Pokroky matematiky, fyziky a astronomie

Vítězslav Línek

Rozdělení t a mnohorozměrná geometrie

Pokroky matematiky, fyziky a astronomie, Vol. 64 (2019), No. 2, 115–119

Persistent URL: http://dml.cz/dmlcz/147805

Rozdělení t a mnohorozměrná geometrie

Vítězslav Línek

Abstrakt. V článku odvozujeme hustotu t rozdělení s využitím n-rozměrné geometrie. Oproti obvyklejším metodám k tomu nepotřebujeme předpoklad normality, postačující je nezávislost mnohorozměrného rozdělení na směru. Kromě základů diferenciálního počtu použijeme k odvození jen vzorec pro povrch n-rozměrné koule. Tento přístup byl inspirován metodami R. A. Fishera.

1. Motivace

Jedním ze základních kamenů matematické statistiky je následující vztah: má-li vektor

$$\boldsymbol{X} = (Y, Z_1, \ldots, Z_n)$$

mnohorozměrné standardizované normální rozdělení, tj. mají-li jeho jednotlivé souřadnice rozdělení $\mathsf{N}(0, 1)$ a jsou nezávislé, pak náhodná veličina

$$T = \frac{Y}{\sqrt{\sum_{i=1}^{n} Z_i^2 \Big/ n}}$$

má tzv. Studentovo neboli t rozdělení s n stupni volnosti, zkráceně $T \sim t_n$, které se používá například v jednovýběrovém t testu o střední hodnotě v normálním rozdělení s neznámým rozptylem. Hustota tohoto rozdělení, daná vzorcem

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad t \in \mathbb{R},$$

se obvykle odvozuje pomocí integrálního počtu (viz např. [1]). Zde nabízíme alternativní cestu, která by mohla zaujmout čtenáře se vztahem k vícerozměrné geometrii. Kromě trochy představivosti a znalosti derivací budeme potřebovat vzorec na výpočet objemu n-sféry o poloměru r:

$$S_n(r) = \frac{2\pi^{\frac{n+1}{2}} r^n}{\Gamma\left(\frac{n+1}{2}\right)}, \quad r > 0$$

(viz např. [2]). Připomeňme, že n-sférou o poloměru r míníme povrch $(n+1)$-rozměrné koule, tj. množinu bodů v $(n+1)$-rozměrném eukleidovském (pod)prostoru, jejichž vzdálenost od daného středu je r. Její objem je tedy n-rozměrná míra.

Na rozdíl od tradičního postupu vlastně ani nepotřebujeme předpoklad normality vektoru $\boldsymbol{X}$; bude postačovat, když rozdělení tohoto vektoru nebude závislé na směru,

Mgr. Vítězslav Línek, Ph.D., FZŠ Mezi Školami 2322, 155 00 Praha 13,
e-mail: vitek.linek@seznam.cz

tj. jeho hustota bude funkcí vzdálenosti od počátku souřadnic. Mnohorozměrné standardizované normální rozdělení tuto definici zřejmě splňuje, neboť jeho hustota je

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{\|\boldsymbol{x}\|^2}{2}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Poznamenejme, že z existence hustoty uvažovaného rozdělení plyne vztah $\mathsf{P}\,[\boldsymbol{X} = \boldsymbol{0}] = 0$.

2. T jako funkce úhlu

Položme

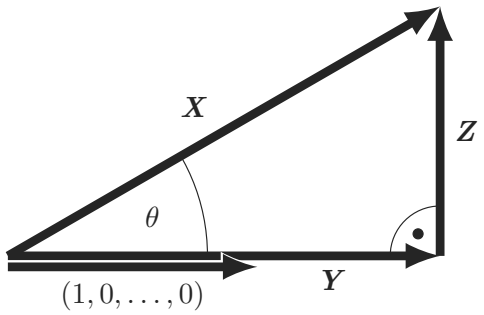
$$\boldsymbol{Y} = (Y, 0, \ldots, 0), \quad \boldsymbol{Z} = (0, Z_1, \ldots, Z_n),$$

takže můžeme psát

$$T = \operatorname{sgn} Y \sqrt{n} \frac{\|\boldsymbol{Y}\|}{\|\boldsymbol{Z}\|}.$$

Vektory $\boldsymbol{Y}$, $\boldsymbol{Z}$ tvoří odvěsny pravoúhlého trojúhelníku, jehož přeponou je vektor $\boldsymbol{X}$. Odhlédneme-li tedy od znaménka a konstanty $\sqrt{n}$, představuje hodnota T kotangentu úhlu mezi vektory $\boldsymbol{Y}$ a $\boldsymbol{X}$. Vezmeme-li znaménko a konstantu v potaz, můžeme T považovat za $\sqrt{n}$-násobek kotangenty úhlu, který svírá vektor $\boldsymbol{X}$ s vektorem $(1, 0, \ldots, 0)$. Označme tento úhel symbolem θ (viz obr. 1), máme tedy

$$T = \sqrt{n}\ \operatorname{cotg} \theta.$$



Obr. 1. Rozložíme-li vektor $\boldsymbol{X} = (Y, Z_1, \ldots, Z_n)$ na jeho kolmé složky $\boldsymbol{Y} = (Y, 0, \ldots, 0)$ a $\boldsymbol{Z} = (0, Z_1, \ldots, Z_n)$, můžeme T interpretovat jako funkci úhlu θ, který svírá vektor $\boldsymbol{X}$ s vektorem $(1, 0, \ldots, 0)$

Uvažujme nyní libovolnou hodnotu $t \in \mathbb{R}^+$: množina všech realizací náhodného vektoru $\boldsymbol{X}$ odpovídajících případu $T = t$ je tvořena právě těmi vektory, které svírají s vektorem $(1, 0, \ldots, 0)$ úhel

$$\theta = \operatorname{arccotg} \frac{t}{\sqrt{n}}.$$

Infinitezimálnímu přírůstku t o hodnotu $\mathrm{d}t$ odpovídá změna úhlu θ

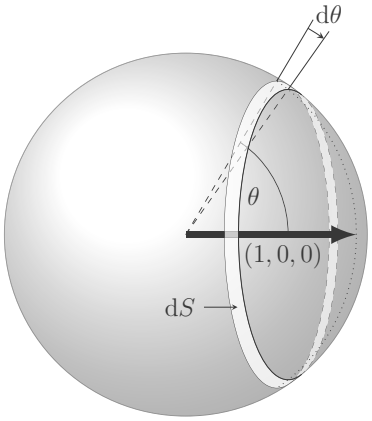
$$\mathrm{d}\theta = -\frac{\mathrm{d}t}{\sqrt{n}\left(1 + \frac{t^2}{n}\right)}.$$

3. Trojrozměrný případ

Abychom si trochu usnadnili další výklad, popišme nejdříve případ $n = 2$, kdy realizacemi vektoru $\boldsymbol{X}$ jsou prvky prostoru $\mathbb{R}^3$. Uvažujme jednotkovou kouli se středem v počátku soustavy souřadnic. Vektory svírající s vektorem $(1, 0, 0)$ úhel dané velikosti θ protínají povrch této koule v kružnici o poloměru $r = \sin\theta$. Tato kružnice je vlastně sféra dimenze 1, jejíž délka je

$$l = 2\pi \sin\theta = S_1(\sin\theta).$$

Zvětšíme-li nyní t o infinitezimální hodnotu $\mathrm{d}t$, změní se θ o hodnotu $\mathrm{d}\theta$. Jevu $T \in (t;\ t + \mathrm{d}t)$ tedy odpovídá množina realizací, které protínají povrch jednotkové koule v pásu délky l a infinitezimální šířky $-\mathrm{d}\theta$ (viz obr. 2).



Obr. 2. Je-li $n = 2$, realizace náhodného vektoru $\boldsymbol{X}$ jsou prvky $\mathbb{R}^3$. Jelikož je T klesající funkcí úhlu θ mezi vektorem $\boldsymbol{X}$ a vektorem $(1, 0, 0)$, vymezují všechny realizace $\boldsymbol{X}$ odpovídající případu $T \in (t; t + \mathrm{d}t)$ na povrchu jednotkové koule se středem v počátku soustavy souřadnic množinu tvaru pásu o obsahu $\mathrm{d}S = -2\pi \sin\theta \cdot \mathrm{d}\theta$, kde θ i $\mathrm{d}\theta$ jsou funkcí t a $\mathrm{d}t$

Pointou našich úvah je to, že jelikož rozdělení náhodného vektoru $\boldsymbol{X}$ nezávisí na směru, můžeme pravděpodobnost tohoto jevu vypočítat jako podíl povrchu tohoto pásu $\mathrm{d}S$ vůči povrchu celé koule $S_2(1)$:

$$\mathrm{d}\mathsf{P} = \frac{\mathrm{d}S}{S_2(1)} = \frac{-l\ \mathrm{d}\theta}{S_2(1)} = -\frac{S_1(\sin\theta)}{S_2(1)} \cdot \mathrm{d}\theta.$$

Dosadíme za θ, $\mathrm{d}\theta$ a po jednoduché úpravě[1] dostáváme

$$\mathrm{d}\mathsf{P} = \frac{2\pi \sin\left(\operatorname{arccotg} \frac{t}{\sqrt{2}}\right)}{4\pi} \cdot \frac{\mathrm{d}t}{\sqrt{2}\left(1 + \frac{t^2}{2}\right)} = \ldots = 2^{-\frac{3}{2}} \left(1 + \frac{t^2}{2}\right)^{-\frac{3}{2}} \mathrm{d}t;$$

část před diferenciálem je hledaná hustota rozdělení t_2.

[1] Připomeňme vztahy $\sin(\operatorname{arccotg} t) = \dfrac{1}{\sqrt{1+t^2}}$, $\cos(\operatorname{arccotg} t) = \dfrac{t}{\sqrt{1+t^2}}$.

4. Zobecnění

V obecném případě $n \in \mathbb{N}$ stačí nahradit jednotkovou kouli jednotkovou hyperkoulí, jejíž povrch tvoří n-sféra o objemu $S_n(1)$, a kružnici na jejím povrchu $(n-1)$-sférou o poloměru $r = \sin\theta$, jejíž „délka“ je $l = S_{n-1}(\sin\theta)$. Po dosazení dostáváme

$$\mathrm{d}\mathsf{P} = \frac{-S_{n-1}(\sin\theta)\cdot \mathrm{d}\theta}{S_n(1)} = \ldots = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}\ \mathrm{d}t,$$

kde část před diferenciálem je hledaná hustota rozdělení t_n.

5. Historické pozadí

Výše uvedená úvaha byla inspirována krátkou poznámkou v životopise R. A. Fishera [3], který napsala jeho dcera. Autorka se v ní na straně 123 snaží vysvětlit, jak Fisher odvodil test známý dnes jako jednovýběrový t-test. Uvažuje jednoduchý model lineární regrese bez konstanty s pouhými třemi dvojicemi pozorovaných dat, tj. model

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \beta \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix},$$

stručněji $\boldsymbol{Y} = \beta\boldsymbol{x} + \boldsymbol{\epsilon}$, kde náhodné veličiny ϵ_i jsou nezávislé a mají normální rozdělení a stejný rozptyl. Nulová hypotéza H_0: $\beta = 0$ pak představuje předpoklad, že platí $\boldsymbol{Y} = \boldsymbol{\epsilon}$, tj. že vektor $\boldsymbol{Y}$ leží v libovolném směru se stejnou pravděpodobností.[2] Tato hypotéza bude zřejmě zamítnuta tehdy, když bude směr získané realizace náhodného vektoru $\boldsymbol{Y}$ „podezřele“ blízký směru vektoru $\boldsymbol{x}$, tj. když bude úhel mezi vektory $\boldsymbol{Y}$ a $\boldsymbol{x}$ malý.

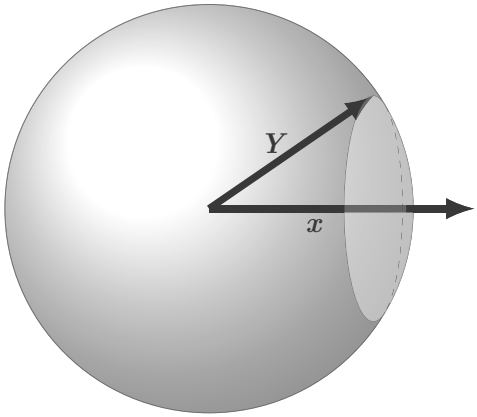
To se ovšem může stát i pouhou náhodou, přestože H_0 platí. Je tedy třeba nalézt vhodné kritérium tak, aby v případě platnosti H_0 došlo k omylu nanejvýš v 5 % případů, tj. aby nebyla překročena obvykle požadovaná pravděpodobnost chyby prvního druhu. Necháme tedy – tak jako to údajně udělal Fisher – získanou realizaci $\boldsymbol{Y}$ rotovat kolem vektoru $\boldsymbol{x}$, čímž opíše na sféře se středem v počátku a poloměrem $\|\boldsymbol{Y}\|$ hranici vrchlíku (viz obr. 3).

Je-li relativní obsah tohoto vrchlíku vůči obsahu celé sféry menší než 5 %, zamítneme nulovou hypotézu. Po zobecnění na více rozměrů představuje tato myšlenka podstatu klasického jednovýběrového t-testu.

Přeformulujme tuto myšlenku vlastními slovy: zmiňovaný relativní podíl obsahu vrchlíku je v podstatě *dosažená hladina významnosti* (p-hodnota) t-testu ověřujícího nulovou hypotézu $\beta = 0$. Kladná hodnota její derivace podle t tedy musí být hustota T – a to je právě způsob, jakým jsme k ní dospěli. Obešli jsme však p-hodnotu a místo toho vyjádřili přímo její derivaci.[3]

[2] Takto formuluje své úvahy J. F. Box; vhodnější vyjádření by bylo, že vektor $\boldsymbol{Y}/\|\boldsymbol{Y}\|$ má rovnoměrné rozdělení na jednotkové sféře.

[3] Stojí za zmínku, že z hlediska statistické praxe je to právě p-hodnota, kterou je třeba vyčíslit v závislosti na t. To je ale úkol mnohem obtížnější než vyjádření hustoty.



Obr. 3. Získáme-li za platnosti trojrozměrného modelu $\boldsymbol{Y} = \beta\boldsymbol{x} + \boldsymbol{\epsilon}$ realizaci náhodného vektoru $\boldsymbol{Y}$, zamítneme nulovou hypotézu H_0: $\beta = 0$ tehdy, když budou vektory $\boldsymbol{Y}$ a $\boldsymbol{x}$ ležet přibližně ve stejném směru. Necháme-li rotovat získanou realizaci vektoru $\boldsymbol{Y}$ kolem vektoru $\boldsymbol{x}$, opíše na sféře se středem v počátku soustavy souřadnic a poloměrem $\|\boldsymbol{Y}\|$ hranici vrchlíku; podíl obsahu tohoto vrchlíku vůči povrchu celé koule představuje hladinu významnosti tohoto testu, tj. pravděpodobnost, že by vektory $\boldsymbol{Y}$ a $\boldsymbol{x}$ mohly ležet tak blízko sebe nebo blíže pouhou náhodou

V obecnějším případě, kdy je model vícerozměrný, je analogická úvaha základem F-testu jako hlavního výsledku analýzy rozptylu; úhel θ zde však nepředstavuje odchylku od vektoru $(1, 0, \ldots, 0)$, nýbrž odchylku od lineárního podprostoru dimenze $m < n$, což klade o něco větší nároky na geometrickou představivost. Případné zájemce o řešení tohoto problému odkazujeme na dizertační práci [6].

Nebyli jsme schopni nalézt ve Fisherově díle žádné přímé potvrzení zde prezentované myšlenky. Avšak náš geometrický způsob uvažování, pomocí kterého jsme hustotu rozdělení t odvodili, je nápadně podobný postupům, které Fisher aplikoval v pracích [4] a [5]. Obzvláště článek [4] je fascinující ukázkou použití geometrie v matematické statistice a rozhodně stojí za hlubší studium.

L i t e r a t u r a

[1] Anděl, J.: *Základy matematické statistiky.* MatfyzPress, Praha, 2005.

[2] Blumenson, L. E.: *A derivation of n-dimensional spherical coordinates.* Amer. Math. Monthly 67 (1960), 63–66.

[3] Box, J. F.: *R. A. Fisher. The life of a scientist.* John Wiley, New York, 1978.

[4] Fisher, R. A.: *A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error.* Monthly Notices Roy. Astronom. Soc. 80 (1920), 758–770.

[5] Fisher, R. A.: *Note on Dr Burnside's recent paper on error of observation.* Proc. Cambridge Philos. Soc. 21 (1923), 655–658.

[6] Línek, V.: *Geometrie lineárního modelu.* Dizertační práce. MFF UK, Praha, 2016.