Rendszerintegráció és -felügyelet laboratórium (VIMIM309)

Kísérlettervezési alapok és mérési eredmények kiértékelése

Mérési segédlet

Készítette: Salánki Ágnes

Utolsó módosítás: 2015. április 30.

Verzió: 1.0

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

1 Bevezető

A labor során a méréseket végző hallgató a gyakorlatban is megismerkedik a rendszerintegráció és rendszerfelügyelet során használatos módszerekkel és eszközökkel. Végigköveti egy elosztott alkalmazás megvalósításának és felügyeletének legfontosabb lépéseit, ipari környezetben használt integrációs köztes réteg (middleware) technológiák és felügyeleti eszközök használatával. A mérések a következő témakörökhöz kapcsolódnak:

1. Munkafolyamatok megvalósítása Java nyelven
2. Megbízható üzenetküldés IBM WebSphere MQ alapon
3. Kommunikáció JMS és JMX technológia segítségével
4. OSGi szolgáltatások fejlesztése
5. Aktor modell konkurens alkalmazások készítésére: Akka
6. Rendszerfelügyelet támogatása komplexesemény-feldolgozással
7. Kísérlettervezési alapok és mérési eredmények kiértékelése

A mérések egy részén a hallgatók különféle elosztott szolgáltatásintegráció megoldásokkal valósítják meg egy példarendszer üzleti logikáját. Jelen mérés során az egyes lépések késleltetését szimuláljuk és elemezzük majd, a szimuláció bemenetének előállításához és az adott mérés konkrét kiértékeléshez használt módszereket adja meg röviden az alábbi útmutató.

2 Az adatelemzés jellemző lépései

Szokásosan az adatelemzést két, egymással iteratívan újrafuttatott fázisra osztjuk: ezek a nagyrészt a vizuális eszközöket alkalmazó felderítő elemzés (EDA – Exploratory Data Analysis) és a hipotézisteszteket használó megerősítő elemzés (CDA – Confirmatory Data Analysis).

Az EDA célja, hogy az elemezni kívánt adatról és így az általa leírt rendszerről egy általános képet kapjunk, olyan kérdések megválaszolására használjuk, mint *Hány rekordunk van? Azokat milyen attribútumok jellemzik? Hogyan alakul az egyes változók eloszlása? Milyen kapcsolatok sejthetőek az változópárokra vetített adathalmazban?* stb. A felderítő fázis végén megfogalmazzuk az adatra vonatkoztatott sejtéseinket, amelyeket a CDA fázisban ellenőrzünk majd le.

A CDA az előző fázisban megfogalmazott állítások ellenőrzésére szolgál, ezek az ellenőrző lépések a statisztikai elemzés során hipotézisesztek formájában jelennek meg. A statisztikai tesztek természetesen csak a rögzített adatot és nem az elméleti modellt látják, így nem tudják teljes biztonsággal megítélni egy-egy állítás abszolút igazságát, de sokszor megelégszünk egy "elég biztos" állítással is.

Általában a fázisok egymástól nem különíthetőek el élesen, ha egy biztosan és gyorsan ellenőrizhető sejtést fogalmazunk meg, akkor azt általában le is ellenőrizzük, majd az igazolt vagy cáfolt információval bővítve a modellünket térünk vissza a felderítő fázishoz.

2.1 A felderítő elemzés eszközei

A felderítő elemzés kihívását a megfelelő vizualizációs lépések kiválasztása adja. Ökölszabályként általában az fogalmazható meg, hogy a kisebb dimenziós plotoktól haladunk a sokdimenziósokig:

- a diszkrét egyváltozós eloszlásokat oszlopdiagramon,
- a numerikus egyváltozós eloszlásokat hisztogramon és boxploton,
- a diszkrét változópárok együttes eloszlását mozaikdiagramon,
- a numerikus változópárok együttes eloszlását pedig pont-pont diagramon ábrázoljuk.

Jelen mérés elvégzésekor legfeljebb kétdimenziós adatok vizuális elemzésére lesz szükség, így itt a sokdimenziós vizualizációs technikákat (párhuzamos koordináták, biplot) nem tárgyaljuk. A felsorolt vizualizációs technikák felfrissíthetőek Kocsis Imre Felügyeleti adatok vizuális elemzése című mérési útmutatójából, külön figyelmet fordítva a pont-pont diagramokra és a hisztogramokra, jelen mérés során ennek a két típusnak a készség szintű ismerete elvárt.

2.2 A hipotézistesztelés alapjai

A hipotézistesztelés folyamata az alábbi módon képzelhető el:

1. megfigyeljük egy adott változó egy jellegzetes tulajdonságát (pl. hogy olyan normális eloszlást követ, amelynek várható értéke 42-nél nagyobb);
2. kiszámítjuk annak a valószínűségét, hogy a sejtésünk hibás (pl. hogy a változónk eloszlása olyan normális eloszlást követ, amelynek várható értéke 42-nél kisebb), majd ha ez a valószínűség megfelelően kicsi, indirekt módon bebizonyítottnak tekintjük eredeti állításunkat. Ennek lépései a következők:
    a. megfogalmazzuk a sejtésünk ellenkezőjét: pl. a változónk eloszlása normális eloszlást követ, amelynek várható értéke 42-nél kisebb, ez a *nullhipotézis*.
    b. megfogalmazzuk a sejtésünket, mint *alternatív hipotézis*.
    c. számolunk egy ún. *teszt statisztikát* – ez egyetlen olyan numerikus érték, amely eloszlásunkat úgy jellemzi, hogy a nullhipotézis igazsága mellett általában csak egy szűken megadott intervallum értékeit vehet föl.
    d. ha a teszt statisztika értéke az adott intervallumon kívül esik, akkor a nullhipotézist elvetjük. Fontos, hogy a nullhipotézist elfogadni „nem szokás", hiszen egy jó teszt statisztika előállhat véletlenszerűen is. Ez analóg az „Ezer példa nem bizonyít, de egy ellenpélda már cáfol" elvével.

A hipotézistesztelés menete tehát hasonló az indirekt bizonyításokhoz (feltesszük a konklúzió ellentétét, majd arra a következtetésre jutunk, hogy a premisszában megfogalmazott, tényszerű állítást is cáfolnunk kell), azzal a különbséggel, hogy itt csak valószínűségi alapon tudunk bármilyen kijelentést is tenni.

Például: tudjuk, hogy a populáció intelligencia hányadosa normális eloszlást követ, aminek várható értéke 100, szórása pedig 15. Ekkor Kaszparov 190-es IQ-ja az átlagtól nagyon eltérőnek tűnik, hiszen annak az esélye, hogy egy normál eloszlás által generált pont annak várható értékétől 6 szórásnyi távolságra essen, kb. 1 az 1 milliárdhoz.

Ebben az esetben a statisztikai eljárásunk nagyjából így néz ki:

Nullhipotézis: Kaszparov teljesen átlagos

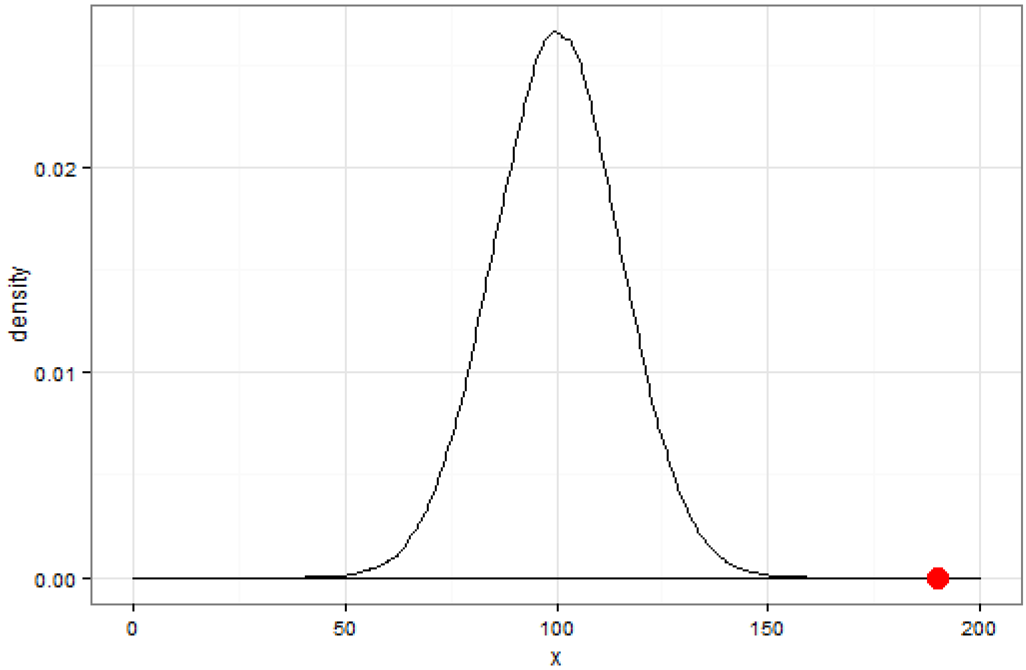
Alternatív hipotézis: Kaszparov nem annyira átlagos

Teszt statisztika: 6 (hány szórásnyira vagyunk a várható értéktől)

p-érték: 10^{-9}.

p-érték: annak a valószínűsége, hogy az adott teszt statisztika előáll, ha a nullhipotézis igaz. Valójában annak a valószínűségét adja meg, hogy tévedünk, ha elvetjük a tesztünk végén a nullhipotézist (elsőfajú hiba valószínűsége). Jelen esetben tehát $(1 - 10^{-9})$ bizonyossággal állíthatjuk, hogy Kaszparov IQ-ja különleges a teljes populációt tekintve.

A mérés során a teszt statisztika kiszámítási módja adott lesz, tehát a beugrón csak az alapfogalmak megértését kérjük majd számon.



Néhány parancs, ami a mérésen hasznos lehet

```
########################### Normális eloszlás ##################################
## Szimuláció
x <- rnorm(n = 1000000, mean = 100, sd = 15)
## Mekkora a valószínűsége, hogy egy normális eloszlás által generált érték
## nem éri el a 190-et?
pnorm(q = 190, mean = 100, sd = 15)
## Hol van az adatok felső egy milliomod részének a vágópontja?
qnorm(0.999999, mean = 100, sd = 15)

## Rajzoláshoz
require(ggplot2)
x <- data.frame(x = rnorm(n = 100000, mean = 100, sd = 15))
base <- ggplot(x)
base + geom_density(aes(x = x)) +
  theme_bw() + xlim(c(0, 200)) +
  geom_point(aes( x = 190, y = 0), col = "red", size = 5)
```

Típusuk szerint megkülönböztetünk paraméteres és nemparaméteres próbákat:

- a paraméteres próbák esetén egy konkrét eloszlás adott paraméterét próbáljuk megbecsülni, pl. az adott mintát generáló normális eloszlás várható értéke 42.
- a nemparaméteres próbák esetén két eloszlásról úgy nyilatkozunk, hogy valójában a konkrét elméleti eloszlásukról nincs fogalmunk. Pl. összehasonlítunk két mintát és megállapítjuk, hogy azok azonos eloszlásból lettek-e generálva (anélkül, hogy a konkrét eloszlásokat megneveznénk).

A mérés során egymintás u-próbát végzünk majd, illetve khi-négyzet próbával és qqplotos módszerrel vetjük össze két adatsor eloszlását. Az említett módszerek közül egyik ismerete sem elvárt a mérést megelőzően.

3 A mérés menete

A mérés két részfeladatot tartalmaz majd.

Egyrészt, az egyes lépésekben eltöltött időt véletlenszerűen generáljuk és ezeket az időket eltároljuk, majd hipotézisteszttel leellenőrizzük, hogy a mintában az elméleti paraméterek megfelelően jelennek-e meg a végeredményben.

Másrészt, egy kívülről kapott generált adatsorról fogjuk megsejteni és leellenőrizni, hogy az milyen eloszlás szerint lett generálva.

A mérés előtt érdemes átismételni a következő valószínűségszámítási alapfogalmakat.

Ezek:

- várható érték, átlag, empirikus szórás és szórásnégyzet, kvantilis;
- minta, populáció;
- valószínűségi változó, annak eloszlásfüggvénye és sűrűségfüggvénye;
- a klasszikus folytonos eloszlások (egyenletes, exponenciális, normális, Poisson) sűrűség- és eloszlásfüggvényeinek formája.

A mérés előtt érdemes felkészíteni a folyamatot a következőkre:

- az egyes tevékenységek idejét egy külső Generate osztály next() függvényétől kérdezzük le, aminek a belsejében kezdetben a klasszikus Java random szám generátor kerüljön;
- tudjuk a tokenhez tartozó egyes időket eltárolni valahol permanensen (érdemes csv vagy valami szeparált txt állományban gondolkodni a későbbi feldolgozás céljából);
- a teljes tevékenységet tudni automatikusan lefuttatni tetszőlegesen sokszor (valószínűleg ezres nagyságrendben mozog majd a kísérletek száma)

4 Ellenőrző kérdések

- A definíció szerint az M ásványvíz töltő űrtartalma 1.5 liter. Megmértünk 20 palackot és azt találtuk, hogy az átlagos űrtartalom 1.495 liter, ebből arra következtetünk, hogy a 1.5 liter állítása csalás. A biztonság kedvéért lefuttatunk egy tesztet, hogy megállapítsuk, az 1.495 valóban furcsa-e. Hogyan fogalmazzuk meg tesztünkhöz a nullhipotézist és az alternatív hipotézist?
- Mit látunk, ha két, különböző szórású, de azonos várható értékű normális eloszlásból generált 10000 pont mind a 100 percentilisét ábrázoljuk egymáshoz képest?
- Mik a paraméteres és nemparaméteres próbák? Mondjunk mindkettőre példát!
- Hogyan néz ki egy 12 várható értékű, de 3 szórású normális eloszlás eloszlás- és sűrűségfüggvénye? Hogyan néznek ki egy ugyanilyen paraméterezésű egyenletes eloszlás nevezetes függvényei?