OBZORY MATEMATIKY, FYZIKY A INFORMATIKY 2/2022 (51)

Gaussovo (normální) rozdělení

Zdeněk Půlpán, Ondřej Slavíček

Abstract [Gaussian distribution]: In geodetic measurements, Gauss used the experience that the distribution of the measured quantities corresponds to the previously known bell curve to estimate the accuracy of the measurement. Here it is first shown how the analytical expression of such a curve can be easily obtained and then attention is drawn to some applications in technology, psychology and pedagogy.

Key words: Gaussian distribution, estimation of in-plane measurement accuracy, technical and psychological measurements.

Souhrn: Při geodetických měřeních Gauss používal k odhadu přesnosti měření zkušenost, že rozdělení naměřených veličin odpovídá již dříve známé zvonovité křivce. Zde je nejprve ukázáno, jak lze k analytickému vyjádření takové křivky snadno dojít a pak je upozorněno na některé aplikace v technice, v psychologii a pedagogice.

Klíčová slova: Gaussovo rozdělení, odhad přesnosti měření v rovině, technická a psychologická měření.

MESC: A30, B50, C70, D20, E10, F70.

Úvod

K. F. Gauss (1777 – 1855) byl sice „čistým“ matematikem, měl však i cit pro aplikace. Přispěl k pokroku v numerické matematice, v teorii elektromagnetismu i v geodézii (organizoval rozsáhlá geodetická měření). Již před Gaussem se typickou zvonovitou křivkou zabýval Francouz Abraham de Moivre (1667 – 1704). Gauss však pro tuto zvonovitou křivku odvodil v praxi použitelný vzorec a prozkoumal další její vlastnosti. Zvonovitá funkce se mu hodila pro odhad chyby při geodetických měřeních. Ukážeme, proč je uvedená křivka pro geodety (a nejen pro ně) tak důležitá a jaký je rozdíl v popisu chyby jednorozměrné od chyby dvourozměrné. Gaussovo rozdělení je limitním rozdělením pro mnohá jiná rozdělení a vyhovuje jisté extremální podmínce pro Shannonovu neurčitost. P. S. Laplace (1749 – 1827) předvedl metodu výpočtu některých integrálů, souvisejících s onou zvonovitou křivkou. Proto se někdy Gaussovo rozdělení nazývá také Gauss-Laplaceovo.

1 Přesnost geodetických měření

Uvažujme o identifikaci bodu A v rovině se souřadnicemi x, y (viz obr. 1), kde počátek soustavy souřadné budeme mít v bodě O. Identifikace bodu A je závislá na náhodě a jeho souřadnice jsou hodnotami dvourozměrné náhodné veličiny (X, Y). Předpokládejme, že měřená souřadnice x bodu A padne do intervalu $\langle x, x + \Delta x\rangle$ s pravděpodobností $f(x) \cdot \Delta x$, kde $f(x)$ je spojitá funkce hustoty pravděpodobnosti v intervalu $(-\infty; +\infty)$ se střední hodnotou v bodě 0 a také podobně i pro souřadnici y v intervalu $\langle y, y + \Delta y\rangle$ s pravděpodobností $f(y) \cdot \Delta y$, kde $f(y)$ je hustota pravděpodobnosti měřené veličiny Y také se střední hodnotou v bodě 0. Předpokládáme, že chyby ve směru souřadné osy x jsou nezávislé na chybách v ose y, pravděpodobnost určení bodu A ve vyšrafovaném diferenciálním obdélníku se stranami $\Delta x, \Delta y$ (viz obr. 1) je proto rovna součinu

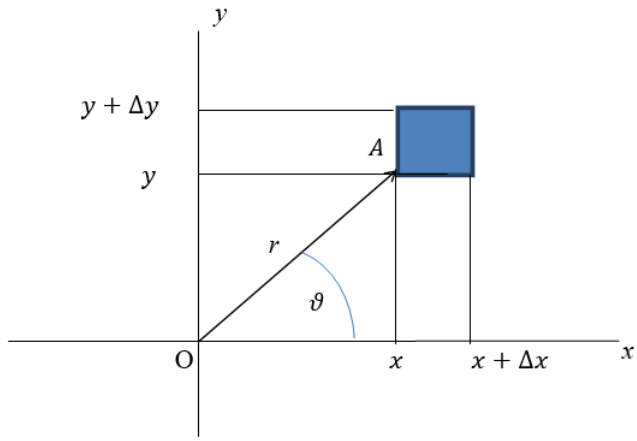
$$f(x) \cdot \Delta x \cdot f(y) \cdot \Delta y. \tag{1}$$

Protože předchozí pravděpodobnost nezávisí na směru (úhlu ϑ), ale pouze na vzdálenosti od bodu O, můžeme ji psát také pomocí funkce $g(r)$, kde r je vzdálenost bodu O od vyšrafovaného obdélníčku:

$$g(r) \cdot \Delta x \cdot \Delta y. \tag{2}$$

Vztahy (1) a (2) vyjadřují tutéž pravděpodobnost, proto musí platit

$$g(r) = f(x) \cdot f(y). \tag{3}$$



Obrázek 1. Přesnost geodetických měření

Protože levá strana předchozího vztahu nezávisí na úhlu ϑ, pak derivováním dostaneme:

$$\frac{\partial g(r)}{\partial \vartheta} = 0 = f(x)\frac{\partial f(y)}{\partial \vartheta} + f(y)\frac{\partial f(x)}{\partial \vartheta}. \tag{4}$$

Užitím polárních souřadnic pro $r > 0, \vartheta \in \langle 0, 2\pi \rangle$ a jejich transformace na kartézské souřadnice x, y ve tvaru

$$\begin{aligned} x &= r \cdot \cos \vartheta \\ y &= r \cdot \sin \vartheta, \end{aligned} \tag{5}$$

můžeme vztah (3) po derivaci podle úhlu ϑ přepsat nejdříve na

$$0 = f(x)\frac{\partial f(y)}{\partial y}\frac{\partial y}{\partial \vartheta} + f(y)\frac{\partial f(x)}{\partial x}\frac{\partial x}{\partial \vartheta}, \tag{6}$$

a pak na

$$0 = f(x) \cdot f'(y) \cdot x + f(y) \cdot f'(x) \cdot (-y). \tag{7}$$

Pro $x \neq 0, y \neq 0, f(x) \neq 0, f(y) \neq 0$ dostáváme

$$\frac{f'(x)}{x \cdot f(x)} = \frac{f'(y)}{y \cdot f(y)}. \tag{8}$$

Protože jsou podle předpokladu identifikace bodu A veličiny X a Y nezávislé, předchozí vztah znamená, že obě části rovnosti (8) jsou rovny blíže neurčené konstantě K, tj.

$$\frac{f'(x)}{x \cdot f(x)} = K = \frac{f'(y)}{y \cdot f(y)}. \tag{9}$$

$$\begin{aligned} \tfrac{f'(x)}{f(x)} &= K \cdot x, \\ \tfrac{f'(y)}{f(y)} &= K \cdot y. \end{aligned} \tag{10}$$

Integrací předchozích vztahů (10) dostaneme

$$\ln f(x) = \frac{Kx^2}{2} + C, \text{ nebo také } f(x) = A \cdot e^{\frac{1}{2}Kx^2}, \text{ kde } A = e^C. \tag{11}$$

Předpokládáme-li, že větší chyby jsou méně pravděpodobné, musí být $K < 0$. Volíme proto konstantu $K = -2k^2, k > 0$. Takže máme

$$f(x) = Ae^{-k^2x^2}, f(y) = Ae^{-k^2y^2}. \tag{12}$$

Proto

$$g(r) = A^2 \cdot e^{-k^2(x^2+y^2)} = A^2 \cdot e^{-k^2r^2}. \tag{13}$$

Protože $g(r)$ je hustota pravděpodobnosti v polárních souřadnicích, musí platit

$$\int_0^{2\pi} \int_0^{+\infty} g(r) dr d\vartheta = 1. \tag{14}$$

Dosazením za hustotu $g(r)$ dostaneme normovací podmínku:

$$A^2 \cdot 2\pi \int_0^{+\infty} r \cdot e^{-k^2 r^2} dr = A^2 \cdot \pi \left[\frac{e^{-k^2 r^2}}{-k^2} \right]_0^{+\infty} = \frac{\pi A^2}{k^2} = 1 \tag{15}$$

a z toho pro A platí

$$A = \frac{k}{\sqrt{\pi}}. \tag{16}$$

Dostáváme tak normální (Gaussovo) rozdělení s hustotou

$$f(x) = \frac{k}{\sqrt{\pi}} \cdot e^{-k^2 x^2}. \tag{17}$$

Pro Laplaceův integrál platí $\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \mathrm{d}x = \sqrt{2\pi}$. Proto musí být $k = \frac{1}{\sqrt{2}}$ (integrál $\int_{-\infty}^{+\infty} f(x) \mathrm{d}x = 1$). Normální rozdělení náhodné veličiny X tvaru

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

označujeme $N(0;1)$ a nazýváme normovaným normálním rozdělením. Pro náhodnou veličinu se střední hodnotou $EX = \mu$ a rozptylem $DX = \sigma^2$ má hustota normálního (Gaussova) rozdělení tvar

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{18}$$

a označujeme ho $N(\mu; \sigma^2)$.

Jako důležité se ukazuje normální rozdělení v souvislosti ve statistice velmi užívané Lindeberg – Lévyho věty, která říká, že součet vzájemně nezávislých náhodných veličin, které jsou stejně distribuovány (rozděleny) s konečnou střední hodnotou a konečným rozptylem, má pro dosti velký počet měření přibližně normální rozdělení.

Měříme-li náhodnou veličinu X, zjistíme v i-tém měření hodnotu $X_i = x_i$, kterou si můžeme v mnoha případech představit jako součet správné hodnoty X_p a chyby ϵ_i, tedy $X_i = X_p + \epsilon_i$. Předpokládáme-li, že pravděpodobnost určité kladné chyby je stejná jako záporné (s toutéž absolutní hodnotou), musí být střední hodnota chyby

$E\epsilon_i = 0$. Dále můžeme v nevelkém rozsahu měřených hodnot předpokládat konstantnost rozptylu chyby měření (nezávislost na i), tedy $DX_i = D\epsilon_i = \sigma^2$. Jsou-li jednotlivá měření na sobě nezávislá a i navíc se stejným rozdělením s konečnou střední hodnotou a konečným rozptylem, pak podle věty Lindeberg - Lévy-ho má $\sum_{i=1}^{n}(X_i - X_p)\frac{1}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\epsilon_i$ pro $n \to \infty$ normované normální rozdělení $N(0;1)$. Důležité je, že limitní přechod k normálnímu rozdělení je pro velkou třídu rozdělení dosti rychlý (například pro binomická rozdělení s parametry n, p pro běžnou potřebu stačí, že $n \cdot p \cdot (1-p) > 9$).

Při současném měření dvou vzájemně nezávislých veličin X, Y bude-li v prvním rozměru pro meze $x, x + \mathrm{d}x$ pravděpodobnost rovna $f_1(x)\mathrm{d}x$, v druhém pro meze $y, y + dy$ pravděpodobnost rovna $f_2(y)dy$, pak pro pravděpodobnost, že výsledek měření bude ve čtverci s vrcholy v bodech (x, y), $(x, x + \mathrm{d}x)$, $(x + \mathrm{d}x, y + dy)$, $(x, y + dy)$, musí platit $f_1(x) \cdot f_2(y) \cdot \mathrm{d}xdy$. Je-li rozdělení obou veličin normální, první s rozdělením $N(\mu_x; \sigma_x^2)$, druhé s rozdělením $N(\mu_y; \sigma_y^2)$, pak

$$f_1(x) \cdot f_2(y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{t^2}{2}}.$$

Tam, kde je hustota konstantní, musí být $\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = t^2$, kde $t > 0$ je konstanta. Množina všech bodů, které splňují předchozí rovnici, je elipsa o poloosách $t\sigma_x, t\sigma_y$. Snadno určíme pravděpodobnost, že chyba bude ležet v intervalu eliptického prstence v mezích poloos $t\sigma_x, (t + \mathrm{d}t)\sigma_x$ a $t\sigma_y, (t + \mathrm{d}t)\sigma_y$. Tato pravděpodobnost bude součinem plochy diferenciálního prstence, příslušného změně proměnné t o $\mathrm{d}t$ a pravděpodobnosti

$$\frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{t^2}{2}}.$$

Protože plocha elipsy je $F = \pi ab = \pi\sigma_x\sigma_y t^2$, je $dF = 2\pi\sigma_x\sigma_y t \cdot \mathrm{d}t$. Pak ale bude pravděpodobnost, že chyba bude ležet v ploše diferenciálního prstence rovna

$$P_t^{t+\mathrm{d}t} = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{t^2}{2}} \cdot 2\pi\sigma_x\sigma_y t \cdot \mathrm{d}t = t \cdot e^{-\frac{t^2}{2}} \cdot \mathrm{d}t.$$

Pro dvojrozměrnou chybu je pro obvod elipsy hustota $f(t)$ podle předchozího výsledku daná vztahem

$$f(t) = t \cdot e^{-\frac{t^2}{2}}. \tag{19}$$

Vztah (19) se liší od vztahu (17), resp. (18), pro chybu jednorozměrnou. U dvojrozměrných chyb neexistují chyby záporné, graf hustoty je asymetrický, jak je možné se snadno přesvědčit.

2 Metoda nejmenších čtverců (MNČ)

Je také Gaussovou zásluhou, že se od roku 1795 začala používat MNČ (Gauss ji použil pro eliminaci chyb geodetického vyměřování).

Mějme n-krát nezávisle opakované měření veličiny X jehož výsledkem jsou hodnoty $x_1, x_2, \ldots, x_n$. Přitom správná hodnota není známa, označme ji x. Odhadem jediné hodnoty, která by mohla zastupovat správnou hodnotu, Gauss stanovil takové číslo x_0, pro nějž je součet čtverců odchylek $Q = Q(x) = \sum_{i=1}^{n} (x_i - x)^2$ minimální. Minimum lze zde najít derivováním, když derivaci $dQ/\mathrm{d}x$ položíme rovnu nule:

$$\frac{dQ}{\mathrm{d}x} = -2 \sum_{i=1}^{n} (x_i - x) = 0.$$

Z předchozího výrazu dostaneme pro x_0 výpočet $x_0 = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$. Že se jedná skutečně o minimum, plyne z toho, že $\frac{d^2Q}{\mathrm{d}x^2} = -2 \sum_{i=1}^{n} (-1) = 2n > 0$.

Uvažované měření, při kterém jsme získali n hodnot, lze tedy zastupovat za určitých podmínek hodnotou jedinou, která je ve smyslu metody nejmenších čtverců aritmetickým průměrem $\bar{x}$. Ten zastupuje střední hodnotu veličiny X. Nyní předpokládejme, že měříme veličinu X, která má normální rozdělení $N(\mu; \sigma^2)$ s neznámými parametry μ, σ^2. Budeme nyní hledat takové odhady parametrů μ a σ^2, které maximalizují součin

$$L = L\left(\mu, \sigma^2\right) = \prod_{i=1}^{n} f(x_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot e^{-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}}.$$

(Hodnota L odpovídá hustotě n-rozměrného měření jehož výsledek je $(x_1, x_2, \ldots, x_n)$ za daného předpokladu.)

Maximu L odpovídá maximum $\ln L$ a opět použijeme derivování. Protože v tomto případě je L funkcí dvou proměnných, použijeme parciálních derivací:

$$\ln L = \sum_{i=1}^{n} \ln f(x_i) = -\frac{n}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2,$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{i=1}^{n} (x_i - \mu) = 0,$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 = 0.$$

Z první derivace dostaneme známý odhad $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, z druhé odhad pro parametr σ^2 ve tvaru $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = s_n^2$.

Odhad metodou maximální věrohodnosti L (funkce L se nazývá věrohodnostní funkce) pro normálně rozdělenou náhodnou veličinu vede k odhadu metodou nejmenších čtverců; jejím jedním výsledkem je výběrový rozptyl s_n^2. (Odhad momentovou metodou vede k odhadu rozptylu ve tvaru $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Jeho výhodou je, že je nestranný, pro střední hodnotu je $ES_{n-1}^2 = \sigma^2$ na rozdíl od odhadu s_n^2, pro který platí $ES_n^2 = E\frac{n-1}{n-1} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}\sigma^2$, který je pouze asymptoticky nestranný.)

3 Aplikace v psychologii

Pomocí normálního rozdělení nemodelujeme jen rozdělení chyb technických měření. Ukazuje se také (experimentálně to bylo potvrzeno), že toto rozdělení popisuje i biologickou variabilitu. Od 19. století se předpokládalo, že většina přímo měřených veličin jakéhokoliv původu má toto rozdělení (odtud také jeho název „normální“). Sem například patřilo rozdělení lidí podle výšky, hmotnosti či některých částí lebky či jiných kostí. Histogramy mnoha takových měření měly tvar, podobající se typické zvonovité křivce. Na tomto základě se předpokládalo, že i řada psychologických měření má v populaci normální rozdělení odhadovaných vlastností, například lidských schopností. Podle toho se v psychologii vytvářely dotazníky, které měly charakterizovat testovaného v reakčních schopnostech, odolnosti vůči stresu, úroveň jeho únavy po určité zátěži, numerické zběhlosti, prostorové představivosti, atd. Výsledky všech takových šetření měly být také normálně rozděleny (tj. v tehdejším pojetí měly mít histogramy všech takových měření zvonovitý tvar). Ti, kteří v testech vybraných schopností uspěli s výsledkem nadprůměrným, byli označováni jako „nadprůměrně inteligentní“, příslušné dotazníky pak jako „testy inteligence“. První test inteligence vytvořil na začátku 20. století A. Binet (1857 – 1911). Testy inteligence se lišily svým obsahem (dotazníkovými položkami), proto každý měřil něco jiného, věřilo se však tomu, že změna dotazníkových položek nemá zásadní vliv na výsledek dotazníku. Původně inteligenční kvocient IQ znamenal podíl mentálního věku a chronologického věku vynásobený 100. Mentální věk se určoval pomocí dotazníku, kdy respondent jeho vyřešením se zařadil mezi referenční populaci se stejným výkonem určitého chronologického věku. Takto na jednorozměrné a spojité stupnici byla populace pomocí určitého dotazníku klasifikována (například populace branců). Teprve později stupnice pro měření inteligence pomocí dotazníku (testu) byla více propojena se statistikou (a příslušným dotazníkem). To je spojeno se jmény Ch. Spearmena, R. Cattella, L. L. Thurstona a P. Guilforda. Ukážeme si princip metody.

Mějme dotazník s celkem n položkami (otázkami, úkoly, …), které jsou homogenní v tom smyslu, že na každou z nich bude respondent odpovídat správně s pravděpodobností p. O pravděpodobnosti p se předpokládá, že je určitou mírou inteligence respondenta. Odpoví-li respondent na k dotazníkových položek správně a lze-li ještě předpokládat, že položky jsou nezávislé, můžeme si výsledný skór představit jako součet n nezávislých náhodných veličin X_i s alternativním rozdělením ($X_i = 1$ když respondent odpověděl na i-tou položku správně s pravděpodobností p, $X_i = 0$ když odpověděl špatně s pravděpodobností $1 - p$), kde

$$\begin{aligned} EX_i &= 1 \cdot p + 0 \cdot (1-p) = p, \\ DX_i &= EX_i^2 - (EX_i)^2 = p - p^2 = p(1-p). \end{aligned} \tag{20}$$

Veličina $X = \sum_{i=1}^{n} X_i$, představující celkový počet správně zodpovězených položek, má pak binomické rozdělení a platí

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{21}$$

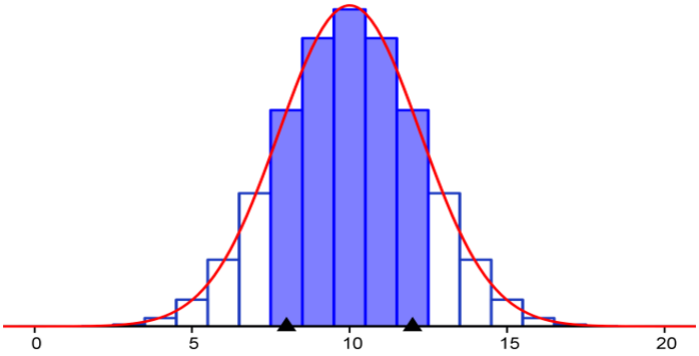
Můžeme-li ještě předpokládat, že $0,1 < p < 0,9$ a počet dotazníkových položek je velký ($n > 30$), lze náhodnou veličinu X aproximovat normálním rozdělením prostřednictvím Moivreovy – Laplaceovy věty

$$\begin{aligned} \lim_{n\to\infty} P(X \le k|p) &= \lim_{n\to\infty} \left(\tfrac{X-np}{\sqrt{np(1-p)}} \le \tfrac{k-np}{\sqrt{np(1-p)}} \right) \\ \sim \Phi\left(\tfrac{k-np}{\sqrt{np(1-p)}} \right) &= \int_{-\infty}^{\frac{k-np}{\sqrt{np(1-p)}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \cdot \mathrm{d}t = P(k|p). \end{aligned} \tag{22}$$

(Přitom jsme označili $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \cdot \mathrm{d}t$.)

Diskrétní binomické rozdělení se zde aproximuje spojitým normálním rozdělením tak, že když například $P(k|p) = 0,5$, pak IQ $= 100$, když $P(k|p) = 0,25$, pak IQ $= 50$, obecně IQ $= 2 \cdot P(k|p) \cdot 100$. Stupnice pro IQ je tedy nerovnoměrná (viz Obr. 2) vzhledem ke k.

Parametr p můžeme z experimentu odhadnout hodnotou $\widehat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n}$ (za předpokladu statistické *nezávislosti* a *stejné obtížnosti* p všech dotazníkových položek). Pro účely aproximace binomického rozdělení rozdělením normálním, které jsme nastínily, je lepší odhadovat přímo střední hodnotu np, resp. směrodatnou odchylku binomického rozdělení $\sqrt{np(1-p)}$ do vztahu (22) pro funkci $\Phi(x)$ (jejíž hodnoty pro každé normované x jsou tabelovány) z empirické střední hodnoty, resp. empirické



Obrázek 2. Vztah binomického rozdělení s parametry $n = 20$ a $p = 0,5$ a rozdělení normálního se střední hodnotou 10 a rozptylem 5. Tmavěji označená plocha je mírou pravděpodobnosti $P(8 \leq X \leq 12)$ binomické náhodné veličiny, plocha pod grafem normálního rozdělení mezi vyznačenými trojúhelníky na ose x je mírou jejího odhadu.

směrodatné odchylky měřené veličiny X. Zkušenost ukázala, že ne každý dotazník měl v uvažované populaci normálně rozdělené výsledky (které byly počtem pozitivně zodpovězených dotazníkových položek). Usuzovalo se, že příčinou jsou nedostatky v obsahu a konstrukci dotazníku. Snaha byla co nejlépe vyhovět podmínkám Moivreovy – Laplaceovy věty. Pro praktické používání dotazníku jako měrného prostředku s aspoň přibližně normálně rozdělenými výsledky je obtížné, a proti účelu dotazování, konstruovat položky tak, aby jich bylo dostatečné množství a přitom všechny měly stejnou obtížnost. Statistickou nezávislost položek je také obtížné splnit (je rovněž komplikované uvedené předpoklady statisticky ověřovat, i slabá statistická závislost položek má na celkový výsledek dotazníku velký vliv). Přitom psychologové trvali v některých šetřeních (např. v testech inteligence) na tom, aby výsledky dotazníku byly v populaci normálně rozloženy (podobně jako je to v technických měřeních) a mohlo se tak používat statistické intepretace výsledků podobně jako v technice. Tak byl zaveden *psychometrický postulát* o potřebě konstruovat některé dotazníky (speciálně testy inteligence) tak, aby výsledky v populaci, pro kterou byl dotazník určen, byly normálně rozdělené. Každý takový dotazník (zvláště měl-li měřit inteligenci) se proto musel nejprve na výběrovém reprezentativním souboru respondentů kalibrovat. To znamená, nejen to, že z pokusů na vybraném reprezentativním souboru respondentů se pro určitý dotazník (který má zde funkci „metru“) odhadnou parametry pro normální rozdělení pomocí výběrové střední hodnoty a výběrového rozptylu z empirických počtů správně zodpovězených položek jednotlivých respondentů (tj. empiricky odhadované veličiny X), ale hlavně to, že se musí vhodně měnit nebo upravovat počet dotazníkových položek. Dotazník měl mít dostatečný počet položek

a doba jeho řešení respondenty nemohla být příliš krátká (vyplňování takového dotazníku někdy mělo trvat respondentu i několik hodin). A priori nelze předpokládat (nejsou pro to ani formální důvody statistické), že psychický výkon, měřený dotazníkem (např. testem inteligence) má v populaci normální rozdělení. Normalita výsledků byla tedy kalibrací příslušnému dotazníku (testu inteligence) vnucena záměrně volbou určitých položek i formou. Pro psychometrická měření tedy neplatilo to, co pro velkou třídu měření fyzikálních. Psychometrický dotazník je originálním měřícím prostředkem a měření pomocí něho závisí na tom, co a jak bylo do něho autorem dotazníku vloženo. Neuniverzálnost měřícího prostředku (dotazníku) znamená, že měření podle různých dotazníků *nemusí být* s hlediska významu (interpretace) vzájemně porovnatelná. Přímá měření fyzikální se realizují univerzálním prostředkem (např. metrem) a velmi často jejich výsledkem je normálně rozdělená náhodná veličina s jednoznačným významem (je to vždy když měření je výsledkem působení součtu nezávislých, stejně rozdělených náhodných chyb s konečnou střední hodnotou a konečným rozptylem).

Výsledek X (chápaný jako počet správně označených dotazníkových položek) a interpretace dotazníkového šetření jsou závislé na obsahu dotazníku (ale i konstrukci a způsobu zadávání, proto se příprava takového dotazníku postupně stávala komplikovaným úkolem) místo pouze na osobnosti. Krom toho zkoumaný jedinec je komplikovaný i ve vztahu k aktu dotazování a měřené psychické vlastnosti jsou ve své podstatě složité a měření by mělo být spíše vícerozměrné povahy (má několik psychologicky vymezených dimenzí). Zde pomohla i faktorová nebo shluková analýza a i jiné statistické techniky. V dnešní době je také zpochybňováno měření inteligence jako aproximace jistých spojitých latentních veličin (některé vlastnosti člověka nemusí být měřitelné na spojité stupnici a mění se v čase, a to skokem, příkladem je situace, označovaná jako „vhled“, problémem je, že tato měření jsou v podstatě nepřímá, určitou proměnnou měříme prostřednictvím jiných, jí ovlivňovaných proměnných).

Ještě obtížnější je zjišťování *znalosti* (něčeho) pomocí dotazníku (vědomostní dotazník velmi často je jen jednorázovou akcí a jeho dlouhodobější příprava se podceňuje). Podceňuje se dosud i představa znalosti jako vícerozměrné veličiny (psychologové již na to přišli).

4 Jedna z fyzikálních aplikací

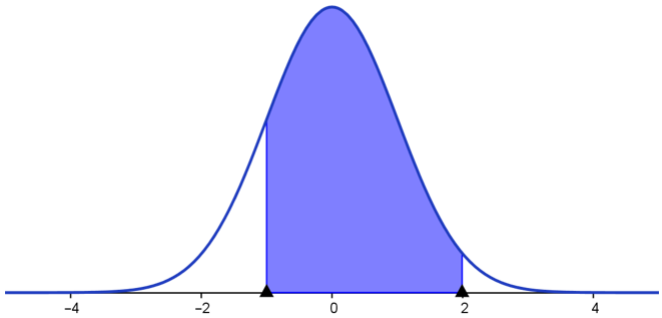
Uvažujme jiný příklad. Máme-li jednorozměrnou verzi fyzikům známé difuzní rovnice popisující difuzi fyzikální veličiny $u(x,t)$ v množině reálných čísel x (obyčejně

značí polohu) a t (obyčejně značí čas) ve tvaru

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} \tag{23}$$

s počáteční podmínkou $u(x,0) = \delta(x)$, kde δ je Diracova funkce (ta má nulové hodnoty všude, kromě $x = 0$ a její integrál přes všechna x je jedna), pak její řešení je pro každé pevné t funkce typu (17):

$$u(x,t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{x^2}{4Dt}}. \tag{24}$$



Obrázek 3. Graf normovaného normálního rozdělení (se střední hodnotou 0 *a rozptylem* 1*); vybarvená část představuje pravděpodobnost, že gaussovská (normálně rozdělená) náhodná veličina bude mít hodnotu mezi* $-0,5$ *a* 2 *(plocha pod grafem omezená zdola osou* x *má velikost* 1*).*

Kdy nabývá entropie své maximální hodnoty?

V práci [4] jsme definovali (krom jiného) i Shannonovu entropii pro charakterizaci neurčitosti rozhodování v pravděpodobnostním prostoru $[\Omega; A_\Omega; P]$ s konečným základním prostorem Ω.

Mějme náhodnou veličinu X s diskrétním rozložením na množině $\Omega = \{x_i;\ i = 1, 2, \ldots, n\}$ s pravděpodobnostní funkcí $P(X = x_i) = p_i, 0 \leq p_i \leq 1;\ i = 1, 2, \ldots, n; \sum_i p_i = 1$. Entropií náhodné veličiny X nazýváme hodnotu výrazu

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \ \text{ [bit] } \ (\text{klademe } \ 0 \cdot \log_2 0 = 0) \tag{25}$$

Entropie náhodné veličiny X je míra neurčitosti apriorní identifikace její polohy na číselné ose při libovolné realizaci.

Pro $H(X)$ pak platí $0 \leq H(X) \leq \log_2 n$, největší hodnota $H(X)$ přísluší rozdělení rovnoměrnému s pravděpodobnostní funkcí $p_i = \frac{1}{n}$, nejmenší hodnota entropie přísluší náhodné veličině s rozdělením u něhož pro jednu hodnotu, např. x_j, je $p_j = 1$ (pro zbývající hodnoty x_i je pak $p_i = 0$). Důkaz tohoto tvrzení vyplývá z následujících úvah:

$$H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -\log_2 \frac{1}{n} = \log_2 n. \tag{26}$$

Typ rozdělení náhodné veličiny X při maximální hodnotě entropie $H(X)$ najdeme metodou Lagrangeových multiplikátorů. Lagrangeova funkce L má tvar

$$L(p_1, p_2, \ldots, p_n, \lambda) = -\sum_{i=1}^{n} p_i \log_2 p_i + \lambda\Big(\sum_{i=1}^{n} p_i - 1\Big), \tag{27}$$

a když její derivace podle všech proměnných položíme rovny nule

$$\frac{\partial L}{\partial p_i} = -\log_2 p_i - \frac{1}{\ln 2} + \lambda = 0 \quad \frac{\partial L}{\partial \lambda} = \sum_{i=1}^{n} p_i - 1 = 0, \tag{28}$$

dostaneme, že extrémní hodnoty nabývá entropie pro $p_i = \frac{1}{n}$. V našem případě je touto extrémní hodnotou maximum $H(X) = \log_2 n$.

Pro případ spojitého rozdělení náhodné veličiny X (na pravděpodobnostním prostoru s $\Omega = (-\infty; +\infty)$ a systémem borelovských podmnožin A_Ω) s hustotou $f(x) \geq 0$, definovanou pro $x \in \langle a; b\rangle$, $\int_a^b f(x)\mathrm{d}x = 1$ kde a, b jsou reálná čísla, určujeme entropii $H(X)$ vztahem

$$H(X) = -\int_a^b f(x) \cdot \log_2 f(x)\mathrm{d}x. \tag{29}$$

Pro takto definovanou entropii pak platí nerovnost $0 \leq H(X) \leq \log_2 (b - a)$ a maximální hodnoty je dosaženo pro rovnoměrné rozložení na intervalu $\langle a; b\rangle$. To můžeme ukázat následujícím postupem.

Nejdříve si všimneme, že pro speciální volbu $f(x) = \frac{1}{b-a}$ pro $x \in \langle a; b\rangle$ dostaneme $H(X)$ ve tvaru

$$-\int_a^b \frac{1}{b-a} \log_2 \frac{1}{b-a}\mathrm{d}x = \log_2 (b - a). \tag{30}$$

Nyní řešíme extremální úlohu: hledáme hustotu $f(x)$ definovanou na $\langle a; b\rangle$ takovou, aby pro ni integrál $-\int_a^b f(x) \cdot \log_2 f(x)\mathrm{d}x$ nabýval maximální hodnoty za omezujících podmínek $f(x) \geq 0$, $\int_a^b f(x)\mathrm{d}x = 1$. Sestrojíme opět Lagrangeovu funkci

$L(f, \lambda)$ ve tvaru

$$L(f, \lambda) = -\frac{f \cdot \ln f}{\ln 2} + \lambda \cdot f, \tag{31}$$

odtud dostáváme podmínku

$$\frac{\partial L}{\partial f} = -\frac{\ln f}{\ln 2} - \frac{1}{\ln 2} + \lambda = 0, \tag{32}$$

a z ní

$$\ln f = \lambda \cdot \ln 2 - 1, \tag{33}$$

tedy

$$f = e^{\lambda \cdot \ln 2 - 1} > 0. \tag{34}$$

Dále pak je

$$\frac{\partial^2 L}{\partial f^2} = -\frac{1}{f \cdot \ln 2} = -\frac{1}{\ln 2 \cdot e^{\lambda \cdot \ln 2 - 1}} < 0, \tag{35}$$

proto maxima se dosahuje pro funkci nezávislou na x ve tvaru $f(x) = e^{\lambda \cdot \ln 2 - 1}$. Z toho, že f má být hustota pravděpodobnosti však plyne, že $e^{\lambda \cdot \ln 2 - 1} = \frac{1}{b-a}$, a tedy funkcionál $H(X)$ nabývá svého maxima pro funkci $f(x) = \frac{1}{b-a}$. Dokázali jsme tedy, že hodnota $\log_2 (b - a)$ udává maximální entropii pro hustotu $f(x) = \frac{1}{b-a}$.

Nyní ukážeme, že maximální hodnoty entropie

$$H(X) = -\int_{-\infty}^{-\infty} f(x) \cdot \log_2 f(x) \mathrm{d}x$$

pro spojitou náhodnou veličinu X s hustotou $f(x) \geq 0$ na intervalu $(-\infty; +\infty)$ se střední hodnotou $EX = \mu$ a rozptylem $DX = \sigma^2$ dosahuje normálně rozložená náhodná veličina právě s uvedenými parametry.

Opět sestrojíme Lagrangeovu funkci, tentokráte pro tři druhy vazeb:

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1, \quad \int_{-\infty}^{+\infty} x \cdot f(x)\mathrm{d}x = \mu, \quad \int_{-\infty}^{+\infty} x^2 \cdot f(x)\mathrm{d}x = \sigma^2 + \mu^2. \tag{36}$$

Ta je ve tvaru

$$L(f, \lambda_1, \lambda_2, \lambda_3) = -\frac{f \cdot \ln f}{\ln 2} + \lambda_1 \cdot f + \lambda_2 \cdot x \cdot f + \lambda_3 \cdot x^2 \cdot f. \tag{37}$$

Derivováním L podle f a položením této derivace nule dostaneme

$$\begin{aligned} &\tfrac{\partial L}{\partial f} = -\tfrac{\ln f}{\ln 2} - \tfrac{1}{\ln 2} + \lambda_1 + \lambda_2 \cdot x + \lambda_3 \cdot x^2 = 0, \\ &f(x) = e^{1 + \left(\lambda_1 + \lambda_2 \cdot x + \lambda_3 \cdot x^2\right) \ln 2} = K \cdot e^{ax^2 + bx + c}. \end{aligned} \tag{38}$$

Dosazením předchozího vztahu do tří vazbových podmínek dostaneme výsledný tvar, odpovídající hustotě normálního rozdělení:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{39}$$

Závěr

Snažili jsme se ukázat, že normální rozdělení je nejdůležitějším typem rozdělení náhodné veličiny (je limitním pro řadu jiných rozdělení, zmínili jsme se o binomickém rozdělení), uplatňuje se nejen v teorii pravděpodobnosti a statistice (zvláště v teorii měření), ale také je podstatou i některých teorií fyzikálních a psychologických. Nezmiňovali jsme se o některých teoriích biologických, kde podstatou je náhoda (například množení, růst, …), která má svůj vnější projev v normálním rozdělení některých biologických veličin (je původcem tzv. biologické variability).

Literatura – References

[1] Hamming, R., W.: *Numerical Methods for Scientists and Engineers*, Mc Graw–Hill, New York 1962.

[2] Mareš, M. : *Příběhy matematiky*, Pistorius, 2. vydání, Příbram 2013.

[3] Britton, Nicholas, F.: *Essential Mathematical Biology*, Springer – Verlag London Limited 2003, University of Bath, Bath BA2 7AY, UK, 3rd printing 2005.

[4] Půlpán, Z.: Neurčitosti pro fuzzy množiny a intuicionistické fuzzy množiny *Obzory matematiky, fyziky a informatiky*, č. 4/2021(50), str.9 – 19.

[5] Guilford, J. P.: *Psychometrics methods*,McGraw-Hill, New York, 1936.

[6] Půlpán, Z.: *Odhad informace z dat vágní povahy*, ed. Gerstner, Academia, Praha 2012.

[7] Rao, C. R.: *Lineární metody statistické indukce a jejich aplikace*, Academia, Praha 1978.

Adresy autorů:

Dopravní fakulta Jana Pernera, Univerzita Pardubice, Studentská 95, 532 10 Pardubice

e-mail: zdenek.pulpan@upce.cz, ondrej.slavicek@upce.cz