Statisztika II.

Házidolgozat

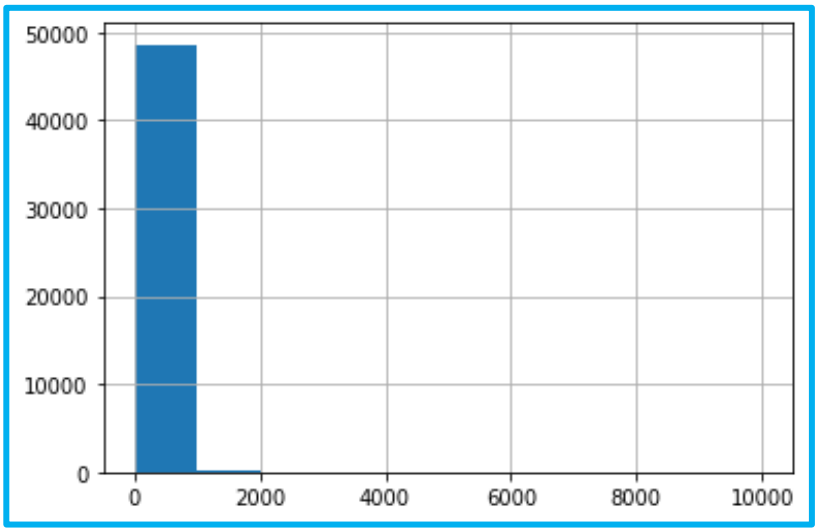
2022/2023 Tavasz

Páli András (VBII48)

II. témakör: FAE, EV és AR mintavételi hibák összehasonlítása

1. Az adatbázis[1] 2019-es adatokat tartalmaz New York-i Airbnbk-ről. A sokaság 48,895 sort tartalmaz és 16 mennyiségi, illetve 6 minőségi ismérvet. Ezekből kiválasztottam 1 mennyiségi és 3 minőségi ismérvet, amik a következők:
    - **price**: az éjszaka ára dollárban
    - **neighbourhood_group**: az Airbnb-hirdetés **régiója** New York City-ben
    - **neighbourhood**: az Airbnb-hirdetés **körzetének** meghatározása
    - **room_type**: az Airbnb-hirdetésben szereplő szobatípus

    A hisztogram alapján láthatjuk, hogy a sokaság eloszlása exponenciális.



2. Gyakorlatokon írt kód alapján megcsináltam a 200 db 100 elemű mintavételeket. Alap „paraméterekből”, ami fontos lehet az a ***random_state*** megadása, hogy minden futtatáskor ugyanaz legyen a minta, ezáltal a további elemzések is igazak lesznek. A különbség a két mintavételi mód között ebben a kódban az a ***sample*** függvényben a ***replace*** paraméter **True** vagy **False** beállítása. Ha **True** akkor FAE, egyébként EV.

3. A jelenleg a sokaság heterogén, ezt kéne valamelyik minőségi ismérv szerint homogén rétegekre bontani. Mivel a vizsgálat szempontjából fontos tényező az, hogy New York melyik körzetében van az adott Airbnb vagy az, hogy milyen a szoba típusa, ezért a mintát úgy töltöm fel, hogy mindegyik részcsoportból legyen benne egyed. Mivel AR minta kell most ezért ki kell számolnunk az egyes rész csoportok arányát, mind a három minőségi ismérv szerint és eldöntjük melyik alapján lesz a mintavétel. Az még fontos, hogy a rétegekből EV mintát veszek.

    A körzetekből 221 féle van, nagyon kicsi arányokkal. A régiókból összesen 5 db, a szoba típusokból pedig 3. A **régiókat** fogom választani, habár valószínű, hogy a körzeteknek nagyobb a varianciahányadosa, tehát jobban befolyásolja az árat. Körzetek arányai annyira kicsik, hogy nem tudtam belőlük 0-nál különböző egész számot kerekíteni, és a **sample** függvény **n** paraméterébe nincs értelme **float** típust rakni, mert nem lehet „tört” egyed. Nagyobb mintaszám (**n**) kéne ahhoz, hogy ez alapján rétegezzünk.

neighbourhoodGroup	rate
Brooklyn	0.411167
Manhattan	0.443011
Queens	0.115881
Staten Island	0.00762859
Bronx	0.0223131

[1] https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data?resource=download

4. AR mintavételhez először az arányokat szedtem ki egy tömbbe, felszoroztam százzal őket, kerekítettem és szerencsémre kijött összegre a 100. Ebből az **aranyok_lista** tömbből lesz adagolva a **sample** függvénynek a **n** paramétere. Aztán létrehoztam egy üres DF-et a mintáknak, fontos, hogy ez eltér az előzőktől abban az értelemben, hogy 2 db oszlop fog megadni 1 db mintavételt - FAE-ben ,EV-ben 1 sor volt 1 db mintavétel – és 400 oszlopa lesz mert a későbbiekben szükség lesz az értékhez tartozó rétegre is a MSE számolásnál, ahhoz kell egy belső szórás. **A random state AR-nél 40**.

Index	Value
0	41
1	44
2	12
3	1
4	2

aranyok_lista

Első **for** ciklus 200-ig megy, ennyi mintavétel kell. Egy darab AR mintához létrehozok egy üres DF-et. Belső ciklus annyiszor fut le ahány réteg lesz, jelen esetben 5. Kiszedi a sokaságból azokat a sorokat, amik az adott réteghez tartoznak egy DF-be. Ezekből már lehet mintát venni, hiszen biztos, hogy csak egyféle réteghez tartozó egyed lesz benne. Itt kell az **aranyok_lista,** az **n** paraméter megkapja a megfelelő az egész, 0-100 közötti számot. (A **uniqueNeighbourhoodGroups** és az **aranyok_lista** ugyanolyan sorrendben tartja az 5 réteget). Aztán az egy darab AR mintát tartalmazó DF-hez hozzáfüzöm az új rétegből a „rész” mintát. Ezzel rétegezéssel az **egy_darab_AR_minta** úgy fog feltöltődni, hogy először 41 db elem lesz benne az első rétegből, 44 db a másodikból, 12 a harmadikból és így tovább. Tehát meglesz 1 db 100 elemű AR minta, amit hozzáfűzök a 200 db mintát tároló DF-hez úgy, hogy minden egyes új minta két új oszlopba fog kerülni.

Index	neighbourhood_group	price0	neighbourhood_group1	price1	neighbourhood_group2	price2	jhbourhoo
97	Staten Island	50	Staten Island	175	Staten Island	85	Staten Is
85	Queens	125	Queens	150	Queens	130	Queens
86	Queens	40	Queens	69	Queens	70	Queens
87	Queens	50	Queens	70	Queens	59	Queens

airbnb_price_AR_minta

Egy új oszlopba kiszámoltatom a mintaátlagokat mind a három mintavételi módhoz. AR-nél arra figyelni kell hogy oszlopokban vannak a minták ezért itt egy új sorban lesznek a mintaátlagok.

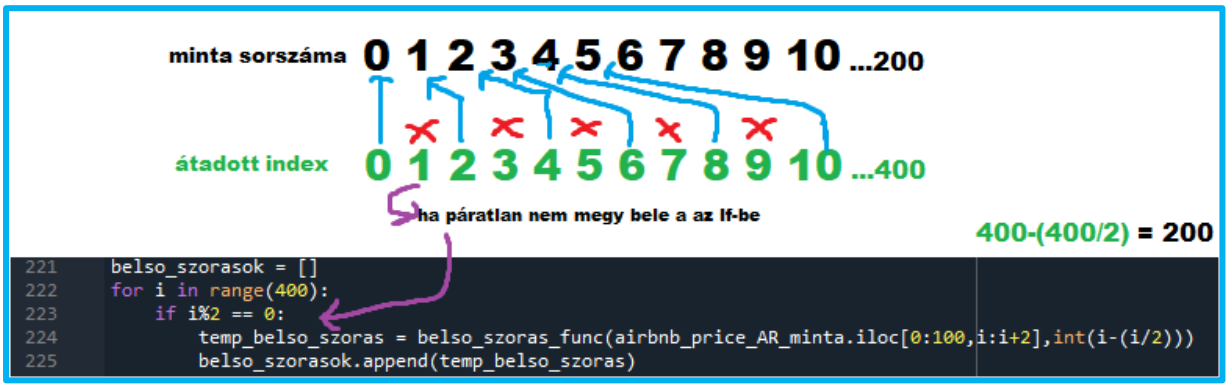
5. Az árak sokasági átlag 152.7 dollár, szórása 240 dollár. **Elvi** átlagos négyzetes hibákra jött a ”papírforma”, FAE > EV > AR. A torzítások elenyészőek.

```
{'FAE Bias': -0.944, 'EV Bias': 0.305, 'AR Bias': 0.532}
```

```
{'ELVI FAE MSE': 577.619, 'ELVI EV MSE': 575.642, 'ELVI AR MSE': 559.684}
```

6. Tapasztalati MSE-hez ki kell számolni a minta átlal becsült szórásokat. Ehhez először kellettek a korrigálatlan varianciák, abból a korrigált. Ezeknek az átlagának a gyöke megadja a becsült sokasági szórást. AR-nél azonban nem a teljes szórás kell, hanem csak a belső, hiszen ilyenkor nem kell számolni a régiók közötti ár szórással, mivel ezt a külső szórást kezeltük a régiónkénti mintavétellel. Ehhez írtam egy külön függvényt, aminek egy mintavételt kell adni, jelen esetben egy két oszlopos DF-et, és a **groupby**-os technikával visszaadja ahhoz az egy mintavételhez tartozó belső szórást. A függvény meghíváskor azonban figyelni kell, hogy hogyan adjuk meg neki

a 2 oszlopos DF-et. Mivel 200 db minta van így 200 db belső szórás kell, viszont 400 oszlopunk van. Ezért a **for** 400-ig fog menni, és csak akkor hívja meg a függvényt ha az **i** páros szám, mert egy db minta úgy áll össze, hogy a **réteget** tartalmazó oszlop mindig páros indexű, a hozzátartozó **ár** oszlop pedig páratlan. Így garantálom, hogy nem lesz olyan animália, hogy a 0-dik mintavételhez tartozó árak lesznek párosítva az első mintavételhez tartozó rétegekkel. Illetve, csak úgy az **i**-ket sem adhatjuk át, mert akkor a páratlan sorszámú mintavételeken átugrana, ezért az **i**-ből ki kell vonni **i/2**-t.



Ezeket egy tömbbe rakom, és az MSE számolásánál a 200 db kvázi „rész” belső szórások átlagát használom. Itt a torzítást nem vettem figyelembe, mert ha az átlagok átlagából kivonjuk a becsült átlagot, akkor az 0 lesz, hiszen ez a kettő ugyanaz, ha nem ismerjük a sokasági jellemzőket. Ez miatt az lesz, hogy az MSE tulajdonképpen standard hiba a négyzeten.

```
{'TAPASZTALATI FAE MSE': 520.133, 'TAPASZTALATI EV MSE': 548.38, 'TAPASZTALATI AR MSE': 330.159}
```

7. Legpontosabb becslést az AR minta adta, mert a rétegzés miatt a külső szórást kiejtjük, így az standard hiba csökkeni fog, mert csak a belső szórást használjuk a kiszámolásához.
Standard hibák különségéből azt láthatjuk, hogy az EV és FAE mintavételnél alig lett jobb a tapasztalt standard hiba, azonban AR-nél relatív sokat javult a becslés, tehát az AR becslése átlagban kisebb mértékben szóródik a többi mintavételhez képest.

Egy AR mintának az átlaga átlagosan **18,17** dollárral tér el a sokasági átlagtól.

```
{'FAE elvi-tapasztalati': 1.209,
 'EV elvi-tapasztalati': 0.573,
 'AR elvi-tapasztalati': 5.481}
```

TAPASZTALATI AR SH	18.17
TAPASZTALATI EV SH	23.42
TAPASZTALATI FAE SH	22.81

Elméletben a következő szabály áll fent: SH_fae >= SH_ev >= SH_ar.
Most a gyakorlatban EV-FAE tekintetében 102.7%-ra nőtt a standard hiba. Ami azt jelenti, hogy kicsivel rosszabb lett az EV becslése mint a FAE.

AR-t FAE-hoz nézve, azonban jelentős javulás történt, 79.7%-ra csökkentette a standard hibát a rétegzés.

```
sh_viszonyulasa_FAE_SHhoz
{'AR/FAE SH': 0.797, 'EV/FAE SH': 1.027}
```

EV és AR standard hiba képlete annyiban különbözik, hogy a teljes szórás helyett AR-ban csak a belső szórást használjuk fel. Ezért eltérést tapasztalunk, méghozzá pontosabb becslést.
Az, hogy mennyivel lesz jobb az AR mintavétel, az EV-hez képest, a külső szóráson múlik. A külső szórás mértéke attól függ hogy a régiók hány százalékban határozzák meg a árak alakulását. Ez lesz a varianciahányados, ami 2.81% lett. Relatív hatásfokot számolhatunk ennek a segítségével ami egyszerűen $1 - H^2$, tehát 97.2%-ra csökkenti a standard hiba négyzetét az arányos rétegzés.