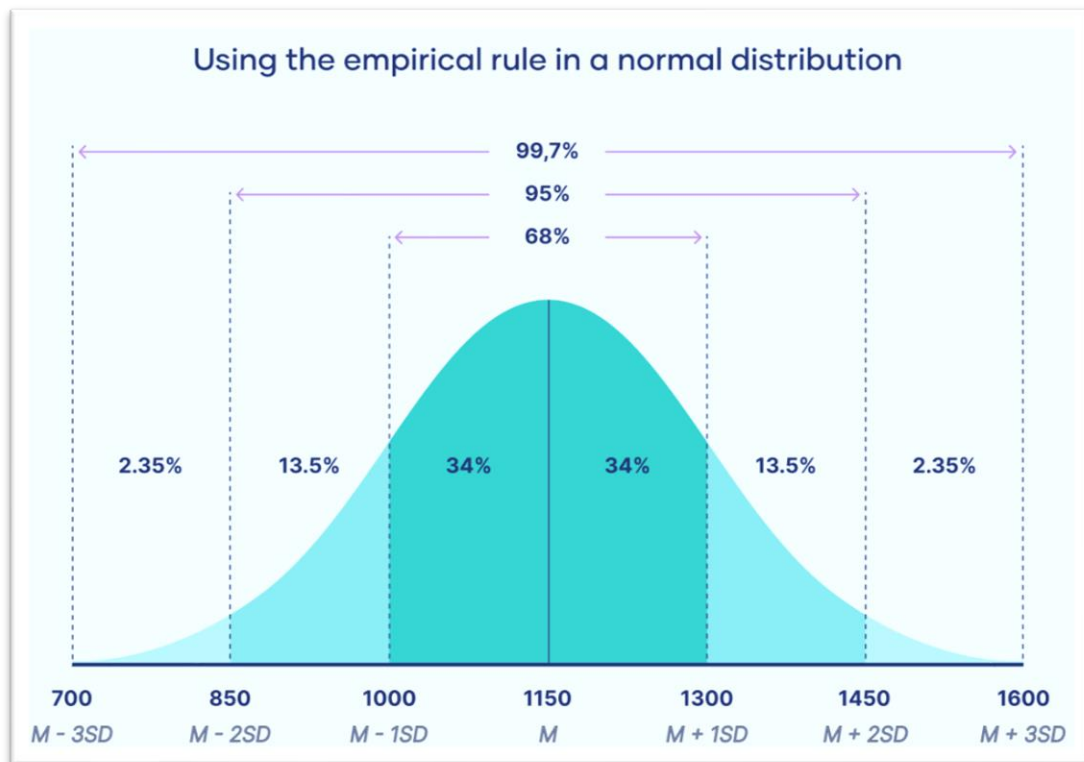


## NORMALITA V STATISTICE

Pojem normalita v statistice znamená, že rozdělení hodnot sledované proměnné v populaci (nebo ve vzorku) se přibližuje k normálnímu rozdělení, také nazývanému Gaussovo rozdělení, což je symetrické rozdělení, které se vyskytuje v mnoha přirozených i umělých procesech a je charakterizováno střední hodnotou a rozptylem.



Normalita ve statistice je důležitá z toho důvodu, že mnoho statistických metod předpokládá, že data mají právě normální rozdělení. Důležitost normality v různých statistických metodách ale může být značně rozdílná a mnoho statistických metod je robustní vůči odchylkám od normálního rozdělení. Existují také alternativní metody pro analýzu dat s „ne-normálním“ rozdělením, jako jsou například neparametrické testy.

## NÁSTROJE PRO OVĚŘOVÁNÍ NORMALITY

Nástroje využívané pro ověření normality lze rozdělit do tří skupin.

1. Popisná statistika
  - a. Koeficient šikmosti
  - b. Koeficient špičatosti
2. Grafické metody
  - a. Histogram
  - b. P-P plot
  - c. Q-Q plot
3. Testy normality
  - a. Kolmogorov- Smirnov
  - b. Shapiro-Wilk

## KOEFIICIENT ŠIKMOSTI (SKEWNESS)

Koeficient šikmosti je míra asymetrie, která ukazuje, zda jsou hodnoty proměnné kulminované na jedné nebo druhé straně od střední hodnoty. Pokud jsou hodnoty proměnné symetricky rozděleny kolem střední hodnoty tak je koeficient šikmosti roven 0. Pokud jsou hodnoty shromážděny více na pravé straně od střední hodnoty pak je koeficient šikmosti kladný. Pokud na levé, pak je koeficient záporný. Obecně může nabývat hodnoty z intervalu  $(-\infty, \infty)$ .

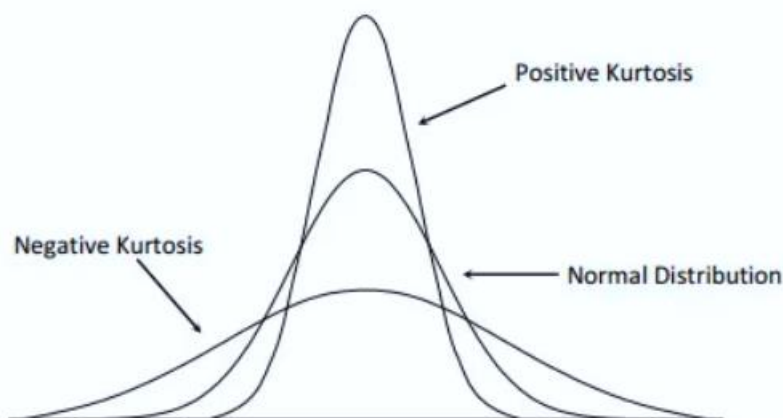


V případě, že bychom chtěli na základě hodnoty tohoto koeficientu tvrdit, že je proměnná normálně rozdělena, měl by se podle některých zdrojů pohybovat v intervalu  $(-1,1)$ . Interpretovat ale jen samotnou hodnotu koeficientu není úplně správný postup. Častěji se doporučuje postup, podle kterého by se měla hodnota koeficientu vydělit jeho směrodatnou chybou a pro tento podíl by mělo platit, že je v intervalu  $(-1,96;1,96)$ .

*Toto pravidlo je vhodné hlavně pro menší datové soubory (méně než 100 pozorování). Pro větší datové soubory není popsán postup vhodný a je lepší normalitu posuzovat například na základě histogramu.*

## KOEFICIENT ŠPIČATOSTI (KURTOSIS)

Koeficient špičatosti měří, jak moc jsou hodnoty proměnné koncentrovány v blízkosti průměru. V případě kladné hodnoty koeficientu špičatosti je vrchol rozdělení výraznější (vyšší) než u normálního rozdělení, to znamená, že hodnoty proměnné jsou koncentrovány blízko průměru a mají menší rozptyl v porovnání s normálním rozdělením.



V případě záporné hodnoty koeficientu šikmosti jsou hodnoty proměnné více rozptýleny a méně koncentrovány v blízkosti průměru v porovnání s normálním rozdělením.

Příkladem proměnné s negativní šikmostí může být například rozdělení příjmů v zemi, kde většina populace má nízké příjmy a pouze menší část populace má vysoké příjmy. To způsobuje, že rozdělení má dlouhý "ocas" směrem k nižším hodnotám a kratší "ocas" na straně vyšších hodnot.

Existuje několik různých kritérií pro to, jaký koeficient šikmosti lze považovat za stále relativně blízký normálnímu rozdělení. Jedním takovým kritériem je, že koeficient šikmosti by neměl překročit hodnotu 2. Tento limit je však spíše orientační a závisí na konkrétní aplikaci a kontextu.

Stejně jako u koeficientu šikmosti, tak i v tomto případě by se měla hodnota koeficientu špičatosti nejprve vydělit jeho směrodatnou chybou a pro tu by mělo platit, že by se měla pohybovat v intervalu  $(-1,96; 1,96)$ . Rovněž ale platí, že toto pravidlo není vhodné používat při větším počtu pozorování.

**Hodnoty zmíněných koeficientů (i s jejich standardními chybami) lze spočítat v záložce Analyze-Descriptive Statistics-Explore. Do okna Dependent List stačí zadat proměnné, pro které chceme tyto koeficienty spočítat a potvrdit tlačítkem OK. Oba koeficienty najdeme v tabulce Descriptive Statistics pod názvy Skewness a Kurtosis.**

## HISTOGRAM

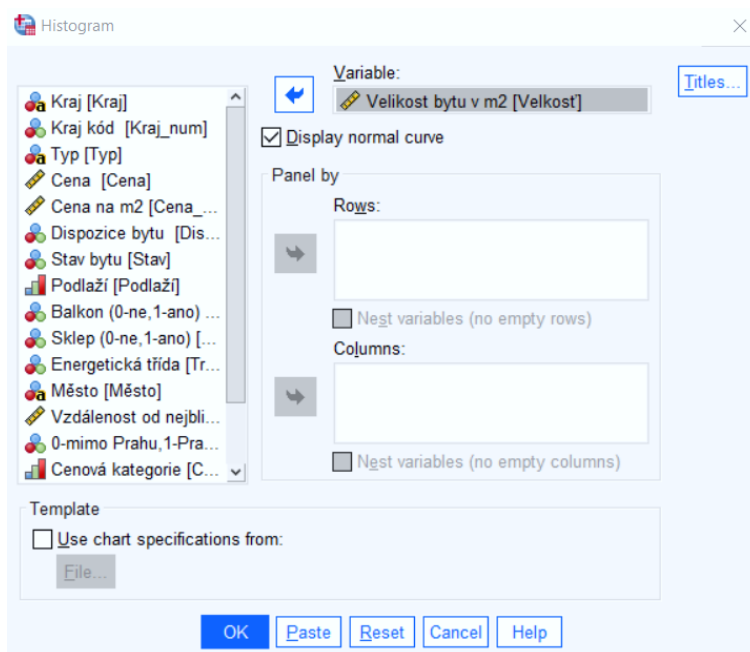
Obvykle se používá pro vizualizaci rozdělení proměnné. Histogram je složen z řady sloupců, přičemž každý sloupec odpovídá určitému rozmezí hodnot vykreslované proměnné. Výška sloupce představuje četnost výskytu hodnot v daném intervalu. Histogram je užitečným nástrojem pro vizuální analýzu dat a může pomoci odhalit vlastnosti rozdělení, jako je například střední hodnota, rozptyl, šikmost a zešikmení. Histogram se často používá k posouzení normality dat.

Při hodnocení normality je třeba zkontrolovat, zda se hodnoty rozložení shodují se základními charakteristikami normálního rozdělení. To zahrnuje zjištění, zda je histogram symetrický, zda má zvonovitý tvar a zda jsou střední hodnota a medián blízké.

## PŘÍKLAD

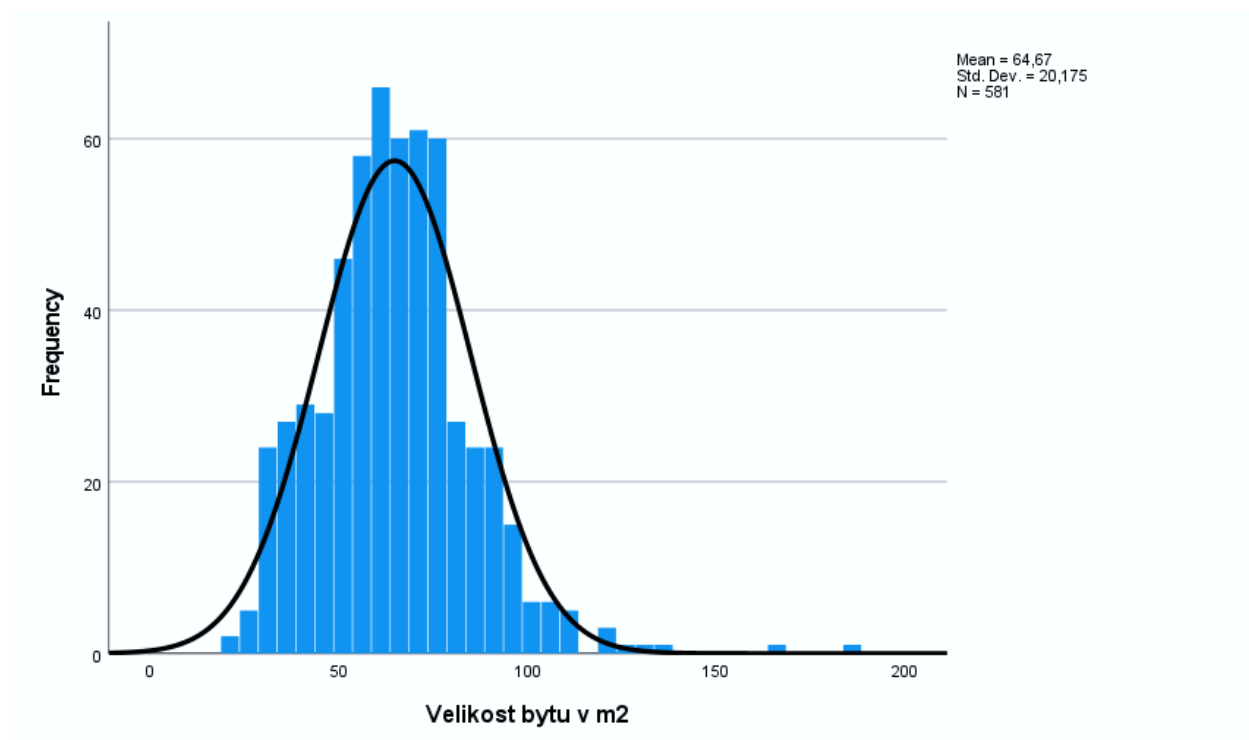
## DATOVÝ SOUBOR – BYTY.SAV

Pomocí histogramu se pokusíme ověřit, zda je proměnná Velikost bytu normálně rozdělena. Otevřeme nabídku *Graphs-Legacy dialogs -Histogram*.



Do okna *Variable* zadáme proměnnou *Velikost bytu v m2*. Dále zaškrtneme možnost *Display normal curve*. Tato volba automaticky do histogramu vykreslí křivku normálního rozdělení, což zjednoduší posouzení normality dané proměnné.

Na histogramu lze pozorovat, že rozdělení Velikosti bytu na m<sup>2</sup> se přibližuje černé křivce reprezentující normální rozdělení. Odchytky, které v grafu vidíme, nejsou nijak zásadní a pro danou proměnnou tedy bude možné použít i některé z parametrických statistických metod.



## P-P PLOT (PROBABILITY-PROBABILITY PLOT)

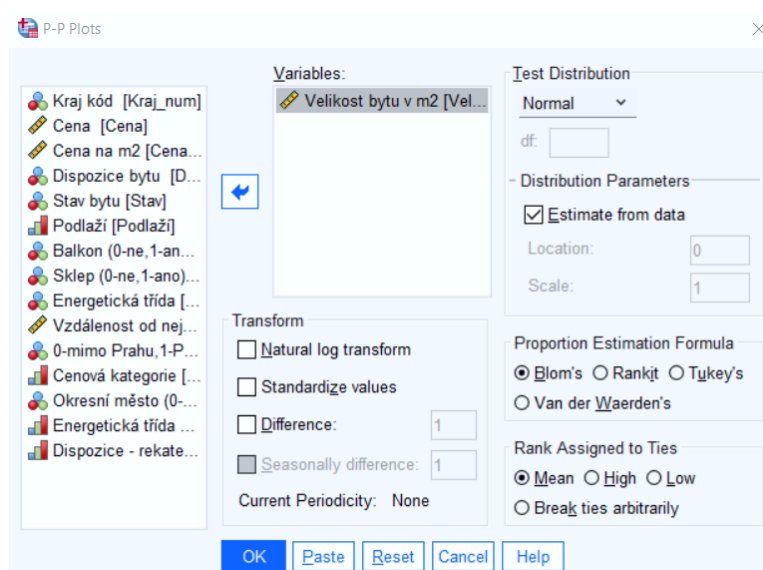
P-P plot (Probability-Probability plot) je grafická metoda, která slouží k posouzení, zda jsou dvě proměnné vybrány ze stejného rozdělení pravděpodobnosti. V SPSS se hodnoty proměnných nejprve transformují pomocí kvantilů a následně pomocí hodnot empirické pravděpodobnosti a kumulativní distribuční funkce standardního normálního rozdělení. Tento postup transformace následně umožňuje snadnou vizuální kontrolu, zda data pocházejí například z normálního rozdělení. Pokud je proměnná normálně rozdělená, body v P-P ploty by měli být přibližně rovnoměrně rozmístěny podél diagonální přímky, která představuje očekávané kvantily standardního normálního rozdělení.

### PŘÍKLAD

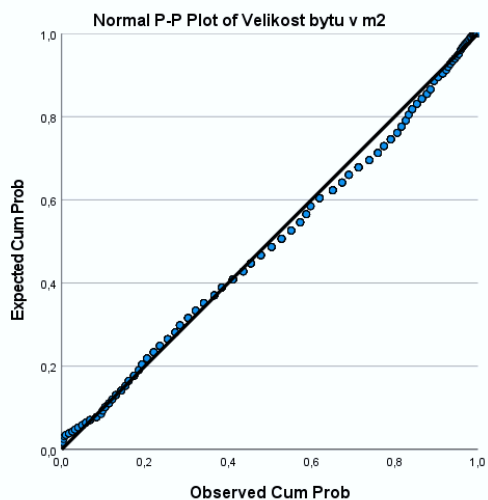
#### DATOVÝ SOUBOR – BYTY.SAV

Pomocí P-P plotu se pokusíme ověřit, zda je proměnná Velikost bytu normálně rozdělena. Otevřeme nabídku *Analyze-Descriptive Statistics-P-P plots*.

Do okna *Variables* přesuneme proměnnou velikost bytu, jejíž rozdělení chceme porovnat s normálním rozdělením. V nastavení *Test Distribution* je již toto rozdělení zvoleno defaultně (*Normal*). Kromě normálního rozdělení ale máme na výběr také jiné možnosti, které mají již více specifické využití. Ponecháme tedy nastavení *Normal* a klikneme na OK.



Při interpretaci P-P plotu je třeba sledovat, jak body v grafu následují diagonální přímku, která zobrazuje kvantily normálního rozdělení. Pokud je proměnná normálně rozdělena, body v grafu by měly být přibližně rovnoměrně rozmístěny podél diagonální přímky. Pokud jsou body v grafu od diagonální přímky vzdálené, může to naznačovat, že data nejsou normálně rozdělena.



## Q-Q PLOT (QUANTILE-QUANTILE PLOT)

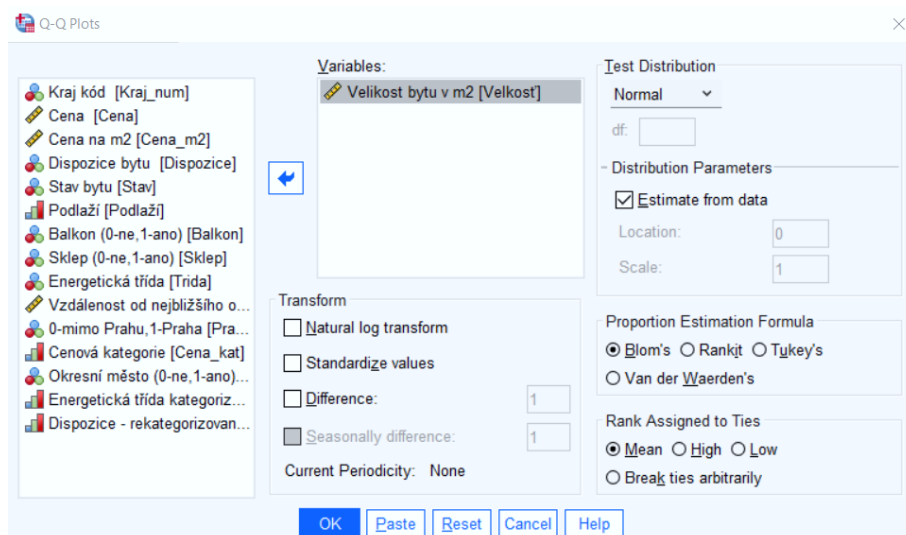
Podobně jako P-P plot, i Q-Q plot porovnává empirické kvantily dat s teoretickými kvantily normálního rozdělení, ale místo empirických pravděpodobností používá kvantily dat. Stejně jako u P-P plotu, může Q-Q plot poskytnout informace o tom, jaký typ zkreslení má zkoumaná proměnná, pokud se odchyluje od normálního rozdělení.

### PŘÍKLAD

#### DATOVÝ SOUBOR: BYTY.SAV

Pomocí Q-Q plotu se pokusíme ověřit, zda je proměnná Velikost bytu normálně rozdělena. Otevřeme nabídku *Analyze-Descriptive Statistics-Q-Q plots*.

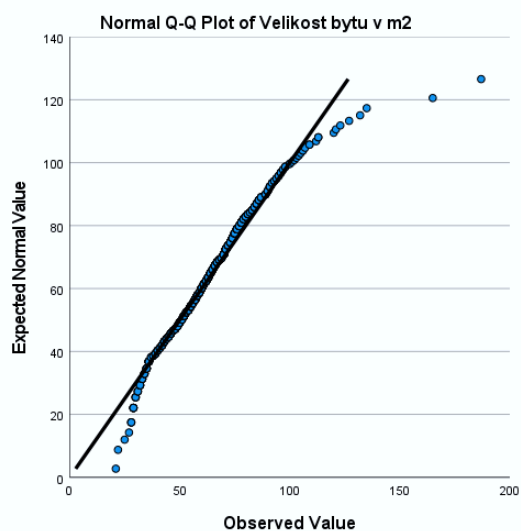
Do okna *Variables* přesuneme proměnnou velikost bytu, jejíž rozdělení chceme porovnat s normálním rozdělením. V nastavení *Test Distribution* je již toto rozdělení zvoleno defaultně (*Normal*). Kromě normálního rozdělení ale máme na výběr také jiné možnosti, které mají již více specifické využití. Ponecháme tedy nastavení *Normal* a klikneme na OK.



Obecně platí, že pokud je proměnná normálně rozdělená měli by být body na grafu přibližně na přímce. Pokud jsou body na grafu zakřivené, znamená to, že se rozdělení liší od normálního.



Je důležité si uvědomit, že Q-Q plot (také i P-P plot) může být citlivý na velikost vzorků a že malé rozdíly v rozděleních mohou být statisticky významné, i když jsou na grafu téměř nepatrné. Proto by měl být Q-Q plot vždy používán v kombinaci s dalšími statistickými metodami pro ověření rozdílů mezi datovými soubory.



## KOLMOGOROV-SMIRNOVŮV TEST PRO OVĚŘENÍ NORMALITY

Jednovýběrový K-S test testuje nulovou hypotézu, že pozorovaná data pocházejí z normálního rozdělení s určitým průměrem a rozptylem. Testování se provádí tím, že se porovnávají hodnoty empirické distribuční funkce (distribuční funkce testované proměnné) s hodnotami distribuční funkce normálního rozdělení s odpovídajícím průměrem a rozptylem. Pokud jsou pozorovaná data normálně rozdělena, pak by měly být hodnoty empirické distribuční funkce blízké k hodnotám distribuční funkce normálního rozdělení.

$H_0$ : Data pochází z normálního rozdělení

$H_1$ : Data nejsou normálně rozdělena

Je důležité si uvědomit, že K-S test je citlivý na velikost vzorku a také na to, jakým způsobem jsou data získávána. Proto je při posuzování normality důležité zohlednit i jiné nástroje, jako například histogram.

## DALŠÍ TESTY NORMALITY

### SHAPIRO-WILK

Test se opírá o statistiku  $W$ , která měří míru shody mezi daty a normálním rozdělením. Shapiro-Wilkův test je obecně považován za citlivější na odchylky od normality než Kolmogorov-Smirnovův test. To znamená, že pokud jsou data mírně odchýlena od normálního rozdělení, může být Shapiro-Wilkův test schopen detekovat tuto odchylku lépe než Kolmogorov-Smirnovův test. Nulová hypotéza pro tento test říká, že data jsou normálně rozdělena. Alternativní hypotéza tvrdí, že data nejsou normálně rozdělena.

### ANDERSON-DARLING

Tento test se zaměřuje na hodnoty vzdálenosti od teoretického normálního rozdělení a měří odchylku mezi teoretickým normálním rozdělením a rozdělením testované proměnné. Nulová hypotéza pro tento test říká, že data jsou normálně rozdělena. Alternativní hypotéza tvrdí, že data nejsou normálně rozdělena.

## SHRNUTÍ

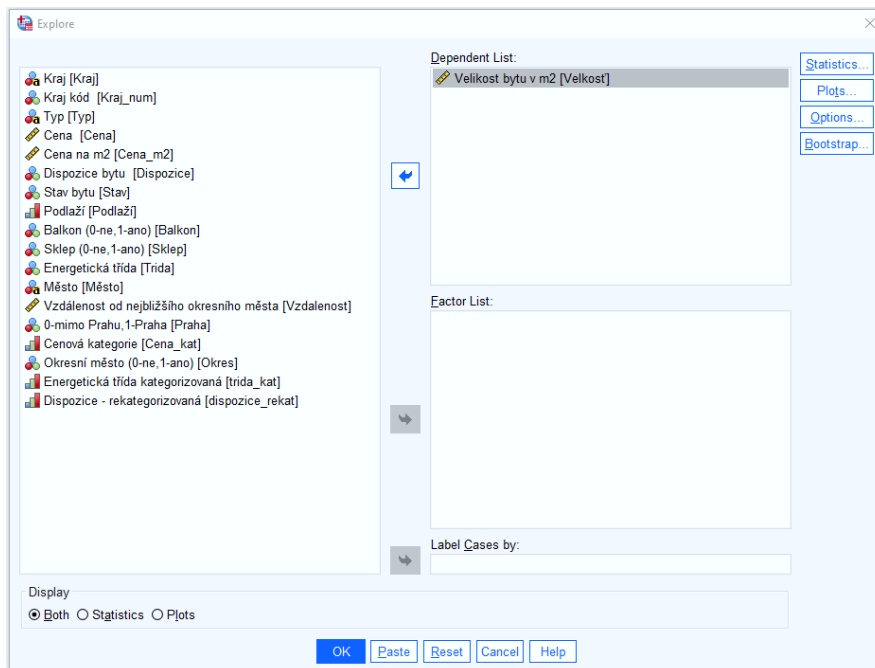
Kolmogorov-Smirnovův je často používán pro testování normality dat, ale může být také použit pro testování shody s jinými rozděleními. Další výhodou Kolmogorov-Smirnovova testu je jeho robustnost proti chybám výběru, což znamená, že test bude poskytovat spolehlivé výsledky, i když jsou data mírně odchýlena od testovaného teoretického rozdělení. Nicméně, je důležité si uvědomit, že Kolmogorov-Smirnovův test je méně citlivý na odchylky od testovaného rozdělení ve srovnání s jinými testy, jako je Shapiro-Wilkův test nebo Anderson-Darlingův test. Také není vhodný pro velmi malé vzorky dat ( $n < 30$ ). Celkově lze říci, že volba testu na testování shody dat s teoretickým rozdělením by měla být založena na konkrétní situaci a vlastnostech dat, která chcete testovat.

## Příklad

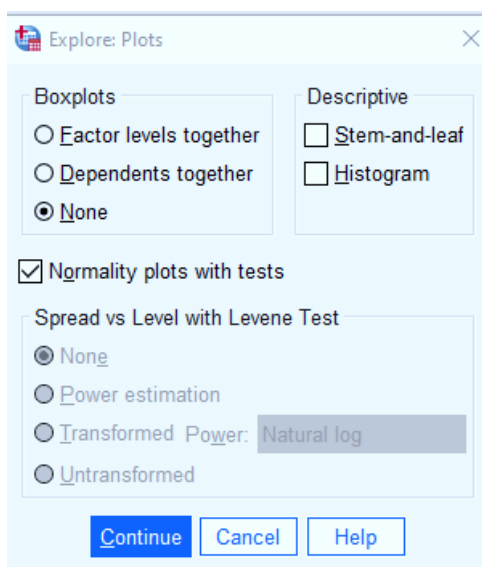
### DATOVÝ SOUBOR BYTY.SAV

Pomocí ANOVY (parametrického testu o střední hodnotě) bychom chtěli otestovat, zda se liší průměrná velikost bytů v jednotlivých krajích ČR. Pro ověření této hypotézy musíme ověřit předpoklad o normálním rozdělení proměnné.

Klikneme na ANALYZE-DESCRIPTIVE STATISTICS-EXPLORE. Do pole *Dependent List* zadáme proměnnou Velikost bytu v m2.



Dále klikneme na záložku *Plots*, kde si zvolíme testy normality (*Normality plots with tests*). Zároveň můžeme zrušit volbu grafů *Stem-and-leaf* a *boxplot*.



## INTERPRETACE VÝSLEDKU

V tabulce **Test of Normality** najdeme údaj o hodnotě testové statistiky (*Statistics*) Kolmogorov Smirnovova testu, počet stupňů volnosti (*df*) signifikanci (*Sig.*)

Tests of Normality						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Velikost bytu v m2	,057	581	<,001	,961	581	<,001

a. Lilliefors Significance Correction

**Pro interpretaci testu stačí porovnat hodnotu signifikance se zvolenou hladinou významnosti.**

**Na 5 % hladině významnosti se zamítá nulová hypotéza o normálním rozdělení proměnné Velikost bytu v m2 (protože signifikance (0,001) je menší než hladina významnosti (0,05)).**

Na základě tohoto testu bychom tedy nemohli pro ověření hypotézy o stejné průměrné velikosti bytů v jednotlivých krajích ČR použít parametrický test ANOVA.