Zpracování vědecko výzkumných dat

trubka Znojil

zpracoval Aleš Křenek

únor – duben 1995

Obsah

1 Základní pojmy

První dojem o povaze naměřených dat lze získat z histogramu s počtem sloupců a měřítkem přiměřeným datovému souboru (o to se většinou postará software). Základní spočitatelné hodnoty vypovídající o datech jsou

1. *α-kvantil* pro $0 < \alpha < 1$ je taková hodnota měřené veličiny, že 100α % naměřených hodnot v daném souboru je menších.
2. *Medián* je 0.5-kvantil, tj. prostřední z naměřených hodnot. Při symetrickém rozložení splývá s průměrem.
3. *Kvartily* jsou 0.25- a 0.75-kvantily.

Při zpracování jde většinou o to získat nějakou transformací z naměřených dat alespoň přibližně normální rozložení; pak se tomu dá věřit. O tom, že je všechno špatně, se lze přesvědčit metodou *hradeb*. Naneseme na obě strany od mediánu nějaký násobek (1.5, 2.5, 3.5...) vzdálenosti mezi kvartily a pokud do tohoto intervalu padnou všechny naměřené hodnoty, rozdělení nelze považovat za normální. Další možností je *Kolgomorovův test,* naměřenými daty se proloží co nejlépe normální rozložení a měří se maximální odchylka.

2 Momenty a rozdělení

Propracovanější hodnoty, které něco řeknou jsou *momenty*

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - p)^k$$

tzv. k-tý moment vůči hodnotě p (pro $k = 1$, $p = 0$ je to průměr) a

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \qquad \text{kde } \bar{x} \text{ je průměr}$$

tzv. *k-tý centrální moment.* Vychází $\mu_1 = 0$ a standardně značíme $s^2 = \mu_2$, kde s je standardní odchylka. To ovšem pouze v kontextu momentů. Jedná-li se o odhad přesnosti měření, používáme $s^2 = 1/(n-1)\sum(x_i - p)^2$ čistě z toho důvodu, že rozptyl má o jeden stupeň volnosti méně, nemá smysl mluvit o rozptylu u jedné naměřené hodnoty.

Normovanými centrálními momenty nazýváme čísla $\nu_k = \mu_k/s^k$ (v tomto kontextu tedy $s = \sqrt{\mu_2}$). Vychází $\nu_1 = 0$, $\nu_2 = 1$, smysl mají až ν_3, tzv. *asymetrie,* a ν_4, tzv. *špičatost.* Je-li $\nu_3 > 0$, znamená to, že průměr je větší než medián, a tedy histogram je oproti normálnímu rozložení deformován doleva, pro $\nu_3 < 0$ naopak.

Pro normální rozdělení vychází $\nu_4 = 3$, zavádí se *excess* (přebytek) jako $\nu_4 - 3$. Vypovídá jednak o skutečné „špičatosti" křivky hustoty rozložení v maximu (pro rovnoměrné rozdělení vychází -1.2, naopak pro Laplaceovo (symetrické záporné exponenciální) $+3$, ale hlavně o rychlosti konvergence k 0 v extrémech. Proto mluvíme o rozděleních *s těžkými konci* pro kladný excess, analogicky *s lehkými konci* pro záporný.

Případ velmi nepříjemný je Cauchyho rozdělení, speciální případ studentova. Vznikne například jako $(x-\bar{x})/(y-\bar{y})$, kde x a y jsou normální. Pro hodnoty blízké průměru (a těch je pro normální rozdělení nemálo) dostáváme limitu typu $0/0$ a ta může nabývat celkem libovolných hodnot. Cauchyho rozdělení má tedy velmi těžké konce.

Jsou-li naměřená data ve skupinách, má smysl vynést rozptyl, asymetrii a špičatost do grafu v závislosti na průměru skupiny. Lze tak odhadnout závislost těchto veličin na průměru. V praxi je velmi časté, že např. rozptyl je tolik a

tolik procent naměřené hodnoty, celkové rozdělení je pak asymetrické, ale po logaritmické transformaci (tj. místo x uvažujeme $\ln x$) už ano. Jindy může pomoci odmocninová transformace, případně i nějaká složitější. Na druhé straně není dobré transformace přehánět, nevhodným použitím může dojít ke značnému zkreslení. Pokud už není patrná žádná závislost, lze hovořit o *homogenních datech* a je možné aplikovat základní testovací metoty, které fungují právě na data normálně rozdělená.

3 Testovací kritéria

Při experimentu získáme obvykle dvě sady dat. Měření nějakého faktoru před provedením experimentu (tzv. kontrolní data) a po provedení (experimentální data), resp. po dvou různých experimentech (např. aplikace nějakého léku, případně dvou různých léků). Na základě statistiky můžeme z těchto dat říci, že s takovou a takovou pravděpodobností měl experiment na měřenou veličinu takový vliv. Standardním postupem, při stejném počtu měření v obou sadách, je sestavení náhodných párů a vypočtení rozdílu. Vypovídajícím kritériem je pak číslo, tzv. t-test

$$t = \frac{\bar{y}}{s/\sqrt{n}}$$

kde y jsou právě zmíněné rozdíly, s je standardní odchylka y (tentokrát s $n-1$), a n počet měření. Udává práh pravděpodobnosti, se kterou můžeme tvrdit, že efekt experimentu, který jsme zjistili (tedy na zmíněném příkladě po aplikaci nového léku zemře více pacientů), nenastal náhodou, ale je zákonitý.

Díky omezeným možnostem statistiky se můžeme dostat do takovýchto problémů

1. Prohlásíme, že experiment měl kýžený efekt (s pravděpodobností např. 95%), ale není to ve skutečnosti pravda, právě kvůli zbývajícím 5%.

2. Ve skutečnosti daný efekt nastává, ale ze statistiky zjistíme, že nikoli.

Chyba druhého druhu je způsobena příliš vysoko nasazeným požadavkem na jistotu odpovědi vzhledem k přesnosti měření. Předexperimentem lze odhadnout tzv. *sílu testu.* Je nutné stanovit, jaký rozdíl v naměřené hodnotě hledáme a přispůsobit buď požadovanou jistotu anebo počet měření.

4 Optimalizace

Optimalizační úlohy ve smyslu statistiky znamenají nalézt nějakou teoretickou, většinou analyticky jednoduše vyjádřenou funkci, která se dostatečně dobře blíží experimentálním hodnotám. Pro jednoduchost předpokládejme funkci vracející jednu hodnotu. Lze ji vyjádřit jako

$$y = F(x_1, \ldots, x_n, q_1, \ldots, q_m)$$

kde vektor x_i jsou nezávislé proměnné (hodnoty zkoumaných faktorů) a q_j parametry analytického vyjádření funkce F, které hledáme. Optimalizace znamená nalézt minimum výrazu

$$\sum_j \left| F(x_1^j, \ldots, x_n^j, q_1, \ldots, q_m) - y^j \right|^c$$

– Protože se psal s "w", byl pan Swoboda rakušák jako dělo

kde x_i^j je hodnota i-tého faktoru a y^j experimentální výsledek pro j-té měření, pro každý parametr q. Konstanta c udává řád optimalizace, pro $c = 1$ hovoříme o lineární optimalizaci, která sice nedává nejlepší přiblížení, ale je méně citlivá na velké chyby, případ $c = 2$ je v praxi nejpoužívanější tzv. metoda *nejmenších čtverců.* Počítá se pochopitelně položením parciální derivace podle jednotlivých q rovné nule, tedy systém

$$\frac{\partial \left(\sum_j F\left(x_1^j, \ldots, x_n^j, q_1, \ldots, q_m\right) - y^j \right)^2 \right)}{\partial q_i} = 0$$

Dostaneme tak systém rovnic v proměnných q. Má-li mít celá metoda smysl, je nutné, aby počet měření byl podstatně vyšší, než počet parametrů, jinak se potýkáme se špatně definovanými veličinami.

Nejčastěji používané optimalizační (nebo také regresní) funkce jsou polynomiální, logaritmické, exponenciální a goniometrické.

5 Analýza variance

Představme si, že provádíme výzkum, kolik škopků vypaří vybraná reprezentativní skupina respondentů za večer. Na výsledek tohoto experimentu může mít zcela jistě vliv mnoho faktorů, například kde se nachází experimentální hospoda, kdy je zavírací hodina, jaké pivo se tam točí, kolik stojí, který den v týdnu experiment provedeme atd. Zajímá nás, které z těchto faktorů jsou statisticky významné, případně jakým způsobem.

Z takto získaných měření získáme obecně multidimenzionální tabulku, pro názornost budeme uvažovat dále jen dva faktory. Hodnotu jednoho prvku tabulky (resp. průměrnou hodnotu, je-li měření pro danou kombinaci faktorů více) můžeme, s ohledem na transformace, vyjádřit jako

$$x_{ij} = X + x_i' + x_j'' + x_{ij}'''$$

kde X je nějaká konstantní složka, x_i' řádkový faktor, x_j'' sloupcový faktor a x_{ij}''' interakce i-tého řádku a j-tého sloupce, tzv. *synergetický efekt,* samotná znalost působení jednotlivých faktorů ještě nemusí vypovídat nic o působení jejich kombinace (když pařím škopky, jsem ožralej, když piju mlíko, je mi špatně, ale pokud to smíchám, je to mnohem horší, než bych mohl z dílčích efektů očekávat).

V případě pouze jednoho pokusu pro danou kombinaci faktorů nelze o interakci vůbec uvažovat, nedokázali bychom rozhodnout, zda se jedná o interakci nebo o chybu měření. Při obecně n faktorech lze v tomto případě mluvit pouze o interakci nanejvýš $n - 1$ faktorů.

Nahradíme-li naměřené hodnoty rozdílem od průměru celé tabulky, zbavíme se konstantní složky X. Dále spočítáme průměry po řádcích a po sloupcích, z nich potom zbytkové rozptyly v řádcích a sloupcích. Nejsou-li data homogenní, tj. rozptyly se příliš liší, je nutné aplikovat nějakou vhodnou transformaci a počítat od začátku. Rozptyly porovnáme s vlivem jednotlivých faktorů (pozná se z průměrných hodnot pro řádky a sloupce), a tak zjistíme, který je významný. K takovému porovnání slouží například

- *Barletův test.* Vyžaduje alespoň 6 hodnot v políčku, jinak je velmi citlivý na hodnoty třeba náhodně blízké.
- *Leveneův test,* pro málo měřených hodnot podstatně robustnější.

Pokud víme dopředu, že interakce je nemožná, a přesto nějaká vyšla, je vhodné opět se nad daty zamyslet a zkusit aplikovat vhodnou transformaci.

– Statistika je přesné počítání s nepřesnými čísly

Jakmile zjistíme, že vliv nějakého faktoru je významný, lze se ptát konkrétněji, tedy jaké rozdíly jsou mezi jednotlivými hodnotami. Problematický v tuto chvíli je počet různých porovnání a úroveň pravděpodobnosti, při např. 6 řádcích tj. 15 porovnáních už zřejmě nějaká závislost vyjde na alespoň 95%, i když to nemusí být pravda. Proto je nutné aplikovat nějaké restrikce, např. *Holmův postup.* Seřadíme t-testem získané pravděpodobnosti, že dané porovnání vyšlo náhodou, od nejmenší k největším. Stanovíme přípustný práh pravděpodobnosti chyby p a nejmenší pravděpodobnost porovnáváme s p/n, kde n je počet porovnání. Pokud vyjde pravděpodobnost chyby menší, prohlásíme rozdíl za významný a postupujeme dále. Porovnáváme druhou hodnotu s $p/(n-1)$ atd. dokud nenarazíme na opačnou nerovnost. Tím prohlásíme všechny další porovnání za statisticky nevýznamná.

Pokud je z nějakého důvodu některé porovnání nezajímavé, ve zmíněném postupu ho neuvažujeme, kriteria na ostatní jsou tak trochu mírnější, přitom je postup statisticky stále korektní.

Může se stát, že některé údaje v tabulce chybí. Z klasického hlediska se nelze hnout z místa, ale regresními metodami (kapitola 4) lze spočítat teoretickou hodnotu, kterou za chybějící experimentální dosadíme.

6 Zpětná analýza variance

V této části se budeme zabývat metodou „co se nehodí, to se zahodí[1]". Princip je takový, že teoretický model vytvoříme z více možných vlivů a na základě analýzy dat vylučujeme potom ty nepodstatné. Narozdíl od klasické analýzy variance je tato metoda podstatně robustnější vůči chybějícím údajům a může bez větších problémů zahrnout i vlivy spojité veličiny. Na druhé straně může být nebezpečná co se týče homogenity dat a nedostatečné definovanosti, vždy je potřeba dát pozor na počet stupňů volnosti počítané veličiny. Opět to znamená, že jakkoli zakuklený počet koeficientů v daném teoretickém modelu musí být podstatně menší než počet získaných experimentálních hodnot.

Počítejme například závislost výnosů chmele (pokus lze pochopitelně provést i pro ječmen) na srážkách v daném roce. Výnos modelujeme funkcí

$$y = a_1 x_1 + a_2 x_2 + \ldots + bz$$

kde a_i jsou neznámé koeficienty udávající kvalitu jednotlivých pokusných ploch, právě jedna z hodnot x_i je prodané měření rovna 1, ostatní 0; to udává, odkud data pocházejí a konečně b je počítaný koeficient vlivu srážek, z potom množství srážek. Metodou nejmenších čtverců, případně jinou optimalizací, spočítáme, co nás zajímá, v tomto případě koeficient b.

Podobně jako při klasické analýze variance může být tabulka vícerozměrná, znamená to tedy, že vedle $a_i x_i$ budou v modelu vystupovat další alternativy $b_i y_i$ atd. Analogicky porovnáváním rozptylů v rámci jednotlivých alternativ lze určit jejich statistickou významnost. Pokud například rozptyl při položení $a_i = 0$ vyjde podle nějakého testu blízký rozptylu uvnitř jedné skupiny měření (tj. se stejnými hodnotami x_i, y_i atd.) je faktor daný alternativami x_i nevýznamný.

Na druhé straně nasadit příliš složitou hypotézu také není optimální řešení, nemusí totiž vyjít, vzhledem k nedostatečnému počtu stupňů volnosti, jako statisticky významné vůbec nic. Je-li potenciálních vlivů příliš mnoho, je možné další řešení. Z datového souboru vybereme náhodně přibližně 40% (lze teoreticky odvodit) a nasadíme komplikovanou hypotézu s nenáročným kritériem na statistickou významnost jednotlivých faktorů. Zjistíme, že některé by významné být mohly, získáme tak podstatně jednodušší hypotézu a tu ověříme na celém datovém souboru.

[1] HA HA HA

– Lepší statistické peky to možná spočítají, ale stejně to bude špatně

7 Korelace

Často je potřebné zjistit, zda mezi nějakými naměřenými veličinami existuje určitý vztah. Máme-li hypotézu, jak by vztah mohl vypadat, tj. jeho teoretickou funkční závislost s nějakými obecnými koeficienty, lze je spočítat nějakou optimalizační metodou a testem rozptylu naměřených hodnot oproti teoretickým zjistit, jak daná hypotéza vyhovuje.

Alternativním přístupem jsou koeficienty korelace. Jako zobecnění momentů zmíněných v části 2 definujeme smíšené (centrální) momenty pro více náhodných veličin, tzv. *korelační koeficienty* jako

$$\mu_{m_1,m_2,\ldots} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^{m_1}(y_i - \bar{y})^{m_2}\ldots$$

Analogicky se definují i normované korelační koeficienty

$$r_{m_1,m_2,\ldots} = \frac{\mu_{m_1,m_2,\ldots}}{\sigma_1^{m_1}\sigma_2^{m_2}\ldots}$$

Nejčastěji počítaný je $r_{1,1}$, udává lineární závislost dvou veličin. Pokud jsou veličiny nezávislé, vyjdou faktory jednou kladné, podruhé záporné, dohromady se to při dostatečně velkém n vyruší a vyjde $r_{1,1} \doteq 0$. Naopak při silné lineární závislosti, například kladné, znamená $x_i > \bar{x}$ většinou i $y_i > \bar{y}$, obrácená nerovnost analogicky, a tedy většina členů sumy je kladných, vzhledem k normování potom vyjde $r_{1,1} \doteq 1$. Analogicky pro zápornou závislost $r_{1,1} \doteq -1$.

Nulová hodnota $r_{1,1}$ znamená, že neexistuje lineární závislost, o jiné závislosti nic nevypovídá. Pro závislosti vyšších řádů je třeba nasadit vyšší korelační koeficienty, testy na ně jsou potom ale podstatně komplikovanější.

Je-li zkoumaných veličin více, lze z nich sestavit síť korelací a počítat *korigované korelace*, tj. takové, které nepovažují dvě veličiny za korelované, pokud mají společnou příčinu. Výzkumy totiž například ukazují paradoxní skutečnost, že u kuřáků je riziko infarktu nižší. Způsobeno je to ale faktem, že lidé náchylní k infarktu raději přestanou kouřit dříve. Korigované korelace ovšem má smysl počítat pouze v případě, že síť korelací není příliš hustá, tj. existují alespoň nějaké dvojice nekorelovaných veličin.

8 Neparametrické testy

Většina v praxi se vyskytujících dat není rozdělena normálně. Naopak většina statistických metod ve svém teoretickém odvození počítá s normálním rozdělením. Je tedy nutný jeden z následujících postupů

- Upravení rozdělení na zhruba normální
- Modifikace metod na nenormální rozdělení

Do první kategorie spadá už zmíněná transformace. Pokud jsou měření zatížena předpokládanými chybami, lze vypustit určitou malou část dat v extrémních hodnotách. To je ale nebezpečné, pokud extrémní hodnoty nebyly způsobeny chybou, může tak, zejména při relativně malém množství dat, dojít ke značnému zkreslení.

Do druhé kategorie patří tzv. *neparametrické testy*. Nejjednodušší případ nastává, pokud nás zajímá jen srovnání experimentů (tj. výsledky jednoho jsou větší než druhého) a nežádáme žádné podrobnější kvantitativní vyjádření. Potom stačí z naměřených hodnot sestavit náhodné páry a porovnat. Pokud vyjdou všechny pozitivní, lze tvrdit, že pokus o n měřeních má daný efekt s pravděpodobností $1 - 1/2^n$ atd. Tato metoda mnoho nevypovídá, je ale robustní vůči prakticky čemukoli.

– Jeden obrázek lže víc než tisíc čísel, já mám nejraději tabulky

Další neparametrické testy vycházejí z uspořádání naměřených hodnot a nahrazení vlastních hodnot jejich pořadím. Pokud například máme porovnat několik experimentálních skupin, spočteme pro každou součet pořadí jejích členů (v rámci všech naměřených hodnot). Problém nastává při počítání kritérií testů, je nutné vyčíslit funkce dané většinou větvícím se rekurentním předpisem podle počtu měření. Většinou je nějakým způsoben nutné spočítat počet možných kombinací, jak mohla daná situace nastat. Algoritmy pro výpočet takových funkcí jsou exponenciální, tedy rozumně spočitatelné pouze pro hodnoty zhruba do 10. Pokud je počet naměřených hodnot vyšší (od 50 nahoru), lze funkce přijatelně aproximovat spojitými, nejproblematičtější tedy je interval 10–50, kam bohužel spadá většina praktických exparimentů. Pro hodnoty do 20 si lze ještě pomoci jistými předpočítanými tabulkami, ale jinak ...

Analogiským způsobem lze vyhodnocovat závislost dvou veličin. Koeficient korelace počítáme opět z pořadí hodnot místo hodnot samých. Tímto způsobem určíme nanejvýš pozitivní nebo negativní závislost veličin, tj. že se vzrůstající první roste nebo klesá druhá, případně absenci takové závislosti, o konkrétní podobě závislosti původních veličin nelze pochopitelně usozovat nic.

Při porovnávání dvou veličin je opět podstatné, jakým způsobem položíme otázku. Je rozdíl, ptáme-li se, zda veličina vzrostla anebo se změnila. V obou případech bude při stejném prahu pravděpodobnosti a stejném původním rozložení požadovaná naměřená odchylka různá. Při dané pravděpodobnosti musíme mít stejnou plochu pod křivkou rozložení, pokud nás zajímá jen na jedné straně, stačí menší odchylka od průměru, aby bylo možné konstatovat, že naměřená veličina je patrně větší.

9 Multivariační metody

Dalším možným statistickým případem je situace, kdy máme vyhodnotit velké množství atributů dané množiny objektů (například ekologické výzkumy, data jsou výskyty desítek až stovek druhů v nějaké množině lokalit).

Hlavní úlohou je zde redukce nepříjemně mnoha dimenzí celé úlohy. Základní metodou je *analýza hlavních komponent* (**P**rincipal **C**omponent **A**nalysis). Zkoumané objekty (v „ekologickém" případě lokality) chápe jako body v hyperprostoru o souřadnicích podle hodnot jejich atributů. Takové vytvářejí jakýsi prostorový útvar, PCA předpokládá, že je to přibližně zkosený hyperelipsoid a na základě modifikované analýzy variance se snaží určit jeho statisticky významné osy (tj. takové, kolem nichž je nezanedbatelný rozptyl). Těch už bývá podstatně méně, úloha se tak stává čitelnější.

Základní PCA funguje uspokojivě pouze tehdy, pouze jsou-li četnosti v jednotlivých dimenzích „rozumně" rozložené. Pokud vyjde útvar tvořený body komplikovanější, PCA nedá důvěryhodné výsledky. Hodně lze napravit modifikací metriky na jednotlivých osách, a to do té míry, že výsledná metrika je euklieidovská jen lokálně. Takto pracuje *detrendovaná analýza korespondence.*

10 Shluková a diskriminační analýza

Shluková analýza je statistická metoda, jejímž cílem je uspořádat získaná data – body ve zmíněném multidimenzionálním prostoru do tzv. shluků, tedy podmnožin, které mají přibližně společné vlastnosti. Existuje celá řada způsobů, jak se takového rozdělení dopočítat, mnohdy se vychází z korelační matice, tj. matice korelačních koeficientů všech dvojic měřených veličin a různě se cvičí s metrikami. Při shlukování lze postupovat shora dolů i sdola nahoru.

Zajímavá visualizační metoda využívá ke tvorbě shluků přirozených lidských schopností. Transformuje jednotlivé měřené veličiny do rysů lidského obličeje a výsledné ksichtíky zobrazí. Je na obsluze, aby rozhodla, které jsou si podobné, což jsme u tváří zvyklí posuzovat, a tak to jde poměrně snadno a přesně.

– Matkou faktorové analýzy byla psychologie

Diskriminační analýza si navíc všímá vlivu jednotlivých faktorů na rozdělení do shluků. Dále zavádí tzv. *aposteriální pravděpodobnost,* která určuje, jak moc se lze spoléhat na to, že daný objekt skutečně přísluší do určeného shluku. Nachází-li se objekt těsně u hraniční nadroviny dvou shluků a četnost v obou je přibližně stejná, je pravděpodobnost, že jsme ho zařadili správně, vyšší proti případu, kdy je četnost ve shluku na druhé straně nadroviny podstatně vyšší.

Diskriminační analýza je metoda relativně robustní vůči nadměrné složitosti modelu, přidáme-li navíc dimenze, které jsou ve skutečnosti nevýznamné, metodu většinou nezmatou a vyjdou také jako bezvýznamné.

– I v renomovaných časopisech se občas objeví takové statistické nesmysly