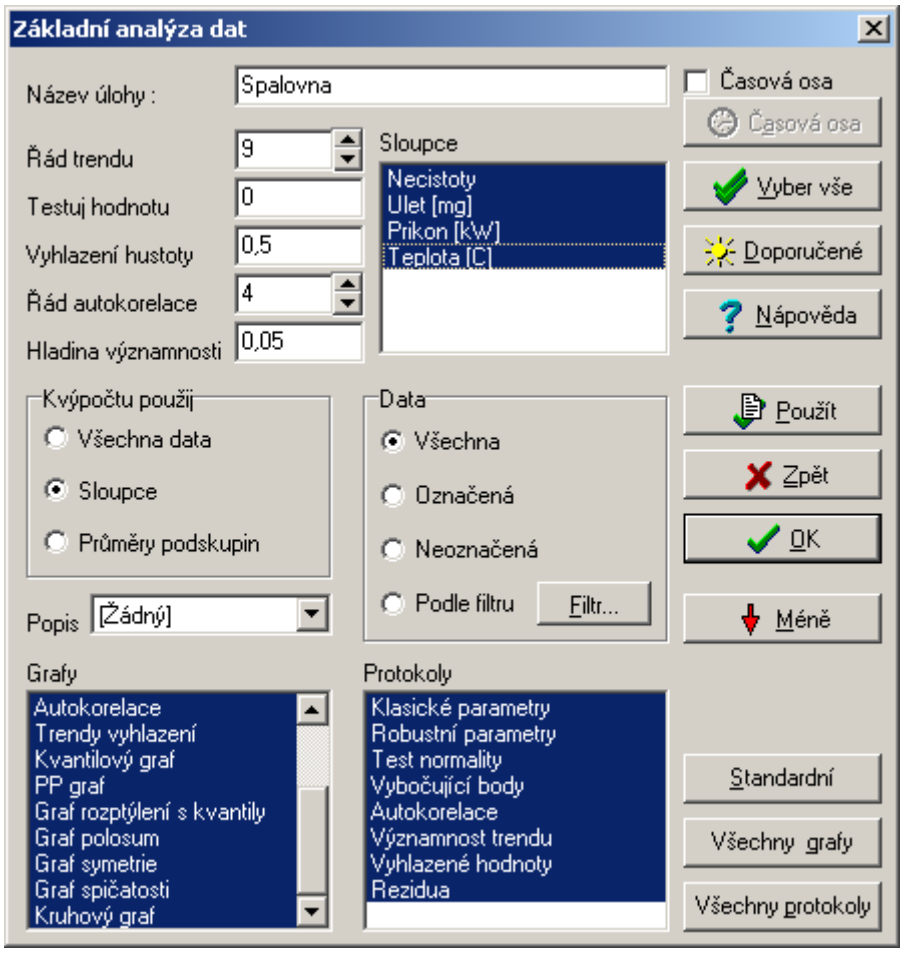
Modul Základní statistika

Menu:	QCExpert	Základní statistika

Základní statistika slouží k předběžné analýze a diagnostice dat, testování předpokladů (vlastností dat), jejichž splnění je nutné pro použití dalších metod, a k podrobnému seznámení s daty. K základním předpokladům o datech patří normalita, nezávislost a homogenita, tedy nepřítomnost vybočujících měření, odlehlých dat a hrubých chyb.



Obrázek 1 Zadání parametrů pro základní statistiku

Data a parametry

Data pro výpočet jsou organizována do sloupců (proměnných). V prvním řádku jsou vždy názvy sloupce. Jsou-li ve vstupním dialogovém panelu vybrány „Všechna data“ nebo „Sloupce“, mohou být délky jednotlivých sloupců různé. Chybějící data jsou ignorována. Má-li se výpočet provést pro „Průměry podskupin, měly by všechny sloupce mít stejnou délku. Minimální počet sloupců je 1. Minimální počet dat je 3. Výběr sloupců stejně jako zadání dalších parametrů se provede v okénku *Sloupce* na panelu *Základní analýza dat*, viz Obrázek 1.

Řád trendu určuje z kolika po sobě jdoucích dat budou počítány klouzavé průměry a klouzavé mediány. Hodnota by měla být menší než polovina počtu dat.
Testuj hodnotu Zde se zadává hodnota pro t-test. Program testuje na zadané hladině významnosti, zda tato hodnota může být shodná se střední hodnotou dat.
Vyhlazení hustoty udává šířku jádra pro jádrové vyhlazení pro graf hustoty pravděpodobnosti. Čím je větší, tím bude křivka hustoty pravděpodobnosti hladší a naopak. Hodnota musí být větší než nula, doporučuje se hodnota kolem 0.5.

Řád autokorelace udává do kterého řádu se budou počítat autokorelační koeficienty. Hodnota musí být alespoň o 2 menší než počet platných dat.
Hladina významnosti udává spolehlivost pro intervaly spolehlivosti a statistické testy. Musí být větší než nula a menší než 0.5. Vynásobena 100 udává hodnotu v procentech. Obvyklá hodnota je 0.05 (tedy 5%).
K výpočtu použij:
Všechna data všechny vybrané sloupce budou brány jako jediný sloupec.
Sloupce Výpočet se provede pro každý z vybraných sloupců zvlášť.
Průměry podskupin Výpočet se provede pro řádkové průměry z vybraných sloupců. Pokud nejsou sloupce stejně dlouhé, nebo obsahují chybějící data, provede se výpočet jen pro úplné řádky. Tento výpočet má význam především pro diagnostiku dat pro regulační diagramy typu x-průměr.
Časová osa, okénko zaškrtneme, obsahují-li data i sloupec s časovým údajem. Příslušný sloupec s časem pak specifikujeme stisknutím tlačítka *Časová osa*.
Vyber vše vybere k analýze všechny sloupce v aktivním listu.
Doporučené přepíše hodnoty v dialogovém panelu na obvyklé hodnoty. Toto tlačítko lze použít, nejsme-li si jisti, zda jsme zadali korektní hodnoty.
Více / Méně zobrazí/schová rozšíření panelu pro specifikaci výstupů. Zde můžeme definovat, které informace chceme mít v protokolu, a které grafy chceme konstruovat v grafickém okně. Tlačítko *Standardní* označí obvyklý zkrácený výstup s nejdůležitějšími informacemi, tlačítka *Všechny grafy* a *Všechny protokoly* označí kompletní výstup.

Poznámka: Velikost položky *Vyhlazené hodnoty* a *Rezidua* v protokolu závisí na počtu dat a pro veliké soubory může zaplnit výstupní list.

Protokol

Sloupec	Název sloupce.
Počet řádků	Celkový počet řádků v datech.
Počet platných dat	Celkový počet platných dat v datech.
Počet chybějících hodnot	Počet prázdných buněk v datech.
Klasické parametry	
Aritmetický průměr	Odhad střední hodnoty pro normálně rozdělená data.
Spodní mez	Spodní mez intervalu spolehlivosti aritmetického průměru na zadané hladině významnosti.
Horní mez	Horní mez intervalu spolehlivosti aritmetického průměru na zadané hladině významnosti.
Rozptyl	Odhad rozptylu.
Směrodatná odchylka	Druhá odmocnina z rozptylu.
Spodní mez	Spodní mez intervalu spolehlivosti směrodatné odchylky na zadané hladině významnosti.
Horní mez	Horní mez intervalu spolehlivosti směrodatné odchylky na zadané hladině významnosti.
Robustní směr.odch.	Odhad robustní směrodatné odchylky (*MAD*) $\sigma_{MAD} = K.\text{median}\left(\left\|x_i - \text{median}\left(\mathbf{x}\right)\right\|\right)$, kde $K = F^{-1}(0.75) = 1.482602$
Detrendovaná směr.odch. (MR)	Odhad směrodatné odchylky na základě klouzavých diferencí prvního řádu, $s_{MR} = \frac{1}{d_2\left(N-1\right)}\sum_{i=2}^{N}\left\|x_{i-1} - x_i\right\|$, kde d_2 je korekce vychýlení d_2=1.128. Tento odhad charakterizuje směrodatnou odchylku pro krátkodobou variabilitu bez případné autokorelace a používá se například v Shewhartových diagramech a pro výpočet způsobilosti *Cp*.
Šikmost	Odhad třetího statistického momentu, šikmosti.

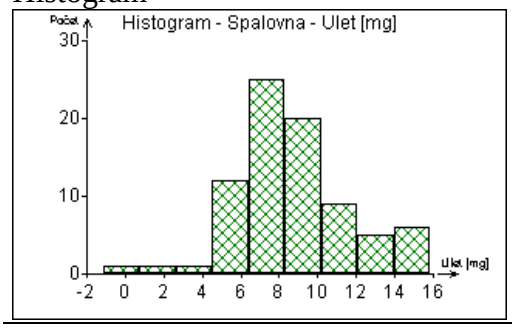
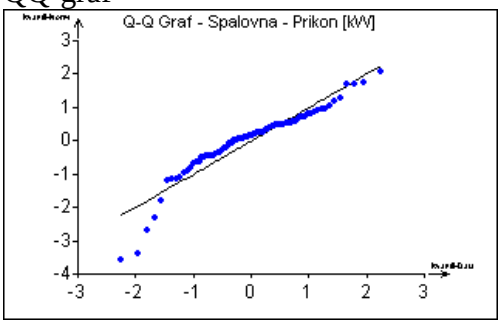
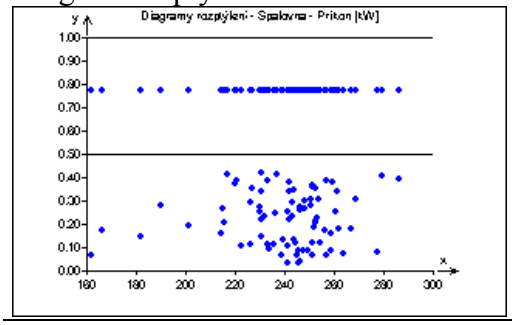
Rozdíl od 0	Normální a každé symetrické rozdělení má šikmost nulovou. Je-li hodnota šikmosti statisticky významně odlišná od 0, nelze data považovat za symetrická. Spolehlivější je však test normality.
Špičatost	Odhad čtvrtého statistického momentu, špičatosti.
Rozdíl od 3	Normální rozdělení má špičatost 3. Je-li hodnota špičatosti statisticky významně odlišná od 3, lze předpokládat, že data neodpovídají normálnímu rozdělení. Spolehlivější je však test normality.
Polosuma	Odhad polosumy, tedy středu nejmenší a největší hodnoty.
Modus	Odhad modu rozdělení, tedy maxima na křivce hustoty pravděpodobnosti.
Geometrický průměr	Geometrický průměr $\bar{x}_G = \left[\prod_{i=1}^{N} x_i\right]^{1/N} = \frac{1}{N}\exp\left(\sum_{i=1}^{N} \ln x_i\right)$ je definován pouze pro kladné hodnoty x_i.
Harmonický průměr	Harmonický průměr $\bar{x}_H = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i}}$ je definován pouze pro kladné hodnoty x_i.
t-test	
Předpokládaná hodnota	Hodnota zadaná v poli „Testuj hodnotu“ panelu „Základní analýza dat“ .
Rozdíl od předpokl. hodnoty	Slovní vyjádření, zda je rozdíl střední hodnoty dat od předpokládané hodnoty statisticky významný na dané hladině významnosti. Předpokládanou hodnotu lze zadat v dialogu Základní analýza dat, Obrázek 1.
Vypočtený	Vypočtená testovací statistika.
Teoretický	Příslušný kvantil t-rozdělení.
Pravděpodobnost	Pravděpodobnost nevýznamnosti tohoto rozdílu odpovídající vypočtené statistice. Je-li menší než zvolená hladina významnosti, rozdíl je významný.
Robustní parametry	
Medián	Odhad mediánu, tedy 50% kvantilu. Tento odhad střední hodnoty je spolehlivější než aritmetický průměr v případě porušení normality dat nebo přítomnosti vybočujících bodů.
IS spodní	Spodní mez intervalu spolehlivosti mediánu na zadané hladině významnosti.
IS horní	Horní mez intervalu spolehlivosti mediánu na zadané hladině významnosti.
Mediánová směr. odchylka	Odhad směrodatné odchylky na základě mediánu.
Mediánový rozptyl	Odhad rozptylu na základě mediánu.
10% uřezaný průměr	Aritmetický průměr pro symetrickém uřezání 10% dat, tedy 5% nejmenších a 5% největších hodnot. Tento robustní odhad střední hodnoty se doporučuje v případě podezření na vybočující body.
IS spodní	Spodní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
IS horní	Horní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
Rozptyl	Odhad rozptylu na základě mediánu.
Směr. odchylka	Odhad směrodatné odchylky na základě mediánu.

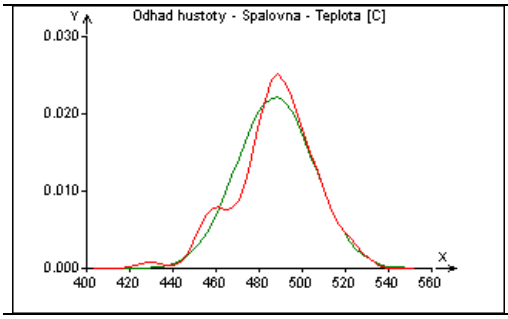
20% uřezaný průměr	Aritmetický průměr pro symetrickém uřezání 20% dat, tedy 10% nejmenších a 10% největších hodnot. Tento robustní odhad střední hodnoty se doporučuje v případě podezření na vybočující body.
IS spodní	Spodní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
IS horní	Horní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
Rozptyl	Odhad rozptylu na základě mediánu.
Směr. odchylka	Odhad směrodatné odchylky na základě mediánu.
40% uřezaný průměr	Aritmetický průměr pro symetrickém uřezání 40% dat, tedy 20% nejmenších a 20% největších hodnot. Tento robustní odhad střední hodnoty se doporučuje v případě podezření na velký počet vybočujících bodů.
IS spodní	Spodní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
IS horní	Horní mez intervalu spolehlivosti uřezaného průměru na zadané hladině významnosti.
Rozptyl	Odhad rozptylu na základě mediánu.
Směr. odchylka	Odhad směrodatné odchylky na základě mediánu.
Znaménkový test	Testuje náhodnost ve střídání hodnot vyšších a nižších než průměr. Je-li toto střídání příliš pravidelné a vyskytují-li se nepravděpodobně dlouhé sekvence po sobě jdoucích dat nad nebo pod průměrem, jsou data podezřelá a jsou označena jako závislá.
Analýza malých výběrů	Odhady střední hodnoty a její interval spolehlivosti na základě postupu podle Hornové založeného na kvantilech. Tento odhad se doporučuje pro malé výběry od 3 dat, kdy obvykle poskytuje správnější hodnoty než prostý průměr. Tyto odhady by se neměly používat při N>20.
N	Počet dat.
Střední hodnota	Odhad střední hodnoty.
Spodní mez	Spodní mez intervalu spolehlivosti.
Horní mez	Horní mez intervalu spolehlivosti.
Test normality	Kombinovaný test normality založený na shodě šikmosti a špičatosti s normálním rozdělením. Ve výstupu se opakují hodnoty klasických parametrů.
Test normality	Slovní závěr testu na zadané hladině významnosti.
Momentový	
Vypočtený	Vypočtená testovací statistika.
Teoretický	Příslušný kvantil t-rozdělení.
Pravděpodobnost	Pravděpodobnost odpovídající vypočtené statistice.
Test normality D'Agostino	Zdokonalený test podle D'Agostina posuzuje výběrové momenty dat. Obecně je tento test podstatně citlivější, než předchozí jednoduchý momentový test.
Test pro menší výběry	Verbální výsledek testu vhodného spíše pro menší výběry (zhruba N<100)
p-hodnota	p-hodnota (p-value) testu. Je-li menší, než zadaná hladina významosti, normalita je zamítnuta.
Test pro větší výběry	Verbální výsledek testu vhodného spíše pro menší výběry (zhruba N>100)

p-hodnota	p-hodnota (p-value) testu. Je-li menší, než zadaná hladina významosti, normalita je zamítnuta.
Test normality Kolmogorov-Smyrnov	Test normality založený na rozdílu teoretické a výběrové distribuční funkce korigovaný pro neznámé (odhadované) μ a σ. Posuzuje se detailní rozdílnost distribuční funkce dat pro případ, že by nenormální rozdělení dat mělo náhodou podobné momenty jako normální rozdělení a nenormalita by tedy nebyla detekována v předchozích testech.
Kritický kvantil chi2(22)	Kritická hodnota maximálního rozdílu distribučních funkcí D z rozdělení $chi^2(\nu=22)$
Testové kriterium D	Maximální dozdíl mezi dstribučními fukcemi
p-hodnota	p-hodnota (p-value) testu. Je-li menší, než zadaná hladina významosti, normalita je zamítnuta.
Normalita	Verbální výsledek testu (Přijata / Zamítnuta) na zadané hladině významnosti.
Vybočující body	Robustní test na přítomnost vybočujících měření založený na kvantilovém odhadu vnitřních mezí dat.
Homogenita	Slovní závěr testu, nejsou-li v datech vybočující měření, je předpoklad homogenity přijat.
Počet vybočujících bodů	Počet případných měření přesahujících přípustné meze, které je možno považovat za vybočující.
Dolní hranice	Dolní hranice, pod níž je možno data považovat za vybočující.
Horní hranice	Horní hranice, nad níž je možno data považovat za vybočující.
Autokorelace	Odhady autokorelačních koeficientů a jejich významnost na zadané hladině významnosti.
Řád autokorelace	Řád autokorelace.
Koeficient	Hodnota autokorelačního koeficientu, formálně odpovídá párovému korelačnímu koeficientu a má stejné vlastnosti.
Pravděpodobnost	Pravděpodobnost nevýznamnosti autokorelační-ho koeficientu; je-li menší než zvolená hladina významnosti, je autokorelace významná.
R0Krit	Kritická hodnota autokorelačního koeficientu, nad níž se korelace považuje za významnou.
Výsledek	Slovní vyjádření významnosti autokorelace.
Test významnosti trendu	Test významnosti lineárního trendu v datech. Nevýznamnost lineárního trendu nemusí znamenat, že v datech není jiný trend jako kolísání, oscilace, nebo jiný nelineární trend. Ten je obvykle indikován testem autokorelace nebo znaménkovým testem.
Směrnice	Hodnota směrnice přímky proložené daty
Významnost	Slovní vyjádření statistické významnosti trendu
Teoretický	Příslušný kvantil t-rozdělení.
Pravděpodobnost	Pravděpodobnost toho, že je lineární trend nevýznamný; vyjde-li menší než zadaná hladina významnosti (tedy obvykle 0.05), považuje se trend za významný.
Vyhlazené hodnoty	Hodnoty klouzavých průměrů a mediánů. Při vhodné volbě řádu trendu lze získat po odečtení těchto hodnot od vstupních dat detrendovaná data, která lze použít pro konstrukci regulačních diagramů v případě, že je v datech příliš silný trend.
Rezidua	Odchylky (rozdíly) naměřených a vyhlazených hodnot, (naměřená -

	vyhlazená).

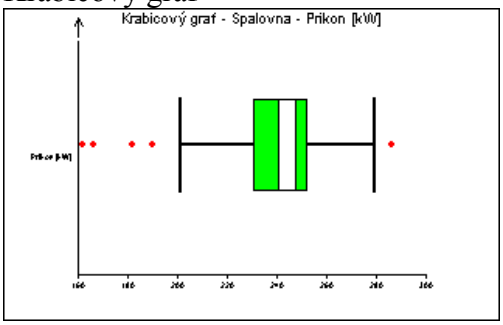
Grafy

Histogram 	Histogram četností dat v jednotlivých třídách s konstantní šířkou; optimální počet tříd je stanovován automaticky s ohledem na počet dat. Kliknutí pravým tlačítkem na dynamický graf zobrazí četnost pro daný sloupec (třídu) a meze třídy.
QQ-graf 	Graf pro diagnostiku normality a odlehlých měření; pro normální data bez odlehlých měření má tvar přímky; pro normální data s odlehlými měřeními má tvar přímky s koncovými body ležícími mimo tuto přímku; pro systematicky sešikmená data s kladnou šikmostí (např. rozdělení lognormální, exponenciální) má nelineární konvexní tvar  . Pro systematicky sešikmená data se zápornou šikmostí má nelineární konkávní tvar  . Pro data s vyšší špičatostí než odpovídá normálnímu rozdělení, tedy s vysokou koncentrací dat kolem střední hodnoty (např. Laplaceovo rozdělení) má tvar konkávně-konvexní  . Pro data s nižší špičatostí než odpovídá normálnímu rozdělení, tedy s malou koncentrací dat kolem střední hodnoty (např. rovnoměrné rozdělení) má tvar konvexně-konkávní  . Proti statistikám má QQ-graf výhodu v možnosti vizuálně posoudit, zda je nelinearita způsobena jen několika body, nebo všemi daty. Ze zkušenosti většinou platí: protíná-li spojnice mezi body přímku více než 4krát, přikloníme se spíše k normalitě, protíná-li spojnice mezi body přímku méně než 4krát, přikloníme se spíše k porušení normality.
Diagram rozptýlení 	Zobrazuje všechna data ve skutečném měřítku na ose X. Popis osy Y nemá význam. Aby nedošlo ke splývání shodných nebo blízkých dat, jsou ve spodní polovině grafu zobrazena táž data, ale náhodně rozmítnutá („rozházená“) na ose Y.
Hustota pravděpodobnosti	Porovnání průběhu hustoty pravděpodobnosti normálního rozdělení (plná zelená čára) s jádrovým odhadem hustoty vypočítaným na základě dat (přerušovaná červená čára). Jádrový odhad používá Gaussovské jádro. Hladkost křivky jádrového odhadu je dána parametrem “Vyhlazení hustoty” v dialogovém



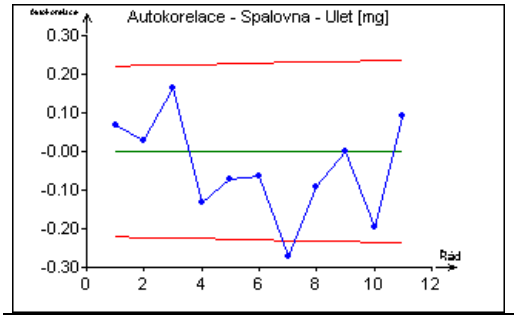
panelu, viz Obrázek 1, čím je parametr menší, tím podrobnější je průběh. Jsou-li data nehomogenní a tvoří shluky, může se objevit více maxim na jádrovém odhadu. V případě normality a většího množství dat jsou si obě křivky blízké. Je však třeba si uvědomit, že při příliš malém parametru vyhlazení hustoty se objeví maxima pro každá data.

Krabicový graf



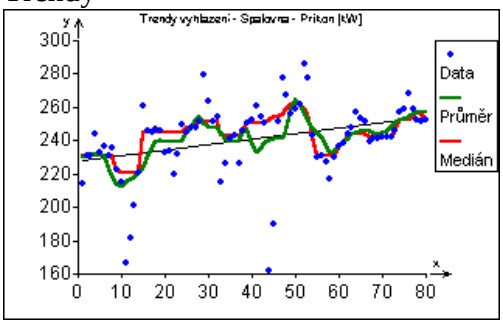
Standardní diagnostický graf. Větší obdélník ohraničuje vnitřních 50% dat, horní okraj zeleného (vyšrafovaného) obdélníku odpovídá 75% kvantilu, spodní okraj zeleného obdélníku odpovídá 25% kvantilu, střed bílého pruhu v zeleném obdélníku odpovídá mediánu, šířka pruhu odpovídá intervalu spolehlivosti mediánu, dva černé proužky jsou tzv. vnitřní hradby. Data mimo vnitřní hradby jsou označena červeným bodem a lze je považovat za vybočující měření.

Autokorelace



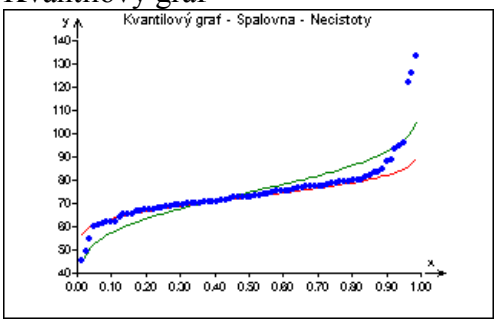
Graf autokorelačních koeficientů až do řádu autokorelace zadaného v dialogovém panelu, viz Obrázek 1. Červené meze ohraničují interval, v němž jsou koeficienty statisticky nevýznamné na zadané hladině významnosti. Překročí-li některý autokorelační koeficient tyto meze, je třeba považovat za závislá.

Trendy



Grafické znázornění trendu v datech pomocí klouzavého průměru (plná křivka) a klouzavého mediánu (přerušovaná křivka) na základě "Řádu trendu" zadaného v dialogovém panelu, viz Obrázek 1. Čím je řád trendu větší, tím jsou křivky hladší, méně citlivé na lokální poruchy a zachycují spíše globální dlouhodobý trend. Menší řád zachycuje spíše lokální chování dat. Klouzavý medián je méně citlivý (robustní) na lokální poruchy v datech a jednotlivá odlehlá měření a tedy vhodnější pro tyto případy. Je-li v datech významný lineární trend, je v grafu ještě zobrazena příslušná regresní přímka.

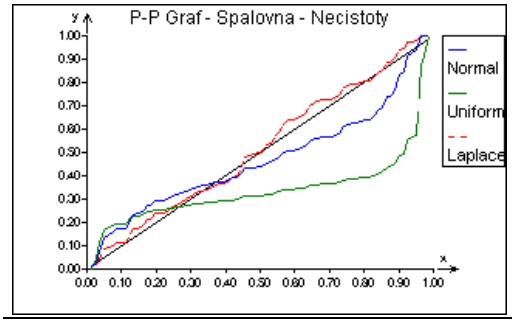
Kvantilový graf



Zobrazuje empirické kvantily dat proložené kvantilovou funkcí normálního rozdělení. Zelená křivka odpovídá funkci s klasickým průměrem a rozptylem (nerobustní), červená křivka odpovídá mediánu a mediánové odchylce (robustní). Podle toho, která z křivek lépe prokládá data, je vhodné zvolit jako odhad střední hodnoty průměr nebo medián.

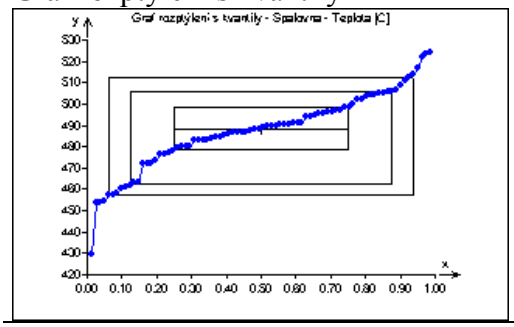
PP-graf

Porovnává data s normálním (modrá křivka plná), Laplaceovým (zelená křivka) a rovnoměrným



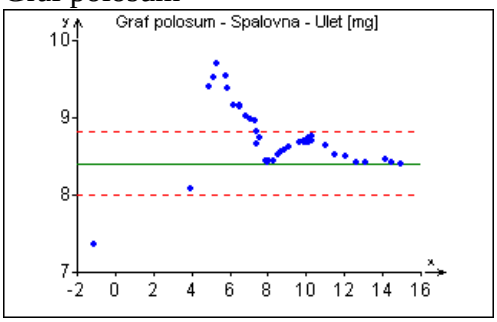
rozdělením pomocí teoretické a empirické distribuční funkce. Která křivka leží nejblíže černé přímce $y = x$, to rozdělení odpovídá experimentálním datům. Graf slouží pro rozlišení symetrických rozdělení podle špičatosti. Podobnost rovnoměrnému rozdělení ukazuje na možné vyloučení vysokých a nízkých hodnot, podobnost s Laplaceovým rozdělením ukazuje na možnou nekonstantnost rozptylu dat.

Graf rozptýlení s kvantily



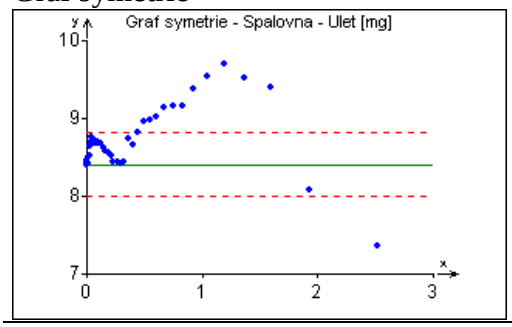
Body grafu jsou vizuálně i významově shodné s kvantilovým grafem. Vzájemná poloha obdélníků odpovídá symetrii, resp. asymetrii rozdělení. Vodorovná příčka uprostřed nejmenšího obdélníku označuje medián, svislá úsečka na příčce odpovídá intervalu spolehlivosti mediánu.

Graf polosum



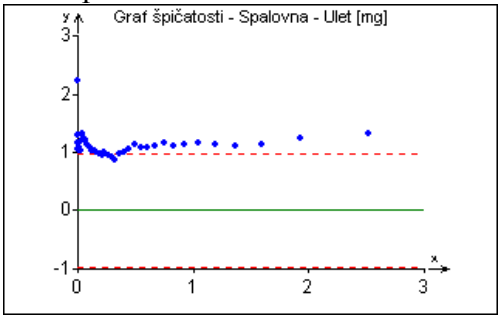
Citlivý indikátor asymetrie rozdělení. V ideálním případě leží body na horizontální přímce. Horizontální přímka, na níž leží poslední bod, představuje medián a červené přerušované meze jeho interval spolehlivosti. V případě asymetrického rozdělení vykazují body výrazný trend (rostoucí pro zápornou šikmost, nebo klesající, pro kladnou šikmost) výrazně překračující přerušované meze. Body jsou konstruovány ze dvojic dat (první, poslední; druhý, předposlední; atd.), proto označením bodu jsou označena dvě příslušná data v tabulce.

Graf symetrie



Má podobný význam jako předchozí graf polosum. Směrnice případného trendu je úměrná šikmosti. V případě asymetrického rozdělení vykazují body výrazný trend (rostoucí pro zápornou šikmost, nebo klesající, pro kladnou šikmost) výrazně překračující přerušované meze. Body jsou konstruovány ze dvojic dat (první, poslední; druhý, předposlední; atd.), proto označením bodu jsou označena dvě příslušná data v tabulce.

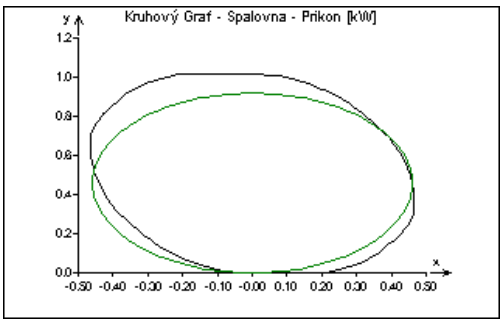
Graf špičatosti



Má podobný význam jako dva předchozí grafy. Směrnice případného trendu je úměrná odchylky špičatosti od 3. V případě výrazně nenormální špičatosti rozdělení vykazují body výrazný trend. Body jsou konstruovány ze dvojic dat (první, poslední; druhý, předposlední; atd.), proto označením bodu jsou označena dvě příslušná data v tabulce.

Kruhový graf

Slouží ke komplexnímu vizuálnímu posouzení normality na základě kombinace šikmosti a špičatosti.



Zelený kruh (elipsa) je optimální tvar pro normální rozdělení, černý „kruh“ představuje data. V případě normálních dat se obě křivky téměř kryjí.