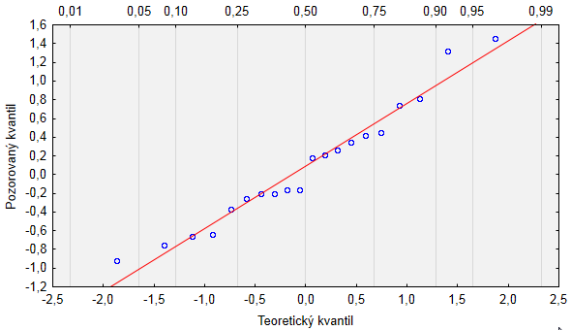
StatSoft

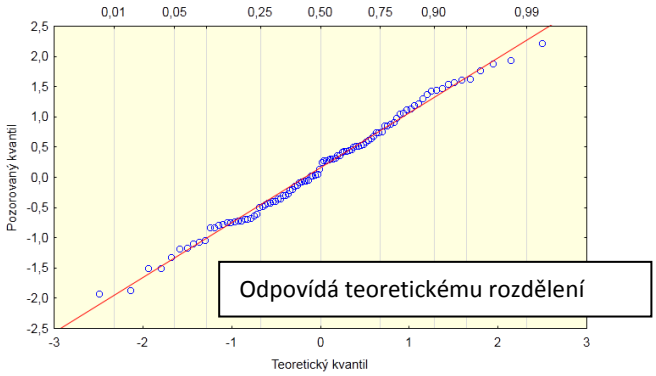
# Jak se pozná normalita pomocí grafů?

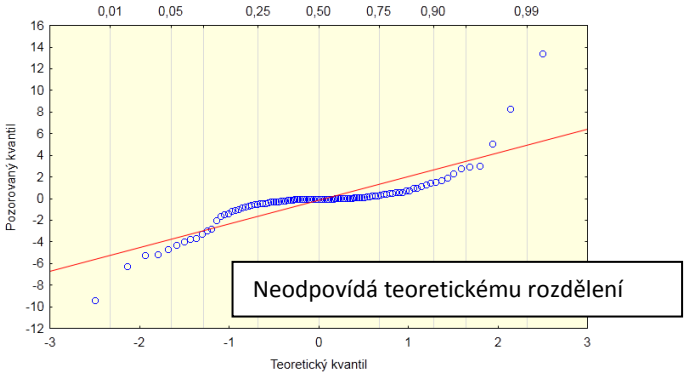



Dnes se podíváme na zoubek speciální třídě grafů, podle názvu článku a případně i ilustračního obrázku vpravo jste jistě již odhadli, že půjde o třídu pravděpodobnostních grafů (včetně Q-Q a P-P grafů). Protože jsou tyto grafy významným a hojně využívaným pomocníkem pro vizuální kontrolu předpokladů na rozdělení zkoumané náhodné veličiny, byla by škoda nenapsat o nich pár slov. Čeká Vás článek, kde se dozvíte, co přesně body v grafech znamenají a co všechno Vám může průběh grafu říci o rozdělení náhodné veličiny.

Úloha zjistit nebo ověřit, z jakého rozdělení pocházejí zkoumaná data, je velmi běžná a trápí nás stále znovu. Proto také byly navrženy grafické nástroje, které nám v tomto mohou pomoci (základní informace o tom, co je pravděpodobnostní rozdělení a jak jednotlivá rozdělení vypadají, naleznete **zde**). Nejjednodušším prostředkem pro hodnocení tvaru rozdělení je jistě histogram, nicméně ten má jednu zásadní vadu – musí se určit vhodný počet dělení (sloupců), aby z něj bylo možné vůbec něco vyčíst a pokud si nejste jistě počtem dělení, těžko můžeme z jeho tvaru něco vyvozovat. Mnohem vhodnější jsou grafy založené na porovnání kvantilů teoretického rozdělení a naměřených kvantilů (Q-Q grafy).

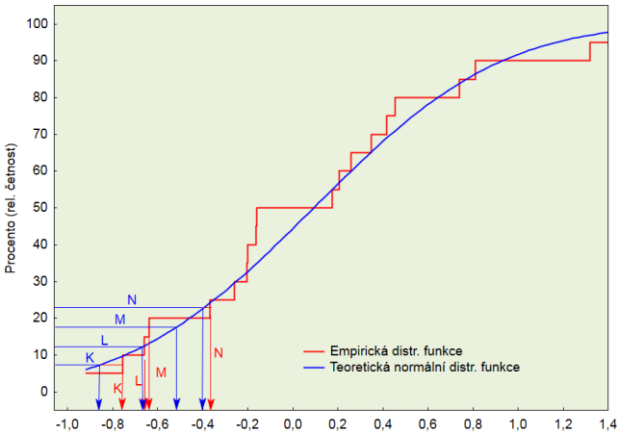
Vykreslování kvantilů proti sobě je samo o sobě dobrý nápad a zní celkem logicky, nicméně z naší zkušenosti víme, že se jedná pro uživatele spíše o tajemný graf, kde není zcela jasné, co se do grafu přesně vynáší, nejsou to prostě klasické body, které se dají jednoduše představit. Věříme, že po přečtení následujících odstavců bude už jasno.




# Co se do grafů vlastně vynáší

Pro jednoduchost bychom nejdříve popsali Q-Q graf a níže si poté vysvětlíme i grafy podobného typu a to P-P graf a takzvaný normální p-graf. Q-Q graf je zkratka pro kvantil-kvantil graf. Do něj se vykreslují proti sobě teoretické vůči opravdu naměřeným (pozorovaným) kvantilům (připomeňme, že $\alpha$ procentní kvantil je hodnota, pod kterou je $\alpha$ procent dat). Příklad takového grafu je vidět v motivaci či obrázcích výše. Jde tedy o jednoduché porovnání toho, co máme v datech s tím, co očekáváme.



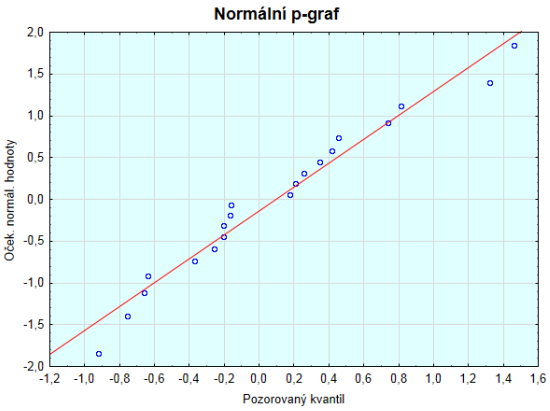

K sestavení grafu nám stačí určit, které kvantily vykreslíme. Určitě by nás napadlo interval (0,1) rozdělit na $n$ rovnoměrně vzdálených bodů ($n$ je počet dat a každé pozorování má stejnou důležitost). Nejlogičtější je asi volit $(i-0{,}5)/n$ kde $i = 1,2,\dots,n$ (pro 5 bodů by to bylo 0,1; 0,3; 0,5; 0,7 a 0,9). Když už víme, které kvantily chceme, stačí je vypočítat. Abychom pochopili situaci lépe, pomožme si vizualizací - na obrázku vpravo s teoretickou a empirickou distribuční funkcí (abychom měli teoretickou distribuční funkci, musíme na začátku mít zvoleno teoretické rozdělení, my zde volíme normální). Na obrázku je vidět vše, co jsme si popsali (i když pro větší $n$=20) - body K, L, M a N budou mít první souřadnici určenou modrou a druhou červenou šipkou příslušnou danému písmenu (povšimněme si, že červená šipka ukazuje přesně hodnoty v našem datovém souboru – jsou to skoky v empirické distribuční funkci).

*Poznámka:* volba $(i-0{,}5)/n$ se v praxi příliš nepoužívá, za to se používá spousta nejrůznějších variant ve tvaru $(i)/(n+1)$ nebo $(i-a)/(n+1-2a)$, v zásadě jde jen o to, mít vždy vybrané hodnoty v intervalu mezi $(i-1)/n$ a $(i)/n$. Nicméně, člověk si s tím nemusí příliš lámat hlavu, všechny možnosti poslouží svému účelu velmi podobně. Software *STATISTICA* využívá pro Q-Q graf defaultně $a = 0{,}375$, ale lze vše nastavit i manuálně. Pro normální p-graf je pak defaultně nastaveno $a = 1/3$. (Více detailů najdete v nápovědě softwaru v sekci „Conceptual Overviews").

Pokud bychom měli vyjádřit vzorcem, jaké hodnoty se ve *STATISTICE* vynášejí, pak to budou body se souřadnicemi: $\left(u_{(j)}, \mathrm{F}^{-1}\left(\frac{3j-1}{3n+1}\right)\right)$, kde $u_{(j)}$ je $j$-tá nejmenší hodnota v datech a $\mathrm{F}$ je distribuční funkce zkoumaného rozdělení.

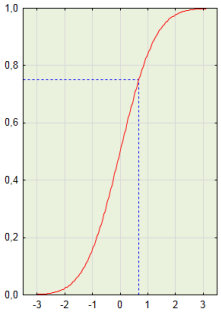
*Poznámka:* přesné odhady parametrů složitějších rozdělení se dají zvolit ručně uživatelem nebo jsou vybrány programem automaticky pomocí algoritmů pro toto určených (více detailů najdete v nápovědě softwaru v sekci „Conceptual Overviews").

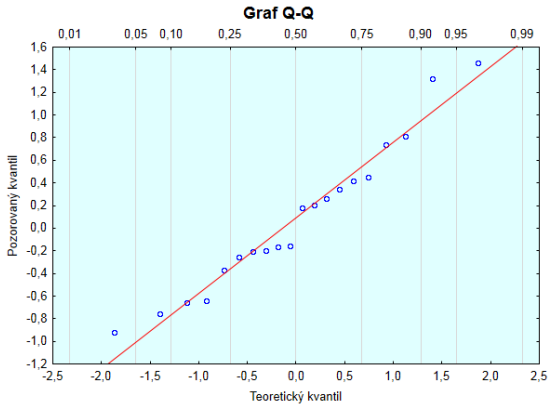
# Rozdíl mezi Q-Q, P-P a p-grafem



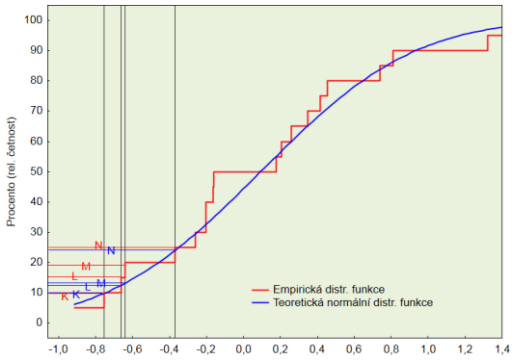

Podívejte se na následující obrázky (s modrým pozadím), jsou zde vykresleny všechny 3 typy grafů pro normální rozdělení na stejných datech. Zaměříme se na rozdíly mezi grafy.

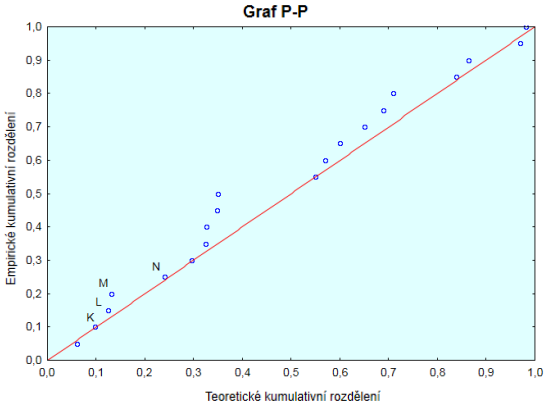
Ihned je jasné, že mezi normálním p-grafem (ten je ve *STATISTICE* definován jen pro normální rozdělení) a Q-Q grafem je rozdíl pouze ve výměně os, body mají stejné souřadnice - tedy pozorovaný a teoretický kvantil. U Q-Q grafu si můžeme nahoře všimnout pomocného druhého měřítka osy x. Toto měřítko je nerovnoměrné a vyjadřuje, u kterého kvantilu se s hodnotami pohybujeme – osa x má měřítko podle naměřených hodnot (použitou transformaci naznačuje následující obrázek distribuční funkce normálního rozdělení). Pokud přetransformujeme hodnoty horního měřítka tak, aby byly rovnoměrné, zjistíme, že dostaneme hodnoty v měřítku os v P-P grafu. Zatímco Q-Q znamená kvantil-kvantil, P-P znamená pravděpodobnost-pravděpodobnost.






U P-P grafu je vykresleno kumulativní rozdělení. Jednoduše: vezmou se hodnoty dat a stanoví se, který těmto bodům přísluší teoretický a pozorovaný kvantil a ty se poté vykreslí do P-P grafu. Souřadnice bodů na svislé ose tedy rostou rovnoměrně a jsou tedy rovny $i/n$). Názorně je princip vidět na grafu distribuční funkce níže (srovnejte body K, L, M, N v tomto a P-P grafu).




*Poznámka:* v softwaru lze zvolit alternativní vykreslení a chování těchto grafů v případě, že máme v datech několik stejných hodnot. Vysvětleme si následující zaškrtávátko:

☑ Neurčovat prům. pozici svázaných pozorování

Zatímco zaškrtnutí vykreslí všechny datové body, při zrušení zaškrtnutí vykreslí místo stejných hodnot jeden bod, přičemž pořadí, z kterého se počítá teoretický kvantil, bude průměrem pořadí stejných hodnot.

# Použití

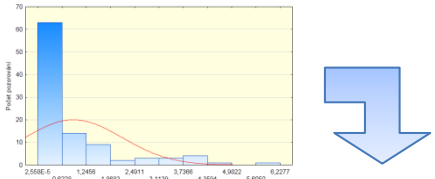
A teď to jednodušší, když už víme, co přesně v grafu vidíme, začne nám být jasné, co nám body v grafu říkají. Poněvadž jsou v grafu vykresleny teoretické a naměřené kvantily, jde vlastně o to, že pokud se data chovají stejně jako naše zamýšlené teoretické rozdělení, pak budou body ležet okolo přímky (teoretické a pozorované kvantily jsou si blízké). Čím blíže jsou body na přímce, tím blíž jsme k teoretickému rozdělení (u všech zmiňovaných grafů jde o to stejné – mít body blízko přímky). Pokud tedy chceme například mít normálně rozdělená data, pak nás potěší, když budou body v normálním pravděpodobnostním grafu ležet okolo přímky. Je třeba upozornit, že jde jen o vizualizační techniku, pokud potřebujeme statisticky otestovat shodnost s daným rozdělením, musíme se uchýlit ke statistickým testům, kterým se zde nyní věnovat nebudeme. V následující kapitole Vám zprostředkujeme příklady dat, kdy data normálně rozdělená nejsou, tím se naučíte grafy lépe číst a interpretovat.

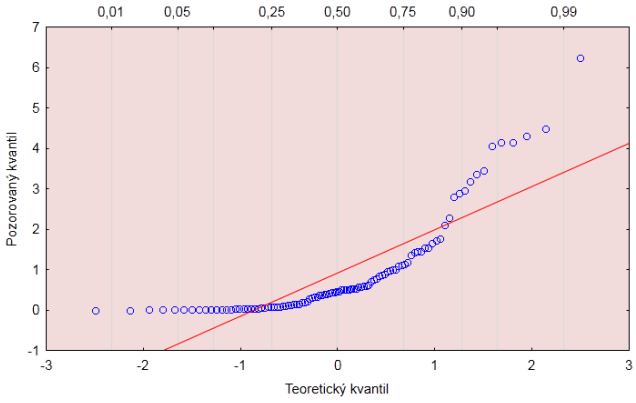
# Jak z Q-Q grafu vyčíst ještě více

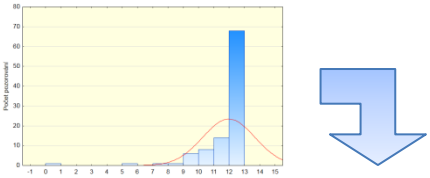
Na následujících příkladech uvidíte, že z grafu lze vyčíst o rozdělení veličiny mnohem víc než jen to, jestli souhlasí nebo nesouhlasí s hledaným rozdělením. Vše budeme ukazovat pro přehlednost jen na Q-Q grafech a normálním teoretickým rozdělením ve všech grafech. Uvědomme si také, že všechny grafy tedy slouží jako jakési porovnání s normálním rozdělením. Následují jednotlivé typy dat, která neodpovídají normálnímu rozdělení.

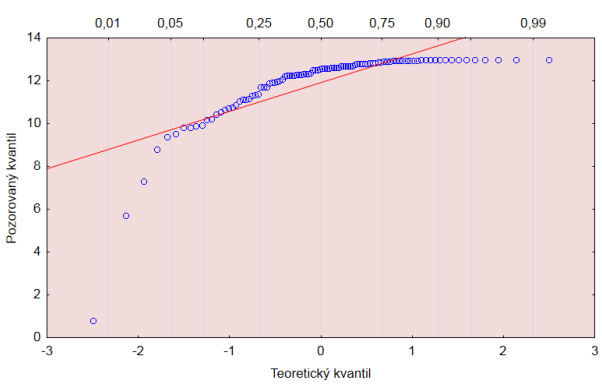
### Zešikmená (nesymetrická) data

Data v Q-Q grafu leží spíše než na přímce na nějaké parabole, to jasně svědčí o nesymetričnosti (zešikmení) – kvantily reálných dat jdou na jedné straně od průměru nahoru pomaleji a na druhé rychleji, což způsobí ve vykreslení proti normálním kvantilům prohnutí (na obrázku vpravo jsou data generovaná z Chí-kvadrát rozdělení).



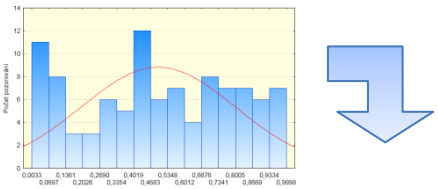


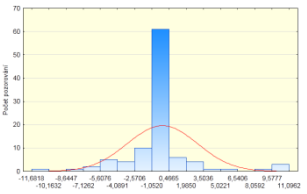



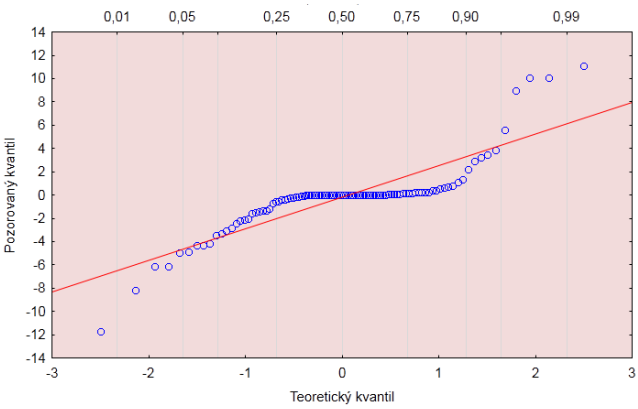


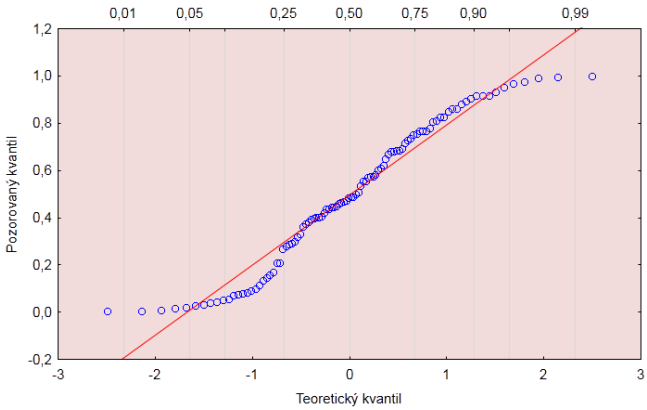
## Jiná špičatost oproti normálnímu rozdělení

Jinou špičatost jasně identifikuje S tvar vykreslených bodů. Data mají buď lehčí nebo těžší chvosty oproti normálnímu rozdělení, což vyvolává S tvar. Na obrázku vpravo je výběr s menší, výběr dole s větší špičatostí než normální rozdělení.



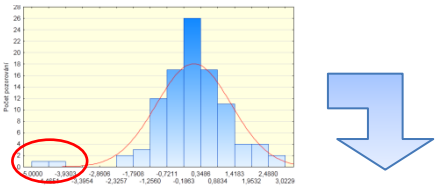


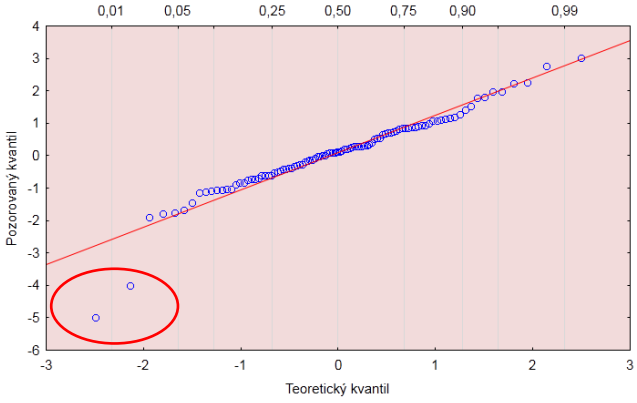



## Odlehlé hodnoty

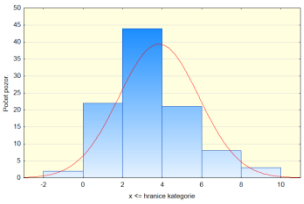
Odlehlé (vybočující) hodnoty jsou hodnoty, které se zjednodušeně chovají jinak než většina dat, mohou to být například chyby měření nebo pozorování, která jsou svou podstatou nějak abnormální. Na Q-Q grafu vidíme, že většina dat se chová slušně a vykazují normální rozdělení, zatímco dva body jsou úplně mimo a v tomto případě jasně znamenají odlehlé hodnoty.

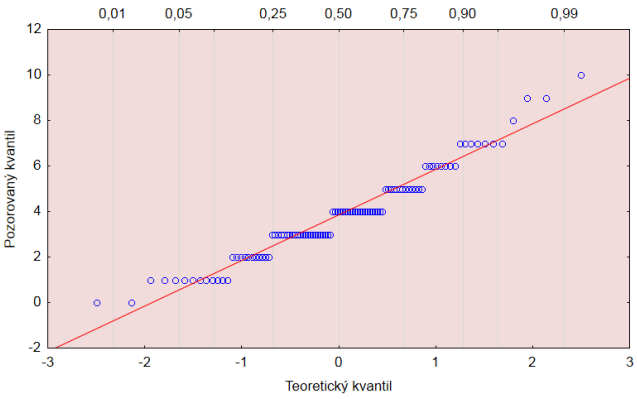




### Diskrétní rozdělení

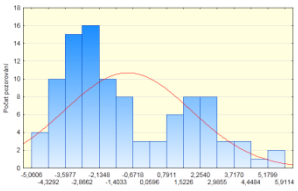
Pokud má veličina jen několik různých hodnot (což už samo o sobě naznačuje, že data nepocházejí z normálního rozdělení), pak vypadá Q-Q graf následovně – jasně vidíme diskrétnost grafu na svislé ose. Zde si můžeme všimnout také nevýhody histogramu, který se zdá být v pořádku, ale podstata dat není vidět kvůli nízkému počtu sloupců v grafu (data pocházejí z Poissonova rozdělení).

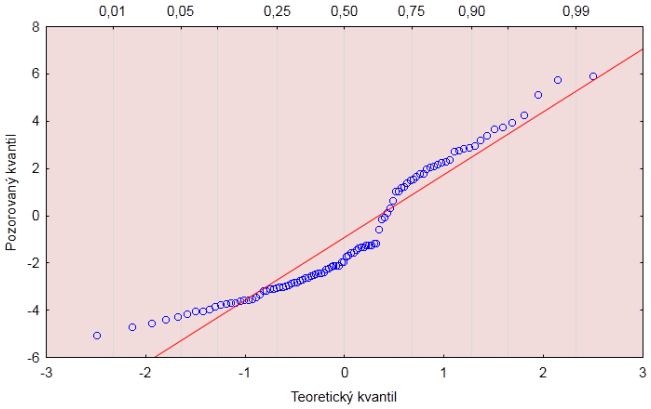





### Směs rozdělení (bimodální rozdělení)

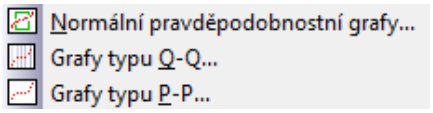
Pokud máme data, která vznikla jako směs dvou rozdělení (na obrázku směs normálních rozdělení) či se jedná o bimodální rozdělení, pak v Q-Q grafu uvidíme oddělené seskupení bodů, oproti problému s jinou špičatostí je rozdíl v tom, že ve střední části mezi seskupeními je malý počet bodů, u problému se špičatostí je většinou těchto bodů v centrální oblasti více a „ujíždějí" spíše kraje.






## Závěrem

V tomto článku zmíněné grafy lze v programu *STATISTICA* vytvořit v sekci *Grafy->2D grafy* nebo také jako výstup u různých analýz, kde jsou tyto grafy jako pomocné při testování předpokladů.




StatSoft
STATISTICA

**StatSoft CR s. r. o.**
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
**t** +420 233 325 006, **f** +420 233 324 005
**e** info@statsoft.cz