evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

TÉMA 2
PRŮZKUMOVÁ ANALÝZA DAT - EDA

CO BYSTE MĚLI PO PROSTUDOVÁNÍ TOHOTO TÉMATU UMĚT?

1. Podstata, význam, výhody a nevýhody metod průzkumové analýzy dat
2. Grafické metody (význam, hlavní typy průzkumových grafů včetně jejich konstrukce a interpretace, možnosti použití pro data určitých vlastností)
3. Početní metody a testy (význam, testy normality, testy odlehlých hodnot, testy nezávislosti dat (autokorelace))
4. Praktický výpočet a interpretace grafických metod ve Statistice

OSNOVA

1. Teorie průzkumové analýzy dat
2. Výpočet krabicového grafu v programu Statistica (v Excelu nelze základními grafickými nástroji provést)
3. Výpočet kvantil-kvantilového grafu v programu Statistica (v Excelu nelze základními grafickými nástroji provést)
4. Výpočet histogramu v programu Statistica
5. Vzorové příklady včetně interpretace výsledků
6. Příklady na procvičení

TEORIE PRŮZKUMOVÉ ANALÝZY DAT

Odkaz na literaturu:
http://user.mendelu.cz/drapela/Statisticke_metody/teorie%20text%20II.pdf
Teorie text II, Strany 1- 28
Odkaz na prezentaci:
http://user.mendelu.cz/drapela/Statisticke_metody/Prezentace/zakladni/EDA.ppt

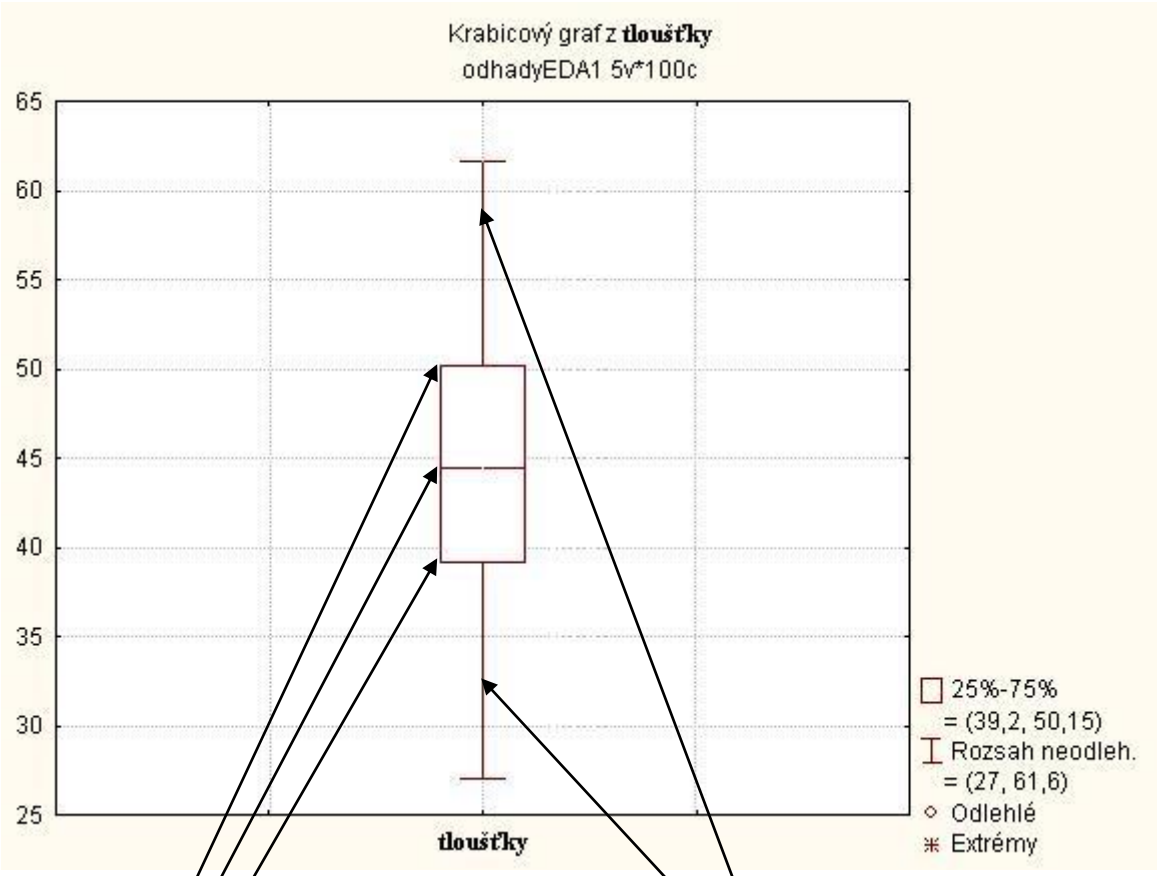
VZOROVÉ PŘÍKLADY

http://user.mendelu.cz/drapela/Statisticke_metody/Data_do_cviceni/Statistica/odhadyEDA1.sta

KRABICOVÝ GRAF

Příklad 1:

Vytvořte krabicový graf pro soubor „tloušťky“ a výsledky interpretujte.

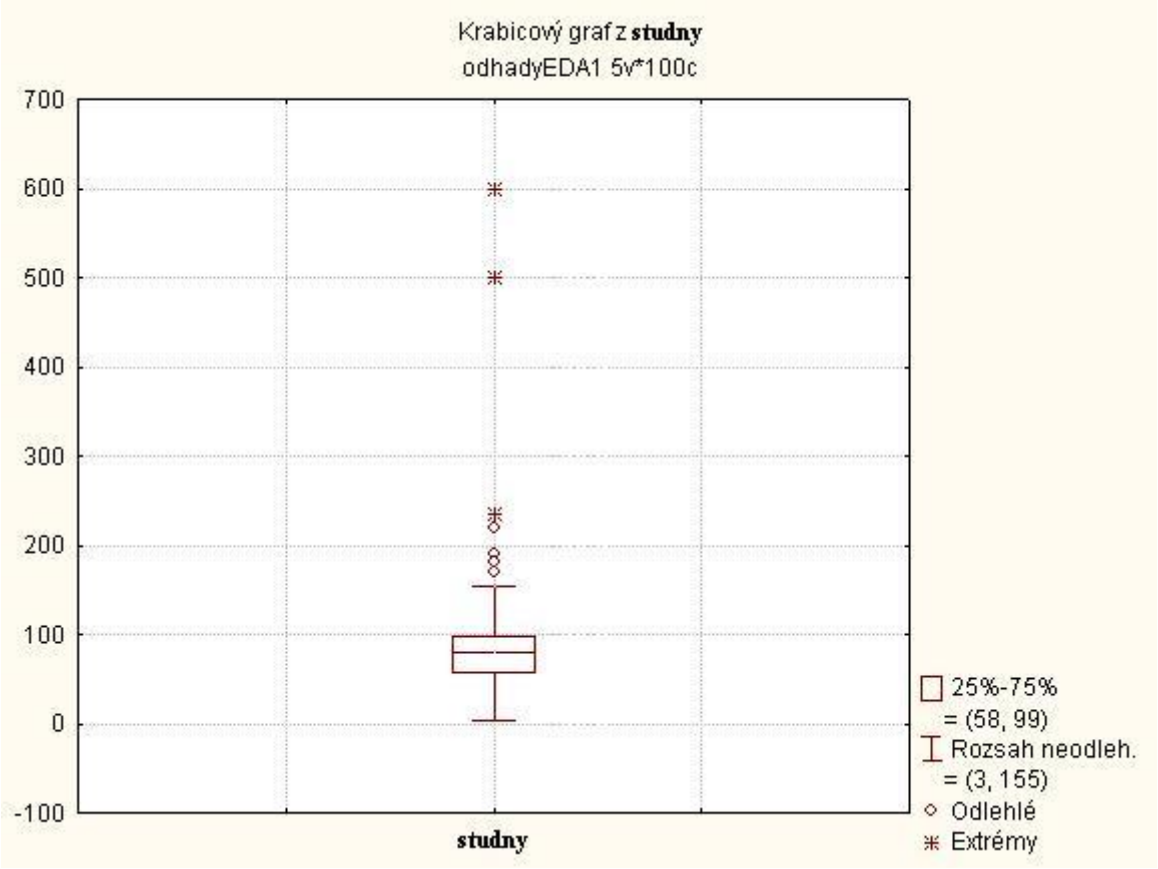


Interpretace:

Hodnota mediánu je přibližně 44, dolního kvartilu 39 a horního kvartilu 50 (přesné hodnoty jsou v legendě). Rozsah nevybočujících hodnot je v rozmezí 27 – 61,6. Ve výběru se nenacházejí žádné odlehlé ani extrémní hodnoty. Medián leží ve středu krabice a zároveň oba vousy jsou stejně dlouhé, což znamená, že rozdělení dat je souměrné. Je nutné si uvědomit, že spodní část grafu představuje levou stranu a horní část grafu pravou stranu – graf je pootočen o 90° doleva vůči grafu v prezentaci.

Příklad 2:

Vytvořte krabicový graf pro soubor „studny“ a výsledky interpretujte.

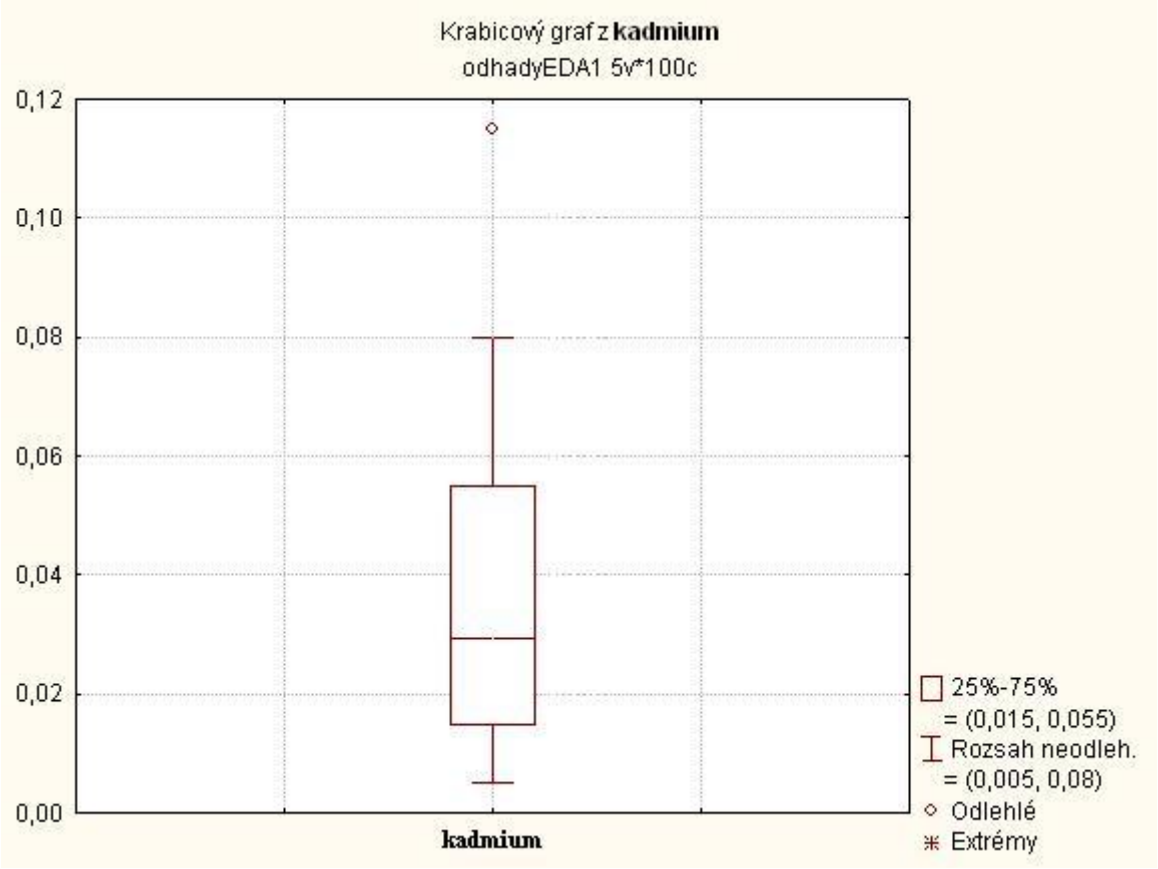


Interpretace:

Hodnota mediánu je přibližně 80, dolního kvartilu 60 a horního kvartilu 100 (přesné hodnoty jsou v legendě). Rozsah nevybočujících hodnot je v rozmezí 3 – 155. Ve výběru se nacházejí čtyři odlehlé (170, 180, 190, 220) a tři extrémní hodnoty (230, 500, 600). Medián leží ve středu krabice a zároveň oba vousy jsou stejně dlouhé, ale je nutné si uvědomit, že krabice (to je 50% hodnot) plus vousy je výrazně na levé straně vzhledem k rozsahu všech 100% hodnot (3 – 600), což znamená, že rozdělení dat je levostranné.

Příklad 3:

Vytvořte krabicový graf pro soubor „kadmium“ a výsledky interpretujte.

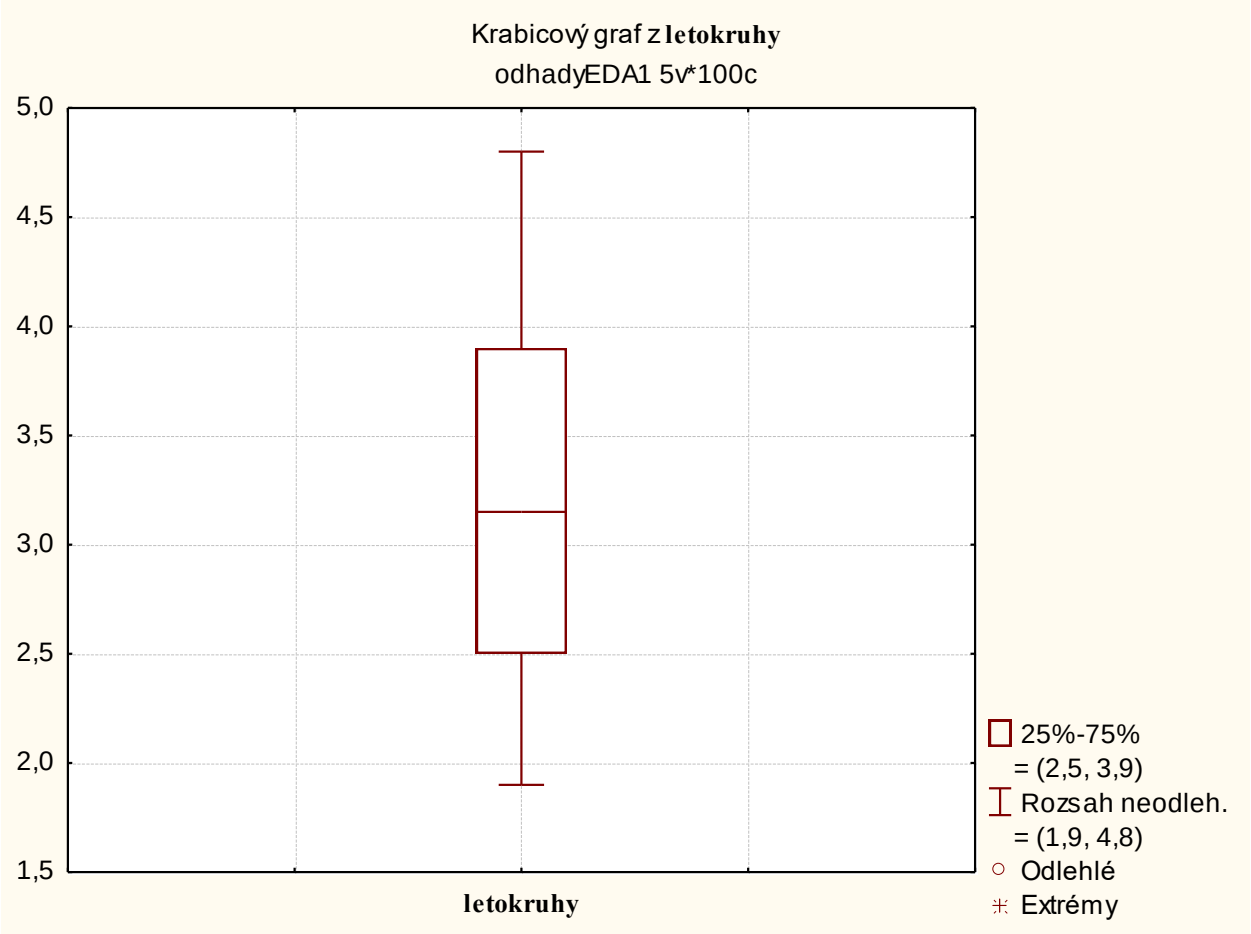


Interpretace:

Hodnota mediánu je přibližně 0,03, dolního kvartilu 0,015 a horního kvartilu 0,055 (přesné hodnoty jsou v legendě). Rozsah nevybočujících hodnot je v rozmezí 0,005 – 0,08. Ve výběru se nachází jedna odlehlá hodnota (0,115) a žádné extrémní hodnoty. Medián leží v levé části krabice a zároveň levý vous je kratší než pravý a i celá krabice s vousy leží v levé části rozsahu všech 100% hodnot, což znamená, že rozdělení dat je levostranné.

Příklad 4:

Vytvořte krabicový graf pro soubor „letokruhy“ a výsledky interpretujte.



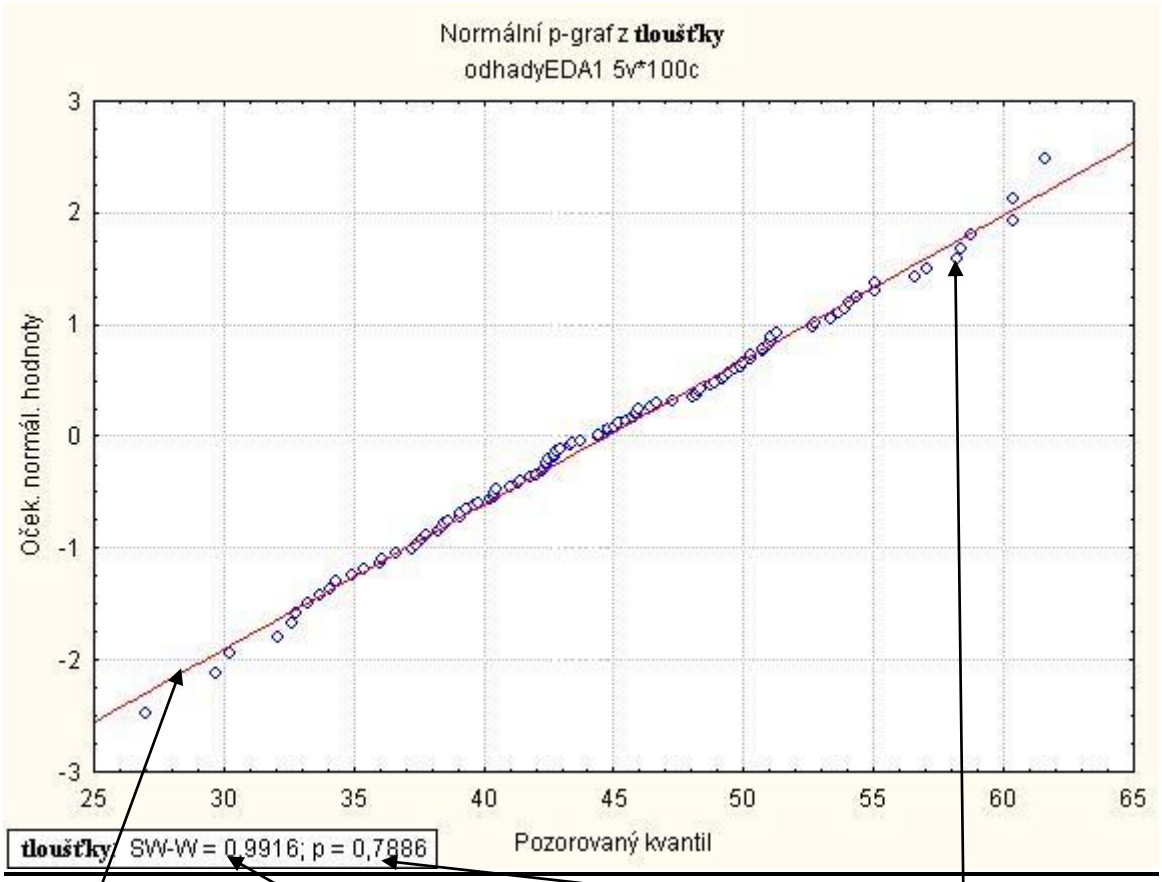
Interpretace:

Hodnota mediánu je přibližně 3,15, dolního kvartilu 2,5 a horního kvartilu 3,9 (přesné hodnoty jsou v legendě). Rozsah nevybočujících hodnot je v rozmezí 1,9 – 4,8. Ve výběru se nenachází žádné odlehlé ani extrémní hodnoty. Medián leží v mírně levé části krabice a zároveň levý vous je nepatrně kratší než pravý což znamená, že rozdělení dat je mírně levostranné.

KVANTIL-KVANTILOVÝ GRAF

Příklad 5:

Vytvořte kvantil-kvantilový graf pro soubor „tloušťky“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.



p hodnota

výsledek Shapiro-Wilksova testu

přímka ideální shody mezi měřeným a modelovým rozdělením

Měřené hodnoty

Interpretace:

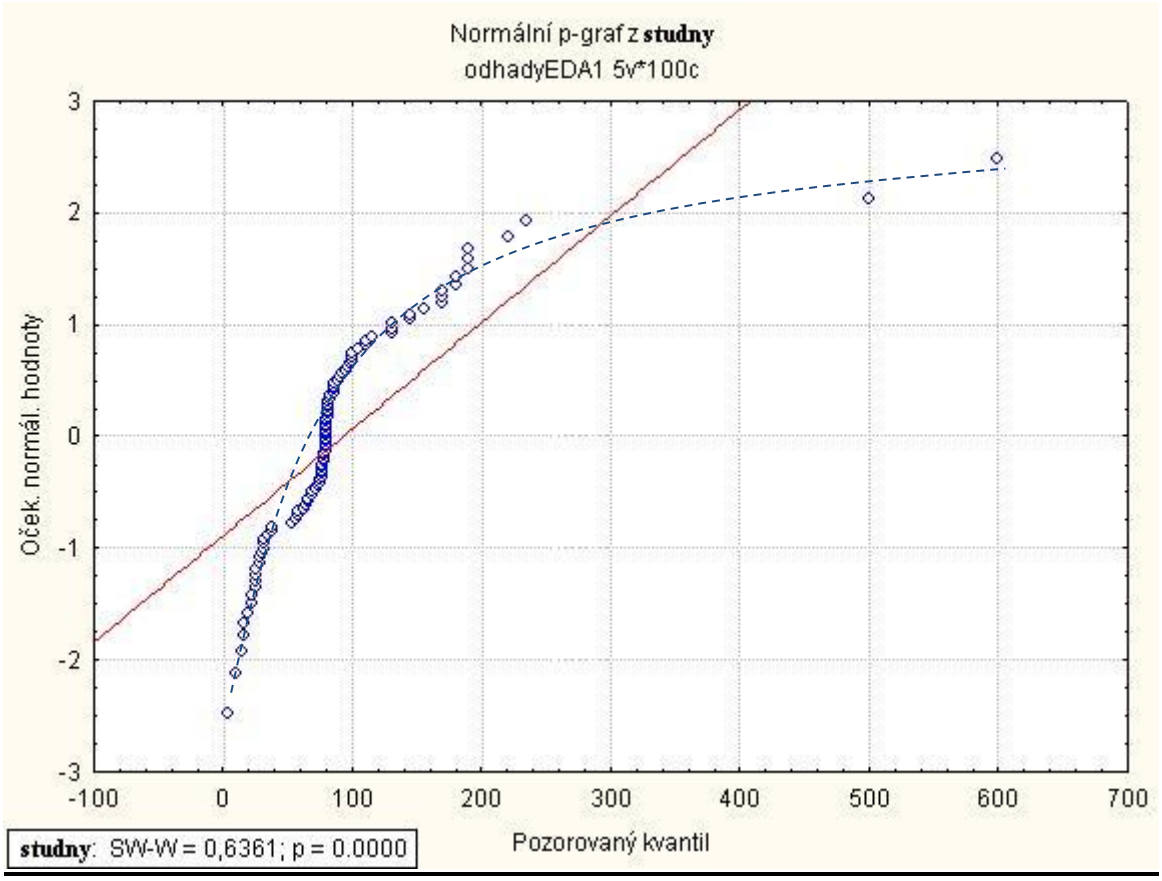
Shapiro-Wilksův test: Jedná se o test shody (normality), ve kterém srovnáváme dané rozdělení dat s rozdělením modelovým (referenčním) – v našem případě se bude vždycky jednat o rozdělení normální. Pro vyhodnocení testu potřebujeme p hodnotu a hladinu významnosti alfa. (teorie p hodnoty a hladiny významnosti alfa je uvedena v tématu 5 – Statistické testy). P hodnota je uvedena ve výsledném grafu a hladina významnosti je vždy nastavena na 0,05, pokud není uvedeno jinak. U každého statistického testu testujeme tzv. nulovou hypotézu (teorie viz téma 5), která v tomto případě zní, že výběr pochází ze základního souboru s normálním rozdělením. Nulová hypotéza obecně není zamítnuta (platí), pokud je p hodnota větší než hladina významnosti alfa. V opačném případě je nulová

hypotéza zamítnuta (neplatí). V tomto případě je p hodnota 0,7886; což je větší než hladina významnosti alfa 0,05, takže nulová hypotéza je nezamítnuta. Předpokládáme tedy, že výběr tedy pochází ze základního souboru s normálním rozdělením.

QQ graf: Červená přímka udává hodnoty kvantilové funkce normálního (modelového) rozdělení. Modrá kolečka jsou reálně naměřené hodnoty. Podle vzájemné polohy měřených a modelových hodnot můžeme posuzovat tvar daného rozdělení – 4 základní diagnostické obrazce jsou v prezentaci. V tomto případě je rozdělení měřených hodnot téměř shodné s hodnotami modelovými, takže můžeme konstatovat, že se jedná o souměrné, normálně zahrocené rozdělení dat. Je důležité si pamatovat, že diagnostické obrazce uvedené v prezentaci platí pouze, pokud jsou měřené hodnoty na ose X a modelové na ose Y!!! Pokud je postavení hodnot na osách zaměněno, otáčí se i interpretace diagnostických obrazců.

Příklad 6:

Vytvořte kvantil-kvantilový graf pro soubor „studny“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.



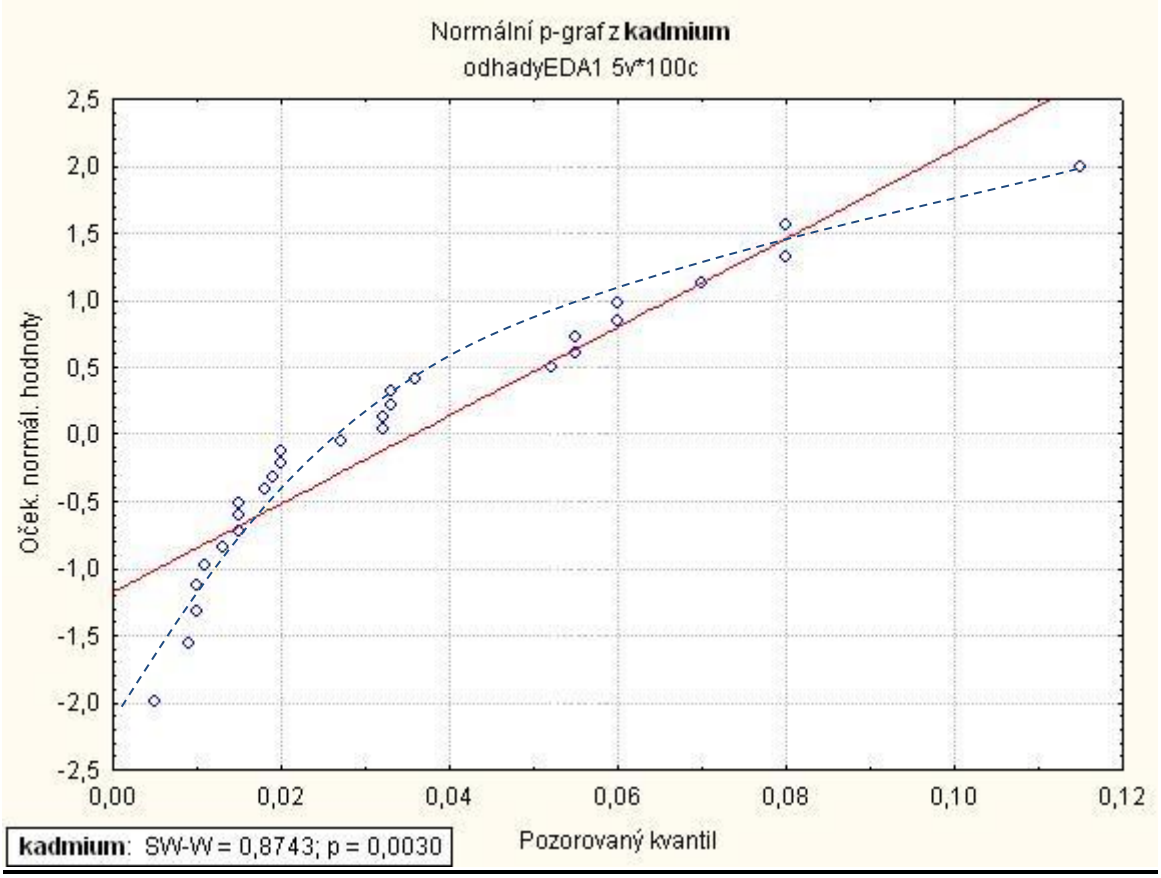
Interpretace:

Shapiro-Wilksův test: V tomto případě je p hodnota 0,0000; což je menší než hladina významnosti alfa 0,05, takže nulová hypotéza je zamítnuta. Výběr tedy nepochází ze základního souboru s normálním rozdělením. Příčiny, proč není rozdělení normální, jsou uvedeny níže.

QQ graf: V tomto případě tvoří měřené hodnoty jakýsi pomyslný oblouk (modrá čárkovaná čára je přidána pro názornost, ale není v grafu běžně znázorňována), který je vypouklý doleva za přímku normálního rozdělení, takže můžeme konstatovat, že se jedná o levostranné, rozdělení dat. Také je v grafu možné vidět dvě hodnoty, které jsou výrazně vzdáleny od ostatních (500, 600). Tyto hodnoty jsou extrémní. Důvody, proč výběr nepochází ze základního souboru s normálním rozdělením, jsou tedy levostranné rozdělení a výskyt extrémních hodnot.

Příklad 7:

Vytvořte kvantil-kvantilový graf pro soubor „kadmium“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.



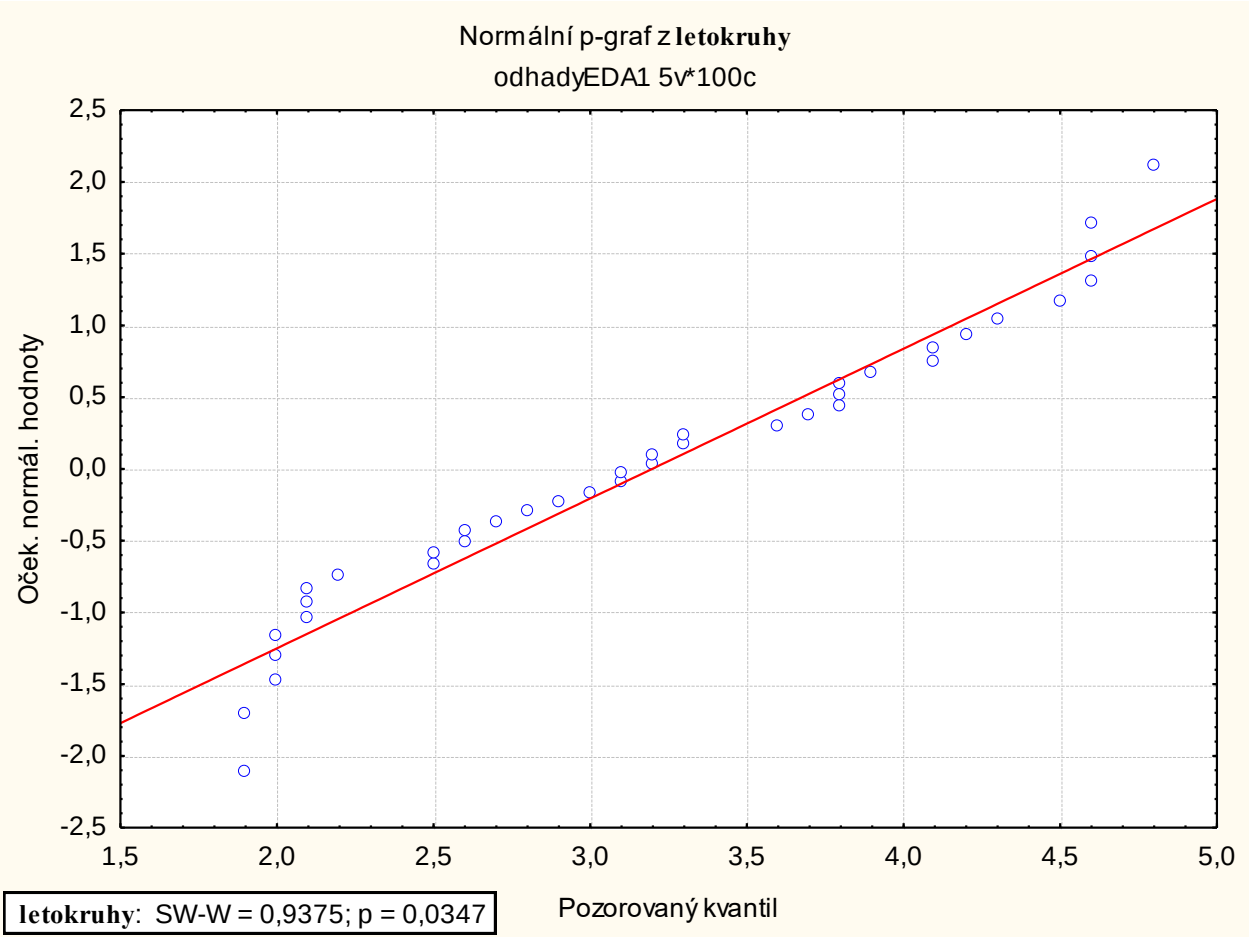
Interpretace:

Shapiro-Wilksův test: V tomto případě je p hodnota 0,003; což je menší než hladina významnosti alfa 0,05, takže nulová hypotéza je zamítnuta. Výběr tedy nepochází ze základního souboru s normálním rozdělením. Příčiny, proč není rozdělení normální, jsou uvedeny níže.

QQ graf: V tomto případě tvoří měřené hodnoty jakýsi pomyslný oblouk (modrá čárkovaná čára je přidána pro názornost, ale není v grafu běžně znázorňována), který je vypouklý doleva za přímku normálního rozdělení, takže můžeme konstatovat, že se jedná o levostranné, rozdělení dat. Také je v grafu možné vidět jednu hodnotu, která je výrazně vzdálena od ostatních (0,115). Tato hodnota je extrémní. Důvody, proč výběr nepochází ze základního souboru s normálním rozdělením, jsou tedy levostranné rozdělení a výskyt extrémní hodnoty.

Příklad 8:

Vytvořte kvantil-kvantilový graf pro soubor „letokruhy“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.



Interpretace:

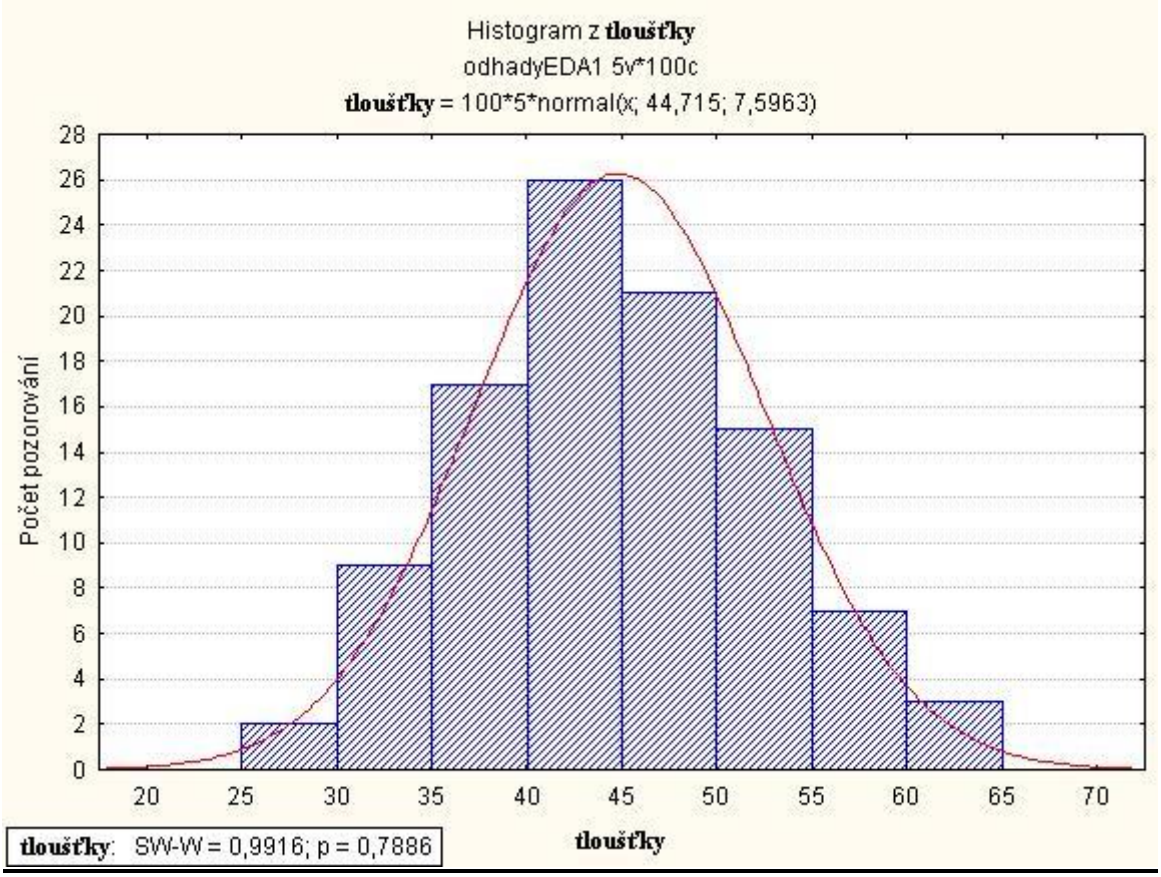
Shapiro-Wilksův test: V tomto případě je p hodnota 0,0347; což je menší než hladina významnosti alfa 0,05, takže nulová hypotéza je zamítnuta. Výběr tedy nepochází ze základního souboru s normálním rozdělením. Příčiny, proč není rozdělení normální, jsou uvedeny níže.

QQ graf: V tomto případě tvoří měřené hodnoty jakési pomyslné obrácené písmeno S, takže můžeme konstatovat, že se jedná o ploché rozdělení dat. Hlavním důvodem, proč výběr nepochází ze základního souboru s normálním rozdělením, je tedy ploché rozdělení, jinak je soubor téměř souměrný. V tomto případě, i když je normalita formálně zamítnuta, můžeme použít momentové odhady (průměr, směrodatnou odchylku apod.), protože hlavním problémem při použití momentových charakteristik je nesouměrnost a extrémní hodnoty.

HISTOGRAM

Příklad 9:

Vytvořte histogram pro soubor „tloušťky“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.

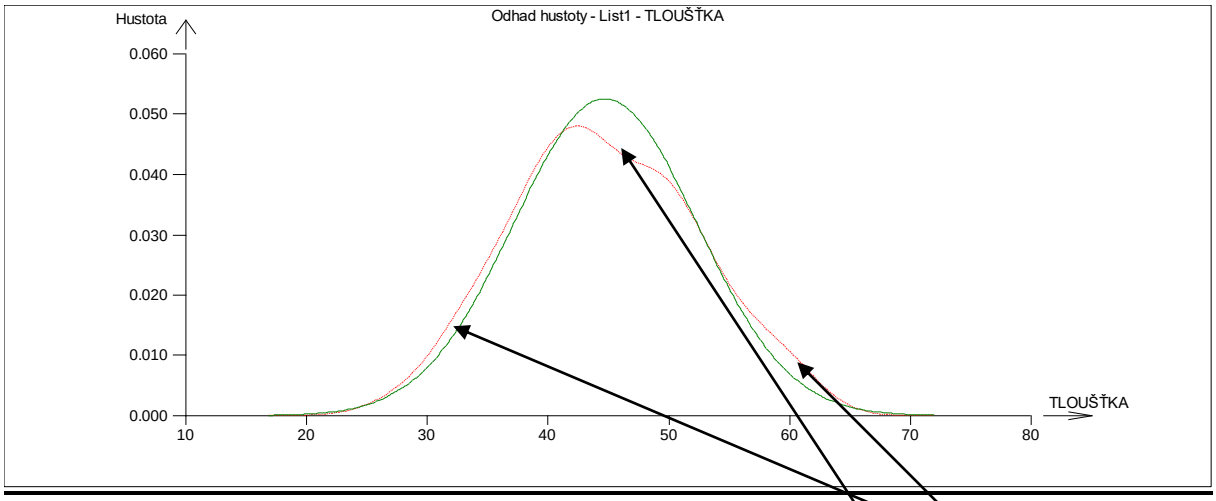


Interpretace:

Shapiro-Wilksův test: viz kvantil kvantilový graf

Histogram: Sloupce v histogramu zobrazují četnosti hodnot v jednotlivých třídách. Pokud bychom tyto sloupce opsali pomyslnou křivkou, dostali bychom jádrový odhad hustoty daného spojitého rozdělení. Červená křivka v histogramu znázorňuje frekvenční funkci modelového (normálního) rozdělení. Také podle vzájemné polohy těchto dvou křivek posuzuje tvar daného rozdělení. V tomto případě můžeme vidět, že modelové a reálné rozdělení jsou si dosti podobné, což znamená, že se jedná o souměrné, normálně zahrocené rozdělení.

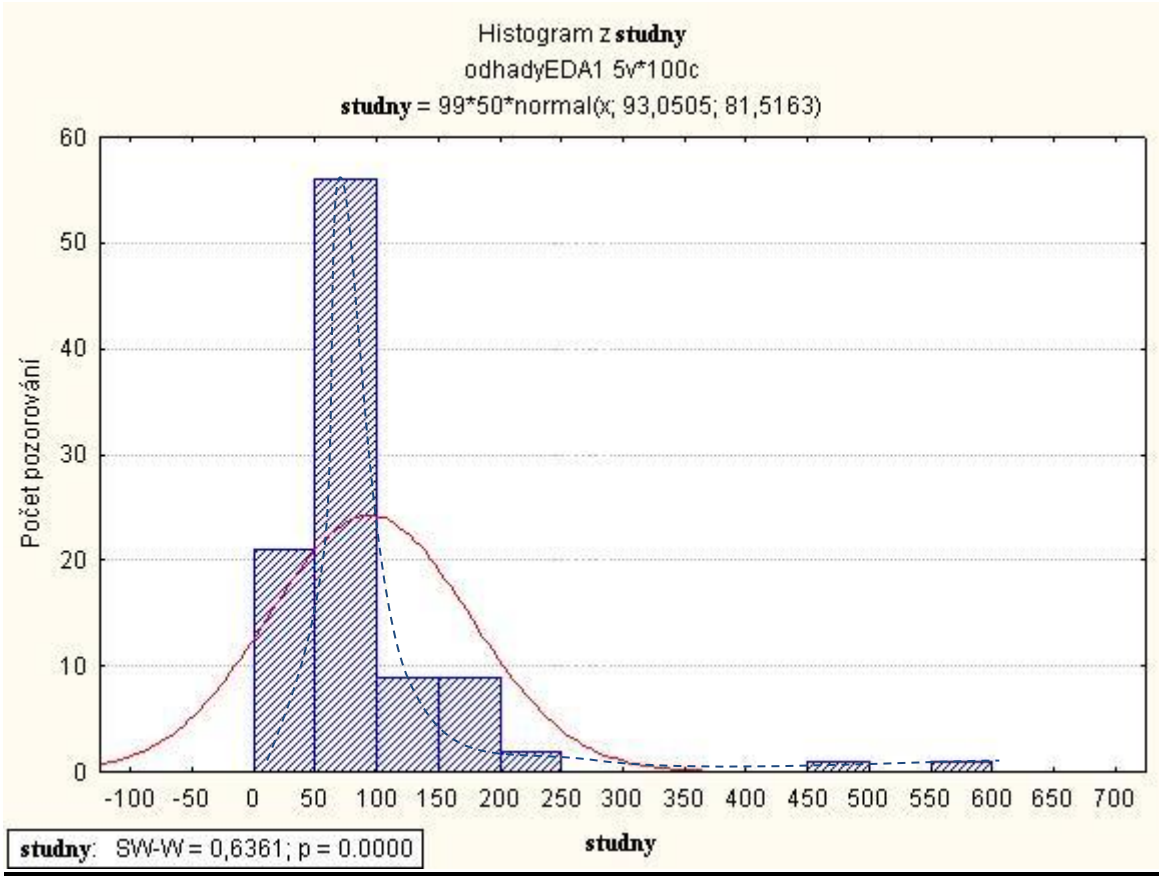
Porovnání tvaru rozdělení měřených hodnot (červená čára) a modelovým normálním rozdělením (zelená čára) pomocí jádrového odhadu hustoty:



Graf potvrzuje výbornou shodu obou rozdělení, v pouze je zde méně „nejčastějších“ hodnot, což způsobuje mírnou „plochost“ rozdělení, naopak u okrajů hodnoty „přebývají“ (plochy pod oběma křivkami – červenou i zelenou musí být stejné, jedná se o pravděpodobnost 1)

Příklad 10:

Vytvořte histogram pro soubor „studny“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.

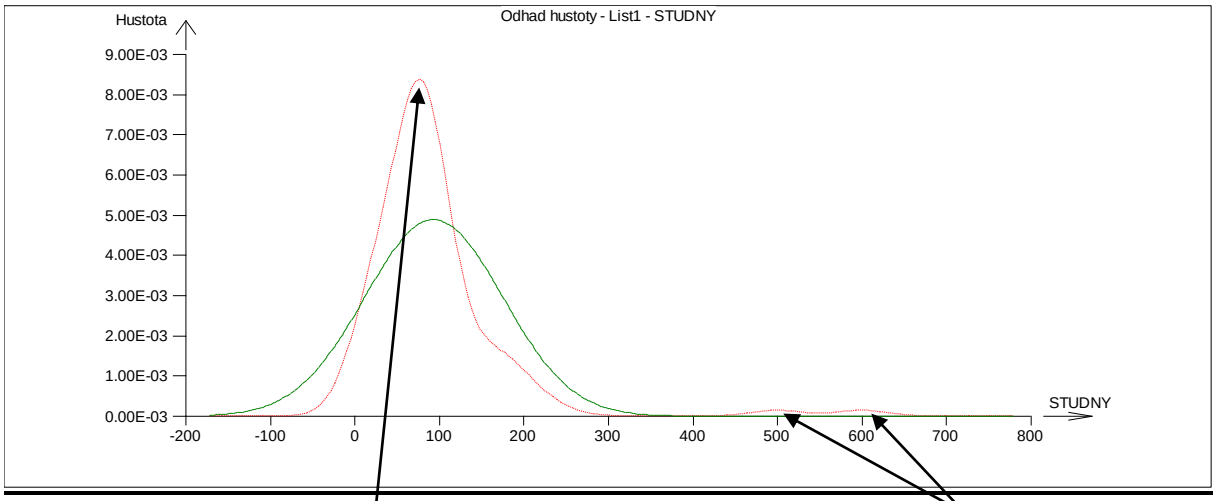


Interpretace:

Shapiro-Wilksův test: viz kvantil kvantilový graf

Histogram: Do tohoto histogramu jsme vložili jádrový odhad hustoty daného rozdělení (není součástí grafického výstupu). Je vidět, že křivka je výrazně vyšší než modelová křivka a většina hodnot je koncentrovaná v levé části funkce, což znamená, že se jedná levostranné, špičaté rozdělení. V tomto histogramu je vidět, že dva sloupce jsou výrazně vzdáleny od ostatních a mají velmi malou četnost, takže se bude jednat o extrémní hodnoty. Důvody, proč výběr nepochází ze základního souboru s normálním rozdělením, jsou tedy levostranné, špičaté rozdělení a výskyt extrémních hodnot.

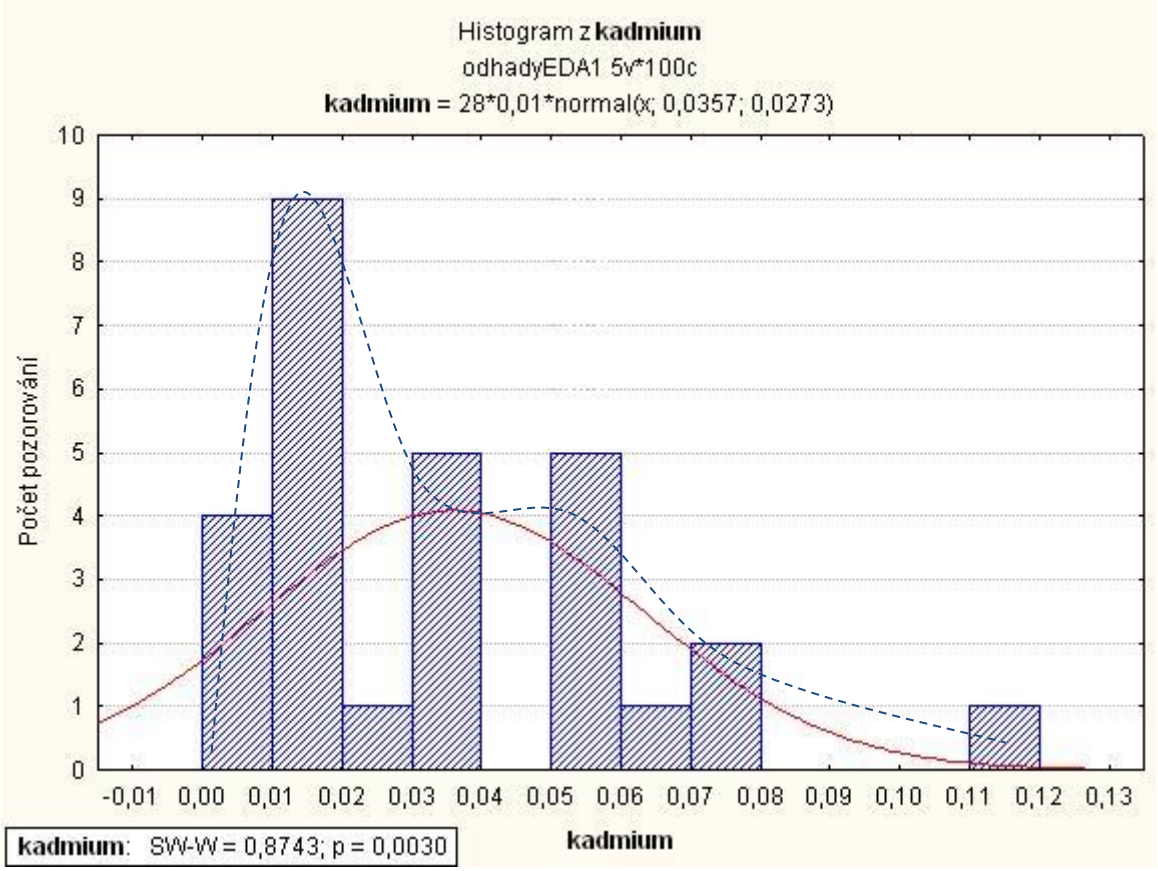
Porovnání tvaru rozdělení měřených hodnot (červená čára) a modelovým normálním rozdělením (zelená čára) pomocí jádrového odhadu hustoty:



Graf potvrzuje výraznou špičatost a levostranné sešikmení dat s výskytem extrémních hodnot.

Příklad 11:

Vytvořte histogram pro soubor „kadmium“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.

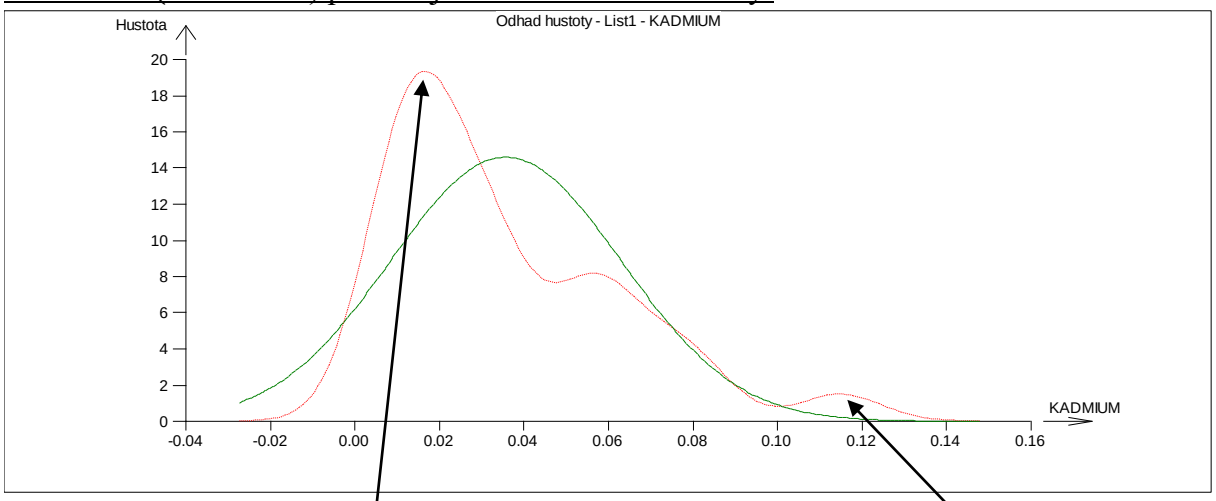


Interpretace:

Shapiro-Wilksův test: viz kvantil kvantilový graf

Histogram: Do tohoto histogramu jsme vložili jádrový odhad hustoty daného rozdělení (není součástí grafického výstupu). Je vidět, že křivka je vyšší než modelová křivka a většina hodnot je koncentrovaná v levé části funkce, což znamená, že se jedná levostranné, špičaté rozdělení. V tomto histogramu je vidět, že jeden sloupec je výrazněni vzdálen od ostatních a má četnost pouze jedna, takže se bude jednat o jednu extrémní hodnotu. Důvody, proč výběr nepochází ze základního souboru s normálním rozdělením, jsou tedy levostranné, špičaté rozdělení a výskyt extrémní hodnoty.

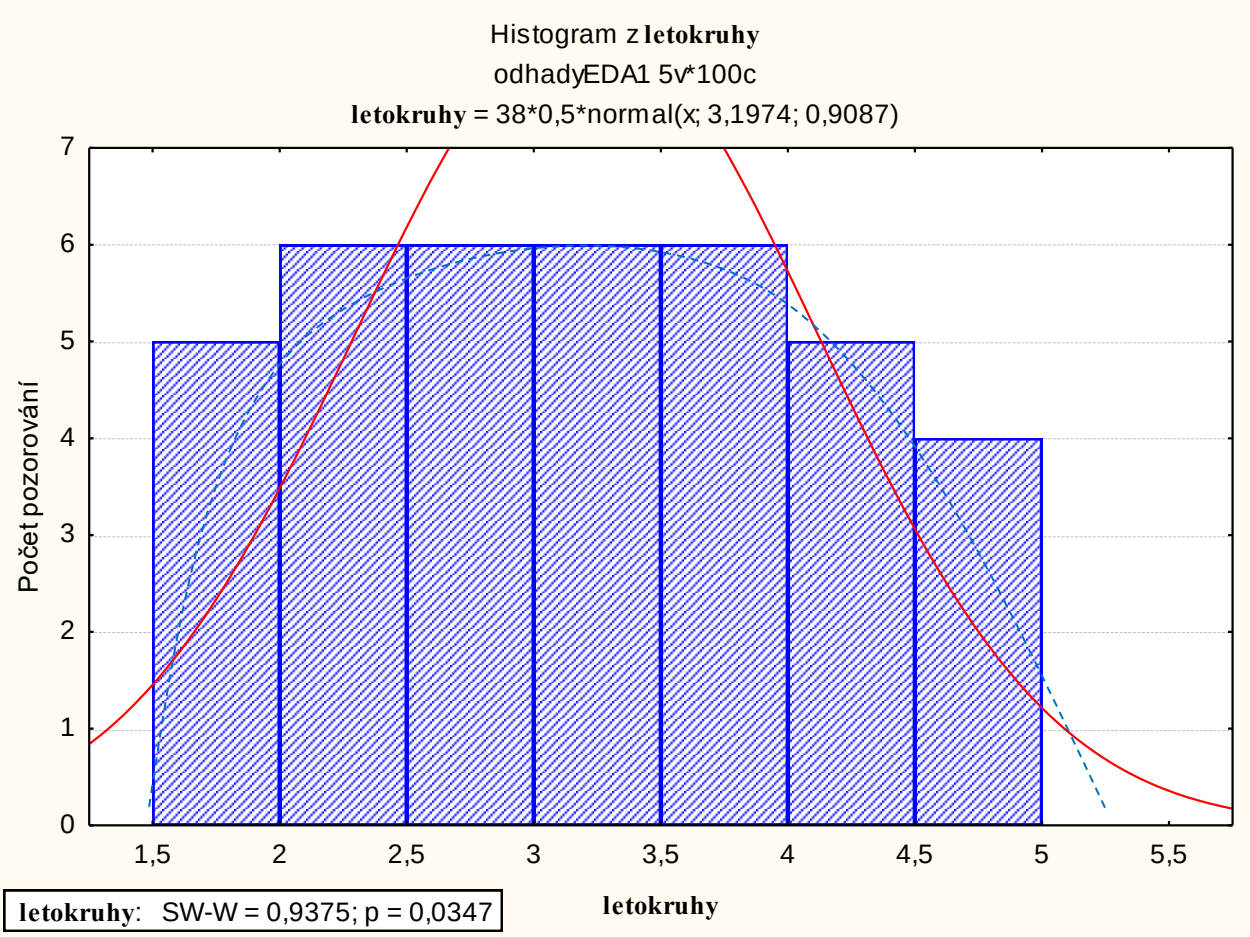
Porovnání tvaru rozdělení měřených hodnot (červená čára) a modelovým normálním rozdělením (zelená čára) pomocí jádrového odhadu hustoty:



Graf potvrzuje výraznou špičatost a levostranné sešikmení dat s výskytem extrémních hodnot.

Příklad 12:

Vytvořte histogram pro soubor „letokruhy“ včetně Shapiro-Wilksova testu normality a výsledky interpretujte.

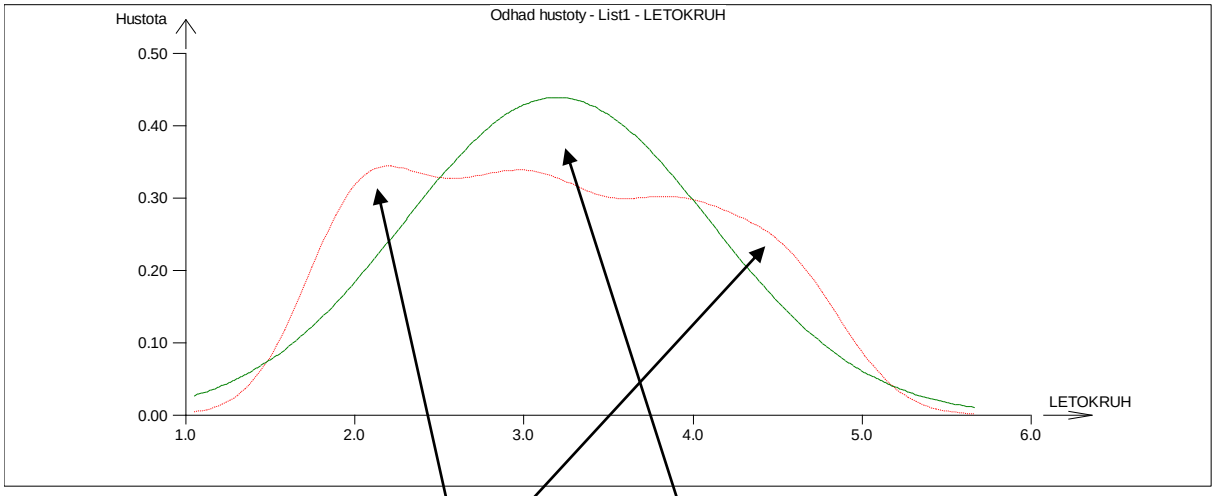


Interpretace:

Shapiro-Wilksův test: viz kvantil kvantilový graf

Histogram: Do tohoto histogramu jsme vložili jádrový odhad hustoty daného rozdělení (není součástí grafického výstupu). Je vidět, že křivka je výrazně nižší než modelová křivka, což znamená, že se jedná o ploché rozdělení. Důvod, proč výběr nepochází ze základního souboru s normálním rozdělením, je tedy levostranné rozdělení.

Porovnání tvaru rozdělení měřených hodnot (červená čára) a modelovým normálním rozdělením (zelená čára) pomocí jádrového odhadu hustoty:



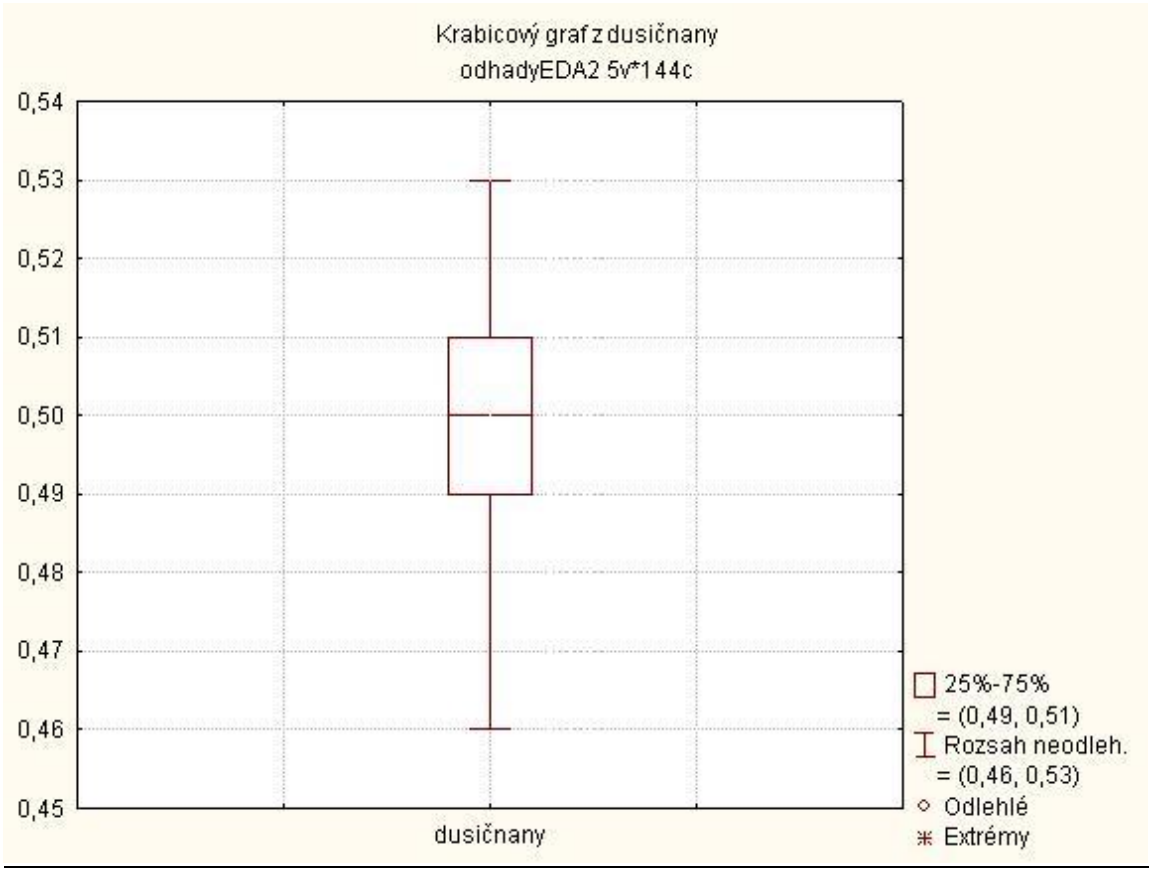
Graf potvrzuje výrazně ploché rozdělení, kde uprostřed (pro teoreticky nejčastější hodnoty) data „chybí“, naopak na okrajích „přebývají“ – měřená data jsou tedy rozložena daleko „rovnoměrněji“ než předpokládá model normálního rozdělení.

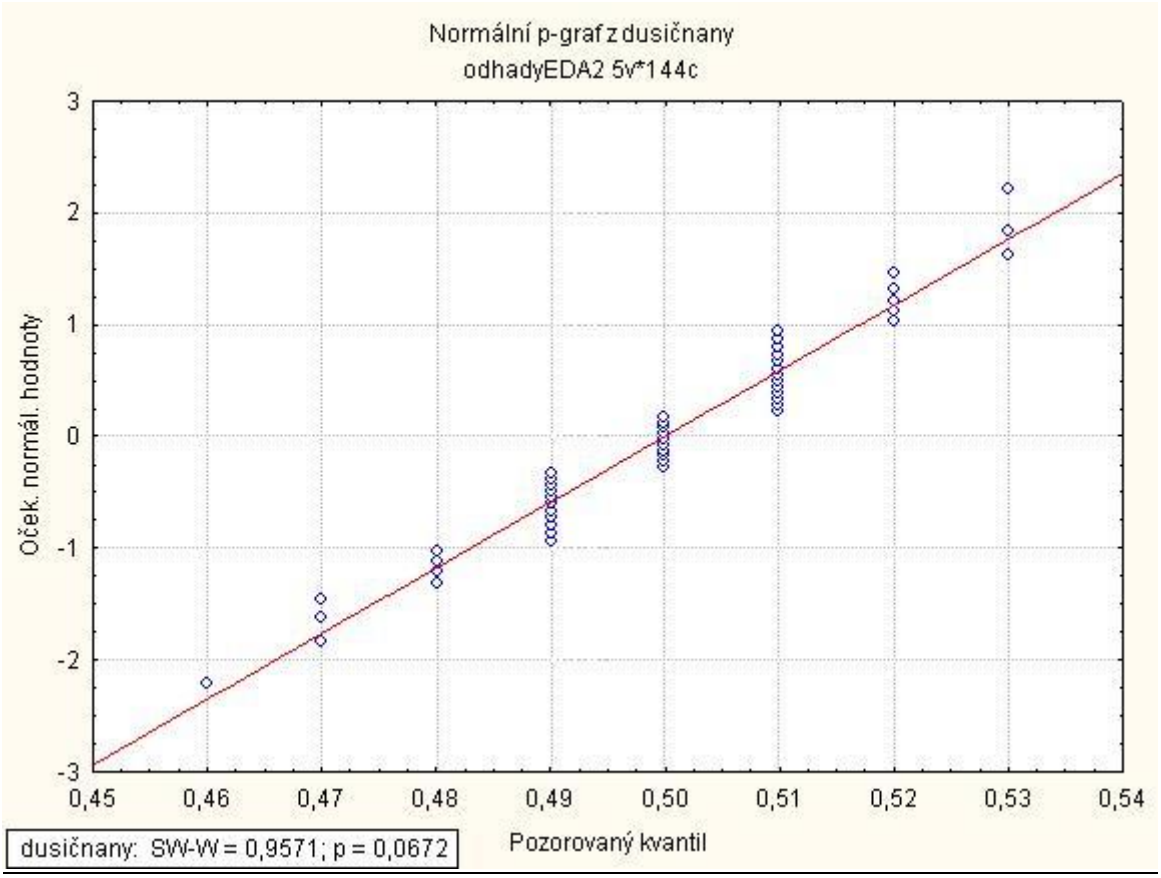
PŘÍKLADY NA PROCVIČENÍ

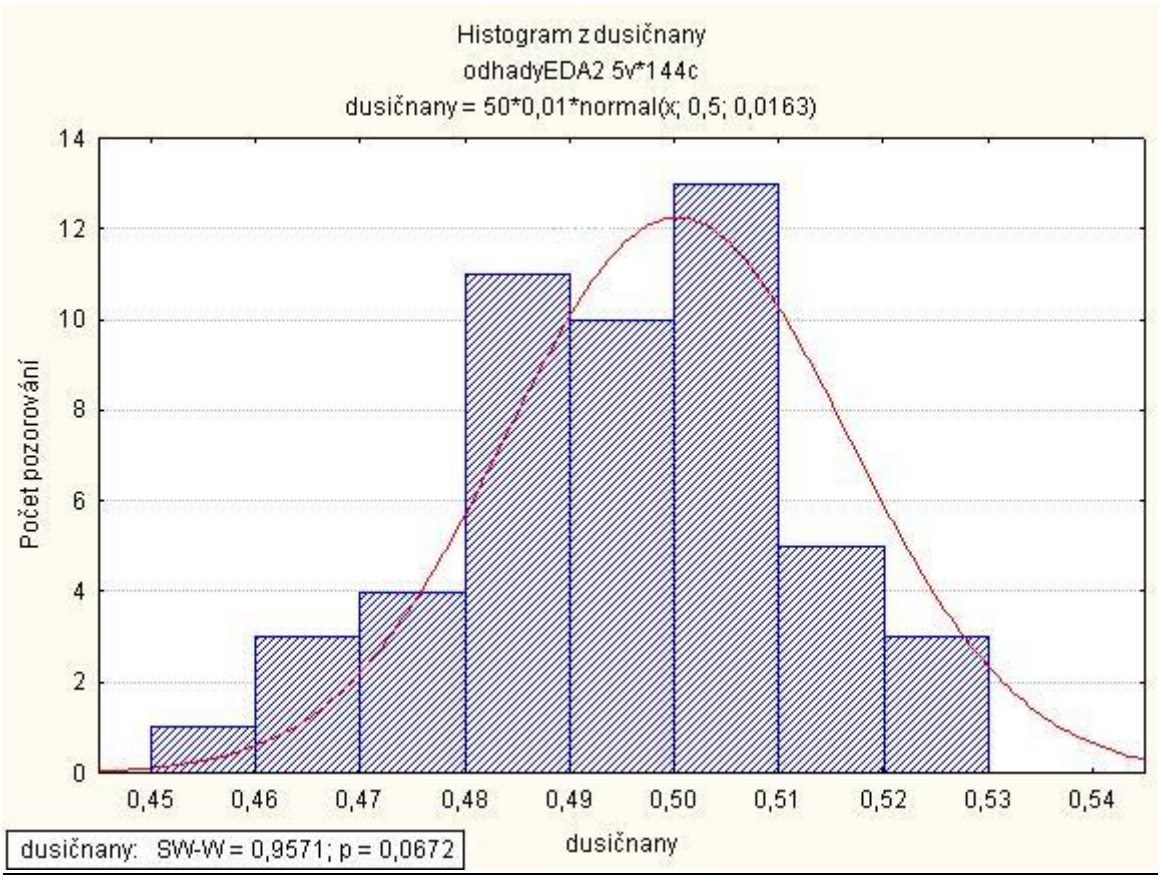
Odkaz na procvičování průzkumové analýzy dat ve Statistice:
http://user.mendelu.cz/drapela/Statisticke_metody/Data_do_cviceni/Statistica/odhadyEDA2.sta

Pro Vaši kontrolu jsou zde uvedeny výstupní grafy všech příkladů a jejich interpretace.

Soubor Dusičnany



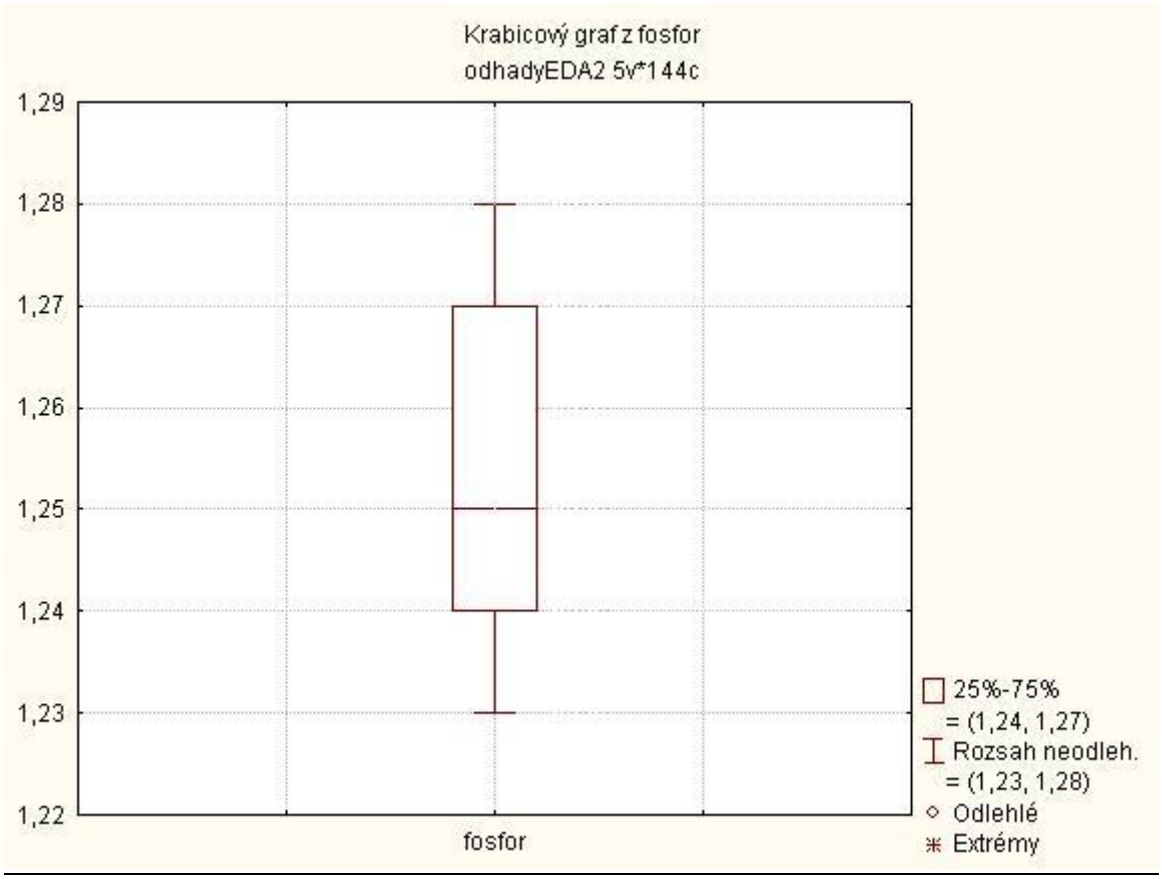

Normální p-graf z dusičnany
odhadyEDA2 5v*144c
Oček. normál. hodnoty
Pozorovaný kvantil
dusičnany: SW-W = 0,9571; p = 0,0672

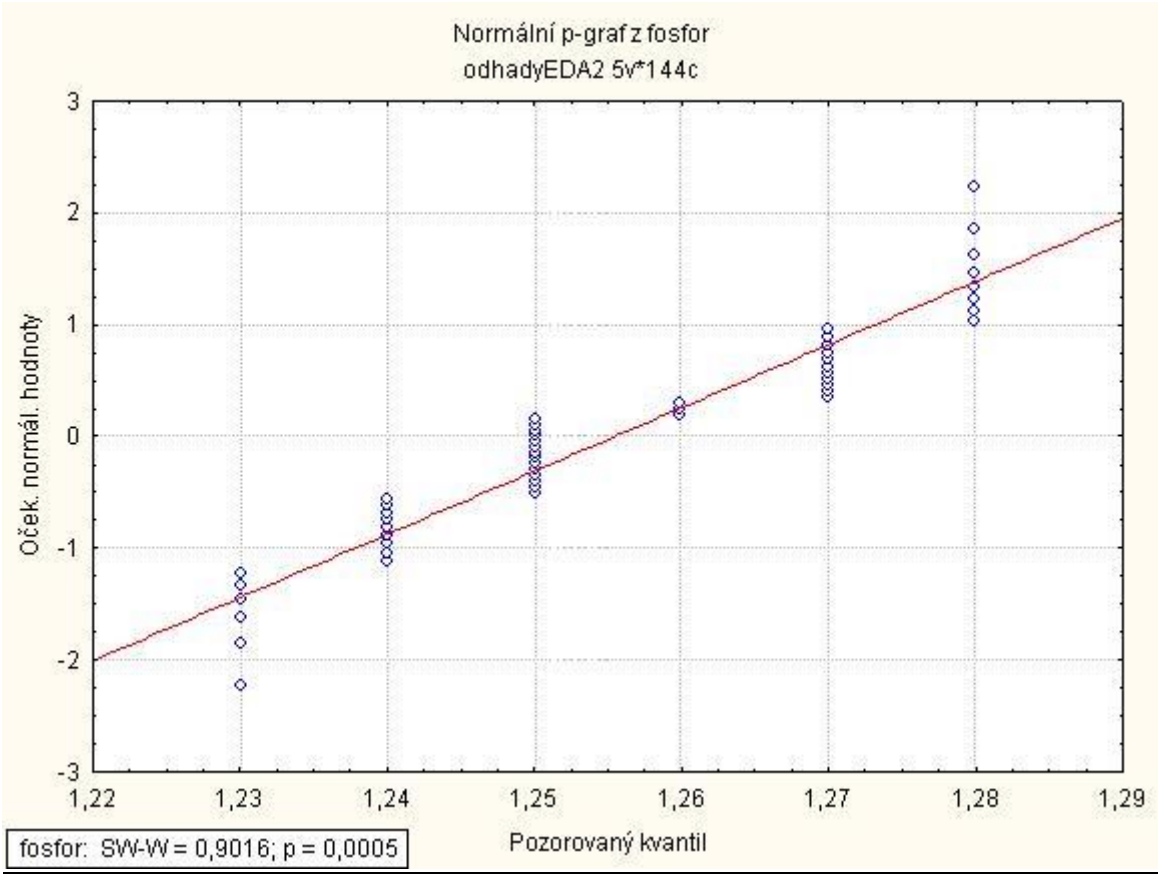

Histogram z dusičnany
odhadyEDA2 5v*144c
dusičnany = 50*0,01*normal(x; 0,5; 0,0163)
Počet pozorování
dusičnany
dusičnany: SW-W = 0,9571; p = 0,0672

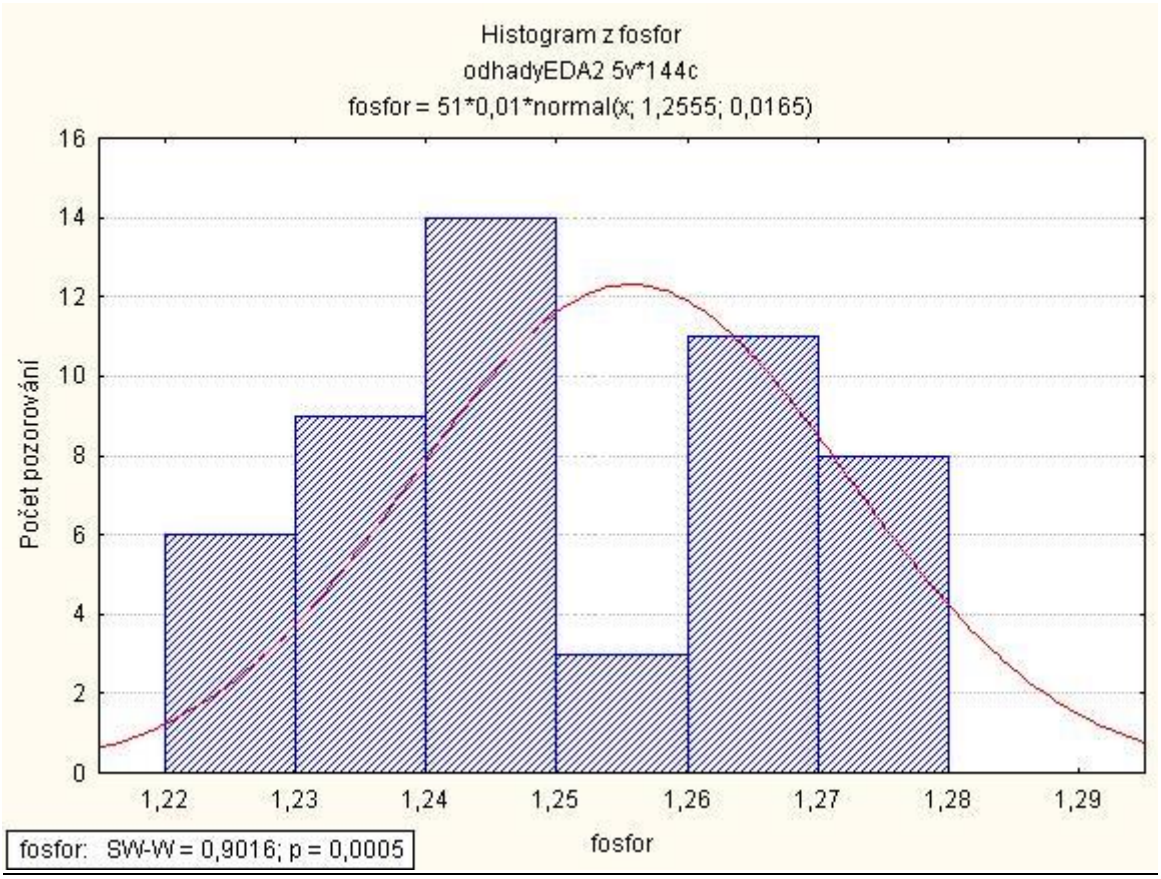
Souhrnná interpretace všech grafů:

Výběr pochází ze základního souboru s normálním rozdělením. Tvar rozdělení je pouze mírně odchýlen do pravostranného a plochého rozdělení, ale tyto odchylky nejsou tak výrazné, aby se nejednalo o normální rozdělení. Ve výběru se nevyskytují žádné odlehlé ani extrémní hodnoty.

Soubor Fosfor



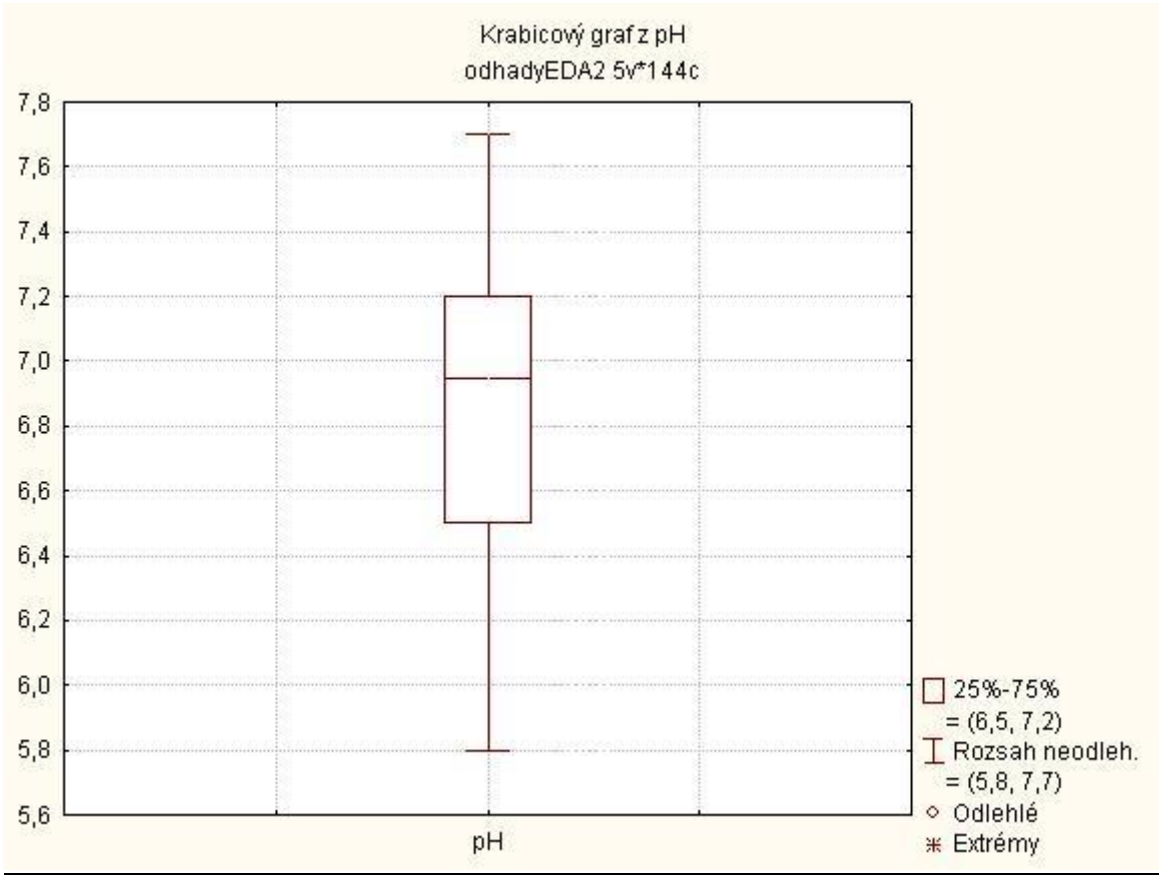

Normální p-graf z fosfor
odhadyEDA2 5v*144c
Oček. normál. hodnoty
Pozorovaný kvantil
fosfor: SW-W = 0,9016; p = 0,0005

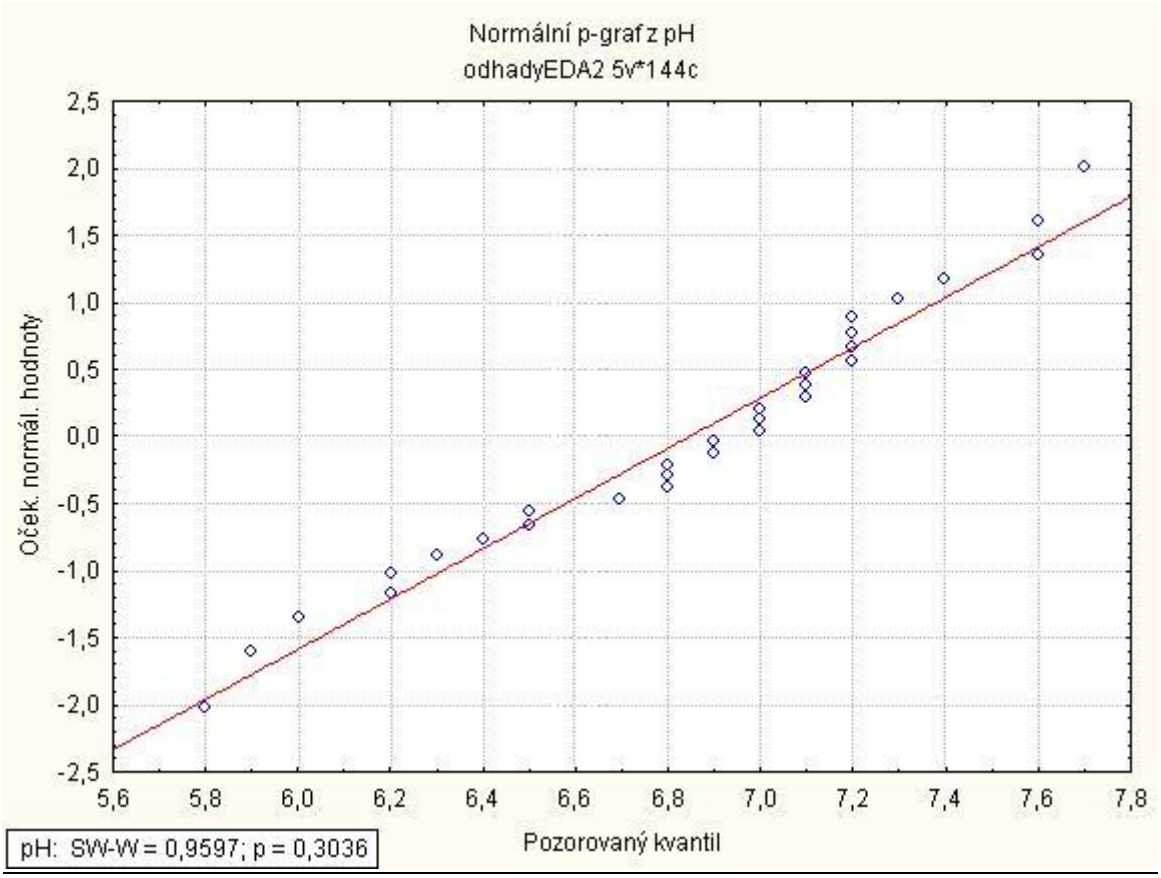

Histogram z fosfor
odhadyEDA2 5v*144c
fosfor = 51*0,01*normal(x; 1,2555; 0,0165)
Počet pozorování
fosfor
fosfor: SW-W = 0,9016; p = 0,0005

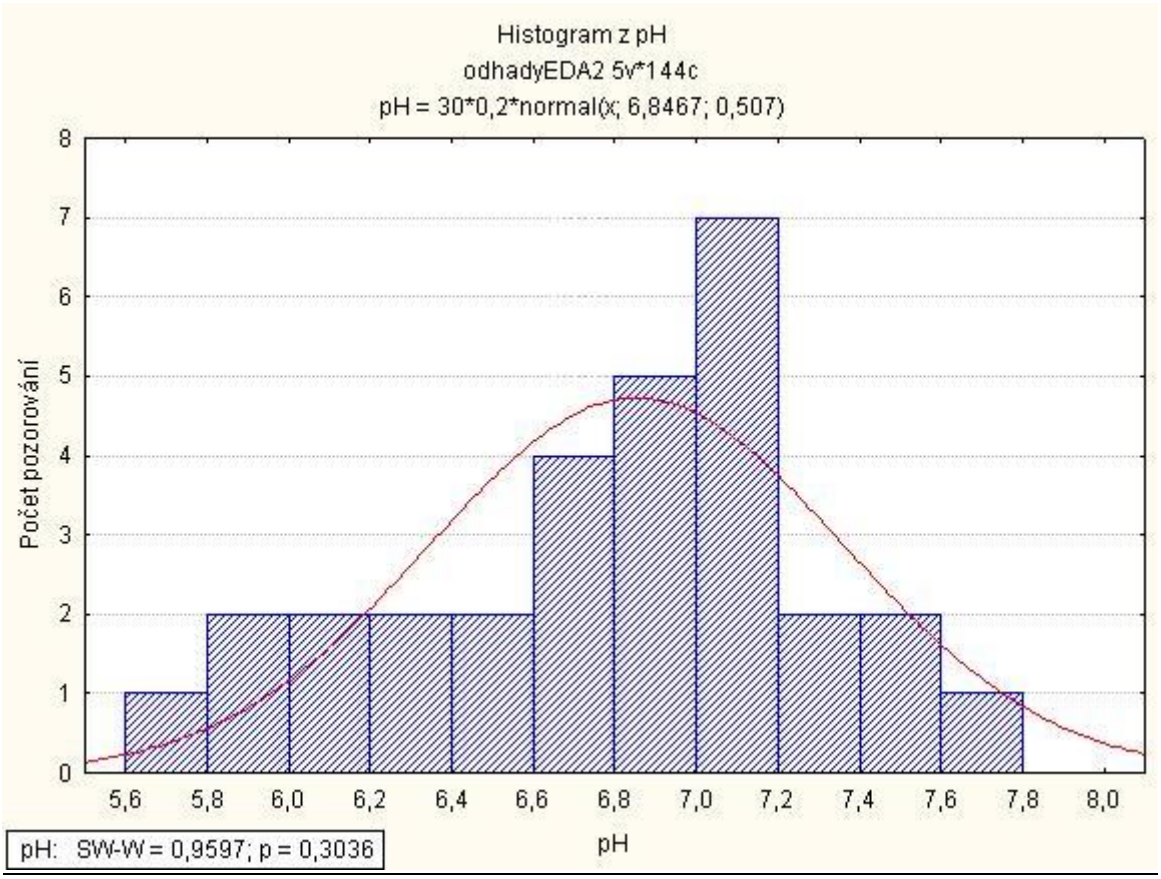
Souhrnná interpretace všech grafů:

Výběr nepochází ze základního souboru s normálním rozdělením. Výběr pochází ze základního souboru, který má ploché a mírně levostranné rozdělení a nevyskytují se v něm žádné odlehlé ani extrémní hodnoty.

Soubor pH



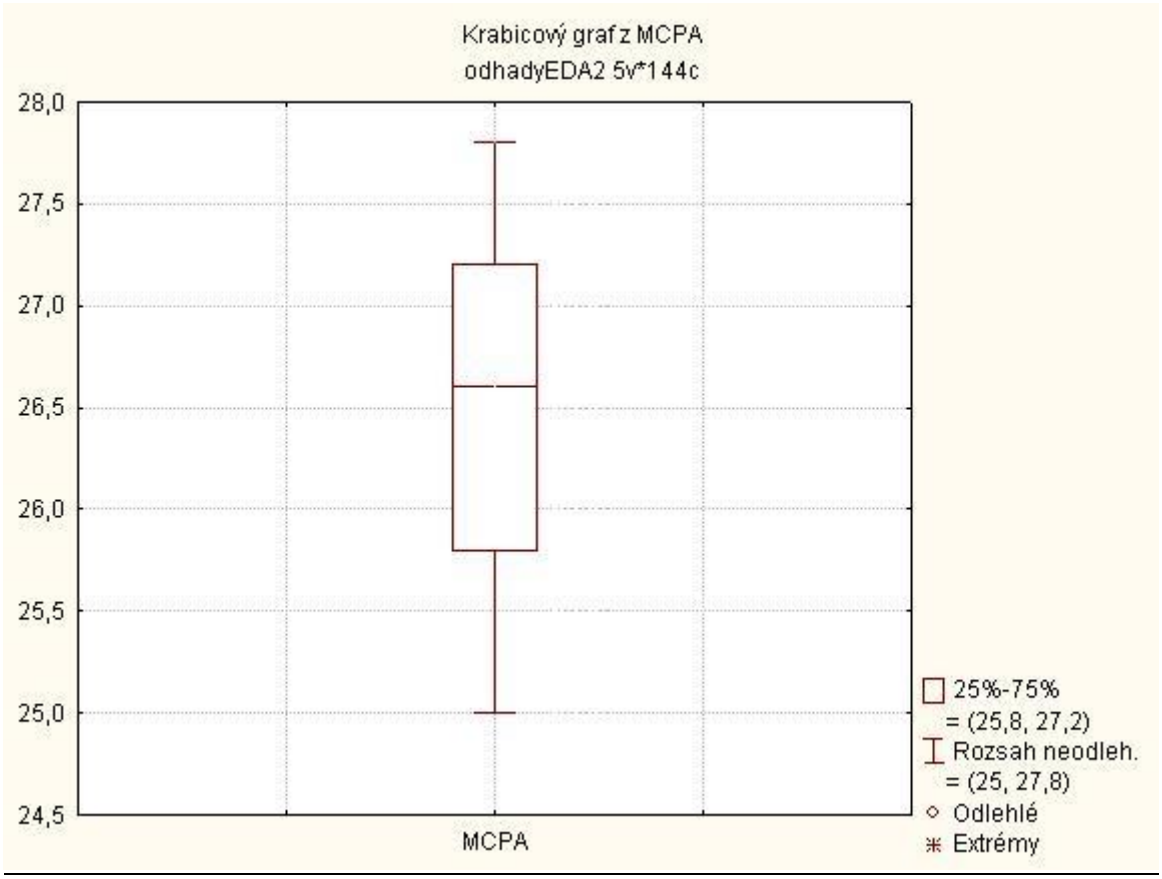

Normální p-graf z pH
odhadyEDA2 5v*144c
Oček. normál. hodnoty
Pozorovaný kvantil
pH: SW-W = 0,9597; p = 0,3036

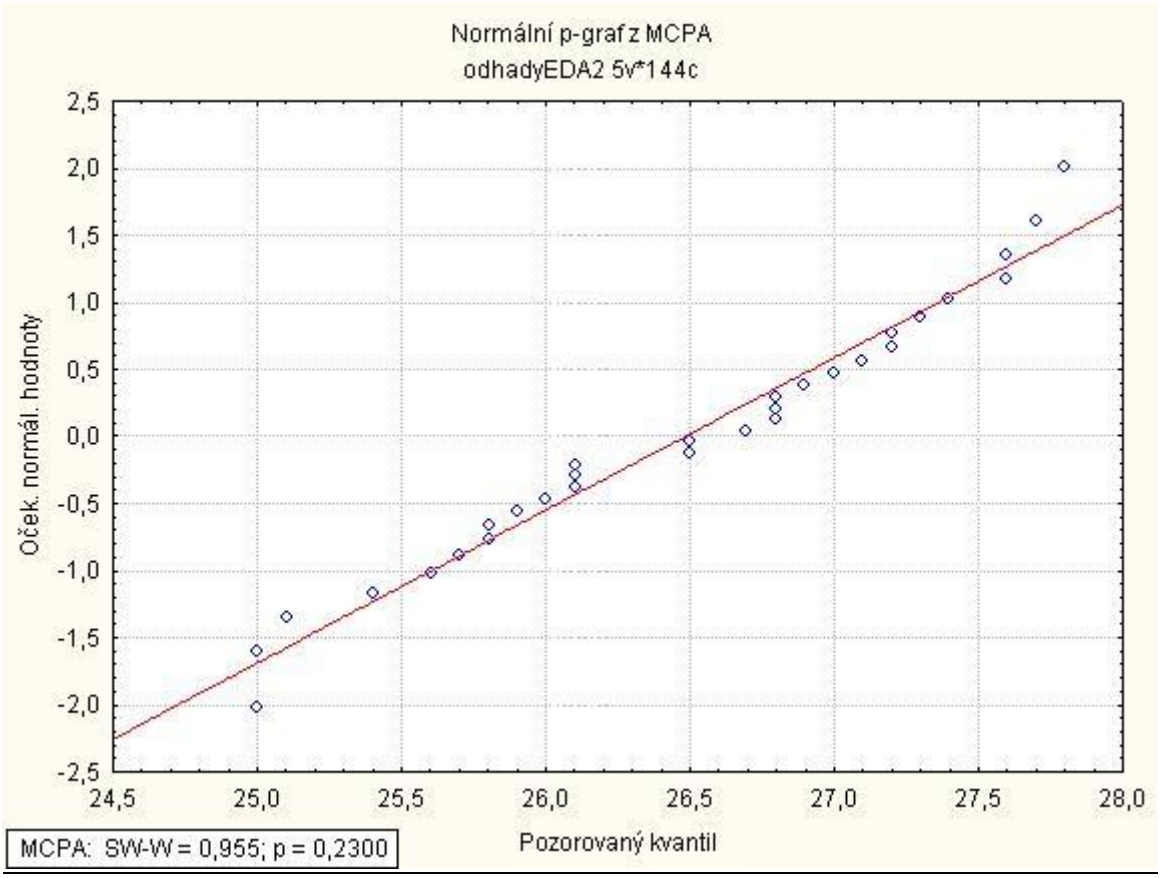

Histogram z pH
odhadyEDA2 5v*144c
pH = 30*0,2*normal(x; 6,8467; 0,507)
Počet pozorování
pH
pH: SW-W = 0,9597; p = 0,3036

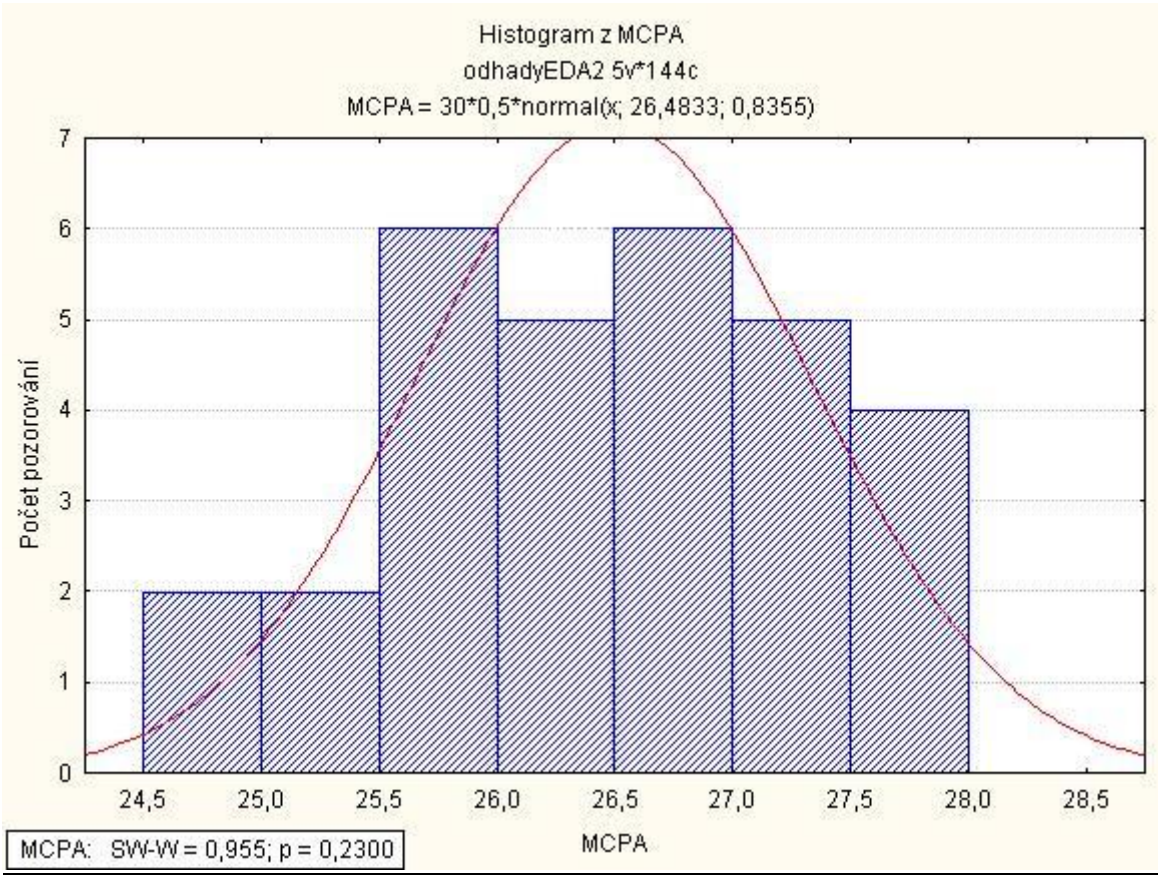
Souhrnná interpretace všech grafů:

Výběr pochází ze základního souboru s normálním rozdělením. Tvar rozdělení je pouze mírně odchýlen do pravostranného a plochého rozdělení, ale tyto odchylky nejsou tak výrazné, aby se nejednalo o normální rozdělení. Ve výběru se nevyskytují žádné odlehlé ani extrémní hodnoty.

Soubor MCPA



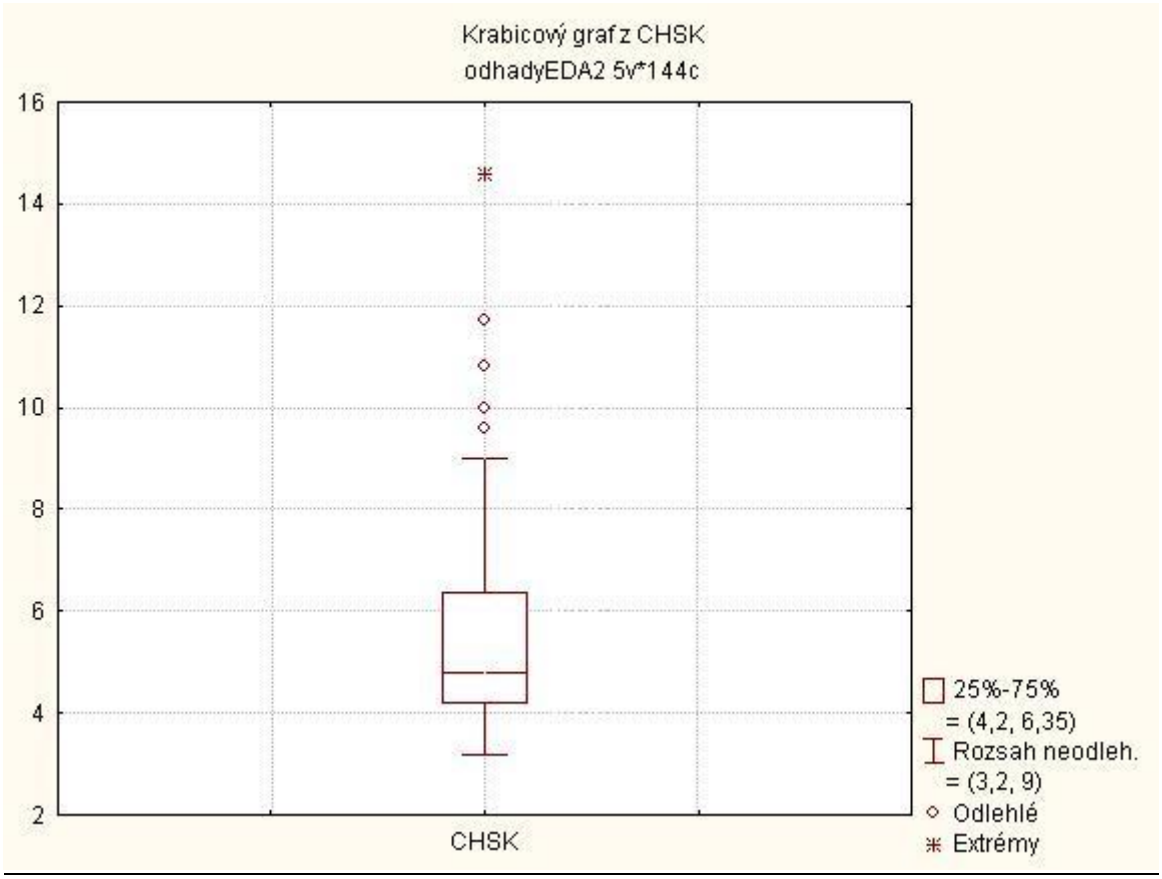

Normální p-graf z MCPA
odhadyEDA2 5v*144c
Oček. normál. hodnoty
Pozorovaný kvantil
MCPA: SW-W = 0,955; p = 0,2300

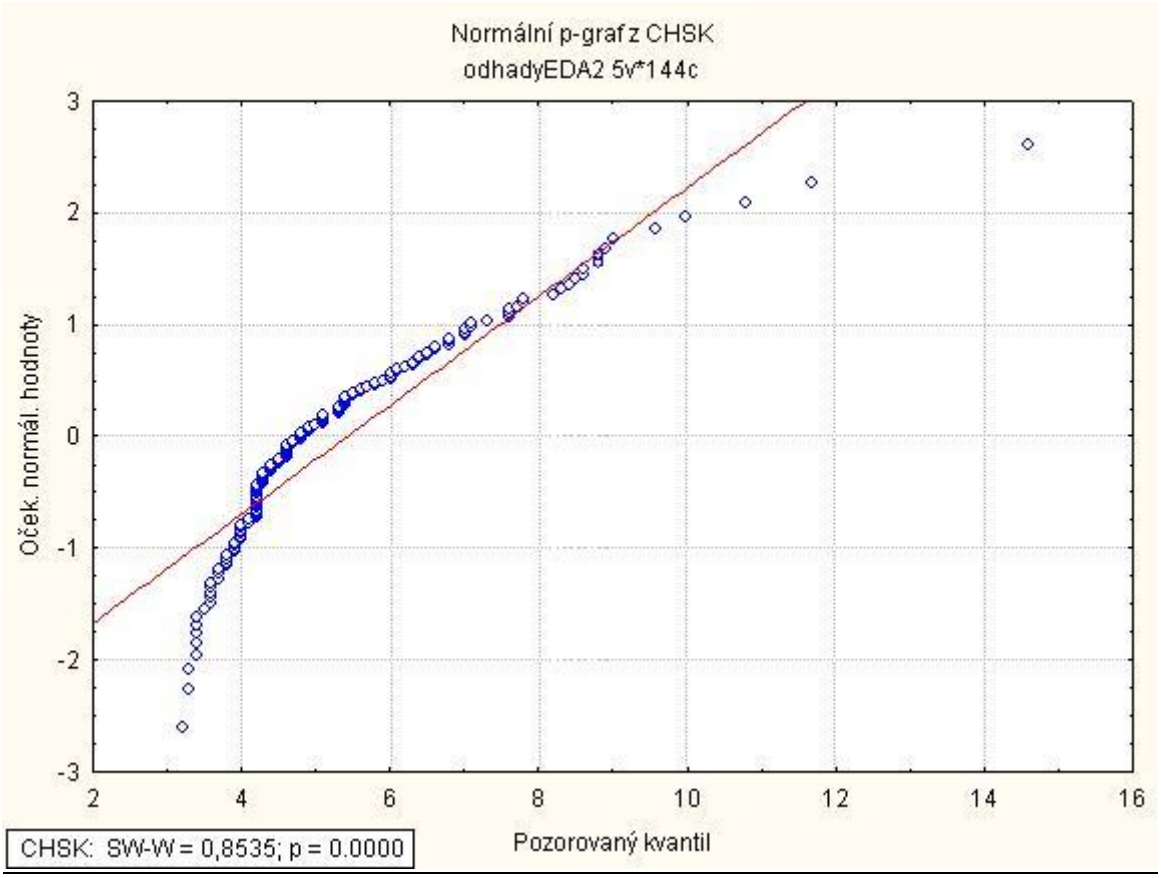

Histogram z MCPA
odhadyEDA2 5v*144c
MCPA = 30*0,5*normal(x; 26,4833; 0,8355)
Počet pozorování
MCPA
MCPA: SW-W = 0,955; p = 0,2300

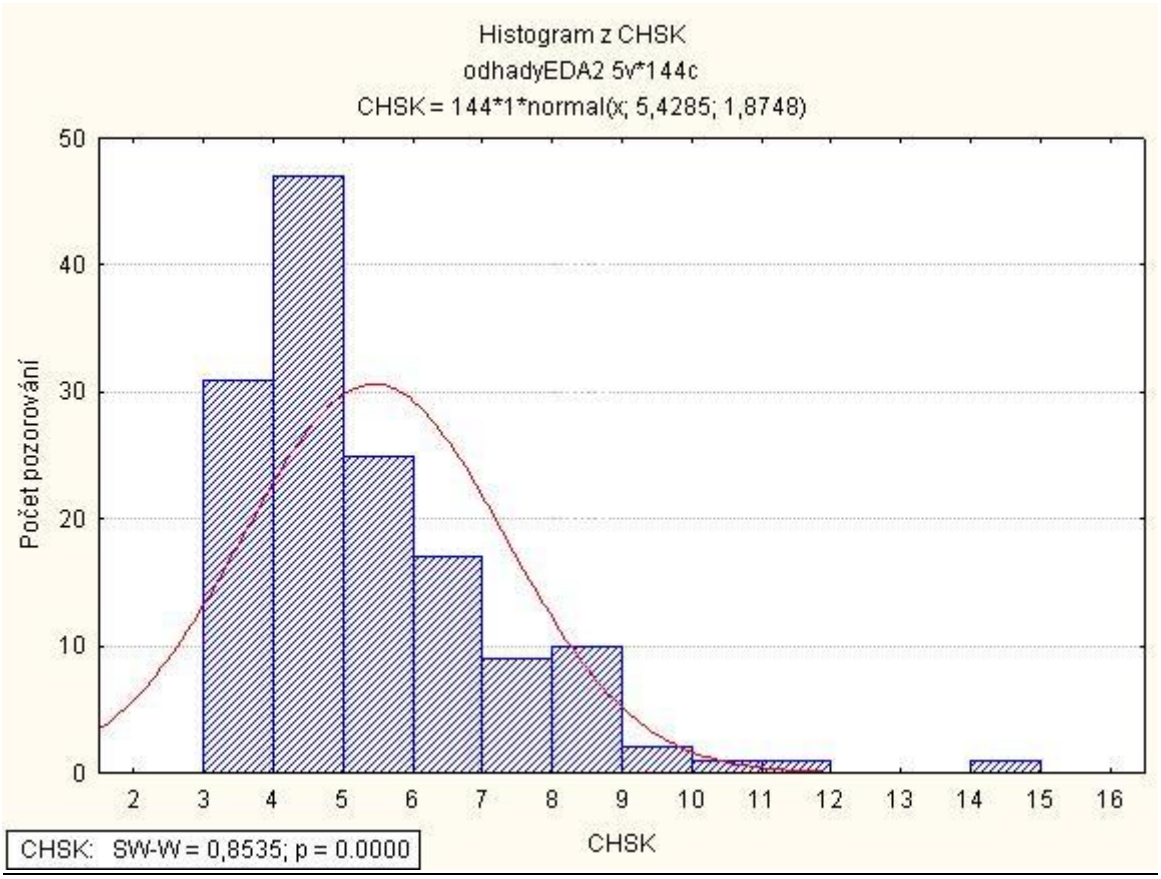
Souhrnná interpretace všech grafů:

Výběr pochází ze základního souboru s normálním rozdělením. Tvar rozdělení je pouze mírně odchýlen do pravostranného a více do plochého rozdělení, ale tyto odchylky nejsou tak výrazné, aby se nejednalo o normální rozdělení. Ve výběru se nevyskytují žádné odlehlé ani extrémní hodnoty.

Soubor CHSK




Normální p-graf z CHSK
odhadyEDA2 5v*144c
Oček. normál. hodnoty
Pozorovaný kvantil
CHSK: SW-W = 0,8535; p = 0.0000


Histogram z CHSK
odhadyEDA2 5v*144c
CHSK = 144*1*normal(x; 5,4285; 1,8748)
Počet pozorování
CHSK
CHSK: SW-W = 0,8535; p = 0.0000

Souhrnná interpretace všech grafů:

Výběr nepochází ze základního souboru s normálním rozdělením. Výběr pochází ze základního souboru, který má výrazně špičaté a výrazně levostranné rozdělení a vyskytují se v něm čtyři odlehlé a jedna extrémní hodnota.