7. gyakorlat – Felderítő adatelemzés és kísérlettervezés – Megoldások

1. Kísérlet kiértékelése

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?
Megoldás
Vajon mitől lehetnek ilyenek az értékek?
A tapasztalati átlag:

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

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):

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

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:

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

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) 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?
Megoldás
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):

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

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.)

2. Kísérlettervezés

Egy modellezett folyamat átbocsátóképességére szimuláció alapján szeretnénk egy közelítő értéket és hozzá tartozó konfidencia-intervallumot meghatározni.

a) Hány szimuláció mérési eredményeiből számoljunk átlagot?
Megoldás
Ha még nincsenek közelítéseink a szórásra, az ökölszabály szerint legalább 100 megfigyelést kell végezni.

b) Az így elvégzett mérési eredmények tapasztalati közepe 500 kérés/s; a tapasztalati szórás 10%. Szeretnénk, hogy 95% konfidencia mellett egy legfeljebb 40 kérés/s széles intervallumba essen az átbocsátóképesség. Hány mérést végezzünk még?
Megoldás
A tapasztalati szórás 10 s, azaz $d = 10\% \cdot 500$ kérés/sec $= 50$ kérés/sec.
Ha a 95%-os (vagyis 2 szórásnyi sugarú) konfidenciaintervallum szélessége (sugár kétszerese) maximum 40 kérés/sec, akkor a normális eloszlás szórása maximum 10 kérés/sec lehet. (És a normális eloszlásra $t \geq 100$ megfigyelésnél már jól simul a $t-1$ szabadságfokú Student-féle t-eloszlás, amivel igazából számolni kéne.) Ez a tapasztalati átlag mint valószínűségi változó szórása a tényleges várható érték körül, értéke:

$$\sigma = \frac{s}{\sqrt{t}} \approx \frac{50 \text{ kérés/s}}{\sqrt{t}} \leq 10 \text{ kérés/s}$$

Innen $5 \leq \sqrt{t}$, azaz legalább 25 megfigyelésből kell számolni az átlagot. Persze az egész normális eloszlásos közelítés az ökölszabály szerint csak 100 megfigyelés fölött alkalmazható – de ennyi megfigyelést már el is végeztünk, tehát nem kell még többet mérni. A jelenlegi kísérlet eredménye alapján:

$$\sigma = \frac{s}{\sqrt{t}} \approx \frac{50 \text{ kérés/s}}{\sqrt{100}} = 5 \text{ kérés/s}$$

Tehát a 2σ sugarú intervallum szélessége 20 kérés/sec, így kétszeres pontosságot garantálhatunk 95% konfidencia mellett. (Avagy 40 kérés/sec széles (négy szórásnyi sugarú) intervallumot is garantálhatnánk 99,9936% konfidenciával.)

3. Szerver teljesítménye

Egy szerveren az alábbi teljesítményjellemzőket mértük:

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

a) A rendelkezésre álló adatok alapján a szerver melyik erőforrása tűnik a szűk keresztmetszetnek?
Megoldás
A HDD kihasználtsága a legnagyobb. A terhelés felskálázásával először a HDD fog telítődni.

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

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

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

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

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, tehát az átbocsátás mediánja $\frac{18}{0{,}1} = 180 \ \frac{\text{kérés}}{\text{s}}$.
A p kvantilis definíció szerint az a szám, amelynél az elemek p-ed része kisebb vagy egyenlő.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. A kapcsolódó átbocsátási ráták 110 $\frac{\text{kérés}}{\text{s}}$ és 120 $\frac{\text{kérés}}{\text{s}}$

c) Ezen 5 mérés alapján milyen becslést tudunk adni az egyszerre kiszolgálás alatt lévő kérések átlagos számára?
Megoldás
Az utolsó 100 ms alatt feldolgozott kérések számából és az átlagos kiszolgálási időből adódik. Mivel az átlagos kiszolgálási idő különböző elemszámú adathalmazokból került kiszámításra, egyszerű átlagolásuk helyett a feldolgozott kérésekkel súlyozott átlagukat kell vennünk.

$$T = \frac{\sum_{i=1}^{n} k_i t_i}{\sum_{i=1}^{n} k_i} = \frac{11 \cdot 15 + 12 \cdot 20 + 21 \cdot 21 + 18 \cdot 25 + 20 \cdot 27}{11 + 12 + 21 + 18 + 20} = 22{,}39 \text{ ms}$$

A rendszer egyensúlyi állapotban van, ezért a b) feladatban kiszámolt átlagos átbocsátással alkalmazhatjuk a Little-törvényt:

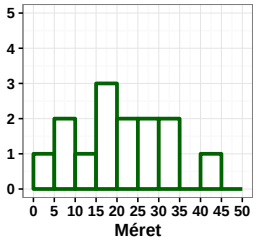
$$N = \bar{X} \cdot T = 164 \ \frac{1}{\text{s}} \cdot 22{,}39 \text{ ms} = 164 \ \frac{1}{\text{s}} \cdot 0{,}02239 \text{ s} = 3{,}67196$$

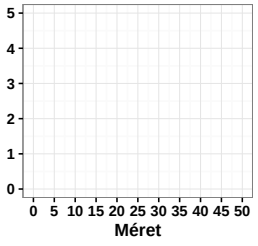
d) Vajon mely mért jellemzők között sejthető ok-okozati viszony?
Megoldás
Ahogy az ábrán is látjuk, az átbocsátás hatással van az erőforrások kihasználtságra. A szűk keresztmetszetnek számító erőforrás (HDD – ld. korábbi gyak) magas kihasználtsága meg is látszik a megnyúlt válaszidőkön.

4. Képgaléria – adatelemzés

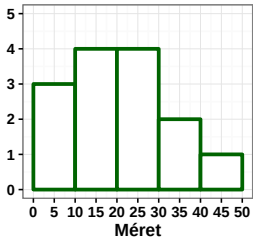
Online képgalériánkban a felhasználók keresés alapján megjeleníthetnek a keresőkifejezésre illeszkedő képeket.

a) Az alábbi 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 (szintén a 0 mérettől kezdve felszámítva az oszlopokat). Rajzoljuk meg az ábrát!

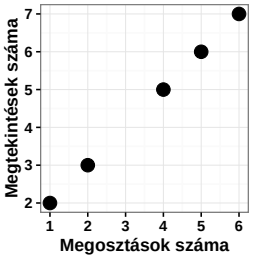


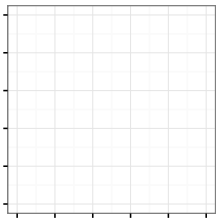


Megoldás

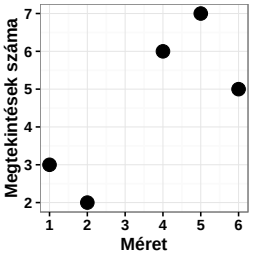


b) Pont-pont diagramon (scatterploton) ábrázoltuk 5 kiválasztott album méretét illetve megtekintési számát a megosztási számmal összehasonlításban. Igaz-e, hogy minél nagyobb az album, annál többen tekintik meg? Válaszolja meg a kérdést egy harmadik pont-pont diagramon, amely a megtekintések számát a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





Megoldás



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

Az értékek $2, 3, 5, 6, 7$; tehát az átlag $\frac{23}{5} = 4{,}6$, míg a medián 5.

Ha hozzávesszük a 40-et, akkor az átlag $\frac{63}{6} = 10{,}5$, míg a medián $\frac{5+6}{2} = 5{,}5$.

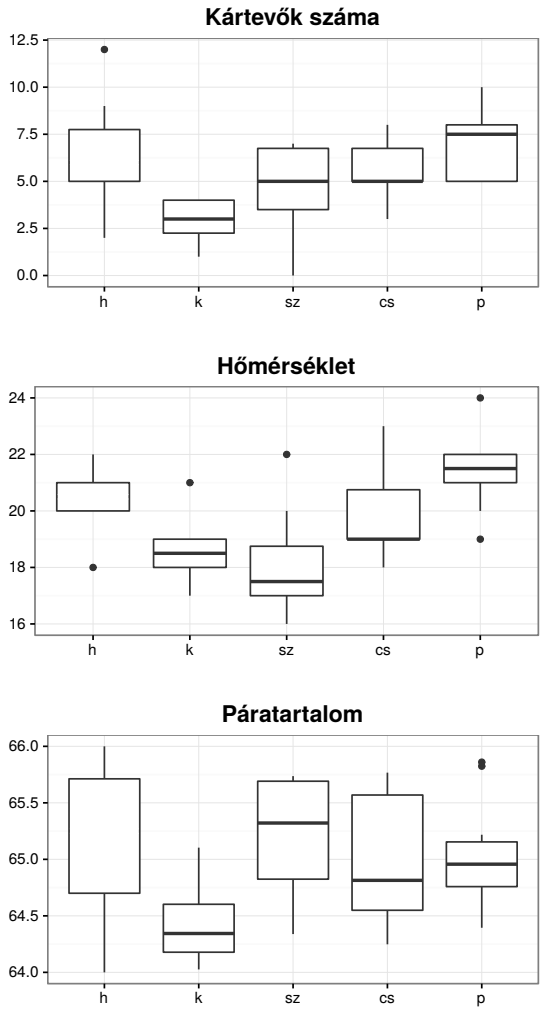
Tanulság: az átlag érzékenyebb a kiugró értékekre, a medián kevésbé.

5. Szenzorhálózat (zh)

Adott egy mezőgazdasági szenzorhálózat, amellyel a szabadföldes, üvegházi, ill. fóliasátras területeink állapotát követjük nyomon a mért értékek (hőmérséklet, páratartalom, fényerősség, szélsebesség stb.) alapján.

Dátum	Hőm. [°C]	Pára. [%]	Kártevők [db]
2015. 05. 04. 08:00	18	66,00	3
2015. 05. 04. 09:00	20	65,75	6
2015. 05. 04. 10:00	20	65,75	8
2015. 05. 04. 11:00	20	65,50	9
2015. 05. 04. 12:00	20	65,50	5
2015. 05. 04. 13:00	21	65,00	12
2015. 05. 04. 14:00	21	64,70	5
2015. 05. 04. 15:00	21	64,70	6
2015. 05. 04. 16:00	21	64,60	7
2015. 05. 04. 17:00	22	64,00	2

a) Sajnos a május 4. hétfői középértékek (medián) lemaradtak az ábráról, rajzoljuk őket be a táblázatban található adatok alapján!
b) Értelmezze a diagramokat: mely változó(k) első kvartilisei mutat(nak) szigorúan monoton változást az idő folyamán?
c) (Kiegészítő feladat.) Szeretnénk párhuzamos koordináta diagramon összevetni a hétfői hőmérsékleti értékeket a detektált kártevők számával.

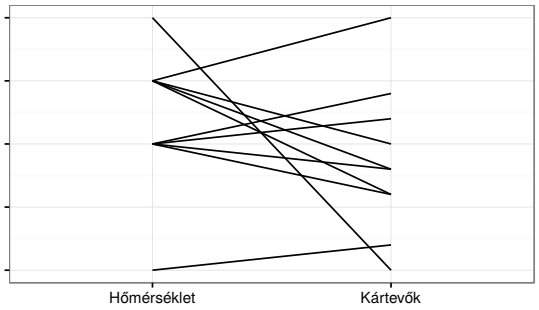


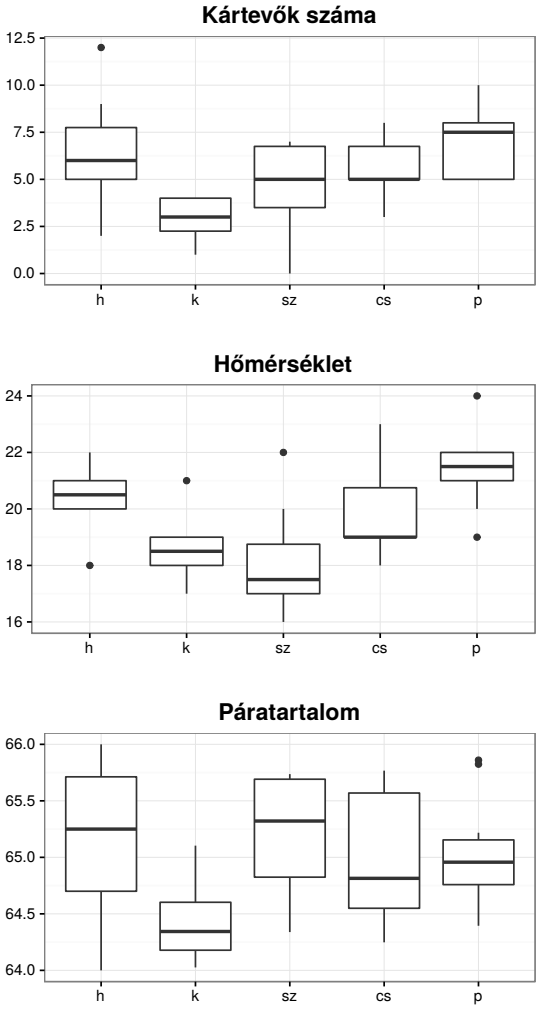
Megoldás

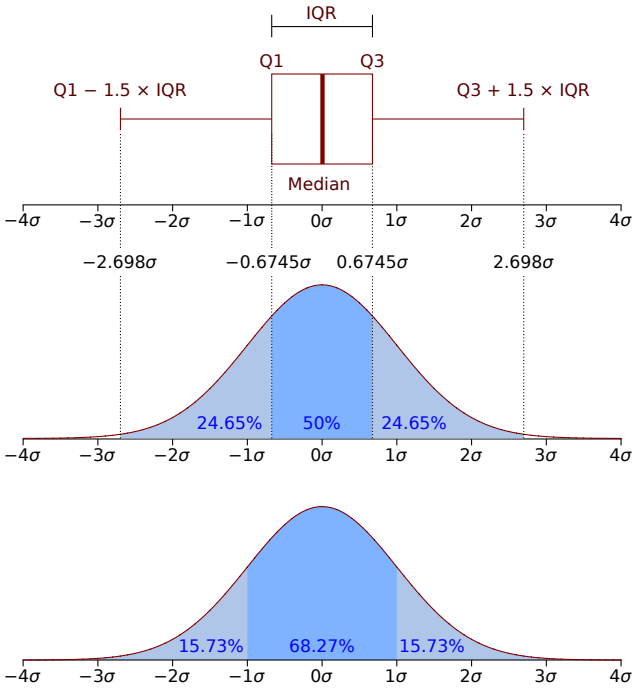
a) Rajzoljuk be a medián értékeket. Mivel páros számú értékünk van, ezért a középső kettő átlaga lesz a medián. Az első két oszlop rendezett, ezért pont a középső két érték átlaga: $\frac{20+21}{2} = 20{,}5$, ill. $\frac{65{,}5+65}{2} = 65{,}25$.
A harmadik oszlop sorbarendezve $2, 3, 5, 5, 6, 6, 7, 8, 9, 12$, így a medián $\frac{6+6}{2} = 6$.

b) Egyik sem, hiszen a „dobozok” alja nem mutat sehol szigorúan monoton változást.
A boxplot főbb jellemzőit az 1. ábra mutatja be. A $\pm 1.5 \times$ IQR-en kívül eső értékeket ponttal jelöljük.
Érdekességként megjegyezzük, hogy a 1.5 konstans használata egy statisztikai konvenció, amely analóg a a normális eloszlású adathalmazok $\pm 3\sigma$ elvével.

c) Az értékeket az alábbi párhuzamos koordináta diagramon ábrázoljuk.







1. ábra. A boxplot főbb jellemzői