

Pravděpodobnostní modely

Menu: QCExpert Pravděpodobnostní modely

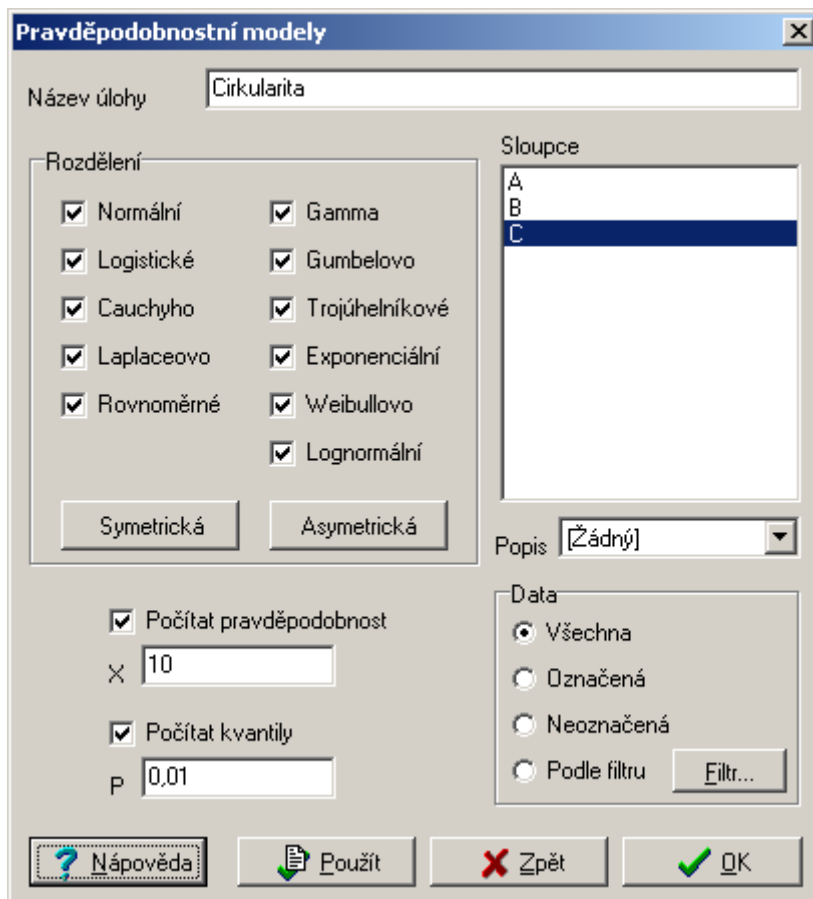
Modul hledá metodou maximální věrohodnosti (MLE, *Maximum Likelihood Estimate*) statistický model (rozdělení), který nejlépe popisuje data. Je přitom k dispozici celkem 11 jednorozměrných rozdělání, 5 symetrických a 6 asymetrických. Modul vypočítá parametry těchto rozdělání pro zadaná data a vyhodnotí nejlepší rozdělání podle dvou kritérií: věrohodnostní funkce a linearity (korelačního koeficientu) P-P grafu. Parametry jsou vypočítány numerickou maximalizací věrohodnostní funkce, resp. jejího logaritmu L ,

$$L(A, B, C, \mathbf{x}) = \sum_{i=1}^n \ln f(A, B, C, x_i),$$

kde A, B, C jsou parametry daného rozdělání, $\mathbf{x}$ je vektor n naměřených dat a f je hustota pravděpodobnosti daného rozdělání, viz dále. Kromě parametrů jednotlivých modelů a vybraných odpovídajících statistik se počítají rovněž zadané kvantily a pravděpodobnosti a konstruuji diagnostické grafy. Modul má použití všude, kde se k popisu dat nehodí normální rozdělání, kde jsou pochybnosti o vhodnosti určitého rozdělání, nebo naopak kdy se rozdělání dat řídí předem známým negaussovským modelem.

Data a parametry

Modul je určen pro analýzu dat **symetrických** a dat **s kladnou šikmostí**. Data se zápornou šikmostí je třeba před analýzou například vynásobit (-1) . Vstupními daty jsou naměřené hodnoty v jednom sloupci, který se vybere v poli *Sloupce*. V poli *Rozdělání* lze vybrat, které modely chceme zahrnout do výpočtu. Vlevo jsou rozdělání symetrická, vpravo asymetrická. Implicitně se počítá všech 11 rozdělání. Tlačítka *Symetrická* a *Asymetrická* slouží k vypnutí a zapnutí celé skupiny symetrických, resp. asymetrických rozdělání. Vybráno musí být vždy alespoň jedno rozdělání. Pokud nechceme uvažovat asymetrické rozdělání, je vhodné asymetrická rozdělání nevybírat, výpočet asymetrických rozdělání trvá déle. Pokud jsou data symetrická, či mají zápornou šikmost, může výpočet asymetrických modelů selhat, což se projeví zprávou „Není k dispozici“, nebo „Chyba“. V poli *Data* lze zvolit, zda se má výpočet provést se všemi daty, nebo pouze s daty označenými, případně neoznačenými. Je-li zaškrtnuto políčko *Počítat pravděpodobnost*, je nutné zadat požadovanou hodnotu (kvantil) do pole X , pro kterou se má pravděpodobnost (tedy odpovídající hodnota p příslušné distribuční funkce) vypočítat. V protokolu je pak pro tuto hodnotu vypočítána u všech rozdělání pravděpodobnost, že bude naměřena hodnota menší než zadané X . Je-li zaškrtnuto políčko *Počítat kvantily*, je nutno zadat pravděpodobnost (hodnotu mezi 0 a 1, $0 < p < 1$), pro níž se mají kvantily počítat. V protokolu jsou pak vypočítány ke každému rozdělání kvantily odpovídající pravděpodobnosti p a $1 - p$.



Obrázek 1 Dialogový panel pro Pravděpodobnostní modely

Dále uvádíme hustoty pravděpodobnosti všech počítaných rozdělení s případnými omezeními pro hodnoty parametrů. Hustota pravděpodobnosti je označena $f(x)$, distribuční funkce $F(x)$, kvantilová funkce $F^{-1}(x)$.

Normální rozdělení:

$$f(x) = \frac{1}{B\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-A}{B}\right)^2\right), \quad A \in \mathbb{R}, B > 0$$

Rovnoměrné rozdělení:

$$f(x) = \begin{cases} \frac{1}{B-A} & \text{pro } A \leq x \leq B \\ 0 & \text{jinak} \end{cases}, \quad A, B \in \mathbb{R}, A < B$$

Laplaceovo rozdělení:

$$f(x) = \frac{1}{2B} \exp\left(-\frac{|x-A|}{B}\right), \quad A \in \mathbb{R}, B > 0$$

Logistické rozdělení:

$$f(x) = \frac{1}{B} \frac{\exp\left(\frac{x-A}{B}\right)}{\left[1 + \exp\left(\frac{x-A}{B}\right)\right]^2}, \quad A \in \mathbb{R}, B > 0$$

Cauchyho rozdělení:

$$f(x) = \frac{1}{\pi B \left[1 + \left(\frac{x-A}{B} \right)^2 \right]}, \quad A \in R, B > 0$$

Asymetrická rozdělení

Exponenciální rozdělení:

$$f(x) = \frac{1}{B} \exp\left(-\frac{x-A}{B}\right), \quad x \geq A, B > 0$$

Gamma rozdělení:

$$f(x) = \frac{1}{B\Gamma(C)} \left(\frac{x-A}{B}\right)^{C-1} \exp\left(-\frac{x-A}{B}\right), \quad x > A, B > 0, C > 0$$

Gumbelovo rozdělení:

$$f(x) = \frac{1}{B} \exp\left(-\frac{x-A}{B}\right) \exp\left(-\exp\left(-\frac{x-A}{B}\right)\right), \quad A \in R, B > 0$$

Lognormální rozdělení:

$$f(x) = \frac{1}{C(x-A)\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\ln(x-A)-B}{C}\right)^2\right), \quad x > A, B \in R, C > 0$$

Trojúhelníkové rozdělení:

$$f(x) = \begin{cases} \frac{2(x-A)}{(B-A)(C-A)} & \text{pro } x < C \\ \frac{2(B-x)}{(B-A)(B-C)} & \text{pro } x \geq C \end{cases}, \quad x > A, B > C$$

Weibullovo rozdělení:

$$f(x) = \frac{C}{B} \left(\frac{x-A}{B}\right)^{C-1} \exp\left(-\left(\frac{x-A}{B}\right)^C\right), \quad x > A, B > 0, C > 0$$

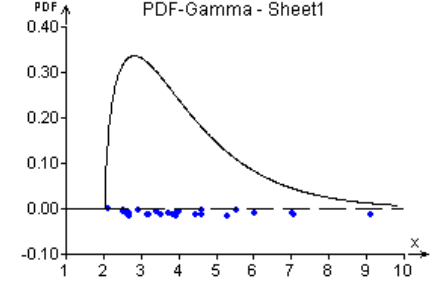
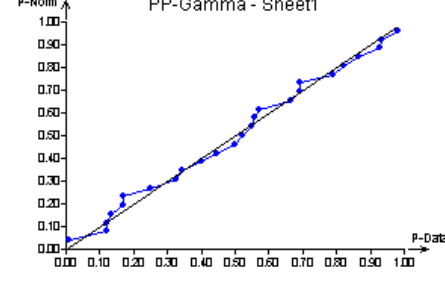
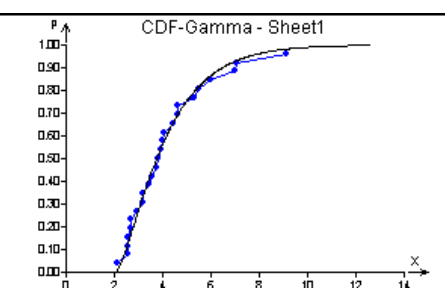
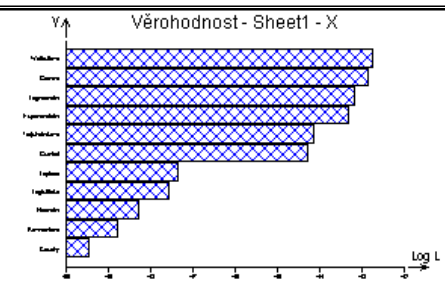
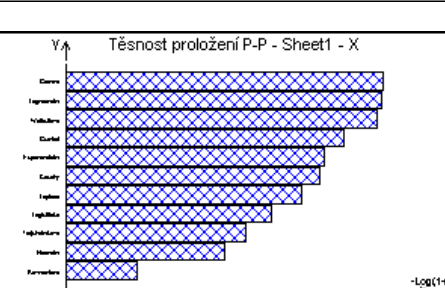
Protokol

Pravděpodobnostní modely, Metoda maximální věrohodnosti (MLE)	
Název úlohy :	Název sešitu s daty
Seznam analyzovaných rozdělení	Tento odstavec obsahuje seznam všech počítaných rozdělení, které byly zaškrtnuty v dialogovém okně. Modely jsou rozděleny do dvou skupin Symetrické modely a Asymetrické modely. U každého rozdělení jsou uvedeny vypočítané hodnoty logaritmu věrohodnosti, korelačního koeficientu P-P přímky a maximálně věrohodné odhady parametrů rozdělení.
Věrohodnost	Vypočítaná hodnota L logaritmu maximální věrohodnosti rozdělení. Věrohodnost je doporučena jako hlavní kritérium pro posouzení, které rozdělení nejlépe odpovídá naměřeným datům. Větší hodnota L odpovídá vhodnějšímu rozdělení.
P-P korelace	Korelační koeficient grafu P-P přímky r_P , která je obdobou Q-Q grafu. Platí, že čím je korelační koeficient větší (blíže k jedničce), tím více se pravděpodobnosti teoretické $F(x_i)$ blíží empirickým $i/(n+1)$. Hodnotu lze použít jako alternativní kritérium posouzení vhodnosti rozdělení,

	které nemusí dávat stejné výsledky jako věrohodnost.
Parametry	Parametry rozdělení získané metodou MLE, tedy tzv. maximálně věrohodné parametry A, B, C, význam parametrů pro jednotlivá rozdělení plyne z definic v odstavci 0.
Výběrové momenty	Výběrové odhady momentů.
Průměr	Aritmetický průměr dat. $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Rozptyl	Výběrový rozptyl dat. $s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$
Šikmost	Výběrová šikmost dat. $a = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$
Špičatost	Výběrová špičatost dat. $b = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$
Medián	Výběrový medián, $\tilde{x} = (x_{(n/2)} + x_{(n/2+1)})/2$ pro sudé n , $\tilde{x} = x_{((n+1)/2)}$ pro liché n .
Modelové momenty	Momenty, medián a modus analyticky vypočítané z parametrů jednotlivých modelů. Není-li k dispozici analytický výpočet, je uvedena pomlčka.
Střední hodnota	$\mu = \int_{-\infty}^{\infty} x dF(x)$
Rozptyl	$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 dF(x)$
Šikmost	$g_1 = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} (x - \mu)^3 dF(x)$
Špičatost	$g_2 = \frac{1}{\sigma^4} \int_{-\infty}^{\infty} (x - \mu)^4 dF(x)$
Medián	$F^{-1}(0.5)$
Modus	Poloha maxima na hustotě pravděpodobnosti $\max_x f(x)$.
Kvantily a pravděpodobnosti	Hodnota pravděpodobnosti pro zadaný kvantil X a hodnoty kvantilů pro zadanou pravděpodobnost p a pravděpodobnost $(1 - p)$ pro počítaná rozdělení. Tato tabulka se generuje jen je-li zadána některá z hodnot X , P v dialogovém okně (Obrázek 1).
Pravd(X)	Pravděpodobnost, že bude naměřena hodnota menší než X za předpokladu platnosti daného modelu pravděpodobnosti, $F(X)$.
Kvant(p)	Hodnota, která nebude překročena s pravděpodobností p (tedy která bude překročena s pravděpodobností $(1 - p)$), $F^{-1}(p)$.
Kvant($1 - p$)	Hodnota, která bude překročena s pravděpodobností p , $F^{-1}(1 - p)$.

Grafy

Pro každé z vybraných rozdělení se generuje graf hustoty pravděpodobnosti, P-P graf a graf distribuční funkce. Pro lepší vizuální hodnocení shody jsou do všech grafů zakreslena i zadaná data. Poslední dva grafy jsou určeny pro souhrnné vyhodnocení všech vybraných rozdělení.

	Graf hustoty pravděpodobnosti $f(x)$ pro každé z vybraných rozdělení. Pod osou x jsou pro lepší posouzení doplněna zadaná data. Vztahy pro hustoty jednotlivých rozdělení byly uvedeny výše. Tento graf nemá přímý diagnostický význam.
	Graf P-P, kde na vodorovnou osu se vynášejí hodnoty $i/(n+1)$ a na svislou osu hodnoty distribuční funkce v jednotlivých zadaných datech $F(x_i)$. V případě ideální shody dat s rozdělením by všechny body ležely na zakreslené přímce $y=x$. Těsnost proložení lze považovat za kritérium shody naměřených dat s daným modelem. Těsnost lze vyjádřit korelačními koeficienty, které jsou uvedeny v protokolu vedle hodnot věrohodností a graficky jsou znázorněny v grafu <i>Těsnost proložení</i>, viz dále.
	Graf distribuční funkce $F(x)$ vybraných rozdělení. Křivky jsou doplněny empirickou distribuční funkcí sestavenou z naměřených dat, jednotlivé body mají souřadnice $(x_i; i/(n+1))$. Těsnost proložení lze považovat za kritérium shody naměřených dat s daným modelem.
	Souhrnný graf věrohodností L pro všechna vybraná rozdělení seřazených podle velikosti. Nejlepší rozdělení z hlediska věrohodnosti je nejvýše a má nejvyšší hodnotu L. Mají-li dvě rozdělení velmi blízké hodnoty L, lze je obvykle považovat za rovnocenné. Samotná hodnota L souvisí s počtem dat a směrodatnou odchylkou rozdělení a není rozhodující. Věrohodnosti pro různě početná data, nebo data s různými směrodatnými odchylkami nelze mezi sebou porovnávat.
	Graf těsnosti proložení porovnává těsnost proložení v P-P grafu pomocí hodnot odvozených z korelačního koeficientu $- \ln(1 - r_p)$. Tato odvozená hodnota je použita proto, že samotné hodnoty korelačních koeficientů r_p by nevedly k dostatečnému vizuálnímu rozlišení. Rozdělení s nejvyšší hodnotou se považují za nejlépe vystihující empirickou distribuční funkci dat. Pořadí rozdělení v grafech věrohodnosti a těsnosti proložení není obvykle shodné.