Semestrální projekt

Semestrální projekt spočívá v nalezení vhodného datového souboru a jeho statistické analýze s využitím metod probíraných v rámci předmětu.

Data

Data lze využít vlastní (laboratorní měření, provozní data, ...), z internetu nebo z jakéhokoliv jiného zdroje. Dbejte na to, aby data byla náhodným výběrem. Musí tedy jít o nezávislá pozorování, měření apod. Počet dat není nijak omezen. Je však potřeba, aby data měla určitý minimální rozsah, zejména tehdy, pokud to vyžadují předpoklady zvolené metody. Velmi obecně lze doporučit minimální rozsah okolo 30 dat. Vždy ale velmi záleží na povaze dat a zvolené metodě.

Volbu datového souboru by měl student provést až po probrání učiva o testování hypotéz. Datový soubor, který budete vyhodnocovat v semestrálním projektu, musí být výběrovým souborem (vzorkem) z nějaké širší množiny tzv. základního souboru neboli populace.

- Počet prvků výběru: minimálně 30.
- Počet prvků základního souboru: řádově tisíce, teoreticky nekonečně mnoho.

V projektu byste měli specifikovat základní a výběrový soubor. Pokud mají data např. 200 hodnot a nejsou náhodným výběrem z nějaké populace, nelze je použít. Jednalo by se o tzv. vyčerpávající šetření, u něhož pozbývá statistická indukce smysl.

Zvolte si data, která již mají charakter náhodného výběru. Není vhodné, aby student prováděl „náhodný výběr“ sám, neboť v rozsahu tohoto předmětu není analýza metod sběru dat, které zajistí skutečně „náhodnost“ výběru. Pouze v případě, kdy si sestavujete vlastní anketu, je povolena výjimka, kdy je na volbě studenta zajistit si intuitivními metodami „náhodnost“ volby respondentů v dotazníku. Při přejímání dat v „hotové podobě“ není vhodné ze zadaných dat nijak „vybírat“, ale je třeba si dát pozor, zda datový soubor je skutečně „vzorkem“ z nějakého základního souboru neboli populace. Pokud datový soubor není výběrovým souborem z nějaké populace, nelze jej ke zpracování použít!

Dále je vhodné zvolit si alespoň jednu spojitou číselnou proměnnou (statistický znak). Tedy proměnnou, která může nabývat teoreticky nekonečně mnoha hodnot buď na libovolný počet desetinných míst v rámci uzavřeného intervalu (např. naměřená délka s libovolnou přesností) anebo různých celočíselných hodnot v rámci intervalu velké délky (např. věk).Pokud je v zadaném souboru více proměnných než chcete vyhodnocovat, okomentujte tuto situaci, a dále s přebytečnými proměnnými nepracujte. Závěry pak vyslovte pouze pro statistické znaky, které jste analyzovali (je tedy povoleno redukovat sloupce nikoli řádky zvolené tabulky).

Metody

Analýza datového souboru by měla vždy obsahovat explorační analýzu a alespoň jednu z níže uvedených skupin metod statistické indukce:

- ANOVA, resp. Kruskalův-Wallisův test + intervalové odhady středních hodnot, resp. mediánů,
- test nezávislosti v kontingenční tabulce + intervalové odhady pravděpodobnosti (nezapomeňte, že v tomto případě by součástí explorační analýzy měly být i míry kontingence),
- lineární regrese (včetně predikce).

Projekt je vhodné začít zpracovávat až poté, co si rozmyslíte základní otázku (resp. otázky), které budete ověřovat. Úvodem krátce popište svá data, jejich zdroj, popř. můžete uvést krátkou ukázku dat. Uveďte, jak v daném případě vypadá populace, tj. množina objektů, o nichž lze dělat závěry na základě statistické indukce. Při řešení jednotlivých otázek vždy nejprve proveďte explorační analýzu, okomentujte výsledky takto získané a následně pomocí metod statistické indukce (intervalové odhady, testování hypotéz, regrese) zobecněte závěry na sledovanou populaci. Není vhodné, aby byl projekt rozdělen na nesouvisející části Explorační analýza a Statistická indukce. Jednotlivé části projektu by měly odpovídat řešeným otázkám. Práce má tvořit kompaktní celek. Závěrem krátce shrňte získané výsledky.

Explorační analýza

Pod explorační analýzu spadá:

- základní popis zkoumaných proměnných a jejich vztahů (číselné charakteristiky, grafické zobrazení),
- identifikace odlehlých pozorování (nestačí odlehlá pozorování identifikovat, musíte se rovněž rozhodnout, jak s případnými odlehlými pozorováními naložíte a své rozhodnutí zdůvodnit),
- grafické posuzování normality numerických spojitých proměnných (normalitu ověřujte pouze v případě, že je předpokladem pro použití metod statistické indukce).

Poznámky k vybraným metodám statistické indukce

Pro testování hypotéz je vždy vhodnější používat parametrické testy, které mají větší sílu testu (schopnost detekce správné alternativní hypotézy) než testy neparametrické. (Síla neparametrických testů klesá z důvodu ztráty původní informace o datech, která jsou nahrazena jejich pořadím.) Použití parametrických testů je však obvykle podmíněno splněním předpokladu normality.

Testy úrovně (střední hodnota, medián)

Jednovýběrové testy

Nejprve otestujte normalitu dat. Je-li splněn předpoklad normality dat, použijte standardní **t-test** pro střední hodnotu, tj. testujte hypotézu $H_0 : \mu = \mu_0$.

Není-li splněn předpoklad normality dat, použijte **Wilcoxonův test** pro ověření úrovně mediánu (místo testování střední hodnoty), tj. testujte hypotézu $H_0 : x_{0,5} = x_{0,5_0}$.

Dvouvýběrové testy

Nejprve otestujte pro oba výběrové soubory normalitu dat. V případě, že jste nezamítli předpoklad normality, je třeba před provedením dvouvýběrových **nepárových testů** provést test shody rozptylů (tzv. homoskedasticity).

- **Nepárové dvouvýběrové testy**

  U nepárových testů je nezbytné, aby náhodné výběry byly **nezávislé**.

  Je-li splněn předpoklad normality obou výběrových souborů i předpoklad homoskedasticity, použijte standardní **dvouvýběrový t-test** pro srovnání středních hodnot dvou základních souborů, tj. testujte hypotézu $H_0 : \mu_1 = \mu_2$.

  Je-li splněn předpoklad normality obou výběrových souborů, avšak předpoklad homoskedasticity je porušen, použijte **Aspinové - Welchův test** shody středních hodnot dvou základních souborů, tj. testujte hypotézu $H_0 : \mu_1 = \mu_2$. *(Ve Statgraphicsu je nutno v Pane Options odškrtnout políčko Assume equal variances.)*

  Jestliže je předpoklad normality porušen (alespoň v jednom z výběrů), použijte některý ze srovnávacích testů mediánů, např. **Mannův - Whitneyův test**, tj. testujte hypotézu $H_0 : x_{0,5_1} = x_{0,5_2}$.

- **Párové testy**

  U párových testů předpokládáme závislost náhodných výběrů, hodnoty jsou zadány „v páru“. Každé dvě párové hodnoty se týkají vždy téže statistické jednotky, např. zátěžová a klidová tepová frekvence naměřená u téhož pacienta nebo ojetí pravé a levé přední pneumatiky zjišťována u téhož auta.

  Je-li splněn předpoklad normality obou výběrových souborů, použijte **párový t-test** pro testování úrovně střední hodnoty diferencí (rozdílů párových hodnot), tj. testujte hypotézu $H_0 : \mu_d = \mu_0$. (Očekáváte-li shodu párových hodnot, testujte hypotézu $H_0 : \mu_d = 0$.)

  Jestliže je předpoklad normality porušen (alespoň v jednom z výběrů), použijte některý z mediánových testů pro testování úrovně diferencí: např. **mediánový test**, tj. testujte hypotézu o mediánu rozdílů párových hodnot ($H_0 : x_{0,5_d} = x_{0,5_0}$).

.

Vícevýběrové testy (testy shody úrovně pro $k \geq 3$ výběry)

- **Jednofaktorová analýza rozptylu (ANOVA)**

  Jednofaktorová ANOVA je rozšířením nepárových dvouvýběrových testů shody středních hodnot, resp. mediánů. Použití ANOVy je podmíněno **nezávislosti výběrů**, což je třeba zajistit již při plánování experimentu. Pokud by náhodné výběry nebyly nezávislé, nebylo by možné ANOVu provést ani v neparametrické podobě.

  Předpokládejme, že máme k dispozici nezávislé výběry. Pak o tom, zda použijeme k analýze parametrickou či neparametrickou ANOVu rozhoduje splnění předpokladů normality a homoskedasticity.

  Je-li splněn předpoklad normality u všech výběrových souborů, tzv. tříd, i předpoklad homoskedasticity, použijte standardní, tj. **parametrickou, podobu analýzy rozptylu ANOVa** pomocí Fisherova F-testu, tj. ověřte shodu středních hodnot všech základních souborů ($H_0 : \mu_1 = \mu_2 = ... = \mu_k$). Dojde-li k zamítnutí nulové hypotézy, použijte k post-hoc analýze (vícenásobnému porovnávání) **Tukeyho korigovaný test**, resp. **LSD test s Bonferroniho korekcí**, resp. **Schéffého test** (v případě výběru malých rozsahů).

  Jestliže je porušen předpoklad normality (alespoň v jednom z výběrů) nebo předpoklad homoskedasticity, použijte neparametrickou podobu analýzy rozptylu ANOVa, tzv. **Kruskalův - Wallisův**

test, kdy testujeme rovnost mediánů základních souborů, tj. testujeme hypotézu $H_0 : x_{0,5_1} = x_{0,5_2} = ... = x_{0,5_k}$. Dojde-li k zamítnutí nulové hypotézy, použijte k post-hoc analýze (vícenásobnému porovnávání) **Dunnové test**, resp. v případě vyváženého třídění **Nemeneyiho test**. (*POZOR! Statgraphics nenabízí post-hoc analýzu pro Kruskalův-Wallisův test! V případě potřeby můžete použít např. excelovský výpočetní applet dostupný zde.*).

- **Friedmanův test**

  Jestliže jsou náhodné výběry závislé (obdoba párových dat), použijte tzv. **Friedmanův test**. Tímto testem ověřujete rovnost mediánů základních souborů, tj. testujete hypotézu $H_0 : x_{0,5_1} = x_{0,5_2} = ... = x_{0,5_k}$. Dojde-li k zamítnutí nulové hypotézy, použijte **Friedmanův test pro post-hoc analýzu**. (*POZOR! Statgraphics nenabízí post-hoc analýzu pro Friedmanův test! V případě potřeby můžete použít např. excelovský výpočetní applet dostupný zde.*).

.

Ověřování předpokladů parametrických testů:

Předpoklady testů (normalita, shodu rozptylů) je vhodné ověřovat jak pomocí metod explorační analýzy, tak pomocí exaktních testů.

.

Ověřování normality

Normalitu lze orientačně posoudit např. dle **histogramu, odhadu hustoty pravděpodobnosti, p-p grafu, q-q grafu**...

Při exaktním testu testujeme nulovou hypotézu: H_0: *výběrový soubor je realizací náhodného výběru z normálního rozdělení.* Ve většině běžných situací (rozsah výběru je mezi 10 a 2000) se doporučuje pro ověřování normality používat **Shapirův-Wilkův test** (ČSN 010225). K dalším známým testům patří **Andersonův-Darlingův test, Lilieforsův test, χ^2-test dobré shody, kombinovaný test šikmosti a špičatosti**, ... Při uvádění výsledků testu normality uvádějte vždy i název použitého testu!

Problémem nastává v případě malých ($n < 10$) nebo velkých ($n > 2000$) výběrů. Je doloženo, že testy normality vykazují pro malé výběru nízkou sílu testu (pravděpodobnost detekování nenormality). Je-li výsledkem testu (pro malý výběr) zamítnutí normality, je téměř jisté, že výběr nepochází z normálního rozdělení. Pokud test ukazuje na nezamítnutí normality, znamená to, že nemáme dostatek důkazu pro to, abychom mohli normalitu zamítnout. Naopak při ověřování normality velkých výběrů je většinou již malý odklon od normality považován za statisticky významný (příčinou je příliš vysoká síla testu). V těchto případech se doporučuje posuzovat normalitu spíše na základě exploratorních grafů.

- **Jak postupovat v případě zamítnutí normality?**

  Při zamítnutí hypotézy o normalitě dat je možné provést buď transformaci dat a přiblížit se tak normalitě nebo přejít na neparametrické testy.V případě, že se rozhodneme pro transformaci, je zřejmé, že půjde o transformaci nelineární, neboť lineární transformace by zachovala původní tvar rozdělení. Použitelné algoritmy jsou:

  - **odmocninová transformace** $t = \sqrt{x}$, mají-li data charakter četností,
  - **logitová transformace** $t = \frac{1}{2} \ln \frac{x}{1-x}$, jde-li o podíly (relativní četnosti),
  - **logaritmická transformace** $t = \ln x$, jsou-li data výběrem z logaritmicko-normálního rozdělení.

  V mnoha případech výše uvedené transformace nepomohou a musí se vyzkoušet náročnější způsoby - např. **Boxův-Coxův systém transformací** nebo plošnou (nelineární) transformací.

.

Ověřování homoskedasticity

Předpokládejme, že máme k **nezávislých výběrů**. Pro orientační posouzení shody rozptylů v populacích, z nichž tyto výběry pocházejí, lze využít pravidlo, které říká, že v případě homoskedasticity (shody rozptylů) by poměr mezi největším a nejmenším výběrovým rozptylem neměl být větší než 2.

Při exaktním posouzení homoskedasticity testujeme nulovou hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ vůči alternativě, že alespoň jedna dvojice rozptylů se liší.

Je-li splněn předpoklad normality ve všech třídách, pak použijeme standardní **dvouvýběrový F-test** (je-li $k = 2$) nebo **Bartlettův test** (je-li $k > 2$).

Pokud je předpoklad normality (alespoň v jedné třídě) porušen, použijeme **Leveneův test**. (*POZOR! V případě, že je splněn předpoklad normality ve všech třídách, preferujeme dvouvýběrový F-test, resp. Bartlettův test, neboť tyto mají větší sílu testu než neparametrický test Leveneův.*)

.

Analýza závislosti v kontingenční (resp. asociační) tabulce

Při této analýze nejdříve posuďte míru závislosti na základě exploratorní analýzy. Využijte vhodné **grafy** (např. mozaikový graf) a **míry kontingence** (Cramerovo V, koeficient kontingence, ...), resp. **míry asociace** (poměr šancí, relativní riziko - *POZOR! Statgraphics míry asociace neumí určit.*).

V rámci statistické indukce testujte hypotézu

H_0: Diskrétní veličiny X, Y jsou nezávislé.

Předpoklady pro použití χ^2 testu dobré shody jsou:

- všechny **očekávané** četnosti jsou alespoň 2,
- alespoň 80% všech **očekávaných** četností je větších než 5.

Pokud jsou splněny předpoklady testu, použijte χ^2 test dobré shody (je implementován ve Statgraphicsu). V případě, že předpoklady testu splněny nejsou, pokuste se nejdříve sloučit sousední řádky nebo sloupce tabulky (minimální rozměr tabulky musí být 2x2). Pokud ani tak nezajistíte splnění předpokladů, vyslovte závěr, že nezávislost veličin X, Y nelze exaktně testovat.

V případě, že ověřujete nezávislost v asociační tabulce (tabulka 2x2), využijte pro statistickou indukcí rovněž intervalové odhady měr asociací (*POZOR! Statgraphics míry asociace neumí určit.*)

Nástroje

Využít lze jakýkoliv vhodný software.

- **Statgraphics** - ve verzi 5 dostupný na počítačích VŠB-TUO, 30 denní trial verze Statgraphicsu Centurion je stažitelná z www.statgraphics.com
- **MS Excel, LibreOffice apod.** - vhodný pro jednoduchou analýzu, zpracování grafů
- **R - open source** skriptovací jazyk, prostředí pro statisické výpočty, dostupný pro Win, Linux, Mac OS, domovská stránka: www.r-project.org
- **Matlab, Octave** - pro běžné metody statistické analýzy lze využít i univerzální matematický software

Doporučení

- Je-li to možné, najděte si data, která jsou pro Vás něčím zajímavá.
- Vyhněte se datům z internetových anket typu www.vyplnto.cz. Vzhledem k tomu, že anketa nepatří mezi metody náhodného výběru, bude prakticky nemožné určit populaci.
- Buďte struční! Projekt nemá minimální požadovaný rozsah. Pokud nepoužíváte metody nebo nástroje, které nejsou náplní výuky, neuvádějte definice nebo vysvětlení použitých pojmů a nástrojů.
- Nezapomeňte v každé fázi práce stručně okomentovat získané výsledky.

Přehled častých chyb

1. **Časté chyby v obsahové stránce:**
   - malý rozsah souboru (doporučeno je 30 statistických jednotek v každé třídě!),
   - u nominální proměnné jsou chybně uvedeny kumulativní četnosti a kumulativní relativní četnosti,
   - výsečové grafy jsou zobrazeny bez udání absolutní četnosti,
   - výsečové grafy jsou zobrazeny pro příliš velký počet kategorií,
   - při testování hypotéz není uvedeno, co přesně testujete (chybí nulová a alternativní hypotéza),
   - nejsou ověřeny předpoklady testů,
   - metody statistické indukce nejsou voleny v závislosti na výsledku ověření předpokladů,
   - metoda ANOVA je aplikována na příliš velký počet tříd (více než 15),
   - u metody ANOVA je chybně uvedena post-hoc analýza i v případě nezamítnutí nulové hypotézy,
   - chybná predikce hodnot u regrese (nevhodná extrapolace).
2. **Časté chyby ve formální stránce:**
   - není uveden zdroj dat,
   - není uveden použitý software (včetně čísla verze),
   - mnoho obecných teoretických komentářů, málo komentářů k vlastním datům,
   - chybné nastaveni fontů v popisu grafů ve Statgraphicsu (font, který umí ve Statgraphicsu češtinu je Středoevropský),
   - používání angličtiny v českém textu,
   - velké množství pravopisných chyb a překlepů (použijte alespoň korektor pravopisu)