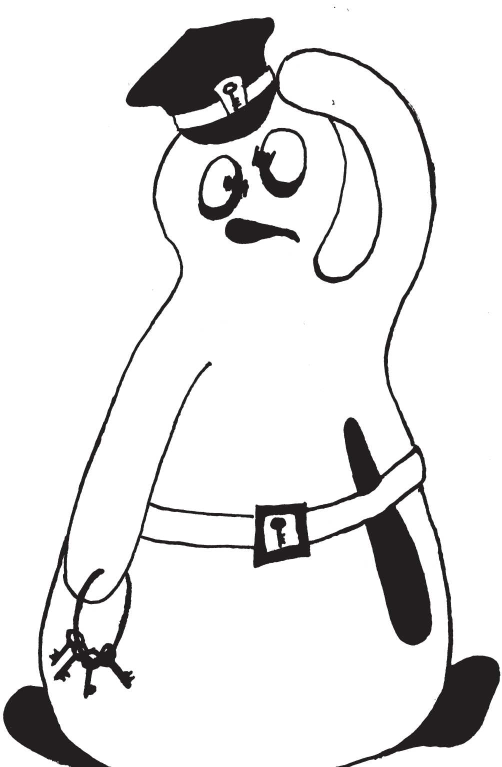


"Zajisté," odvětí strážce." (Str. 110)

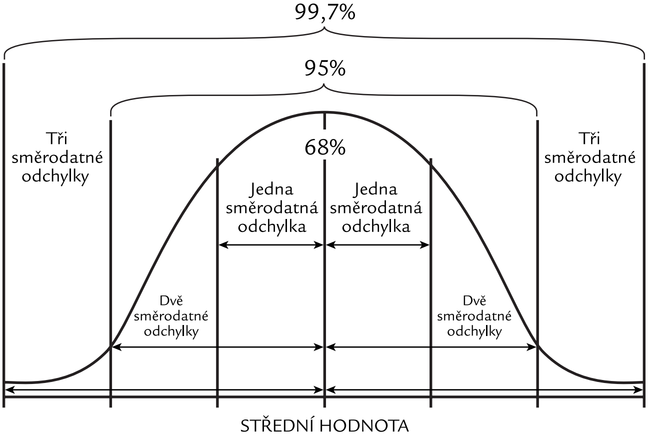
KAPITOLA 17

Normální rozdělení

Nejdůležitější pravděpodobnostní rozdělení se nazývá normální či Gaussovo. Má zajímavou historii. To druhé jméno dostalo na počest slavného německého matematika Carla Friedricha Gausse (1777–1855), který učinil řadu nesmírně důležitých objevů v čisté matematice i v dalších přírodních vědách. Gauss skutečně používal normální rozdělení a byl považován za jeho objevitele. Ale v roce 1908 anglický statistik Karl Pearson nalezl historické spisy dokazující, že normální rozdělení ve skutečnosti objevil již o století dříve anglický matematik Abraham de Moivre (1667–1754).

De Moivre byl francouzský hugenot, který před perzekucí ze strany katolíků uprchl do Anglie. Zde se seznámil s Newtonem i Halleyem (jehož jméno nese známá kometa), a dokonce se stal i členem Královské společnosti. Jelikož přesto nebyl nijak zámožný, příležitostně si přivydělával vyučováním matematiky.

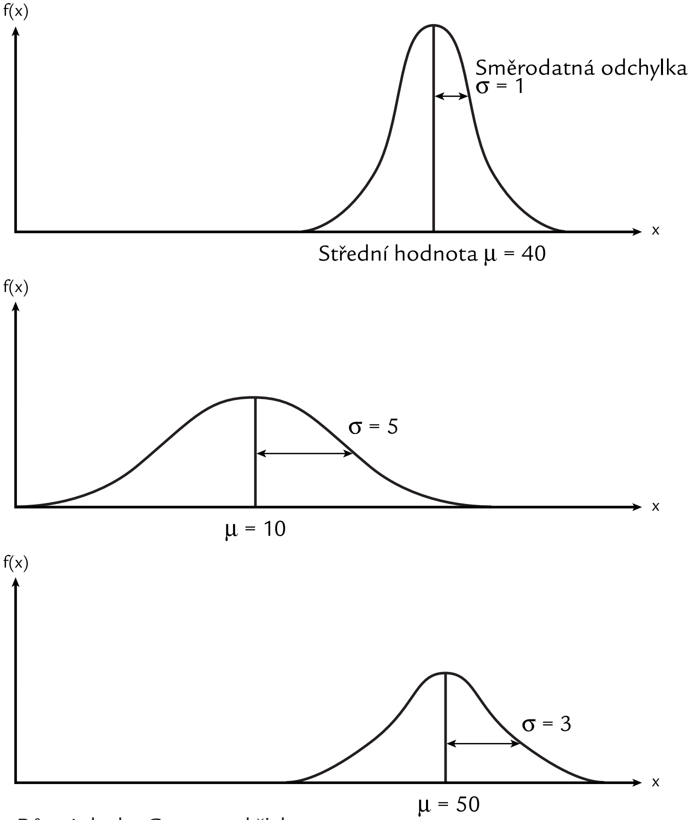
De Moivre objevil „zákon chyb“. Zjistil, že když se nashromáždí mnoho nezávislých náhodných faktorů, pak vytvoří křivku zvonového tvaru, kde méně běžné hodnoty najdeme na obou koncích křivky, zatímco běžnější hodnoty se soustřeďují směrem ke středu. Normální rozdělení, reprezentované Gaussovou (zvonovou) křivkou, vidíme na následujícím obrázku.



Gaussova křivka

Tato křivka má *střední hodnotu* (průměr, nebo centrum hmoty, pokud oblasti vytvořené křivkou přisoudíme určitou „váhu") a *směrodatnou odchylku*, jež je měřítkem šířky křivky (směrodatná odchylka je jednotka míry, kterou uplatňujeme ve vodorovném směru přes křivku). Jelikož jak střední hodnota, tak i směrodatná odchylka může být jakékoli číslo, existuje nekonečně mnoho možných Gaussových křivek. Některé jsou znázorněny dále.

Zdá se, že se Gaussova křivka „objevuje zčistajasna" pokaždé, když se nashromáždí mnoho náhodných faktorů, a proto jí lidé již od jejího objevení přisuzovali téměř mystické vlastnosti. Gaussova křivka svým způsobem vyjadřuje „božský zákon přírody", kterým se řídí spousta jevů v reálném světě. Například výšky lidí mají tendenci být normálně rozloženy kolem svého průměru a totéž platí pro řadu dalších proměnných, s nimiž se setkáváme v přírodě. V roce 1960 Eugene



Různé druhy Gaussovy křivky

Fama z chicagské univerzity ukázal, že procentové změny v cenách akcií a dalších aktiv prodávaných na dobře fungujících trzích se řídí normálním rozdělením. Tento objev vedl k založení kvantitativní finanční analýzy. Později se ukázalo, že i devizové výplatní poměry měn lze modelovat pomocí normálního rozdělení.

Normální rozdělení je užitečné, protože jeho hodnoty a pravděpodobnosti jsou spočítány a nalezneme je v tabulkách, a tak nám umožňují provést pravděpodobnostní výpočty pro libovolnou náhodnou proměnnou, která se řídí tímto rozdělením. Uveďme si jeho klíčové parametry.

Normální rozdělení

1. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením se neliší od střední hodnoty o více než jednu směrodatnou odchylku, je přibližně 68 %.
2. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením se neliší od střední hodnoty o více než dvě směrodatné odchylky, je přibližně 95 %.
3. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením se neliší od střední hodnoty o více než tři směrodatné odchylky, je přibližně 99,7 %.
4. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením se neliší od střední hodnoty o více než čtyři směrodatné odchylky, se blíží jistotě.

Jestliže se rozložení výšky mužů v nějaké populaci řídí normálním rozdělením s průměrem (střední hodnotou) 178 centimetrů a směrodatnou odchylkou 10 centimetrů, pak užitím výše uvedených platností dostáváme:

1. Pravděpodobnost, že výška náhodně vybraného muže bude v intervalu mezi 168 a 188 centimetry, je přibližně 68 %.
2. Pravděpodobnost, že výška náhodně vybraného muže bude v intervalu mezi 158 a 198 centimetry, je přibližně 95 %.

3. Pravděpodobnost, že výška náhodně vybraného muže bude v intervalu mezi 148 a 218 centimetry, je přibližně 99,7 %.

Chvost normálního rozdělení je oblast napravo od nějakého bodu na pravé straně od střední hodnoty nebo symetricky oblast nalevo od nějakého bodu na levé straně od střední hodnoty. Název „chvost“ má původ v myšlence W. Gossetta, který v roce 1908 nakreslil tento vtipný obrázek:

Dva chvosty Gaussovy křivky

Protože celková pravděpodobnost je vždy rovna jedné, neboli 100 %, můžeme na základě již uvedených pravidel spočítat chvostové pravděpodobnosti.

1. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením bude napravo od střední hodnoty o více než jednu směrodatnou odchylku, je přibližně 16 % (a podobně nalevo od střední hodnoty o více než jednu směrodatnou odchylku).

2. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením bude napravo od střední hodnoty o více než dvě směrodatné odchylky, je přibližně 2,5 % (a podobně nalevo od střední hodnoty o více než dvě směrodatné odchylky).

3. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením bude napravo od střední hodnoty o více než tři směrodatné odchylky, je přibližně 0,15 % (a podobně nalevo od střední hodnoty o více než tři směrodatné odchylky).

4. Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením bude napravo od střední hodnoty o více než čtyři směrodatné odchylky, je téměř nulová (a podobně nalevo od střední hodnoty o více než čtyři směrodatné odchylky).

A na závěr si uvedeme ještě jeden příklad. Denní směnný kurz eura k dolaru se po jistou dobu řídí podle normálního rozdělení se střední hodnotou 1,2 dolaru za 1 euro a směrodatnou odchylkou 0,05 dolaru. S jakou pravděpodobností dostaneme zítra za 1 euro 1,3 dolaru nebo více? Protože 1,3 se právě rovná střední hodnotě plus dvěma směrodatným odchylkám, vidíme z výše uvedených platností, že pravděpodobnost, že hodnota bude 1,3 nebo vyšší (totiž že bude ležet ve chvostu křivky), je 2,5 %.

KAPITOLA 18

Jak porozumět průzkumům veřejného mínění před prezidentskými (a jinými) volbami

Z předchozí kapitoly je nejužitečnější následující poznatek, který bychom si měli zapamatovat:

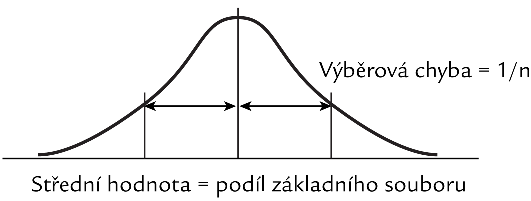
> Pravděpodobnost, že hodnota náhodné veličiny s normálním rozdělením padne do intervalu dvou směrodatných odchylek kolem střední hodnoty, je přibližně 95 %.

(Pro 95 % *přesně* stačí 1,96 směrodatné odchylky, takže užitím rámce dvou směrodatných odchylek dostáváme o trochu větší pravděpodobnost, přesně 95,44 %.)

Tento fakt nám pomůže pochopit průzkumy veřejného mínění a většinu dalších statistických namátkových průzkumů. Při průzkumu veřejného mínění se snažíme získat informace o celé populaci, například o populaci všech voličů v USA, na základě informací získaných z relativně malého *náhodného výběru* z tohoto základního souboru. Pokud všemi možnými způsoby zajistíme, že výběr je skutečně náhodný, můžeme se poté odvolávat na pravděpodobnostní zákony, neboť výběry z namátkových průzkumů se řídí podle normálního rozdělení.

Střední hodnota a směrodatná odchylka

Řekněme, že chceme provést průzkum veřejného mínění zjišťující, jakou veřejnou podporu má určitý politik. Rozsah náhodného výběru, který hodláme použít, je samozřejmě menší než velikost základního souboru. Všechny představitelné výsledky, které mnohé a mnohé různé výběry – totiž potenciální namátkové průzkumy – mohou dát, se řídí normálním rozdělením (viz obrázek) s většinou soustředěnou kolem středu a s několika extrémními případy na každé straně.



Normální rozdělení výběrového podílu

Nyní si zavedeme termín *výběrový podíl*, abychom vyjádřili, kolik procent z vybrané skupiny je pro určitou možnost (například zeptáme-li se deseti lidí, zda podporují určitého politika, a šest z nich odpoví ano, pak výběrový podíl je 60 %). Budeme také používat termín *výběrová propor*ce, což je vlastně výběrový podíl vyjádřený desetinným číslem (jinými slovy, vydělíme výběrový podíl číslem 100; v našem případě je tedy výběrová proporce 0,6). Nakonec ještě budeme používat termín *podíl základního souboru* – v hrubých rysech aktuální část základního souboru – která vybírá určitou odpověď (jestliže tedy ve skutečnosti 63 % základního souboru podporuje daného politika, pak podíl základního souboru, jak čtenář určitě uhodne, se rovná 63 %).

Střední hodnota normálního rozdělení výběrových podílů by se měla rovnat podílu základního souboru; čím více výběrů provedeme, s tím větší pravděpodobností se přiblížíme podílu základního souboru (když se podíváme na „zvonovou křivku“ na obrázku, vidíme, jak to dává smysl). Směrodatná odchylka normálního rozdělení výběrových podílů se rovná:

$$\sqrt{\frac{(\textit{výběrová proporce}) \times (1 - \textit{výběrová proporce})}{\textit{rozsah náhodného výběru}}}\,.$$

Předpokládá se, že rozsah náhodného výběru je mnohem menší než velikost základního souboru. Násobek pod odmocninou v uvedeném vzorci nabývá maxima pro výběrovou proporci rovnu 0,5 (druhý činitel je pak také 0,5) – násobek maximalizujeme, abychom pokryli všechny případy, a vzali tak do úvahy největší možnou chybu. Definujeme *výběrovou chybu* průzkumu, jíž se jinak též říká *mezní chyba*, jako *dvojnásobek* uvedené směrodatné odchylky. Dvojnásobek nám dává 95% pravděpodobnost, že předpověď našeho průzkumu bude správná. Nyní počítejme:

$$2 \times \sqrt{\frac{(0{,}5) \times (1 - 0{,}5)}{\textit{rozsah náhodného výběru}}}\,.$$

Což se rovná:

$$\frac{1}{\sqrt{\textit{rozsah náhodného výběru}}}\,.$$

Tato veličina, a priori – tedy předtím než provedeme výběr – maximalizuje naši výběrovou chybu, a použije se tedy jako výběrová chyba.

Výběrová chyba (mezní chyba) průzkumu veřejného mínění, který je správný s 95% pravděpodobností, je rovna

$$\frac{1}{\sqrt{\textit{rozsah náhodného výběru}}}\,.$$

Výsledek průzkumu veřejného mínění se prezentuje například takto: „65 % pro kandidáta s výběrovou chybou plus minus 5 %." V typické zprávě tohoto typu je to třeba chápat tak (není-li řečeno jinak), že výsledek je vždy správný *s 95% pravděpodobností*, a výsledek nám říká, že aktuální podíl základního souboru pro daného kandidáta v celém základním souboru, z něhož byl proveden výběr, je *kdekoli v intervalu kolem stanoveného podílu základního souboru plus minus stanovená výběrová chyba*. V našem příkladě se s 95% pravděpodobností nachází aktuální podíl voličů daného kandidáta v intervalu mezi 60 % a 70 %. Kvůli výběrové chybě je zde stále ještě 2,5% pravděpodobnost, že aktuální podíl pro kandidáta je pod 60 %; a 2,5% pravděpodobnost, že aktuální podíl pro kandidáta leží nad 70 %.

Můžete si také zkontrolovat, zda prezentované průzkumy v médiích říkají pravdu, anebo ne. (Podle mě tomu tak vždy není.) Jednoduše vezměte udávanou velikost náhodného výběru užitého při sledovaném průzkumu a dosaďte do předchozího vzorečku. Předpokládejte například, že uvedený průzkum se zakládal na náhodném výběru jednoho tisíce lidí. V tomto případě se výběrová chyba rovná:

$$\frac{1}{\sqrt{1000}} = 0{,}0316\,.$$

což je 3,16 %, zaokrouhleně tedy 3 %, a ne 5 %, jak průzkum udával. V opravdu seriózních průzkumech bude udaná výběrová chyba odpovídat našemu vzorci.