Pokroky matematiky, fyziky a astronomie

Jaroslav Janko
Teorie ověřování statistických hypothes [Dokončení]

Pokroky matematiky, fyziky a astronomie, Vol. 3 (1958), No. 2, 125--139

Persistent URL: http://dml.cz/dmlcz/137407

POKROKY MATEMATIKY, FYSIKY A ASTRONOMIE
ROČNÍK III — ČÍSLO 2

MATEMATIKA

TEORIE OVĚŘOVÁNÍ STATISTICKÝCH HYPOTHES

Prof. J. JANKO, Praha

(dokončení)

9. Stanovení hladiny významnosti

Je třeba ještě nyní věnovat pozornost rozsahu kritického oboru čili pravděpodobnosti chyby typu I, neboli hladině významnosti α. Bude účelné uvažovat konkrétně, že máme za úkol rozhodnout, zda jsou dva průměry normálních základních souborů různé, posouzením významnosti rozdílu mezi výběrovými průměry $\bar{x}_1$ a $\bar{x}_2$, když rozptyly základních souborů jsou stejné a známé. Jedním z obvyklých postupů rozhodneme, že není významného rozdílu mezi průměry $\bar{x}_1$ a $\bar{x}_2$, jestliže čtverec pozorovaného rozdílu je menší než určitý násobek rozptylu, čili $(\bar{x}_1 - \bar{x}_2)^2 < t\sigma^2$. Součinitel t určíme pomocí výběrového rozdělení náhodné proměnné $\dfrac{(\bar{x}_1 - \bar{x}_2)^2}{\sigma^2}$ a určitého čísla α, které jsme nazvali hladinou významnosti. Pravděpodobnost $P\{(\bar{x}_1 - \bar{x}_2)^2 > t\sigma^2\}$, že odchylka $(\bar{x}_1 - \bar{x}_2)^2$ překročí určitou hodnotu $t\sigma^2$, je možno udělat tak malou jak chceme, jestliže se vezme $t\sigma^2$ dostatečně velké. Zvolí se tedy tak, že $P\{(\bar{x}_1 - \bar{x}_2)^2 > t\sigma^2\} = \alpha$, kde α je tak malé, abychom byli přesvědčeni, že je prakticky jisté, že určitý jev o pravděpodobnosti α se nevyskytne v jednom jediném pokusu. Když tedy vypočítáme z pozorovaných hodnot $(\bar{x}_1 - \bar{x}_2)^2$ a shledáme, že tento čtverec rozdílu výběrových průměrů je větší než $t\sigma^2$, znamená to, že jev, jehož pravděpodobnost je α, nastal. Podle naší hypotesy však takový jev má být prakticky nemožný v jednom jediném pokusu, takže musíme dojíti k závěru, že v tomto případě nebyla naše hypotesa potvrzena pozorovanou skutečností. Říkáme, že odchylka je významná. Když však shledáme, že $(\bar{x}_1 - \bar{x}_2)^2 \leqq t\sigma^2$, jsme ochotni přijmout hypotesu jako vhodnou interpretaci pozorovaných dat aspoň do té doby, než získáme další informace pozorováním skutečnosti. V tomto případě usuzujeme, že odchylka mohla vzniknout náhodnou měnlivostí a že pozorování jsou ve shodě s tou hypotesou.

Jestliže pozorovaná odchylka překročí hranici významnosti $t\sigma^2$, považujeme hypotesu za vyvrácenou na základě zkušenosti získané pozorováním, což ovšem není nikterak ekvivalentní s vyvrácením logickým. I když je hypotesa pravdivá, může se ve výjimečném případě vyskytnout jev $(\bar{x}_1 - \bar{x}_2)^2 > t\sigma^2$, jehož pravděpodobnost je α; ale je-li α dostatečně malé, jsme přesvědčeni, že je prakticky oprávněno tuto možnost neuvažovat. Obráceně zase, když se vyskytne jedna hodnota $(\bar{x}_1 - \bar{x}_2)^2 \leqq t\sigma^2$, není to důkazem pravdivosti hypotesy. Je tím pouze ukázáno, že na základě speciálního užitého testu je tu postačující shoda mezi theorií a pozorováním. Praktického potvrzení statistické hypotesy lze dosáhnout opakovanými testy různého druhu. V praktických aplikacích různých testů významnosti se užívá obvykle 5%, 1%, 0,1% hladin významnosti. Záleží ovšem na speciálních okolnostech dotyčného případu, které hladiny se má užít. Je však ve statistice stavem málo uspokojivým, jestliže se postupuje podle určité úmluvy, že pozorovaná hodnota, která překročí 5% hladinu, ale nikoli 1%, je skoro významná, hodnota mezi 1% a θ,1% hranicí je významná a hodnota překračující 0,1% hranici je vysoce významná.

V našem konkretním případu by měl výzkumník zvolit hladinu na základě své představy o tom, jaké risiko může převzít tím, že udělá chybu, když řekne, že průměry jsou významně rozdílné. Velikost risika je možno v mnohých případech měřit očekávanou objektivní hospodářskou ztrátou, která nastane v důsledku špatného rozhodnutí. Potom musí býti výzkumník veden snahou zařídit postup rozhodování tak, aby tato ztráta byla co nejmenší, čili postupovat podle zásady *minimální hospodářské škody*. Bývá to možné aspoň tam, kde důsledky špatného rozhodnutí nemají za následek ztráty lidských životů, neboť v těchto případech je problém mnohem složitější.

Je zřejmě žádoucí, aby statistikovy představy risika měly pokud možno reálný základ.

Touto cestou jde theorie rozhodovacích funkcí. Když statistik prozkoumá uspořádání pokusu a cíle, pro něž byl pokus navržen, může vzít v úvahu různá rozhodnutí, která může chtít udělat po provedení pokusu. Tento souhrn rozhodnutí se nazývá prostorem rozhodnutí. Určité zvláštní rozhodnutí označme d. Za účelem určitého rozhodnutí je třeba se pokusit říci něco o neznámé situaci, před kterou je statistik postaven, a v modelu, který sestrojil, je tato situace dána parametrem Θ. Když statistik má určitý plán, který mu řekne, které rozhodnutí má udělat při každém možném výsledku pokusu, má tak zvanou rozhodovací funkci $d(x)$, která pro každý výsledek x pokusu předepisuje rozhodnutí $d(x)$. Když zvolil určitou rozhodovací funkci, pak má již plán pro jednání, jakmile získal pozorování z výsledku pokusu. Statistikovou snahou v každém pokusu je najít určitou rozhodovací funkci $d(x)$, která je v nějakém smyslu dobrá nebo nejlepší. Aby mohl posuzovat rozhodovací funkce, musí míti nějakou představu o relativních výhodách, resp. škodách různých rozhodnutí d v každé situaci Θ, v níž se může ocitnout. Předpokládá se tudíž, že statistik může měřiti ztrátu hospodářskou, kterou utrpí, když udělá rozhodnutí d v situaci Θ. Dostane tudíž ztrátovou funkci $W(d, \Theta)$, na niž se klade požadavek $W(d, \Theta) \geqq 0$. Je zřejmo, že se očekává, že statistik neutrpí ztrátu, když d je správné rozhodnutí v nějaké situaci Θ, čili v takovém případě $W(d, \Theta) = 0$. Statistik nyní může zkoumat účinek rozhodovací funkce $d(x)$ v různých situacích Θ. Ovšem ztráta, kterou utrpí, nezávisí jen na Θ a na té zvláštní funkci, které užívá, ale také na pozorované hodnotě x dotyčné náhodné veličiny. Postupuje se nyní tak, že se zkoumá očekávaná čili průměrná ztráta, která se nazývá risikem a je tedy vyjádřena relací $R_{d(x)}(\Theta) = E_\Theta\{W(d(x), \Theta)\}$. Interpretujeme-li tento výraz pomocí relativní četnosti, pak můžeme říci, že představuje průměrnou ztrátu v řadě opakování pokusu. Připojený index značí, že risiko závisí na funkci $d(x)$, nikoli na zvláštních hodnotách té funkce. Statistik by potřeboval najít tu rozhodovací funkci, která by mu dala minimální hodnotu risika pro každé Θ. Ale jen v málo případech se stává, že funkce $d(x)$, pro kterou má risiko minimum pro jedno Θ, je funkcí, která dává minimum i pro jiné hodnoty Θ. Na základě úvah z theorie her je navrženo toto řešení: určí se maximální ztráta, která se může vyskytnout, užije-li se funkce $d(x)$, a statistik vybere tu rozhodovací funkci $d^*(x)$, která minimalisuje maximální risiko,

$$\max_{\Theta} R_{d^*(x)}(\Theta) \leqq \max_{\Theta} R_{d(x)}(\Theta) \text{ pro všechna } d(x) .$$

Tímto tak zvaným minimaxovým řešením se vybere rozhodovací funkce, která poskytuje ochranu proti nejnepříznivější situaci, jaká může vzniknout.

Různé části statistické indukce, jako odhad, testování hypotes, obory spolehlivosti, obory toleranční a j., je možno považovat za příklady obecné theorie rozhodování.

10. Sekvenční testy hypotes

Sekvenční metoda poskytuje techniku testování hypotes nebo odhadu parametrů v takových případech, kdy rozsah výběru není předem pevně stanoven, nýbrž se určuje v průběhu provádění pokusu, na základě kriteria, které závisí na výsledcích pozorování, jak je postupně dostáváme při pokusu.

Uvedli jsme (str. 13, č. 1), že nejlepší kritický obor pro testování hypotesy $H = \Theta$ proti alternativní hypotese $\overline{H} = \Theta'$ je ten, který tvoří všechny náhodné výběry $(x_1, \dots, x_n)$, splňující nerovnost

$$\frac{f_1(x_1)\, f_1(x_2) \dots f_1(x_n)}{f_0(x_1)\, f_0(x_2) \dots f_0(x_n)} > A ,$$

kde konstanta A se určí tak, aby

$$\int_{l_n > A} \dots \int f_0(x_1) \dots f_0(x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n = \alpha , \qquad (10.1)$$

kde $l_n = \prod_{i=1}^{n} \frac{f_1(x_i)}{f_0(x_i)}$ a kritický obor pro zamítnutí hypotesy H je obor $l_n > A$. Pro tento kritický obor je pravděpodobnost β chyby typu II (přijetí H, když pravdivá hypotesa je $\overline{H}$) minimální.

Sekvenční test užívá poměru věrohodnosti l_n a dvou kladných čísel $A > 1$ a $B < 1$. Jak postupně přicházejí pozorované hodnoty $x_1, x_2, \dots$, počítají se poměry $l_1, l_2, l_3, \dots$ tak dlouho, pokud

$$B < l_n < A . \qquad (10.2)$$

Jestliže je pro některé n hodnota l_n menší nebo rovna B, přijme se hypotesa H a test je skončen. Je-li na některém stupni l_n větší nebo rovno A, zamítne se H a test je skončen. Postup trvá tedy tak dlouho, až l_n padne mimo interval (10.2), kdy se ve výběru již nepokračuje.

Lze ukázati, že tento test jednou skončí, čili že je pravděpodobnost 1, že sekvenční test poměrem věrohodnosti skončí, ať je rozdělení x jakékoliv.

Zaveďme $z = \lg \left\{ \frac{f_1(x)}{f_0(x)} \right\}$, pak z bude mít nějakou funkci hustoty $g(z)$, která je určena funkcí hustoty proměnné x. Posloupnost pozorovaných hodnot $x_1, x_2, \dots$ určuje posloupnost pozorovaných hodnot $z_1, z_2, \dots$ proměnné z. Posloupnost nerovností (10.2) pak přejde v

$$\lg B < \sum_{i=1}^{n} z_i < \lg A , \qquad (10.3)$$

kde $\lg A$ je číslo kladné a $\lg B$ je číslo záporné. Označme $c = \lg A - \lg B$ a $p = \int\limits_{-c}^{c} g(z)\,dz$. Padne-li některá z hodnot z_i mimo interval od $-c$ do $+c$, je jedna z nerovností v (10.3) nutně porušena na tomto stupni, nebyla-li již porušena na některém z dřívějších. Jestliže vztah (10.3) má platit pro všechna n, musí aspoň každé z_i padnout mezi $-c$ a $+c$. Mohou být ovšem porušeny tyto nerovnosti i když všechna z padnou do tohoto intervalu. Poněvadž ta z_i jsou nezávislá, je pravděpodobnost, že každé z prvních n pozorovaných z_i padne do toho intervalu, rovna p^n, a tato pravděpodobnost se blíží nule, když n roste nade všechny meze, ježto $p < 1$. Z toho plyne, že vztah (10.3) nemůže býti splněn do nekonečna. Kdyby funkce $g(z)$ byla rovna nule mimo interval $\langle -c, +c \rangle$, docílilo by se zavedením nové proměnné, aby nenulový obor její funkce hustoty nepadl do tohoto intervalu.

Abychom určili hranice A a B, najdeme základní vztahy mezi veličinami α, β, A, B.

Pravděpodobnost α, že hypotesa H bude zamítnuta, když je pravdivá, se vypočítá jako pravděpodobnost, že $l_n > A$ dříve než $l_n < B$, čili

$$\alpha = P\{l_1 \geqq A\} + P\{B < l_1 < A, l_2 \geqq A\} + \\ + P\{B < l_1 < A, B < l_2 < A, l_3 \geqq A\} + \dots \tag{10.4}$$

Pravděpodobnost β, že hypotesa H bude přijata, když pravdivá hypotesa je $\overline{H}$, se pak stanoví jako pravděpodobnost, že $l_n < B$ dříve než $l_n > A$.

$$\beta = P\{l_1 \leqq B\} + P\{B < l_1 < A, l_2 \leqq B\} + \\ + P\{B < l_1 < A,\ B < l_2 < A,\ l_3 \leqq B\} + \dots \tag{10.5}$$

Bylo by možno všechny tyto pravděpodobnosti vypočítat pro určité dvě funkce hustoty, takže ve výrazu (10.4) na pravé straně bychom užili $f_0(x)$ a v (5) funkce $f_1(x)$. Z toho je patrno, že veličiny α a β jsou známými funkcemi A a B. Jestliže se tedy stanoví α a β předem, pak je možno z rovnic (10.4) a (10.5) určit hranice A a B. Tento výpočet by byl ovšem příliš složitý, takže se ho v praxi neužívá. Lze však dosáhnouti přibližných a zcela vyhovujících výrazů touto úvahou:

Představme si, že by l_n byla spojitá funkce spojité proměnné n, takže může býti znázorněna křivkou. Pohybujeme-li se nyní po ose n, dojdeme k bodu, kde se l_n po prve rovná buď A nebo B; to znamená, test se provádí tak dlouho, dokud je vztah (10.2) splněn, a přestane, jakmile je buď $l_n = B$ (kdy se hypotesa H přijme) nebo $l_n = A$ (kdy se přijme alternativní hypotesa $\overline{H}$). Ve všech bodech $(x_1, x_2, \dots)$ prostoru, kde je hypotesa H přijata, je věrohodnost $\overline{H}$, kterou označíme λ_1, rovna přesně B-násobné věrohodnosti H, kterou označíme λ_0, neboť v těchto bodech je poměr věrohodností $l = \dfrac{\lambda_1}{\lambda_0} = B$. Je tudíž integrál funkce λ_1 přes tyto body právě roven B-násobku integrálu funkce λ_0 přes tyto body. První integrál je však β, neboť je to pravděpodobnost, že H bude přijata, když $\overline{H}$ je pravdivá, a druhý je $1 - \alpha$, neboť je to pravděpodobnost, že H bude přijata, když je pravdivá. Kdyby bylo možno provádět výběr spojitě, bylo by přesně $\beta = B(1 - \alpha)$, takže

$$B = \frac{\beta}{1 - \alpha}. \tag{10.6}$$

Poněvadž n není spojitá proměnná, platí tento vztah přibližně, ale užijeme-li ho, chyba je velmi malá.

Podobnou úvahu můžeme provésti, je-li $l_n = A$; potom $1 - \beta = A\alpha$, takže

$$A = \frac{1-\beta}{\alpha}\,. \tag{10.7}$$

Pomocí rovnic (10.6) a (10.7) je provádění sekvenčního testu velmi jednoduché, neboť se zvolí podle povahy konkretního problému vhodně α a β, vypočítá se A a B, načež se ihned přikročí k provádění testu.

Abychom mohli posoudit práci sekvenčního testu spočívajícího na poměru věrohodností musíme mít *obraz jeho silofunkce*. Budeme pro jednoduchost uvažovat případ jednoho neznámého parametru Θ a zavedeme označení funkce hustoty $f(x, \Theta)$. Budeme testovat nulovou hypotesu $\Theta = \Theta_0$ proti alternativní hypotese $\Theta = \Theta_1$. Budeme zkoumat sílu testu $P(\Theta)$ čili pravděpodobnost, že hypotesa Θ_0 bude zamítnuta, když jiná hodnota parametru Θ je pravdivá. Jsou dány pravděpodobnosti α a β, takže

$$P(\Theta_0) = \alpha\,, \quad P(\Theta_1) = 1 - \beta\,.$$

Naznačíme zde jen přibližné odvození bez důkazu jeho správnosti. Θ bude v dalším značit libovolnou pevnou hodnotu parametru. Užívá se k odvození důvtipného obratu, uvažujícího výraz $\left[\frac{f(x, \Theta_1)}{f(x, \Theta_0)}\right]^h$. Je třeba, aby existovalo číslo $h \neq 0$ takové, že $g(x, \Theta) = \left[\frac{f(x, \Theta_1)}{f(x, \Theta_0)}\right]^h f(x, \Theta)$ je funkcí hustoty čili $\int_{-\infty}^{\infty} g(x, \Theta)\, \mathrm{d}x = 1$.

Pro $h = 0$ bude ovšem $g(x, \Theta)$ funkcí hustoty, poněvadž $f(x, \Theta)$ je funkce hustoty. Abychom ukázali, že existuje takové číslo h různé od nuly, uvažujeme střední hodnotu veličiny $\left[\frac{f(x, \Theta_1)}{f(x, \Theta_0)}\right]^u$ jako funkci u, tedy

$$\psi(u) = \int_{-\infty}^{\infty} \left[\frac{f(x, \Theta_1)}{f(x, \Theta_0)}\right]^u f(x, \Theta)\, \mathrm{d}x\,.$$

Je zřejmě $\psi(u)$ stále kladná a $\psi(0) = 1$. Když $u \to \infty$ v kladném nebo záporném směru, $\psi(u) \to \infty$. Poněvadž totiž $f(x, \Theta_1) \neq f(x, \Theta_0)$, bude existovat interval nebo množina intervalů, kde poměr těchto dvou funkcí je větší než jedna; na takových intervalech nabývá integrovaná funkce velikých hodnot, když u roste. Podobně budou intervaly, kde převratná hodnota toho poměru je větší než jedna, takže integrovaná funkce nabývá velkých hodnot pro velké záporné hodnoty u. Ve výjimečných případech mohou existovat intervaly jen jednoho druhu, a to tehdy, když nejsou splněny určité dosti obecné podmínky ([18], str. 158). Tím je tedy existence h objasněna. Kdyby funkce $\psi(u)$ měla v $u = 0$ minimum, pak by v tomto případě h neexistovalo, ale to se může přihodit jen ve výjimečných případech, kdy nejsou splněny výše uvedené podmínky. Může ovšem existovat minimum funkce $\psi(u)$ pro kladné u nebo pro záporné u, takže h je buď kladné nebo záporné. Vidíme tedy, že existuje obecně $h \neq 0$ takové, že $\psi(h) = 1$ a tedy $g(x, \Theta)$ je funkce hustoty.

Budeme nyní provádět sekvenční test nulové hypotesy H', že funkce hustoty je $f(x, \Theta)$ proti alternativní hypotese $\overline{H}'$, že funkce hustoty je $g(x, \Theta)$. Předpokládáme, že nulová hypotesa je pravdivá. Jako hranice poměru věrohodností vezmeme A^h a B^h. V provádění testu se tedy pokračuje, pokud

$$B^h < \frac{g(x_1, \Theta)\, g(x_2, \Theta) \ldots g(x_n, \Theta)}{f(x_1, \Theta)\, f(x_2, \Theta) \ldots f(x_n, \Theta)} < A^h \tag{10.8}$$

a přestane se, jakmile se tento poměr rovná či překročí některou z hranic. Předpokládejme, že $h > 0$; kdyby $h < 0$ zaměnily by se A a B.

Vzhledem k definici funkce $g(x, \Theta)$ je patrno, že test (10.8) je zcela ekvivalentní původnímu sekvenčnímu testu $B < l_n < A$.

Hypotesa H se tudíž zamítne tehdy a jen tehdy, když se zamítne H'. Můžeme však vypočítat pravděpodobnost, že H' bude zamítnuta, když je pravdivá, t. j. když $f(x, \Theta)$ je pravdivá funkce hustoty. Tato pravděpodobnost je pak také rovna pravděpodobnosti, že hypotesa H bude zamítnuta, je-li $f(x, \Theta)$ pravdivá funkce hustoty, t. j. rovná-li se hodnotě silofunkce $P(\Theta)$. Hypotesa H' bude zamítnuta, když je pravdivá s pravděpodobností α' a přijata, když je $\overline{H}'$ pravdivá s pravděpodobností β', takže ve shodě s (6) a (7) je $B^h \doteq \frac{\beta'}{1 - \alpha'}$, $A^h \doteq \frac{1 - \beta'}{\alpha'}$. Vypočítáme-li z těchto rovnic α', dostaneme $\alpha' = P(\Theta) \doteq \doteq \frac{1 - B^h}{A^h - B^h}$. Máme-li najít pořadnici silofunkce v bodě Θ, musíme nejprve najít $\psi(u)$ pro tuto hodnotu Θ, pak položíme $\psi(u) = 1$ a řešíme podle u. Kořen různý od nuly je pak to číslo h, které určí $P(\Theta)$.

Stanovme nyní průměr výběrového rozsahu při sekvenčním testu. Výběrový rozsah n je tu náhodnou proměnnou, pro jejíž střední hodnotu $E(n)$ chceme najít přibližný výraz. Mějme náhodnou proměnnou $z = \lg \frac{f(x, \Theta_1)}{f(x, \Theta_0)}$ a budiž n nejmenší celé číslo, pro které součet $z_1 + z_2 + \ldots + z_n = Z_n$ nesplňuje nerovnosti

$$\lg B < Z_n < \lg A\,. \tag{10.9}$$

Ukážeme, že střední hodnota proměnné Z_n, která závisí na náhodných proměnných z a náhodné proměnné n, je

$$E(Z_n) = E(n)\, E(z)\,. \tag{10.10}$$

Nechť N je velmi velká, ale pevná hodnota proměnné n. Nebudeme přihlížet k té části zákona rozdělení n, která je na pravo od hodnoty N. Chybu, která tím vznikne, můžeme učinit libovolně malou, vezmeme-li N dosti velké. Poněvadž N je pevné, platí

$$E(Z_N) = NE(z)\,. \tag{10.11}$$

Proměnnou Z_N píšeme ve tvaru $Z_N = Z_n + W_n$, čímž definujeme novou proměnnou W_n a vzhledem ku (10.10) platí

$$E(Z_n + W_n) = NE(z)\,. \tag{10.12}$$

Potíž dospět k rovnici (10.10) přímo spočívá v tom, že obor proměnné z_i závisí na indexu i. Je-li $i > n$, pak $E(z_i) = E(z)$, ale pro $i < n$ je obor z_i omezen vztahy (10.9). V rovnici (10.12) obsahuje proměnná W_n jen ta z, pro která $i > n$, takže střední hodnota každého z ve W_n je $E(z)$.

Je tudíž střední hodnota

$$E(W_n) = E(z)\,E(N - n)\,, \tag{10.13}$$

kde druhý faktor na pravé straně závisí jen na rozdělení n. Z rovnic (10.12) a (10.13) plyne

$$\begin{aligned} NE(z) &= E(Z_n) + E(W_n) \\ &= E(Z_n) + E(z)\,[N - E(n)]\,, \end{aligned}$$

což se shoduje s (10.10), odkud pak dostáváme

$$E(n) = \frac{E(Z_n)}{E(z)}\,. \tag{10.14}$$

Na základě této rovnice můžeme odvodit jednoduchou přibližnou formuli pro střední hodnotu výběrového rozsahu. Proměnná Z_n nabývá jen hodnot větších než $\lg A$ a menších než $\lg B$. Nepřihlížíme-li k hodnotám, o které Z_n překročí $\lg A$ nebo o které je menší než $\lg B$, můžeme říci, že Z_n nabývá skutečně jen dvou hodnot, $\lg A$ a $\lg B$. Je-li pravdivý zákon rozdělení $f(x, \Theta)$, pak pravděpodobnost, že Z_n nabude hodnoty $\lg A$, je $P(\Theta)$, a pravděpodobnost, že nabude hodnoty $\lg B$, je $1 - P(\Theta)$. Bude tudíž přibližně střední hodnota $E(Z_n) \doteq P(\Theta) \lg A + [1 - P(\Theta)] \lg B$, takže

$$E_\Theta(n) \doteq \frac{P(\Theta) \lg A + [1 - P(\Theta)] \lg B}{E_\Theta(z)}\,, \tag{10.15}$$

kde v $E(n)$ a $E(z)$ bylo vyznačeno, že jde o podmíněnou střední hodnotu. Pomocí tohoto výsledku můžeme srovnávat sekvenční testy s testy o pevném výběrovém rozsahu.

Příklad:

Vezmeme v úvahu test hypotesy $\mu = 0$ proti alternativní hypotese $\mu = 1$ pro normální základní soubor $N(\mu, 1)$. Zvolíme $\alpha = 0{,}01$ a $\beta = 0{,}01$. Pak bude podle (10.7) $A = 99$ a podle (10.6) $B = \dfrac{1}{99}$. Předpokládejme dále, že pravdivá hodnota parametru je nula, takže $P(\Theta)$ v rovnici (10.15) je 0,01. Musíme vypočítat střední hodnotu proměnné $z = \lg \dfrac{e^{-\frac{(x-1)^2}{2}}}{e^{-\frac{x^2}{2}}} = x - \dfrac{1}{2}$, která je při předpokládaném pravdivém rozdělení $-\frac{1}{2}$. Po dosazení do (10.15) tedy bude $E(n) \doteq \dfrac{0{,}01 \lg 99 + 0{,}99 \lg \frac{1}{99}}{-\frac{1}{2}} \doteq 1{,}96 \lg 99 \doteq 9$. Proveďme nyní srovnání tohoto výsledku s náhodným výběrem o pevném rozsahu. Nejlepší test o stejných pravděpodobnostech chyb obojího typu dostaneme, zvolíme-li určité číslo c a přijmeme hypotesu $\mu = 0$, je-li $\bar{x} < c$, a zamítneme ji při $\bar{x} > c$. Pravděpodobnost α, že hypotesa H bude zamítnuta (při $\mu = 0$), je dána výrazem

$$\alpha = \sqrt{\frac{n}{2\pi}} \int_c^\infty e^{-\frac{n}{2}\bar{x}^2}\, d\bar{x} = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{n}\,c}^\infty e^{-\frac{t^2}{2}}\, dt\,,$$

takže pro $\alpha = 0{,}01$ je

$$c\sqrt{n} = 2{,}326\,. \tag{10.16}$$

Pravděpodobnost β, že hypotesa H bude přijata při pravdivé hypotese $\overline{H}$ (t. j. $\mu = 1$), je

$$\beta = \sqrt{\frac{n}{2\pi}} \int_{-\infty}^{c} e^{-\frac{n}{2}(\bar{x}-1)^2}\, d\bar{x} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{n}(c-1)} e^{-\frac{t^2}{2}}\, dt\,,$$

takže pro $\beta = 0{,}01$ je

$$(c - 1)\sqrt{n} = -\,2{,}326 \tag{10.17}$$

Z rovnic (10.16) a (10.17) najdeme $n = 22$. Vidíme tedy, že sekvenční test naší nulové hypotesy by vyžadoval průměrně jen $\frac{9}{22}$ počtu pozorování, jež vyžaduje výběr s pevným rozsahem čili jen 41 procent.

11. Sekvenční test poměrem pravděpodobností, je-li počet pozorování shora omezen

Sekvenční test poměrem pravděpodobností, odvozený Waldem, má tu optimální vlastnost, že minimalisuje očekávaný počet pozorování potřebných k rozhodnutí. Další výhoda, o níž jsme se již zmínili, je, že hranice pro poměr pravděpodobností můžeme dostat aspoň s postačující přibližností bez řešení nějakého distribučního problému, takže není třeba, abychom se omezovali na typy rozdělení, jichž se obvykle užívá v běžných testovacích postupech. Cílem sekvenčního testu poměrem pravděpodobností je rozhodnout mezi dvěma alternativními hypotesami H a $\overline{H}$, při čemž se ve výběru pokračuje tak dlouho, až je dosaženo rozhodnutí ve prospěch jedné nebo druhé alternativní hypotesy. Jsou však případy, které vyžadují, aby bylo učiněno rozhodnutí dříve, než se vezme určitý předem stanovený počet pozorování, kde je tedy nutná úprava v tomto problému tak, že se určí horní mez N pro počet pozorování. Pro tento případ navrhl Wald tuto metodu kusého sekvenčního postupu. Je-li $\frac{p_{1N}}{p_{0N}}$ poměr pravděpodobností při N-tém pozorování a nedosáhlo-li se rozhodnutí před N-tým pozorováním při kriteriu s hranicemi A a B, které byly vhodně stanoveny pro obecný sekvenční postup, použije se rozhodovacího pravidla takového, že se přijme $\overline{H}$, je-li $\frac{p_{1N}}{p_{0N}} \geqq 1$ a hypotesa H se přijme, je-li $\frac{p_{1N}}{p_{0N}} < 1$. Wald vypočítal také horní mez chyby, které se dopustíme při takovém přerušení [18].

Problém tohoto kriteria lze studovati hlouběji a položit otázku, zda je možno nahraditi hranice A a B užšími hranicemi, když je známo, že nebude vzato víc než N pozorování. Označíme příslušné hranice v tomto případě $A(N)$ a $B(N)$, takže Waldovým hranicím by odpovídaly symboly $A(\infty)$ a $B(\infty)$. Užší hranice mají tu výhodu, že vzrostla do určité míry četnost případů, v nichž se dostane před N-tým pozorováním rozhodná odpověď, t. j. zamítnutí nulové hypotesy s jistou spolehlivostí, zatím co četnost nesprávných rozhodnutí zůstává na téže hladině.

Uvažujme tedy kusý sekvenční postup, který končí při N-tém pozorování a vede k rozhodnutí pro přijetí testované hypotesy H nebo pro přijetí alternativní hypotesy $\overline{H}$ nebo pro nepřijetí žádné z nich. Výklad podáme podle [11]. Označíme zase

α pravděpodobnost přijetí $\overline{H}$, když je H pravdivá

β pravděpodobnost přijetí H, když je $\overline{H}$ pravdivá.

Poněvadž tento sekvenční postup může skončit bez přijetí jedné z obou hypotes, je třeba, abychom zavedli

δ_0 pravděpodobnost, že sekvenční postup skončí, aniž se vyjádří pro některou hypotesu, když hypotesa H je pravdivá,

δ_1 pravděpodobnost téhož jevu, když hypotesa $\overline{H}$ je pravdivá.

Tyto dvě pravděpodobnosti nejsou předem určeny, ale lze počítati jejich hodnoty přiřazené určitému postupu testu.

Označíme p_{0n} resp. p_{1n} hustoty pravděpodobnosti prvních n pozorování, které odpovídají hypotesám H resp. $\overline{H}$.

Kusý sekvenční test s novými hranicemi $A(N)$ a $B(N)$ bude definován rozhodovacím pravidlem, které předpisuje:

1. pokračovat ve výběru tak dlouho, pokud jsou splněny nerovnosti

$$B(N) < \frac{p_{1n}}{p_{0n}} < A(N) \quad \text{pro} \quad n < N\,;$$

2. přijmout hypotesu $\overline{H}$, je-li $\dfrac{p_{1n}}{p_{0n}} \geqq A(N)$ na každém stupni $n \leqq N$ a přerušit výběr;

3. přijmout hypotesu H, je-li $\dfrac{p_{1n}}{p_{0n}} \leqq B(N)$ na každém stupni $n \leqq N$ a přerušit výběr;

4. prohlásit, že hypotesy jsou nerozeznatelné, když poměr pravděpodobností leží mezi $B(N)$ a $A(N)$ až do N-tého stupně. Na tomto stupni by se mohla udělat rozhodnutí prozatímní místo rozhodnutí 4, která by mohla být užitečná při plánování budoucích pokusů. Rozhodovací pravidlo upravené může pak v bodě 4 předpisovat

4a. přijmout hypotesu $\overline{H}$, prozatímně, je-li $1 < \dfrac{p_{1N}}{p_{0N}} < A(N)$;

4b. přijmout hypotesu H, prozatímně, je-li $B(N) < \dfrac{p_{1N}}{p_{0N}} < 1$ a jestliže je $p_{1N} = p_{0N}$, užít poměru $p_{1,N-1}$ a $p_{0,N-1}$.

Hledejme nyní vztahy mezi veličinami α, β, δ_0, δ_1 a hranicemi $A(N)$ a $B(N)$. Množiny bodů $(x_1, x_2, \ldots, x_N)$, které vedou k rozhodnutím 2, 3, 4., označme w_1, w_2, w_3 resp. a prostorový element pišme krátce dv. Potom platí tyto vztahy

$$\int_{w_1} p_{1N}\,dv \geqq A(N) \int_{w_1} p_{0N}\,dv \quad \text{nebo} \quad 1 - \beta - \delta_1 \geqq \alpha A(N)\,,$$

$$\int_{w_2} p_{1N}\,dv \leqq B(N) \int_{w_2} p_{0N}\,dv \quad \text{nebo} \quad \beta \leqq (1 - \alpha - \delta_0)\,B(N)\,,$$

takže

$$A(N) \leqq \frac{1 - \beta - \delta_1}{\alpha}\,, \quad B(N) \geqq \frac{\beta}{1 - \alpha - \delta_0}\,.$$

Zvolíme-li tedy jako hranice hodnoty $A(N)$ a $B(N)$ vyplývající z těchto vztahů při znaménku rovnosti, vidíme, že jsou užší než meze obecné A a B, neboť $\frac{1-\beta}{\alpha} > \frac{1-\beta-\delta_1}{\alpha}$ a $\frac{\beta}{1-\alpha} < \frac{\beta}{1-\alpha-\delta_0}$. Určit přesné hodnoty δ_0 a δ_1 je nesnadné, ale můžeme k nim dostat dolní meze d_0 a d_1 přibližně a to skoro ve všech případech, kdy N je velké. Potom jsou upravené hranice

$$A(N) = \frac{1-\beta-d_1}{\alpha} \geqq \frac{1-\beta-\delta_1}{\alpha}, \tag{11.1}$$

$$B(N) = \frac{\beta}{1-\alpha-d_0} \leqq \frac{\beta}{1-\alpha-\delta_0}. \tag{11.2}$$

Uvažme způsob, kterým bychom mohli určit d_1 a d_0. Kdybychom testovali hypotesu H proti alternativě $\overline{H}$ výběrem pevného rozsahu N obvyklým postupem, byl by nejlepší kritický obor w rozsahu α určen z podmínek $\int_w p_{0N}\,\mathrm{d}v = \alpha$, $\int_w p_{1N}\,\mathrm{d}v = \gamma_1$, kde γ_1 znamená maximální pravděpodobnost, že bude zamítnuta hypotesa H, když je pravdivá hypotesa $\overline{H}$. Takový kritický obor je určen vztahem $p_{1N} > kp_{0N}$; kde k je zvoleno tak, aby splňovalo podmínku první.

Podle horního sekvenčního postupu je pravděpodobnost, že bude zamítnuta hypotesa H, když je pravdivá hypotesa $\overline{H}$, rovna $1-\beta-\delta_1$, a tato nemůže překročit pravděpodobnost γ_1 patřící nejlepšímu kritickému oboru. Tudíž

$$1-\beta-\delta_1 \leqq \gamma_1 \quad \text{čili} \quad \delta_1 \geqq 1-\beta-\gamma_1 .$$

Můžeme tedy určit d_1 v případě, že $1-\beta-\gamma_1 > 0$ tak, že položíme $d_1 = 1-\beta-\gamma_1$, $d_1 = 0$ pak, jestliže je $1-\beta-\gamma_1 < 0$. Podobně, označíme-li γ_0 maximální hodnotu integrálu $\int_w p_{0N}\,\mathrm{d}v$ za podmínky $\int_w p_{1N}\,\mathrm{d}v = \beta$, položíme $d_0 = 1-\alpha-\gamma_0$, je-li výraz na pravé straně této rovnice kladný, v ostatních případech položíme $d_0 = 0$. Závisí tedy určení hodnot d_1 a d_0 na γ_1 a γ_0, jejichž výpočet nemusí být obecně snadný. Výpočty jsou však velmi jednoduché, je-li N velké a pozorování nezávislá, neboť pak lze považovati $\lg \frac{p_{1N}}{p_{0N}}$ za proměnnou s normálním rozdělením.

Příklad: Jako příklad vezmeme v úvahu základní soubor s normálním rozdělením $N(\mu, 1)$. Průměr za nulové hypotesy H nechť je $\mu = 0$ a za alternativní hypotesy $\overline{H}$ je $\mu = 0{,}2$. Je-li počet pozorování nejvýše $N = 100$, jest rozhodnout, který je nejlepší sekvenční test.

Je zřejmé, že nejlepší test pro hypotesu H na základě výběru o rozsahu $N = 100$ pozorování je $\bar{x} \geqq \frac{1{,}645}{\sqrt{N}}$, je-li hladina významnosti pětiprocentní a $\bar{x}$ značí průměr 100 pozorování. Pravděpodobnost, že zamítneme testovanou hypotesu, když je pravdivá hypotesa alternativní, je dána plochou normální křivky $N(0,1)$ napravo od pořadnice v bodě $\bar{x} = 1{,}645 - \frac{0{,}2}{\frac{1}{\sqrt{100}}} = -0{,}355$, která je přibližně rovna 0,64.

Můžeme tedy položit na základě vztahu (11.1) $A\,(100) = \frac{\gamma_1}{\alpha} = \frac{0{,}64}{0{,}05} = 12{,}8$ a podobně pro $\beta = 0{,}05$ bude podle (11.2) $B\,(100) = \frac{\beta}{\gamma_0} = \frac{0{,}05}{0{,}64} = 0{,}078$. Tyto hranice jsou užší než obecné hranice $A = 19$, $B = 0{,}053$.

Dotkneme se ještě chyby, které se dopustíme, když děláme prozatímní rozhodnutí v případech, kdy poměr pravděpodobností leží mezi hranicemi $B(N)$ a $A(N)$ až do N-tého stupně. Označme $\varrho_0(N)$ pravděpodobnost podle H, že sekvenční postup nevede k rozhodné odpovědi a že poměr $\frac{p_{1N}}{p_{0N}}$ leží v intervalu 1 až $A(N)$. Celková chyba, které se dopustíme, když přijmeme alternativní hypotesu $\overline{H}$, je-li hypotesa H pravdivá, včetně chyby z prozatímního rozhodnutí, je [11] $\alpha + \varrho_0(N) = \alpha'$. Budiž $\overline{\varrho}_0(N)$ pravděpodobnost, že $1 < \frac{p_{1N}}{p_{0N}} < A(N)$, když H je pravdivá hypotesa. Vzhledem k tomu, že v tomto jsou zahrnuty případy, kde poměr pravděpodobností $\frac{p_{1N}}{p_{0N}}$ nemusí ležet v intervalu $\langle B(N), A(N)\rangle$ pro všechna $n < N$, vyplývá, že $\overline{\varrho}_0(N) > \varrho_0(N)$. Bude tedy horní mez celkové chyby $\alpha + \overline{\varrho}_0(N)$. Naše aproximace je velmi hrubá; skutečná hodnota α' je asi mnohem menší než horní mez, kterou jsme našli. V obecném případě není vyčíslení $\overline{\varrho}_0(N)$ snadné. Je-li však N velké, lze užíti k určení aproximace normálním rozdělením, takže

$$\overline{\varrho}_{0N} = P\{0 < \lg p_{1N} - \lg p_{0N} < \lg A(N)\}\,,$$

jak ukázal Wald a užil A místo $A(N)$, takže je pak $\overline{\varrho}_{0N}$ nadceněno. Podobně je možno vyčíslit chybu, které se dopouštíme, přijmeme-li hypotesu H, když je $\overline{H}$ pravdivá.

V příkladě, který jsme nahoře uvažovali je

$$\overline{\varrho}_{0N} = P\{0 < 0{,}2\sum_{i=1}^{100} x_i - 2 < \lg A(N)\}$$

takže α' je menší než 0,153.

12. Významnost „statistická“ a „materiální“

Zmíníme se jen krátce o otázce ověřování přibližné platnosti statistických hypotes. Formuluje se na př. testovaná hypotesa tak, že určitý základní soubor má normální zákon rozdělení a při tom se uznává jako skutečnost, že přírodní základní soubor není nikdy přesně normální. V takovém případě chceme zamítat testovanou normalitu jen tehdy, když odchylka skutečného rozdělení od normální křivky je tak velká, že se v našem zkoumání jeví vskutku závažnou. Jiná hypotesa, kterou máme testovat je, že dvě dodávky výrobků jsou vzhledem k určitému znaku úplně stejné. Jsme přesvědčeni, že nemohou býti ve skutečnosti přesně stejné, ale hypotesu o stejnosti chceme zamítnout jen tehdy, projeví-li se dodávky různými do takové míry, že ta různost má praktický význam. Z těchto příkladů je zřejmo, že chceme obyčejně zamítat testované statistické hypotesy teprve tehdy, když se jeví nepravdivými v rozsahu, který je významný pro zkoumaný předmět. Poněvadž testy užívané ve statistické praxi nemají tolik síly, aby vznikalo opravdové nebezpečí zamítnutí hypotesy, když její nesprávnost je jen malého dosahu, nepůsobí tato okolnost velkých nesnází; může tomu však tak býti, pracuje-li se s rozsáhlými výběry. Proto se jeví vhodným rozlišovat [4] při formulaci problému mezi významností statistickou a t. zv. materiální. Výzkumník zvolí vhodně pro svůj úkol nulovou hypotesu H_0. Shromažďuje pak data, na jejichž základě tuto hypotesu buď zamítne nebo přijme. Existuje však třída nulových hypotes, které nejsou v rozporu se získanými daty na zvolené hladině významnosti. Tuto třídu je možno nazvati fiduciální množinou hypotes. Uvažujme obvyklý případ, že tyto hypotesy se vztahují k hodnotám určitého parametru Θ, takže ta fiduciální množina bude míti povahu intervalu. Bude pak vhodné zavésti do parametrického prostoru míru „vzdálenosti" Θ od H_0, kterou označme třeba $\Delta(\Theta)$. Tato míra bude vhodnou stupnicí vyjadřovat aspoň zhruba, jak jsou odchylky od H_0 podstatné čili materiální. Fiduciální množina hypotes může pak být určena jako množina těch Θ, pro která $\Delta(\Theta)$ nepřekročí určitou stanovenou hodnotu Δ_0. Při volbě Δ_0 budeme postaveni před problémy podobné těm, s nimiž se setkáváme, když máme volit alternativu, pro kterou se má dostat určitě stanovená síla. Taková formulace vyžaduje ovšem úpravy obvyklých testů, která byla pro některé parametrické problémy provedena v práci [4], kde byla také s tohoto hlediska uvažována úprava klasického Studentova problému a testu dobré přiléhavosti.

13. Problémy klasifikace

Od problémů ověřování hypotes je vhodné a užitečné rozlišovat problémy klasifikace, třebaže se jeví často snaha řešit oba tyto problémy na stejném základě. Při ověřování hypotesy máme na jedné straně jasně stanovenou nulovou hypotesu a na druhé straně poměrně neurčitou množinu alternativ. Hlavní váha je položena na nulovou hypotesu, která má býti zamítnuta nebo prozatímně přijata. Když je nulová hypotesa zamítnuta, nečiní se již rozhodnutí o skutečné alternativní hypotese. V problémech klasifikace, které se také nazývají problémy *diskriminace*, však máme množinu alternativních hypotes, z nichž má býti jedna vybrána. Otázkou v prvním problému je zamítnutí nulové hypotesy s určitým daným risikem. Otázkou v druhém problému je odvážení a volba mezi špatnými a správnými rozhodnutími. Nulová hypotesa je jedna; výzkumník ji zvolil vhodně k svému výzkumnému problému. Shromáždí-li se v průběhu experimentální práce dosti informace proti této nulové hypotese, zamítne ji výzkumník a nepokouší se odvažovat mezi informacemi poskytnutými z pozorovaných dat a volit různé alternativy. V druhém problému nemá nulová hypotesa vynikající úlohu. Tento problém vzniká v pokusech laboratorních i ve zkoumání průmyslových výrobků; můžeme jej předběžně charakterisovat takto: Prvky určitého základního souboru jsou postupně podrobeny pokusům, z nichž každý vede k určitému číslu x (nebo skupině m reálných čísel). O každém prvku se předpokládá, že patří do jedné z n tříd. Tyto třídy jsou charakterisovány n hustotami pravděpodobnosti $p_i(x)$ $i = 1, 2, \dots, n$. Na základě pozorované hodnoty x jest rozhodnouti, do které třídy dotyčný prvek patří, a toto rozhodnutí jest učinit s nejmenším možným risikem chyby.

Ukažme explicitní řešení jednoduchého problému, kdy je položena otázka, zda určitý prvek patří do jednoho ze dvou základních souborů. Zde existují zřejmě dvě alternativní hypotesy, z nichž je nutno jednu zvolit. Je tudíž třeba nějakého pravidla pro postup, pomocí něhož by mohl být individuální prvek přiřazen jedné nebo druhé skupině. Při užívání žádného takového pravidla nelze se vyhnout chybám.

Je-li α_1 zlomek špatných klasifikací pro prvky z první skupiny a α_2 zlomek špatných klasifikací pro prvky z druhé skupiny, pak zlomek špatných klasifikací v celku je $\pi_1\alpha_1 + \pi_2\alpha_2$, když považujeme prvky za náhodná pozorování z nějakého základního souboru, který obsahuje prvky těch dvou skupin v poměru $\pi_1 : \pi_2$, při čemž $\pi_1 + \pi_2 = 1$. Je jasno, že ten postup je nejlepší, který dává nejmenší možnou hodnotu výrazu $\pi_1\alpha_1 + \pi_2\alpha_2$.

Označme $f_1(x|\Theta_1)$ a $f_2(x|\Theta_2)$ hustoty pravděpodobnosti určitého měřitelného znaku (vlastnosti) ve dvou skupinách. Můžeme si představit, že symbol x zastupuje všechny dosažitelné měřitelné znaky a Θ všechny parametry. Potom nejlepší řešení, které dává minimum výrazu $\pi_1\alpha_1 + \pi_2\alpha_2$ je to, které přiřazuje určitý prvek první skupině S_1, jestliže

$$\pi_1 f_1(x|\Theta_1) \geqq \pi_2 f_2(x|\Theta_2) \tag{13.1}$$

a druhé skupině S_2, jestliže

$$\pi_1 f_1(x|\Theta_1) \leqq \pi_2 f_2(x|\Theta_2)\,. \tag{13.2}$$

Platí-li rovnost, lze tento případ rozhodnout na základě výsledku vrhu mincí.

Ztráta, vyplývající z toho, že se přiřadí určitý prvek z první skupiny druhé skupině, budiž Z_1, a ztráta Z_2, přiřadí-li se prvek z druhé skupiny do první. Pak nejlepší řešení je to, které dává minimum očekávané ztrátě $\pi_1\alpha_1 Z_1 + \pi_2\alpha_2 Z_2$. Nejlepší řešení, které minimalisuje očekávanou ztrátu je to, které přiřadí prvek první skupině, jestliže platí $\pi_1 z_1 f_1(x|\Theta_1) \geqq \pi_2 z_2 f_2(x|\Theta_2)$, a druhé, jestliže $\pi_1 z_1 f_1(x|\Theta_1) \leqq \pi_2 z_2 f_2(x|\Theta_2)$. Tato řešení jsou odvozena v práci Raově [10]. V každém daném problému lze veličiny z_1 a z_2, které vystupují v nejlepším řešení, snadno zjistit; apriorní pravděpodobnosti však jsou buď neznámé nebo odhadnuté z dat těžko získatelných. Ale i hrubý odhad jejich je užitečný v takových problémech klasifikace.

Postup při klasifikaci tudíž spočívá obvykle na náhodném výběru, na jehož základě se sestrojí určitá diskriminační funkce, pomocí níž se roztřídění prvků provede. Chceme-li docílit nejlepší diskriminace, je zřejmé, že budeme žádat, aby rozdíl středních hodnot diskriminační funkce v těch dvou skupinách byl co největší a rozptyl co nejmenší. Nechť tedy závisí klasifikační kriterium na r pozorovaných vlastnostech, takže f_1 a f_2 jsou hustoty r-rozměrného normálního zákona rozdělení s touž kovarianční maticí (μ_{ij}), k níž inversní matice je (μ^{ij}). Označme střední hodnoty i-té vlastnosti v první skupině μ_{i1} a ve druhé μ_{i2} jejich rozdíl $\mu_{i1} - \mu_{i2} \doteq \delta_i$. Potom vzhledem ku (13.1) a (13.2) musí na hranici platit

$$\frac{\pi_2}{\pi_1} = \frac{f_1}{f_2} = \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r}(x_i - \mu_{i1})(x_j - \mu_{j1})\,\mu^{ij}\right\}}{\exp\left\{-\frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r}(x_i - \mu_{i2})(x_j - \mu_{j2})\,\mu^{ij}\right\}}\,.$$

Pro logaritmy tudíž platí

$$2\lg\frac{\pi_2}{\pi_1} = \sum_{i=1}^{r}\sum_{j=1}^{r}\mu^{ij}\,[(x_i - \mu_{i2})(x_j - \mu_{j2}) - (x_i - \mu_{i1})(x_j - \mu_{j1})] =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r}\mu^{ij}\,[(\mu_{i1} - \mu_{i2})\,x_j + (\mu_{j1} - \mu_{j2})\,x_i + \mu_{i2}\mu_{j2} - \mu_{i1}\mu_{j1}]\,.$$

Matice (μ^{ij}) je symetrická, takže hledaná rovnice pro hranici je

$$\sum_{i=1}^{r}(\mu^{1i}\delta_1 + \mu^{2i}\delta_2 + \ldots + \mu^{ri}\delta_r)\,x_i = \text{konst.}$$

Nejlepší diskriminační funkce má tedy tvar $X = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_r x_r$, kde koeficienty λ_j splňují vztahy $\sum_{j=1}^{r} \lambda_j \mu_{ij} = \delta_i$, $i = 1, 2, \ldots, r$. Stanovíme nyní odhady l_j koeficientů λ_j na základě náhodného výběru, takže budeme uvažovat lineární funkci

$$X = l_1 x_1 + l_2 x_2 + \ldots + l_r x_r \,. \tag{13.3}$$

Kriterium, na němž založíme své rozhodování, bude spočívat na tom, že hodnoty l_j musejí býti zvoleny tak, aby se dosáhlo maxima zlomku, v jehož čitateli jsou rozdíly mezi výběrovými průměry a ve jmenovateli směrodatné odchylky v těch dvou skupinách.

Nechť tedy značí d_i rozdíl průměrů obou skupin v i-té vlastnosti, čili $d_i = x_{i1} - x_{i2}$. Pak rozdíl středních hodnot diskriminační funkce (13.3) bude $D = \sum_{i=1}^{r} l_i d_i$. Rozptyl V lineární funkce (13.3) bude dán výrazem $V = \sum_{i=1}^{r} \sum_{j=1}^{r} l_i l_j a_{ij}$, kde a_{ij} značí kovarianci proměnných x_i a x_j, o níž předpokládáme, že je táž pro oba základní soubory. Na základě těchto úvah bude tudíž vhodné určit koeficienty l_j tak, aby bylo dosaženo maxima zlomku $\dfrac{D^2}{V} = \dfrac{\left(\sum_{i=1}^{r} l_i d_i\right)^2}{\sum_{i=1}^{r} \sum_{j=1}^{r} l_i l_j a_{ij}}$.

Pro každé l tedy bude platit rovnice

$$\frac{1}{V^2}\left(2D \frac{\partial D}{\partial l} V - \frac{\partial V}{\partial l} D^2\right) = 0 \,,$$

což vede k soustavě rovnic

$$\frac{1}{2}\frac{\partial V}{\partial l_i} = \frac{V}{D}\frac{\partial D}{\partial l_i} \qquad i = 1, 2, \ldots, r \,. \tag{13.4}$$

Vzhledem k tomu, že

$$\frac{\partial D}{\partial l_i} = d_i \,, \qquad \frac{\partial V}{\partial l_i} = 2(l_1 a_{1i} + l_2 a_{2i} + \ldots + l_r a_{ri})$$

můžeme psáti soustavu (13.4) ve tvaru

$$\begin{aligned} l_1 a_{11} + l_2 a_{12} + \ldots + l_r a_{1r} &= \frac{V}{D} d_1 \,, \\ l_1 a_{21} + l_2 a_{22} + \ldots + l_r a_{2r} &= \frac{V}{D} d_2 \,, \\ \vdots \qquad \vdots \qquad\qquad \vdots \quad &\qquad \vdots \\ l_1 a_{r1} + l_2 a_{r2} + \ldots + l_r a_{rr} &= \frac{V}{D} d_r \,. \end{aligned} \tag{13.5}$$

Budeme psáti $\dfrac{V}{D} = h$. Poněvadž se diskriminační schopnost nezmění, užijeme-li místo dané diskriminační funkce libovolné její lineární funkce, můžeme položit $h = 1$. Potom řešení systému (13.5) dává hodnoty koeficientů l_i ve tvaru

$$l_i = a^{1i} d_1 + a^{2i} d_2 + \ldots + a^{ri} d_r, \quad i = 1, 2, \ldots, r \,.$$

Welch ukázal dále, že obecná diskriminační funkce v případě dvou alternativ je poměr věrohodností dvou hypotes a je možno odvodit ji buď z věty Bayesovy s danými apriorními pravděpodobnostmi nebo pomocí základní věty Neymanovy-Pearsonovy, kdy jsou chyby pro ty dvě hypotesy minimalisovány v určitém daném poměru.

Literatura

[1] W. Feller: *An introduction to probability theory and its applications.* Vol. 1. New York, John Wiley & Sons (1950).

[2] Fisher R. A. (1936): *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, London 7, str. 179.

[3] Fisher R. A. (1940): *The precision of discriminant functions.* Annals of Eugenics, London 10, str. 422.

[4] Hodges J. L. and Lehmann E. L. (1954): *Testing the approximate validity of statistical hypotheses.* Journal of the royal statistical society. Series B, Vol. 16, str. 261—268.

[5] Lehman E. L. (1950): *Some principles of the theory of testing hypotheses.* Annals of mathematical statistics 21, str. 1—26.

[6] Neyman J. (1935): *Su un teorema concernente le cosiddette statistiche sufficienti.* Giornale dell Istituto Italiano degli Attuari. Vol. VI, No 4.

[7] Neyman J. (1941): *On a statistical problem arising in routine analyses and in sampling inspections of mass production.* Annals of mathematical statistics 12, str. 46—76.

[8] Neyman J. and Pearson E. S. (1933): *On the problem of the most efficient tests of statistical hypotheses.* Phil. Trans. Roy. Soc. London, Series *A*, 231, str. 289—337.

[9] Neyman J. and Pearson E. S. (1936): *Sufficient statistics and uniformly most powerful tests of statistical hypotheses.* Statistical research memoirs I, str. 113—137.

[10] Rao R. (1948): *The utilization of multiple measurements in problems of biological classification.* Journal of the Royal Statistical Society, Series B, Vol. 10, str. 159.

[11] Rao R. (1950): *Statistical inference applied to classificatory problems.* Sankhyā 10, str. 229—256.

[12] Rao R. (1951): *The discriminant function approach in the classification of time series.* Sankhyā 11, str. 265.

[13] Roy S. N. (1947): *Notes on testing of composite hypotheses.* Sankhyā 8, str. 257—270.

[14] Roy S. N. (1948): *Notes on testing of composite hypotheses II.* Sankhyā 9, str. 19—38.

[15] Wald A. (1939): *Contributions to the theory of statistical estimation and testing hypotheses* Annals of Math. Stat. 10, str. 299.

[16] Wald A. (1942): *On the principles of statistical inference.* Notre Dame, Indiana.

[17] Wald A.: *Notes on the theory of statistical estimation and testing hypotheses.* Columbia university.

[18] Wald A. (1947): *Sequential analysis.* J. Wiley — New York.

[19] Wald A. (1950): *Statistical decision functions.* J. Wiley — New York.

[20] Welch B. L. (1939): *Note on discriminant functions.* Biometrika 31, str. 218.

ZOBRAZOVACÍ METODY V DESKRIPTIVNÍ GEOMETRII

Prof. dr A. Urban

(dokončení)

Všechna v běžné technické praxi se vyskytující dvojobrazová zobrazení jsou odvozena buď z dvojobrazového zobrazení v projektivním prostoru, v němž pomocné průmětny splývají (případ *c*), str. 30, č. 1) nebo ze zobrazení, v němž pomocné průmětny i jejich uzly jsou různé (případ *a*), str. 30 č. 1), ale hlavní střed S a pomocné středy ${}^1S$, ${}^2S$ jsou kolineární (obr. 13). Za předpokladu kolineárnosti středů promítání všech tří basí splývají uzlové body ${}^2S_1$, ${}^1S_2$ nákresny v jediný bod (neležící na základnici), který označujeme S_0 a často mu říkáme *hlavní bod* zobrazení. Také sdružené uzlové přímky nákresny splývají v jedinou přímku procházející ovšem hlavním bodem; říkáme jí