Učitel matematiky

Petr Emanovský

Jak viktoriánský polyhistor objevil korelační analýzu

Učitel matematiky, Vol. 30 (2022), No. 2, 65–76

Persistent URL: http://dml.cz/dmlcz/150453

JAK VIKTORIÁNSKÝ POLYHISTOR OBJEVIL KORELAČNÍ ANALÝZU

Petr Emanovský[1]

Úvod

16. únor 2022 představuje 200. výročí narození nadšeného přírodovědce Francise Galtona (1822–1911). Život a dílo této renesanční osobnosti jsou opravdu obdivuhodné. Galton je právem považován za průkopníka nových vědeckých metod, zejména v biologii, psychologii a statistice. Jeho široký vědecký zájem však zahrnoval i řešení problémů antropologie, meteorologie a kriminalistiky. Galtonovo jméno je známé především v souvislosti s jeho významným vynálezem – Galtonovou deskou (fazolový stroj, quincunx). Toto zařízení Galton sestrojil k ilustraci platnosti centrální limitní věty a zejména faktu, že při dostatečně rozsáhlém vzorku se binomické rozdělení blíží normálnímu (Stigler, 1989; Kalina & Soukup, 2019). Galton však přispěl k rozvoji statistiky celou řadou dalších originálních myšlenek. K jeho významným objevům v rámci statistiky patří bezesporu zavedení pojmů regrese a korelace. Je poněkud s podivem, že k objevu tak fundamentálních a široce užívaných pojmů došlo až na konci 19. století (Stigler, 1989). Pravdou také je, že dodnes používaný Galtonův klasický párový korelační koeficient je pojmenován po Karlu Pearsonovi, který v roce 1920 Galtonovy myšlenky upřesnil.

[1]Článek vznikl za podpory projektu Univerzity Palackého Olomouc „Algebraické a geometrické struktury“ IGA PrF 2021 030.

Analýza závislosti dvou náhodných veličin

Častým problémem statistické analýzy dat je zkoumání vzájemných vztahů dvou nebo i více náhodných veličin. Při studiu závislosti dvou proměnných využívá statistika metod regresní a korelační analýzy. Regresní analýza se snaží určit vztah mezi dvěma proměnnými (lineární, kvadratický apod.) a popsat jej graficky nebo funkčním předpisem. Hlavní motivací je při tom usuzovat z hodnot jedné veličiny (nezávisle proměnné) na hodnoty druhé veličiny (závisle proměnné). Korelační analýza se zaměřuje na vzájemnou sílu vztahu dvou veličin, aniž by zkoumala jeho kauzalitu. Vztahy mezi proměnnými jsou zkoumány graficky a vyjadřovány pomocí tzv. korelačních koeficientů. Pro různé situace statistického uvažování máme různé korelační koeficienty. K nejznámějším patří Pearsonův korelační koeficient, Spearmanův korelační koeficient pořadí a Kendallův koeficient pořadové korelace (Hendl, 2004). V praxi však dochází k častému prolínání metod regresní a korelační analýzy.

Pearsonův korelační koeficient

V případě jednorozměrného statistického souboru charakterizuje variabilitu n naměřených výsledků $x_1, \ldots, x_n$ náhodné veličiny X variance (rozptyl) určená vzorcem

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Pro dvourozměrný statistický soubor náhodných veličin X, Y definujeme analogicky tzv. kovarianci vztahem

$$\operatorname{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Uvažujme dvě náhodné veličiny X, Y, které mají obě alespoň přibližně normální rozdělení. Máme-li k dispozici náhodný výběr n párů měření (x_1, y_1), $\ldots$, (x_n, y_n) těchto veličin, můžeme určit

výběrové průměry $\bar{x}$, $\bar{y}$ a výběrové rozptyly s_X^2, s_Y^2. Nyní lze vypočítat výběrový (Pearsonův) korelační koeficient:

$$r_{XY} = \frac{\mathrm{cov}(X, Y)}{s_X s_Y}.$$

Hodnota výběrového korelačního koeficientu vyjadřuje sílu závislosti veličin X, Y v intervalu od -1 do 1. Hodnoty blízké nule indikují velmi slabý vzájemný vztah veličin, naopak hodnoty koeficientu blízké -1 nebo 1 odpovídají silnému lineárnímu vztahu. Záporná korelace ukazuje, že vysokým hodnotám jedné proměnné odpovídají nízké hodnoty druhé veličiny a naopak. Při kladné korelaci vysokým hodnotám odpovídají opět vysoké hodnoty (Komenda, 1994). Poznamenejme, že právě uvedený způsob zavedení korelačního koeficientu se zřejmě nejčastěji vyskytuje v základních učebnicích statistiky (Komenda, 1994; Hendl, 2004; Chráska, 2007; Hindls et al., 2007). K tomuto pojmu však lze dospět mnoha jinými způsoby (Rodgers & Nicewander, 1988). Po zavedení pojmu korelace zpravidla v učebnicích následuje vysvětlení základních pojmů regresní analýzy, většinou s důrazem na lineární regresi. Použitím metody nejmenších čtverců lze dokázat, že ze známých hodnot veličiny X lze nejlépe predikovat neznámé hodnoty veličiny Y pomocí lineární regresní funkce s předpisem $y = a + b_{YX}x$, kde $b_{YX} = r_{XY}\frac{s_Y}{s_X}$ je regresní koeficient (směrnice) regresní přímky a $a = y - b_{YX}x$ je funkční hodnota regresní přímky pro $x = 0$. Standardizací veličin X, Y můžeme dosáhnout toho, že $s_X = s_Y = 1$. Potom $b_{YX} = r_{XY}$ a korelační koeficient je přímo roven směrnici regresní přímky. Dále je zřejmé, že pro $r_{XY} = 0$ je také $b_{YX} = 0$, tedy nulové korelaci odpovídá regresní přímka rovnoběžná s osou x (Komenda, 1994).

Příklad. Pomocí Pearsonova koeficientu korelace zjistíme těsnost vztahu mezi výsledky testu z matematiky a z fyziky deseti studentů (viz tabulka 1).

Pro výpočet korelačního koeficientu je výhodné použít následující vzorec (Chráska, 2007):

$$r_{XY} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}.$$

S ohledem na tvar tohoto vzorce připravíme tabulku s dílčími výpočty:

Tab. 1: Výsledky testu z matematiky a fyziky (Chráska, 2007)

Žák i	Matematika x_i	Fyzika y_i	$x_i y_i$	x_i^2	y_i^2
1	7	12	84	49	144
2	10	14	140	100	196
3	1	8	8	1	64
4	9	15	135	81	225
5	6	12	72	36	144
6	4	9	36	16	81
7	8	13	104	64	169
8	3	13	39	9	169
9	8	16	128	64	256
10	5	13	65	25	169
	$\sum$ 61	$\sum$ 125	$\sum$ 811	$\sum$ 445	$\sum$ 1617

$$r_{XY} = \frac{10 \cdot 811 - 61 \cdot 125}{\sqrt{(10 \cdot 445 - 61^2)\,(10 \cdot 1617 - 125^2)}} \approx 0{,}77.$$

Vypočtená hodnota Pearsonova korelačního koeficientu indikuje vysokou závislost zkoumaných veličin.

Život Francise Galtona

Francis Galton se narodil 16. února 1822 v Birminghamu jako nejmladší z devíti dětí. Většinu svého života prožil v období vlády královny Viktorie (1837–1901) a byl vychováván jako typický viktorián. Do pěti let byl vzděláván doma svou invalidní sestrou a projevoval se jako geniální dítě. Později vystřídal různé školy.

Na přání rodiny začal studium medicíny, které však nedokončil. Zaujala ho matematika, kterou začal v roce 1840 intenzivně studovat v Cambridge. Studium však opět nezvládl a v roce 1844 se vrátil k medicíně. Po smrti otce však studia zanechal a stal se nadšeným cestovatelem. Mimo jiné podnikl tři významné cesty – do východní Evropy (1840), na Blízký východ (1845–1846) a do jihozápadní Afriky (1850–1852). Tyto cesty významně ovlivnily jeho pohled na svět a vzbudily u něj velký zájem o přírodní vědy (Bulmer, 2003). Již v roce 1854 se stal za svou práci v Africe uznávaným vědcem a členem Royal Geographical Society. V roce 1853 se Galton oženil s Luisou Butler (1822–1897), jejich manželství však bylo bezdětné. V roce 1909 byl Galton za svůj přínos vědě pasován na rytíře a o dva roky později, v 88 letech, zemřel.

Francis Galton všestranný přírodovědec

Francis Galton byl vskutku nadšeným polyhistorem, který se zajímal o vše, co souvisí s přírodou a člověkem. Svou vědeckou kariéru prožil jako prezident geografické sekce British Association (1862–1872), později působil jako prezident sekce antropologické (1877–1885). Výsledky svého bádání pravidelně publikoval formou krátkých článků, kterých se dochovalo více než 300 (Bulmer, 2003). Jednou z významných oblastí Galtonova vědeckého zájmu byla meteorologie. V roce 1862 zpracoval meteorologickou mapu Evropy a zabýval se problematikou předpovědi počasí. V roce 1863 publikoval stěžejní práci *Meteorographica*, která se stala významným dílem rozvíjející se moderní meteorologie. Grafické a statistické metody, které při zpracování této knihy použil, představují zřejmě důležitou přípravu na Galtonův pozdější výzkum dědičnosti (Bulmer, 2003). V rámci genetiky byl Galton silně ovlivněn prací svého bratrance Charlese Darwina. Nezajímal se však příliš o genetiku zvířat a rostlin, ale především o genetiku člověka. Své myšlenky shrnul v dílech *Hereditary Talent and Character* (1865) a *Hereditary Genius* (1869). Galton vytvořil vlastní originální teorii dědičnosti, která však byla formulována tak, že ji nebylo možné jednoduše experimentálně verifikovat (Gillham, 2001). Po určitém počátečním tápání dospěl po roce 1875 ke statistické teorii

dědičnosti, která již umožňovala seriózní vědecký výzkum. Vytvořil tzv. teorii dědičných mechanizmů, která ovšem byla později překonána objevy Gregora Johanna Mendela (1822–1884). Galton je rovněž považován za jednoho z prvních experimentálních psychologů. Zabýval se mimo jiné řešením některých netradičních problémů, studoval zejména individuální rozdíly lidských schopností. Zkoumal například korelaci mezi velikostí hlavy a inteligencí, dědičnost geniality nebo typy nadání. Galton byl přímo posedlý měřením a kvantifikací. Vyvinul řadu testů a měřicích metod, kterými obohatil experimentální vědu. Měl spoustu originálních nápadů, často je však nedotáhl do konce a přenechal je jiným. Poněkud méně známé, ale o to zajímavější, jsou některé Galtonovy objevy v oblasti kriminalistiky. Uveďme alespoň metodu překrývání fotografií a metodu identifikace osob pomocí otisků prstů. Poněkud kontroverzní je Galtonem vytvořená vědní disciplína zvaná eugenika, která je založena na myšlence zdokonalování lidské rasy pomocí řízené evoluce. Tato teorie brzy získala řadu příznivců i odpůrců po celém světě, později však byla zneužita německými nacisty. Dnes jsou některé Galtonovy názory na uspořádání společnosti hodnoceny jako zcela nepřijatelné (Kalina, 2011; Kalina & Soukup, 2019).

Francis Galton objevitel regrese a korelace

Galton dospěl k pojmům regrese a korelace při studiu dědičnosti. První myšlenky o regresi se objevují v jeho díle *Hereditary Genius* z roku 1869. Inspirován prací svého bratrance Charlese Darwina zkoumal a porovnával velikost a hmotnost dvou generací hrachu. Tento výzkum postupně vedl k odhalení pojmu regrese. Galton zjistil, že první generace extrémně velkých hrachů produkuje druhou generaci hrachů, které již nejsou tak extrémně velké a vykazují tedy „návrat k průměru“ (Hendl, 2004; Stanton, 2001). Později tuto regresi k průměru zaznamenal i při mezigeneračním srovnávání fyzických parametrů tisíců jedinců. Tento koncept upřesňoval následujících 16 let a v roce 1885 poprvé statisticky popsal pojem regrese k průměru. Tento pojem demonstroval na empirickém příkladu regrese výšky dětí k průměrné výšce jejich rodičů.

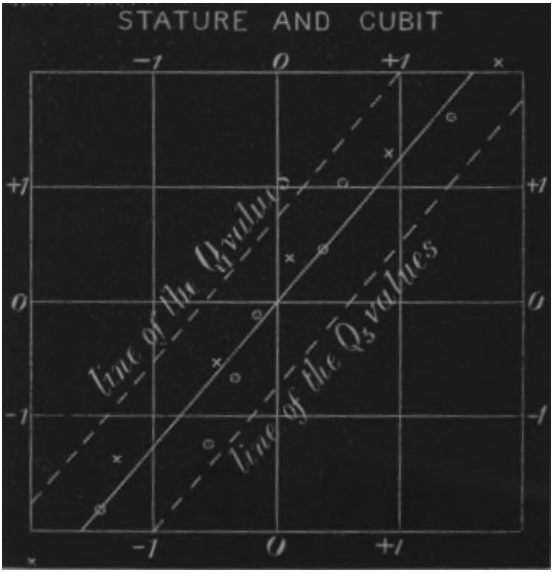
V roce 1888 Galton definoval související pojem korelace. Regresní a korelační analýza byly následně rozvíjeny dalšími statistiky. Největší měrou k tomuto rozvoji přispěli zejména Francis Ysidro Edgeworth (1845–1926), Karl Pearson (1857–1936) a George Udny Yule (1871–1951) (Fienberg, 1992). Galton navrhl korelační koeficient jako míru lineární regrese a definoval jej jako směrnici regresní přímky pro standardizované proměnné (Militký & Meloun, 2000; Fancher, 1989). Nutno poznamenat, že Galton pracoval převážně na úrovni deskriptivní statistiky, své teorie rozvíjel především na základě grafů a tabulek s minimem výpočtů. Jako míru centrální tendence používal medián místo aritmetického průměru a jako míru variability mezikvartilové rozpětí místo směrodatné odchylky (Pearson, 1920). To představovalo určitou výhodu zjednodušených výpočtů, na druhé straně to však pro výpočty znamenalo jisté omezení. Nebylo takto například možné pracovat s pojmem kovariance. Galtonova matematická erudice však zřejmě nebyla na takové úrovni, aby si tuto nevýhodu uvědomil. Bohužel, tento nedostatek zřejmě způsobil určité zpoždění rozšíření regresní analýzy do vědeckého výzkumu (Stanton, 2001). Galtonovy koncepty regresní a korelační analýzy dále rozvinuli další statistici, zejména Karl Pearson. Poznamenejme, že pro odhad regresních funkcí byla vyvinuta metoda nejmenších čtverců, kterou poprvé publikoval v roce 1805 A. M. Legendre (1752–1833), ale pravděpodobně ji použil již v roce 1795 C. F. Gauss (1777–1855) (Kotoučková, 2019).

Později Galton zaměřil svůj výzkum na širší uplatnění regresní metody, což postupně vedlo mimo jiné k zavedení pojmu korelace. Hlavní myšlenky týkající se této problematiky publikoval v roce 1888 v článku *Co-relations and their Measurement*. V roce 1890 pak napsal o tomto objevu populární stať s názvem *Kinship and correlation*. Ve své práci byl Galton silně motivován zejména dvěma okolnostmi. První se týkala snahy vyřešit dlouholetý antropologický problém určení výšky neznámého jedince na základě znalosti délky některé jeho kosti. Druhý motivační faktor souvisel s nově se rozvíjející metodou Bertillona umožňující identifikaci osob na základě jejich antropometrických dat (Bulmer, 2003).

V článku (Galton, 1888) je uveden příklad určení korelací dat uvedených v tabulce 2. Galton analyzoval data zachycující výšku postavy (proměnná y) a délku levého předloktí (proměnná x) u 348 dospělých mužů (tab. 2). Příslušné dvojice sobě odpovídajících hodnot zakreslil do dvojrozměrného grafu a zjistil lineární regresní závislost obou proměnných (obr. 1). Z grafu odhadl směrnici regresní přímky $b_{YX} = 2{,}5$. Tedy každému palci přírůstku délky předloktí odpovídá 2,5 palce přírůstku výšky postavy. Pro graf inverzní funkce zjistil hodnotu směrnice $b_{XY} = 0{,}26$, tj. každému palci přírůstku tělesné výšky odpovídá 0,26 palce přírůstku délky předloktí. Na základě tohoto pozorování Galton učinil závěr, že mezi oběma proměnnými existuje pozitivní vztah (korelace). Zbývalo dořešit problém, jak změřit sílu tohoto vztahu. Galton si uvědomoval, že klíčovou roli pro vyjádření síly tohoto vztahu hraje směrnice regresní přímky. Velikost této směrnice však závisí na variabilitě obou proměnných. S rostoucí variabilitou proměnné Y se směrnice zvětšuje a s klesající variabilitou proměnné Y se směrnice zmenšuje. Řešení tohoto problému založil Galton na standardizaci obou proměnných, tedy na jejich vyjádření v nových jednotkách závislých na jejich rozptylech (Stanton, 2001). Věděl, že směrodatná odchylka výšky je $s_Y = 1{,}75$ palců a směrodatná odchylka délky předloktí je $s_X = 0{,}56$ palců. Nové standardizované proměnné $\bar{X}$, $\bar{Y}$ získal vydělením původních proměnných jejich směrodatnými odchylkami. Tedy $\bar{X} = \frac{X}{s_X}$ a $\bar{Y} = \frac{Y}{s_Y}$. Jednotkou standardizované proměnné $\bar{X}$ tedy již není palec, ale 0,56 palce a podobně jednotkou proměnné $\bar{Y}$ je 1,75 palce. Standardizací získáme stejné rozptyly obou veličin, tj. $s_{\bar{X}} = s_{\bar{Y}} = 1$. V důsledku toho máme i stejné regresní koeficienty standardizovaných proměnných, tedy $b_{\bar{Y}\bar{X}} = b_{YX} \cdot \frac{s_X}{s_Y} = 2{,}5 \cdot \frac{0{,}56}{1{,}75} = 0{,}80$ a $b_{\bar{X}\bar{Y}} = b_{XY} \cdot \frac{s_Y}{s_X} = 0{,}26 \cdot \frac{1{,}75}{0{,}56} \approx 0{,}80$. Společná hodnota standardizovaných koeficientů umožňuje vyjádřit korelaci r (tab. 3). Směrodatná odchylka reziduované chyby standardizované regresní přímky je $s_{\bar{y}\bar{x}} = \sqrt{1 - r^2}$. Tedy pro standardizované hodnoty vychází směrodatná odchylka 0,60. Pro nestandardizované je to $0{,}60 s_y = 1{,}05$ (Galton, 1888; Bulmer, 2003).

Tab. 2: Tělesná výška a délka levého předloktí 348 mužů (Galton, 1888)

Stature in inches.	Length of left cubit in inches, 348 adult males.								Total cases.
	Under 16·5.	16·5 and under 17·0.	17·0 and under 17·5.	17·5 and under 18·0.	18·0 and under 18·5.	18·5 and under 19·0.	19·0 and under 19·5.	19·5 and above.	
71 and above .	..	..	..	1	3	4	15	7	30
70.....	..	..	..	1	5	13	11	..	30
69.....	..	1	1	2	25	15	6	..	50
68.....	..	1	3	7	14	7	4	2	48
67.....	..	1	7	15	28	8	2	..	61
66.....	..	1	7	18	15	6	..	..	48
65.....	..	4	10	12	8	2	..	..	36
64.....	..	5	11	2	3	..	..	..	21
Below 64.....	9	12	10	3	1	..	..	..	34
Totals	9	25	49	61	102	55	38	9	348



Obr. 1: Galtonův graf regresních přímek (Galton, 1888)

Tab. 3: Výsledky korelační analýzy tělesných parametrů 348 mužů (Galton, 1888)

Subject.	Relative.	In units of Q.		In units of ordinary measure.	
		r.	$\sqrt{(1-r^2)} = f$.	As 1 to	f.
Stature	Cubit	0·8	0·60	0·26	0·45
Cubit.....	Stature			2·5	1·4
Stature	Head length.....	0·35	0·93	0·38	1·63
Head length....	Stature			3·2	0·17
Stature	Middle finger	0·7	0·72	0·06	0·10
Middle finger...	Stature			8·2	1·26
Middle finger...	Cubit	0·85	0·61	3·13	0·34
Cubit.....	Middle finger			0·21	0·09

Závěr

Francis Galton byl bezesporu výjimečnou vědeckou osobností, která obohatila lidské poznání v mnoha směrech. K rozvoji statistických metod přispěl nejen objevem regresní a korelační analýzy, ale zavedl některé další důležité statistické pojmy, které se používají dodnes. Jmenujme například Galtonovu křivku kumulativní četnosti (ogive curve), kvartilovou odchylku, medián nebo percentilový systém (Fitzpatrick, 1960). Velkou měrou se rovněž zasloužil o rozvoj grafických metod ve statistice. Statistické metody se snažil aplikovat v nejrůznějších oblastech vědy, zejména v biologii a psychologii.

Literatura

[1] Bulmer, M. (2003). *Francis Galton Pioneer of Heredity and Biometry.* The John Hopkins University Press.

[2] Fancher, R. E. (1989). Galton on Examinations: An Unpublished Step in the Invention of Correlation. *The University of Chicago Press on behalf of The History of Science Society, 80*(3), 446–455.

[3] Fienberg, S. E. (1992). A brief history of statistics in three and one-half chapters: a review essay. *Statistical Science, 7*(2), 208–225.

[4] Fitzpatrick, P. J. (1960). Leading British Statisticians of the Nineteenth Century. *Journal of American Statistical Association, 55*(289), 38–70.

[5] Galton, F. (1888). *Co-relations and their Measurement, chiefly from Anthropometric Data.* Dostupné z `https://royalsocietypublishing.org/doi/pdf/10.1098/rspl.1888.0082`.

[6] Gillham, N. W. (2001). *A Life of Sir Francis Galton (From African Exploration to the Birth of Eugenics).* Oxford University Press.

[7] Hendl, J. (2004). *Přehled statistických metod – zpracování dat.* Portál.

[8] Hindls, R., Hronová, S., Seger, J., & Fischer, J. (2007). *Statistika pro ekonomy.* Professional Publishing.

[9] Chráska, M. (2007). *Metody pedagogického výzkumu.* Portál.

[10] Kalina, J. (2011). 100. výročí úmrtí Francise Galtona. *Pokroky matematiky, fyziky a astronomie, 56*(1), 54–57.

[11] Kalina, J., & Soukup, L. (2019). Průkopníci statistiky ve vědách o člověku v 19. století. *Bulletin České statistické společnosti, 30*(3), 1–15.

[12] Kotoučková, H. (2019). Předchůdci metody nejmenších čtverců. *Pokroky matematiky, fyziky a astronomie, 64*(1), 55–63.

[13] Komenda, S. (1994). *Biometrie.* VUP.

[14] Militký, J., & Meloun, M. (2000). *Korelace, její využití a zneužití.* Dostupné z `https://meloun.upce.cz/docs/publication/085.pdf`

[15] Pearson, K. (1920). Notes on the History of Correlation. *Biometrika, 13*(1). 25–45.

[16] Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen Ways

to Look at the Correlation Coefficient. *The American Statistician, 42*(1), 59–66.

[17] Stanton, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education, 9*(3), 1–13.

[18] Stigler, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science, 4*(2), 73–86.

Abstract

The regression and correlation analysis is among the most commonly used methods of statistical data processing. The discovery of this method is associated with the name of the genius thinker of the 19th century Francis Galton. The article focuses on the life and work of this important scientist and especially on his ideas when creating the basic concepts of the regression and correlation analysis.

Petr Emanovský
Přírodovědecká fakulta Univerzity Palackého v Olomouci
17. listopadu 1192/12
771 46 Olomouc
e-mail: petr.emanovsky@upol.cz