$
1	$2 \cdot 12 \cdot 36 / (52)^2$	$2 \cdot (12 \cdot 3 + 1 \cdot 36) / (52)^2$	$2 \cdot 3 / (52)^2$
2	$(12/52)^2$	$2 \cdot 12 / (52)^2$	$(1/52)^2$

A peremeloszlások

- (X,Y) eloszlásából (elnevezés: *együtteloszlás*) következtethetünk az egyes változók eloszlására: $P(X=1)=P(X=1, Y=0)+P(X=1, Y=1)+P(X=1, Y=2)$
- Az egyes koordináták (alacsonyabb dimenziós vektorok) eloszlása: *peremeloszlás*.
- Az együttes eloszlás tehát meghatározza a peremeloszlást (a megfordítás nem igaz).



Együttes eloszlásfüggvény

- Az $F_X(z):=P(X < z)$ $\mathbf{R}^d \rightarrow \mathbf{R}$ függvény az X valószínűségi vektorváltozó együttes eloszlásfüggvénye.
- Az egymenziós esettel analóg tulajdonságai:
 - $0 \leq F_X(z) \leq 1$
 - $F_X(z)$ minden koordinátájában monoton növő
 - $\lim_{z_j \rightarrow \infty} F_X(z) = 1$, ha z minden koordinátájára $z_j \rightarrow \infty$
 - $\lim_{z_j \rightarrow -\infty} F_X(z) = 0$ ha z legalább egy koordinátájára $z_j \rightarrow -\infty$
 - $F_X(z)$ minden koordinátájában balról folytonos.

Az együttes eloszlásfüggvény további tulajdonságai

- Téglatestek valószínűségei:
 - $P(a \leq X < b) \geq 0$ minden $a < b \in \mathbf{R}^d$ -re. Ez kifejezhető az X eloszlásfüggvényével:
 - $d=2$ -re: $P(a \leq X < b) = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2)$.
- Tetszőleges, a felsorolt összes tulajdonsággal rendelkező F-hez létezik X d-dimenziós vektorváltozó, aminek F az eloszlásfüggvénye.
- A peremeloszlások meghatározása

$$\lim_{x \rightarrow \infty} F_{X,Y}(x,y) = F_Y(y), \quad \lim_{y \rightarrow \infty} F_{X,Y}(x,y) = F_X(x)$$

Sűrűségfüggvény

- Ha létezik $f: \mathbf{R}^d \rightarrow \mathbf{R}$ függvény, hogy F előáll f integrálfüggvényének:

$$F(z) = \int_{-\infty}^z f(t) dt$$

akkor azt mondjuk, hogy F abszolút folytonos, f sűrűségfüggvényével. Az integrál most d-dimenziós, értelmezése:

$$F(z) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \dots \int_{-\infty}^{z_d} f(t_1, t_2, \dots, t_d) dt_d \dots dt_2 dt_1$$

A peremeloszlások sűrűségfüggvénye

- Legyen $d=2$. Ha (X,Y) abszolút folytonos, $f(x,y)$ együttes sűrűségfüggvény, akkor X sűrűségfüggvénye

$$g_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$$

- Bizonyítás.

$$\int_{-\infty}^{z_1} \int_{-\infty}^{z_2} f_{X,Y}(x,y) dy dx = F_{X,Y}(z_1, z_2) = P(X < z_1, Y < z_2)$$

- Ugyanígy Y sűrűségfüggvénye

$$h_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$$

A függetlenség esete

- Ha X koordinátái függetlenek, akkor definíció szerint
- $F_X(z) = P(X_1 < z_1, X_2 < z_2, \dots, X_d < z_d) = F_1(z_1)F_2(z_2)\dots F_d(z_d)$ (minden $z \in \mathbb{R}^d$ -re). Meg is fordítható: F szorzatelőállításából következik a függetlenség.
- Deriválva: a függetlenség abszolút folytonos változókra ekvivalens a sűrűségfüggvény $f_X(z) = f_1(z_1)f_2(z_2)\dots f_d(z_d)$ alakú előállításával is.
- Példa: az egységnégyzeten egyenletes eloszlás sűrűségfüggvénye ($f(z)=1$ ha $0 < z < 1$) előáll $f_1(z_1)f_2(z_2)$ alakban, ahol $f_i(z_i)=1$, ha $0 < z_i < 1$ ($i=1,2$), ez éppen a $[0,1]$ intervallumon egyenletes eloszlás.

Kovariancia

- Definíció. Az X és Y kovarienciája:
 $\text{cov}(X,Y) := E[(X-E(X))(Y-E(Y))]$
- Kiszámítása: $\text{cov}(X,Y) = E[XY-XE(Y)-YE(X)+E(X)E(Y)] = E(XY)-E(X)E(Y)$
- A korábban független diszkrét változókra látott, $E(XY)=E(X)E(Y)$ egyenlőség abszolút folytonosakra is igaz – ennek értelmében $\text{cov}(X,Y)=0$, ha X és Y függetlenek.
- Megj.: Abból, hogy $\text{cov}(X,Y)=0$ nem következik, hogy függetlenek: legyen X szimmetrikus a $0=ra$ (pl. $P(X=1)=P(X=-1)=P(X=0)=1/3$) és $Y=X^2$. Ekkor $\text{cov}(X,Y)=E(X^3)-E(X)E(X^2)=0-0$, hiszen $E(X^3)=E(X)=0$.
- A kovariancia szimmetrikus: $\text{cov}(X,Y) = \text{cov}(Y,X)$
- $\text{cov}(X,X) = D^2(X)$

Valószínűségszámítás és statisztika előadás Info. BSC B-C szakosoknak

20018/2019 1. félév
Zempléni András
5. előadás

Kovariancia

- Definíció. Az X és Y kovarienciája:
 $\text{cov}(X,Y) := E[(X-E(X))(Y-E(Y))]$
- Kisázmítása: $\text{cov}(X,Y) = E[XY] - E[X]E[Y] - YE[X] - E[X]E[Y] = E[XY] - 2E[X]E[Y]$
- A korábban független diszkrét változókra látott, $E(XY) = E(X)E(Y)$ egyenlőség abszolút folytonosakra is igaz – ennek értelmében $\text{cov}(X,Y) = 0$, ha X és Y függetlenek.
- Megj.: Abból, hogy $\text{cov}(X,Y) = 0$ nem következik, hogy függetlenek: legyen X szimmetrikus a 0=ra (pl. $P(X=1)=P(X=-1)=P(X=0)=1/3$) és $Y=X^2$. Ekkor $\text{cov}(X,Y) = E(X^3) - E(X)E(X^2) = 0 - 0$, hiszen $E(X^3) = E(X) = 0$.
- A kovariancia szimmetrikus: $\text{cov}(X,Y) = \text{cov}(Y,X)$
- $\text{cov}(X,X) = D^2(X)$

Összeg szórásnégyzete

- Láttuk: $D^2(X+Y) = D^2(X) + D^2(Y)$, ha X és Y függetlenek (elég, hogy $\text{cov}(X,Y) = 0$).
- Általánosan: $D^2(X+Y) = D^2(X) + D^2(Y) + 2\text{cov}(X,Y)$
- n tagú összegre:
$$D^2(X_1 + \dots + X_n) = \sum_{i=1}^n D^2(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{cov}(X_i, X_j)$$
- Spec.: $D^2(X_1 + X_2 + \dots + X_n) = D^2(X_1) + D^2(X_2) + \dots + D^2(X_n)$, ha a tagok páronként függetlenek.

Korrelációs együttható

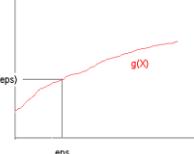
- A kovariancia skálafüggő: $\text{cov}(aX, bY) = ab \cdot \text{cov}(X, Y)$
- A változók közötti lineáris kapcsolat erősségeit mérő mennyiségek a *korrelációs együttható*:
$$R(X, Y) = \frac{\text{cov}(X, Y)}{D(X)D(Y)}$$
- Tulajdonságai:
 - $R(X, Y) = 0$, ha X és Y függetlenek (ez sem fordítható meg)
 - Ez alapján definíció szerint legyen $R(X, Y) = 0$, ha X vagy Y elfajult eloszlású.
 - $R(X, aX+b) = 1$, ha $a > 0$, mert $\text{cov}(X, aX+b) = aD^2(X)$.

A korreláció tulajdonságai

- $|R(X, Y)| \leq 1$ és $|R| = 1$ akkor és csak akkor, ha $X = aY + b - 1$ valószínűséggel ($a \neq 0, b \in \mathbb{R}$).
- Ehhez: $X^* = \frac{X - E(X)}{D(X)}, Y^* = \frac{Y - E(Y)}{D(Y)}$
 a standardizált változók. $E(X^*) = E(Y^*) = 0$,
 $D(X^*) = D(Y^*) = 1$. $R(X, Y) = E(X^*Y^*)$.
 $0 \leq E(X^*Y^*)^2 = E(X^{*2}) + 2E(X^*Y^*) + E(Y^{*2}) = 2 \pm 2E(X^*Y^*)$, tehát $|R(X, Y)| \leq 1$.
 Ebből: $R=1$ akkor és csak akkor, ha $0 = E(X^*Y^*)^2$, azaz $X^* = Y^*$ 1 valószínűséggel. Ekkor $X = aY + b$, $a > 0$.
 $R=-1$ akkor és csak akkor, ha $0 = E(X^*Y^*)^2$, azaz $X^* = -Y^*$ 1 valószínűséggel. Ekkor $X = aY + b$, $a < 0$.

Markov-típusú egyenlőtlenségek

- Legyen $X \geq 0$ valószínűségi változó,
 $g: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ monoton növő. Ekkor
 $P(X \geq \varepsilon) \leq E(g(X))/g(\varepsilon)$.
- Bizonyítás.
 $E(g(X)) \geq g(\varepsilon)P(X \geq \varepsilon)$
 mert $X \geq \varepsilon$ eseményen
 $g(X) \geq g(\varepsilon)$



Alkalmazások

- $g(x)=x$: Ha $X \geq 0$ valószínűségi változó, akkor $P(X \geq \varepsilon) \leq E(X)/\varepsilon$ (ezt nevezik Markov egyenlőtlenségnek).
- $g(x)=x$, X helyett $(X-E(X))^2$ -re alkalmazva:
 $P((X-E(X))^2 \geq \varepsilon^2) \leq E(X-E(X))^2/\varepsilon^2$, ami egyszerűsítve $P(|X-E(X)| \geq \varepsilon) \leq D^2(X)/\varepsilon^2$ (elnevezés: Csebisev egyenlőtlenség).
- Megjegyzés. A fenti egyenlőtlenségek élesek, azaz minden ε -ra megadható olyan valószínűségi változó, amelyre $P(X \geq \varepsilon) = E(X)/\varepsilon$. (Két értéket vesz fel: 0, ε).

Alkalmazások/2

- Az eredmények a gyakorlatban mégsem adnak kellően pontos becslést, mert a tényleges eloszlások tulajdonságait nem veszi figyelembe. Ezért, ha ismerjük az adott változó eloszlását, minden abból adjuk meg a valószínűségek értékét.
- Példa: hányszor kell egy szabályos érmét feldobni, hogy a fejek relatív gyakorisága legalább 0.99 valószínűséggel ne terjen el 0.05-nél jobban 0.5-től? Csebisev egyenlőtlenségből:
 $P(|X-0.5| \geq 0.05) \leq D^2(X)/\varepsilon^2 = 400/4n \leq 0.01$ elég, amiből $n \geq 10000$ adódik. A binomiális eloszlásból adódó pontos érték: 670. [Szimuláció](#)

Nagy számok törvényei

- Legyenek X_1, X_2, \dots, X_n független, azonos eloszlású valószínűségi változók. Tegyük fel, hogy $\sigma^2 = D^2(X)$ véges ($m := E(X)$). Ekkor minden $\varepsilon, \delta > 0$ -hoz megadható olyan n_0 , hogy $n > n_0$ esetén $P(|(X_1 + X_2 + \dots + X_n)/n - m| \geq \varepsilon) \leq \delta$.
- Bizonyítás. $E(X_1 + X_2 + \dots + X_n)/n = m$, és $D^2[(X_1 + X_2 + \dots + X_n)/n] = \sigma^2/n$. A Csebisev egyenlőtlenség miatt $P(|(X_1 + X_2 + \dots + X_n)/n - m| \geq \varepsilon) \leq \sigma^2/\varepsilon^2 n$, ami 0-hoz tart, azaz elég nagy n -re kisebb lesz δ -nál.
- Elnevezés: $(X_1 + X_2 + \dots + X_n)/n \rightarrow m$ sztochasztikusan (az ilyen konverenciát bizonyító tételeket gyenge tételeknek nevezünk).

Megjegyzések.

- A tétel feltételei gyengíthetők: elég, ha a független, azonos eloszlású változók várható értéke véges.
- Az állítás is erősíthető: 1 valószínűségű konvergencia is bizonyítható. Ez azt jelenti, hogy $P\{\omega: (X_1 + X_2 + \dots + X_n)/n \rightarrow m\} = 1$.
- Ha $\Omega = [0, 1]$ és $P(A)$ az A "hossza", akkor az 1 valószínűségű konvergencia lényegében a szokásos pontonkénti konvergencia. Ez nem következik a sztochasztikus konverenciából.
- A törvény nem jelenti azt, hogy az eddig nem szerepelt értékek a jövőben a vártól gyakoribbak lesznek, hanem csupán az eloszlás szerint kapott nagyszámú érték állítja helyre a várt gyakoriságokat.

További tételek

- A nagy számok törvényének legelső verzióját még Bernoulli bizonyította, indikátorváltozóra: eszerint azonos körülmények között elvégzett független kísérleteknél tetszőleges esemény relatív gyakorisága tart az esemény valószínűségéhez. (Az előző speciális esete: X indikátorváltozó.)
- [Szimuláció](#)
- Gyengén összefüggő valószínűségi változókra is átvihető a téTEL

A nagy számok törvényének néhány alkalmazása

- Korlátos valószínűségi változóra teljesül a nagy számok erős törvénye.
- Monte Carlo módszerek: véletlen számokat használnak
 - A $[0, 1]$ intervallumon egyenletes eloszlású, egymástól független véletlen számokat szimulálunk: X_1, X_2, \dots és ezek segítségével közelíthetünk például integralokat:
$$\frac{f(X_1) + f(X_2) + \dots + f(X_n)}{n} \rightarrow E(f(X)) = \int_0^1 f(x) dx$$
 - Az egyenletes eloszlásból más eloszlások is megkaphatók

Konvolúció

- Független valószínűségi változók összegének eloszlása
- Példák diszkrét eloszlásokra
 - Független, azonos paraméterű indikátorok konvolúciója binomiális.
 - Azonos p paraméterű binomiálisok konvolúciója is binomiális (a kísérletszámok összeadódnak)
 - Poisson eloszlások konvolúciója is Poisson (a paraméterek összeadódnak)
 - Pascal eloszlások konvolúciója negatív binomiális
- Példák folytonos eloszlásokra
 - Normális eloszlások konvolúciója is normális (a várható érték és a szórásnégyzet is összeadódik)
 - Azonos paraméterű exponenciális eloszlások konvolúciója: [gamma eloszlás](#)

Valódi határeloszlások

- Kérdés: lehet-e nemelfajult valószínűségi változó a határérték?
- Tétel: ha X_1, X_2, \dots független valószínűségi változók, b_n számsorozat, melyre $b_n \rightarrow \infty$ és $(X_1 + X_2 + \dots + X_n)/b_n \rightarrow X$ 1 valószínűsséggel, akkor X 1 valószínűsséggel állandó.

Gyenge (eloszlásbeli) konvergencia

- Definíció. $X_n \rightarrow X$ gyengén, ha az eloszlásfüggvényeikre teljesül: $F_n(z) \rightarrow F(z)$ az F minden folytonossági pontjában.
- Megjegyzés. Ez a konvergencia nem mond semmit a valószínűségi változók közelsgéről. $\Omega = [0, 1]$, $P = \text{"hosszúság"}$, $X_n = I_{[0, 0.5]}$ esetén $F_n(z) = F(z)$, azaz teljesül a gyenge konvergencia.
- A gyenge konvergencia következik a sztochasztikusból.
- A fentiekből az is látszik, hogy a határértéknek csak az eloszlása érdekes.

Tulajdonságok

- Azt nem célszerű megkövetelni, hogy F minden pontjában teljesüljön a konvergencia:
- $X_n = \delta_{1/n}$ (az 1/n értéket felvező elfajult eloszlású v.v.) esetén $X_n \rightarrow X = \delta_0$ 1 valószínűsséggel. $F_n(0) = 1$, de $F(0) = 0$ (F balról folytonos). A többi pontban teljesül a konvergencia: $F_n(z) \rightarrow 0$, ha $z < 0$, $F_n(z) = 1$, ha $z > 0$.
- Ha $X_n \rightarrow X$ sztochasztikusan, akkor $X_n \rightarrow X$ gyengén is.

Centrális határeloszlás téTEL

- Legyenek $X_1, X_2, \dots, X_n, \dots$ független, azonos eloszlású valószínűségi változók. Tegyük fel, hogy $\sigma^2 = D^2(X)$ véges ($m := E(X)$). Tekintsük a standardizált összegüket:

$$Z_n := \frac{X_1 + \dots + X_n - nm}{\sqrt{n}\sigma}$$

Ekkor Z_n gyengén konvergál a standard normális eloszláshoz, azaz

$$P\left(\frac{X_1 + \dots + X_n - nm}{\sqrt{n}\sigma} < z\right) \rightarrow \Phi(z)$$

ahol Φ a standard normális eloszlás eloszlásfüggvénye.

Általánosítások

- Ha nem azonos eloszlásúak a tagok, további feltételekre (pl. magasabb momentumok létezése, hasonló nagyságrendű összeadandók) van szükség.
- Gyenge összefüggőség esetére is általánosítható a téTEL.

Stabilis eloszlások

- A normális eloszlás kitüntetett szerepe azon alapult, hogy teljesítette az ún. *stabilitást*: ha X, Y független, F eloszlásfüggvényük, akkor tetszőleges a, b esetén megadhatók α, β számok, hogy $aX+bY$ eloszlásfüggvénye $F(\alpha z+\beta)$. (Azaz $aX+bY$ ugyanabba az eloszláscsaládba tartozik, mint az összeadandók.)
- Belátható, hogy független, azonos eloszlású változók összegének normális utáni határeloszlása csak stabilis lehet. Ugyanakkor minden ilyen stabilis eloszlás elői is áll határeloszlásként.
- A centrális határeloszlás tétele következménye, hogy nincs más véges szórású stabilis eloszlás. Ugyanakkor nem véges szórású van: pl. a Cauchy eloszlás, melynek sűrűségfüggvénye $f(x)=1/\pi(1+x^2)$

A centrális határeloszlás tétele lokális változata

- Legyenek $X_1, X_2, \dots, X_n, \dots$ független, azonos eloszlású valószínűségi változók. Tegyük fel, hogy $\sigma^2 = D^2(X) = E(X^2) - E(X)^2$ véges ($n := E(X_i)$) valamint, hogy abszoluton folytonosak, szakaszonként folytonos sűrűségfüggvények. Tekintsük a standardizált összegüket ($Z_n = \frac{X_n - \mu}{\sigma}$) és tegyük fel, hogy legalább egy n -re Z_n sűrűségfüggvénye korlátos. Ekkor a Z_n változó f_n sűrűségfüggvénye (mely a tagok abszoluton folytonosságára miatt létezik) konvergál a standard normális eloszlás sűrűségfüggvényéhez, azaz

$$f_n(t) \rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

Az indikátorváltozók esete

- Tétel (Moivre-Laplace) Legyenek $X_1, X_2, \dots, X_n, \dots$ független, azonos p paraméterű indikátorváltozók. Ekkor $Y = X_1 + X_2 + \dots + X_n$ binomiális eloszlású (n, p) paraméterrel.

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}$$

- A közelítés egyenletesen jó olyan k értékekre, melyek legfeljebb $(np(1-p))^{1/2}$ -vel térik el np -től (azaz a binomiális eloszlás centrális tagjai jól közelíthetők a megfelelő normális eloszlás sűrűségfüggvényével).
- [binomiális eloszlás](#)

Valószínűségszámítás és statisztika előadás info. BSC/B-C szakosoknak

6. előadás
október 16.

A matematikai statisztika tárgya

- Következtetések levonása adatok alapján
 - Ipari termelés
 - Mezőgazdaság
 - Szociológia (közvéleménykutatások)
 - Természettudományok
 - Meteorológia (pl. klímaváltozás)
 - Genetika (chiptechnológia)
 - Pénzügyi adatok stb.

Történet

- Táblázatokat a biztosítók már többszáz éve használnak
- Maga a tudomány fiatal tudomány, alig 100 éves a múltja
 - Angliai mezőgazdasági alkalmazások voltak az elsők
- Fejlődése felgyorsult az utóbbi évtizedekben (számítógépek jóvoltából)

Populáció

- Az a sokaság, aminek a jellemzőire kiváncsiak vagyunk.
- Példák:
 - Gyártmányok
 - Magyarország szavazópolgárai
 - A Ft/Euro árfolyam napi változásai
- Legtöbbször nincs mód teljes körű (100%-os) adatfelvételre.

Minta

- A populációból kiválasztott részhalmaz, amelyre vonatkozóan az adatok rendelkezésre állnak. Mivel a mintavétel véletlen, ezért a mintaelemek valószínűségi változók.
- Fontos szempont a reprezentativitás.
- Gyakorlatban legtöbbször feltesszük, hogy a mintaelemek függetlenek.

Adatok

- Mintavétel a populációból: eredménye a (statisztikai) minta
- A mintavétel módja is lényeges (legegyszerűbb eset: bármelyik elem ugyanakkora valószínűséggel kerül a mintába)
- Példa: Nem jó, ha a büfében kérdezzük meg a diákokat az előadásról (nem lesz reprezentatívv)
- A mintavétel eredménye: (statisztikai) minta: x_1, x_2, \dots, x_n számsorozat, az X_1, X_2, \dots, X_n valószínűségi változó-sorozat realizációja.

Matematikai statisztika helye a tudományok között

- Matematikai tudomány, mert a valószínűségszámítás eredményeire épül.
- Ugyanakkor a statisztika minden nap alkalmazása nem mindig kellően precíz (teljesülnek-e a feltételek?) Ezért lényeges, hogy a valószínűségszámítási eredményeket alkalmazva fogalmazzuk meg következtetéseinket.

Példák

1. Egy hónapban 10 hurrikánt figyeltünk meg. Mit gondolunk, mennyi hurrikán lesz jövőre ugyanebben a hónapban?
2. Egy közvéleménykutatás során azt kaptuk, hogy 1000 emberből 400 választaná az adott pártot. Mások szerint a párt 50%-ot fog kapni. Előfordulhat-e ez? Mekkora eséllyel?

Statisztikai elemzés lépései

- Tervezés (mit vizsgálunk, hogyan gyűjtjük az adatokat)
- Adatgyűjtés
- Kódolás (ha szükséges)
- Ellenőrzés: leíró statisztikákkal
- Elemzés: matematikai statisztika módszereivel

Leíró statisztika

- Nem a véletlen hatását vizsgálja, hanem a konkrét minta
 - megjelenítése,
 - jellemzőinek kiszámításaa feladata.
- Adatok elrendezhetők táblázatban (fontos: forrás feltüntetése), illetve ábrázolhatók grafikusan.

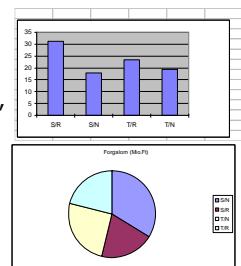
Adatok típusai (skálák)

- Nominális: csak gyakoriságot tudunk számolni (pl. nem, nemzetiség)
- Ordinális (rendezett): pl. értékelés szavakkal (rossz-közepes-jó), sorrend egyértelmű, kvantilisek számolhatók
- Intervallum (pl. hőmérséklet): különbség egyértelmű, de hányados nem)
- Arány (itt minden matematikai művelet értelmes), ez szerencsére a leggyakoribb

Grafikus megjelenítés

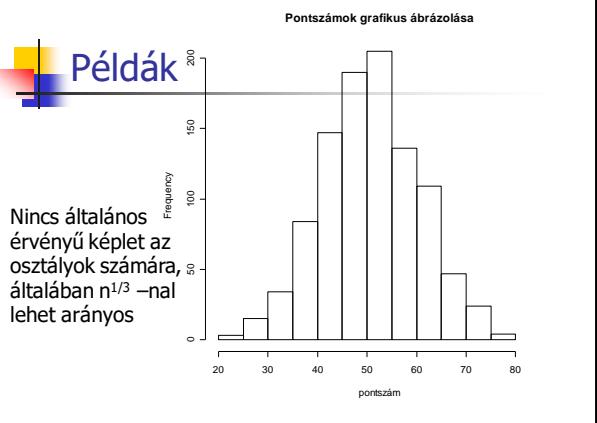
- Ne legyen túl bonyolult!
- Példák:
 - oszlopdiagram
 - X tengely: csoportok, típusok
 - Y tengely: gyakoriságok, értékek
- kördiagram

Heti forgalom, MFt, XXZZ áruház

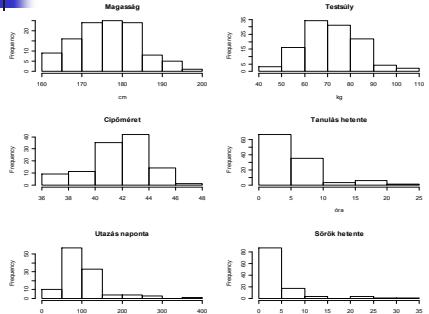


Hisztogram

- Adatainkat osztályokba soroljuk (mindegyiket pontosan egybe, pl. az i-edik osztály: $a_i \leq x < a_{i+1}$), a csoportok relatív gyakoriságai megegyeznek az osztály fölött rajzolt téglalap területével.
- Összterület: 1



Hallgatói adatok

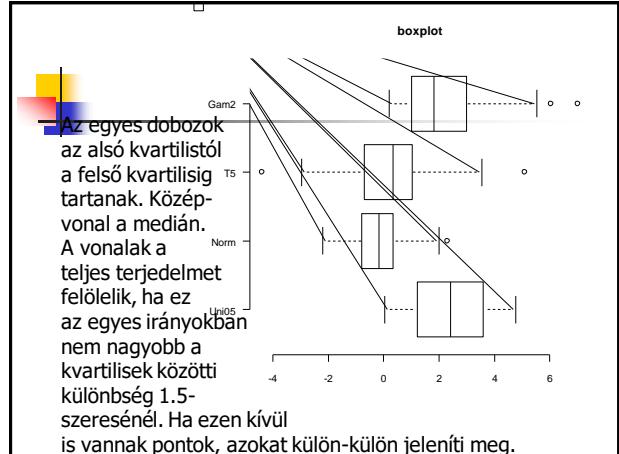


Középértékek

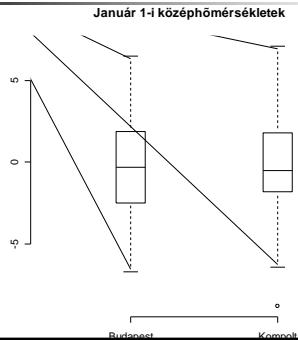
- Mintaátlag: $\bar{x} := \frac{x_1 + \dots + x_n}{n}$
- ha az egyes értékek (f_i) gyakoriságai (f_i) adottak: $\bar{x} := \frac{f_1 l_1 + \dots + f_k l_k}{\sum f_i}$
- Medián: a sorbarendezett minta középső eleme (ha páros sok eleme van: a két középső átlaga).
- Kvartilisek: negyedelőpontok (1/4-3/4, illetve 3/4-1/4 arányban osztják fel a rendezett mintát)
- Az átlag érzékeny a kiugró értékekre, a medián viszont nem.

Hallgatói adatok

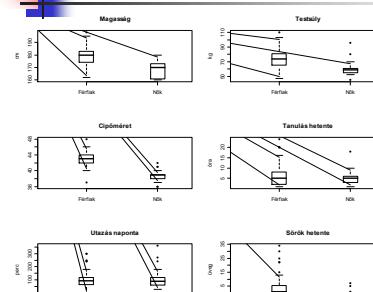
V1	V2	V3	V4	V5
Min. : 160.0	Min. : 45.00	Min. : 36.00	F: 95	Min. : 1.000
1st Qu.: 172.0	1st Qu.: 64.00	1st Qu.: 41.00	N: 17	1st Qu.: 2.000
Median : 178.0	Median : 72.00	Median : 43.00		Median : 5.000
Mean : 177.2	Mean : 72.18	Mean : 42.28		Mean : 6.036
3rd Qu.: 182.0	3rd Qu.: 80.25	3rd Qu.: 44.00		3rd Qu.: 8.000
Max. : 198.0	Max. : 110.00	Max. : 48.00		Max. : 24.000
V6	V7			
Min. : 0.0	Min. : 0.000			
1st Qu.: 60.0	1st Qu.: 0.000			
Median : 92.5	Median : 1.000			
Mean : 104.1	Mean : 3.527			
3rd Qu.: 120.0	3rd Qu.: 5.000			
Max. : 360.0	Max. : 34.000			



Példa adatbázis: Napi középhőméréséletek 1951-1988 között



A hallgatói adatok nemenkénti bontásban



Vajon melyik esetben szignifikáns az eltérés?

Matematikai statisztika

- A minta: valószínűségi változó-sorozat realizációja. A belőle számolt statisztikák eloszlásának vizsgálatához magukat a valószínűségi változókat használjuk, nem pedig a realizationánál kapott számtéreket.
- Statisztika: a minta függvénye (val.változó).
- Példák statisztikára:
 - minimum, maximum, mintaátlag
 - terjedelem: $X_{(n)} - X_{(1)}$

Becslések

- A mintából kiszámolt értékek tekinthetők a vizsgált populációra vonatkozó közelítéseknek.
- Ezek tulajdonságait (mennyire pontosak/megbízhatóak) a valószínűségszámítás eszközeivel tudjuk vizsgálni.

Becslések tulajdonságai

- *Torzítatlanság.* θ valós paramétert becslünk a $T(\underline{X})$ statisztikával. Ez torzítatlan, ha ,
 $E_\theta(T(\underline{X})) = \theta$
 minden θ paraméterértékre.
- Példák torzítatlan becslésekre:
 - Valószínűség becslése relatív gyakorisággal.
 - Várható érték becslése mintaátlaggal
 - Poisson eloszlás paramétere: mintaátlag

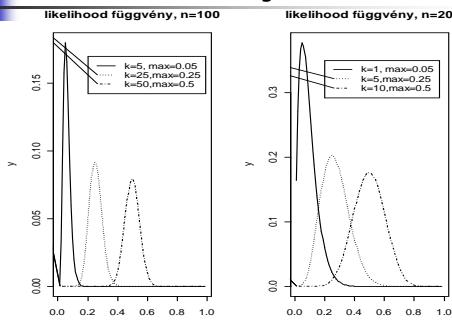
Becslési módszerek

- Eddig: „ad hoc” módszerek
- Általános eljárás kellene
 - Példa: valószínűség becslése, n kísérletből. Jelölje k a sikeresek számát (X_i $i=1,\dots,n$ indikátor minta)

$$P\left(\sum_{i=1}^n X_i = k\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

Most p függvényében nézzük, k rögzített (elnevezés: likelihood függvény).

A likelihood függvény maximumhelye logikus választás a valószínűség becslésének



A módszer általánosan

$$L(\theta; \underline{x}) = f_{\theta}(\underline{x}) = \prod_{i=1}^n f_{\theta}(x_i)$$

(a likelihood függvény) maximumhelye lesz a θ paraméter maximum likelihood becslése.

Ha a függvény deriválható, a loglikelihood függvény

$$l(\theta; \underline{x}) = \ln f_{\theta}(\underline{x}) = \sum_{i=1}^n \ln f_{\theta}(x_i)$$

maximumhelye deriválással

$$\frac{\partial}{\partial \theta} l(\theta; \underline{x}) = \frac{\partial}{\partial \theta} \ln f_{\theta}(\underline{x}) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f_{\theta}(x_i) = 0$$

megoldásaként megtalálható

Példák

- valószínűségre: relatív gyakoriság
- Poisson eloszlás paraméterére: \bar{x}
- Exponenciális eloszlás paraméterére: $1/\bar{x}$

Valószínűségszámítás és statisztika előadás info. BSC/B-C szakosoknak

7. előadás
November 6.

Becslések tulajdonságai

- *Torzítatlanság.* θ valós paramétert becslünk a $T(\underline{X})$ statisztikával. Ez torzítatlan, ha, $E_{\theta}(T(\underline{X})) = \theta$
minden θ paraméterértékre.
- Példák torzítatlan becslésekre:
 - Valószínűség becslése relatív gyakorisággal
 - Várható érték becslése mintaátlaggal
 - Poisson eloszlás paraméterére: mintaátlag

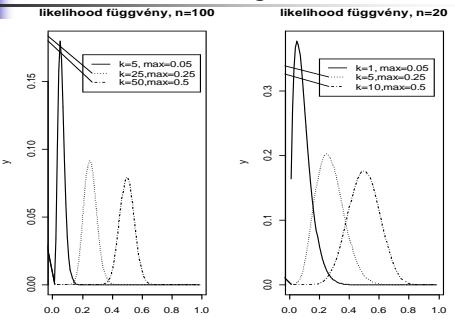
Becslési módszerek

- Eddig: „ad hoc” módszerek
- Általános eljárás kellene
 - Példa: valószínűség becslése, n kísérletből. Jelölje k a sikeresek számát (X_i , $i=1, \dots, n$ indikátor minta)

$$P\left(\sum_{i=1}^n X_i = k\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

Most p függvényében nézzük, k rögzített (elnevezés: likelihood függvény).

A likelihood függvény maximumhelye logikus választás a valószínűség becslésének



A módszer általánosan

$$L(\theta; \underline{x}) = f_{\theta}(\underline{x}) = \prod_{i=1}^n f_{\theta}(x_i)$$

(a likelihood függvény) maximumhelye lesz a θ paraméter maximum likelihood becslése.

Ha a függvény deriválható, a loglikelihood függvény

$$l(\theta; \underline{x}) = \ln f_{\theta}(\underline{x}) = \sum_{i=1}^n \ln f_{\theta}(x_i)$$

maximumhelye deriválással

$$\frac{\partial}{\partial \theta} l(\theta; \underline{x}) = \frac{\partial}{\partial \theta} \ln f_{\theta}(\underline{x}) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f_{\theta}(x_i) = 0$$

megoldásaként megtalálható

Példák

- valószínűségre: relatív gyakoriság
- Poisson eloszlás paraméterére: \bar{x}
- Exponenciális eloszlás paraméterére: $1/\bar{x}$

További példák ML becslésre

- Normális eloszlás várható értékére: \bar{x}
- A módszer többdimenziós paraméter becslésére is használható: $N(m, \sigma^2)$ esetén
$$(\bar{x}, \sum (x_i - \bar{x})^2 / n)$$
a maximum likelihood becslés.

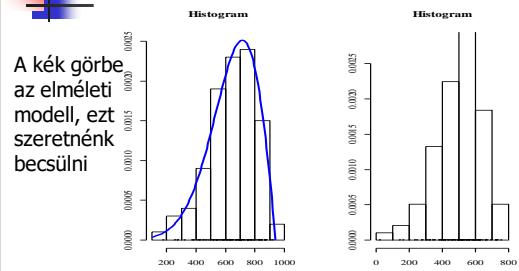
Tulajdonságok

- Nem minden torzítatlan
- Ha $T(x)$ a θ paraméter maximum likelihood becslése, akkor $\psi(T(x))$ a $\psi(\theta)$ paraméter maximum likelihood becslése.
- Nem minden lehet deriválással meghatározni: példa egyenletes eloszlás a $[0, \theta]$ intervallumon.

Aszimptotikus tulajdonságok

- Ha a likelihood függvény teljesít bizonyos regularitási feltételeket, akkor a maximum likelihood becslés
 - létezik
 - aszimptotikusan torzítatlan
 - aszimptotikusan hatásos
 - aszimptotikusan normális eloszlású.

Sűrűségfüggvény becslése hisztogrammal



Hátránya: az intervallumbeosztás szubjektív, nem a pontos értékek szerepelnek benne

Parzen-Rosenblatt módszer/1

- Tapasztalati eloszlásfüggvény nem deriválható, de ezen segíthetünk, ha az egyes megfigyeléseket nem pontszerűnek, hanem az adott érték körül kicsi szórású folytonos eloszlásúnak képzeljük (ez az eloszlás a magfüggvény).
- Ennek a folytonos keverékeloszlásnak a deriváltja jól közelíti a sűrűségfüggvényt.

Parzen-Rosenblatt módszer/2

- Tétel. Ha a mintánk egy $f(x)$ sűrűségfüggvényű eloszlásból származik, a $k(y)$ magfüggvény egyenletesen korlátos és $yk(y)$ határértéke a végtelenben 0, valamint h_n olyan számsorozat, melyre $\lim h_n = 0$ és $\lim nh_n = \infty$, akkor

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^n k\left(\frac{x - X_i}{h_n}\right)$$

aszimptotikusan torzítatlan, konzisztenzs becslés az $f(x)$ minden folytonossági pontjában.

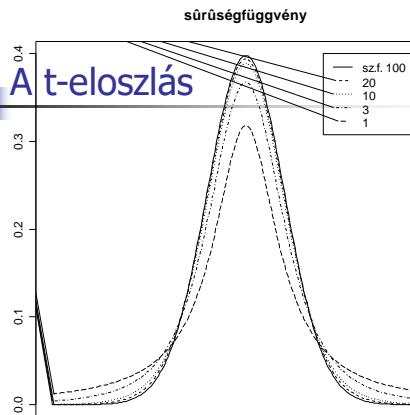
Konfidenciaintervallum

- Olyan intervallum, mely legalább $1-\alpha$ valószínűséggel tartalmazza a keresett paramétert:
$$P_\theta(T_1(X) < \theta < T_2(X)) \geq 1 - \alpha$$
- Példa: normális eloszlás várható értékére (m , ismert szórás esetén)
$$P\left(m \in \left(\bar{X} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

ahol $z_{1-\alpha/2}$ a standard normális eloszlás $1 - \alpha/2$ kvantilise

Konfidenciaintervallum a normális eloszlás várható értékére

- Ha a szórás nem ismert, becsüljük
 - Tulajdonság: normális eloszlású minta esetén a mintaátlag és a tapasztalati szórás független
 - $n-1$ szabadságfokú t (Student) eloszlás:
- $$\frac{\bar{X}_0}{\sqrt{(\bar{X}_1^2 + \dots + \bar{X}_{n-1}^2)/(n-1)}}$$
- eloszlása, ahol $\bar{X}_0, \bar{X}_1, \dots, \bar{X}_{n-1}$ független azonos, standard normális eloszlásúak



Konfidenciaintervallum a normális eloszlás várható értékére/2

- $$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{((\bar{X}_1 - \bar{X})^2 + \dots + (\bar{X}_n - \bar{X})^2)/(n-1)}}$$
- eloszlása $n-1$ szabadságfokú t-eloszlás
 - Ebből konfidencia intervallum m -re
- $$\left(\bar{X} - t_{n-1,\alpha/2} \frac{\sqrt{((\bar{X}_1 - \bar{X})^2 + \dots + (\bar{X}_n - \bar{X})^2)/(n-1)}}{\sqrt{n}}; \bar{X} + t_{n-1,\alpha/2} \frac{\sqrt{((\bar{X}_1 - \bar{X})^2 + \dots + (\bar{X}_n - \bar{X})^2)/(n-1)}}{\sqrt{n}} \right)$$
- Ha n nagy, az intervallum az ismert szórású esetnél látott hoz közelít.

Konfidencia intervallum a valószínűségre

- Ebben az esetben az egyes mintaelemek indikátorok: $\sigma^2 = p(1-p)$, tehát p -re a következő intervallumot kapjuk ($1-\alpha$ megbízhatóságú)
$$\left(\bar{X} - \frac{z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right)$$
- ahol $\bar{X} = \hat{p} = \frac{k}{n}$ (a relatív gyakoriság). Ez csak approximativ, n elég nagy kell, hogy legyen (pl. $n > 50$).

Mintaelemszám választás

- Ahhoz, hogy a várható értékre felírt $1-\alpha$ megbízhatóságú intervallum adott d számnál rövidebb legyen: $n \geq \frac{4(z_{1-\alpha/2})^2 \sigma^2}{d^2}$
- A valószínűség esetén: $n \geq \frac{4(z_{1-\alpha/2})^2 \hat{p}(1-\hat{p})}{d^2}$
- Mivel p és becslése is ismeretlen a kutatás tervezésekor, ezért a következő felső becslés használható $n \geq \frac{z_{1-\alpha/2}^2}{d^2}$

Valószínűségszámítás és statisztika előadás info. BSC/B-C szakosoknak

8. előadás
November 13.

Konfidenciaintervallum

- Olyan intervallum, mely legalább $1-\alpha$ valószínűsggyel tartalmazza a keresett paramétert:

$$P_\theta(T_1(X) < \theta < T_2(X)) \geq 1 - \alpha$$

minden θ paramétre

- Példa: normális eloszlás várható értékére (m , ismert szórás esetén)

$$P\left(m \in \left(\bar{X} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

ahol $z_{1-\alpha/2}$ a standard normális eloszlás $1 - \alpha/2$ kvantilise

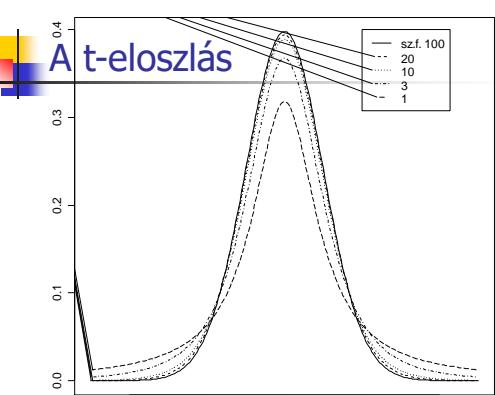
Konfidenciaintervallum a normális eloszlás várható értékére

- Ha a szórás nem ismert, becsüljük
- Tulajdonság: normális eloszlású minta esetén a mintaátlag és a tapasztalati szórás független
- $n-1$ szabadságfokú t (Student) eloszlás:

$$\frac{\bar{X}_0}{\sqrt{(X_1^2 + \dots + X_{n-1}^2)/(n-1)}}$$

eloszlása, ahol X_0, X_1, \dots, X_{n-1} független azonos, standard normális eloszlásúak

sűrűségfüggvény



Konfidenciaintervallum a normális eloszlás várható értékére/2

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)/(n-1)}}$$

- eloszlása $n-1$ szabadságfokú t-eloszlás
 - Ebből konfidencia intervallum m -re
- $$\left(\bar{X} - t_{n-1,\alpha/2} \frac{\sqrt{((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)/(n-1)}}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{\sqrt{((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)/(n-1)}}{\sqrt{n}} \right)$$
- Ha n nagy, az intervallum az ismert szórású esetnél látottthoz közelít.

Konfidencia intervallum a valószínűségre

- Ebben az esetben az egyes mintaelemek indikátorok: $\sigma^2 = p(1-p)$, tehát p -re a következő intervallumot kapjuk ($1-\alpha$ megbízhatóságú)

$$\left(\bar{X} - \frac{z_{1-\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2} \sqrt{p(1-p)}}{\sqrt{n}} \right)$$

- ahol $\bar{X} = \hat{p} = \frac{k}{n}$ (a relatív gyakoriság). Ez csak approximativ, n elég nagy kell, hogy legyen (pl. $n>50$).

Mintaelemszám választás

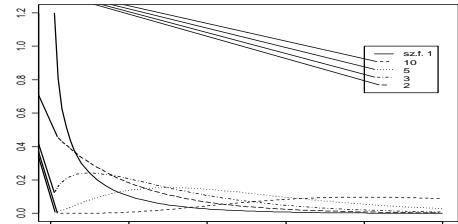
- Ahhoz, hogy a várható értékre felírt $1 - \alpha$ megbízhatóságú intervallum adott d számnál rövidebb legyen: $n \geq \frac{4(z_{1-\alpha/2})^2 \sigma^2}{d^2}$
- A valószínűség esetén: $n \geq \frac{4(z_{1-\alpha/2})^2 \hat{p}(1-\hat{p})}{d^2}$

Mivel p és becslése is ismeretlen a kutatás tervezéskor, ezért a következő felső becslés használható

$$n \geq \frac{z_{1-\alpha/2}^2}{d^2}$$

χ^2 -négyzet eloszlás

- Legyen X_1, \dots, X_n független azonos, standard normális eloszlás
- $X_1^2 + \dots + X_n^2$ eloszlása n szabadságfokú χ^2 eloszlás suruségtüggvény



Konfidencia intervallum a normális elo. szórásnégyzetére

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

eloszlása $n-1$ szabadságfokú χ^2 eloszlás. Ebből adódik konfidenciaintervallum σ^2 -re:

$$P\left(\sigma^2 \in \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{h_{1-\alpha/2,n-1}}, \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{h_{\alpha/2,n-1}} \right)\right) = 1 - \alpha$$

ahol $h_{\alpha/2,n-1}$ és $h_{1-\alpha/2,n-1}$ az $\alpha/2$, illetve $1 - \alpha/2$ kvantilise az $n-1$ szabadságfokú χ^2 -négyzet eloszlásnak.

Hipotézisvizsgálat

- H_0 nullhipotézis (jelezni akarjuk, ha nem igaz) $\theta \in \Theta_0$.
- H_1 ellenhipotézis $\theta \in \Theta_1$.
- Elsőfajú hiba: H_0 igaz, de elutasítjuk
- Másodfajú hiba: H_0 hamis, de elfogadjuk
- Példák:
 - 2 kocka közül melyikkel dobunk? (H_0 : az elsővel)
 - mekkora a fejdobás valószínűsége? (H_0 : $p=0,5$)

Alapfogalmak

- Kritikus tartomány: \mathfrak{N}_k azon X megfigyelés-vektorok halmaza, amelyre elutasítjuk a nullhipotézist.
- Általában egy $T(X)$ statisztika segítségével definiáljuk.
- Elsőfajú hiba valószínűsége: $\alpha = P_\theta(\mathfrak{N}_k)$ ahol $\theta \in \Theta_0$. Alapértelmezés: $\alpha=0,05$
- Másodfajú hiba valószínűsége: $\beta(\theta) = 1 - P_\theta(\mathfrak{N}_k) = P_\theta(\mathfrak{N}_e)$ ahol $\theta \in \Theta_1$ és X_e az elfogadási tartomány.
- Erőfüggvény: $1 - \beta(\theta) = P_\theta(\mathfrak{N}_k)$

Próbák tulajdonságai

- Terjedelem: elsőfajú hiba valószínűségek felső határa.
- Konzisztenca: az erőfv. 1-hez tart (minden $\theta \in \Theta_1$ -re)
- Torzítatlanság: $P_{\theta_0}(\mathfrak{N}_k) \leq P_{\theta_1}(\mathfrak{N}_k)$ (minden $\theta_0 \in \Theta_0$ -ra és $\theta_1 \in \Theta_1$ -re)
- Azonos terjedelmű próbák közül az az erősebb, amelynek az erőfüggvénye egyenletesen (minden $\theta \in \Theta_1$ -re) nagyobb (vagy egyenlő).
- Az a próba a legerősebb, ami minden más azonos terjedelmű próbánál erősebb

Paraméteres próbák

- Lényeg: valamilyen, véges sok valós paraméterrel leírható modellt tételezünk fel a mintáról.
- Példa:
 - Normális
 - Poisson stb.eloszlású minta.
- A feladat: a paraméter(ek)re vonatkozó hipotézis vizsgálata.
- A továbbiakban normális eloszlású mintákkal foglalkozunk

Próbák a normális eloszlás várható értékére: u-próba.

- $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$. Ha ismert a szórás (u-próba):
$$u = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma}$$
- Kritikus tartomány: $|u| > z_{1-\alpha/2}$. ($z_{1-\alpha/2}$ a standard normális eloszlás $1-\alpha/2$ kvantilise)
- Tulajdonságok:
 - torzítatlan
 - konzisztens
- Ha egyoldali az ellenhipotézis, akkor a kritikus tartomány $u > z_{1-\alpha}$ ($\mu > \mu_0$), illetve $u < -z_{1-\alpha}$ alakú ($\mu < \mu_0$). Ezek legerősebb próbák!

Valószínűségszámítás és statisztika előadás info. BSC/B-C szakosoknak

9. előadás
November 20.

Próbák a normális eloszlás várható értékére: t próba.

- $H_0: m=m_0, H_1: m \neq m_0$. Ha nem ismert a szórás (t-próba):
$$t = \sqrt{n} \frac{\bar{X} - m_0}{\hat{\sigma}}$$
ahol
$$\hat{\sigma} = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)}$$
- Kritikus tartomány: $|t| > t_{1-\alpha/2, n-1}$. (H_0 esetén a próbástatisztika $n-1$ szabadságfokú, t-eloszlású.)
- Ha egyoldali az ellenhipotézis, akkor a kritikus tartomány $t > t_{1-\alpha, n-1}$ ($m > m_0$), illetve $t < -t_{1-\alpha, n-1}$ alakú ($m < m_0$). Ezek is legerősebb próbák!

Megjegyzések

- A kétoldali esetre kapott próba nem a legerősebb (ilyenkor nincs is ilyen).
- Ha a minta elemszáma nagy, a t-próba helyett az u-próba is használható (ekkor még a normális eloszlásúságra sincs szükség a centrális határeloszlás tétele miatt).
- A gyakorlatban a számítógépes programok az úgynevezett p-értéket adják meg, ami a legkisebb α elsőfajú hibavalószínűség, amire még elutasítjuk a nullhipotézist.

Kétoldali próbák és konfidencia intervallumok

- A normális eloszlásnál a várható értékre vonatkozó α terjedelmű próbánál láttuk, hogy a $H_0: m=m_0$ hipotézist a $H_1: m \neq m_0$ hipotézissel szemben pontosan akkor fogadjuk el, ha m_0 benne van az $1 - \alpha$ megbízhatóságú konfidencia intervallumban.
- Egyoldali esetre is átvihető (egyoldali konfidencia intervallumokkal)

Kétmintás eset: párosított megfigyelések

- Példa: Van-e különbség Budapest és Cegléd napi átlaghőmérséklete között? Formálisan: $H_0: m_1 = m_2$
- Ha ugyanazon napokról van megfigyelésünk mindenki helyen: nem függetlenek a minták.
- Ekkor a párok tagjai közötti különbséget vizsgálva, az előző egymintás esetre vezethető vissza a feladat.
- $H_0^*: m=0, H_1^*: m \neq 0$ az új hipotézisek.

Kétmintás eset: független minták

Első minta: n elemű, σ_1 szórású, második: m elemű, σ_2 szórású. Ha ismert σ : kétmintás u-próba

$$u = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}}$$

Kritikus tartomány: mint az egymintás esetben
Ha ismeretlenek, de azonosak a szórások:

$$t_{n+m-2} = \sqrt{\frac{nm(n+m-2)}{n+m}} \frac{\bar{X} - \bar{Y}}{\sqrt{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}}$$

A szórás vizsgálata kétmintás esetben: F-próba

- $H_0: \sigma_1 = \sigma_2$
- Két független, n_1 illetve n_2 elemű normális eloszlású minta alapján a próbatestítésztika:
- $F: \max\left(\frac{s_1^2}{s_2^2}, \frac{s_2^2}{s_1^2}\right)$
- F: a korrigált tapasztalati szórásnégyzetek hányszáma
- Kritikus érték: az (n_1-1, n_2-1) szabadságfokú F eloszlás $1-\alpha/2$ kvantilise (n_1 számlálóbeli, n_2 pedig a nevezőbeli minta elemszáma).

Kétmintás t-próba ismét

- Alkalmazható, ha az F-próba elfogadja a szórások azonosságát.

- Ha nem, akkor Welch-próba:

$$t' = \frac{\bar{X} - \bar{Y}}{\sqrt{s_1^2/n + s_2^2/m}}$$

- H_0 esetén közelítőleg t eloszlású v szabadságfokkal, ahol

$$v = \frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{\left(s_1^2/n\right)^2}{n-1} + \frac{\left(s_2^2/m\right)^2}{m-1}}$$

χ^2 -négyzet próba illeszkedésvizsgálatra

- H_0 hipotézis: az A_1, A_2, \dots, A_r teljes eseményrendszerre teljesül $P(A_1)=p_1, P(A_2)=p_2, \dots, P(A_r)=p_r$
- A tesztstatisztika: $\sum_{i=1}^r \frac{(v_i - np_i)^2}{np_i}$ ami aszimptotikusan $r-1$ szabadságfokú χ^2 -négyzet eloszlású, ha igaz a nullhipotézis (v_i az A_i gyakorisága)
- Aszimptotikusan ($n \rightarrow \infty$) a statisztika eloszlása $r-1$ szabadságfokú χ^2 -négyzet eloszláshoz közelít.
- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az $r-1$ szabadságfokú χ^2 -négyzet eloszlás $1-\alpha$ kvantilise, elutasítjuk a nullhipotézist.

χ^2 -négyzet próba függetlenségvizsgálatra

- H_0 hipotézis: az A_1, A_2, \dots, A_r és B_1, B_2, \dots, B_s teljes eseményrendszerre teljesül a függetlenség. $\sum_{i,j} \frac{(v_{ij} - np_i q_j)^2}{np_i q_j}$ ahol v_{ij} az $A_i \cap B_j$ gyakorisága, $p_i = P(A_i), q_j = P(B_j)$.
- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az $(r-1)(s-1)$ szabadságfokú χ^2 -négyzet eloszlás $1-\alpha$ kvantilise, elutasítjuk a nullhipotézist.

Becsüléses esetek

- Illeszkedésvizsgálatnál: ha az illesztendő eloszlást nem ismerjük – csak a családját – becsülik a paramétereit. Ekkor a próbatestítésztika szabadságfoka annyival csökken, ahány paramétert becsültünk.
- Függetlenségvizsgálatnál: itt általában nem ismerjük a teljes eseményrendszer tagjainak valószínűségét, így $r-1+s-1$ valószínűséget kell becsülnünk. A szabadságfok ekkor tehát $r-1+s-2=(r-1)(s-1)$.

Az illeszkedésvizsgálat alkalmazása folytonos eloszlásokra

- A teljes eseményrendszer a számegyenes felosztása révén jön létre.
- Ügyeljünk arra, hogy minden intervallum közel azonos valószínűségű legyen.
- Ha paraméterbecslés szükséges, ML módszer alkalmazható.

Valószínűségszámítás és statisztika előadás info. BSC/B-C szakosoknak

10. előadás
November 27.

Az illeszkedésvizsgálat alkalmazása folytonos eloszlásokra

- A teljes eseményrendszer a számegyenes felosztása révén jön létre.
- Ügyeljünk arra, hogy minden intervallum közel azonos valószínűségű legyen.
- Ha paraméterbecslés szükséges, ML módszer alkalmazható.

Homogenitásvizsgálat khi-négyzet próbával

- Homogenitásvizsgálat: H_0 : a két minta eloszlása azonos
- A próbastatisztika:
$$nm \sum_{i=1}^r \frac{\left(\frac{v_i}{n} - \frac{\mu_i}{m} \right)^2}{\frac{v_i}{n} + \mu_i}$$

(Az v_i gyakorisága v_i az első és μ_i a második mintánál)
- A statisztika aszimptotikusan $r-1$ szabadságfokú χ^2 -négyzet eloszlású, ha igaz a nullhipotézis
- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az $r-1$ szabadságfokú χ^2 -négyzet eloszlás $1-\alpha$ kvantilise, elutasítjuk a nullhipotézist.

További nemparaméteres próbák

- Illeszkedésvizsgálat: Adott eloszlású-e a minta? (Például paraméteres próbához kellhet.)
 - Egymintás Kolmogorov-Szmirnov próba: a tapasztalati és az elméleti eloszlásfüggvény eltérésének maximumán alapul.
 - Kétmintás eset (homogenitásvizsgálat):
$$D_{m,n} = \max_x |F_n(x) - G_m(x)|$$
 - Ugyanerre az eltérésre más próbák is épülnek (Anderson-Darling, Cramér-von Mises), melyek az eltérés (esetleg súlyozott) integrálját használják.

További nemparaméteres tesztek $H_0: P(X>Y)=1/2$ tesztelésére

- Párosított esetre: előjelpróba (a különbségek előjelén alapul). Ez H_0 esetén Binom($n; 1/2$) eloszlású.
- Wilcoxon (Mann-Whitney) próba (rangstatisztika): független mintákra
- Azt számoljuk össze, hogy hány olyan pár van, ahol $X_i > Y_j$. A kapott statisztika aszimptotikusan normális eloszlású, nem érzékeny a kiugró értékekre.

Kovariancia, korreláció becslés

- A kovariancia becslése: tapasztalati kovariancia
$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$
- A korreláció becslése a minta alapján: tapasztalati korrelációs együttható
$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Y közelítése X függvényével

- Gyakori eset, hogy nem ismerjük a számunkra érdekes mennyiséget (Y) pontos értékét (pl. holnapi részvény-árfolyam, vízállás, időjárás). Van viszont információnk hozzá kapcsolódó mennyiségről (X , mai értékek).
- Feladat: olyan f_0 megtalálása, amelyre $f_0(X)$ a lehető legjobb közelítése Y -nak.
- Matematikailag: f_0 a megoldása a $\min_f E(Y - f(X))^2$ szélsőérték-problémának (legkisebb négyzetes becslés).
- Ha az együttes eloszlás ismert (nem teljesen reális, de a megfigyelések alapján közelíthető), akkor megoldható a feladat.

A várható érték optimumtulajdonsága

Állítás. A

$$\min_a E(Y - a)^2$$

feladat megoldása $a = E(Y)$.

Bizonyítás. $E(Y-a)^2 = E(Y^2) - 2aE(Y) + a^2$

a szerint deriválva adódik, hogy valóban $E(Y)$ a minimumhely.

A minimum értéke $D^2(Y)$.

Ugyanígy: X tetszőleges értéke esetén $E(Y|X=x)$ adja a minimumot.

A feltételes várható érték közelítése Nadarajah módszerével

$$\hat{r}_n(x) = \frac{\sum_{i=1}^n Y_i k\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n k\left(\frac{x - X_i}{h_n}\right)}$$

A sűrűségfüggvényre vonatkozó regularitási feltételek esetén ez konziszens becslése az $E(Y|X)$ regressziójának (k a magfüggvény, h_n az ablakszélesség).

Analóg módszer: k legközelebbi szomszéd (itt fix, k számú értékből számolunk, a magfüggvény konstans)

Optimum a lineáris függvények körében

$$\min_{a,b} E[Y - (aX + b)]^2$$

- Egyszerűbben megoldható
- Nem kell az együttes eloszlás
- A megoldás deriválással jön ki:

$$a = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E^2(X)}$$

$$b = E(Y) - \frac{E(XY) - E(X)E(Y)}{E(X^2) - E^2(X)} E(X)$$

A regressziós egyenes tulajdonságai

- Ez a legkisebb négyzetes eltérést adó a lineáris függvények között (a fenti megoldás valóban minimum)
- Elnevezés: regressziós egyenes
- Átmegy az $(E(X), E(Y))$ ponton
- Példa: Kockával dobunk, majd ha k az eredmény, az $1, \dots, k$ cédrulák közül húzunk egyet. Nem tudjuk a húzás eredményét, csak a kockadobásét. Hogyan tippeljünk a húzott száma (a legkisebb négyzetes eltérést adó becslést keressük)? $E(h|K=k) = (k+1)/2$ az univerzálisan legjobb közelítés, tehát a legjobb lineáris közelítés is.

Lineáris modell a gyakorlatban

- (X_i, Y_i) független, azonos eloszlású minta. $i=1, \dots, n$
- $Y_i = aX_i + b + \varepsilon_i$ (X_i a magyarázó változó értéke, ε_i független, azonos eloszlású hiba. $E(\varepsilon_i) = 0$, általában feltessük, hogy normális eloszlású. a, b a becsülendő együtthatók) $\sum (Y_i - aX_i - b)^2 \rightarrow \min$

$$\text{Megoldás: } \hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \hat{b} = \bar{y} - \hat{a}\bar{x}$$

A lineáris regresszió becsléseinek tulajdonságai

Torzítatlanok, $D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, D(\hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$

Az x^* pontban előrejelzett érték $\hat{a}x^* + \hat{b}$

ennek a szórása

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

Az x^* pontbeli megfigyelés

$$\text{szórása } \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

A szórásbecslésnél σ helyett

annak becsült értékét használjuk: $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2}{n-2}$

Valószínűségszámítás és statisztika előadás info. BSC/B-C szakosoknak

11. előadás
December 4.

A lineáris regresszió becsléseinek tulajdonságai

$$\text{Torzítatlanok, } D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, D(\hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Az x^* pontban előrejelzett érték $\hat{a}x^* + \hat{b}$

$$\text{ennek a szórása} \quad \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\text{Az } x^* \text{ pontbeli megfigyelés szórása} \quad \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\text{A szórásbecslésnél } \sigma \text{ helyett annak becsült értékét használjuk: } \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2}{n-2}$$

Hipotézisvizsgálat/1

- $H_0: a=0$ tesztelése t-próbával:

$$t_{n-2} = \frac{\hat{a} \sqrt{(n-2) \sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2}}$$

- Konfidencia intervallum a-ra:

$$\hat{a} \pm t_{n-2, 1-\alpha/2} \hat{D}(a)$$

Hipotézisvizsgálat/2

- $H_0: b=0$ tesztelése t-próbával:

$$t_{n-2} = \frac{\hat{b} \sqrt{n(n-2) \sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^n (y_i - \hat{a}x_i - b)^2} \sqrt{\sum_{i=1}^n x_i^2}}$$

- Konfidencia intervallum b-re:

$$\hat{b} \pm t_{n-2, 1-\alpha/2} \hat{D}(b)$$

Szóródások

- Teljes ingadozás: $\sum_{i=1}^n (y_i - \bar{y})^2$
- Reziduális négyzetösszeg: $\sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- A megmagyarázott variabilitás részaránya:

$$R^2 = \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{éppen a tapasztalati korrelációs együttható négyzete}$$

Többváltozós lineáris modell

- Több magyarázó változót is bevonhatunk a modellbe: $Y = X\beta + \varepsilon$ ahol Y, ε n hosszú vektorok, X nxk-as mátrix (ismert értékekből), β pedig k hosszú (ismeretlen) paramétervektor. $E(Y) = X\beta$.

A legkisebb négyzetek módszere

$$\sum_{i=1}^n \varepsilon_i^2 = (Y - X\beta)'(Y - X\beta) \rightarrow \min$$

- A megoldás:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

A becslés tulajdonságai

- Torzítatlan
- Kovarianciamátrix:
$$E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \sigma^2(X'X)^{-1}$$
- Ha ε normális eloszlású, akkor a legkisebb négyzetes becslés egyúttal ML becslés is.
- Példák: lineáris regresszió, szórásanalízis.

Hipotézisvizsgálat a lineáris modellben

- A vizsgált hipotézis: $H_0: \beta' H' = 0$
ahol H rxk-as mátrix ($r < k$), $\text{rang}(H) = r$.
- A valószínűsséghányados próba statisztika:
$$F = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta}) - (Y - X\hat{\beta})'(Y - X\hat{\beta})}{(Y - X\hat{\beta})'(Y - X\hat{\beta})}$$
- $(n-k)/r F$ a H_0 esetén F eloszlású ($r, n-k$) szabadsági fokkal. (Akkor utasítjuk el H_0 -t, ha F nagy.)

Szóráselemzés

- Egy vagy több faktor különböző „szintjein” mérünk eredményeket (pl. termésátlagokat)
- A kérdés: mely faktor(ok) hatnak
- Ez is lineáris modell: $Y = X\beta + \varepsilon$, ahol X 0-1 mátrix.

Egyszeres osztályozás

- Egy faktor különböző szintjein mérünk. $Y_{ij} = a_j + \varepsilon_{ij}$ (a_j a faktor j-edik szintjének a hatása, $j=1, \dots, k$, $i=1, \dots, n_j$). ε_{ij} független, azonos, $N(0, \sigma)$ eloszlású, $N = n_1 + \dots + n_k$.
- Ez is $Y = X\beta + \varepsilon$ alakú.
- A nullhipotézis: a szintek nincsenek hatással, azaz $a_1 = a_2 = \dots = a_k$
- Ez is $H_0: \beta' H' = 0$ alakú, tehát F-próba végezhető.

A négyzetösszegek felosztása

Az átlagok felbontása:

$$Y_{ij} - \bar{Y} = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \bar{Y})$$

A négyzetösszegek felbontása:

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^k n_j (\bar{Y}_j - \bar{Y})^2$$

A négyzetösszegek felosztása

Másképpen:

$$SS_{\text{össz}} = SS_{\text{csb}} + SS_{\text{csk}}$$

A szabadsági fokok:

$$df_{\text{össz}} = df_{\text{csb}} + df_{\text{csk}}$$

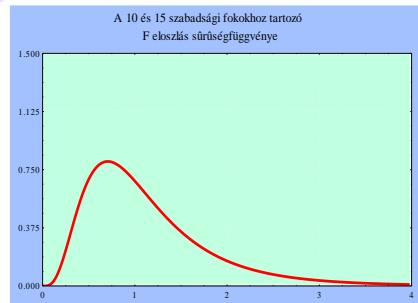
$$N - 1 = (N - k) + (k - 1)$$

Az F-próba

A H_0 (nincs hatás) hipotézis fennállása esetén a "csk" csoportok közötti és "csb" csoporton belüli szórásnégyzetek hányadosa nem túl nagy és az eloszlása ismert (megfelel az általános lineáris modellnél látott próbnak, a kritikus tartomány is ugyanúgy az adott szabadságfokú F eloszlás $1-\alpha$ kvantilisénél nagyobb értéket tartalmazza):

$$\frac{MS_{csk}}{MS_{csb}} = \frac{(n-k)SS_{csk}}{(k-1)SS_{csb}} \text{ eloszlása } F_{df_{csk}, df_{csb}}$$

Az F-eloszlás



Példa

	1. csoport	2. csoport
1. megfigyelés	2	6
2. megfigyelés	3	7
3. megfigyelés	1	5

Átlag	2	6
Négyzetösszeg	2	2

Teljes átlag	4	
Teljes négyzetösszeg	28	

A szóráselemzés táblázata

Forrás	Szórások elemzése				
	SS	df	MS	F	p
Hatás(csk)	24.0	1	24.0	24.0	.008
Hiba (csb)	4.0	4	1.0		

Monte Carlo módszerek

- Véletlen szám generátorok
- Szimulációk
 - Integrálok közelítése
 - Bootstrap
- Markov láncokon alapuló algoritmusok

Véletlen szám generálás

LCG: $X_{n+1} = (aX_n + c) \bmod m$

$$(0 < m, 0 < a < m, 0 \leq c < m, 0 \leq X_0 < m)$$

Jól bevált paraméterválasztások:

1. Borland C/C++ $m=2^{32}$, $a=1664525$, $c=1013904223$
2. Delphi, Pascal $m=2^{32}$, $a=134775813$, $c=1$

Valószínűségszámítás és statisztika előadás info. BSC/B-C szakosoknak

12. előadás
December 11.

Véletlen szám generálás inverz módszerrel

Tétel: Legyen X val. vált., F eloszlásfüggvény, amely szigorúan monoton növekedő és folytonos. Ekkor

- $F(X)$ egyenletes eloszlású $[0,1]$ -en
- Ha $U \sim U(0,1)$ akkor $F^{-1}(U)$ eloszlásfüggvénye F .

Pl.: Ha $X \sim \exp(\lambda)$ $F(x) = 1 - \exp(-\lambda x)$
 $F^{-1}(x) = -\ln(1-x)/\lambda$
 $-\ln(1-U)/\lambda \sim \exp(\lambda)$

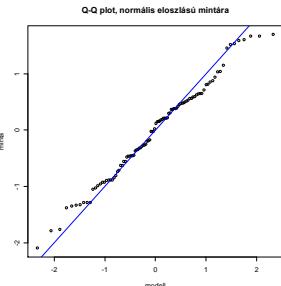
Kiterjesztése: általánosított inverz:
 $F^{-1}(x) = \inf\{x \mid F(x) = y\}$

Eloszlás illeszkedésének vizsgálata: Q-Q plot

A megfigyelt és az illesztett eloszlás kétdimenziós ábrázolása.

Eloszlásfüggvény q-kvantilise: az az érték, amelynél q valószínűséggel kapunk kisebbet: $G^{-1}(q)$
Spec.: $q=1/2$: median

$$\left\{ \left(G^{-1}\left(\frac{k}{n+1} \right), X_k^{(n)} \right) : k = 1, 2, \dots, n \right\}$$



Neumann módszer

Legyen $f(x)$ tetszőleges sűrűségfüggvény, $g(x)$ pedig olyan sűrűségfüggvény, amelyre $f(x) < Mg(x)$, valamely $M > 1$ esetén és $g(x)$ -ból könnyen tudunk mintát venni (tipikus példa az egyenletes eloszlás).

Algoritmus:

- Vegyük mintát: $u \sim U(0,1)$ -ből, $x \sim g(x)$ -ből
- Ha $u < f(x)/Mg(x)$, akkor x -et elfogadjuk
- Különben elutasítjuk, és 1-be lépünk.

Normális eloszlású véletlen szám

Box-Müller módszer

- Legyen U, V független, $E[0;1]$ eloszlású. Ekkor

$$\sqrt{-2 \ln U} \sin(2\pi V), \sqrt{-2 \ln U} \cos(2\pi V)$$

két független standard normális eloszlású változó lesz.

Szimulációk

- Modellek tesztelésére
- Integrálok kiszámítására: a $[0,1]$ intervallumon egyenletes eloszlású véletlen számokból:
$$\frac{g(X_1) + g(X_2) + \dots + g(X_n)}{n} \rightarrow E(g(X)) = \int_0^1 g(x) dx$$
a nagy számok törvénye miatt

Alternatív algoritmus

- $0 < g < 1$, $0 < x < 1$ feltehető (lin. transzformációval ide vihető)
- (U_i, V_i) egyenletes eloszlású az egységnégyzetben ($i=1,\dots,N$)
- I becslése $\#\{i: V_i < g(U_i)\}/N$
- A becslés torzítatlan, szórása $(I(1-I)/4N)^{1/2}$.

A módszerek összehasonlítása

- Első eljárás is torzítatlan és szórásnégyzete $(E(g^2(X))-I^2)/N$, amiből kisebb szórás adódik.
- Tovább is javítható, ha X az $[a,b]$ -re koncentrálódó sűrűségfüggvényű eloszlásból származik:
$$\int_a^b g(x)dx = \int_a^b \frac{g(x)}{f(x)} f(x)dx = E\left(\frac{g(X)}{f(X)}\right)$$
 és ennek még kisebb a szórása, ha $f(x) \approx g(x)$
- Tehát $g(X)/f(X)$ szimulált értékeinek átlaga jó közelítés
- Az is előnye, hogy improprius integrálokra is alkalmazható (ha pl. f a normális eloszlás

Markov láncon alapuló módszerek

- A független esetnél sokszor hatékonyabb, ha iteratívan generáljuk a mintákat.
- Egyszerű struktúra: Markov lánc. A lehetséges értékek (állapotok) halmaza tipikusan véges
- A lényeges feltétel, hogy a következő lépésekben az eloszlás csak az utolsó lépésekben elérő állapottól függ

Sztochasztikus folyamatok

- Valószínűségi változó-sorozatok, a tagok nem függetlenek!
- Markov láncok (X_n a rendszer állapota az n-dik lépésben)
- Poisson folyamat (X_t a $[0,t]$ intervallumon bekövetkező események száma: Poisson eloszlású λt paraméterrel)
- Sorbanállási rendszerek (X_t a rendszerben az igények száma a t időpontban)
- Idősorok statisztikai elemzése (X_t a t időpontban mért érték)

Idősor-elemzés

- Adatok: X_t a t időpontban megfigyelt érték; ezek tipikusan nem függetlenek egymástól
- Stacionaritást feltételezzük
 - Erős stacionaritás: az együttes eloszlások nem függnek az időtől
 - Gyenge stacionaritás: a kovariancia-struktúra állandó
- Ha nem stacionárius: pl. szezonálitás, trend figyelhető meg, akkor azokat előzetesen, regressziós módszerekkel eltávolítjuk.

A stacionárius adatsor elemzése

Idősor (x_1, x_2, \dots, x_n)

1. rendű autokorrelációs együttható :

Korreláció az alábbi $n-1$ pár között $(x_2, x_1), (x_3, x_2), \dots, (x_n, x_{n-1})$

$$r_1 = \frac{\sum_{t=2}^n (x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$$

2. rendű autokorrelációs együttható :

Korreláció az alábbi $n-2$ pár között: $(x_3, x_1), (x_4, x_2), \dots, (x_n, x_{n-2})$

$$r_2 = \frac{\sum_{t=3}^n (x_t - \bar{x})(x_{t-2} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$$

k. rendű autokorrelációs együttható

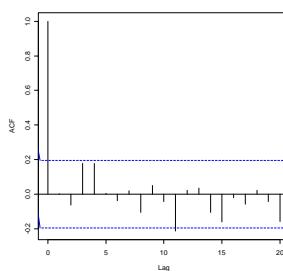
Általában: k. rendű autokorrelációs együttható :
Korreláció az alábbi n-k pár között (x_{k+1}, x_1), (x_{k+2}, x_2), ..., (x_n, x_{n-k})

$$r_k = \frac{\sum_{t=k+1}^n (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2}$$

Grafikus megjelenítés: korrelogramm

X tengely: rend, y tengely: autokorrelációk

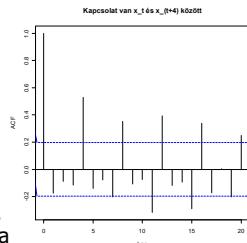
Független, azonos eloszlású változók



a megfigyelések száma
 $n=100$, konfidencia sáv
az $r=0$ teszteléséhez
normális határeloszlásból:
[- $1.96\sqrt{1/n}$, $1.96\sqrt{1/n}$]

Továbblépés

- Ha nem fogadható el a reziduálisok korrelálatlansága:
 - Lehetnek fel nem tárt periódusok
 - De más kapcsolat is fennmaradhat az egymáshoz közel megfigyelések között (pl. időjárási adatok, eltérés a sokévi átlagtól): lineáris modellekkel közelíthetők



Általános, lineáris modell

Autoregressziós folyamatok (AR)

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + a_3 X_{t-3} + \dots + a_p X_{t-p} + \varepsilon_t$$

■ Stacionárius, ha az $1 - (a_1 s + a_2 s^2 + \dots + a_p s^p) = 0$
egyenlet gyökei az egységkörön kívül helyezkednek el

Mozgóátlag folyamatok (MA)

$$X_t = b_0 \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + b_3 \varepsilon_{t-3} + \dots + b_q \varepsilon_{t-q}$$

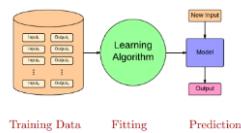
■ Mindig stacionárius

Kombináció: ARMA folyamatok

Modern fogalmak

- Gépi tanulás (az adatokból tanulnak az algoritmusok)
- Statisztikai tanulás (emellett még a bizonytalanságot is próbálja mérni)
- Adattudomány (mindenből egy kicsi...)
- Big data: nemcsak a nagysága miatt más, hanem a struktúrája miatt is:
 - Sok különböző adatstruktúra
 - Erősen hiányos adatbázisok
- A cél az összefüggések feltárása

Felügyelt tanulás



- Hagyományos statisztika: hosszú előkészítés, adattisztítás
- Modern megközelítés: nagy adat sok tulajdonsággal – gépi tanulás segít ezek közül a fontosakat kiválasztani
- Lényeg: az új adatokon működjön a módszer!

Mi is az a „big data”?

- Nincs matematikai definíció
- Nagy (terabyte-petabyte), változó, komplex (pl. szöveg-képek-video)
- A cél az összefüggések kinyerése
- Lehetséges módszerek:
 - Részek elemzése, összesítése
 - Véletlen részminták vizsgálata
 - Mesterséges intelligencia

Példák ipari alkalmazásokra

- Bonyolult rendszerek monitorozása, a potenciális hibaforrások előzetes beazonosítása (pl. olajfúrtornyoknál)
- App a kagylótenyésztő farmereknek a tápanyag- és egyéb faktorok optimalizálására (ha már ez megvan, hasonlók sok helyen elközelhetőek)
- Szállítócégek költségszámítása – jó előre, így könnyű optimalizálni: mit mikor, mivel érdemes szállítani
- Ipar 4.0: szimulációk döntő fontosságúak