

## 7. gyakorlat – Felderítő adatelemzés és kísérlettervezés

### 1. feladat

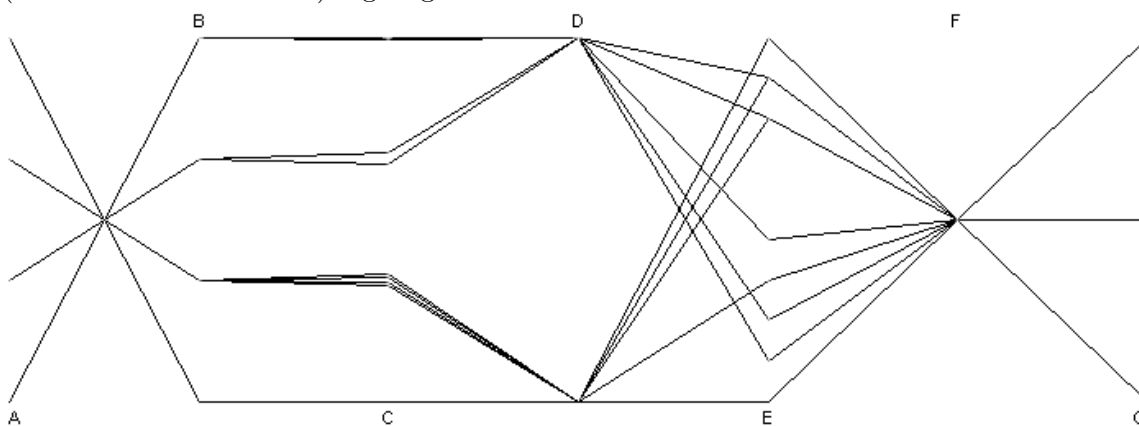
Egy szerveren az alábbi teljesítményjellemzőket mértük:

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

- Ábrázoljuk a feldolgozott kérések számát és a CPU kihasználtságot pontfelhő (scatterplot) diagramon! Értelmezzük a diagramot!
- 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?

### 2. feladat

Alternatív algoritmusok teljesítményjellemzőit mérjük az  $A$ – $G$  algoritmusok többszöri futtatásával. Az alábbi ábrán ezen algoritmusok átlagos memóriai igényét hasonlítjuk össze 10 különböző benchmark (töröttvonalal ábrázolva) segítségével.<sup>1</sup>



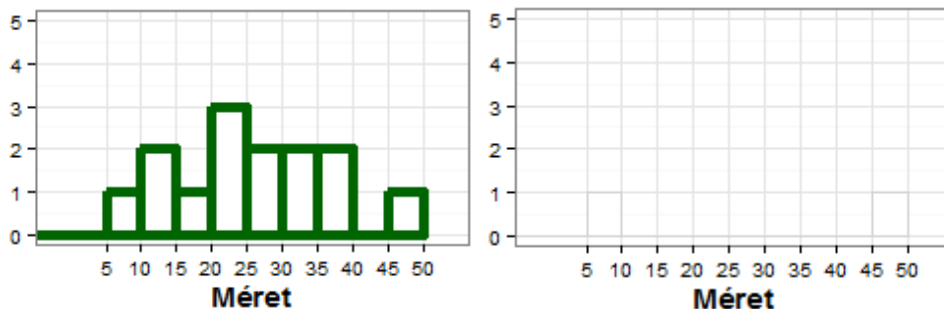
- Hogy hívják ezt a diagramot?
- Mi az  $F$  algoritmus jellegzetessége?
- Hogyan viszonyul egymáshoz az  $A$  és  $B$  algoritmusok memóriai igénye?
- Az  $F$  algoritmuson kívül mely algoritmus(ok) viselkedése tér el a többitől?
- Elegendő információt szolgáltat-e ez a diagram ahhoz, hogy válasszunk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

<sup>1</sup>Egy népszerű benchmark halmaz pl. a <http://benchmarksgame.alioth.debian.org/> címen elérhető *The Computer Language Benchmarks Game*.

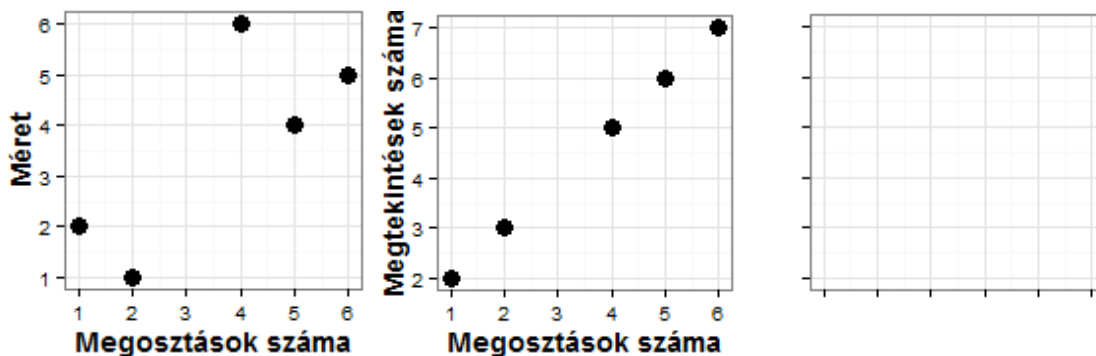
### 3. feladat

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

- 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 képeket 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?

### 4. feladat

Infrastruktúránk méretezését megnehezíti, hogy egy adott feladattípus végrehajtási ideje a körülmények függvényében ingadozik, például lapozás, memória szemétygyűjtés, 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  $\sigma$  szórású normális eloszlás 68%-os konfidenciaintervalluma  $1\sigma$ , a 95%-os  $2\sigma$ , a 99,7%-os  $3\sigma$  sugarú.