Analýza rozptylu v datech s opakovanými měřeními

Roman Kroupa*

1. dubna 1994

Analýza rozptylu je statistická metoda užívaná v biologických, průmyslových a jiných experimentech. Touto metodou zjišťujeme vliv nějakého faktoru na střední hodnotu nějaké náhodné veličiny. Přitom předpokládáme, že každá z těchto náhodných veličin je z normálního rozdělení a že jejich střední hodnoty sice mohou být různé, ale jejich rozptyl musí být stejný. Základy této metody jsou dostatečně popsány v literatuře [1,7].

Analýza rozptylu v datech s opakovanými měřeními je speciální případ, kdy měření provádíme na určitých objektech (případech) opakovaně, takže sledujeme také jejich vývoj (změnu).

Tato statistická metoda je zúžením obecné analýzy rozptylu a zachycuje jakousi korelaci mezi měřeními. Jde o experimenty, ve kterých je případ, na kterém je prováděno měření, křížen minimálně s jedním faktorem, který budeme nazývat opakovací. Faktory, do nichž jsou případy vloženy nazveme skupinové. Analýza rozptylu s opakovaným měřením řeší experimenty tak, že případ považuje za další faktor a to faktor náhodný. Faktor opakovací je potom považován za pevný. Jde tedy vždy o smíšený model s jedním pozorováním v buňce.

Jednofaktorový experiment s opakovanými měřeními

Jde o nejjednodušší uspořádání experimentu analýzy rozptylu s opakovaným měřením a proto má také toto uspořádání asi nejčastější uplatnění. Uveďme si jednoduchý příklad.

Zkoumáme, zda má určitá didaktická metoda vliv na úspěšnost žáka v didaktickém testu. (např. nás zajímá, zda při systematickém provádění skupinového vyučování se zlepšuje schopnost žáků řešit samostatně různé problémy). Tzn. že po určitou dobu budeme na skupinu měřených případů působit daným faktorem (budeme provádět skupinové vyučování) a průběžně budeme měřit žákovu úspěšnost. Výsledky měření můžeme zapsat do tabulky s R řádky (počet žáků) a C sloupci (počet měření). Tzn. že hodnota x_{rc} byla naměřena r-tému žákovi při c-tém testování.

Testové skóre každého žáka při každém měření je náhodná veličina, která je ovlivněna výkonností každého žáka a také působením dané didaktické metody. Tato veličina se bude řídit modelem:

$$x_{rc} = \mu + A_r + B_c + (AB)_{rc} + e_{rc}$$

případy	1	2	3	...	C
1	x_{11}	x_{12}	x_{13}	...	x_{1C}
2	x_{21}	x_{22}	x_{23}	...	x_{2C}
⋮	⋮	⋮	⋮	⋮	⋮
R	x_{R1}	x_{R2}	x_{R3}	...	x_{RC}

A_r $N(0, \sigma_r^2)$ náhodná veličina - vliv úrovně (výkonnosti) žáka

B_c konstanta - vliv didaktické metody, ukazuje, zda se zvyšuje výkonnost celé skupiny

$(AB)_{rc}$ $N(0, \sigma_{rc}^2)$ náhodná veličina - ukazuje, zda je účinnost metody pro jednotlivé žáky stejná

e_{rc} $N(0, \sigma_e^2)$ náhodná veličina - zahrnuje náhodné vlivy, chyby měření ...

Odhady parametrů získáme metodou nejmenších čtverců. Pro reparametrizační rovnice

$$\sum_r A_r = \sum_c B_c = 0$$

$$\sum_r (AB)_{rc} = 0 \qquad \text{pro každé } c = 1, \ldots, C$$

$$\sum_c (AB)_{rc} = 0 \qquad \text{pro každé } r = 1, \ldots, R$$

získáme pro odhady parametrů vztahy

$$\mu = x_{..}$$
$$A_r = x_{r.} - x_{..}$$
$$B_c = x_{.c} - x_{..}$$
$$(AB)_{rc} = x_{rc} - x_{r.} - x_{.c} + x_{..}$$

Pozn.: Tečka v indexu znamená průměr počítaný přes index nahrazený tečkou.

Rozklad součtu čtverců

Součet čtverců odchylek od střední hodnoty můžeme rozdělit na dvě nezávislé složky. První (S_r .. součet čtverců odchylek řádkových průměrů od průměru celkového) vyjadřuje variabilitu mezi případy, druhá (součet čtverců odchylek od řádkových průměrů) vyjadřuje variabilitu uvnitř případu. Tuto druhou složku ještě můžeme rozdělit na dvě části podle jejich zdroje. První odpovídá vlivu sloupců (S_c .. vliv opakovaných měření na průměrnou hodnotu všech případů) a druhá S_{rc} vlivu interakce mezi případy a opakovanými měřeními.

Zdroj rozptylu	Součet čtverců	Stupně volnosti
Mezi případy A	$S_r = \sum_r \sum_c (x_{r.} - x_{..})^2$	$R-1$
Uvnitř případu	$\sum_r \sum_c (x_{rc} - x_{r.})^2$	
B	$S_c = \sum_r \sum_c (x_{.c} - x_{..})^2$	$C-1$
(AB)	$S_{rc} = \sum_r \sum_c (x_{rc} - x_{r.} - x_{.c} + x_{..})^2$	$(R-1)*(C-1)$
Celkový	$\sum_r \sum_c (x_{rc} - x_{..})^2$	$R*C-1$

Zdroj rozptylu	Střední hodnota průměrného čtverce
Mezi případy A	$s_r = \sigma_e^2 + C\sigma_r^2$
Uvnitř případu	
B	$s_c = \sigma_e^2 + \sigma_r^2 + \sigma_{rc}^2 + \frac{R\sum_c B_c^2}{C-1}$
(AB)	$s_{rc} = \sigma_e^2 + \sigma_r^2 + \sigma_{rc}^2$

Základem analýzy rozptylu je srovnání odhadů jednotlivých složek rozptylu. Proto je třeba určit střední hodnotu každého průměrného čtverce. (Průměrný čtverec získáme jako podíl součtu čtverců a odpovídajícího počtu stupňů volnosti.)

Vyjdeme-li z modelu pro náhodnou veličinu x_{rc} můžeme říci, že je z rozdělení

$$x_{rc} \sim N(\mu + B_c, \sigma_e^2 + \sigma_r^2 + \sigma_{rc}^2)$$

Pro veličinu x_{rc} dále platí

$$x_{rc} = x_{r.} + (x_{.c} - x_{..}) + (x_{rc} - x_{r.} - x_{.c} + x_{..})$$

Způsobem popsaným v [7] potom můžeme určit střední hodnoty průměrných čtverců s_c a s_{rc}.

Náhodná veličina $x_{r.}$ je z rozdělení

$$x_{r.} \sim N(\mu, \sigma_e^2/C + \sigma_r^2)$$

a platí pro ni

$$x_{r.} = x_{..} + (x_{r.} - x_{..})$$

Stejně jako v předcházejícím případě určíme střední hodnotu průměrného čtverce s_r.

Tabulka středních hodnot průměrných čtverců jednotlivých složek rozptylu má následující tvar: Porovnáním jednotlivých středních hodnot průměrných čtverců je na první pohled zřejmé, že správný chybový výraz pro testování nulové hypotézy H_0: $B_1 = B_2 = \ldots = B_C = 0$ je střední hodnota průměrného čtverce příslušející interakci (AB). Oba tyto výrazy se totiž liší pouze o $\frac{R\sum_c B_c^2}{C-1}$, který je v případě platnosti nulové hypotézy roven nule a podíl s_c/s_{rc} má Fisherovo rozdělení. Tzn. že platí

$$F_H = (R-1) * \frac{S_C}{S_{RC}}$$

Je-li tato hodnota větší než teoretická, potom zamítáme nulovou hypotézu na dané hladině významnosti.

úrovně faktoru L	1			2		
úrovně faktoru C	1	2	3	1	2	3
případy						
1	x_{111}	x_{121}	x_{131}	x_{112}	x_{122}	x_{132}
2	x_{211}	x_{221}	x_{231}	x_{212}	x_{222}	x_{232}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
R	x_{R11}	x_{R21}	x_{R31}	x_{R12}	x_{R22}	x_{R32}

Dvoufaktorový experiment s opakovanými měřeními

Jde o uspořádání experimentu s dvěma opakovacími faktory a proto také není z hlediska teorie moc složité a v „obecné" analýze rozptylu je představováno třífaktorovým tříděním, ve kterém jsou všechny faktory navzájem křížené a proto je také částečně popsáno v literatuře [1]. V našem případě jde o speciální případ s jedním pozorováním pro každou kombinaci faktorů (jako vždy v analýze rozptylu s opakovaným měřením) s jedním faktorem náhodným a dvěmi pevnými. Také tento případ má časté uplatnění.

Představme si, že během léčení je pacientům měřena teplota. Tato veličina se samozřejmě může během léčení měnit (což chceme zjistit), ale také víme, že tělesná teplota kolísá během dne i u zdravého člověka. Budeme-li provádět u skupiny pacientů, kteří jsou ve stejné fázi onemocnění, měření několik dní po sobě, zjistíme například, že průměrná teplota celé skupiny se v průběhu léčení významně snižuje, ale naopak během dne se zvyšuje.

Tímto způsobem uspořádáme především experimenty z biologie, medicíny, psychologie a v jiných oborech, kde výsledky měření závisí na tzv. biorytmech (viz uvedený příklad). Velice dobré využití má tato metoda i ve sportu, protože zde je jeden ze základních principů princip „zatížení ve vlnách". Tzn. že intenzita zatížení se mění v jednak průběhu každé tréninkové jednotky, jednak během mikrocyklů (většinou jeden týden), mezocyklů atd. Proto také odpovídajícím způsobem kolísá výkonnost každého sportovce.

Výsledky měření zapisujeme do tabulky následujícím způsobem.

Kdyby šlo o případ měření teploty každý den ráno, v poledne a večer, denní doba by byla prvním opakovacím faktorem a dny druhým, potom x_{324} je teplota třetího pacienta v poledne čtvrtého dne.

O náhodné veličině x_{rcl} pak můžeme říci, že se chová podle modelu

$$x_{rcl} = \mu + A_r + B_c + D_l + (AB)_{rc} + (AD)_{rl} + (BD)_{cl} + (ABD)_{rcl} + e_{rcl}$$

A_r	$N\ (0, \sigma_r^2)$ náhodná veličina
$(AB)_{rc}$	$N\ (0, \sigma_{rc}^2)$ náhodná veličina
$(AD)_{rl}$	$N\ (0, \sigma_{rl}^2)$ náhodná veličina
$(ABD)_{rcl}$	$N\ (0, \sigma_{rcl}^2)$ náhodná veličina
e_{rc}	$N\ (0, \sigma_e^2)$ náhodná veličina
B_c, D_l, $(BD)_{cl}$	jsou konstanty

V tomto případě rozkládáme celkovou variabilitu opět na variabilitu uvnitř případu a mezi případy. Variabilitu mezi případy dělíme dále na tři části. Jedna souvisí s vlivem prvního opakovacího faktoru, druhá s vlivem druhého faktoru a třetí s vlivem interakce mezi těmito faktory. Každá tato část se dále dělí na dvě složky. První odpovídá pouze vlivu

zdroj rozptylu	součet čtverců	stupně volnosti
R případy	$S_R = \sum\sum\sum(x_{r..} - x_{...})^2$	$R-1$
C faktor	$S_C = \sum\sum\sum(x_{.c.} - x_{...})^2$	$C-1$
L faktor	$S_L = \sum\sum\sum(x_{..l} - x_{...})^2$	$L-1$
RC inter.	$S_{RC} = \sum\sum\sum(x_{rc.} - x_{r..} - x_{.c.} + x_{...})^2$	$(R-1)(C-1)$
RL inter.	$S_{RL} = \sum\sum\sum(x_{r.l} - x_{r..} - x_{..l} + x_{...})^2$	$(R-1)(L-1)$
CL inter.	$S_{CL} = \sum\sum\sum(x_{.cl} - x_{.c.} - x_{..l} + x_{...})^2$	$(C-1)(L-1)$
RCL inter.	$S_{RCL} = (x_{rcl} - x_{rc.} - x_{r.l} - x_{.cl} + x_{r..} + x_{.c.} + x_{..l} - x_{...})^2$	$(R-1)(C-1)(L-1)$

tohoto faktoru a druhá vlivu interakce mezi daným faktorem a případy. Součty čtverců odlišujeme opět indexem. Tzn. že S_R, S_C, S_L, S_{CL}, S_{RC}, S_{RL}, S_{RCL} jsou součty čtverců příslušející po řadě případům a dále faktorům opakovacím C a L, interakci opakovacích faktorů, interakci případů s faktory C a L a interakci mezi případy a oběma opakovacími faktory. Pro testování významnosti vlivu opakovacích faktorů a jejich interakce je třeba použít těchto poměrů průměrných čtverců:

$$s_c/s_{rc}, \qquad s_l/s_{rl}, \qquad s_{cl}/s_{rcl}$$

Dvoufaktorový experiment s opakovanými měřeními na jednom faktoru

Tento experiment je již složitější než předcházející (smíšený model s jedním vloženým faktorem) a proto může být interpretace výsledků různá podle uspořádání experimentu.

Použijeme např. pro každou skupinu sportovců jinou tréninkovou metodu a opakovaně budeme testovat některou pohybovou schopnost. Tabulka analýzy rozptylu pak má následující tvar:

Náhodná veličina x_{src} se řídí modelem:

$$x_{src} = \mu + A_{sr} + R_r + C_c + (AC)_{src} + (RC)_{rc} + e_{src}$$

A_{sr}	$N(0,\sigma_s^2)$	náhodná veličina
$(AC)_{src}$	$N(0,\sigma_{sc}^2)$	náhodná veličina
e_{src}	$N(0,\sigma_e^2)$	náhodná veličina
R_r, C_c, $(RC)_{rc}$.		jsou konstanty

Součty čtverců mají v podstatě týž význam, jako při jednofaktorové analýze rozptylu s opakovaným měřením. Součet čtverců $S_{s/r}$ však nevyjadřuje vliv interakcí mezi případy

R skupiny	S případy	1	2	3
1	1	x_{111}	x_{112}	x_{113}
	2	x_{211}	x_{212}	x_{213}
	⋮	⋮	⋮	⋮
	9	x_{911}	x_{912}	x_{913}
2	1	x_{121}	x_{122}	x_{123}
	2	x_{221}	x_{222}	x_{223}
	⋮	⋮	⋮	⋮
	7	x_{721}	x_{722}	x_{723}

Zdroj rozptylu	Součet čtverců	Stupně volnosti
Mezi případy	$C\sum_s\sum_r(x_{sr.}-x_{...})^2$	$n-1$
R skupiny	$S_r = C\sum_r(x_{.r.}-x_{...})^2 n_r$	$R-1$
S/R případy	$S_{s/r} = C\sum_s\sum_r(x_{sr.}-x_{.r.})^2$	$n-R$
Uvnitř případu	$\sum_s\sum_r\sum_c(x_{src}-x_{sr.})^2$	$n-1$
C opak. měření	$S_c = n\sum_c(x_{..c}-x_{...})^2$	$C-1$
$R\times C$ interakce	$S_{rc} = \sum_r\sum_c(x_{.rc}-x_{.r.}-x_{..c}+x_{...})^2$	$(R-1)(C-1)$
$C\times S/R$ interakce	$S_{sc} = \sum_s\sum_r\sum_c(x_{src}-x_{sr.}-x_{.rc}+x_{.r.})^2$	$(n-R)(C-1)$
Celkový	$S_T = \sum_s\sum_r\sum_c(x_{src}-x_{...})^2$	$n*C-1$

Zdroj rozptylu	Střední hodnota průměrného čtverce
Mezi případy	
R skupiny	$s_r = \sigma_e^2 + C\sigma_s^2 + \frac{C\sum_r R_r^2 n_r}{R-1}$
S/R případy	$s_{s/r} = \sigma_e^2 + C\sigma_s^2$
Uvnitř případu	
C opak. měření	$s_c = \sigma_e^2 + \sigma_s^2 + \sigma_{sc}^2 + \frac{\sum_c C_c^2}{C-1}$
$R \times C$ interakce	$s_{rc} = \sigma_e^2 + \sigma_s^2 + \sigma_s c^2 + \frac{\sum_c \sum_r (RC)_{rc}^2 n_r}{(R-1)(C-1)}$
$C \times S/R$ interakce	$s_{sc} = \sigma_e^2 + \sigma_s^2 + \sigma_{sc}^2$

a skupinami, ale pouze vliv případů na celkovou variabilitu. Je to proto, že případy jsou vloženy do skupin, tzn. že měření x_{111} a x_{121} nenáleží témuž případu. Totéž platí o interakci $C \times S/R$. Jde o interakci výhradně mezi případy a opakovanými měřeními, nikoliv skupinami.

Srovnáním zjišťujeme, že správný chybový výraz pro testování rozdílů mezi skupinami je střední čtverec $s_{s/r}$, kdežto sloupcové efekty (opakovací faktor) a interakce mezi skupinami a opakováním testujeme srovnáním se středním čtvercem s_{sc}. Tzn., že pro testování vlivu jednotlivých faktorů použijeme tyto poměry:

$$s_r/s_{s/r}, \qquad s_c/s_{sc}, \qquad s_{rc}/s_{sc}.$$

Srovnání výsledků

Naměřené hodnoty popsaných experimentů můžeme zpracovávat několika způsoby. V našich podmínkách je asi nejčastější zpracování dat pomocí statistických paketů SOLO a BMDP. Oba tyto produkty jsou však značně náročné jednak na technické vybavení, ale také na obsluhu. Navíc se mnohdy provozují bez potřebného popisu teoretického pozadí analýzy rozptylu s opakovanými měřeními, takže uživatel ani neví, co vlastně počítá. Proto byl také vytvořen program ANOVA, který zpracovává data popsaných tří experimentů a dále třífaktorový experiment s dvěma skupinovými faktory křížením i s jedním skupinovým faktorem vloženým do druhého (oba skupinové faktory jsou pevné). Základem tohoto programu je prostředí ESTAT [8]. Ovládání programu je ve formě nabídek, přičemž program využívá vlastní tabulkový editor. V mnoha případech pak je možné získat nápovědu. Při testování správnosti programu pak bylo provedeno srovnání s výsledky:
a) řešených úloh práce [3],
b) získanými systémem SOLO [9],
c) získanými systémem BMDP , program 2V [2].

a) Jelikož se v teoretické části vycházelo především z práce [3], bylo provedeno srovnání hlavně s výsledky řešených příkladů této práce (str. 350, 353, 361). Takto jsme postupovali u těchto návrhu experimentu:

- jednofaktorová analýza rozptylu s opakovanými měřeními,
- dvoufaktorová analýza rozptylu s opakovanými měřeními,

- dvoufaktorová analýza rozptylu s opakovanými měřeními na jednom faktoru pro případ vyváženého třídění.

V těchto případech se všechny hodnoty v tabulce výsledků analýzy rozptylu podle očekávání shodovaly s přesností na 0.0001. Takové odchylky je možné vysvětlit zaokrouhlováním při výpočtech.

b) Produktem SOLO byly kontrolovány výsledky u dvoufaktorové analýzy rozptylu s opakovanými měřeními na jednom faktoru. Výsledná tabulka analýzy rozptylu i částečné průměry se shodují zhruba se stejnou přesností jako v případě a).

c) Produktem BMDP byly kontrolovány výsledky u dvoufaktorové analýzy rozptylu s opakovaným měřením na jednom faktoru a u třífaktorové analýzy rozptylu s dvěma kříženými skupinovými faktory. Ve všech případech se velice přesně shodovaly částečné průměry a směrodatné odchylky pro jednotlivé úrovně faktorů. Částečně se však již liší tabulky analýzy rozptylu. Ve všech uspořádáních jsou shodné chybové výrazy a počty skupin volnosti. Z ostatních výrazů jsou shodné pouze součty čtverců (a samozřejmě průměrné čtverce) skupinového faktoru a interakce skupina - opakování u dvoufaktorové analýzy rozptylu. Ostatní součty čtverců (i průměrné čtverce, testové statistiky F.exp. a pravděpodobnosti) se na různé úrovni odlišují. Tyto odchylky již není možné vysvětlit zaokrouhlováním při výpočtech, ale musí být způsobeny tím, že BMDP předpokládá jiné uspořádání experimentu než program ANOVA. K tomuto tvrzení nás opravňují i rozdílné výsledky dvoufaktorové analýzy rozptylu s opakováním na jednom faktoru oproti literatuře [3] nebo výsledkům získaným produktem SOLO, které se naopak s výsledky programu ANOVA shodují.

Výsledky třífaktorového experimentu s jedním skupinovým faktorem vloženým do druhého nebyly nijak konfrontovány, neboť žádný nám dostupný statistický program tento návrh experimentu nedovoluje.

Program ANOVA i podrobné odvození výsledků všech pěti návrhů experimentů [10] je k dispizici u ing. Tvrdíka, CSc. na katedře informatiky a počítačů PřF OU OSTRAVA.

Literatura

[1] Anděl, J. : *Matematická statistika.* SNTL, Praha, 1978.

[2] Dixon, W. J. : *BMDP Statistical Software Manual.* University of California Press, 1990.

[3] Hušková, M. – Dupačová, J. : *Analýza rozptylu.* SPN, Praha, 1978.

[4] Křivý, I. : *Základy matematické statistiky.* Pedagogická fakulta v Ostravě, 1985.

[5] Lamoš, F. – Potocký, J. : *Pravděpodobost a matematická statistika.* ALFA, Bratislava, 1989.

[6] Likeš, J. : *Navrhování průmyslových experimentů.* SNTL, Praha, 1968.

[7] Žváček, J. – Řezanková, H. : *ESTAT, statistické programovací prostředí v Turbo Pascalu.* Sborník Robust 90, JČSMF, 1990.

[8] Kolektiv : *Návod k použití programového produkt SOLO.* UK Praha, 1992.

[9] Kroupa, R. : *Analýza rozptylu v datech s opakovanými měřeními.* diplomová práce PřF OU Ostrava, 1993.