**12 Porovnání empirického a teoretického rozdělení a neparametrické testy**

**A. Porovnání empirického a teoretického rozdělení**

Použití statistických testů je podmíněno předpoklady o typu rozdělení. Proto bychom měli před každým testem zjistit, z jakého rozdělení data pocházejí a podle toho pak zvolit přiměřený test.

***Ověření normality***

Ověření normality je nezbytný krok před každou analýzou jednorozměrných dat.

Proto pro testování, zda výběr pochází právě z normálního rozdělení, byly vyvinuty různé metody a testy. Některé z nich jsou: Shewhartovy regulační diagramy, indexy způsobilosti, grafické znázornění a vizuální posouzení (např. Q-Q graf, jádrové odhady hustoty, kruhový graf aj.), dále pak **testy založené na odhadu šikmosti a špičatosti**, **test dobré shody** (oba viz níže), dále A-D test (Anderson-Darling), R-J test (Ryan-Joiner), S-W test (Shapiro-Wilks), K-S test (Kolmogorov-Smirnov) a jiné.

Zde uvedeme **TEST ŠIKMOSTI A ŠPIČATOSTI** *(pro normální rozdělení)*

Normální rozdělení má oba koeficienty (koeficient šikmosti a standardizovaný koeficient špičatosti) rovny nule. Připomeňme (viz třetí předn.): šikmost: , (stand.) špičatost: .

Z daného výběru se tedy vypočítají odhady těchto koeficientů:

a ověří se významnost jejich odchylky od 0.

Pochází-li výběr z normálního rozdělení, tedy z N(*μ*, σ2), pak při velkém rozsahu výběru:

přibližné rozdělení *a*3 je N(0, D(*a*3)), kde D(*a*3) =

a přibližné rozdělení *a*4 je N(E(*a*4), D(*a*4)), kde E(*a*4) = – D(*a*4) =

Hypotéza normality se tedy zamítá, když nebo

Willcoxon:

2 Ranky

=rank(A1,$A$1:$B$13,1)

Suma ranku

=sum

n1,n2 je count

Expectation

=n1\*(n1+n2+1)/2

chyba

=sqrt(n1\*n2\*(n1+n2+1)/n1)

Stat

=(sum-expectation)/chyba

p-hodnota

1-norm.dist(Stat,0,1,1)

**TEST DOBRÉ SHODY** *(pro libovolné rozdělení, tedy jak normální, tak i jiné)*

***(též Pearsonův* χ2*-test)***

- je založen na skutečnosti, že náhodný výběr s multinomickým rozdělením (tj. vektor pozorovaných četností) lze transformovat na náhodnou veličinu (pivotovou statistiku), která má v případě platnosti nulové hypotézy asymptoticky rozdělení χ2.

***- postup:***

Testujme hypotézu, která tvrdí, že náhodný výběr pochází z konkrétního rozdělení. Test provedeme tak, že porovnáme četnosti naměřených hodnot (resp. - v případě spojité NV - četnosti dat v třídících intervalech) s četností teoretickou, tj. takovou, která by měla být v případě, že výběr opravdu pochází z daného rozdělení.

Je zřejmé, že čím je rozdíl mezi pozorovanými a teoretickými četnostmi větší, tím máme silnější důvod hypotézu zamítnout.

cestující letiště

P(X=5)=BINOM.DIST(5, 200, 0.02, FALSE)

P(X<7)=BINOM.DIST(6, 200, 0.02, TRUE)

1−P(X≤6)=1 - BINOM.DIST(6, 200, 0.02, TRUE)

Někdy je hypotetické rozdělení zadáno zcela přesně (i s parametry), někdy je zadán pouze typ rozdělení a parametry se musí odhadnout.

Označme pozorovanou četnost *j*-tého třídícího intervalu *nj*.

Teoretickou pravděpodobnost, že se při platnosti nulové hypotézy bude NV realizovat v *j*-tém třídícím intervalu označme *pj*. Je-li rozsah náhodného výběru *n*, pak teoretická četnost je *npj*.

H0: Náh. výběr pochází z daného rozdělení. H1: Neplatí H0.

Obor možných hodnot tvoří *r*-tříd. (resp. – v případě spojité NV – obor hodnot rozdělíme

na *r*-třídících intervalů).

Provedeme náhodný výběr *x*1, …, *xn* (o rozsahu *n*) a určíme četnosti tříd *nj*. (kontrola )

Není-li hypotetické rozdělení zadáno zcela přesně, odhadneme z náh. výběru nezadané parametry.

Určíme teoretické pravděpodobnosti *pj*, se kterými by NV měla nabývat dané hodnoty (resp.

hodnoty z daného intervalu).

Realizace pivotové statistiky: *r*0 = (kvantily budeme vybírat z (r – k – 1)

kde *r* je počet tříd. intervalů a *k* je počet odhadovaných parametrů )

Kritický obor je: W = (r – k – 1)

Pozn.: Pro tento test musí být splněna podmínka, že každá hodnota *npj* je rovna aspoň 5. V případě nesplnění této podmínky je potřeba některé intervaly (resp. varianty) sloučit, což vede ke ztrátě informace. Navíc taky je hodnota testové statistiky silně závislá na volbě třídících intervalů.

**B: Neparametrické testy**

Při použití parametrických testů musí být splněny určité předpoklady:

- výběr pochází z konkrétního rozdělení (my jsme probírali jen případy výběrů z normálního rozdělení)

Při větších rozsazích (*n* ≥ 30) v důsledku CLV není nesplnění tohoto předpokladu příliš na závadu.

- homogenita rozptylů

- intervalový, nebo poměrový charakter dat

piliny

P(X>19.5)=1 - (19.5 - 18) / (22 - 18)

P(X<20.5)=(20.5 - 18) / (22 - 18)

P(20≤X≤21.5)=(21.5 - 20) / (22 - 18)

Pokud tyto předpoklady nejsou splněny (např. výběry mají malé rozsahy a/nebo pocházejí z rozdělení, která jsou jiná, než normální), použijeme neparametrické testy.

V dalším se budeme zabývat neparametrickými testy založenými na pořadí hodnot (testy o mediánech) místo na hodnotách samotných. Z tohoto faktu plyne výhoda neparametrických testů, že eliminují vliv ojedinělých odlehlých, či extrémních hodnot.

Další výhodou neparametrických testů je, že jsou výpočetně jednodušší (počítačům je to ale jedno).

Naopak, nevýhodou neparametrických testů je, že oproti parametrickým testům jsou „slabší“, tj. nepravdivou hypotézu zamítají s menší pravděpodobností než parametrické testy.

**ZNAMÉNKOVÝ TEST**

- **jednovýběrový**: je obdobou jednovýběrového t-testu

Nechť X1, X2, …, X*n* je náhodný výběr ze spojitého rozdělení, x0,5 je medián tohoto rozdělení

a *m* je konstanta. Nulovou hypotézou je pak H0: x0,5 = *m*.

- **párový**: je obdobou párového t-testu

Nechť (X1, Y1), (X2, Y2), …, (X*n*, Y*n*) je náhodný výběr ze spojitého dvourozměrného rozdělení.

Testujeme pak rozdíl dvou mediánů, tedy H0: x0,5 – y0,5 = *m* (u párového testu nejčastěji *m* = 0).

Postup: - utvoříme rozdíly: pro **jednovýběrový** D*i* = X*i* – *m*

pro **párový** D*i* = (X*i* – Y*i*) – *m*

! Jsou-li některé rozdíly nulové, pak za *n* bereme jen počet nenulových rozdílů (ozn. *n*red)!

- označme S+ počet kladných rozdílů Di (tj. počet hodnot větších než medián). Při platnosti H0

by tento počet měl být (přibližně) stejný, jako počet

hodnot menších než medián.

Kritický obor pro S+ je: W = ,

kde *k*1 a *k*2 jsou tabulkové kritické hodnoty znaménkového testu pro dané *α* a *nred*.

● Asymptotická varianta: Pro velká *n* ( > 20) lze použít testovací statistiku:

*r*0 = (kvantily budeme vybírat z N(0,1))

Kritický obor pro *r*0 je: W =

============================

**Pořadí a průměrné pořadí**

Pořadí hodnot získáme očíslováním vzestupně uspořádaných hodnot.

Jsou-li některé hodnoty stejné, přiřadí se jim všem průměrné pořadí.

Např.: Datový soubor: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2

vzestupné uspořádání: 1,8 1,8 1,9 2 2 2,1 2,1 2,2 2,3 2,4

pořadí 1 2 3 4 5 6 7 8 9 10

(průměrné) pořadí 1,5 1,5 3 4,5 4,5 6,5 6,5 8 9 10

**WILCOXONŮV TEST**

*(jednovýběrový a párový je přesnější variantou znaménkových testů)*

- **jednovýběrový**: Nechť X1, X2, …, X*n* je náhodný výběr ze spojitého rozdělení s hustotou *f*(x), která je symetrická kolem mediánu, tj. *f*(x0,5 + x) = *f*(x0,5 – x) a *m* je konstanta. Nulovou hypotézou je pak H0: x0,5 = *m*.

(Není-li splněn předpoklad o symetrii *f*(x), lze použít znaménkový test.)

- **párový**: Nechť (X1, Y1), (X2, Y2), …, (X*n*, Y*n*) je náhodný výběr ze spojitého dvourozměrného rozdělení.

Testujeme pak rozdíl dvou mediánů, tedy H0: x0,5 – y0,5 = *m* (u párového testu se nejčastěji počítá *m* = 0).

Postup: - utvoříme rozdíly: pro **jednovýběrový** D*i* = X*i* – *m*

pro **párový** D*i* = (X*i* – Y*i*) – *m*

! Jsou-li některé rozdíly nulové, pak za *n* bereme jen počet nenulových rozdílů (ozn. *n*red)!

- absolutní hodnoty |D*i*| seřadíme vzestupně a stanovíme jejich průměrné pořadí

- označme: S+ = součet pořadí původně kladných rozdílů

- označme: S– = součet pořadí původně záporných rozdílů

- testová statistika *r*0 =

Kritický obor pro *r*0 je: W =

kde je tabulková kritická hodnota pro jednovýběrový Wilcoxonův test.

● Asymptotická varianta: Pro velká *n* ( ≥ 30) lze použít testovací statistiku:

*r*0 = (kvantily budeme vybírat z N(0,1))

Kritický obor pro *r*0 je: W =

**Testování pomocí Spearmanova korelačního koeficientu** *(pro dvourozměrný výběr)*

Určíme zvlášť pořadí *Ri* hodnot náhodné veličiny X1 a zvlášť pořadí *Qi* hodnot náhodné veličiny X2

Spearmanův koeficient pořadové korelace *r*S =

Kritický obor pro *rS* je: W =

kde je tabulková kritická hodnota

● Pro *n* > 20 lze použít (asymptotickou) testovou statistiku *r*0 =

(kvantily budeme vybírat z *t*(n – 2))

Kritický obor pro *r*0 je: W =

● Pro *n* > 30 lze dokonce použít statistiku *r*0 = (kvantily budeme vybírat z N(0, 1))

Kritický obor pro *r*0 je: W =

↓

↓

- **dvouvýběrový Wilcoxonův test**

Nechť X11, X12, …, X1*n*1 o rozsahu *n*1 a X21, X22, …, X2*n*2 o rozsahu *n*2 jsou dva nezávislé náhodné výběry ze dvou spojitých rozdělení, jejichž distribuční funkce se mohou lišit pouze posunutím. Chceme testovat hypotézu, že oba soubory mají stejné rozdělení (jde tedy o hypotézu „silnější“, než u analogických parametrických testů, kde se testovala pouze shoda středních hodnot).

Postup: - všech *n*1 + *n*2 hodnot uspořádáme vzestupně a přiřadíme jim tzv. sdružené průměrné pořadí

- označme T1 součet pořadí pro hodnoty X1*i* ve sdruženém prům. pořadí

- označme T2 součet pořadí pro hodnoty X2*j* ve sdruženém prům. pořadí

- vypočítáme statistiky: U1 = U2 =

- testová statistika *r*0 = min{U1, U2}

Kritický obor pro *r*0 je: W = kde je tabulková kritická hodnota

pro dvouvýběrový Wilcoxonův test.

● Asymptotická varianta: Pro velká *n*1, *n*2 ( > 30) lze použít testovací statistiku:

*r*0 = (kvantily budeme vybírat z N(0,1))

Kritický obor pro *r*0 je: W =