

Normální rozdělení

Normální rozdělení (*Gaussovo rozdělení, Gaussova distribuce, Laplaceovo-Gaussovo rozdělení*) patří mezi nejdůležitější rozdělení pravděpodobnosti spojité náhodné veličiny. Náhodné děje vyskytující se v přírodě či společnosti lze dobře modelovat právě normálním rozdělením. Jako příklad takového náhodného děje, který se řídí normálním rozdělením, může sloužit např. IQ^[pozn. 1] nebo výšky v populaci, vitální kapacity plic nebo třeba chyby měření^[pozn. 2]. Řada dalších rozdělení se při dostatečně velkém vzorku k normálnímu rozdělení číselně blíží, takže lze využít tabelovaných hodnot distribuce např. k výpočtu problémů modelovatelných binomickým rozdělením. Jiná rozdělení lze na normální poměrně snadno transformovat, např. tzv. lognormální rozdělení popisující stáří bílých krvinek v periferní krvi. K normálnímu rozdělení patří často zmiňované náhodné chyby, např. chyby měření, způsobené velkým počtem neznámých a vzájemně nezávislých příčin. Proto bývá normální distribuce také označována jako *zákon chyb*. Tímto zákonem se také řídí rozdělení některých fyzikálních a technických veličin.^[1] Normální rozdělení plně charakterizují dvě konstanty: **střední hodnota** μ a **rozptyl** σ^2 . Pokud chceme napsat, že náhodná veličina X (tedy např. výsledek pokusu) má náhodné rozdělení se střední hodnotou μ a rozptylem σ^2 , obvykle použijeme označení:

$$X \sim N(\mu, \sigma^2)$$

Gaussova křivka

Gaussova křivka (hustota pravděpodobnosti) je vlastně funkcí o dvou parametrech: střední hodnoty μ a rozptylu σ^2 . **Gaussova křivka** je symetrická, střední hodnota μ leží právě pod jejím vrcholem (viz např.^[1]). Tvar křivky s extrémem v místě střední hodnoty vlastně říká to, že při opakování náhodného pokusu řídícího se normálním rozdělením budou nejčastěji vycházet hodnoty v okolí střední hodnoty. Symetrie křivky pak říká to, že výsledky vychýlené nad i pod střední hodnotu budou vycházet zhruba stejně často. Parametr σ^2 určuje, jak těsně se křivka přimyká střední hodnotě; čím nižší je tento parametr, tím je graf „ostřejší“. V praxi se často používá tzv. **pravidlo tří sigma**, někdy i dvou nebo jednoho sigma. Platí totiž, že výsledek náhodného pokusu s rozdělením $N(\mu, \sigma^2)$ leží v intervalu:

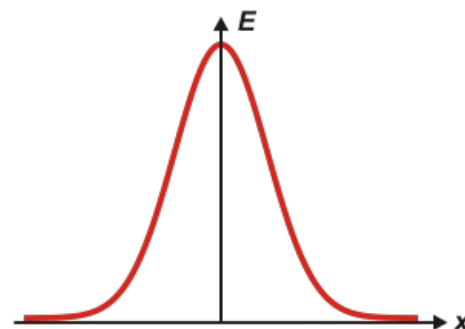
- $(\mu - \sigma, \mu + \sigma)$ s pravděpodobností 68,27 %,
- $(\mu - 2\sigma, \mu + 2\sigma)$ s pravděpodobností 95,45 %,
- $(\mu - 3\sigma, \mu + 3\sigma)$ s pravděpodobností 99,73 %.

Význam Gaussovy křivky

Význam křivky popisující hustotu pravděpodobnosti si lze představit konstrukcí z histogramu, tedy sloupcového diagramu tvořeného obdélníky, které mají pevně zvolenou šířku základny na ose x odpovídající zvolenému intervalu (např. při měření výšky můžeme vytvořit kategorie po pěti centimetrech). Výška obdélníku (tedy rozměr v ose y) pak odpovídá četnosti dané třídy. Pokud budeme mít výsledky z pokusu s normálním rozdělením a budeme šířku jedné kategorie limitně snižovat k nule, dostaneme právě hustotu pravděpodobnosti normálního rozdělení. Předmětem zájmu statistické inference je zjišťování číselných charakteristik rozsáhlých souborů z malých vzorků. Pokud předpokládáme, že má sledovaný znak normální rozdělení, zajímají nás odhady střední hodnoty a rozptylu (nebo směrodatné odchylky, což je vlastně odmocnina z rozptylu). Odhadem střední hodnoty z konečného počtu pokusů je aritmetický průměr, modus i medián, číselně bývají shodné, nejsnáze se počítá aritmetický průměr. Poměrně často je známý jen jeden výběr z populace a jeho aritmetický průměr \bar{x} a nás zajímá, jak dobrý je to odhad střední hodnoty. Jinými slovy nás zajímá, v jakém pásmu kolem zjištěného aritmetického průměru se s předem stanovenou pravděpodobností nachází skutečná střední hodnota. Tento interval se nazývá **interval spolehlivosti**, podle anglického *confidence interval* se někdy hovoří i o konfidenčním intervalu. Pokud je známá směrodatná odchylka σ dané populace, má interval spolehlivosti tvar:

$$\left\langle \bar{x} - y_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + y_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right\rangle$$

Koeficient spolehlivosti $(1 - \alpha)$ je **pravděpodobnost**, že se μ vyskytuje v intervalu spolehlivosti, používá se **95** nebo **99** %; *přesnost* je poloviční délka/šířka intervalu. $y_{1-\frac{\alpha}{2}}$ je kvantil normovaného normálního rozdělení s hodnotami $\mu = 0$ a $\sigma = 1$. Výběrový průměr výběru o rozsahu n pro náhodnou veličinu s normálním rozložením (μ ,



Typický tvar křivky, která popisuje hustotu pravděpodobnosti normálního rozdělení



Normální rozdělení (Gaussova křivka).

Graf znázorňuje hustotu normálního rozdělení se střední hodnotou rovnou μ a směrodatnou odchylkou rovnou σ . Hodnota funkce říká, v jakých oblastech je výsledek náhodného pokusu více pravděpodobný a v jakých méně. Výsledky poblíž střední hodnoty μ jsou pravděpodobnější než odlehlé.

σ^2) je také náhodná veličina s normálním rozložením. Čím je rozsah výběru n větší, tím vyšší a užší je křivka hustoty rozložení pravděpodobnosti. Vysoký štíhlý graf odpovídá užšímu intervalu spolehlivosti. Pokud chceme zmenšit interval spolehlivosti, tedy získat větší jistotu v odhadu skutečné střední hodnoty, je třeba zvyšovat počet měření. Lze postupovat i opačně, totiž že si k předem stanovené šíři intervalu spolehlivosti a hladině významnosti vypočítáme, kolik měření musíme provést. Pokud hodnotu směrodatné odchylky σ neznáme, odhadujeme ji na základě výběru o rozsahu n pomocí výběrové směrodatné odchylky s a v odhadu intervalu spolehlivosti budou vystupovat kvantily Studentova rozdělení. Pokud je však počet měření větší než 30, je rozdíl mezi Studentovým a normálním rozdělením prakticky zanedbatelný.^[2]

Testy normality

 Podrobnější informace naleznete na stránce *Testy normality*.

Řadu statistických metod lze aplikovat pouze na výběrové soubory pocházející z normálního rozdělení. K určení, zda lze rozdělení dat považovat za normální, slouží **testy normality** (např. Shapirův-Wilkův test). Většina statistického softwaru implementuje nějakou formu testů normality. Normalitu rozdělení dat lze hodnotit též graficky, například pomocí histogramu nebo krabicového grafu. Přesnější (a použitelné i pro testování jiných rozdělení) je použití **Q-Q grafu** nebo **P-P grafu**. Velmi hrubý odhad míry normality dat dostaneme prostým porovnáním aritmetického průměru a mediánu; podle některých autorů^[zdroj?] by se neměly lišit o více než 10 %.

Poznámky pod čarou

1. I když existují názory toto zpochybňující, tak dominuje pohled, že měření IQ je náhodný pokus řídicí se normálním rozdělením. To mimo jiné znamená, že relativní četnosti jednotlivých úzkých hodnot IQ by měly dobře vyplnit plochu pod křivkou hustoty pravděpodobnosti.
2. Obvyklým a vesměs splněným předpokladem je to, že chyba měření má normální rozdělení se střední hodnotou nula. To znamená, že ke správné hodnotě se přičte nějaké malé kladné nebo záporné číslo, přičemž průměr těchto čísel s rostoucím počtem měření bude směřovat k nule. To je i podkladem zvyšování přesnosti měření tím, že se měření opakuje.

Odkazy

Související články

- Testy normality

Reference

1. Normální rozdělení (česká Wikipedie)
2. Studentovo rozdělení (česká Wikipedie)

Použitá literatura

- GEIZEROVÁ, Helena, et al. *Epidemiologie : vybrané kapitoly pro seminární a praktická cvičení*. 1. vydání. Praha : Karolinum, 1995. 83 s. ISBN 80-7184-179-X.
- ZVÁROVÁ, Jana. *Základy statistiky pro biomedicínské obory*. 2. vydání. Praha : Karolinum, 2011. 219 s. Biomedicínská statistika; sv. I. ISBN 978-80-246-1931-6.