7. gyakorlat – Felderítő adatelemzés és kísérlettervezés – Megoldások

1. feladat

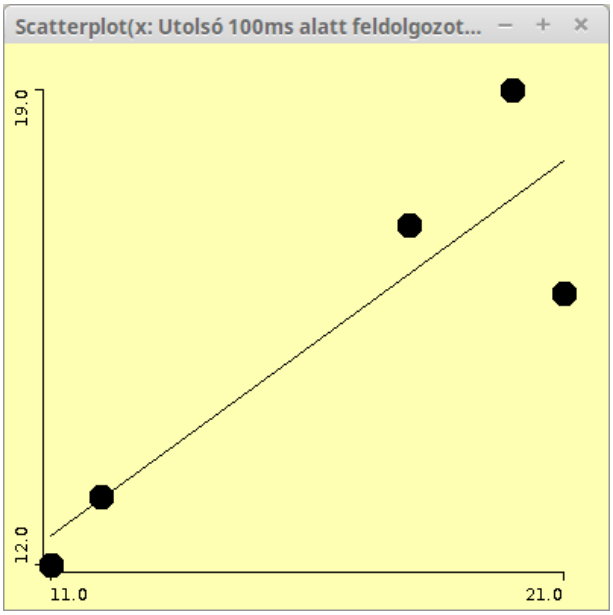
Egy szerveren az alábbi teljesítményjellemzőket mértük:

Mintavétel időpontja [ms]	500	600	700	800	900
Utolsó 100 ms alatt feldolgozott kérések száma [darab]	11	12	21	18	20
Utolsó 100 ms átlagos kiszolgálási ideje [ms]	15	20	21	25	27
Utolsó 100 ms CPU kihasználtság [%]	12	13	16	17	19
Utolsó 100 ms HDD I/O kihasználtság [%]	55	63	87	61	73

a. Ábrázoljuk a feldolgozott kérések számát és a CPU kihasználtságot pontfelhő (scatterplot) diagramon! Értelmezzük a diagramot!
b. Az első mintavétel idején mekkora az átbocsátási ráta értéke? Az 5 mintavétel alapján mekkora az átbocsátási ráta tapasztalati átlaga és mediánja? Mi tartozik a 40%-os kvantilisbe?

Megoldás

a. Két klaszter (csoportosulás) látszik, nagyjából pozitív a korreláció (kb. arányosak az adatok), de nem direkten monoton (valami más is befolyásolhatja az adatokat, ezért ingadozik). Értelmezve a látottakat a CPU átlagos kihasználtsága a feldolgozott kérések számával nő. A bal alsó csoport kisebb terhelésű pillanatokat tartalmaz, míg a jobb felső nagyobbakat. Ez a megfigyelés adott esetben jó alapja lehet a terhelés vizsgálatának (pl. az egyes csoportokhoz tartozó pontok időben is közel vannak-e egymáshoz).



b. A mintavételi időkből látszik, hogy két mintavétel között 100 ms telik el. Ebből

$$X_1 = \frac{k_1}{\Delta t} = \frac{11 \text{ kérés}}{100 \text{ ms}} = \frac{11 \text{ kérés}}{100 \text{ ms}} \left[\frac{1000 \text{ ms}}{1 \text{ s}}\right] = 110 \ \frac{\text{kérés}}{\text{s}}. \tag{1}$$

A tapasztalati átlag kiszámítása történhet a másik négy átbocsátás kiszámításával és átlagolással, vagy a következő módon (kihasználva, hogy Δt végig 100ms):

$$\bar{k} = \frac{\sum_{i=1}^{n} k_i}{n} = \frac{11 + 12 + 21 + 18 + 20}{5} = 16{,}4 \tag{2}$$

Ebből az átlagos átbocsátás $\bar{X} = \frac{\bar{k}}{\Delta t} = \frac{16{,}4}{0{,}1} = 164 \ \frac{\text{kérés}}{\text{s}}$

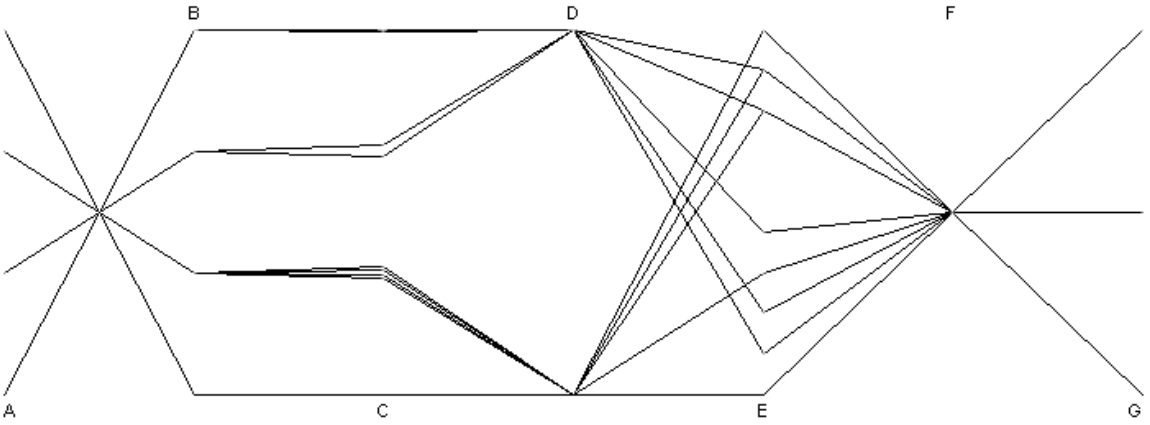
Az elemek sorba állítva $11, 12, 18, 20, 21$, ebből rögtön látszik, hogy a medián 18.

A p kvantilis definíció szerint az a szám, amelynél az elemek p-ed része kisebb vagy egyenlő[1]. A p kvantilisba azok az elemek tartoznak, amelyek kisebb vagy egyenlők a p kvantilisnál. A kvantilis speciálisabb változata a percentilis, amely egész százalékokkal dolgozik, valamint a kvartilis, amely "negyedeli" az adatot. Pl. a 35. percentilis a 35%-os kvantilisnak felel meg (a kvantilis lehetne pl. 35,7% is!), a második kvartilis pedig az 50%-os kvantilisnak.

Itt az elemek legkisebb 40%-a a 11 és a 12, ezért a 40%-os kvantilis értéke a 12 lesz, és a 11, illetve 12 elemek tartoznak bele.

2. feladat

Alternatív algoritmusok teljesítményjellemzőit mérjük az A–G algoritmusok többszöri futtatásával. Az alábbi ábrán ezen algoritmusok átlagos memóriaigényét hasonlítjuk össze 10 különböző benchmark (töröttvonallal ábrázolva) segítségével.[2]



a. Hogy hívják ezt a diagramot?
b. Mi az F algoritmus jellegzetessége?
c. Hogyan viszonyul egymáshoz az A és B algoritmusok memóriaigénye?
d. Az F algoritmuson kívül mely algoritmus(ok) viselkedése tér el a többitől?
e. Elegendő információt szolgáltat-e ez a diagram ahhoz, hogy válasszunk az algoritmusok között?

Megoldás

a. Párhuzamos koordináták diagram (*parallel coordinates plot*).
b. Konstans memóriával dolgozik, mivel minden benchmarkhoz tartozó vonal egy pontban metszi az F-hez tartozó tengelyt. A párhuzamos koordináták jellegzetességei miatt erről az ábráról nem leolvasható, hogy pontosan mennyivel.
c. Fordítottan arányos, mert a két tengely között futó vonalak egy pontban metszik egymást. Ha a vonalak meghosszabbításai a két tengelyen kívül metszik egymást, vagy teljesen párhuzamosak, akkor egyenes arányosságot állapíthatnánk meg (mint pl. B és C között).
d. Az A algoritmus memóriafogyasztása kb. fordítottan arányos a B, C, D algoritmusokkal. Az E és a G algoritmusok futásidejének elemzéséhez interaktív elemzésre lenne szükségünk (pl. Mondrian), mivel most nem látszik, hogy (pl. az F-beli kereszteződés után) melyik szakasz melyik benchmarkhoz tartozik.
e. Általában nem, pl. a futásidőt nem mutatja – ezt pl. egy scatterplottal lehetne megvizsgálni.

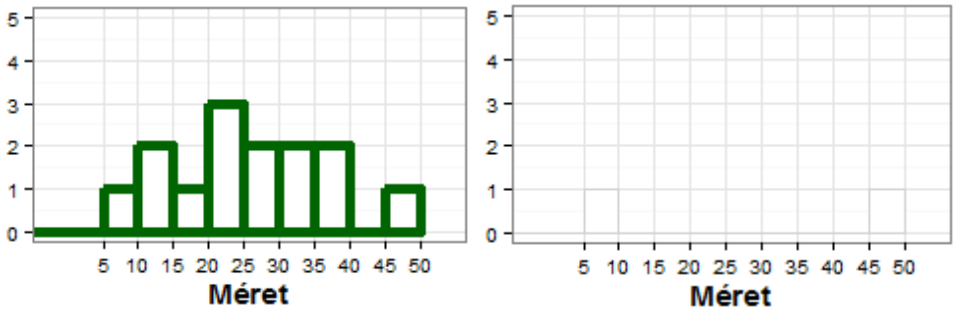
3. feladat

Online képgalériánkban a felhasználók keresés alapján megjeleníthetnek a keresőkifejezésre illeszkedő képeket.

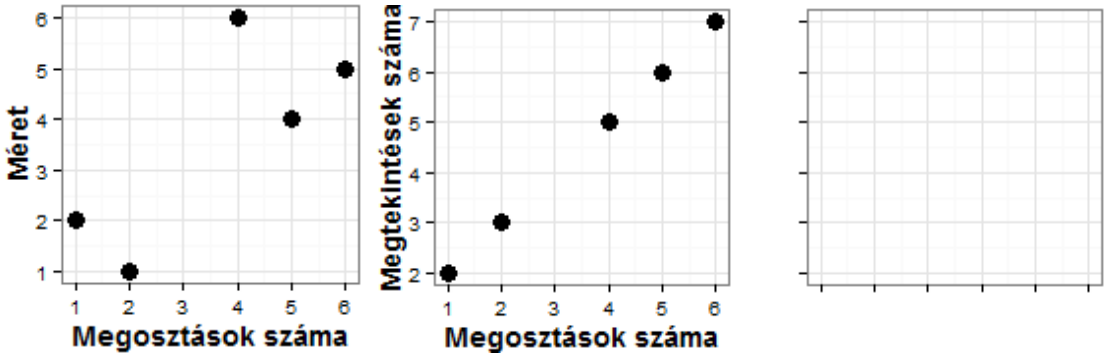
[1] A matematikai statisztika elemei

[2] Egy népszerű benchmark halmaz pl. a http://benchmarksgame.alioth.debian.org/ címen elérhető *The Computer Language Benchmarks Game*.

a. Az oldalt látható hisztogramon ábrázoltuk az albumok méretének eloszlását. Mivel a tárhely hatékony szervezéséhez elég azt tudnunk, hogy hány 10 alatti, 10 és 20 közötti stb. képet tartalmazó albumunk van, az alábbihoz képest kétszeres oszlopszélességű hisztogramot szeretnénk. Rajzoljuk meg az ábrát!



b. Pont-pont diagramon (scatterploton) ábrázoltuk 5 kiválasztott album méretét illetve megtekintési számát a megosztási számmal összehasonlításban. Igaz-e, hogy minél nagyobb az album, annál többen tekintik meg? Válaszolja meg a kérdést egy harmadik pont-pont diagramon, amely a megosztási számot a méret függvényében ábrázolja!



c. Az albumok jellemző népszerűségét szeretnénk meghatározni, emiatt a pont-pont diagram alapján kiszámoltuk a megtekintési számok átlagát és mediánját. Általánosságban megtehető-e ez egy pont-pont diagram alapján? Mennyivel változnak ezen középértékek, ha feltöltünk egy új albumot, amelyet 40-en tekintenek meg?

Megoldás

TODO

4. feladat

Infrastruktúránk méretezését megnehezíti, hogy egy adott feladattípus végrehajtási ideje a körülmények függvényében ingadozik, például lapozás, memória szemétgyűjtés, memória cache találatok stb. változékonysága folytán. Ezért összeállítottunk egy valós munkaterhelést jól jellemző benchmarkot, és ennek többszöri lefuttatása során a futási időket átlagolva szeretnénk meghatározni a rendszer átlagos teljesítményét.

a. Az első tíz futtatás eredményei: 37 s, 34 s, 35 s, 39 s, 57 s, 41 s, 36 s, 35 s, 61 s, 35 s. Mennyi ez alapján a rövid kísérlet alapján a tapasztalati átlag és tapasztalati szórás?
b. Nagyobb léptékben futtatva a kísérletet, a benchmark 10 000 futtatása átlagban 44,3 másodpercig tartott, 11,6 másodperc tapasztalati szórással. Mennyire lehetünk biztosak a kapott eredmény pontosságában?

Emlékeztető: a σ szórású normális eloszlás 68%-os konfidenciaintervalluma 1σ, a 95%-os 2σ, a 99,7%-os 3σ sugarú.

Megoldás

a. Vajon mitől lehetnek ilyenek az értékek?

A tapasztalati átlag:

$$m = \frac{37 + 34 + 35 + 39 + 57 + 41 + 36 + 35 + 61 + 35}{10} = 41 \text{ [s]} \tag{3}$$

Tapasztalati átlagtól eltérések és eltérések négyzetei:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	
x_i	37	34	35	39	57	41	36	35	61	35	[s]
$x_i - m$	−4	−7	−6	−2	+16	0	−5	−6	+20	−6	[s]
$(x_i - m)^2$	16	49	36	4	256	0	25	36	400	36	[s^2]

Tapasztalati átlagtól eltérések: mennyi ezeknek az összege/átlaga? Miért 0?
Tapasztalati átlagtól eltérések négyzete: jól érezhető, hogy a nagy eltérést jobban büntetjük, összes négyzetes eltérés 858 s^2. Innen eltérések négyzetes közepe (átlagos négyzetes eltérés gyöke):

$$s = \sqrt{\frac{(x_1 - E)^2 + ... + (x_t - E)^2}{t}} = \sqrt{\frac{858 \text{ s}^2}{10}} \approx 9{,}26 \text{ s} \tag{4}$$

Ez lenne a sokaság szórása, ha ez a 10 adatpont lenne a teljes sokaság. Mivel ez a 10 adatpont csak a sokaságból vett minta, valójában $(t-1)$-et (jelen esetben 9-et) kell a nevezőbe írni, hogy az úgynevezett korrigált szórást kapjuk:

$$s^* = \sqrt{\frac{(x_1 - E)^2 + ... + (x_t - E)^2}{t - 1}} = \sqrt{\frac{858 \text{ s}^2}{9}} \approx 9{,}76 \text{ s} \tag{5}$$

Ez utóbbi már valóban az eredeti valószínűségi változó szórását közelíti (ha nagyon sokszor ismételnénk meg ezt a 10 hosszú kísérletsorozatot, a tapasztalati szórás várható értéke a szórás lenne).
b. Mivel jóval több mint 100 megfigyelésből állt a kísérlet, a tapasztalati átlag jó közelítéssel normális eloszlással szór a tényleges várható érték körül. Ennek a Gauss-haranggörbének a szórása (az ismeretlen valószínűség változó szórását a kísérletből adódó tapasztalati szórással helyettesítve):

$$\sigma = \frac{s}{\sqrt{t}} \approx \frac{11{,}6 \text{ s}}{\sqrt{10\ 000}} = 0{,}116 \text{ s} \tag{6}$$

Tehát azt mondhatjuk, hogy a benchmark várható végrehajtási ideje 44,3 s ± 0,1 s. A normális eloszlás konfidenciaintervallumairól tanultak alapján 99,7% konfidencia mellett jelenthetjük ki, hogy a végrehajtási idő várható értéke a [44,0 s, 44,6 s] intervallumba esik. (Mivel mi igazából nem ismerjük a tényleges szórást, csak a tapasztalati szórást, valójában a $t - 1 = 9999$ paraméterű Student-féle t-eloszlás konfidencia-intervallumait kéne használni, de az ilyen magas paraméternél nagyon erősen közelíti a normális eloszlást.)