Data1: Naměřené hodnoty studované veličiny, např. výška rostliny.
Data3: V souboru Data1 změněna poslední hodnota na 40.

Data1	Data3
1	1
5	5
7	7
8	8
9	9
10	10
11	11
12	12
12.5	12.5
13	13
14	14
15	15
16	16
17	17
18	18
20	20
24	40

Příklad 5: Explorační analýza dat – vizualizace dat frekvenčními histogramy a krabicovými diagramy, percentile plots.

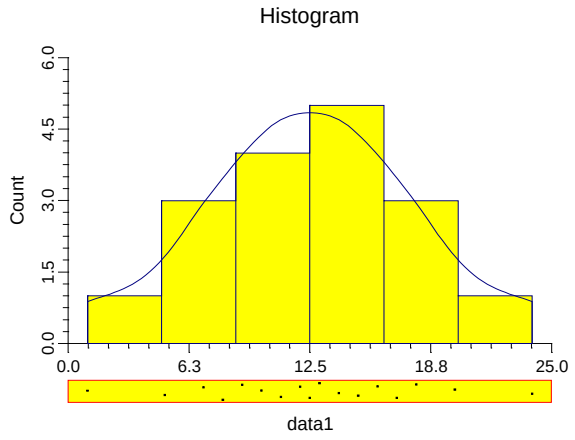
Použité proměnné: Data1, Data3

Smyslem explorační analýzy dat je získat informaci o charakteru rozložení dat. Oproti "suchým" výpočtům (viz předchozí příklady) jsou k tomuto účelu k dispozici vhodná rozličná grafická zobrazení. Nejčastějšími typy grafické vizualizace dat jsou frekvenční histogramy (*Frequency Histogram*) a krabicové diagramy (*Box Plots*).

Frekvenční histogramy znázorňují četnost měření v určitých (vhodně stanovených) intervalech. Výsledkem jsou sloupečky vynesené na ose x (ta postihuje hodnoty studované proměnné), jejichž výška odpovídá četnostem měření v daném rozsahu hodnot každého sloupečku – čím vyšší četnost měření v daném intervalu hodnot tím vyšší sloupeček. V programu NCSS histogramy získáme v **Graphics/Histograms**.

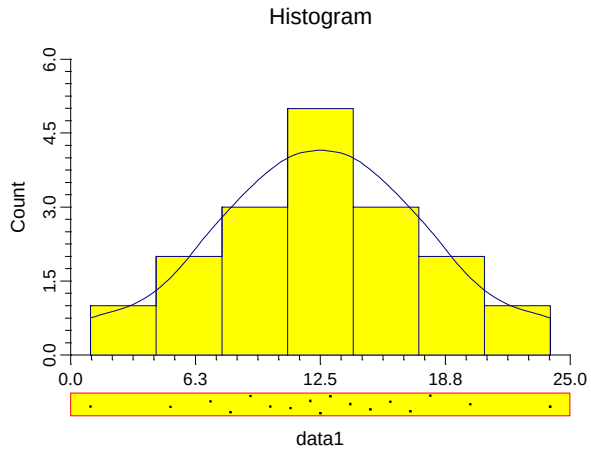
Frekvenční histogram souboru Data1 má šest sloupečků (jejich počet je určen automaticky programem) a je mírně asymetrický, tj. v pravé části je strmější. Žlutý pruh pod osou x je tzv. *jitter plot*, který ukazuje počet měření a zároveň jejich hodnotu v každém intervalu (viz černé body uvnitř).

Data1: Default histogram
Histogram Section



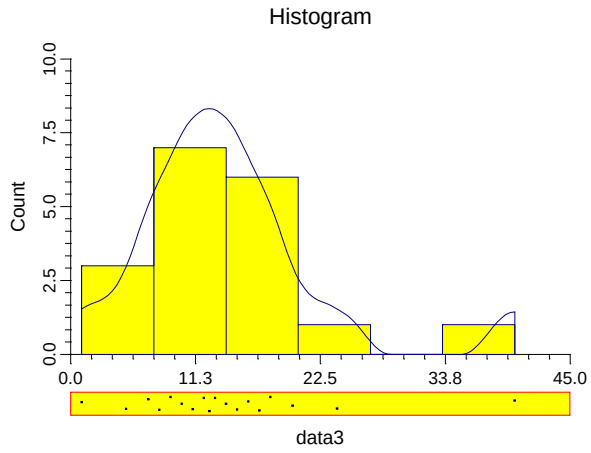
V dialogovém okně zadání histogramu (zelená kostka) můžeme změnit počet sloupečků v oddíle *Bars*, na druhém řádku označeném *Number Bins*. Jestliže Data1 rozdělíme na sedm intervalů, tzn. k předchozímu přidáme jeden sloupeček, výsledný histogram je symetrický. Je tedy třeba mít na paměti, že grafická prezentace, ukazující trend našich dat, je do určité míry ovlivněna použitým měřítkem ("jemností" při rozdělení hodnot na ose x do kategorií), v tomto případě počtem sloupečků histogramu.

Data1: Změna počtu sloupců na 7
Histogram Section



Histogram souboru Data3, kde je zahrnuto jedno odlehlé měření (hodnota 40), je výrazně asymetrický. Tuto asymetrii, která je dána skutečnou nerovnoměrností rozdělení dat, pochopitelně nelze odstrant změnou počtu sloupečků histogramu.

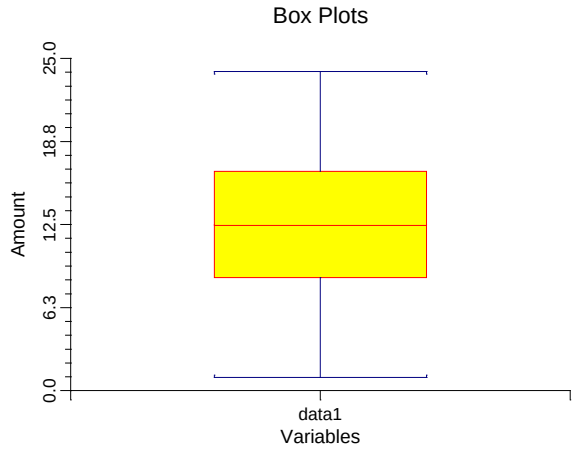
Data3: Default histogram
Histogram Section



Krabicové diagramy jsou jinou formou vizualizace rozdělení dat souboru. V NCSS se tato volba se nalézá v oddíle **Graphics/Box Plots**. Rozměry krabicového diagramu jsou vhodně zvoleny a odpovídají určitým statistickým parametrům zobrazovaného souboru. Výška krabice má velikost mezikvartilového rozpětí (tedy vzdálenosti mezi 25% a 75% percentily), příčná čára uvnitř krabice odpovídá hodnotě mediánu – konkrétní hodnoty jsou pak vyneseny na ose y. Svislé čáry nad a pod krabicí zakončené příčnou úsečkou (tzv. whiskers) znázorňují velikost intervalu, jehož velikost je maximálně rovna 1.5 násobku mezikvartilového rozpětí. Jinými slovy, konce těchto whiskers zasahují do nejkrajnější hodnoty měření spadající do tohoto

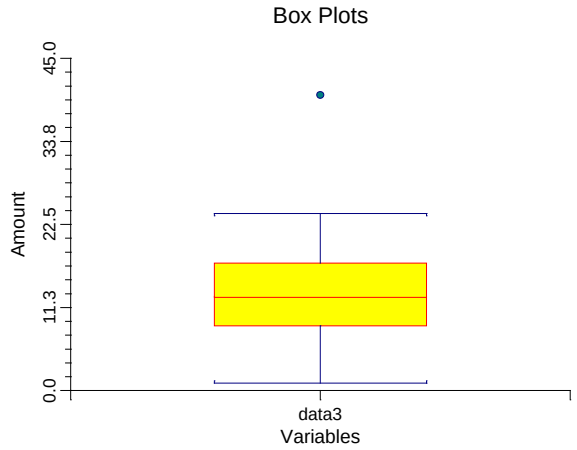
rozmezí. Z krabicového diagramu opět vidíme, že datový soubor je symetrický.

Box Plot Section
Data1



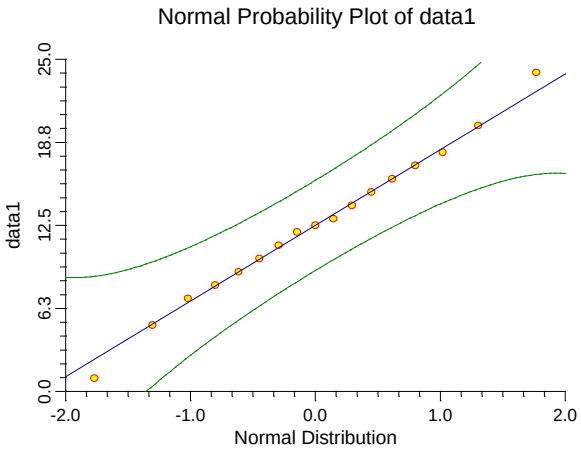
Krabicový diagram souboru Data3 obsahuje jedno odlehlé měření (hodnota 40), které se nalézá za hranicí vymezenou intervalem 1.5 x mezikvartilové rozpětí. Tato odlehlá hodnota je znázorněna jako bod nad vlastním diagramem.

Box Plot Section
Data3



Probability plots (Q-Q-plots) jsou grafickým znázorněním míry, s jakou se rozdělení dat souboru blíží rozdělení normálnímu (programu NCSS tuto možnost nalezneme v oddíle **Graphics/Probability Plots**). Na ose y jsou vyneseny konkrétní hodnoty měření proti odpovídajícím percentilům normálního rozdělení (tzv. NED, normal equivalent deviates) vyneseným na ose x; např. medián datového souboru odpovídá nulové hodnotě NED. Čím více se datový soubor blíží normálnímu rozdělení, tím lépe hodnoty odpovídají modré přímce (ta odpovídá právě normálnímu rozdělení). Z následujícího grafu je zřejmé, že rozdělení hodnot souboru Data1 normálnímu rozdělení odpovídá velmi dobře.

Data1
Probability Plot Section



Narozdíl od předchozího souboru, odlehlá hodnota souboru Data3 se výrazně odchyluje od modré přímky a tudíž evidentně výrazně narušuje normalitu rozdělení dat.

Data3
Probability Plot Section