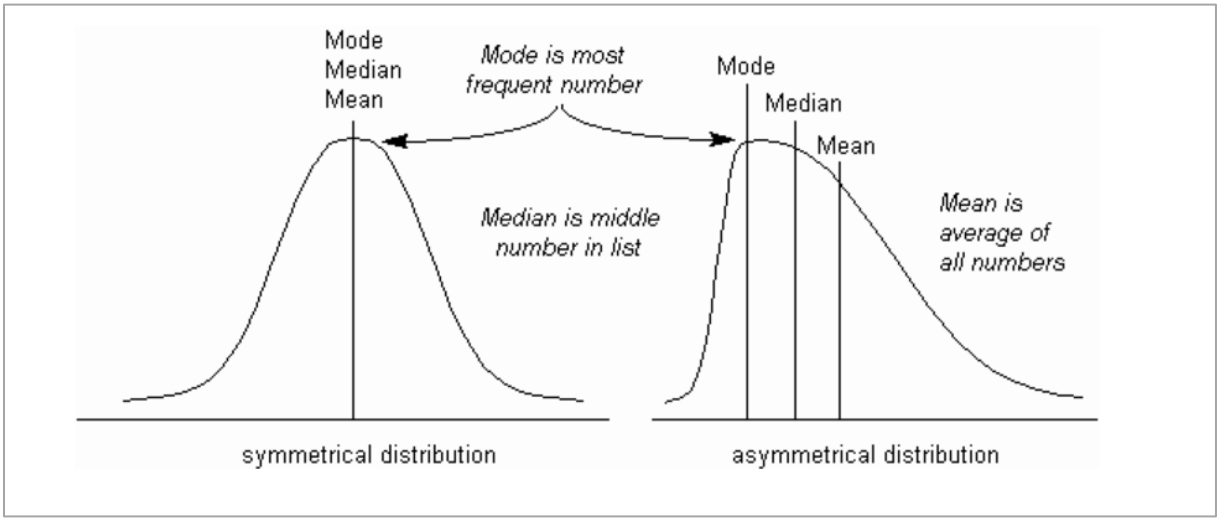
Popis dat

Používá se k určení, z jakého rozdělení data pocházejí, a také k odhalení chybných pozorování (odlehlých hodnot). Při popisu odhadujeme **střední hodnotu** a **variabilitu** rozdělení pravděpodobnosti, ze kterého naše data pocházejí. Je důležité vždy uvádět oba tyto parametry.

Číselné charakteristiky dat

Ukazatele středu:

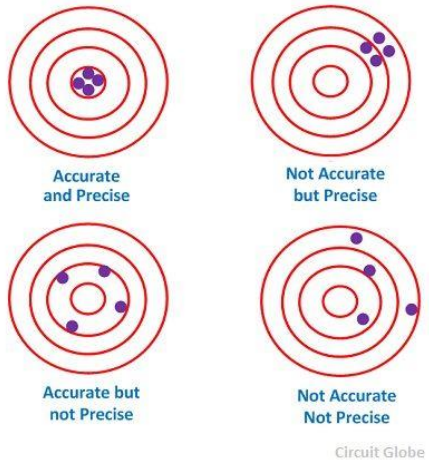
- **Modus** je nejčetnější varianta, je vhodný všechny typy dat (i pro binární a nominální)
- **Medián ($\tilde{x}$)** je hodnota prostředního výsledku (nebo průměr ze dvou prostředních při sudém počtu vzorků), je vhodný pro ordinální a kvantitativní data. Podobně je možné určit např. kvartily – 1. kvartil je výsledek, který leží v čtvrtině, 2. kvartil je medián a 3. kvartil leží ve 3 čtvrtinách výsledků seřazených podle velikosti.
- Aritmetický **průměr ($\bar{x}$)** je součet všech hodnot vydělený jejich počtem. Je vhodný pro kvantitativní data. Ukazatelem středu je nicméně jen tehdy, pokud data mají symetrické rozdělení pravděpodobnosti.



Ukazatele variability:

- **Rozptyl (s^2)** je „suma čtverců odchylek“ lomená (n-1): $\sum(x_i - \bar{x})^2/(n-1)$. Pozn: u rozptylu populace (nebo velkého souboru) lomíme počtem opakování (n)
- **Směrodatná odchylka (s**) je odrazem variability měřené veličiny v celé populaci. Jde vlastně o průměrnou chybu (průměr z rozdílů mezi hodnotami jednotlivých vzorků od průměru celého výběru). Proto není ovlivněna velikostí vzorku.
- **Relativní směrodatná odchylka (s_r**) se udává v procentech průměru: $(s/\bar{x})\times100$
- **Střední chyba průměru (SEM**) je odrazem přesnosti odhadu průměru. Je ovlivněná variabilitou a velikostí výběru. $SEM=s/\sqrt{n}$
- **Interval spolehlivosti** souvisí s SEM. Např. 95% interval spolehlivosti je interval, v němž leží průměr pro 95 výběrů ze 100. Čím větší máme výběr, tím užší bude interval spolehlivosti. 90% interval bude užší než 95%, protože stačí, když se do něj vejdou průměry 90 % pokusů.

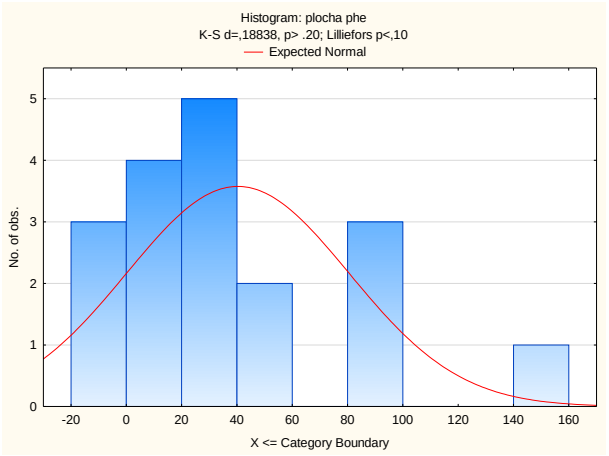
Preciznost měření (**precision**) můžeme zjistit srovnáním několika opakování (vyjadřuje se nejčastěji jako směrodatná odchylka), zatímco **pravdivost** (**trueness**) zjistíme srovnáním našeho průměrného výsledku s ověřenou referenční hodnotou. **Opakovatelnost** je přesnost měření provedených ve stejné laboratoři s časovým odstupem a **reprodukovatelnost** je přesnost měření provedených v různých laboratořích.

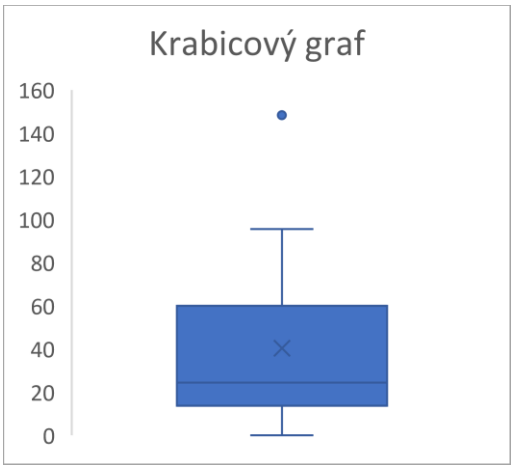


Vizualizace dat

Kvalitativní data vizualizujeme pomocí **sloupcových** nebo **koláčových** grafů. Sloupcové se hodí pro vizualizaci absolutních četností, zatímco koláčové pro vizualizaci procentuálního zastoupení jednotlivých variant.

Kvantitativní data (případně i ordinální) lze vizualizovat pomocí **histogramu**. Spojitá data je před vizualizací potřeba rozdělit do intervalů, počet intervalů bychom měli volit cca √n. Pozor na rozdíl mezi sloupcovým grafem a histogramem – u histogramu jsou naměřené hodnoty na ose x. Histogram se také hodí pro porovnání tvaru hustoty četností s tvarem hustoty pravděpodobnosti očekávaného rozdělení.





Pokud sledujeme spojitá data (i v několika skupinách – kategoriích), je vhodný **krabicový graf**. Tento graf navíc umožňuje posoudit symetrii a variabilitu dat a odhalit případné odlehlé hodnoty. Existuje víc variant krabicového grafu, proto je třeba číst popisky. Na grafu na obrázku vidíme krabicový graf pro data z příkladu. Křížkem je označený průměr, vodorovnou čarou medián, úsečkami extrémní (ale ne odlehlé) hodnoty a bodem odlehlá hodnota.

Pokud jsou data i sledovaný parametr spojité (sledujeme vlastně vztah dvou proměnných), použijeme **bodový graf**.

Identifikace odlehlých hodnot

V našem příkladu krabicového grafu jsou jako odlehlé označeny hodnoty, které jsou od okraje boxu – 1. a 3. kvartilu – vzdálené víc než 1,5×**IQR – kvartilové rozpětí** (rozdíl hodnot 1. a 3. kvartilu). Pro identifikaci odlehlých hodnot je kromě toho možné použít např. **Grubbsův test** (dostupný online).

Odlehlou hodnotu nelze vyloučit při n=3, kdy dvě opakování jsou stejná (a samozřejmě při n<3). Proto je dobré mít vždy alespoň 4 opakování.