Učitel matematiky

František Mošna

Box-ploty v MS Excelu a jejich možné využití ve výuce

Učitel matematiky, Vol. 23 (2015), No. 4, 206–214

Persistent URL: http://dml.cz/dmlcz/149436

BOX-PLOTY V MS EXCELU A JEJICH MOŽNÉ VYUŽITÍ VE VÝUCE

František Mošna

Popisná statistika shromažďuje určitá data, zpracovává je a přehledně prezentuje. Užívá k tomu mimo jiné i grafických prostředků. Mezi nimi najdeme tzv. krabicové grafy či box-ploty, které nejsou v českém prostředí příliš známy. Právě jimi se budeme zabývat, a to z hlediska jejich možného využití ve výuce a tvorby v programu MS Excel.

Charakteristiky a box-ploty

K charakterizaci jistého souboru dat se běžně užívá prostý (aritmetický) průměr $\bar{x}$ (anglicky mean). Vyjadřuje jakýsi střed, kolem kterého jsou hodnoty souboru rozloženy. Někdy se přidává informace o směrodatné odchylce. Ta udává, jak hodně či málo jsou hodnoty kolem tohoto středu rozptýleny či soustředěny. Mnohdy však tyto dvě charakteristiky nedávají dostatečnou představu o souboru dat, neboť nevystihují jeho podstatné vlastnosti.

Mezi další významné charakteristiky souboru dat patří modus $\widehat{x}$ – nejčetnější hodnota a medián $\tilde{x}$ – prostřední hodnota (tedy 50 % hodnot není větších a 50 % není menších než tento medián).

Medián zpravidla doplňujeme ještě o tzv. kvartily. Dolní kvartil $\tilde{x}_{25}$ je takové číslo, pro které 25 % hodnot je menších nebo rovných tomuto číslu a 75 % hodnot je větších nebo rovných tomuto číslu. U horního kvartilu $\tilde{x}_{75}$ je tomu obráceně, tedy 75 % hodnot je menších nebo rovných číslu $\tilde{x}_{75}$ a 25 % hodnot je větších nebo rovných tomuto číslu. Poznamenejme, že ani modus, ani medián, ani kvartily nemusejí být zavedeny jednoznačně. S kvartily souvisí ještě tzv. mezikvartilové rozpětí $\tilde{R} = \tilde{x}_{75} - \tilde{x}_{25}$, udávající

informaci o rozložení souboru dat. Někdy nás zajímá také nejmenší a největší hodnota x_{min}, x_{max}.

Je všeobecně známo, že více než mnoho slov, čísel a vzorců sdělí obrázek. Koncem 70. let minulého století zavedl John W. Tukey nový nástroj k popisu rozložení souboru kvantitativních dat – tzv. box-ploty, viz (Tukey, 1977). Česky je překládáme většinou jako krabicové grafy (mediánové grafy nebo podobně). Box-plot prezentuje 5 námi uvedených charakteristik (medián, kvartily a extrémy) graficky. Celý soubor dat je rozdělen do čtyř částí a box-plot zobrazuje rozsah hodnot v každé čtvrtině. Prostřední část znázorňuje mezikvartilové rozpětí. Nespornou výhodou box-plotů je jejich názornost a relativní objektivita.

V box-plotech se někdy vyznačují také tzv. odlehlé hodnoty. Za odlehlé zpravidla uvažujeme hodnoty větší než $\tilde{x}_{75}+1{,}5\cdot\tilde{R}$ nebo menší než $\tilde{x}_{25}-1{,}5\cdot\tilde{R}$, viz např. (Charamza & Hanousek, 1992). Poznamenejme, že do box-plotů se někdy dosazují místo mediánu, kvartilů a extrémů jiné hodnoty či meze (průměr, směrodatná odchylka nebo její 1,96 násobek).

Volba charakteristik

Vypráví se anekdota, že existuje „lež“, „sprostá lež“ a „statistika“. Některé výpovědi týkající se hromadných dat a užívající statistických pojmů mohou skutečně vypadat poněkud podezřele, nevěrohodně a manipulovaně. Statistika však na vině nebývá. Záleží vždy na zdroji, jaké informace se rozhodne sdělit, uvést, zdůraznit nebo naopak zatajit, potlačit, upozadit a jaké k tomu použije statistické pojmy či prostředky. Miloš Rejchrt v jedné písni zpíval, že „půlka pravdy může taky lhát“. Na druhou stranu poznamenejme, že absolutní objektivity dosáhnout nelze, neboť o každé události či o každém jevu lze vyslovit nekonečně mnoho tvrzení. Toho však žádný člověk při nejlepší vůli není schopen.

Ukážeme si na příkladu význam uvedených statistických nástrojů a to, jak lze jejich cíleným výběrem informovat jednostranně či neobjektivně.

Příklad 1
K běžeckému závodu kategorie A se přihlásilo 11 účastníků. Kromě jiných údajů byl zjišťován také jejich věk:

$$17, 20, 30, 46, 27, 15, 20, 22, 20, 29, 29.$$

Pokusme se charakterizovat věk účastníků závodu kategorie A.

O stáří (lépe řečeno mládí) účastníků lze pronést různé výroky. Například:
„Nejvíce účastníků je ve věku 20 let.“
„Padesát procent účastníků není starších než 22 let.“
„Nejstaršímu účastníkovi je už 46 let.“
„Průměrný věk účastníků 25 let.“
„Prostředních 50 % účastníků je ve věku 20 až 29 let.“

Uvedená tvrzení můžeme jednoduše vyjádřit pomocí zavedených charakteristik $\widehat{x} = 20$, $\tilde{x} \leq 22$, $x_{max} = 46$, $\bar{x} = 25$ a $\tilde{x}_{25} = 20$, $\tilde{x}_{75} = 29$.

Všechna tato tvrzení jsou pravdivá. Každé z nich však zdůrazňuje jinou vlastnost dat. První dva výroky se snaží podtrhnout, že účastníci jsou spíše mladí. Další dva výroky vyznívají naopak spíše ve prospěch vysokého věku účastníků. Poslední výrok vypovídá o rozsahu věku účastníků. Když uvedeme jen některou informaci a jinou zamlčíme, nepodáváme tím úplný obraz. Často tedy nestačí pouze sdělit informaci o průměru (a směrodatné odchylce), je třeba podívat se na soubor dat z více hledisek, uvést více charakteristik a vyjádřit je graficky například pomocí box-plotu.

Podobné příklady jsou uvedeny například v knize (Swoboda, 1977).

Porovnávání souborů

Box-ploty jsou vhodné také pro vizuální porovnávání dvou nebo více souborů dat a dobře doplňují dvouvýběrové, párové testy nebo metody třídění.

Přidejme ještě jeden podobný příklad.

Příklad 2
K běžeckému závodu kategorie B se přihlásilo 13 účastníků. Opět byl zjišťován jejich věk:

$$24, 28, 21, 19, 21, 28, 28, 26, 30, 28, 22, 26, 24.$$

Měli bychom porovnat závody obou kategorií A (příklad 1) a B z hlediska věkových dispozic.

Zkusme vyslovit podobná tvrzení jako u příkladu 1. Všimněme si, že průměry jsou u obou souborů dat A i B stejné. Přesto je charakter dat poněkud odlišný. Liší se u nich modus, medián, kvartily i extrémy. Z box-plotu je hezky patrné, že data souboru B jsou více soustředěna kolem centra, než je tomu u souboru dat A (viz obr. 4).

Nástroje k vytváření box-plotů

S box-ploty si dovede poradit řada statistických programů. Velmi jednoduše je lze získat například pomocí software STATISTICA nebo SPSS. Box-ploty lze také zobrazovat pomocí některých grafických kalkulaček, viz (Robová, 2004).

V MS Excelu neexistuje nástroj pro přímé vytvoření box-plotů ze souboru dat. Můžeme si pořídit doplňky Tech Chart Utility (BoxCharter) od Jona Peltiera (viz `http://peltiertech.com`), s jejichž pomocí lze tyto grafy tvořit přímo, viz (Cihlář, 2008).

Excel má v nabídce několik typů grafů tzv. burzovní grafy, které udávají většinou kurs akcií na počátku, maximum, minimum a kurs na konci období. Lze jich využít i k vytvoření jakéhosi grafu nahrazujícího box-ploty, neboť jsou jim graficky velmi podobné.

Třetí možnost je poněkud složitější a spočívá v užití kombinace standardních grafických nástrojů Excelu – sloupcových grafů a tzv. chybových úseček (grafický prvek Excelu znázorňující intervalové odchylky od nějaké základní hodnoty). Následující text popisuje a shrnuje základní kroky na této pracnější cestě.

Ukážeme si návod k vytvoření box-plotu na datech uvedených v příkladech 1 a 2.

U obou skupin dat zjistíme charakteristiky uváděné v box-plotu, tedy medián, kvartily a extrémy. K tomu se nejlépe hodí

statistická funkce *QUARTIL*, do níž dosadíme příslušná data a zvolíme vhodný parametr. Pro minimum je to parametr 0, užijeme tedy funkci *QUARTIL(data;0)* nebo také *MIN(data)*, pro dolní kvartil *QUARTIL(data;1)*, pro medián *QUARTIL(data;2)* nebo také *MEDIAN(data)*, pro horní kvartil *QUARTIL(data;3)* a pro maximum *QUARTIL(data;4)* nebo také *MAX(data)*. Ponechme stranou problematiku korektnosti a způsobu výpočtu uvedených funkcí v Excelu.

Přípravná tabulka

Pro vytvoření box-plotu si připravíme tabulku obsahující dolní kvartil (Q1), a rozdíly medián – dolní kvartil (Q2 – Q1), horní kvartil – medián (Q3 – Q2), maximum – horní kvartil (Q4 – Q3) a dolní kvartil – minimum (Q1 – Q0), viz tab. 1 vpravo.

	A	B	C	D	E	F	G	H
1	**Prvky box plotu:**				**Tabulka k sestavení box plotu:**			
2		**A**	**B**			**A**	**B**	
3	Min (Quartil 0)	**15**	**19**		Q1	**20**	**22**	
4	Quartil 1	**20**	**22**		Q2-Q1	**2**	**4**	
5	Median (Quartil 2)	**22**	**26**		Q3-Q2	**7**	**2**	
6	Quartil 3	**29**	**28**		Q4-Q3	**17**	**2**	
7	Max (Quartil 4)	**46**	**30**		Q1-Q0	**5**	**3**	
8								

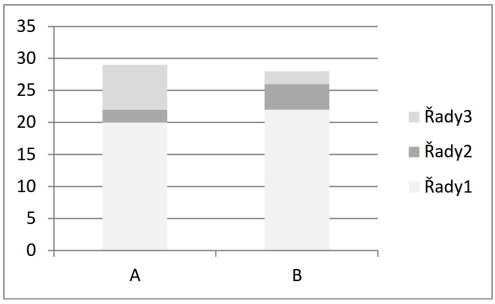
Tab. 1: Tabulky pro sestavení box-plotu

Sloupcový graf

Když máme připravenu tabulku, začneme vytvářet sloupcový graf následujícím postupem:

- označíme do bloku popisky a data z prvních tří řádků připravené tabulky (řádky Q1, Q2 – Q1, Q3 – Q2),
- přes položky *vložení, grafy, sloupcový* (vybereme druhý typ zleva v prvním řádku) vytvoříme sloupcový graf,
- eventuálně vyměníme sloupce za řádky (klikneme pravou myší na graf, *vybrat data, přepnout řádek/sloupec a OK*).

Výsledek vidíme na obr. 1.



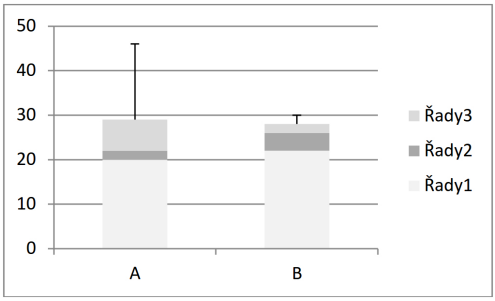
Obr. 1: Sloupcový graf

Chybové úsečky

Následuje přidání horní chybové úsečky:

- označíme horní třetinu u jednoho sloupcového grafu,
- přes záložky *nástroje grafu, rozložení, analýza* zvolíme *chybové úsečky*,
- pokračujeme přes *další možnosti chybových úseček*, volíme zobrazení *plus*, typ chybové hodnoty *vlastní, zadat hodnotu*,
- za kladnou chybovou hodnotu dosadíme data z předposledního řádku připravené tabulky (Q4 – Q3), potvrdíme *OK* a *zavřít*.

Výsledek vidíme na obr. 2.



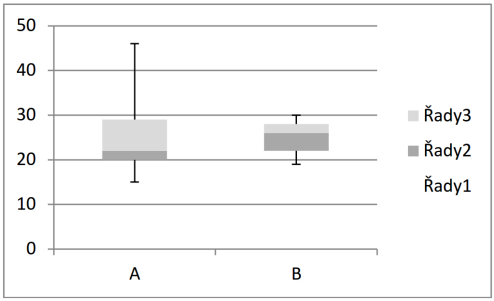
Obr. 2: Horní chybová úsečka

Podobně vytvoříme dolní chybovou úsečku jen s tím rozdílem, že označíme dolní třetinu sloupcového grafu, volíme *minus* a pro zápornou chybovou hodnotu vložíme data z posledního řádku připravené tabulky (Q1 − Q0), kladnou chybovou hodnotou se nezabýváme.

Dále odstraníme výplň dolní třetiny sloupcového grafu:

- klikneme pravou myší na dolní třetinu grafu,
- zvolíme *formát datové řady, výplň, bez výplně, zavřít.*

Výsledek vidíme na obr. 3.



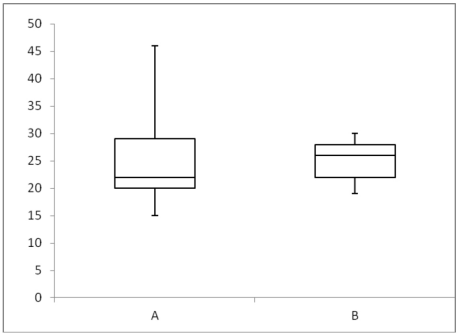
Obr. 3: Dolní chybová úsečka

Závěrečné úpravy

Přidáme ohraničení a odstraníme zbylé výplně:

- provedeme ohraničení horní a prostřední třetiny sloupcového grafu (pravou myší na příslušnou třetinu, dále přes *formát datové řady, barva ohraničení, plná čára* – barvu volíme černou, *styly ohraničení* – šířku volíme 1 bodů, *zavřít*),
- odstraníme výplň horní a prostřední části sloupcového grafu podobně jako v případě dolní třetiny. Na závěr provedeme konečnou úpravu:
- odstranění mřížky, legendy (kliknutím pravou myší na příslušný objekt a *odstranit*), eventuálně vložení popisků, nadpisů, pozadí a podobně.

Výsledný obrázek (obr. 4) je srovnatelný s box-plotem získaným například pomocí software STATISTICA.



Obr. 4: Výsledný box-plot

Závěr

Doporučujeme čtenářům, aby si prověřili postup pro uvedenou sadu dat a pak si vyzkoušeli zpracovat svoje vlastní data. Popsanou konstrukci jsme prováděli v Excelu 10, v jiných verzích je však postup stejný nebo se liší nepatrně.

Podrobnější popis využití box-plotů pro zobrazování dat, který lze s úspěchem využít i na střední škole, najdeme v práci (Cihlář, 2008). Dále lze doporučit aplet Mean and Median na adrese `http://illuminations.nctm.org/Activity.aspx?id=3576`, který na základě zadaných dat box-plot automaticky vykresluje. Žáci se tak mohou soustředit na zkoumání vlivu charakteru dat na jednotlivé statistické charakteristiky spíše než na tvorbu samotného grafu.

Literatura

[1] Cihlář, J. (2008). Krabicový graf (box-and-whisker graf) a jeho využití pro analýzu dat. *Excel Asistent Magazín*, 5(1), 2–21. Dostupné z `http://www.dataspectrum.cz/excelmag/download/eam0108.pdf`

[2] Charamza, P. & Hanousek, J. (1992). *Moderní metody zpracování dat: Matematická statistika pro každého*. Praha: Grada.

[3] Robová, J. (2004). Základy statistiky na grafickém kalkulátoru. *Matematika–fyzika–informatika*, 13(9), 555–563.

[4] Swoboda, H. (1977). *Moderní statistika*. Praha: Svoboda.

[5] Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Abstract

A set of statistical data can be presented using various characteristics (mean, mode, median, quartiles, variance, etc.) and various graphical means, too. The paper analyzes the advantages and disadvantages of various tools and their suitability and possible use in teaching. Namely, it disscuses the meaning of box-plots and provides guidance on how to create these charts using MS Excel program tools.

František Mošna
KMDM, Pedagogická fakulta UK v Praze
M. Rettigové 4
116 39 Praha 1 – Nové Město