Rendszerintegráció és -felügyelet laboratórium (VIMIM309)

Kísérlettervezési alapok és mérési eredmények kiértékelése

Mérési útmutató

Készítette: Salánki Ágnes

Utolsó módosítás: 2015. május 4.

Verzió: 1.0

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

1 Bevezető

A mérés során az alábbi működési módot szimuláljuk le: az egyes lépések a kialakított munkafolyamatban különböző késleltetéssel hajtódnak végre, mi ezeket a késleltetéseket vetjük alá statisztikai jellemzésnek.

Ennek megfelelően a mérés során elvégzendő feladatok a következőek:

1. Instrumentáljuk a kódunkat úgy, hogy az egyes lépések automatikusan hajtódjanak végre, a késleltetési értékeket szimuláljuk a Java pszeudo random számokat generáló függvénye segítségével kétféleképpen!
2. Futtassuk le a folyamatot 1000-szer, 5000-szer és 10000-szer, a tevékenységek átlagos késleltetéseit (akár a beérkezési és továbbadási időpillanatot, akár az abszolút időkülönbséget) rögzítsük le egy-egy csv-ben!
3. Végezzünk felderítő analízist a három számsoron, határozzuk meg a kapott empirikus eloszlás főbb jellemzőit! A centrális határeloszlás tételét alkalmazhatjuk-e itt hipotézisek felállítására és ha igen, milyen módon?
4. Tudjuk-e a centrális határeloszlás tételét itt úgy használni, hogy a várható értékre adott becslésünk pontosabb legyen?
5. Állítsunk fel hipotéziseket a mérések átlagát illetően, majd elemezzük azok kimenetét!
6. QQ-plot és khi-négyzet próba segítségével vessük össze az 5000 és 10000 futtatás esetén kapott átlagos késleltetéseket (homogenitásvizsgálat)!
7. Khi-négyzet próba segítségével végezzünk illeszkedésvizsgálatot az egyenletes diszkrét eloszlásunk vizsgálatára!
8. A kiadott csv-ben található késleltetéseket elemezve határozzuk meg az egyes tevékenységekre jellemző eloszlásokat, valamint azok empirikus átlagát és szórását! A kiadott késleltetések alapján hol van(nak) a rendszer szűk keresztmetszete(i)?

A továbbiakban a mérést segítendő elméleti hátteret részletezzük.

2 Pszeudo random késleltetések

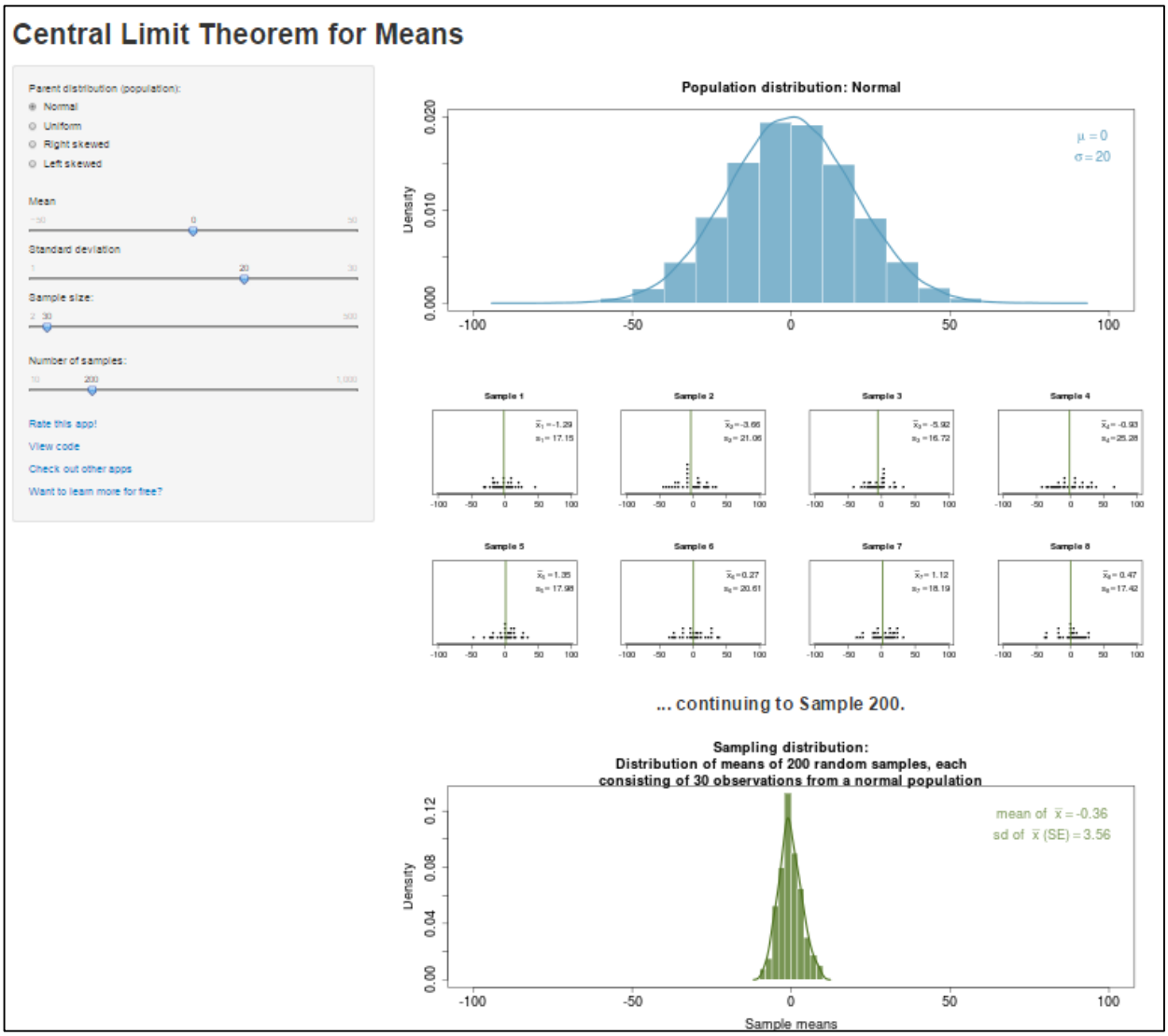
A Java Random osztálya használható egyenletes vagy normális eloszlásból származó minták generálására, előbbi támogatja mind a [0, 1.0] intervallumba eső folytonos, mint például az Integer típusú diszkrét értékeket előállítását. A mérés során próbáljuk ki a folytonos normális eloszlást (nextGaussian()) és a diszkrét egyenletes eloszlást (nextInt(n))!

3 A centrális határeloszlás tétele

Tételezzük fel, hogy néhányszor mintavételezünk egy adott populációból és kiszámoljuk a minták átlagát. Ezek után ezeket az átlagokat összegyűjtjük és meghatározzuk ezek átlagát is, legyen ez m. Ekkor a centrális határeloszlás tétele szerint bizonyos feltételek mellett az átlagok normális eloszlást követnek és a kiszámolt m érték jó becslője a nagy populáció várható értékének, függetlenül a populáció eloszlásától.

Azaz, $\bar{x} \sim N(mean = \mu, SE = \frac{s}{\sqrt{n}})$, ahol $\bar{x}$ jelöli a mintaátlagokat, μ a populáció átlagot, s a populáció szórását és n a mintaméretet.

A centrális határeloszlás tételét szemlélteti az alábbi webes alkalmazás (Ábra 1). A felületen 4 eloszlás közül lehet kiválasztani a populáció eloszlását (ezt megfelelően lehet paraméterezni), majd ki lehet választani a minták méretét és számát. A legalsó panel bemutatja a mintákból számolt átlagokból alakult eloszlást.



Ábra 1 A hivatkozott alkalmazás felhasználói felülete

Amit érdemes megfigyelni:

- Tetszőlegesen választható a populáció eloszlása, az átlagokból képezett eloszlás mindig normális eloszlást fog követni.
- Minél kisebbek mintákat veszünk, annál kevésbé közelítjük a normális eloszlást.
- Minél kevesebb mintát veszünk, a végső becslésünk a várható értéket tekintve annál rosszabb (annál nagyobb az átlagokon számított szórás).

A centrális határeloszlás tétele hatalmas jelentőségű, hiszen lehetővé teszi a nagy populáció egy értékének pontos (behatárolható hibájú) becslését a populációból vett minták értékeiből. Például, a magyar középiskolások IQ-jának várható értéke becsülhető úgy, hogy minden megyéből véletlenszerűen kiválasztunk néhány diákot, azokkal elvégeztetünk egy-egy tesztet, majd az egyes megyék átlagából számolunk egy várható értéket. A reprezentatív mintavétel technikáit jelen jegyzet nem tárgyalja, a becslés pontosságáról alább lesz majd szó.

Milyen feltételek teljesülése szükséges a centrális határeloszlás tételéhez?

- *Függetlenség*: a mintába kerülő elemek egymástól független kiválasztása biztosított kell hogy legyen.
    - véletlenszerű mintavételezéssel
    - ha a mintavételezés visszatevés nélküli, egy mintába nem kerülhet több elem, mint a populáció 10%-a

- *Mintaméret*: ökölszabály szerint ha a mintavételezett populáció nem normális eloszlást követ vagy ha nem szimmetrikus, a tétel csak 30-nál nagyobb méretű mintáknál alkalmazható biztonságosan.

Mennyire becsüljük pontosan a várható értéket ilyenkor?

Arról már volt szó, hogy a mintavétel során az egyes minták átlagából eloszlást képezünk és ennek az eloszlásnak az átlaga jó becslője a populáció átlagának. Ennek az eloszlásnak az empirikus szórásával ugyanakkor azt is meg tudjuk mondani, mennyire becsültünk jól: a populáció szórása biztosan kisebb, mint az eloszlás szórása és a mintaszám gyöke között számolt hányados.

A fenti alkalmazásban tehát 20 szórású normális eloszlásból 400 mintát véve $\frac{20}{\sqrt{400}} = 1$-hez közeli szórású eloszlást kell hogy kapjunk a végén, függetlenül a minták számától és a normális eloszlás beállított várható értékétől.

Az egyenlőség igaz a másik oldalról nézve is: ha legalább SE (mint sampling error) hibával szeretnénk megbecsülni a populáció várható értékét a populáció becsült s szórása mellett, akkor egy-egy mintavételnél legalább $(\frac{s}{SE})^2$mintát kell begyűjtenünk.

4 Egymintás u-próba végzése

Az átlagokból képezett eloszlás (a random generálás módjától függetlenül) várhatóan normális eloszlást fog követni (természetesen ezt ellenőrizzük is le). Ellenőrizzük le egymintás u-próbával, hogy az eloszlás várható értéke lehet-e 0.5, 0.1, illetve teszteljük le az átlagokból kapott eloszlás várható értékét (az átlagok átlagát) is!

Az egymintás u-próba (az angol terminológiában z-test) bemenete egy normális eloszlásúnak feltételezett populáció, valamint egy m érték, amelyet a normális eloszlás várható értékének ("populáció átlag") feltételezünk.

Nullhipotézise: az adott populáció átlaga egy adott m értékkel egyenlő.

Alternatív hipotézise: az adott populáció átlaga nem egyenlő az adott m értékkel.

A próbastatisztika értéke: $u = \frac{\bar{x}-m}{\sigma/\sqrt{n}}$, ahol $\bar{x}$ adja a mintaátlagot, m az adott teszt értéket (a mi esetünkben ez 0.5 illetve 0.1 lesz), a σ a mintából származó szórást, n-nel pedig a minta méretét jelöljük (itt most ez 1000, 5000 stb. lesz).

Egymintás u-tesztet R-ben a BSDA csomag z.test nevű függvényével végzünk, itt adható meg a populáció átlag. A p-értéket általában 0.05 alatt a nulhipotézist cáfoló bionyítéknak tekintjük.

5 Homogenitásvizsgálat qq-plottal

Tételezzük fel, hogy van két változónk és szeretnénk megállapítani ezekről, hogy azonos eloszlás generálta-e őket.

Ahhoz, hogy ezt felderítsük, szokás a két változó ún. qq-plotját felrajzolni, amely az egyes empirikus eloszlások adott kvantiliseit ábrázolják egymáshoz képest.

Emlékeztetőül: *a q. kvantilis* az adott változó azon értéke, melynél kisebb mintaelemek hányada q. A 0.25, 0.5, 0.75. és 1. kvantilisokat *kvartilisoknak* hívjuk, ha %-os arányban fejezzük ki magunkat, akkor *percentilisekről* beszélünk.

Abban az esetben, ha két változót ugyanaz az eloszlás generált, akkor az egyes kvantiliseik hasonló pontokba esnek (feltételezzük, hogy a két mintánk megfelelően nagy), így az egyes kvantiliseket egymáshoz képest ábrázolva egy egyenest kapunk.

A QQ-plot előnye, hogy a különböző eloszlásokat nagyon gyorsan ki tudjuk szűrni, illetve hogy így különböző elemszámú minták összevetésére is lehetőségünk nyílik.

R-ben a qqplot paranccsal tudjuk az x és y bemenő változóként kapott egydimenziós eloszlások QQ-plotját előállítani.

6 Homogenitásvizsgálat és illeszkedésvizsgálat khi-négyzet próbával

A khi-négyzet próba (chi-square test) két diszkrét eloszlás azonosságát cáfolhatja az alábbi módon:

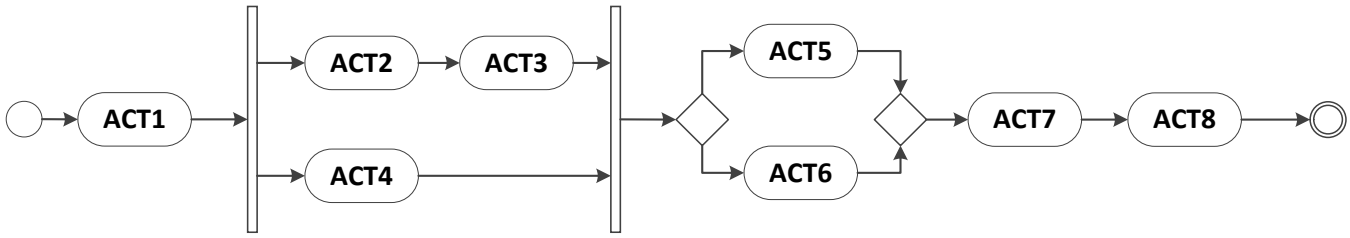
Nullhipotézis: a két vizsgált diszkrét eloszlás homogén
Alternatív hipotézis: a két vizsgált diszkrét eloszlás nem homogén
A tesztstatisztikát nem részletezzük.

R-ben a chisq.test függvény alkalmazásával futtathatunk homogenitásvizsgálatot.

7 Mesterséges adatok

Legyen adott egy üzleti folyamat (Ábra 2), amelyben előbb egy fork-join, majd egy döntéshozatal történik, egyébként az egyes tevékenységek szekvenciálisan követik egymást.

A csatolt csv állományban 10 000 kísérlet késleltetési adatai találhatóak percekben.



Ábra 2 Az elemzendő folyamat

Az elemzés során a következő kérdésekre keressük a választ:

1. Milyen az egyes tevékenységek elvégzési idejének eloszlása?
    a. Végezzen felderítő analízist az egyes késleltetéseket illetően!
    b. Becsülje meg a késleltetések eloszlását és határozza meg azok empirikus paramétereit!
2. A jó közelítéssel egyenletes eloszlással jellemezhető tevékenységeken végezzen khi négyzet próbát sejtésének igazolására!
3. Válasszon ki egy másik diszkrét eloszlást és végezzen arra is egyenletes eloszlású illeszkedésvizsgálatot majd dokumentálja az eredményt!
4. QQ-plottal hasonlítsa össze a folytonos értékeket tartalmazó eloszlásokat, határozza meg sejthetően egy eloszlásból származó tevékenységeket!
5. Egy, a normális eloszlást követő tevékenységet vessen alá egymintás u-próbának!
    a. * Válasszon ki két normális eloszlású tevékenységet és vesse alá őket kétmintás u-próbának!
6. Milyen az eloszlása a teljes workflow futási idejének?
    a. *Egyszerű korrelációanalízissel határozza meg, van-e olyan tevékenység, amelynek késleltetése jelentősen meghatározza az össz futási időt?
7. A join után található elágazáson nem látjuk a feltételt. Meg tudjuk-e tippelni pusztán az első 4 tevékenységi időit elemezve, mi lehet az elágazási feltétel?
8. Meg tudunk-e nevezni konkrétan egy-két szűk keresztmetszetet a tevékenységek között?