Jan Kalina

Ronald Fisher, otec biostatistiky

Pokroky matematiky, fyziky a astronomie, Vol. 57 (2012), No. 3, 186–190

Persistent URL: http://dml.cz/dmlcz/143200

Ronald Fisher, otec biostatistiky

Jan Kalina, Praha

Sir Ronald Aylmer Fisher (1890–1962) byl geniální anglický vědec, od jehož úmrtí si letos připomínáme 50 let. Věnoval se statistice, biologii (zejména genetice a evoluční biologii) i eugenice. Dodnes je právem považován za jednoho z největších statistiků všech dob a za své výsledky ve statistice i biologii byl povýšen do šlechtického stavu. Jeho fenomenální představivost se rozvinula také díky tomu, že měl od dětství velmi slabý zrak. Kvůli tomu byl donucen řešit i nejsložitější problémy zpaměti bez pomoci tužky a papíru.

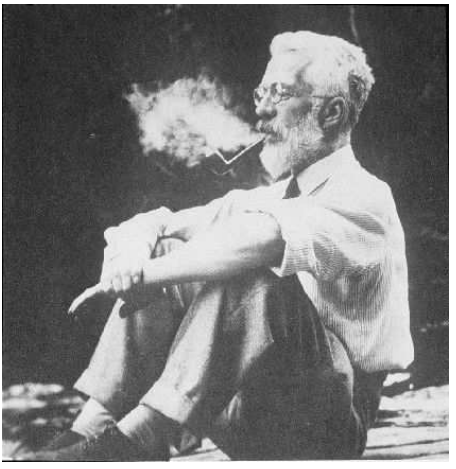
Ze základních statistických metod navrhl metodu maximální věrohodnosti, analýzu rozptylu, lineární diskriminační analýzu, vybudoval teorii navrhování experimentů (systematicky prostudoval randomizované experimenty) a přispěl k teorii testování hypotéz (zavedl permutační testy) i teorii odhadu (definoval Fisherovu informaci).

Fisherův výzkum ve statistice byl výrazně formován jeho osobními spory s věhlasným a populárním statistikem Karlem Pearsonem (1857–1936). V roce 1919 odmítl zaměstnání na University College London na Katedře aplikované statistiky vedené Pearsonem, přestože šlo o první a v té době jedinou statistickou katedru na světě [8]. Raději se stáhl do ústraní a přijal práci v zemědělské laboratoři, kde budoval své teoretické výsledky doslova na koleně. Profesuru na londýnské statistické katedře přijal teprve v roce 1933, když Karl Pearson odešel do důchodu.

R. A. Fisher kritizoval Pearsonovy statistické metody a odmítal je používat. Ilustrativním příkladem je Fisherův postoj k lékařské studii, která prokázala souvislost kouření s výskytem rakoviny za pomoci Pearsonova korelačního koeficientu. Fisher ve vědeckých článcích i v tisku napadl a ironizoval takové použití korelačního koeficientu a zcela obdobně prokázal souvislost mezi množstvím dovezených jablek a počtem rozvodů v Anglii [12]. Přirozeně se zde jedná o úmyslně nesprávnou interpretaci, protože korelace (statistická souvislost) nutně neznamená kauzální vztah mezi oběma veličinami, kdy se jedna z nich chová jako odezva druhé.

Fisher je považován za velmi významného genetika tehdejší doby, kdy stály proti sobě dva tábory biologů s protikladnými názory na zákony dědičnosti. Jedna skupina v čele s Pearsonem uznávala jen postupný evoluční vývoj, který má spojitý charakter. Naproti tomu stoupenci zapomenutého a znovu objeveného Johanna Gregora Mendela (1822–1884) poukazovali na diskrétní charakter dědičnosti, který je dán tím, že konkrétní dědičný znak buď je, nebo není přítomen. Fisher vnímal statistiku jako potenciál pro usmíření obou táborů. Na své vlastní farmě prováděl genetické experimenty s křížením rostlin a živočichů, při nichž prostudoval dědičnost spojitých znaků (výška, hmotnost), a ukázal, že je konzistentní s Mendelovými principy diskrétního

RNDr. Jan Kalina, Ph.D., Oddělení medicínské informatiky, Ústav informatiky AV ČR, v.v.i., Pod Vodárenskou věží 2, 182 07 Praha 8, e-mail: kalina@cs.cas.cz



Obr. 1. Ronald Fisher (1890–1962)

charakteru. V klíčovém Fisherově článku [4] je nejen vůbec poprvé definován rozptyl, ale vedle odvození analýzy rozptylu zde dokázal Fisher i založit kvantitativní genetiku a položit základy celé moderní biologie.

Je třeba také zmínit, že se Fisher snažil aplikovat pojmy selekce, genetická dominance a degenerace na člověka i na celé národy a rasy. To ho přivedlo ke studiu eugeniky, která měla za cíl zlepšení biologických a genetických vlastností člověka [9]. Fisher se stal profesorem eugeniky v Londýně. Byl i členem vlivné Eugenické společnosti, která hledala argumenty ve prospěch diskriminace společensky znevýhodněných. Eugenické představy se však posléze staly ideovou základnou nacistů, kteří je zneužili např. pro zdůvodňování perzekucí (či úplné likvidace) duševně nemocných a tělesně postižených. Fisher však ještě po druhé světové válce kritizoval snahy UNESCO o společenské zrovnoprávnění lidských ras po celém světě. To ovšem působil již jen na málo významné katedře genetiky.

Fisher bývá často nazýván otcem statistiky [11], přestože někteří autoři přisuzují stejný titul jiným statistikům, mezi něž patří zejména Adolphe Quetelet (1796–1874) nebo Gottfried Achenwall (1719–1772), viz [7]. Přitom je třeba říci, že se statistika až do 80. let 19. století zabývala především sbíráním faktů, které popisovaly činnost lidí ve státě [13] či stav společnosti a jejího rozvoje. K tomu patřilo také porovnávání takových faktů mezi různými státy [2]. Fisher má mimořádné zásluhy o vybudování matematických základů moderní statistiky a také o rozvoj statistických metod pro biologické aplikace. Právem proto můžeme považovat R. A. Fishera i za otce biostatistiky, i když se pojem biostatistika během Fisherova života ještě nepoužíval. Biostatistikou se nejčastěji rozumí aplikace statistiky do biologie, zejména navrhování a analýza klinických studií. V posledních letech se pojem biostatistika používá i pro bioinformatiku a výpočetní biologii, do které se někdy zařazují i analýzy genových expresí.

V tomto článku si připomeneme dvě velmi významné metody, které R. A. Fisher navrhl. Ukážeme, že i moderní postupy pro statistickou analýzu genetických dat věrně vycházejí z Fisherových myšlenek. Zbývající kapitoly se proto věnují statistickým úlohám v kvantitativní genetice, tedy oboru, který sám Fisher založil.

Využití Fisherova faktoriálového testu v populační genetice

Hardyovo-Weinbergovo ekvilibrium je základním zákonem populační genetiky [3], který popisuje rovnovážný stav v genetickém chování celé populace. Zde popíšeme statistický test hypotézy o Hardyově-Weinbergově ekvilibriu, který představuje využití původních Fisherových výsledků pro moderní genetické aplikace.

O konkrétním genu řekneme, že vyhovuje Hardyovu-Weinbergovu zákonu, pokud podíl jednotlivých verzí genu zůstává konstantní napříč generacemi [10]. Rovnováhu považujeme za nulovou hypotézu, pro kterou využijeme test založený na klasickém Fisherově faktoriálovém testu. Uvažujme jeden konkrétní gen, jehož dvě různé varianty (alely) označíme jako A, a. Obecně u daného genu dědí každý jedinec jednu alelu po otci a jednu po matce, a proto může mít jedinec jeden z možných genotypů AA, Aa, aa. Jako příklad uvažujme gen, který určuje takzvanou plášťovost srsti u skotu. Alela A odpovídá jednobarevnému zbarvení celého těla, kdežto alela a skvrnitosti. Přitom pouze zvíře s genotypem aa je strakaté, zatímco jedinci s ostatními genotypy AA a Aa jsou po celém těle jednobarevní. Jiným příkladem může být lidský gen HLA-DR5 s alelami A a a, který ovlivňuje produkci thyreotropního hormonu (TSH). Pouze genotyp aa se považuje za dispozici pro autoimunitní onemocnění štítné žlázy [1].

Označíme pravděpodobnost výskytu alely A jako p $(0 < p < 1)$ a pravděpodobnost alely a jako $1 - p$. Hardyova-Weinbergova rovnováha odpovídá nulové hypotéze, že se alely dědí náhodně a nezávisle (alela po otci nezávisí na alele po matce). Za tohoto předpokladu je pravděpodobnost genotypu AA rovna p^2, pravděpodobnost Aa je rovna $2p(1-p)$ a pro pravděpodobnost aa zbývá $(1-p)^2$.

Mějme náhodný výběr jedinců z uvažované populace. Zaveďme značení n_{AA} pro počet jedinců s genotypem AA, n_{Aa} pro počet jedinců Aa a n_{aa} pro počet jedinců aa. Označme četnosti jednotlivých alel v populaci n_{A} a n_{a}. Platí $n_{\mathsf{A}} = 2n_{\mathsf{AA}} + n_{\mathsf{Aa}}$ a $n_{\mathsf{a}} = n_{\mathsf{Aa}} + 2n_{\mathsf{aa}}$. Označme dále $n = n_{\mathsf{AA}} + n_{\mathsf{Aa}} + n_{\mathsf{aa}}$. Můžeme uvažovat čtyřpolní tabulku četností jedinců podle genotypu

	Po otci A	Po otci a
Po matce A	n_{AA}	n_{12}
Po matce a	n_{21}	n_{aa}

.

Zde vystupuje i četnost n_{21} těch jedinců, kteří dědí alelu A po otci a alelu a po matce, a také četnost n_{12} těch jedinců, kteří dědí alelu A po matce a alelu a po otci. Protože jsou však jednotlivé alely nerozlišitelné, hodnoty n_{12} a n_{21} nelze v praxi zjistit a známý je pouze jejich součet n_{Aa}.

Klasický Fisherův test je testem nezávislosti dvou znaků ve čtyřpolní tabulce a je vyčíslen jako pravděpodobnost vzniku dané tabulky za podmínky, že jsou pevné řádkové i sloupcové součty jednotlivých četností. Exaktní test nulové hypotézy, že platí Hardyova-Weinbergova rovnováha, se získá jako přímá aplikace Fisherova faktoriálového testu. Test je založen na výpočtu podmíněné pravděpodobnosti pro výskyt genotypu Aa za podmínky, že jsou pevně dány hodnoty n_{A} a n_{a}. Označme pomocí n_{Aa} pozorovanou četnost genotypu Aa, kterou chápeme jako realizaci náhodné veličiny N_{Aa}.

Můžeme vyjádřit pravděpodobnost

$$\begin{aligned}\mathsf{P}(N_{\mathsf{Aa}} = n_{\mathsf{Aa}}|n_{\mathsf{A}}, n_{\mathsf{a}}) &= \mathsf{P}(N_{12} + N_{21} = n_{\mathsf{Aa}}|n_{\mathsf{A}}, n_{\mathsf{a}}) = \\ &= \sum_j \mathsf{P}(N_{12} = j, N_{21} = n_{\mathsf{Aa}} - j|n_{\mathsf{A}}, n_{\mathsf{a}}) = \\ &= \frac{n!}{n_{\mathsf{AA}}! n_{\mathsf{Aa}}! n_{\mathsf{aa}}!} 2^{n_{\mathsf{Aa}}} \Big/ \binom{2n}{n_{\mathsf{A}}}, \end{aligned} \tag{1}$$

která je podmíněna pevnými hodnotami konstant n_{A} a n_{a}. Test platnosti Hardyova-Weinbergova zákona pro daný gen zamítá nulovou hypotézu, pokud pravděpodobnost $\mathsf{P}(N_{\mathsf{Aa}} \geq n_{\mathsf{Aa}}|n_{\mathsf{A}}, n_{\mathsf{a}})$ vyčíslená nasčítáním pravděpodobností (1) překročí předem zvolenou hladinu 5 %.

Fisherova lineární diskriminační analýza

Lineární diskriminační analýza navržená R. A. Fisherem představuje klasickou metodu mnohorozměrné statistiky, kterou však nelze použít ve většině molekulárně genetických aplikací. Musel by být totiž k dispozici větší počet pozorování než počet proměnných. Teprve nejmodernější klasifikační metody jsou schopny analyzovat desítky tisíc genů na desítkách nebo nejvýše stovkách vzorků. I takové metody přitom přirozeným způsobem vycházejí z Fisherových myšlenek. To ilustrujeme na příkladu metody SCRDA [6], která přímo vychází z lineární diskriminační analýzy.

Fisher navrhl lineární diskriminační analýzu (LDA) v roce 1936 při studiu polyploidie kosatců [5], konkrétně pro ověření hypotézy, že jeden druh kosatce je polyploidním křížencem dvou jiných druhů. Polyploidií se rozumí skutečnost, že kosatce mají zvýšený počet chromozómových sad (více než obvyklé dvě sady), což je právě důsledek křížení.

Předpokládejme, že máme k dispozici celkový počet K nezávislých náhodných výběrů p-rozměrných dat, kde k-tý výběr je tvořen pozorováními $\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kn_k}$. Odhad společné varianční matice spočtený ze všech pozorování označíme jako $\mathbf{S}$. Označme průměr naměřených dat v k-tém výběru pomocí $\bar{\mathbf{X}}_k$.

Klasifikační pravidlo lineární diskriminační analýzy lze zformulovat pomocí lineárních diskriminačních skórů $L_1(\mathbf{Z}), \ldots, L_K(\mathbf{Z})$ ve tvaru

$$L_k(\mathbf{Z}) = -\frac{1}{2}\bar{\mathbf{X}}_k^T \mathbf{S}^{-1} \bar{\mathbf{X}}_k + \bar{\mathbf{X}}_k^T \mathbf{S}^{-1} \mathbf{Z} + \log p_k, \quad k = 1, \ldots, K, \tag{2}$$

kde p_k značí apriorní pravděpodobnost jevu, že nové pozorování pochází z k-tého výběru. Dnes se standardně pro LDA předpokládá mnohorozměrné normální rozdělení dat a shodné varianční matice ve všech skupinách, přestože Fisher původně nepožadoval tyto předpoklady a také uvažoval speciální případ $p_1 = \cdots = p_K$. Klasifikační pravidlo pak zařadí p-rozměrné pozorování $\mathbf{Z}$ do k-té skupiny právě tehdy, když platí $L_k(\mathbf{Z}) \geq L_j(\mathbf{Z})$ pro každé $j = 1, \ldots, K$.

Článek [6] navrhl analogii lineární diskriminační analýzy, která modifikuje vzorec pro lineární diskriminační skóry (2) tak, aby mohl být použit pro vysoce rozměrná data. Metoda vychází z Fisherova vzorce (2), v němž nahrazuje průměry a varianční matice jinými statistickými odhady, které jsou odvozeny v rámci takzvaného *shrinkage* přístupu, což je obecná metodologie pro konstrukci statistických odhadů mnohorozměrných parametrů. Zde se jedná o lineární *shrinkage* (smrštěné) odhady, které mají

tvar lineární kombinace klasického odhadu s nějakým jednodušším odhadem, který je získán za dodatečných ale nesplněných předpokladů.

Klasifikační pravidlo metody SCRDA (*shrunken centroid regularized discriminant analysis*) [6] je založeno na lineárních klasifikačních skórech

$$L_k(\mathbf{Z}) = -\frac{1}{2}\left(\bar{\mathbf{X}}_k^*\right)^T \left(\mathbf{S}^*\right)^{-1} \bar{\mathbf{X}}_k^* + \left(\bar{\mathbf{X}}_k^*\right)^T \left(\mathbf{S}^*\right)^{-1} \mathbf{Z} + \log p_k, \quad k = 1, \ldots, K, \qquad (3)$$

kde smrštěný průměr $\mathbf{X}_k^*$ genové exprese v k-tém výběru je lineární kombinací klasického výběrového průměru a sdruženého průměru exprese daného genu přes všechny výběry. Dále $\mathbf{S}^*$ je smrštěná varianční matice, která má tvar $\mathbf{S}^* = \lambda\mathbf{S} + (1-\lambda)\boldsymbol{\mathcal{I}}$ s parametrem $\lambda \in [0,1]$, kde $\boldsymbol{\mathcal{I}}$ je jednotková matice. Je zaručeno, že matice $\mathbf{S}^*$ je regulární i pro vysoce rozměrná data. Přitom se váha pro kombinaci dvou různých odhadů optimalizuje tak, aby metoda měla maximální schopnost správně klasifikovat nezávislá data.

Získané klasifikační pravidlo je přirozenou modifikací Fisherovy LDA, která využívá moderní statistickou teorii odhadu. Můžeme říci, že současný rapidní rozvoj genetického poznání tak vychází z Fisherových myšlenek a mimo jiné i díky nim je dnes molekulární genetika disciplínou, která patří k nejrychleji se rozvíjejícím vědním oborům.

Poděkování. Příspěvek vznikl s podporou RVO: 67985807.

L i t e r a t u r a

[1] Akamizu, T., Kasuga, M., Davies, T. F.: *The genetics of complex thyroid diseases.* Springer, Tokyo, 2002.

[2] Aldrich, J.: *Mathematics in the London/Royal Statistical Society 1834–1934.* Electronic Journal for History of Probability and Statistics *6* (2010), 1–33.

[3] Emigh, T. H.: *A comparison of tests for Hardy-Weinberg equilibrium.* Biometrics *36* (1980), 627–642.

[4] Fisher, R. A.: *The correlation between relatives on the supposition of Mendelian inheritance.* Trans. Roy. Soc. Edinburgh *52* (1918), 399–433.

[5] Fisher, R. A.: *The use of multiple measurements in taxonomic problems.* Ann. Eugenics *7* (1936), 179–188.

[6] Guo, Y., Hastie, T., Tibshirani, R.: *Regularized discriminant analysis and its application in microarrays.* Biostatistics *8* (2007), 86–100.

[7] Hankins, F. H.: *Adolphe Quetelet as statistician.* Longmans, New York.

[8] Johnson, N. L., Kotz, S.: *Leading personalities in statistical sciences from the seventeenth century to present.* Wiley, New York, 1997.

[9] Kalina, J.: *100. výročí úmrtí Francise Galtona.* PMFA *56* (2011), 54–57.

[10] Katrnoška, F., Křížek, M.: *Genetický kód a teorie monoidů aneb 50 let od objevu struktury DNA.* PMFA *48* (2003), 207–222.

[11] Krishnan, T.: *Fisher's contribution to statistics.* Resonance *2* (1997), 32–36.

[12] Yates, F., Mather, K.: *Ronald Aylmer Fisher.* In Biographical Memoirs of Fellows of the Royal Society of London *9* (1963), 91–120.

[13] Zichová, J.: *Josef Erben a jeho přínos pro pražskou statistiku v 19. století.* PMFA *54* (2009), 57–71.