Design experimentu a statistika - AGA46E

Instructor: M. Maciak (Czech University of Life Sciences, Prague)

Final Project - zadani - Duben 2015

Zakladni instrukce

- Finalni projekt je individualni prace ve ktere kazdy student samostatne analyzuje vlastni datovy soubor. Kazdy student ma k dispozici individualni data vygenerovane na zaklade datumu narozeni.
  Finalni projekt musi byt spracovan za pomoci vhodneho statistickeho softwaru (preferovany je statisticky software `R`, ale mozne jsou i jine varianty). Program **Microsoft Excel neni vhodny statisticky software**!

- Data potrebne k spracovani finalniho projektu jsou simulovana data s fiktivnimi vysledkami hodnoceni studentov Ceske Zemedelske Univerzity v Praze. Datovy soubor obsahuje nekolik tisic fiktivnich studentu a nekolik promennych (diskretnych a spojitych), ktere byli pro kazdeho studenta zaznamenany.

  Datovy soubor finalProjectData2015.RData (k stazeni primo po kliknuti na predchozi odkaz se jmenem souboru) je k dispozici take v univerzitnem systemu Moodle a na webove strance predmetu.

- **Dulezite:** Ke spracovani finalniho projektu nepouzivejte cely datovy soubor, pouze mensi (individualni) datovy soubor vygenerovany pomoci automatickeho `R` scriptu *generator.r*, ktery je stejne soucasti tohohle finalniho projektu (k stazeni v systemu Moodle a take na webove strance predmetu: generator.r). Podrobny postup, jak si kazdy student/studentka samostatne vygeneruje prislusny datovy soubor, je detailne popsany dalsi sekci.

- **Finalni projekt je nutne odovzdat nejpozdeji dva dni pred skouskou! Student muze reseni odovzdat jako "paper copy" nebo ho zaslat emailem na adresu *maciak@af.czu.cz*. V pripade, ze student posle reseni emailem, je nutne, aby byl finalni report ulozeny ve formatu PDF (pouze jeden soubor PDF, ktery obsahuje vsechnu potrebne nalezitosti - popis, obrazky, tabulky, vysledky a finalni interpretaci). Nazev PDF suboru musi byt intuitivni (napr. '*jmenoprijmeni.pdf*').**

- Kazdy projekt bude vyhodnocen individualne. Maximum ohodnoceni je $100\ \%$, co predstavuje 10 bodu do zapoctoveho hodnoceni. Podrobne informace o hodnoceni predmetu a zpusobu ziskani zapoctu jsou v systemu Moodle, nebo na strance predmetu);

Popis promennych

- data obsahuji simulovane vysledky studentu Ceske Zemedelske Univerzity v Praze.

- v celkovem datovem souboru je zaznamenano 6500 studentu a sledovanych je celkove 16 ruznych promennych (spojitych, diskretnych i faktorovych).

- kazdy student dostane ke spracovani pouze mensi cast datovehou souboru, ktery bude automaticky vygenerovany pomoci souboru *generator.r*. Datovy soubor pro kazdeho studenta by mel obsahovat cca 700 pozorovani (radku) a cca 9-10 promennych (sloupcu). Presne hodnoty poctu pozorovani a poctu promennych jsou ruzna pro ruznych studentov.

- popis jednotlivych promennych je uvedeny nizsie. Ne vsechny promenne jsou ale nutne obsazene v souboru, ktery bude mit k dispozicii kazdy student - student pouze dostane vyber nekterych promenych.

  1. *Gender* - pohlavi studenta: 'male' nebo 'female';
  2. *Age* - vek studenta udavany v rocich;
  3. *Nationality* - narodnost studenta - kodovana pouze jako 'czech' pro studenta/studentku ceske narodnosti a 'other' jinak;
  4. *Expenses* - celkove mesacne naklady na zivobyti odhadnute studentem v tisicich korunach;
  5. *Housing* - informace o tom, jesli student bydli na koleji - 'dormitory', nebo s rodici - 'parents', nebo ve vlastnem - 'own';
  6. *Class* - informace o tom, ve kterem rocniku je dany student aktualne zapsany;
  7. *Faculty* - informace o tom, na ktere CZU fakulte student aktualne studuje;
  8. *WeekHours* - priblizny prumerny pocet hodin, ktere je dany student tyzdenne na univerzite/fakulte;
  9. *WorkHours* - priblizny prumerny pocet hodin, ktere student potrebuje tyzdenne na samostudium a pripravu do skoly;
  10. *WeekClasses* - priblizny prumerny pocet vyucovacich hodin za jeden tyzden;
  11. *Score1*, *Score2* a *FinalScore* - celkovy prumerny prospech v procentach za dva ruzne semestre (1 a 2) a pak celkove hodnoceni za vsechny semestre;
  12. *Grade1*, *Grade2* a *FinalGrade* - celkovy prumerny prospech propocitany na vyslednou znamku - uvadene za dva ruzdne semestre (1 a 2) a pak celkove hodnoceni za vsechny semestre dohromady;

Generovani vlastniho datoveho souboru

- k vygenerovani vlasniho datoveho souboru je za potreby samostatny `R`-kovy skript *generator.r*, ktery je k stazeni primo po kliknuti na predchozi odkaz, nebo je k dispozici v systemu Moodle nebo na webove strance predmetu.
- soubor *generator.r* si stahnete a ulozte do pracovniho adresaru. Do stejneho pracovniho adresaru ulozte aj datovy soubor *finalProjectData.RData*.
- otevrte si program `R` a nastavte spravny pracovni adresar (pomoci prikazu `setwd()` a `getwd()` nebo vyklikanim v hlavnim menu programu).
- overte, jestli ste ve spravnem adresaru - zadejte v `R`-ku prikaz

  ```
  > list.files()
  ```

  a meli by ste videt zoznam vsech souboru, ktere se nachazeji v danem pracovnem adresaru. Jestlize je pracovny adresar spravny, musi mezi nimi take byt soubory *generator.r* a *finalProjectData.RData*.
- v `R`-ku zadejte nasledujici prikaz a potvrdte tlacitkem 'enter':

  ```
  > source("generator.r")
  ```

- nasledne postupujte podla instrukci na obrazovke:

  1. zadejte prvnich 6 cisel sveho rodneho cisla (bez lomitka, pouze prvnich 6 cisel) a nasledne potvrdte tlacitkem 'enter';
  2. vyberte si jednu z ponukanych moznosti: zadejte hodnotu 1 kdyz preferujete hokej, hodnotu 2, kdyz preferujete futbal a hodnotu 3, kdyz preferujete tenis a opet potvrdte tlacitkem 'enter';
  3. TO JE VSE!

4. prislusny datovy soubor s jmenem *mojeData* byl automaticky vytvoren v R-ku (napiste napr. prikaz `ls()` a uvidite objekt s nazvem *mojeData*).
5. prislusny datovy soubor bol stejne tak ulozeny do Vaseho pracovniho adresaru - jmeno souboru je *mojeDataVASECISLO.txt*. Tento soubor muzete kedykolvek opakovane nacitat v programu `R` pomoci prikazu `read.table()`.
6. **ke spracovani finalniho projektu pouzite hladinu vyznamnosti, kteru Vam vygeneroval software pri generovani datoveho souboru (prislusna hladina spolehlivosti se automaticky zobrazi na obrazovce)!**
7. generovani datoveho souboru muzete kedykolvek zopakovat (dostanete stejne data), staci, kdyz zadate prvnich 6 cisel rodniho cisla a spravnu volbu sportu, ktery preferujete;

Pozadavky ke spracovani

Spracovani finalniho projektu je zcela individualni. Je ovsem nutne, aby byl finalni projekt ucelenym reportem - uvod, popis dat, hypotezy, metodologie, spracovani, vysledky, a finalni interpretace. Celkovy objem reportu neni stanoveny, bez problemu je mozne odovzdat report na cca 2-3 stranach, a report bude kompletny.

Je ovsem nutne splnit nasledujici:

1. uvest nekolik popisnych statistik, popsat data, udelat nekolik obrazku, vysvetlit, predpokladat mozne zavery...
2. alespon dva intervaly spolehlivosti (podle Vasi volby);
3. alespon dva statisticke testy (opet podle Vasi individualni volby);
4. alespon jeden parovy a jeden neparovy t-test;
5. alespon dvakrat vyuzit metodu analyzy rozptylu interpretovat vysledky;
6. alespon jednou pouzit linearni regresni model;
7. **V uvodu reportu, ktery odevzdate, uvedte 6 cisel rodniho cisla, ktere ste pouzili k simulaci a stejne tak Vasi volbu ohledem preferovaneho sportu.**