Programtervező informatikus BSc, B szakirány
Valószínűségszámítás és statisztika gyakorlat

1. (1-2 hét) Valószínűségek kiszámítása; feltételes valószínűség és Bayes-tétel

Elmélet

Definíció (Ismétlés nélküli permutáció)**.** n (különböző) elem egy sorrendje.

$$n!.$$

Definíció (Ismétléses permutáció)**.** n (nem feltétlen különböző) elem egy sorrendje, ahol az egyforma elemeket nem különböztetjük meg (tfh. az n elem közül $k_1, \ldots, k_r$ darab megegyezik).

$$\frac{n!}{k_1! \cdots k_r!} = \binom{n}{k_1, \ \ldots \ , k_r}.$$

Definíció (Ismétlés nélküli kombináció)**.** n (különböző) elem közül k számú $(k \leq n)$ elem egyszerre történő kiválasztása (sorrend nem számít, nincs visszatevés).

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}.$$

Definíció (Ismétléses kombináció)**.** n (különböző) elem visszatevéses eljárással kiválasztott valamely k számú $(k \leq n)$ elem kiválasztása (sorrend nem számít).

$$\binom{n+k-1}{k}.$$

Definíció (Ismétlés nélküli variáció)**.** n (különböző) elem közül kiválasztott valamely k számú $(k \leq n)$ elem egy sorrendje (nincs visszatevés).

$$\frac{n!}{(n-k)!}.$$

Definíció (Ismétléses variáció)**.** n (különböző) elem közül visszatevéses eljárással kiválasztott valamely k számú $(k \leq n)$ elem egy sorrendje.

$$n^k.$$

A valószínűség a matematikai fogalma annak, hogy mekkora esély van valamire, például egy ászt kihúzni egy kártyapakliból, vagy egy piros golyót kihúzni egy színes golyókkal teli zsákból. A klasszikus valószínűség azokat az eseteket nézi, amikor minden egyes lehetséges kimenetelhez ugyanakkora esély tartozik. Például, ha egy szabályos dobókockával dobunk, akkor ugyanakkora az esély arra, hogy 1, 2, 3, 4, 5, vagy 6-ost dobjunk. Ekkor egy tetszőleges esemény valószínűsége megadható a kedvező kimenetelek és az összes lehetséges kimenetelek számának hányadosával:

Klasszikus valószínűség: Az A esemény valószínűsége megadható úgy, mint $P(A) = \dfrac{\text{kedvező kimenetelek száma}}{\text{összes kimenetel száma}}$.

Természetesen a későbbiekben ennél bonyolultabb esetekkel is fogunk találkozni, de az alapfogalmak megértéséhez ez a véges sok lehetőséget tartalmazó egyszerű modell is elegendő.

Definícó (Feltételes valószínűség).

Ha B bekövetkezett, mi a valószínűsége, hogy A bekövetkezik? $\quad P(A|B) = \dfrac{P(A \cap B)}{P(B)}, \quad$ ha $P(B) \neq 0$

Definícó (Teljes eseményrendszer).

$B_1, B_2, ...$ események teljes eseményrendszert alkotnak, ha **1)** $B_i \cap B_j = \emptyset \quad \forall i \neq j-$re **2)** $\bigcup\limits_{i=1}^{\infty} B_i = \Omega$

Teljes valószínűség tétele:
Legyen $B_1, B_2, ...$ teljes eseményrendszer, A tetszőleges esemény, $P(B_j) > 0$ minden j-re. Ekkor

$$P(A) = \sum_{j=1}^{\infty} P(A|B_j)P(B_j).$$

Bayes-tétel:
Legyen $B_1, ..., B_n, ...$ teljes eseményrendszer, A tetszőleges esemény, $P(B_j) > 0$ minden j-re. Ekkor

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{\infty} P(A|B_j)P(B_j)}.$$

Definíció (Események függetlensége).
A és B események függetlenek, ha
$P(A \cap B) = P(A) \cdot P(B)$ (A esemény bekövetkezése nem befolyásolja B esemény bekövetkezését, és fordítva).

Feladatok

1.1. Feladat. Hányféleképpen lehet 8 bástyát letenni egy sakktáblára, hogy ne üssék egymást?

Megoldás

Az első bástya 64 helyre kerülhet. Ekkor a lefedett mező sorába és oszlopába már nem kerülhet újabb bástya, így a következőt már csak 7 sor és 7 oszlop valamelyikébe tehetjük le, ami 49 lehetőség. Minden újabb bástya letételével még egy újabb sor és oszlop kerül lefedésre. Tehát ezután sorra $36, 25, 16, 9, 4$, és 1 lehetőség van a következő bástyák letételére. Viszont a bástyák letevésének sorrendjét így figyelembe vettük, pedig mind a 8 bástya egyforma, külsőleg nem megkülönböztethető. Így le kell osztanunk a lerakott bástyák permutációinak számával, azaz $8!$-sal. Tehát összesen $\frac{64 \cdot 49 \cdot 36 \cdot 25 \cdot 16 \cdot 9 \cdot 4 \cdot 1}{8!} = 40320 = 8!$ féleképp tehetjük le a bástyákat. A végeredményt közvetlenül is megkaphatjuk, ha oszloponként (ill. soronként) nézzük a bástyák helyét.

1.2. Feladat. Mi a valószínűsége, hogy egy véletlenszerűen kiválasztott 6 jegyű szám jegyei mind különbözőek?

Megoldás

Az első számjegyet az $1, 2, \ldots, 9$ számjegyek közül, a többi számjegyet a $0, 1, 2, \ldots, 9$ számjegyek közül választhatjuk. Így az összes esetek száma $9 \cdot 10^5$. Kedvező esetek száma: $9 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5$, mert itt visszatevés nélkül választunk, a sorrend számít, illetve arra figyelünk, hogy az első számjegy ne lehessen 0. Tehát a keresett valószínűség $\frac{9 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5}{9 \cdot 10^5} = \frac{136080}{900000} = 0{,}1512$.

1.3. Feladat. Ha egy magyarkártya-csomagból (32 lap: piros, zöld, makk, tök) visszatevéssel húzunk három lapot, akkor mi annak a valószínűsége, hogy

a) pontosan egy piros színű lapot húztunk?

b) legalább egy piros színű lapot húztunk?

Megoldás

a) A 3 kihúzott lap közül $\binom{3}{1} = 3$-féleképp dönthetjük el, hogy melyik legyen a piros színű. Ezután feltehető, hogy az első húzott lap piros, a többi nem. Mivel visszatevéses mintavétel, ezért piros lap húzásának valószínűsége mindig $\frac{8}{32}$, nem piros lap húzásának valószínűsége pedig $\frac{24}{32}$. Tehát a keresett valószínűség: $\binom{3}{1} \cdot (\frac{8}{32})^1 \cdot (\frac{24}{32})^2 = 3 \cdot \frac{1}{4} \cdot \frac{9}{16} = \frac{27}{64} = 0{,}4219$.

b) Kényelmesebb most a komplementer esemény valószínűségét kivonni 1-ből. A komplementer esemény: nincsen piros a húzott lapok között. Ennek valószínűsége $\binom{3}{0} \cdot (\frac{8}{32})^0 \cdot (\frac{24}{32})^3 = \frac{27}{64}$. Tehát a keresett valószínűség $1 - \frac{27}{64} = \frac{37}{64} = 0{,}5781$.

1.4. Feladat. Egy zsákban 10 pár cipő van. 4 db-ot kiválasztva, mi a valószínűsége, hogy van közöttük pár, ha

a) egyformák a párok?

b) különbözőek a párok?

Megoldás

a) 10 balos és 10 jobbos cipő van. Mi a valószínűsége, hogy a 4 kihúzott között van balos és jobbos is? Célszerű most is a komplementer esemény valószínűségét kivonni 1-ből. A komplementer esemény: vagy 4 balosat húztunk, vagy 4 jobbosat. Ennek valószínűsége: $2 \cdot \frac{10 \cdot 9 \cdot 8 \cdot 7}{20 \cdot 19 \cdot 18 \cdot 17} = \frac{28}{323}$ vagy $\frac{\binom{10}{4}}{\binom{20}{4}} + \frac{\binom{10}{4}}{\binom{20}{4}} = \frac{28}{323}$. Tehát a keresett valószínűség $1 - \frac{28}{323} = 0{,}9133$.

b) Most is érdemes a komplementer esemény valószínűségét kiszámítani. Komplementer esemény: nincs pár a 4 cipő között. Ha így akarom a cipőket kiválasztani, akkor az elsőt 20-féleképp választhatom ki, a másodikat 18-féleképp (az első és párja kiesik), a harmadikat 16-féleképp és a negyediket 14-féleképp. Összes eset: $20 \cdot 19 \cdot 18 \cdot 17$. Tehát a komplementer esemény valószínűsége $\frac{20 \cdot 18 \cdot 16 \cdot 14}{20 \cdot 19 \cdot 18 \cdot 17} = \frac{224}{323}$ vagy kiválasztunk 10 párból a 4 párat először, majd ezek balosát ill. jobbosát $\frac{\binom{10}{4}\binom{2}{1}\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{20}{4}} = \frac{224}{323}$. Tehát a keresett valószínűség $1 - \frac{224}{323} = 0{,}3065$.

1.5. Feladat. $\star$ n dobozba véletlenszerűen helyezünk el n golyót úgy, hogy bármennyi golyó kerülhet az egyes dobozokba.

a) Mi a valószínűsége, hogy minden dobozba kerül golyó?

b) Annak mi a valószínűsége, hogy pontosan egy doboz marad üresen?

Megoldás

Vegyük észre hogy a probléma kitűzése nem határozza meg teljesen egyértelműen hogy milyen valószínűségi modellt kell használni, ugyanis nem írja elő hogy milyen módon helyezzük a golyókat a dobozokba, s azt sem rögzíti hogy megkülönböztetett vagy azonos golyókról van szó. Mindenesetre feltesszük hogy a dobozok meg vannak különböztetve (habár a feladat kitűzése ezt sem rögzíti).

a) 1. Értelmezés: A golyókat megkülönböztetjük (ez nem feltétlenül jelenti, hogy a golyók fizikailag különbözőek, már az is megkülönböztetés, hogy ha egymás után rakjuk őket a dobozokba, s így első, második stb., golyóról lehet beszélni). Ilyenkor, hacsak a feladat explicite nem ír elő mást, a "véletlenszerűen" szó értelmezése az, hogy minden golyót egymástól függetlenül, azonos $(1/n)$ valószínűséggel helyezünk a dobozokba.
Tekintsük az $n = 2$ esetet, egyszerűség kedvéért. A valószínűségi tér természetes módon egy szorzattér, $\Omega = \{1, 2\} \times \{1, 2\}$, ahol a Descartes szorzat első komponense azt kódolja el, hogy az első golyó az 1-es vagy a 2-es dobozba kerül, a második komponens ugyanezt teszi a második golyóval. Például $\omega = (2, 1)$ azt jelenti, hogy az első golyó a 2-es, a második golyó az 1-es dobozba került. Összesen $2 \cdot 2 = 4$ kimeneti lehetőség van, és a függetlenségi feltevés miatt mindegyik $1/2 \cdot 1/2 = 1/4$ valószínűségű.
Általánosan: n megkülönböztetett golyót n dobozba n^n féleképpen tudjuk betenni (ismétléses variáció). A kedvező esetek száma $n!$, azaz a lehetséges permutációk száma. Így a keresett valószínűség

$$\mathbb{P}(\text{minden dobozban van egy golyó}) = \frac{n!}{n^n}.$$

2. Értelmezés: Ha a golyók nincsenek megkülönböztetve, és a berakási folyamat sem utal rá, akkor úgy is okoskodhatunk, hogy csupán a végeredményt látjuk és a valószínűségi terünket az összes lehetséges kimenet halmazaként definiáljuk. Vegyük észre, hogy az 1. Értelmezéssel ellentétben most mindössze 3 lehetőségünk van:
(a) az első dobozban két golyó, a másodikban semmi;
(b) mindkét dobozban egy golyó;
(c) első dobozban semmi, a másodikban kettő.
Struktúrájában ez a valószínűségi tér nagyon más mint az előző, nemcsak az elemek száma különbözik, de nincs Descartes szorzat struktúrája sem. A "véletlenszerűen" szó elvileg értelmezhető úgy is, hogy a három lehetséges kimenet egyenlő valószínűségű. Így például $1/3$ annak a valószínűsége hogy mindkét dobozba egy-egy golyó került, míg az első értelmezés szerint ugyanez a valószínűség $1/2$.
Általánosan: n nem megkülönböztetett n dobozba $\binom{2n-1}{n}$ féleképpen tudjuk betenni (ismétléses kombináció). [Rendezzük az n dobozt sorba, ekkor $n-1$ válaszfal keletkezik közöttük. Az összes esetek száma az n golyó és az $n-1$ válaszfal sorrendjeinek száma, ami egy ismétléses permutáció: $\frac{\left(n+(n-1)\right)!}{n!\cdot(n-1)!} = \binom{2n-1}{n}$.] A kedvező esetek száma 1, azaz minden dobozba egy golyó kerül. Így a keresett valószínűség

$$\mathbb{P}(\text{minden dobozban van egy golyó}) = \frac{1}{\binom{2n-1}{n}}.$$

A két értelmezés közötti döntés nem matematikai hanem modellezési probléma; sokszor azonban a matematikusnak kell rámutatni a felhasználónál arra, ha esetleg a probléma nincs kellő pontossággal megfogalmazva. Rögzítsük le azonban, hogy az esetek túlnyomó többségében az első értelmezés felel meg a "véletlenszerűen" köznapi fogalmának.

b) Ha a golyókat megkülönbözőztetjük, akkor - mint előbb - az n golyót n dobozba n^n féleképpen tudjuk letenni (ismétléses variáció). A kedvező esetek számát a következőképpen kaphatjuk: az üres dobozt n féleképpen, a dobozt melyben 2 golyó lesz pedig $n-1$ féleképpen választhatjuk ki. Az n golyót $n!$ féleképpen tehetjük le, viszont kétféleképpen is eljuthatunk ugyanahhoz az elrendezéshez, hiszen a 2 golyós dobozban bármelyik jöhetett a most üres dobozból. Így a keresett valószínűség

$$\mathbb{P}(\text{pontosan egy doboz marad üresen}) = \frac{n(n-1)\frac{n!}{2}}{n^n} = \frac{\binom{n}{2}n!}{n^n}.$$

Ha a golyókat nem különbözőztetjük meg, akkor az n golyót n dobozba $\binom{2n-1}{n}$ féleképpen tudjuk betenni (ismétléses kombináció). A kedvező esetek számát a következőképpen kaphatjuk: az üres dobozt n féleképpen, a dobozt melyben 2 golyó lesz pedig $n-1$ féleképpen választhatjuk ki. Így a keresett valószínűség

$$\mathbb{P}(\text{pontosan egy doboz marad üresen}) = \frac{n(n-1)}{\binom{2n-1}{n}}.$$

1.6. Feladat. Egy boltban 10 látszólag egyforma számítógép közül 3 felújított, a többi új. Mi a valószínűsége, hogy ha veszünk 5 gépet a laborba, akkor pontosan 2 felújított lesz közöttük?

Megoldás

A 10 gépből 3 felújított, 7 új. Tehát a 3 felújított gép közül kell 2-t kiválasztani, illetve a 7 új gép közül kell a maradék 3-mat kiválasztani. A kiválasztás sorrendje nem számít, és visszatevés nélküli mintavétel. A kedvező esetek száma: $\binom{3}{2} \cdot \binom{7}{3} = 3 \cdot 35 = 105$. Összes esetek száma: $\binom{10}{5} = 252$. Tehát a keresett valószínűség $\frac{105}{252} = 0{,}4167$. (Hipergeometriai eloszlás $N = 10, M = 3, n = 5$ paraméterekkel.)

1.7. Feladat. Ha a 6 karakteres jelszavunkat véletlenszerűen választjuk a 10 számjegy és a 26 karakter közül, akkor mi a valószínűsége, hogy pontosan 3 szám lesz benne?

Megoldás

$\binom{6}{3} = 20$-féleképp lehet a 6 karakterből a 3 szám helyét kiválasztani. Ezután feltehető, hogy az első 3 karakter szám, az utolsó 3 karakter betű. Számjegy választásának valószínűsége $\frac{10}{36}$, betűé $\frac{26}{36}$. A keresett valószínűség tehát $\binom{6}{3} \cdot (\frac{10}{36})^3 \cdot (\frac{26}{36})^3 = 0{,}1615$. (Binomiális eloszlás $n = 60, p = \frac{10}{36}$ paraméterekkel.)

1.8. Feladat. Az ötöslottónál adjuk meg annak a valószínűségét, hogy egy szelvénnyel játszva öttalálatosunk lesz, illetve hogy legalább négyesünk lesz. Mi a valószínűsége, hogy minden kihúzott szám páros? (Hogy viszonylik ez a visszatevéses esethez?)

Megoldás

Annak a valószínűsége, hogy ötösünk lesz: $\frac{\binom{5}{5}}{\binom{90}{5}} = \frac{1}{\binom{90}{5}}$.
Annak a valószínűsége, hogy legalább négyesünk lesz: $\frac{\binom{5}{5}}{\binom{90}{5}} + \frac{\binom{5}{4}\binom{85}{1}}{\binom{90}{5}}$.
Annak a valószínűsége, hogy minden kihúzott szám páros: $\frac{\binom{45}{5}}{\binom{90}{5}} \approx 0,028$.

A visszatevéses esetben (tehát, mikor egy számot többször is kihúzhatunk) annak a valószínűsége, hogy párosakat húzunk: $\left(\frac{45}{90}\right)^5 = \left(\frac{1}{2}\right)^5 \approx 0,031$. Bár a két érték közel van egymáshoz, a visszatevés nélküli esetben kisebb a valószínűség, mert ott fogynak a páros számok a választás során.

1.9. Feladat. Mennyi a valószínűsége, hogy két kockadobásnál mind a két dobás 6-os, feltéve, hogy tudjuk, hogy legalább az egyik dobás 6-os?

Megoldás

Legyen A esemény az, hogy mindkét dobás hatos, B pedig, hogy legalább az egyik hatos. Ekkor

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{\frac{1}{36}}{\frac{11}{36}} = \frac{1}{11}$$

1.10. Feladat. 41 millió ötöslottó-szelvényt töltenek ki egymástól függetlenül. Mennyi a valószínűsége, hogy lesz legalább egy 5-ös találat?

Megoldás

$$P(\text{legalább egy ötös találat lesz a 41M-ból}) = 1 - P(\text{nem lesz ötös találat a 41M-ból}) \overset{\text{függetlenség}}{=}$$

$$= 1 - P(\text{egy embernek nem lesz ötös találata})^{41 \cdot 10^6} = 1 - \left(1 - \frac{\binom{5}{5}}{\binom{90}{5}}\right)^{41 \cdot 10^6} \approx 0,6066.$$

1.11. Feladat. 100 érme közül az egyik hamis (ennek mindkét oldalán fej található). Egy érmét véletlenszerűen kiválasztva és azzal 10-szer dobva, 10 fejet kaptunk. Ezen feltétellel mi a valószínűsége, hogy a hamis érmével dobtunk?

Megoldás

Jelölje A azt az eseményt, hogy 10 dobásból 10 fej, B_1 azt, hogy jó érmével dobtunk, illetve B_2 azt, hogy hamis érmével dobtunk. Ekkor:

$$P(B_1) = \frac{99}{100}; \qquad P(A|B_1) = \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{0} = \frac{1}{2^{10}}$$
$$P(B_2) = \frac{1}{100}; \qquad P(A|B_2) = 1$$

Alkalmazva a Bayes-tételt:

$$P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} = \frac{1 \cdot \frac{1}{100}}{\frac{1}{1024} \cdot \frac{99}{100} + 1 \cdot \frac{1}{100}} \approx 0.9118.$$

1.12. Feladat. Egy diák a vizsgán p valószínűséggel tudja a helyes választ. Amennyiben nem tudja, akkor tippel (az esélye, hogy eltalálja a helyes választ, ekkor $\frac{1}{3}$). Ha helyesen válaszolt, mennyi a valószínűsége, hogy tudta a helyes választ?

Megoldás

Jelölje A azt az eseményt, hogy helyesen válaszolt, B_1 azt, hogy tudta a választ, illetve B_2, hogy nem tudta a választ. Ekkor:

$$P(B_1) = p; \qquad P(A|B_1) = 1$$
$$P(B_2) = 1 - p; \qquad P(A|B_2) = \frac{1}{3}$$

Alkalmazva a Bayes-tételt:

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} = \frac{1 \cdot p}{1 \cdot p + \frac{1}{3} \cdot (1-p)} = \frac{3p}{2p+1}$$

1.13. Feladat. Egy számítógépes program két független részből áll. Az egyikben $0,2$, a másikban $0,3$ a hiba valószínűsége. Ha a program hibát jelez, akkor mi a valószínűsége, hogy mindkét rész hibás?

Megoldás

Vezessük be a következő jelöléseket:

- A - a program hibát jelez;
- B_1 - egyik rész sem hibás;
- B_2 - pontosan az egyik rész hibás;
- B_3 - mindkét rész hibás.

Ekkor

$$P(B_1) = P(\text{sem az első, sem a második}) = (1-0,2)(1-0,3) = 0,56 \qquad P(A|B_1) = 0$$
$$P(B_2) = P(\text{pontosan az egyik}) = 0,2(1-0,3) + 0,3(1-0,2) = 0,14 + 0,24 = 0,38; \qquad P(A|B_2) = 1$$
$$P(B_3) = 0,06; \qquad P(A|B_3) = 1$$

Alkalmazva a Bayes-tételt:

$$P(B_3|A) = \frac{P(A|B_3)P(B_3)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} = \frac{1 \cdot 0,06}{0 \cdot 0,56 + 1 \cdot 0,38 + 1 \cdot 0,06} = \frac{0,06}{0,44} \approx 0,1364.$$

1.14. Feladat. Egy számítógép processzorát 3 üzemben készítik. 20% eséllyel az elsőben, 30% eséllyel a másodikban és 50% eséllyel a harmadikban. A garanciális hibák valószínűsége az egyes üzemekben rendre 10%, 4%, illetve 1%. Ha a gépünk processzora elromlott, akkor mi a valószínűsége, hogy az első üzemben készült?

Megoldás

Vezessük be a következő jelöléseket:

- A - a processzorunk elromlott;
- B_1 - a processzorunk az első üzemben készült;
- B_2 - a processzorunk a második üzemben készült;
- B_3 - a processzorunk a harmadik üzemben készült.

Ekkor

$$P(B_1) = 0,2; \qquad P(A|B_1) = 0,10$$
$$P(B_2) = 0,3; \qquad P(A|B_2) = 0,04$$
$$P(B_3) = 0,5; \qquad P(A|B_3) = 0,01$$

Alkalmazva a Bayes-tételt:

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} = \frac{0,1 \cdot 0,2}{0,1 \cdot 0,2 + 0,04 \cdot 0,3 + 0,01 \cdot 0,5} \approx 0,5405$$

2. (3-4 hét) Valószínűségi változó, diszkrét eloszlások

Elmélet

A kísérletek megfigyelése során bekövetkező különböző elemi eseményekhez különböző számértékeket rendelünk, például a mérőeszköz által mutatott értéket vagy két kocka dobásakor azoknak az összegét. Ekkor előfordulhat, hogy két különböző elemi eseményhez is rendelhetünk azonos számértéket, pl. két kocka dobásakor a ⚂⚃ és ⚃⚃ dobásokhoz is 8-ast rendelünk. Ezt a hozzárendelést nevezzük valószínűségi változónak, ami tehát egy függvény az elemi események halmazán, és így számszerűsíti a kísérlet eredményét.

Definíció (X valószínűségi változó eloszlásfüggvénye)**.** $F_X(x) = P(X < x)$.

Az eloszlásfüggvény tulajdonságai: $0 \le F_X(x) \le 1$;
monoton növő;
balról folytonos;
$\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.

Állítás Tetszőleges X valószínűségi változó esetén $P(a \le X < b) = F(b) - F(a)$; $P(a < X \le b) = F(b+) - F(a+)$.

Diszkrét eloszlások:

Definíció (Diszkrét valószínűségi változó)**.** Értékkészlete legfeljebb megszámlálhatóan végtelen, azaz $\{x_1, ..., x_n, ...\}$ elemekből áll. Eloszlása: $p_i := P(X = x_i) = P(\omega : X(\omega) = x_i)$

Legyen X és Y diszkrét valószínűségi változók, melyekre $p_i = P(X = x_i)$ és $q_j = P(Y = y_j)$, ekkor $X + Y$ is diszkrét valószínűségi változó, melynek eloszlása:
$P(X + Y = z) = \sum_y P(X = z - y, Y = y) = \sum_y P(X = z - y|Y = y)P(Y = y)$.
Amennyiben fennáll, hogy $P(X = x_i, Y = y_j) = p_i q_j \ \forall\, i, j$ esetén (függetlenség, részletesen a következő fejezetben foglalkozunk ezzel), akkor
$P(X + Y = z) = \sum_y P(X = z - y)P(Y = y) = \sum_{x_i+y_j=z} p_i q_j$,
azaz az összegzés azon p_i, q_j párokra történik, melyekhez tartozó x_i és y_j összege éppen z.

Definíció (Diszkrét valószínűségi változó várható értéke)**.** Jelölése: EX.
Legyen X diszkrét valószínűségi változó, amely az x_1,x_2,... értékeket veszi fel, p_1,p_2,... valószínűségekkel, ekkor
$EX = \sum_{k=1}^{\infty} x_k p_k$, ha a végtelen összeg abszolút konvergens.

Legyen X diszkrét valószínűségi változó, melyre $p_i = P(X = x_i)$, ekkor $EX^2 = \sum_{x_i} {x_i}^2 p_i$.

Állítás Ha EX és EY véges, $a, b \in \mathbb{R}$, akkor
$E(aX + b) = aEX + b$ és
$E(X + Y) = EX + EY$.

Definíció (X szórásnégyzete)**.** $D^2X = E[(X - EX)]^2 = EX^2 - E^2X$

Definíció (X szórása)**.** $DX = \sqrt{D^2X}$

Nevezetes diszkrét eloszlások:

Név (paraméterek)	Értékek (k)	$P(X = k)$	EX	D^2X
Indikátor (p) (= Binomiális (1, p))	0, 1	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomiális (n, p)	$0, 1, ..., n$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Poisson (λ)	$0, 1, ...$	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ
Geometriai vagy Pascal (p) (= Negatív binomiális (1, p))	$1, 2, ...$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negatív binomiális (n, p)	$n, n+1, ...$	$\binom{k-1}{n-1}p^n(1-p)^{k-n}$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Hipergeometriai (N, M, n)	$0, 1, ..., n$	$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right)$

Feladatok

2.1. Feladat. Adjuk meg annak a valószínűségi változónak az eloszlását, ami egy hatgyermekes családban a fiúk számát adja meg. (Tegyük fel, hogy mindig $\frac{1}{2}$ - $\frac{1}{2}$ a fiúk, ill. a lányok születési valószínűsége.)

Megoldás

Jelölje X valószínűségi változó a fiúk számát. Ekkor a feladat visszatevéses mintavételként kezelhetõ, mely paramétereire $p = \frac{1}{2}$ és $n = 6$ teljesülnek. Amiből a kívánt eloszlás:

$$P(X = k) = \binom{6}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{6-k} = \binom{6}{k} \cdot \left(\frac{1}{2}\right)^6.$$

2.2. Feladat. Tegyük fel, hogy az új internet-előfizetők véletlenszerűen választott 20%-a speciális kedvezményt kap. Mi a valószínűsége, hogy 10 ismerősünk közül, akik most fizettek elő, legalább négyen részesülnek a kedvezményben?

Megoldás

Legyen X az a valószínűségi változó, mely megadja a speciális kedvezményt kapó ismerőseink számát. Ekkor ez egy olyan visszatevéses mintavételként kezelhető feladat, mely paramétereire $p = \frac{1}{5}$ és $n = 10$. Így pedig

$$\begin{aligned} P(X \geq 4) &= 1 - P(X < 4) = \\ &= 1 - \left[\binom{10}{0}\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^{10} + \binom{10}{1}\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^9 + \binom{10}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^8 + \binom{10}{3}\left(\frac{1}{5}\right)^3\left(\frac{4}{5}\right)^7\right] \\ &= 1 - \left[\binom{10}{0}4^{10} + \binom{10}{1}4^9 + \binom{10}{2}4^8 + \binom{10}{3}4^7\right]\left(\frac{1}{5}\right)^{10} \approx 0,1209. \end{aligned}$$

2.3. Feladat. Egy tétel áru 1% selejtet tartalmaz. Hány darabot kell találomra kivennünk és megvizsgálnunk, hogy a megvizsgált darabok között legalább 0,95 valószínűséggel selejtes is legyen, ha az egyes kiválasztott darabokat vizsgálatuk után visszatesszük?

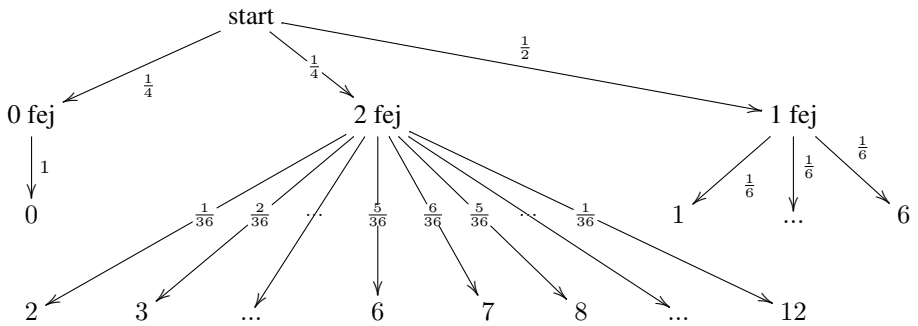
Megoldás

Legyen X = a selejtes áruk száma a vizsgált darabok közt. Ekkor
$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{n}{0} \cdot 0,01^0 \cdot 0,99^n > 0,95 \Rightarrow 0,05 > 0,99^n \Rightarrow n > \frac{\ln 0,05}{\ln 0,99} \approx 298,07 \Rightarrow n \geq 299.$

2.4. Feladat. Dobjunk egy kockával annyiszor, ahány fejet dobtunk két szabályos érmével. Jelölje X a kapott számok összegét. Adjuk meg X eloszlását!

Megoldás

Esetszétbontással érdemes. Annak a valószínűsége, hogy 0,1,2 fejet dobunk rendre 1/4, 1/2, 1/4. Az összegek 0 és 12 közé eshetnek, attól függően, hogy hány fejet dobtunk.



$P(X = 0) = \frac{1}{4} \cdot 1$
$P(X = 1) = \frac{1}{2} \cdot \frac{1}{6}$
$P(X = 2) = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{1}{36}$
$\vdots$
$P(X = 6) = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{5}{36}$
$P(X = 7) = \frac{1}{4} \cdot \frac{6}{36}$
$\vdots$
$P(X = 12) = \frac{1}{4} \cdot \frac{1}{36}$

2.5. Feladat. Jelölje X az ötöslottón kihúzott lottószámok legkisebbikét. Adjuk meg X eloszlását!

Megoldás

Jelentse $X = k$ azt, hogy a legkisebb kihúzott szám k. Ez $1 - 86$-ig bármelyik szám lehet. Ezek alapján, ha tudjuk, hogy k a legkisebb:

$$P(X = k) = \frac{\binom{90-k}{4}}{\binom{90}{5}},$$

mert a maradék kihúzott szám $k + 1$ és 90 közé eshet.

2.6. Feladat. Egy érmével dobva (tfh. p a fej valószínűsége), jelölje X az első azonosakból álló sorozat hosszát. (Azaz pl., ha a sorozat FFI..., akkor $X = 2$.) Adjuk meg X eloszlását!

Megoldás

Tegyük fel, hogy k-szor dobtunk egymás után fejet. Ez akkor lesz pontosan k hosszú sorozat, ha a k fej után közvetlenül írást dobtunk. Ugyanez fordítva is kell, hogy teljesüljön, azaz k írás után 1 fej kell. Ezek alapján az eloszlás:

$$P(X = k) = p^k(1 - p) + (1 - p)^k p$$

2.7. Feladat. Legyenek az X diszkrét valószínűségi változó értékei $-2, 1, 3$, a következő valószínűségekkel:

$$P(-2) = 1/2, \quad P(1) = 1/3, \quad P(3) = 1/6.$$

Rajzolja fel az $F(x)$ eloszlásfüggvényt!

Megoldás

$$F(x) = P(X < x) = \begin{cases} 0, & \text{ha } x \leq -2 \\ \frac{1}{2}, & \text{ha } -2 < x \leq 1 \\ \frac{1}{2} + \frac{1}{3} = \frac{5}{6}, & \text{ha } 1 < x \leq 3 \\ 1, & \text{ha } x > 3 \end{cases}$$

2.8. Feladat. Tegyük fel, hogy a 3 valószínűségszámítás gyakorlatra rendre 15, 20, illetve 25 diák jár. Várhatóan mekkora egy véletlenszerűen kiválasztott diák csoportja?

Megoldás

Legyen X a valószínűségszámítás gyakorlatra járó diákok száma. Ekkor
$P(X = 15) = 15/60 = 1/4$
$P(X = 20) = 20/60 = 1/3$
$P(X = 25) = 25/60 = 5/12$
Így a várható érték $EX = 15 \cdot 1/4 + 20 \cdot 1/3 + 25 \cdot 5/12 = (45 + 80 + 125)/12 = 250/12 = 20,83$.

2.9. Feladat. Két kockával dobunk. Egy ilyen dobást sikeresnek nevezünk, ha van 6-os a kapott számok között. Várhatóan hány sikeres dobásunk lesz n próbálkozásból?

Megoldás

Legyen X a sikeres dobások száma az n dobásból. Ekkor X egy p paraméterű binomiális eloszlást követ, melyre $p = \frac{11}{36}$ a sikeres dobás valószínűsége. Így X várható értéke $EX = np$, azaz várhatóan $\frac{11}{36}n$ sikeres dobásunk lesz.

2.10. Feladat. Tegyük fel, hogy egy dobozban van $2N$ kártyalap, melyek közül kettőn 1-es, kettőn 2-es szám van és így tovább. Válasszunk ki véletlenszerűen m lapot. Várhatóan hány pár marad a dobozban?

Megoldás

Legyen X_i annak az indikátora, hogy mindkét i feliratú lap bent marad az m lap kivétele után, azaz

$$X_i = \begin{cases} 1, & \text{ha mindkét } i \text{ feliratú lap bent marad} \\ 0, & \text{különben.} \end{cases}$$

Ekkor

$$p = P(X_i = 1) = \frac{\binom{2N-2}{m}}{\binom{2N}{m}}. \qquad \left(\text{Legyen} \binom{n}{k} := 0, \text{ ha } n < k.\right)$$

Legyen X a dobozban maradt párok száma az m lap kivétele után. Ekkor $X = X_1 + X_2 + \cdots + X_N$, melynek várható értéke

$$EX = EX_1 + EX_2 + \cdots + EX_N = Np = N\frac{\binom{2N-2}{m}}{\binom{2N}{m}} = \frac{(2N - m)(2N - 1 - m)}{2(2N - 1)}.$$

2.11. Feladat. Mennyi az ötöslottón kihúzott

a) számok összegének várható értéke?

b) páros számok számának várható értéke?

Megoldás

a) Egy húzásnál a várható érték $1 \cdot \frac{1}{90} + 2 \cdot \frac{1}{90} + \cdots + 90 \cdot \frac{1}{90} = \frac{1+2+\ldots+90}{90} = 45,5$. Öt szám kihúzása esetén pedig az összeg várható értéke $5 \cdot 45,5 = 227,5$.

b) A lottón kihúzott (páros és páratlan) számok számának várható értéke 5, azaz $E(\text{párosak száma}) + E(\text{páratlanok száma}) = 5$. Mivel ugyanannyi páros és páratlan szám közül választhatunk, így $E(\text{párosak száma}) = E(\text{páratlanok száma})$. Ez viszont csak akkor teljesülhet, ha a $E(\text{párosak száma}) = 2,5$.

Más megoldás: Jelölje X a kihúzott páros számok darabszámát. Ekkor X hipergeometrikus eloszlást követ $N = 90$, $K = 45$ és $m = 5$ paraméterekkel, így $EX = m\frac{K}{N} = 5\frac{45}{90} = 2.5$.

2.12. Feladat. Egy bükkösben a bükkmagoncok négyzetméterenkénti száma Poisson-eloszlású, $\lambda = 2,5$ db / m^2 paraméterrel. Mi a valószínűsége annak, hogy egy 1 m^2-es mintában

a) legfeljebb egy, ill.

b) több, mint három magoncot találunk?

c) Adja meg a magoncok számanak várható értékét és szórását!

Megoldás

Legyen X a bükkmagoncok négyzetméterenkénti száma. Ekkor $X \sim Poisson(\lambda)$, ahol $\lambda = 2,5$.

a) $P(X \leq 1) = P(X = 0) + P(X = 1) = 1 \cdot e^{-2,5} + 2,5 \cdot e^{-2,5} = (1 + 2,5)e^{-2,5} \approx 0,287$.

b) $P(X > 3) = 1 - P(X \leq 3) = 1 - (P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)) = 1 - (1 \cdot e^{-2,5} + 2,5 \cdot e^{-2,5} + \frac{2,5^2}{2} \cdot e^{-2,5} + \frac{2,5^3}{6} \cdot e^{-2,5}) = 1 - \left(1 + 2,5 + \frac{2,5^2}{2} + \frac{2,5^3}{6}\right) e^{-2,5} \approx 0,242$.

c) $EX = \lambda = 2,5,\ DX = \sqrt{\lambda} = \sqrt{2,5} \approx 1,58$.

2.13. Feladat. Egy adott területről származó talajmintákban a spórák száma Poisson-eloszlású. A minták harmadában egyáltalán nincs spóra. Mi a valószínűsége annak, hogy egy mintában a spórák száma egynél több? Mekkora a spórák számának várható értéke és szórása?

Megoldás

Legyen X a spórák száma a vizsgált mintában. Ekkor $X \sim Poisson(\lambda)$.
$P(X = 0) = e^{-\lambda} = \frac{1}{3}$, így $\lambda = -\ln\frac{1}{3} = \ln 3 \approx 1,099$.
$P(X > 1) = 1 - P(X \leq 1) = 1 - (P(X = 0) + P(X = 1)) = 1 - (1 \cdot e^{-\ln 3} + \ln 3 \cdot e^{-\ln 3}) \approx 0,3$.
$EX = \lambda = \ln 3$ és $DX = \sqrt{\ln 3} \approx 1,048$.

3. (5-6 hét) Abszlút folytonos eloszlások, függetlenség, egyenlőtlenségek, aszimptotikus tulajdonságok)

Elmélet

Abszolút folytonos eloszlások:

Definíció (Abszolút folytonos valószínűségi változó)**.** Ha létezik olyan $f(x)$ függvény, amelyre $F(x) = \int_{-\infty}^{x} f(t)\,dt$. Ilyenkor $f(x)$-et sűrűségfüggvénynek hívjuk. (Megjegyzés: Az f sűrűségfüggvény létezéséhez szükséges (de nem elégséges), hogy F folytonos legyen (azaz $P(X = x) = 0 \quad \forall x$-re).)

Tétel. *Legyen X abszolút folytonos eloszlású. Ekkor $f(x) = F'(x)$; $f(x) \geq 0$; $\int_{-\infty}^{\infty} f(x)\,dx = 1$; $P(X = x) = 0$ $\forall x$-re;*
$P(a < X \leq b) = P(a \leq X < b) = F(b) - F(a)$.

Definíció (Várható érték)**.** Legyen X abszolút folytonos valószínűségi változó $f(x)$ sűrűségfüggvénnyel, ekkor
$EX = \int_{-\infty}^{\infty} xf(x)\,dx$, ha az integrál létezik.

Nevezetes abszolút folytonos eloszlások:

Név (paraméterek)	Értékek	Eloszlásfüggvény (F)	Sűrűségfüggvény (f)	EX	D^2X
Standard normális $N(0,1)$	$(-\infty,\infty)$	$\Phi(x) =$ táblázatban	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad x \in \mathbb{R}$	0	1
Normális N(m,σ^2)	$(-\infty,\infty)$	visszavezethető $\Phi(x)$-re	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}} \quad x \in \mathbb{R}$	m	σ^2
Egyenletes E$[a,b]$	$[a,b]$	$\begin{cases} 0 & \text{ha } x \leq a \\ \frac{x-a}{b-a} & \text{ha } a < x \leq b \\ 1 & \text{ha } b < x \end{cases}$	$\begin{cases} \frac{1}{b-a} & \text{ha } a < x \leq b \\ 0 & \text{különben} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciális Exp(λ)	$(0,\infty)$	$\begin{cases} 1-e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\begin{cases} \lambda e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma $\Gamma(\alpha,\lambda)$	$(0,\infty)$	nincs zárt elemi képlet	$\begin{cases} \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x} & \text{ha } x \geq 0 \\ 0 & \text{különben} \end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$

Normális eloszlás standardizálása: Legyen $X \sim N(m,\sigma^2)$, ekkor $\frac{X-m}{\sigma} \sim N(0,1)$.

Függetlenség:

Definíció (Valószínűségi változók függetlesége)**.** Az $X_1, X_2, \ldots X_n$ valószínűségi változók függetlenek, ha bármely $I_1, I_2, \ldots, I_n$ intervallumra $P(X_1 \in I_1, \ldots, X_n \in I_n) = \prod_{i=1}^n P(X_i \in I_i)$
Megjegyzés: Független valószínűségi változók függvényei is függetlenek lesznek.

Tétel (Valószínűségi változók függetlensége)**.** *(i) Az $X_1, X_2, \ldots X_n$ valószínűségi változók pontosan akkor függetlenek, ha együttes eloszlásfüggvényük megegyezik eloszlásfüggvényeik szorzatával, azaz $F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n F_{X_i}(x_i)$ $\forall x$-re.*
(ii) Az $X_1, X_2, \ldots X_n$ diszkrét valószínűségi változók pontosan akkor függetlenek, ha
$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$ $\forall x_i$*-re.*
(iii) Az $X_1, X_2, \ldots X_n$ abszolút folytonos valószínűségi változók pontosan akkor függetlenek ha
$f(\mathbf{x}) = \prod_{i=1}^n f_{X_i}(x_i)$ $\forall x_i$*-re.*

Definíció (X és Y kovarianciája)**.** $cov(X,Y) = E(XY) - EXEY$

Definíció (X és Y korrelációja)**.** $R(X,Y) = \frac{cov(X,Y)}{DXDY}$

Ha X és Y függetlenek $\Rightarrow cov(X,Y) = 0$, de fordítva nem igaz.
$D^2(aX+b) = a^2D^2X, \quad D^2(X+Y) = D^2(X) + D^2(Y) + 2cov(X,Y)$

Egyenlőtlenségek:

Tétel (Markov-egyenlőtlenség)**.** *Legyen $g : \mathbb{R} \to \mathbb{R}$ monoton növő pozitív függvény, $X \geq 0$ valószínűségi változó, melyre $EX < \infty$ és $\varepsilon > 0$ tetszőleges. Ekkor*

$$P(X \geq \varepsilon) \leq \frac{E\left(g(X)\right)}{g(\varepsilon)}$$

Spec., ha $g(x) = x$, akkor

$$P(X \geq \varepsilon) \leq \frac{EX}{\varepsilon}$$

Tétel (Csebisev-egyenlőtlenség)**.** *Legyen X tetszőleges valószínűségi változó, melyre $D^2X < \infty$ és $\varepsilon > 0$ tetszőleges. Ekkor*

$$P(|X - EX| \geq \varepsilon) \leq \frac{D^2X}{\varepsilon^2}$$

Aszimptotikus tulajdonságok:

Tétel (Nagy számok törvénye (NSZT))**.** *Legyenek $X_1, X_2, \ldots$ i.i.d. valószínűségi változók, $EX_1 = m < \infty$. Ekkor*

$$\frac{X_1 + \ldots + X_n}{n} \overset{n\to\infty}{\longrightarrow} m \qquad \textit{1 valószínűséggel.}$$

Tétel (Centrális határeloszlás tétel (CHT))**.** *Legyenek $X_1, X_2, \ldots$ i.i.d. valószínűségi változók, $EX_1 = m$, $D^2X_1 = \sigma^2 < \infty$. Ekkor*

$$\frac{X_1 + \ldots + X_n - nm}{\sqrt{n}\sigma} \overset{n\to\infty}{\longrightarrow} N(0,1) \qquad \textit{gyengén,}$$

azaz

$$P\left(\frac{X_1 + \ldots + X_n - nm}{\sqrt{n}\sigma} < x\right) \overset{n\to\infty}{\longrightarrow} \Phi(x)$$

Feladatok

3.1. Feladat. Tegyük fel, hogy egy számítógép meghibásodási időpontja 0 és 10 év között van és itt geometriai modellel írható le. Határozzuk meg a jelenség eloszlásfüggvényét!

Megoldás

Legyen a ξ valószínűségi változó a meghibásodás időpontja, azaz ξ a [0,10] intervallumból veheti fel értékeit. Ekkor $P(\xi < 0) = 0$, mivel a meghibásodás időpontja nem lehet negatív. Hasonlóan $P(\xi < 10) = 1$, mivel a számítógép 10 éven túl nem üzemelhet. Ha viszont $0 < x < 10$, akkor $P(\xi < x) = \frac{x}{10}$, mivel a meghibásodás valószínűsége arányos a szakasz hosszával.

Ekkor az eloszlásfüggvény a következő alakú: $F(x) = P(\xi < x) = \begin{cases} 0 & \textit{ha } x \leq 0 \\ \frac{x}{10} & \textit{ha } 0 < x \leq 10 \\ 1 & \textit{ha } 10 < x \end{cases}$

Az ilyen eloszlásfüggvényű valószínűségi változót egyenletes eloszlásúnak nevezzük a $[0, 10]$ intervallumon.

3.2. Feladat. Legyen $0 < Y < 3$ valószínűségi változó. Eloszlásfüggvénye ezen az intervallumon $F(x) = cx^3$. Mennyi c és $P(-1 < Y < 1)$?

Megoldás

Mivel $Y < 3$, így $P(Y = 3) = 0$, tehát $F(x)$ folytonos az $x = 3$ pontban. Az eloszlásfüggvénynek monoton növekedőnek kell lennie és legfeljebb 1 lehet, vagyis c pozitív lehet csak és $x = 3$-ban már 1, vagyis

$$1 = \max_{x\in(0,3]} cx^3 = c \cdot 3^3 = 27c \quad \Rightarrow \quad c = \frac{1}{27}.$$

Tudjuk, hogy -1-ben az eloszlásfüggvény 0-át vesz fel, emiatt $P(-1 < Y < 1) = F(1) - F(-1) = \frac{1}{27} - 0$.

3.3. Feladat. Legyen X egy abszolút folytonos valószínűségi változó a $[0, c]$ intervallumon, sűrűségfüggvénye:

$$f(x) = \begin{cases} \frac{1}{9}x^2, & \text{ha } 0 \leq x < c \\ 0, & \text{ha } x < 0 \text{ vagy } x \geq c. \end{cases}$$

Határozza meg c-t és X eloszlásfüggvényét!

Megoldás

Mivel a sűrűségfüggvény integrálja $= 1$ *a* $[0, c]$ *intervallumon, így* $1 = \int_0^c \frac{1}{9} t^2 \, dt = \frac{1}{9} \left[\frac{t^3}{3} \right]_0^c = \frac{1}{9} \frac{c^3}{3}$, *amiből* $c = 3$.

Felhasználva, hogy az eloszlásfüggvény a sűrűségfüggvény integrálja:

$$F(x) = \int_0^x \frac{1}{9} t^2 \, dt = \left[\frac{1}{9} \frac{t^3}{3} \right]_0^x = \frac{x^3}{27} \quad 0 < x \leq 3, \text{ így } F(x) = P(X < x) = \begin{cases} 0, & \text{ha } x \leq 0 \\ \frac{x^3}{27}, & \text{ha } 0 < x \leq 3 \\ 1, & \text{ha } x > 3 \end{cases}$$

3.4. Feladat. Az X valószínűségi változó a $[0, c]$ intervallumon veszi fel értékeit és ott sűrűségfüggvénye $4e^{-2x}$. Határozzuk meg c értékét és annak valószínűségét, hogy $\frac{1}{4} < X < \frac{1}{2}$!

Megoldás

Mivel az eloszlásfüggvény a sűrűségfüggvény integrálja, így

$$F(x) = \int_0^x 4e^{-2t} \, dt = \left[\frac{e^{-2t}}{-2} 4 \right]_0^x = -2e^{-2x} + 2 \qquad 0 < x \leq c,$$

és $F(c) = 1$*-ből következik, hogy*

$$\begin{aligned} -2e^{-2c} + 2 &= 1 \\ e^{-2c} &= \frac{1}{2} \\ -2c &= \ln \frac{1}{2} \\ -2c &= -\ln 2 \end{aligned}$$

azaz $c = \dfrac{\ln(2)}{2} \approx 0,35$.

$P\left(\dfrac{1}{4} < X < \dfrac{1}{2} \right) = F\left(\dfrac{1}{2} \right) - F\left(\dfrac{1}{4} \right) = 1 - (-2e^{-2\frac{1}{4}} + 2) = \dfrac{2}{\sqrt{e}} - 1 \approx 0,21.$

3.5. Feladat. Véletlenszerűen választunk egy pontot az $x^2 + y^2 < 1$ kör belsejében. Jelölje Z a távolságát a középponttól. Adjuk meg Z eloszlás- és sűrűségfüggvényét valamint várható értékét!

Megoldás

Legyen Z *a középponttól való távolság. Ekkor* $0 \leq Z \leq 1$*, így itt* $F(r) = P(Z < r) = \frac{r^2 \pi}{1^2 \pi} = r^2$ *, így*

$$F(r) = \begin{cases} 0, & \text{ha } r \leq 0 \\ r^2, & \text{ha } 0 < r \leq 1 \\ 1, & \text{ha } 1 < r \end{cases}$$

Ebből deriválással adódik, hogy $f(r) = F'(r) = 2r$ *a* $[0, 1]$*-en, így*

$$f(r) = \begin{cases} 0, & \text{ha } r \leq 0 \text{ és } r > 1 \\ 2r, & \text{ha } 0 < r \leq 1 \end{cases}$$

$$EZ = \int_0^1 r \cdot 2r \, dr = \left[\frac{2r^3}{3} \right]_0^1 = \frac{2}{3}.$$

3.6. Feladat. Legyen X sűrűségfüggvénye $\dfrac{c}{x^4}$ ha $x > 1$, és 0 különben.

a) $c = ?$

b) $EX = ?$

Megoldás

a) Mivel $1 = \int_1^\infty \frac{c}{x^4}dx = \lim_{t\to\infty}\int_1^t \frac{c}{x^4}dx = \lim_{t\to\infty}\left[\frac{c}{-3\cdot x^3}\right]_1^t = \lim_{t\to\infty}\left[\frac{c}{-3\cdot t^3} - \frac{c}{-3\cdot 1^3}\right] = 0 + \frac{c}{3} = \frac{c}{3}$

$\left(\textit{egyszerűbb jelöléssel: } 1 = \int_1^\infty \frac{c}{x^4}dx = \left[\frac{c}{-3\cdot x^3}\right]_1^\infty = \frac{c}{3}\right)$, *így következik, hogy* $c = 3$.

b) $EX = \int_1^\infty x\frac{3}{x^4}\,dx = \left[\frac{-3}{2\cdot x^2}\right]_1^\infty = 1,5$

3.7. Feladat. Tapasztalatok szerint az út hossza, amit egy bizonyos típusú robogó megtesz az első meghibásodásáig exponenciális eloszlású valószínűségi változó. Ez a távolság átlagosan 6000 km. Mi a valószínűsége annak, hogy egy véletlenszerűen kiválasztott robogó

a) kevesebb, mint 4000 km megtétele után meghibásodik?

b) több, mint 6500 km megtétele után hibásodik meg?

c) 4000 km-nél több, de 6000 km-nél kevesebb út megtétele után hibásodik meg?

d) Legfeljebb mekkora utat tesz meg az első meghibásodásig a robogók leghamarabb meghibásodó 20%-a?

Megoldás

Legyen X *az első meghibásodásig megtett út. Ekkor* $X \sim Exp(\lambda)$, *ahol* $\lambda = \frac{1}{6000}$.

a) $P(X < 4000) = 1 - e^{-\frac{1}{6000}4000} \approx 0,4866$

b) $P(X > 6500) = 1 - P(X < 6500) = e^{-\frac{1}{6000}6500} \approx 0,3385$

c) $P(4000 < X < 6000) = P(X < 6000) - P(X < 4000) = 1 - e^{-\frac{1}{6000}6000} - (1 - e^{-\frac{1}{6000}4000}) \approx 0,1455$

d) $0,2 = P(X < c) = 1 - e^{-\frac{1}{6000}c}$, *azaz* $0,8 = e^{-\frac{1}{6000}c}$, *amiből* $c = -6000\ln(0,8) \approx 1338,86$.

3.8. Feladat. Egy tehén napi tejhozamát normális eloszású valószínűségi változóval, $m = 22,1$ liter várható értekkel és $\sigma = 1,5$ liter szórással, modellezzük.

a) Mi annak a valószínűsége, hogy egy adott napon a tejhozam 23 és 25 liter közé esik?

b) Mekkora valószínűséggel esik a napi tejhozam $m - \sigma$ es $m + \sigma$ közé?

($\Phi(0,6) = 0,7257, \Phi(1,93) = 0,9732, \Phi(1) = 0,8413$)

Megoldás

Legyen X *a napi tejhozam. Ekkor* $X \sim N(22,1; 1,5^2)$.

a) $P(23 < X < 25) = P(X < 25) - P(X < 23) = P\left(\frac{X-m}{\sigma} < \frac{25-m}{\sigma}\right) - P(X < 23) = P\left(\frac{X-22,1}{1,5} < \frac{25-22,1}{1,5}\right) - P(X < 23) = \Phi\left(\frac{25-22,1}{1,5}\right) - \Phi\left(\frac{23-22,1}{1,5}\right) = \Phi(1,93) - \Phi(0,6) = 0,9732 - 0,7257 = 0,2475.$

b) $P(m-\sigma < X < m+\sigma) = P(X < m+\sigma) - P(X < m-\sigma) = \Phi\left(\frac{(m+\sigma)-m}{\sigma}\right) - \Phi\left(\frac{(m-\sigma)-m}{\sigma}\right) = \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1 = 2\cdot 0,8413 - 1 = 0,6826.$

3.9. Feladat. Mennyi garanciát adjunk, ha azt szeretnénk, hogy termékeink legfeljebb 10%-át kelljen garanciaidőn belül javítani, ha a készülék élettartama 10 év várható értékű és 2 év szórású normális eloszlással közelíthető.

Megoldás

Legyen X *egy termék meghibásodásának ideje. Ekkor* $X \sim N(10, 2^2)$

$$0,1 = P(X < c) = P\left(\frac{X-10}{2} < \frac{c-10}{2}\right) = \Phi\left(\frac{c-10}{2}\right)$$

$$c = 2\cdot\Phi^{-1}(0,1) + 10 = 2\cdot(-\Phi^{-1}(0,9)) + 10 = -2\cdot 1,28 + 10 = 7,44.$$

Standard normáslis eloszás eloszlásfüggvényének értékei: `http://www.cs.elte.hu/~kovacsa/stdnormelo.pdf`

3.10. Feladat. Tegyük fel, hogy egy populációban az intelligenciahányados (IQ) normális eloszlású 110 várható értékkel és 10 szórással. Mi a valószínűsége, hogy egy véletlenszerűen kiválasztott ember IQ-ja 120 feletti?
/ $\Phi(1) = 0,8413$ /

Megoldás

Legyen X egy véletlenszerűen kiválasztott ember IQ-ja. Ekkor $X \sim N(110, 10^2)$.

$$P(X > 120) = 1 - P(X < 120) = 1 - P\left(\frac{X - 110}{10} < 1\right) = 1 - \Phi(1) = 1 - 0,8413 \approx 16\%$$

3.11. Feladat. Legyen X sűrűségfüggvénye $\frac{c}{x^4}$ ha $1 < x$, és 0 különben. Mi a c konstans értéke és mennyi D^2X?

Megoldás

$$1 = \int_1^\infty \frac{c}{x^4} dx = \left[\frac{cx^{-3}}{-3}\right]_1^\infty = 0 - \left(-\frac{c}{3}\right) = \frac{c}{3}, \text{ így } c = 3$$

$$EX = \int_1^\infty x\frac{3}{x^4} dx = \int_1^\infty 3x^{-3} dx = \left[-\frac{3}{2}x^{-2}\right]_1^\infty = 0 - \left(-\frac{3}{2}\right) = \frac{3}{2}$$

Mivel $E(g(x)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$, *így*

$$EX^2 = \int_1^\infty x^2\frac{3}{x^4} dx = \int_1^\infty 3x^{-2} dx = \left[-3x^{-1}\right]_1^\infty = 0 - (-3) = 3$$

$$D^2X = EX^2 - E^2X = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4}$$

3.12. Feladat. Legyen X egyenletes eloszlású az $[1, 4]$ intervallumon Számítsuk ki $(X - 1)^2$ várható értékét!

Megoldás

Ha $X \sim Egyenletes[1, 4]$, akkor $Y = X - 1 \sim Egyenletes[0, 3]$. Ekkor

$$E(X - 1)^2 = EY^2 = \int_0^3 y^2\frac{1}{3}\,dy = \frac{1}{3}\left[\frac{y^3}{3}\right]_0^3 = 3$$

Más megoldás:

$$E(X - 1)^2 = D^2(X - 1) + E^2(X - 1) = \frac{(3 - 0)^2}{12} + \left(\frac{0 + 3}{2}\right)^2 = \frac{3}{4} + \frac{9}{4} = 3$$

3.13. Feladat. Legyen X és Y független valószínűségi változók mindkettő 0 várható értékkel és 1 szórással. Legyen $W = X - Y$. Számítsa ki W várható értékét és szórását!

Megoldás

$EW = EX - EY = 0$ *és* $DW = \sqrt{D^2X + D^2Y} = \sqrt{2}$

3.14. Feladat. Adjon meg véges sok értéket felvehető (X) ill. végtelen sok értéket felvehető (Y) diszkrét valószínűségi változókat melyeknek szórása 1!

Megoldás

Például: Legyen $P(X = -1) = \frac{1}{2}, P(X = 1) = \frac{1}{2}$ *ill.* $Y \sim Poisson(1)$.

3.15. Feladat. Legyen $X \sim N(2, \sqrt{5}^2)$ és $Y \sim N(5, 3^2)$ függetlenek és legyen $W = 3X - 2Y + 1$. Számítsa ki
a) EW-t és D^2W-t, ill.
b) $P(W \leq 6)$-ot!
$(\Phi(1) = 0,8413)$

Megoldás

a) $EW = 3EX - 2EY + 1 = 6 - 10 + 1 = -3$ *és*
$D^2W = D^2(3X - 2Y) = D^2(3X) + D^2(-2Y) = 3^2D^2X + (-2)^2D^2Y = 9D^2X + 4D^2Y = 45 + 36 = 81$
b) Mivel független normális eloszlású valószínűségű változók összege is normális eloszlású, és $3X \sim N(6,\, 3^2 \cdot \sqrt{5}^2)$ *továbbá* $-2Y \sim N(-10,\, (-2)^2 \cdot 3^2)$*, így* $W \sim N(-3, 9^2)$.

$$P(W \leq 6) = P\left(\frac{W-(-3)}{9} < \frac{6-(-3)}{9}\right) = \Phi(1) = 0,8413$$

3.16. Feladat. Legyen X egy véges szórású valószínűségi változó és legyen $a, b \in \mathbb{R}$.
a) Mutassa meg, hogy $aX + b$ és X kovarianciája egyenlő a-szor X szórásnégyzetével!
b) Számolja ki $aX + b$ és X korrelációját ($a \neq 0$)!

Megoldás

a)

$$cov(aX + b, X) = cov(aX, X) + cov(b, X) = acov(X, X) = aD^2(X)$$

b)

$$corr(aX + b, X) = \frac{cov(aX + b, X)}{D(aX + b)DX} = \frac{aD^2X}{\sqrt{a^2}DXDX} = \begin{cases} 1, & ha\ \ a > 0 \\ -1, & ha\ \ a < 0 \end{cases}$$

3.17. Feladat. Legyen X és Y független valószínűségi változók, melyre $D^2X < \infty$ és $D^2Y < \infty$.
a) Mutassa meg, hogy $X + Y$ és X kovarianciája egyenlő X szórásnégyzetével!
b) Számolja ki $X + Y$ és X korrelációját!

Megoldás

a)

$$cov(X + Y, X) = E(X + Y)X - E(X + Y)EX = EX^2 + E(YX) - E^2X - EYEX =$$
$$= EX^2 - E^2X + E(YX) - EYEX = cov(X, X) + cov(Y, X) = D^2(X)$$

b)

$$corr(X + Y, X) = \frac{cov(X + Y, X)}{D(X + Y)DX} = \frac{D^2X}{\sqrt{D^2X + D^2Y}DX} = \frac{DX}{\sqrt{D^2X + D^2Y}}$$

3.18. Feladat. Tegyük fel, hogy egy tábla csokoládé tömege normális eloszlású 100g várható értékkel és 3g szórással. Legalább hány csokoládét csomagoljunk egy dobozba, hogy a dobozban levő táblák átlagos tömege legalább 0,9 valószínűséggel nagyobb legyen 99,5 g-nál, ha feltételezzük, hogy az egyes táblák tömege egymástól független? ($\Phi(1{,}28) = 0{,}8997$)

Megoldás

Legyen X *egy tábla csokoládé tömege,* $X \sim N(100, 3^2)$*. Ekkor* n *tábla csokoládé átlagos tömege* $\overline{X} \sim N(100, \frac{9}{n})$*, mivel* $D^2(\overline{X}) = D^2\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n D^2(X_i) = \frac{n \cdot 9}{n^2} = \frac{9}{n}$.

$$0{,}9 = P(\overline{X} > 99{,}5) = 1 - P(\overline{X} < 99{,}5) = 1 - P\left(\frac{\overline{X}-100}{\frac{3}{\sqrt{n}}} < \frac{-0{,}5 \cdot \sqrt{n}}{3}\right) = 1 - \Phi\left(\frac{-\sqrt{n}}{6}\right) = \Phi\left(\frac{\sqrt{n}}{6}\right)$$

Mivel tudjuk, hogy $\Phi(1{,}28) = 0{,}8997 \approx 0{,}9$*, így* $1{,}28 = \frac{\sqrt{n}}{6}$*. Ebből következik, hogy* $n = (6 \cdot 1{,}28)^2 = 58{,}9$*, azaz legalább* 59 *csokit kell egy dobozba csomagolni.*

3.19. Feladat. Egy scannelt kép átlagos mérete 600KB, 100KB szórással. Mi a valószínűsége, hogy 80 ilyen kép együttesen 47 és 48MB közötti tárhelyet foglal el, ha feltételezzük, hogy a képek mérete egymástól független?
($\Phi(1{,}12) = 0{,}8686$)

Megoldás

Jelölje X *egy kép eloszlását* $\mu = 600$*KB várható értékkel és* $\sigma = 100$*KB szórással. Legyen* S_n n *db ilyen valószínűségi változó összege (*$n = 80$*). A centrális határeloszlás tétel szerint*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \to Z \text{ ha } n \to \infty, \text{ ahol } Z \sim N(0, 1).$$

Tehát

$$P(47000 \leq S_n \leq 48000) = P\left(\frac{47000 - 80 \cdot 600}{\sqrt{80} \cdot 100} \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq \frac{48000 - 80 \cdot 600}{\sqrt{80} \cdot 100}\right) \approx$$

$$\approx P(-1, 12 \leq Z \leq 0) = \Phi(0) - \Phi(-1, 12) = 0, 5 - (1 - \Phi(1, 12)) = 0, 5 - (1 - 0, 8686) = 0, 3686 = 36, 9\%$$

3.20. Feladat. Egy szoftver frissítéséhez 68 file-t kell installálni, amik egymástól függetlenül 10mp várható értékű és 2mp szórású ideig töltődnek.
a) Mi a valószínűsége, hogy a teljes frissítés lezajlik 12 percen belül?
b) A cég a következő frissítésnél azt ígéri, hogy az már 95% valószínűséggel 10 percen belül betöltődik. Hány file-ból állhat ez a frissítés?
($\Phi(2,42) = 0,992, \Phi(1,645) = 0,95$)

Megoldás

Legyen X egy fájl telepítési ideje $\mu = 10$ mp várható értékkel és $\sigma = 2$ mp szórással. Jelölje S_n n db fájl telepítési idejének az összegét ($n = 68$).
a) Használva a Centrális Határeloszlás Tételt,

$$P(\text{teljes frissítés lezajlik 12 percen belül}) = P(S_n < 720) = P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} < \frac{720 - 680}{2\sqrt{68}}\right) \approx \Phi(2,42) = 99,2\%$$

b)

$$0,95 = P(S_n < 600) = P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} < \frac{600 - 10n}{2\sqrt{n}}\right) \approx \Phi\left(\frac{600 - 10n}{2\sqrt{n}}\right)$$

Mivel tudjuk, hogy $\Phi(1,645) = 0,95$, így $1,645 = \dfrac{600 - 10n}{2\sqrt{n}}$, vagyis

$$\begin{aligned} 3{,}29\sqrt{n} &= 600 - 10n \ / \ y := \sqrt{n} \\ 3{,}29y &= 600 - 10y^2 \\ 10y^2 + 3{,}29y - 600 &= 0 \\ &\to y_1 = 7{,}58, y_2 = -7{,}91 \\ y = \sqrt{n} \geq 0 &\Rightarrow \sqrt{n} = 7{,}58 \end{aligned}$$

Így következik, hogy $n = 57{,}51$, azaz legfeljebb 57 fájlból állhat a frissítés.

3.21. Feladat. Legyen egy X pozitív valószínűségi változó várható értéke $EX = 3$ és szórása $DX = 3$. Számítsuk ki, hogy legfeljebb mekkora valószínűséggel vesz fel a változó 13-at vagy annál nagyobb értéket! Mennyi a valószínűség pontos értéke, ha feltesszük, hogy az eloszlás exponenciális?

Megoldás

Markov-egyenlőtlenséggel: $P(X \geq 13) \leq \dfrac{EX}{13} = \dfrac{3}{13} \approx 0{,}23$
A Csebisev-egyenlőtlenséget $\varepsilon = 10$ értékre használva

$$P(X \geq 13) = P(X - 3 \geq 13 - 3) = P(X - 3 \geq 10) \leq P(|X - 3| \geq 10) \leq \frac{D^2X}{10^2} = \frac{9}{100} = 0,09$$

Ha X exponenciális eloszlású, akkor eloszlásfüggvénye $F(x) = 1 - e^{-\frac{1}{3}x}$, így

$$P(X \geq 13) = 1 - P(X < 13) = 1 - (1 - e^{-\frac{13}{3}}) = e^{-\frac{13}{3}} = 0,013$$

3.22. Feladat. Egy elektromos vezetékgyártó cég 40 m-es vezetékeket gyárt $0,2$ m szórással. Legfeljebb mennyi annak a valószínűsége, hogy a vezeték hossza legalább 1 m-rel eltér a várható 40 m-es értéktől?

Megoldás

A Csebisev-egyenlőtlenséget $\varepsilon = 1$ értékre használva

$$P(|X - 40| \geq 1) \leq \frac{D^2X}{1^2} = \frac{0,2^2}{1^2} = 0,04$$

Vagyis legfeljebb 0,04 annak a valószínűsége, hogy a vezeték rövidebb, mint 39 m ill. hosszabb, mint 41 m.

4. (7-8 hét) Leíró statisztikák, statisztikai alapfogalmak: becslések (maximum likelihood, momentum)

Elmélet

Definíció (Minta)**.** $X_1, ..., X_n$ valószínűségi változó sorozat. A továbbiakban feltesszük, hogy függetlenek és azonos eloszlásúak. Realizációja: $x_1, ..., x_n$

Definíció (Statisztika)**.** A minta valamely függvénye, pl.:

Mintaátlag v. átlag: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Tapasztalati szórás: $S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ *(az átlagtól való átlagos abszolút eltérés)*

Korrigált tapasztalati szórás: $S_n^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Szórási együttható (vagy relatív szórás): $V = \frac{S_n}{\overline{X}} = \frac{S_n}{\overline{X}}100\%$ *(az átlagtól való átlagos eltérés százalékban)*
/megjegyzés: lehet a korrigált tapasztalati szórással számolni/

k-adik tapasztalati momentum ($k \geq 1, k \in \mathbb{Z}$)*:* $m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

Tapasztalati módusz: a legtöbbször előforduló érték
Rendezett minta: $X_1^* \leq ... \leq X_n^*$ *a mintaelemek nem csökkenő sorrendben*

Tapasztalati medián: $X^*_{\frac{n+1}{2}}$*, ha* n *páratlan és* $\frac{X^*_{\frac{n}{2}} + X^*_{\frac{n}{2}+1}}{2}$*, ha* n *páros*

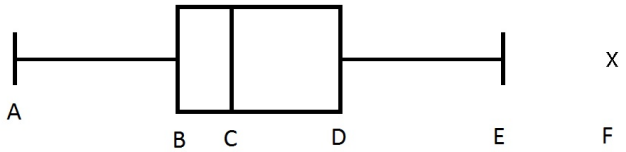
Terjedelem: $R = X_n^* - X_1^*$ *(legnagyobb − legkisebb mintaelem)*
z-kvantilis: $q_z = \inf\{x : F(x) \geq z\}$*. Ha* F *invertálható, akkor* $q_z = F^{-1}(z)$*.*
Tapasztalati z-kvantilis: q_z *értelmezése: a mintaelemek* z*-ed része legfeljebb a* q_z*, (*$1-z$*)-ed része pedig legalább a* q_z *értéket veszi fel (*$0 < z < 1$*); sokféleképpen számolható, pl. interpolációs módszerrel: először megállapítjuk a sorszámot:* $(n+1)z = e + t$ *(e: egészrész, t: törtrész), majd kiszámoljuk a z-kvantilist:* $q_z = X_e^* + t(X_{e+1}^* - X_e^*)$*.*
Kvartilisek: Speciális kvantilisek, alsó (vagy első) kvartilis: $Q_1 = q_{\frac{1}{4}}$*,*
medián: $Q_2 = q_{\frac{1}{2}}$*,*
felső (vagy harmadik) kvartilis: $Q_3 = q_{\frac{3}{4}}$
Interkvartilis terjedelem: $IQR = q_{\frac{3}{4}} - q_{\frac{1}{4}} = Q_3 - Q_1$

Tapasztalati eloszlásfüggvény: $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < x)$

ahol $I(X_i < x) = \begin{cases} 1 & \text{ha } X_i < x \\ 0 & \text{ha } X_i \geq x \end{cases}$ *indikátor függvény*

Tétel (Glivenko-Cantelli)**.** *Az* $F_n(x)$ *tapasztalati eloszlásfüggvény és az* $F(x)$ *elméleti eloszlásfüggvény közötti eltérés maximuma 1 valószínűséggelen 0-hoz konvergál, ami azt jelenti, hogy elég nagy minta esetén* $F_n(x)$ *éréke minden* x*-re tetszőleges közel van* $F(x)$ *értékéhez és* n*-et növelve mindenütt annak közelében marad.*

Definíció (Boxplot)**.**



$A = \max\{x_1^*, Q_1 - 1,5 \cdot IQR\}$, $B = Q_1$, $C = Q_2$, $D = Q_3$, $E = \min\{x_n^*, Q_3 + 1,5 \cdot IQR\}$
F: kieső értékek, azokat tüntetjük fel pontokként, amik A-n vagy E-n kívülre esnek

Legyenek $X_1, X_2, \ldots, X_n$ független, azonos eloszlású valószínűségi változók (minta) egy ϑ paraméterrel és legyen $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. A becslés a minta eloszlásának ismeretlen paraméterét közelíti a minta segítségével.

Definíció (Torzítatlan becslés)**.** A ϑ valós paraméter $T(\mathbf{X})$ becslése torzítatlan, ha $E(T(\mathbf{X})) = \vartheta$ minden ϑ paraméterértékre.

Definíció (Likelihood függvény)**.** $L(\vartheta; \mathbf{x}) = f_\vartheta(\mathbf{x}) = \prod_{i=1}^{n} f_\vartheta(x_i)$, ha az eloszlás abszolút folytonos

$L(\vartheta; \mathbf{x}) = P_\vartheta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} P_\vartheta(X_i = x_i)$, ha az eloszlás diszkrét

Definíció (Log-likelihood függvény)**.** $l(\vartheta; \mathbf{x}) = \ln(L(\vartheta; \mathbf{x}))$

Paraméterbecslési módszerek:

Maximum likelihood módszer (ML-módszer):

Azt a paraméterértéket keressük, ahol a likelihood függvény a legnagyobb értéket veszi fel (azaz diszkrét esetben az ismeretlen paraméter azon értéket keressük, amely mellett a bekövetkezett eredmény maximális valószínűségű): $\max_\vartheta L(\vartheta; \mathbf{x})$. Ez nyilván megegyezik azzal a paraméterértékkel, ahol a log-likelihood függvény veszi fel a legnagyobb értéket, azaz: $\max_\vartheta l(\vartheta; \mathbf{x})$.

Amennyiben a függvény deriválható ϑ szerint, akkor a maximumot kereshetjük a szokásos módon, a deriváltak segítségével, azonban a feladatunkat jelentősen megnehezíti, hogy olyan n-szeres szorzatot kellene deriválni, amelyiknek minden tagjában ott van az a változó, ami szerint deriválnunk kellene. Ezért likelihood függvény helyett a log-likelihood függvény maximumhelyét keressük.

Ha ϑ 1 dimenziós, akkor $\partial_\vartheta l(\vartheta, \mathbf{x}) = 0$, míg ha $\vartheta = (\vartheta_1, ..., \vartheta_p)$ p dimenziós, akkor $\partial_{\vartheta_i} l(\vartheta, \mathbf{x}) = 0$ megoldásából kapjuk a becslést. (A második deriváltak segítségével ellenőrizzük, hogy valóban maximum.)

Tétel (ML-becslés invariáns tulajdonsága)**.** *Ha ϑ ML-becslése $\hat{\vartheta}$, akkor tetszőleges g függvény esetén $g(\vartheta)$ ML-becslése $g(\hat{\vartheta})$.*

Momentum módszer:

A mintából számítható tapasztalati momentumokat ($m_i := \frac{1}{n}\sum_j x_j^i$) egyenlővé tesszük az elméleti momentumokkal ($M_i(\vartheta) := E_\vartheta X^i$), mégpedig annyit, amennyiből a paramétereket meg tudjuk határozni. p darab ismeretlen paraméter esetén tipikusan p ismeretlenes egyenletrendszert oldunk meg ϑ-ra: $M_1(\vartheta) = m_1, \ldots M_p(\vartheta) = m_p$ (megjegyzés: $m_1 = \overline{x}$)

Feladatok

4.1. Feladat. Legyen $X_1, \ldots, X_n$ független, azonos eloszlású valószínűségi változók m várható értékkel. Célunk az ismeretlen m paraméter becslése. Tekintsük az alábbi statisztikákat és állapítsuk meg, hogy melyek torzítatlanok! Amelyik nem torzítatlan, hogyan tudnánk torzítatlanná tenni?

$T_1(\mathbf{X}) = X_8, \qquad T_2(\mathbf{X}) = \dfrac{X_9 + X_{19}}{9}, \qquad T_3(\mathbf{X}) = \overline{X}$

Megoldás

$E(T_1(\mathbf{X})) = E(X_8) = m$, *így* T_1 *torzítatlan*

$E(T_2(\mathbf{X})) = E\left(\dfrac{X_9 + X_{19}}{9}\right) = \dfrac{E(X_9) + E(X_{19})}{9} = \dfrac{2m}{9}$, *így* T_2 *nem torzítatlan, viszont* $\dfrac{9}{2}T_2$ *már igen*

$E(T_3(\mathbf{X})) = E(\overline{X}) = E\left(\dfrac{1}{n}\sum_{1=1}^{n} E(X_i)\right) = m$, *így* T_3 *torzítatlan*

4.2. Feladat. Adjon torzítatlan becslést a független, azonos $E[0, \vartheta]$ eloszlású $X_1, \ldots, X_n$ minta ϑ paraméterére a mintaátlag segítségével!

Megoldás

Mivel $X_1, \ldots, X_n \sim E[0, \vartheta]$, *így* $E(X_i) = \frac{\vartheta}{2}$. *Ekkor* $E(\overline{X}) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{n}{n}E(X_1) = \frac{\vartheta}{2}$, *tehát* $E(2\overline{X}) = \vartheta$, *vagyis* $2\overline{X}$ *torzítatlan becslése* ϑ*-nak.*

4.3. Feladat. Legyen az alábbi gyakorisági tábla egy 20 elemű minta, a következő diszkrét eloszlásból:
$P(X_i = -1) = c, P(X_i = 1) = 3c, P(X_i = 2) = 1 - 4c$ $\left(i = 1, \dots 20 \text{ és } c \text{ az ismeretlen paraméter, } 0 < c < \frac{1}{4}\right)$.

érték	-1	1	2
gyakoriság	4	10	6

Határozza meg c ML-becslését és c becslését a momentum módszerrel!

Megoldás

c ML-becslése:

$$L(c, \mathbf{x}) = P(X_1 = x_1, \dots, X_{20} = x_{20}) = c^4(3c)^{10}(1-4c)^6$$

$$lnL(c, \mathbf{x}) = 4\ln(c) + 10\ln(3c) + 6\ln(1-4c)$$

$$(lnL(c, \mathbf{x}))'_c = \frac{4}{c} + \frac{10}{c} - \frac{6 \cdot 4}{1-4c}$$

Átrendezve a $(lnL(c, \mathbf{x}))'_c = 0$ *egyenletet, kapjuk, hogy*

$$\begin{aligned} \frac{14}{c} - \frac{24}{1-4c} &= 0 \\ 14(1-4c) - 24c &= 0 \\ 14 &= 80c \end{aligned}$$

így $\hat{c} = \frac{7}{40} = \frac{21}{120}$. *Ez valóban maximum, mivel* $(lnL(c, \mathbf{x}))''_c$*-t kiértékelve a* $\hat{c}$ *helyen* $(lnL(c; \mathbf{x}))''_c = -\frac{14}{c^2} - \frac{96}{(1-4c)^2} < 0$.

c becslése momentum-módszerrel:

$$M_1(c) = EX = -1 \cdot c + 1 \cdot 3c + 2 \cdot (1-4c) = 2 - 6c, \qquad m_1 = \frac{1}{20}(-1 \cdot 4 + 1 \cdot 10 + 2 \cdot 6) = 0,9$$

így az $M_1(c) = m_1$ *egyenletet c-re megoldva kapjuk, hogy* $\hat{c} = \frac{2 - 0,9}{6} = \frac{11}{60} = \frac{22}{120}$

4.4. Feladat. Legyenek $X_1, X_2, \dots, X_n$ független azonos eloszlású valószínűségi változók az alábbi eloszlásokból. Számolja ki az ismeretlen paraméter ML-becslését!
a) $Bin(m, p)$ binomiális eloszlás, ahol $m \in \mathbb{N}$ adott és p a paraméter
b) $Exp(\lambda)$ exponenciális eloszlás
c) $N(\mu, \sigma^2)$ normális eloszlás, ahol $\sigma \in \mathbb{N}$ adott és μ a paraméter

Megoldás

(Továbbá lehet, hogy érdemes megjegyezni, hogy az $\overline{x} = m$ *eset külön meggondolást igényel.)*

a)

$$L(m, p; \mathbf{x}) = \prod_{k=1}^{n} \binom{m}{x_k} p^{x_k}(1-p)^{m-x_k} \qquad (x_k = 0, 1, \dots, m)$$

$$lnL(m, p; \mathbf{x}) = \sum_{k=1}^{n} \ln \binom{m}{x_k} + \ln p \sum_{k=1}^{n} x_k + \ln(1-p) \sum_{k=1}^{n} (m - x_k)$$

$$(lnL(m, p; \mathbf{x}))'_p = \frac{1}{p}\sum_{k=1}^{n} x_k + \frac{-1}{1-p}\sum_{k=1}^{n}(m - x_k) = \frac{1}{p}\sum_{k=1}^{n} x_k + \frac{-1}{1-p}\left(nm - \sum_{k=1}^{n} x_k\right) = \frac{1}{p}n\overline{x} + \frac{-1}{1-p}(nm - n\overline{x})$$

Átrendezve a $(lnL(m, p; \mathbf{x}))'_p = 0$ *egyenletet, kapjuk, hogy*

$$\begin{aligned} \frac{1}{p}n\overline{x} + \frac{-1}{1-p}(nm - n\overline{x}) &= 0 \\ \frac{\overline{x}}{p} - \frac{m - \overline{x}}{1-p} &= 0 \\ \overline{x} - p\overline{x} - pm + p\overline{x} &= 0 \end{aligned}$$

így $\hat{p} = \dfrac{\overline{X}}{m}$. *Ez valóban maximum, mivel* $(lnL(m,p))''_p$*-t kiértékelve a* $\hat{p}$ *helyen* $(lnL(m,p;\mathbf{x}))''_p = \dfrac{-n\overline{x}}{p^2} + \dfrac{-n(m-\overline{x})}{(1-p)^2} = -n\left(\dfrac{\overline{x}}{p^2} + \dfrac{m-\overline{x}}{(1-p)^2}\right) < 0$.

b)

$$L(\lambda;\mathbf{x}) = \prod_{k=1}^{n} \lambda e^{-\lambda x_k} \qquad (x_k > 0)$$

$$lnL(\lambda;\mathbf{x}) = \sum_{k=1}^{n} \ln \lambda e^{-\lambda x_k} = \sum_{k=1}^{n} \ln \lambda + \sum_{k=1}^{n} \ln e^{-\lambda x_k} = n\ln\lambda - \lambda \sum_{k=1}^{n} x_k = n\ln\lambda - \lambda n\overline{x}$$

$$(lnL(\lambda;\mathbf{x}))'_\lambda = \frac{n}{\lambda} - \sum_{k=1}^{n} x_k = \frac{n}{\lambda} - n\overline{x}$$

Átrendezve a $(lnL(\lambda;\mathbf{x}))'_\lambda = 0$ *egyenletet, kapjuk, hogy* $\hat{\lambda} = \dfrac{1}{\overline{X}}$. *Ez valóban maximum, mivel* $(lnL(\lambda))''_\lambda = -\dfrac{n}{\lambda^2} < 0$.

c)

$$L(\mu,\sigma^2;\mathbf{x}) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum\limits_{k=1}^{n}(x_i-\mu)^2}$$

$$lnL(\mu,\sigma^2;\mathbf{x}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_i-\mu)^2$$

$$\left(lnL(\mu,\sigma^2;\mathbf{x})\right)'_\mu = -\frac{1}{2\sigma^2}(-2)\sum_{k=1}^{n}(x_i-\mu)$$

Átrendezve a $\left(lnL(\mu,\sigma^2;\mathbf{x})\right)'_\mu = 0$ *egyenletet, kapjuk, hogy* $\hat{\mu} = \dfrac{\sum\limits_{k=1}^{n} X_i}{n} = \overline{X}$. *Ez valóban maximum, mivel* $\left(lnL(\mu,\sigma^2;\mathbf{x})\right)''_\mu = -\dfrac{n}{\sigma^2} < 0$.

4.5. Feladat. Határozza meg az ismeretlen paraméter ML-becslését, ha a minta $E[a,1]$ eloszlású!

Megoldás

A paraméter függvényében nem deriválható a likelihood függvény (ugrik):

$$L(a;\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{1-a} I(a \le x_i \le 1) = \frac{1}{(1-a)^n} I(a \le x_1, x_2, ..., x_n \le 1) =$$

$$= \frac{1}{(1-a)^n} I(a \le x_1^* \le ... \le x_n^* \le 1) = \frac{1}{(1-a)^n} I(a \le x_1^*) I(x_n^* \le 1)$$

Az $I(a \le x_1^*)I(x_n^* \le 1)$ *rész 0 vagy 1 lehet, tehát úgy kell megválasztani a paramétereket, hogy 1 legyen:* $a \le x_1^*$ *és* $x_n^* \le 1$ *teljesüljön. Mivel a* $(-\infty, x_1^*]$ *intervallumon az* $\dfrac{1}{(1-a)^n}$ *függvény maximuma az* $a = x_1^*$ *pontban van, így* $\hat{a} = X_1^*$.

4.6. Feladat. Legyenek $X_1, X_2, \ldots, X_n$ független azonos $E[a,b]$ eloszlású valószínűségi változók. Számolja ki az ismeretlen paraméterek becslését a momentum módszerrel!

Megoldás

$M_1(a,b) = E(X) = \frac{a+b}{2}, \qquad m_1 = \overline{x}$

$M_2(a,b) = E(X^2) = D^2(X) + E(X)^2 = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2, \qquad m_2 = \frac{1}{n}\sum\limits_{k=1}^{n} x_k^2$

Így $M_1(a,b) = m_1$ *és* $M_2(a,b) = m_2$*-ből kapjuk, hogy*

$\frac{a+b}{2} = m_1$

$\frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2 = m_2$

és ezt oldjuk meg a, b*-re, először* m_1 *és* m_2*-vel kifejezve. Átrendezve a fenti adja, hogy* $\frac{(b-a)^2}{12} = m_2 - m_1^2$*, így*

$b - a = \sqrt{12(m_2 - m_1^2)}$

$b + a = 2m_1$.

Ezeket összeadva kapjuk, hogy $b = m_1 + \sqrt{3(m_2 - m_1^2)}$ *és* $a = m_1 - \sqrt{3(m_2 - m_1^2)}$. *Azaz a paraméterek becslése a momentum módszerrel:*

$$\hat{a} = \overline{X} - \sqrt{3\left(\frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}^2\right)} = \overline{X} - \sqrt{3}S_n \quad \textit{és} \quad \hat{b} = \overline{X} + \sqrt{3\left(\frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}^2\right)} = \overline{X} + \sqrt{3}S_n$$

5. (9-10 hét) Konfidenciaintervalluok, paraméteres próbák

Elmélet

Definíció (Konfidenciaintervallum a normális eloszlás várható értékére)**.** Legyenek $X_1, X_2, \ldots, X_n \sim N(m, \sigma^2)$ független azonos eloszlású valószínűségi változók (tfh. σ ismert). Ekkor az $(1-\alpha)100\%$-os konfidenciaintervallum m-re: $\overline{X} \pm u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, ahol $u_{1-\frac{\alpha}{2}}$ a standard normális megfelelő kvantilisét jelöli.

Hipotézisvizsgálat

Hipotézis: állítás, aminek igazságát vizsgálni szeretnénk
Statisztikai próba: eljárás aminek a segítségével döntést hozhatunk a hipotézisről

Legyen $\boldsymbol{X} = (X_1, \ldots, X_n)$ független, azonos eloszású minta. Jelölje $\mathcal{X}$ a mintateret, azaz a minta lehetséges értékeinek halmazát.

Nullhipotézis: $H_0 : \vartheta \in \Theta_0$
Ellenhipotézis: $H_1 : \vartheta \in \Theta_1$

Paramétertér: $\Theta = \Theta_0 \cup \Theta_1$
Döntés: $T(\boldsymbol{X})$ statisztika ($T : \mathcal{X} \to \mathbb{R}$ próbastatisztika) segítségével, melynek ismerjük az eloszlását a nullhipotézis fennállása esetén

Mintateret két részre bontjuk: $\mathcal{X} = \mathcal{X}_e \cup \mathcal{X}_k$ és $\mathcal{X}_e \cap \mathcal{X}_k = \emptyset$
$\mathcal{X}_k$: kritikus tartomány − azon $\boldsymbol{X}$ megfigyelések halmaza, amikre elutasítjuk a nullhipotézist
$\mathcal{X}_e$: elfogadási tartomány − azon $\boldsymbol{X}$ megfigyelések halmaza, amikre elfogadjuk a nullhipotézist
Kritikus érték: c (függ α-tól, ld. alább)
$\mathcal{X}_k = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{x}) \geq c\}$ vagy $\mathcal{X}_k = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{x}) \leq c\}$ vagy $\mathcal{X}_k = \{\boldsymbol{x} \in \mathcal{X} : |T(\boldsymbol{x})| \geq c\}$
$\mathcal{X}_e = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{x}) < c\}$ $\mathcal{X}_e = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{x}) > c\}$ $\mathcal{X}_e = \{\boldsymbol{x} \in \mathcal{X} : |T(\boldsymbol{x})| < c\}$

	Döntés	
Valós állapot	*H_0-t elfogadjuk ($\mathcal{X}_e$)*	*H_0-t elvetjük ($\mathcal{X}_k$)*
H_0 igaz ($\vartheta \in \Theta_0$)	*helyes döntés ($1-\alpha$)*	*elsőfajú hiba (α)*
H_0 hamis ($\vartheta \in \Theta_1$)	*másodfajú hiba (β)*	*helyes döntés ($1-\beta$)*

Elsőfajú hiba valószínűsége:
Egyszerű hipotézis (Θ_0 halmaz egyelemű) esetén: $\mathbb{P}_{\vartheta_0}(\boldsymbol{X} \in \mathcal{X}_k) = \alpha \quad \vartheta_0 \in \Theta_0$ */ = ℙ(elvetjük H_0-t | H_0 igaz) /*
Összetett hipotézis (Θ_0 halmaz több elemű) esetén: $\mathbb{P}_{\vartheta}(\boldsymbol{X} \in \mathcal{X}_k) \leq \alpha \quad \forall \vartheta \in \Theta_0$

*Próba (pontos) terjedelme vagy **szignifikanciaszintje**:* $\alpha = \sup\{\mathbb{P}_{\vartheta}(\boldsymbol{X} \in \mathcal{X}_k) : \vartheta \in \Theta_0\}$
***Megbízhatósági (konfidencia-) szint**:* $1-\alpha$ */ = ℙ(elfogadjuk H_0-t | H_0 igaz) /*

A próba meghatározása: előre rögzített α terjedelemhez azt a c értéket keressük, amire a próba pontos terjedelme éppen α.

Másodfajú hiba valószínűsége:
$\beta(\vartheta) = \mathbb{P}_{\vartheta}(\boldsymbol{X} \in \mathcal{X}_e) = 1 - \mathbb{P}_{\vartheta}(\boldsymbol{X} \in \mathcal{X}_k) \quad \vartheta \in \Theta_1$ */ = ℙ$_\vartheta$(elfogadjuk H_0-t | H_0 hamis) /*

Erőfüggvény: $\psi(\vartheta) = 1 - \beta(\vartheta)$ */ = ℙ(elvetjük H_0-t | H_0 hamis) /*
Minél erősebb a próba, annál nagyobb valószínűséggel veti el a hamis nullhipotézist. Vagyis a próba ereje annak a valószínűsége, hogy egy adott különbséget adott mintanagyság és terjedelem mellett egy statisztikai próba kimutat. (Kísérletek tervezésekor az erő nagyságának előre meghatározott értékéből határozható meg a mintanelemszám.) A próba erejét addig nem tudjuk kiszámolni, ameddig az ellenhipotézis egy értékét nem rögzítjük ill. nem mondjuk meg a különbég nagyságát, amit ki szeretnénk mutatni.

***p-érték**: annak a valószínűsége, hogy igaz H_0 esetén a tapasztalt eltérést vagy annál nagyobb eltérést kapunk. Ha egy próbát számítógép segítségével végzünk el, rendszerint a p-érték révén tudunk dönteni: ha p-érték$< \alpha$, akkor elvetjük H_0-t.*

A hipotézisek nem egyenrangúak. H_0-t csak indokolt esetben szeretnénk elutasítani, így az elsőfajú hiba súlyosabbnak számít, mint a másodfajú hiba. Általában az elsőfajú hiba legnagyobb valószínűségét adjuk meg, de a másodfajú hiba csökkentésére is törekszünk (pl. mintanagyság növelésével).
***H_0 elfogadása**: statisztikailag nem találtunk komoly bizonyítékot arra, hogy H_0 nem lenne igaz; vagyis H_0 elfogadása esetén sem lehet állítani, hogy H_0 teljesül*

***H_0 elvetése**: statisztikailag komoly bizonyítékot találtunk arra, hogy a H_0 nem igaz, azaz H_1 igaz*

Próbák normális eloszlás várható értékére

Egymintás próbák

$X_1, \dots, X_n \sim N(m, \sigma^2)$

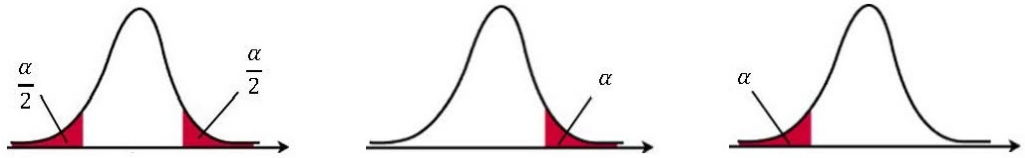
$H_0 : m = m_0$	$H_0 : m \le m_0$	$H_0 : m \ge m_0$
$H_1 : m \ne m_0$	$H_1 : m > m_0$	$H_1 : m < m_0$

Egymintás u-próba (σ ismert)

Próbastatisztika: $T(\boldsymbol{X}) = u = \dfrac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$

Kritikus tartományok:

$\mathcal{X}_k = \{\mathbf{X} : |u| > u_{1-\frac{\alpha}{2}}\}$ $\quad$ $\mathcal{X}_k = \{\mathbf{X} : u > u_{1-\alpha}\}$ $\quad$ $\mathcal{X}_k = \{\mathbf{X} : u < u_{\alpha}\}$



Kapcsolat a konfidenciaintervallummal (az alábbi lépések ekvivalensek):

$$\boxed{|u| > u_{1-\frac{\alpha}{2}}} \Leftrightarrow u > u_{1-\frac{\alpha}{2}} \text{ vagy } u < -u_{1-\frac{\alpha}{2}} \Leftrightarrow \frac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}} > u_{1-\frac{\alpha}{2}} \text{ vagy } \frac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}} < -u_{1-\frac{\alpha}{2}} \Leftrightarrow$$

$$\overline{X} - m_0 > u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \text{ vagy } \overline{X} - m_0 < -u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Leftrightarrow \boxed{m_0 \notin \left(\overline{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)}$$

Vagyis a null hipotézist pontosan akkor utasítjuk el, ha a $(1-\alpha)100\%$-os konfidenciaintervallum nem tartalmazza m_0-t.

Egymintás t-próba (σ ismeretlen)

Próbastatisztika: $T(\boldsymbol{X}) = t = \dfrac{\overline{X} - m_0}{\frac{s_n^*}{\sqrt{n}}} \overset{H_0 \text{ esetén}}{\sim} t_{n-1}$

Kritikus tartományok:

$\mathcal{X}_k = \{\mathbf{X} : |t| > t_{n-1,1-\alpha/2}\}$ $\quad$ $\mathcal{X}_k = \{\mathbf{X} : t > t_{n-1,1-\alpha}\}$ $\quad$ $\mathcal{X}_k = \{\mathbf{X} : t < t_{n-1,\alpha}\}$

Kétmintás próbák

$X_1, \dots, X_n \sim N(m_1, \sigma_1^2)$, $\quad Y_1, \dots, Y_m \sim N(m_2, \sigma_2^2)$ $\quad$ *függetlenek*

$H_0 : m_1 = m_2$	$H_0 : m_1 \le m_2$	$H_0 : m_1 \ge m_2$
$H_1 : m_1 \ne m_2$	$H_1 : m_1 > m_2$	$H_1 : m_1 < m_2$

<table>
<tr><th></th><th colspan="2">a két minta
független</th><th>a két minta páronként összetartozó,
nem független</th></tr>
<tr><td>σ_1 és σ_2 ismert</td><td colspan="2">Kétmintás u-próba</td><td>Egymintás u-próba
a különbségekre</td></tr>
<tr><td rowspan="3">σ_1 és σ_2 ismeretlen</td><td colspan="2">előzetes F-próba</td><td rowspan="3">Egymintás t-próba
a különbségekre</td></tr>
<tr><td>$\sigma_1 = \sigma_2$</td><td>$\sigma_1 \ne \sigma_2$</td></tr>
<tr><td>Kétmintás t-próba</td><td>Welch-próba</td></tr>
</table>

előzetes F-próba (σ_1, σ_2 ismeretlen)

$H_0 : \sigma_1 = \sigma_2$

$H_1 : \sigma_1 \ne \sigma_2$

Próbastatisztika:

$$F = \begin{cases} \frac{(s_1^*)^2}{(s_2^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{n-1,m-1} & \text{ha } s_1^* > s_2^* \\ \frac{(s_2^*)^2}{(s_1^*)^2} \overset{H_0 \text{ esetén}}{\sim} F_{m-1,n-1} & \text{ha } s_2^* > s_1^* \end{cases}$$

Kétmintás u-próba (σ_1, σ_2 ismert)

Próbastatisztika: $u = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} N(0,1)$

Kétmintás t-próba ($\sigma_1 = \sigma_2$ ismeretlen)

Próbastatisztika: $t = \sqrt{\dfrac{nm}{n+m}} \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)(s_1^*)^2 + (m-1)(s_2^*)^2}{n+m-2}}} \overset{H_0 \text{ esetén}}{\sim} t_{n+m-2}$

Welch-próba ($\sigma_1 \ne \sigma_2$ ismeretlen)

Próbastatisztika: $t' = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{(s_1^*)^2}{n} + \frac{(s_2^*)^2}{m}}} \overset{H_0 \text{ esetén}}{\sim} t_f$ *, ahol* $f \approx \dfrac{\left(\frac{(S_1^*)^2}{n} + \frac{(S_2^*)^2}{m}\right)^2}{\frac{\left(\frac{(S_1^*)^2}{n}\right)^2}{n-1} + \frac{\left(\frac{(S_2^*)^2}{m}\right)^2}{m-1}}$

Feladatok

5.1. Feladat. Legyen X_1, X_2, X_3, X_4 független azonos $N(\mu, 2^2)$ eloszlású minta. A megfigyelt értékek a következők:
14,8; 12,2; 16,8; 11,1
a) Adjon 95%-os megbízhatóságú konfidenciaintervallumot μ-re!
b) Hány elemű mintára van szükség, ha azt szeretnénk, hogy a konfidenciaintervallum legfeljebb $1{,}6$ hosszúságú legyen?
($u_{0{,}975} = 1{,}96$)

Megoldás

a) Adatok: $n = 4$
$\overline{x} = \frac{14{,}8+12{,}2+16{,}8+11{,}1}{4} = 13{,}725$
$\sigma = 2$
$\alpha = 0{,}05$
Ekkor $u_{0{,}975} = 1{,}96$*, így az* $(1-\alpha)100\%$ *megbízhatóságú konfidenciaintervallum* μ*-re:*

$$\left(\overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = \left(13{,}725 - u_{0{,}975}\frac{2}{\sqrt{4}}, 13{,}725 + u_{0{,}975}\frac{2}{\sqrt{4}}\right) = (11{,}765; 15{,}685)$$

R-kód:
```
minta<-c(14.8, 12.2, 16.8, 11.1)
n<-length(minta)
sigma<-2
alpha<-0.05
mean(minta)-qnorm(1-alpha/2)*sigma/sqrt(n)
mean(minta)+qnorm(1-alpha/2)*sigma/sqrt(n)
```

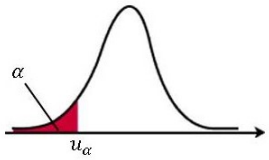
b) A konfidenciaintervallum hossza: $2u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 2u_{0{,}975}\frac{2}{\sqrt{n}} < 1{,}6$*, így* $n > \left(\frac{4u_{0{,}975}}{1{,}6}\right)^2 = \left(\frac{4 \cdot 1{,}96}{1{,}6}\right)^2 \approx 24{,}01$
tehát legalább 25 *elemű mintára van szükség.*

R-kód (folytatás):
```
hossz<-1.6
(2*qnorm(1-alpha/2)*sigma/hossz)^2
```

5.2. Feladat. Azt szeretnénk vizsgálni, hogy a napi középhőmérséklet október 18-án Budapesten $15°$C alatt volt-e? Az elmúlt 4 év napi középhőmérsékletei a következők voltak: $14,8; 12,2; 16,8; 11,1\ °C$, valamint tegyük fel, hogy az adatok normális eloszlásból származnak.
a) Írjuk fel a null- és ellenhipotézist!
b) Tegyük fel, hogy a napi középhőmérséklet szórása $\sigma = 2$. Tesztelje a fenti hipotézist $\alpha = 0.05$ terjedelem mellett! Adja meg a kritikus tartományt és p-értéket! Mi a döntés?
c) Tesztelje a hipotézist úgy is, hogy nem használja a szórásra vonatkozó előzetes információt!
d) Milyen hipotézist írjunk fel, ha azt szeretnénk vizsgálni, hogy a napi középhőmérséklet október 18-án Budapesten $15°$C-tól különböző volt? Teszteljük a fenti adatok segítségével!
($u_{0{,}05} = -1{,}645,\ \Phi(1{,}275) = 0{,}899,\ t_{3;0{,}05} = -2{,}353\ u_{0{,}975} = 1{,}96$)

Megoldás

a) Legyen m *a napi középhőmérséklet október 18-án Budapesten. Ekkor*
$H_0 \colon m \geq 15$
$H_1 \colon m < 15$
b) Adatok: $n = 4$
$\overline{x} = \frac{14{,}8+12{,}2+16{,}8+11{,}1}{4} = 13{,}725$
$\sigma = 2$
$\alpha = 0{,}05$
Milyen próbát használjunk? Egymintás, egyoldali u-próbát.
Próbastatisztika: $U = \frac{\overline{X} - 15}{\frac{\sigma}{\sqrt{n}}} \overset{H_0 \text{esetén}}{\sim} N(0,1)$*, melynek értéke:* $u = \frac{13{,}725 - 15}{\frac{2}{\sqrt{4}}} = -1{,}275$
U *mely értékeire utasítjuk el* H_0*-t?*
H_0 *esetén* $P(U < u_\alpha) = \Phi(u_\alpha) = \alpha$*, azaz* $\Phi(u_{0{,}05}) = 0{,}05$ *tehát* $u_{0{,}05} = -1{,}645$ *így a kritikus tartomány* $= \{\mathbf{x} \in \chi : U < u_\alpha\} = \{\mathbf{x} \in \chi : U < -1{,}645\}$.
Mivel most $u = -1{,}275 > -1{,}645$*, nem utasítjuk el* H_0*-t. Azaz nincs elég bizonyítékunk, hogy a napi középhőmérséklet október 18-án Budapesten* $15°C$ *alatt lenne.*



A hipotézist a p-érték α-val való összehasonlításával is tesztelhetjük:

p-érték $= \Phi(-1,275) = 1 - \Phi(1,275) = 1 - 0,899 = 0,101 > \alpha = 0,05$*, így nem vetjük el* H_0*-t.*

c) Adatok: $n = 4$

$\overline{x} = \frac{14,8+12,2+16,8+11,1}{4} = 13,725$

$s_n^* = \sqrt{\frac{(14,8-13,725)^2+(12,2-13,725)^2+(16,8-13,725)^2+(11,1-13,725)^2}{3}} = \sqrt{6,6092} = 2,57$

$\alpha = 0,05$

Milyen próbát használjunk? Egymintás, egyoldali t-próbát.

Próbastatisztika: $T = \frac{\overline{X} - 15}{\frac{s_n^*}{\sqrt{n}}} \overset{H_0 esetén}{\sim} t_{n-1}$*, melynek értéke:* $t = \frac{13,725 - 15}{\frac{2,57}{\sqrt{4}}} = -0,99$

T *mely értékeire utasítjuk el* H_0*-t?*

H_0 *esetén* $P(T < t_{n-1;\alpha}) = \alpha$*, azaz* $P(T < t_{3;0,05}) = 0,05$ *tehát* $t_{3;0,05} = -2,353$ *így a kritikus tartomány* $= \{\mathbf{x} \in \chi : T < t_{n-1;\alpha}\} = \{\mathbf{x} \in \chi : T < -2,353\}$.

Mivel most $t = -0,99 > -2,353$*, nem utasítjuk el* H_0*-t. Azaz nincs elég bizonyítékunk, hogy a napi középhőmérséklet október 18-án Budapesten* $15°C$ *alatt lenne.*

A hipotézist a p-érték α-val való összehasonlításával is tesztelhetjük:

p-érték $= P(t_3 < -0,99) = 0.198 > \alpha = 0,05$*, így nem vetjük el* H_0*-t.*

d) Legyen m *a napi középhőmérséklet október 18-án Budapesten. Ekkor*

$H_0 \colon m = 15$

$H_1 \colon m \neq 15$

Adatok: $n = 4$

$\overline{x} = \frac{14,8+12,2+16,8+11,1}{4} = 13,725$

$\sigma = 2$

$\alpha = 0,05$

Milyen próbát használjunk? Egymintás, kétoldali u-próbát.

Próbastatisztika: $U = \frac{\overline{X} - 15}{\frac{\sigma}{\sqrt{n}}} \overset{H_0 esetén}{\sim} N(0,1)$*, melynek értéke:* $u = \frac{13,725 - 15}{\frac{2}{\sqrt{4}}} = -1,275$

U *mely értékeire utasítjuk el* H_0*-t?*

H_0 *esetén* $P(|U| > u_{1-\frac{\alpha}{2}}) = \alpha$*, azaz* $P(U < u_{0,975}) = 0,975$ *tehát* $u_{0,975} = 1,96$ *így a kritikus tartomány* $= \{\mathbf{x} \in \chi : |U| > u_{1-\frac{\alpha}{2}}\} = \{\mathbf{x} \in \chi : |U| > 1,96\}$.

Mivel most $|u| = 1,275 < 1,96$*, nem utasítjuk el* H_0*-t. Azaz nincs elég bizonyítékunk, hogy a napi középhőmérséklet október 18-án Budapesten* $15°C$*-tól különböző lenne.*

A hipotézist a p-érték α-val való összehasonlításával is tesztelhetjük:

p-érték $= 2 \cdot (1 - \Phi(1,275)) = 0.202 > \alpha = 0,05$*, így nem utasítjuk el* H_0*-t.*

A hipotézist a várható értékre vonatkozó 95%*-os konfidenciaintervallum segítségével is tesztelhetjük:*

A konfidenciaintervallum $(11,765; 15,685)$ *(Feladat 1.) tartalmazza a* 15*-öt, így nem utasítjuk el* H_0*-t.*

5.3. Feladat. Az alábbi két minta két különböző gyáregységben tapasztalt selejtarányra vonatkozik (ezrelékben). Állítható-e, hogy az „A” gyáregység jobban dolgozott? (Feltételezhetjük, hogy a minták normális eloszlásúak, függetlenek.)

A	11,9	12,1	12,8	12,2	12,5	11,9	12,5	11,8	12,4	12,9
B	12,1	12,0	12,9	12,2	12,7	12,6	12,6	12,8	12,0	13,1

$(F_{9,9;0,975} = 4,026,\ t_{18;0,05} = -1,734)$

Megoldás

Jelölje m_A *az „A” és* m_B *az „B” gyáregység selejtarányát. Ekkor*

$H_0 \colon m_A \geq m_B$

$H_1 \colon m_A < m_B$

Adatok: $n_A = 10$, $n_B = 10$

$\overline{x}_A = \frac{11{,}9+\cdots+12{,}9}{10} = 12{,}3$

$\overline{x}_B = \frac{12{,}1+\cdots+13{,}1}{10} = 12{,}5$

$s_A^{*2} = \frac{(11{,}9-12{,}3)^2+\cdots+(12{,}9-12{,}3)^2}{9} = \frac{132}{900} = 0{,}147$

$s_B^{*2} = \frac{(11{,}9-12{,}5)^2+\cdots+(12{,}9-12{,}5)^2}{9} = \frac{142}{900} = 0{,}158$

$\alpha = 0{,}05$

Van különbség a szórások közt? Előzetes F-próbával tesztelünk.

$H_0 \colon \sigma_A^2 = \sigma_B^2$

$H_1 \colon \sigma_A^2 \neq \sigma_B^2$

$F = \frac{s_B^{*2}}{s_A^{*2}} = \frac{142}{132} = 1{,}076 <$ *kritikus érték* $= F_{9,9;0,975} = 4{,}026$*, tehát nem utasítjuk el H_0-t, így nincs rá okunk, hogy a két szórást különbözőnek tekintsük.*

Milyen próbát használjunk? Kétmintás, egyoldali t-próbát.

$T = \sqrt{\frac{n_A n_B}{n_A+n_B}} \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{(n_A-1)s_A^{*2}+(n_B-1)s_B^{*2}}{n_A+n_B-2}}} \overset{H_0 \text{ esetén}}{\sim} t_{n+m-2}$*, melynek értéke:* $t = \sqrt{\frac{10\cdot 10}{10+10}} \frac{12{,}3-12{,}5}{\sqrt{\frac{9\cdot 0{,}147+9\cdot 0{,}158}{10+10-2}}} = -1{,}13$

T mely értékeire utasítjuk el H_0-t?

H_0 esetén $P(T < t_{n_A+n_B-2;\alpha}) = \alpha$*, azaz* $P(T < t_{18;0,05}) = 0{,}05$ *tehát* $t_{18;0,05} = -1{,}734$ *így a kritikus tartomány* $= \{\mathbf{x} \in \chi : T < t_{n_A+n_B-2;\alpha}\} = \{\mathbf{x} \in \chi : T < -1{,}734\}$.

Mivel most $t = -1{,}13 > -1{,}734$*, nem utasítjuk el H_0-t, azaz nincs elég bizonyítékunk arra, hogy az „A” gyáregység jobban dolgozott.*

5.4. Feladat. Két szervert hasonlítottunk össze. Az elsőn 30 futás átlagos ideje 6,7 mp volt, míg ettől függetlenül a másodikon 20 futásé 7,2 mp. Vizsgáljuk meg, hogy van-e szignifikáns különbség a két szerver sebessége közt, ha a futási idők szórása mindkét gépen 0,5 volt?
($u_{0,975} = 1{,}96$)

Megoldás

Jelölje m_1 és m_2 az első illetve a második szerveren való futás idejét. Ekkor

$H_0 \colon m_1 = m_2$

$H_1 \colon m_1 \neq m_2$

Adatok: $n_1 = 30$, $n_2 = 20$

$\overline{x}_1 = 6.7$

$\overline{x}_2 = 7.2$

$\sigma_1 = \sigma_2 = 0.5$

Milyen próbát használjunk? Kétmintás, kétoldali u-próbát (szórások ismertek).

$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \overset{H_0 \text{esetén}}{\sim} N(0,1)$*, melynek értéke:* $u = \frac{6.7-7.2}{\sqrt{\frac{0.5^2}{30}+\frac{0.5^2}{20}}} = -3.464$

U mely értékeire utasítjuk el H_0-t?

H_0 esetén $P(|U| > u_{1-\frac{\alpha}{2}}) = \alpha$*, azaz* $P(U < u_{0,975}) = 0{,}975$ *tehát* $u_{0,975} = 1{,}96$ *így a kritikus tartomány* $= \{\mathbf{x} \in \chi : |U| > u_{1-\frac{\alpha}{2}}\} = \{\mathbf{x} \in \chi : |U| > 1{,}96\}$.

Mivel most $|u| = 3{,}464 > 1{,}96$*, elutasítjuk H_0-t, azaz a két szerver futási ideje közt szignifikáns különbség van.*

5.5. Feladat. Az alábbi két minta 10 forgalmas csomópont levegőjében található szennyezőanyag koncentrációra vonatkozó két adatsort tartalmaz. Az első sorban a november 15-i, a másodikban a november 29-i számok szerepelnek. Szignifikánsan változott-e a légszennyezettség?

november 15.	20,9	17,1	15,8	18,8	20,1	15,6	14,8	24,1	18,9	12,5
november 29.	21,4	16,7	16,4	19,2	19,9	16,6	15,0	24,0	19,2	13,2

($t_{9;0,975} = 2{,}262$)

Megoldás

Jelölje m_1 és m_2 a november 15-i illetve a november 29-i légszennyeződés várható értékét. Ekkor

$H_0 \colon m_1 = m_2$

$H_1 \colon m_1 \neq m_2$

Mivel ugyanazokon a helyeken mérték a légszennyezettséget, a minták páronként összetartozóak (egymástól nem független megfigyeléseink vannak). A légszennyeződés változására vonatkozó információ a két mérési eredmény különbségében rejlik.

november 29. - november 15.	*0,5*	*-0,4*	*0,6*	*0,4*	*-0,2*	*1,0*	*0,2*	*-0,1*	*0,3*	*0,7*

Jelöljük m*-mel a november* 29*-én és a november* 15*-én mért légszennyeződés várható értékének különbségét, azaz* $m = m_2 - m_1$*. Ekkor a fenti hipotézis a következőképpen fogalmazható meg:*

$H_0 \colon m = 0$
$H_1 \colon m \neq 0$

Adatok: $n = 10$
$\overline{x} = \frac{0{,}5+\cdots+0{,}7}{10} = 0{,}3$
$s_n^* = \sqrt{\frac{(0{,}5-0{,}3)^2+\cdots+(0{,}07-0{,}3)^2}{10-1}} = 0{,}435$
$\alpha = 0{,}05$

Milyen próbát használjunk? Egymintás, kétoldali t-próbát.

$T = \frac{\overline{X} - 0}{\frac{s_n^*}{\sqrt{n}}} \overset{H_0 \text{ esetén}}{\sim} t_{n-1}$*, melynek értéke:* $t = \frac{0{,}3 - 0}{\frac{0{,}435}{\sqrt{10}}} = 2{,}18$

T *mely értékeire utasítjuk el* H_0*-t?*

H_0 *esetén* $P(|T| > t_{n-1;1-\frac{\alpha}{2}}) = \alpha$*, azaz* $P(T < t_{9;0{,}975}) = 0{,}975$ *tehát* $t_{9;0{,}975} = 2{,}262$ *így a kritikus tartomány* $= \{\mathbf{x} \in \chi : |T| > t_{n-1;1-\frac{\alpha}{2}}\} = \{\mathbf{x} \in \chi : |T| > 2{,}262\}$.
Mivel most $t = 2{,}18 < 2{,}262$*, nem utasítjuk el* H_0*-t, azaz nincs elég bizonyítékunk, hogy különbség lenne a november* 15*-i és* 29*-i légszennyeződés mértéke közt.*

Viszont vegyük észre, hogy a próbastatisztika értéke közel van a kritikus értékhez. Ezt a p-érték α*-hoz közeli értékéből is látjuk: p-érték* $= P(|t_9| > 2{,}18) = 0{,}057$*. Ez utóbbi azt mutatja, hogy az* $\alpha = 0{,}05$ *szignifikanciaszinten nem utasítjuk el a nullhipotézist, viszont egy* $0{,}057$*-nél magasabb szinten már igen.*

6. (11-12 hét) Nemparaméteres próbák, egyszerű lineáris regresszió

Elmélet

Nemparaméteres próbák:

Diszkrét illeszkedésvizsgálat

Legyen $X_1, \dots, X_n$ egy n elemű minta és tegyük fel, hogy a mintaelemek r különböző x_j $(j = 1, \dots r)$ értéket vehetnek fel. Továbbá jelölje $\nu_j (j = 1, \dots r)$ az egyes értékek megfigyelt gyakoriságát, azaz n független megfigyelést osztályozunk valamilyen szempont szerint, r páronként diszjunkt osztályba. Az egyes osztályok feltételezett valószínűségei rendre $p_1, \dots p_r$.

Osztályok	*1*	*2*	*...*	*r*	*Összesen*
Értékek	x_1	x_2	...	x_r	
Gyakoriságok	ν_1	ν_2	...	ν_r	*n*
Valószínűségek	p_1	p_2	...	p_r	*1*

Azt vizsgáljuk, hogy a minta eloszlása megegyezik-e a feltételezett eloszlással. Ismert eloszlás esetén tiszta illeszkedésvizsgálatot végzünk. Ha viszont az eloszlás paraméteres és csak az eloszláscsaládot ismerjük, a paraméter(eke)t viszont nem (pl. az a kérdés, hogy származhatnak-e az adatok p paraméterű binomiális eloszlásból), akkor becsléses illeszkedésvizsgálatot végzünk.

Tiszta illeszkedésvizsgálat:

$H_0 : P(X_i = x_j) = p_j \quad j = 1, \dots, r$
$H_1 : \exists$ *legalább egy* j *melyre* $P(X_i = x_j) \neq p_j$

Próbastatisztika: $T_n = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$ *Kritikus tartomány:* $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Becsléses illeszkedésvizsgálat:

Legyen θ egy s dimenziós paramétervektor, valamint legyen $\hat{\theta}$ a θ paramétervektor ML-becslése, és legyen $\hat{p}_j = p_j(\hat{\theta})$.

$H_0 : P(X_i = x_j) = \hat{p}_j \quad j = 1, \dots, r$
$H_1 : \exists$ *legalább egy* j *melyre* $P(X_i = x_j) \neq \hat{p}_j$

Próbastatisztika: $T_n = \sum_{j=1}^{r} \frac{(\nu_j - n\hat{p}_j)^2}{n\hat{p}_j} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-s-1}$ *Kritikus tartomány:* $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-s-1,1-\alpha}\}$

Megjegyzés: Mivel a próba aszimptotikus, vigyáznunk kell arra, hogy a minta elemszáma elég nagy legyen. Konyhaszabályként meg szokás követelni, hogy az ún. elméleti gyakoriság (np_j) legalább 5 legyen. Ha ez nem teljesül, akkor a kis várt gyakoriságokkal rendelkező eseményeket összevonjuk.

Függetlenségvizsgálat

n független megfigyelést két szempont szerint osztályozunk, az 1. szempont szerint r osztály, míg a 2. szempont szerint s osztály van. Annak a valószínűsége, hogy egy megfigyelést az 1. szempont szerint az i-edik, a második szerint pedig a j-edik osztályba sorolunk, p_{ij}. Az ilyen tulajdonságú megfigyelések számát pedig ν_{ij}-vel jelöljük. Az osztályozási eljárás eredményét ún. kontingenciatábla formájában szokás megadni:

		2. szempont					
		1	...	*j*	...	*s*	*Sorösszegek*
1. szempont	*1*	ν_{11}	...	ν_{1j}	...	ν_{1s}	$\nu_{1\bullet}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	i	ν_{i1}	...	ν_{ij}	...	ν_{is}	$\nu_{i\bullet}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	r	ν_{r1}	...	ν_{rj}	...	ν_{rs}	$\nu_{r\bullet}$
Oszlopösszegek		$\nu_{\bullet 1}$	...	$\nu_{\bullet j}$	...	$\nu_{\bullet s}$	n

ν_{ij} *= megfigyelések gyakorisága az* (i, j) *osztályban*

$\nu_{i\bullet} = \sum_{j=1}^{s} \nu_{ij} \quad \nu_{\bullet j} = \sum_{i=1}^{r} \nu_{ij}$

Hasonlóan $p_{i\bullet}$ ill. $p_{\bullet j}$ a marginális eloszlást jelölik, tehát a $[p_{ij}]$ mátrix sor-, illetve oszlopösszegei: $p_{i\bullet} = \sum_{j=1}^{s} p_{ij} \quad p_{\bullet j} = \sum_{i=1}^{r} p_{ij}$

H_0: a két szempont független egymástól, azaz $p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \;\; 1 \leq i \leq r, \, 1 \leq j \leq s$

H_1: a két szempont nem független, azaz $p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j}$ legalább egy (i,j) párra

Próbastatisztika: $T_n = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\nu_{ij} - \frac{\nu_{i\bullet}\nu_{\bullet j}}{n}\right)^2}{\frac{\nu_{i\bullet}\nu_{\bullet j}}{n}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{(r-1)(s-1)}$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$

Megjegyzés: Ha $r = s = 2$, akkor a próbastatisztika a következőképpen leegyszerűsödik:
$T_n = n \cdot \frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1\bullet}\nu_{2\bullet}\nu_{\bullet 1}\nu_{\bullet 2}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_1$.

Homogenitásvizsgálat

Van két független mintánk (adatsorunk) az egyikben n, a másikban m megfigyeléssel. Valamilyen szempont szerint r, páronként diszjunkt osztályba soroljuk a megfigyeléseket. Az i-edik osztály valószínűsége p_i az 1. minta és q_i a 2. minta esetén $(i = 1, 2, ..., r)$. Legyenek az egyes osztályok gyakoriságai $\nu_1, \ldots, \nu_r$ az 1. minta és $\mu_1, \ldots, \mu_r$ a 2. minta esetén.

Osztályok	*1*	*2*	*...*	*r*	*Összesen*
1. minta					
Gyakoriságok	ν_1	ν_2	...	ν_r	n
Valószínűségek	p_1	p_2	...	p_r	1
2. minta					
Gyakoriságok	μ_1	μ_2	...	μ_r	m
Valószínűségek	q_1	q_2	...	q_r	1

Azt vizsgáljuk, hogy a két minta ugyanolyan eloszlás szerint sorolódik-e be az egyes osztályokba:

H_0 : a két eloszlás megegyezik, azaz $p_i = q_i$ $i = 1, \ldots r$
H_1 : a két eloszlás nem megegyezik meg, azaz $\exists$ legalább egy i, hogy $p_i \neq q_i$

Próbastatisztika: $T_{n,m} = nm \sum_{i=1}^{r} \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$ *Kritikus tartomány:* $\mathcal{X}_k = \{\mathbf{x} : T_{n,m}(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Korrelácó:

Legyenek $X_1, \ldots, X_n$ és $Y_1, \ldots, Y_n$ n elemű minták. A korreláció becslése a minták alapján:

Tapasztalati korrelációs együttható: $r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$

Egyszerű lineáris regresszió:

Adott $(x_1, y_1), \ldots, (x_n, y_n)$ számpárokra szeretnénk egyenest illeszteni.

Modell: $y_i = a + bx_i + \varepsilon_i$, *ahol* $E\varepsilon_i = 0$ *és* $D^2\varepsilon_i = \sigma^2 < \infty \quad (i = 1, \ldots, n)$

Cél: a és b becslése

Módszer: legkisebb négyzetek: $\min \sum_{i=1}^{n} (y_i - (a + bx_i))^2$

Megoldás: $\hat{b} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$, *ennek szórásnégyzete:* $D^2(\hat{b}) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}$

$\hat{a} = \overline{y} - \hat{b}\overline{x}$, *ennek szórásnégyzete:* $D^2(\hat{a}) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right)$

Reziduálisok: $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i) \quad (i = 1, \ldots, n)$

Reziduális szórásnégyzet becslése: $\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2}$

Determinációs együttható: $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} = r^2$ *Az R^2 mutatja meg, hogy X változékonysága mennyire magyarázza Y változékonyságát. Értéke 0 és 1 között lehet, minél nagyobb, annál jobban teljesít a model, azaz annál jobban illeszkedik az egyenes.*

Feladatok

6.1. Feladat. Egy gyárban egy termék minőségét 4 elemű mintákat véve ellenőrzik, havonta 300 mintavétellel. Megszámolták, hogy a legutóbbi hónapban hányszor volt selejtes a minta, melynek eredményiet az alábbi táblázat tartalmazza:

Selejtesek száma	0	1	2	3	4
Darabszám	80	113	77	27	3

Modellezhető a mintákban levő selejtesek száma
a) $(4; 0,25)$, ill.
b) $(4; p)$ paraméterű binomiális eloszlással $(\alpha = 0,05)$? $\left(\chi^2_{3;0,95} = 7,81,\ \chi^2_{2;0,95} = 5,99\right)$

Megoldás

a) Tiszta illeszkedésvizsgálat

H_0: $X_i \sim Bin(4;\ 0,25)$.
H_1: X_i *nem ilyen eloszlású*

Vegyük észre, hogy az utólsó oszopra vonatkozóan $np_5 = 300 \cdot \binom{4}{4} \cdot 0,25^4 \cdot 0,75^0 = 1,2 < 5$*, így az utolsó két oszlopban levő eseményeket vonjuk össze. A módosított táblázat a következő:*

Selejtesek száma	*0*	*1*	*2*	*3 vagy 4*
Darabszám	*80*	*113*	*77*	*30*

$r = 4$, $n = 300$

Határozzuk meg az egyes selejtes termékekre vonatkozó valószínűségeket, illetve ezek alapján gyakoriságokat:

$p_1 = P(X_j = 0) = \binom{4}{0} \cdot 0,25^0 \cdot 0,75^4 = 0,3164 \Rightarrow n \cdot p_1 = 300 \cdot 0,3164 = 94,9$
$p_2 = P(X_j = 1) = \binom{4}{1} \cdot 0,25^1 \cdot 0,75^3 = 0,4219 \Rightarrow n \cdot p_2 = 300 \cdot 0,4219 = 126,6$
$p_3 = P(X_j = 2) = \binom{4}{2} \cdot 0,25^2 \cdot 0,75^2 = 0,2109 \Rightarrow n \cdot p_3 = 300 \cdot 0,2109 = 63,3$
$p_4 = P(X_j \geq 3) = 1 - p_1 - p_2 - p_3 = 0,0508 \Rightarrow n \cdot p_4 = 300 - 94,9 - 126,6 - 63,3 = 15,2$

Selejtesek száma	*0*	*1*	*2*	*3 vagy 4*
Darabszám v. gyakoriságok (ν_j)	*80*	*113*	*77*	*30*
Valószínűségek (p_j)	*0,3164*	*0,4219*	*0,2109*	*0,0508*
Elméleti gyakoriságok (np_j)	*94,9*	*126,6*	*63,3*	*15,2*

Próbastatisztika: $T_n = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$*, melynek értéke*

$\frac{(80-94,9)^2}{94,9} + \frac{(113-126,6)^2}{126,6} + \frac{(77-63,3)^2}{63,3} + \frac{(30-15,2)^2}{15,2} = 2,339 + 1,461 + 2,965 + 14,411 = 21,176$

A próbastatisztika mely értékeire utasítjuk el H_0*-t?*

Mivel a szabadsági fok $r - 1 = 3$*, így* $\chi^2_{3;0,95} = 7,81$*, azaz a kritikus tartomány* $= \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\} = \{\mathbf{x} : T_n(\mathbf{x}) > 7,81\}$*. Mivel most* $T_n = 21,176 > 7,81$*, így elutasítjuk* H_0*-t, azaz mondhatjuk, hogy a selejtes termékek száma nem* $Bin(4;\ 0,25)$ *eloszlást követ.*

A hipotézist a p-érték α*-val való összehasonlításával is tesztelhetjük:*

p-érték $= P(\chi^2_3 > 21,176) = 0,0001 < 0,05$*, így elutasítjuk* H_0*-t.*

b) Becsléses illeszkedésvizsgálat

H_0: $X_i \sim Bin(4; p)$ *valamilyen* p*-re*
H_1: X_i *nem ilyen eloszlású*

Először meg kell becsülni az ismeretlen p *paramétert ML-módszerrel. (Egy paramétert becslünk, így* $s = 1$*.) A 8. gyakorlat 6 a) feladata alapján tudjuk, hogy* $Bin(m, p)$ *eloszlású minta esetén* (m *ismert) a* p *ML-becslése* $\hat{p} = \frac{\overline{X}}{m}$*. Mivel* $\overline{x} = \frac{0 \cdot 80 + 1 \cdot 113 + 2 \cdot 77 + 3 \cdot 27 + 4 \cdot 3}{300} = \frac{360}{300} = 1,2$*, így* $\hat{p} = \frac{1,2}{4} = 0,3$*. Vegyük észre, hogy az utolsó oszopra vonatkozóan* $np_5 = 300 \cdot \binom{4}{4} \cdot 0,3^4 \cdot 0,7^0 = 2,43 < 5$*, így az utolsó két oszlopban levő eseményeket vonjuk össze. A módosított táblázat a következő:*

Selejtesek száma	*0*	*1*	*2*	*3 vagy 4*
Darabszám	*80*	*113*	*77*	*30*

$r = 4$, $n = 300$

Határozzuk meg az egyes selejtes termékekre vonatkozó valószínűségeket, illetve ezek alapján gyakoriságokat:

$\hat{p_1} = \hat{P}(X_j = 0) = \binom{4}{0} \cdot 0,3^0 \cdot 0,7^4 = 0,2401 \Rightarrow n \cdot \hat{p_1} = 300 \cdot 0,2401 = 72$
$\hat{p_2} = \hat{P}(X_j = 1) = \binom{4}{1} \cdot 0,3^1 \cdot 0,7^3 = 0,4116 \Rightarrow n \cdot \hat{p_2} = 300 \cdot 0,4116 = 123,5$
$\hat{p_3} = \hat{P}(X_j = 2) = \binom{4}{2} \cdot 0,3^2 \cdot 0,7^2 = 0,2646 \Rightarrow n \cdot \hat{p_3} = 300 \cdot 0,2646 = 79,4$

$\hat{p_4} = \hat{P}(X_j \geq 3) = 1 - \hat{p_1} - \hat{p_2} - \hat{p_3} = 0,0837 \Rightarrow n \cdot \hat{p_4} = 300 - 72 - 123,5 - 79,4 = 25,1$

Selejtesek száma	*0*	*1*	*2*	*3 vagy 4*
Darabszám v. gyakoriságok (ν_j)	*80*	*113*	*77*	*30*
Valószínűségek $(\hat{p_j})$	*0,2401*	*0,4116*	*0,2646*	*0,0837*
Elméleti gyakoriságok $(n\hat{p_j})$	*72*	*123,5*	*79,4*	*25,1*

Próbastatisztika: $T_n = \sum_{j=1}^{r} \frac{(\nu_j - n\hat{p_j})^2}{n\hat{p_j}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-s-1}$, *melynek értéke*

$\frac{(80-72)^2}{72} + \frac{(113-123,5)^2}{123,5} + \frac{(77-79,4)^2}{79,4} + \frac{(30-25,1)^2}{25,1} = 0,889 + 0,893 + 0,073 + 0,957 = 2,812$

A próbastatisztika mely értékeire utasítjuk el H_0*-t?*

Mivel 1 *paramétert becsültünk* ($s = 1$)*, a szabadsági fok* $r - s - 1 = 2$*, így* $\chi^2_{2;0,95} = 5,99$*, azaz a kritikus tartomány* $= \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-s-1,1-\alpha}\} = \{\mathbf{x} : T_n(\mathbf{x}) > 5,99\}$*. Mivel most* $T_n = 2,812 < 5,99$*, így nem utasítjuk el* H_0*-t, tehát tekinthetjük a selejtes termékek számát* $Bin(4, p)$ *eloszlásúnak.*

6.2. Feladat. Az alábbi kontingencia-táblázat mutatja, hogy egy 100 éves időszakban egy adott napon a csapadék mennyisége és az átlaghőmérséklet hogyan alakult:

Hőmérséklet | Csapadék	kevés	átlagos	sok
hűvös	15	10	5
átlagos	10	10	20
meleg	5	20	5

A cellákban az egyes esetek gyakoriságai találhatóak. $\alpha = 0,05$ mellett tekinthető-e a csapadékmennyiség és a hőmérséklet függetlennek? $(\chi^2_{4;0,95} = 9,49)$

Megoldás

Függetlenségvizsgálat

H_0: *a csapadék és a hőmérséklet függetlenek*

H_1: *nem függetlenek*

Egészítsük ki a táblázatot egy "összesen" sorral és oszloppal:

Hőmérséklet | Csapadék	*kevés*	*átlagos*	*sok*	*Összesen*
hűvös	*15*	*10*	*5*	$\nu_{1\bullet} = 30$
átlagos	*10*	*10*	*20*	$\nu_{2\bullet} = 40$
meleg	*5*	*20*	*5*	$\nu_{3\bullet} = 30$
Összesen	$\nu_{\bullet 1} = 30$	$\nu_{\bullet 2} = 40$	$\nu_{\bullet 3} = 30$	$n = 100$

A várt gyakoriságok $\hat{\nu}_{ij} = n \cdot \frac{\nu_{i\bullet}}{n} \cdot \frac{\nu_{\bullet j}}{n} = \frac{\nu_{i\bullet} \cdot \nu_{\bullet j}}{n}$ *táblázatban:*

Hőmérséklet | Csapadék	*kevés*	*átlagos*	*sok*	*Összesen*
hűvös	$\frac{30 \cdot 30}{100} = 9$	$\frac{40 \cdot 30}{100} = 12$	$\frac{30 \cdot 30}{100} = 9$	$\nu_{1\bullet} = 30$
átlagos	$\frac{30 \cdot 40}{100} = 12$	$\frac{40 \cdot 40}{100} = 16$	$\frac{30 \cdot 40}{100} = 12$	$\nu_{2\bullet} = 40$
meleg	$\frac{30 \cdot 30}{100} = 9$	$\frac{30 \cdot 40}{100} = 12$	$\frac{30 \cdot 30}{100} = 9$	$\nu_{3\bullet} = 30$
Összesen	$\nu_{\bullet 1} = 30$	$\nu_{\bullet 2} = 40$	$\nu_{\bullet 3} = 30$	$n = 100$

Próbastatisztika: $T_n = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\nu_{ij} - \frac{\nu_{i\bullet}\nu_{\bullet j}}{n}\right)^2}{\frac{\nu_{i\bullet}\nu_{\bullet j}}{n}} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{(r-1)(s-1)}$ *(*r *az oszlopok,* s *a sorok száma), melynek értéke*

$\frac{(15-9)^2}{9} + \frac{(10-12)^2}{12} + \frac{(5-9)^2}{9} + \frac{(10-12)^2}{12} + \frac{(10-16)^2}{16} + \frac{(20-12)^2}{12} + \frac{(5-9)^2}{9} + \frac{(20-12)^2}{12} + \frac{(5-9)^2}{9} =$

$= 4 + 0,333 + 1,778 + 0,333 + 2,25 + 5,333 + 1,778 + 5,333 + 1,778 = 22,916$

A próbastatisztika mely értékeire utasítjuk el H_0*-t?*

Mivel a szabadsági fok $(r-1)(s-1) = 2 \cdot 2 = 4$*, így* $\chi^2_{4;0,95} = 9,49$*, azaz a kritikus tartomány* $= \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{(r-1)(s-1),1-\alpha}\} = \{\mathbf{x} : T_n(\mathbf{x}) > 9,49\}$*. Mivel most* $T_n = 22,916 > 9,49$*, így elutasítjuk* H_0*-t, tehát állíthatjuk, hogy a csapadék és a hőmérséklet nem függetlenek.*

A hipotézist a p-érték α*-val való összehasonlításával is tesztelhetjük:*

p-érték $= P(\chi^2_4 > 22,916) = 0,0001 < 0,05$*, így elutasítjuk* H_0*-t.*

6.3. Feladat. Két dobókockával dobva az alábbi gyakoriságokat figyeltük meg:

Dobások	1	2	3	4	5	6
1. kocka	27	24	26	23	18	32
2. kocka	18	12	15	21	14	20

$\alpha = 0,05$ mellett döntsünk arról, hogy tekinthető-e a két eloszlás azonosnak! $\left(\chi^2_{5;0,95} = 11,1\right)$

Megoldás

Homogenitásvizsgálat

H_0 *a két eloszlás megegyezik*
H_1 *a két eloszlás nem egyezik meg*

Egészítsük ki a táblázatot egy „összesen” oszlopplal:

Dobások	*1*	*2*	*3*	*4*	*5*	*6*	*Összesen*
1. kocka (ν_i)	*27*	*24*	*26*	*23*	*18*	*32*	$n = 150$
2. kocka (μ_i)	*18*	*12*	*15*	*21*	*14*	*20*	$m = 100$

$r = 6$

Próbastatisztika: $T_{n,m} = nm \sum_{i=1}^{r} \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i} \overset{H_0 \text{ esetén}}{\sim} \chi^2_{r-1}$ *melynek értéke*

$$T_{150,100} = 150 \cdot 100 \left(\frac{\left(\frac{27}{150} - \frac{18}{100}\right)^2}{27+18} + \frac{\left(\frac{24}{150} - \frac{12}{100}\right)^2}{24+12} + \frac{\left(\frac{26}{150} - \frac{15}{100}\right)^2}{26+15} + \frac{\left(\frac{23}{150} - \frac{21}{100}\right)^2}{23+21} + \frac{\left(\frac{18}{150} - \frac{14}{100}\right)^2}{18+14} + \frac{\left(\frac{32}{150} - \frac{20}{100}\right)^2}{32+20} \right) =$$
$$= 0 + 0,67 + 0,20 + 1,09 + 0,19 + 0,05 = 2,2$$

A próbastatisztika mely értékeire utasítjuk el H_0*-t?*

Mivel a szabadsági fok $r - 1 = 5$*, így* $\chi^2_{5;0,95} = 11,1$*, azaz a kritikus tartomány* $= \{\mathbf{x} : T_{n,m}(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\} =$
$= \{\mathbf{x} : T_{n,m}(\mathbf{x}) > 11,1\}$*. Mivel most* $T_{150,100} = 2,2 < 11,1$*, így nem utasítjuk el* H_0*-t, ami nem mutat elentmondást a két eloszlás azonosságával.*