Matematika B4
I. gyakorlat

2006. február 16.

1. Egy-dimenziós adatrendszerek

Van valamilyen adatrendszer (számsorozat), amelyről szeretnénk kiszámolni bizonyos dolgokat. Az egyes értékeket jelöljük z_i-vel, a számukat pedig n-nel. Nézzünk egy konkrét példát: 1, 3, 2, 1, 4. A valószínűségszámításban és statisztikában legfontosabb mennyiségek a következők:

1. Átlag:
   Az adatok átlaga: $\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$.
   A példánkban: $\overline{z} = \frac{1}{5}(1+3+2+1+4) = 2.2$

2. Átlagos abszolút eltérés az átlagtól:
   $\frac{1}{n}\sum_{i=1}^{n} |z_i - \overline{z}|$
   A példánkban: $\frac{1}{5}(|1-2.2| + |3-2.2| + |2-2.2| + |1-2.2| + |4-2.2|) = 1.04$

3. Momentumok:
   Az adatrendszer k-adik momentuma: $\frac{1}{n}\sum_{i=1}^{n} {z_i}^k$. Minden adatrendszerre a nulladik momentum 1, az első momentum pedig megegyezik az átlaggal.
   A fenti példában a második momentum 6.2, a harmadik momentum 20.2.

4. Valamely c pontra vonatkozó momentumok
   $\frac{1}{n}\sum_{i=1}^{n}(z_i - c)^k$

5. Szórásnégyzet:
   Az adatok átlagos négyzetes eltérése az átlagtól. $\sigma^2 = \frac{1}{n}\sum(z_i - \overline{z})^2 = (\frac{1}{n}\sum_{i=1}^{n} {z_i}^2) - \overline{z}^2$.
   A szórásnégyzet nem más, mint az átlagra vonatkozó második momentum.
   A példánkban: $\sigma^2 = \frac{1}{5}(1^2+3^2+2^2+1^2+4^2) - 2.2^2 = 1.36$

6. Szórás:
   A szórásnégyzet gyöke. $\sigma = \sqrt{\sigma^2}$.
   Esetünkben $\sigma = \sqrt{1.36} = 1.17$

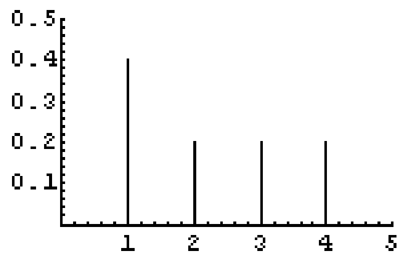
7. Medián:
   Rendezzük nagyság szerint sorba az adatokat. Ha n páratlan, akkor a medián a középső elem, ha páros, akkor a két középső elem átlaga. A mediántól nagyobb és kisebb elemből is ugyanannyi van.
   A fenti példában a medián a 2.

8. Relatív gyakoriság:

   a) Egy elem relatív gyakoriságát úgy kapjuk meg, hogy előfordulási számát elosztjuk az adatrendszer összes elemeinek számával.
      Esetünkben az 1 relatív gyakorisága $\frac{2}{5}$, a 2, 3, 4 relatív gyakorisága $\frac{1}{5}$.

   b) Szokás az X-tengelyt intervallumokra is felosztani, ilyenkor egy intervallum relatív gyakoriságát úgy kapjuk meg, hogy az intervallumba eső adatok számát osztjuk az összes adat számával.
      Például a $[0, 0.25]$ intervallum relatív gyakorisága $\frac{3}{5}$, a $(2.5, 5]$ relatív gyakorisága $\frac{2}{5}$.

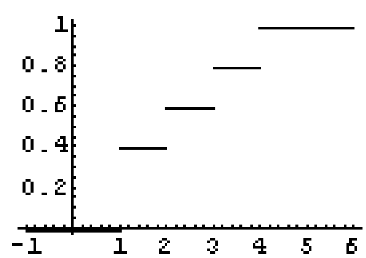
9. Szemléltetés pálcikákkal
   A relatív gyakoriságok grafikonja, ahol az X-tengelyen ábrázoljuk az előforduló értékeket, az Y-tengelyen pedig a hozzájuk tartozó relatív gyakoriságának megfelelő hosszúságú pálcikát.



1. ábra.

10. Hisztogram:
    Ha az X-tengelyt diszjunkt intervallumokra bontjuk, akkor az intervallumok fölé rajzolhatunk téglalapokat úgy, hogy területük az intervallum relatív gyakoriságával egyezzen meg.

11. Eloszlásfüggvény:
    Tetszőleges x érték esetén az $F(x)$ definíciója: $F(x) = \frac{\text{x-nél szigorúan kisebb elemek száma}}{n}$.
    Az eloszlásfüggvény grafikonja mindig egy olyan lépcsős függvény, amely vízszintes darabokból és ugrásokból áll. Ugrások ott lesznek, ahol az adatok vannak, az ugrás nagysága az adott relatív gyakorisággal egyenlő.
    A fenti példa eloszlásfüggvénye:



2. ábra.

Vegyünk 10 db egész számot 1 és 6 között. Ezek lehetnek kitalált számok, indexbeli jegyek, de használhatunk dobókockát is ezek generálásához. Ezek a számok alkotják az 1. adatrendszert.
A 2. adatrendszer elemei legyenek az 1. adatrendszer elemeinek kétszerese.
A 3. adatrendszer elemeit úgy kapjuk meg, hogy az 1. adatrendszer elemeihez hozzáadunk 10-et.
A 4. adatrendszer elemeit úgy kapjuk meg, hogy az 1. adatrendszer elemeit emeljük négyzetre.
A 5. adatrendszer elemeit úgy kapjuk meg, hogy az 1. adatrendszer elemeinek vesszük a reciprokát. Például:

1. adatrendszer	3	1	6	2	3	3	3	1	4	5
2. adatrendszer	6	2	12	4	6	6	6	2	8	10
3. adatrendszer	13	11	16	12	13	13	13	11	14	15
4. adatrendszer	9	1	36	4	9	9	9	1	16	25
5. adatrendszer	$1/3$	1	$1/6$	$1/2$	$1/3$	$1/3$	$1/3$	1	$1/4$	$1/5$

2. Feladatok

1. Számoljuk ki az így kapott adatsorok átlagát, szórását, második és harmadik momentumát. Mit állapíthatunk meg az átlagról és a szórásról? Hogyan változnak az egyes sorozatok esetén? Igazoljuk az észrevételeket általá-

nos esetre!
Szemléltessük az adatrendszert

a) "dróton"!
b) "ablakpárkányon"!

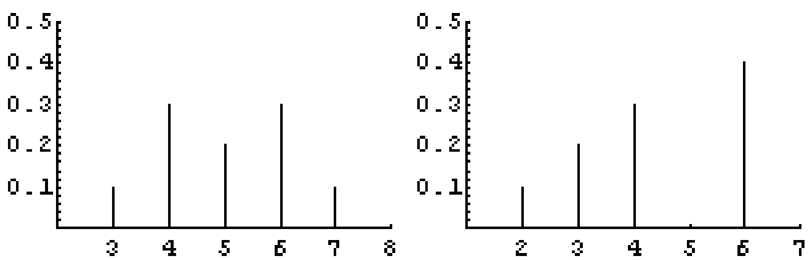
2. A fenti sorozatok relatív gyakoriságait

    a) szemléltessük pálcikákkal
    b) rajzoljuk fel az eloszlásfüggvényeket
    c) rajzoljuk fel a hisztogramjaikat
    d) határozzuk meg a mediánt.

3. Bizonyítsuk be, hogy minden adatrendszerben az átlagtól való átlagos (előjeles) különbség 0.

4. Mit jelent az, hogy ha a szórás 0?

5. Az alábbi két ábrán pálcikákkal ábrázoltuk a relatív gyakoriságot. Számoljuk ki a mediánt és átlagot mindkét esetben!



3. ábra.

6. Készítsünk olyan adatsort, ahol

    a) a medián megegyezik az átlaggal!
    b) a medián nagyobb az átlagtól!
    c) a medián kisebb az átlagtól!

7. Lehet-e nagyobb az átlagos abszolút eltérés a szórástól? És fordítva?

8. 1000 adatot gyűjtünk össze. Az első 600 szám átlaga 9, a maradék 400 átlaga 5. Mennyi a teljes adatrendszer átlaga?

9. Vegyük a következő adatrendszert: $5, 5, 10, 25, 35$.

    a) Mi az a szám, amit minden elemhez hozzáadva 20 lesz az átlag?
    b) Mennyivel kell megszorozni a számokat, hogy az új adatrendszer szórása 6 legyen?
    c) Centralizáljuk az adatrendszert, azaz toljuk el az elemeket úgy, hogy az új adatrendszer átlaga 0 legyen.
    d) Standarizáljuk a fenti számsorozatot, azaz miután eltoltuk úgy, hogy 0 legyen az átlag szorozzuk be az elemeket egy számmal, hogy az új adatrenszer szórása 1 legyen.

10. Adottak a következő számok: $x_1 = 3$, $x_2 = 7$, $x_3 = 21$. Keressük meg, hogy mely c-re lesz a $\sum_{i=1}^{3} |z_i - c|$ kifejezés minimális! Mi a megoldás általános esetben?

11. Oldjuk meg az elő feladatot abszolút érték helyett négyzettel, azaz mely c érték mellett lesz a $\sum_{i=1}^{3} (x_1 - c)^2$ minimális? Mi az általános megoldás?

12. Generáljunk 20 db véletlen számot a zsebszámológép segítségével (az RND gomb használatával). Osszuk fel

    a) 2 egyenlő részre az intervallumot.
    b) 4 egyenlő részre az intervallumot.
    c) 10 egyenlő részre az intervallumot.

    Ábrázoljuk az adatrendszert dróton, ablakpárkányon, számoljuk ki az eloszlásfüggvényét. Ábrázoljuk a relatív gyakoriságokat pálcikákkal és hisztogrammal is.

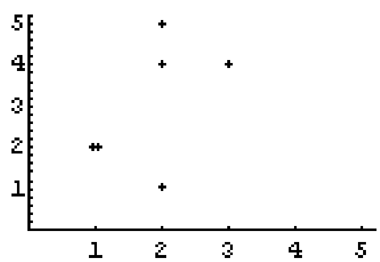
13. Legyen RND_1 és RND_2 két számológéppel generált szám. A fenti feladatot számoljuk végig úgy is, hogy minden egyes adatot $\frac{RND_1+RND_2}{2}$ utasítással definiálunk.

14. Egy adatrendszerben csak két féle érték szerepel: 0 és 1. Tudjuk, hogy az átlag 0.3, adjuk meg a relatív gyakoriságokat!

3. Többdimenziós adatrendszerek

Kétdimenziós adatrendszerek esetén külön számolhatunk átlagot és szórást koordinátánként, de felmerül, hogy mi a kapcsolat ezen koordináták között. u_i-vel és v_i-vel jelölöm az adatok koordinátáit, például u_i az i. hallgató HáRe jegye, v_i pedig a B4 jegye.

HáRe	1	2	2	3	2	1
B4	2	4	5	4	1	2



4. ábra.

1. Kovariancia:

$$c = \frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v})$$

   A fenti példában a B4 és HáRe jegyek kovarianciája pozitív, 0.5. Ez intuitíve azt jelenti, hogy a nagyobb jobb HáRe jegyhez "általában" jobb B4 jegy is párosul.

2. Korrelációs együttható:

$$r = \frac{\text{c}}{\sigma_U \cdot \sigma_V}$$

   Esetünkben a korrelációs együttható: 0.51, mivel a HáRe esetén a szórás 0.69, B4 esetén 1.41.

4. Feladatok

Adott a következő táblázat:

Név	Hal Otmár	K. Bea	Lepi Pál	Saaanyi
Magasság (U)	182	165	171	195
Súly (V)	87	62	95	117

15. Számoljuk ki a magasság és súly kovarianciáját és korrelációs együtthatóját.

16. Keressük meg azt a(z)

    a) origón átmenő

    b) konstans

    c) általános

    $v = l(u)$ egyenest, amelyre a $\sum_{i=1}^{n}(v_i - l(u_i))^2$ érték minimális lesz! (Ezek az egyenesek lesznek a négyzetes értelemben vett legjobb becslések U ismeretében V-re, ha figyelembe vesszük a feltételeket.)

17. Oldjuk meg az előző feladatot úgy is, hogy U és V szerepét felcseréljük (azaz olyan, mintha V ismeretében szeretnénk becsülni U-et)!

18. Készítsünk 3 adatból álló adatrendszert, amelyben a korrelációs együttható értéke:

    a) $r = 0$

    b) $r = 1$

    c) $r = -1$

    d) $r = 0.5$

    e) $r = -0.5$

    f) $r = 0.8$

    g) $r = -0.8$

19. Ellenőrizzük az eddigi példákon, hogy $c^2 \leq \sigma_U \sigma_V$ és vele együtt $|r| \leq 1$ egyenlőtlenség mindig teljesül.
    Próbáljuk meg belátni, hogy $c^2 \leq \sigma_U \sigma_V$ mindig fennáll.
    Ötlet: A tétel Cauchy-Schwarz egyenlőtlenség felhasználásával ($\sum_{i=1}^{n}(a_i b_i)^2 \leq \sum_{i=1}^{n} {a_i}^2 \sum_{i=1}^{n} {b_i}^2$) könnyű, a Cauchy-Scwarz pedig egy kis számolás után kijön a $0 \leq \sum(a_i - t \cdot b_i)^2$ kibontásából és diszkriminánsa vizsgálatából.