



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



Univerzita Jana Evangelisty Purkyně v Ústí nad Labem
Pedagogická fakulta
Katedra preprimárního a primárního vzdělávání



VYBRANÉ KAPITOLY ZE STATISTICKÉ ANALÝZY EMPIRICKÝCH DAT

Trahorsch, Petr (Ed.)

Chytrý, Vlastimil

Nováková, Alena

Pavlátová, Věra

Ústí nad Labem 2019

Anotace: Učební materiál s názvem Vybrané kapitoly ze statistické analýzy empirických dat představuje vybrané metody kvantitativně orientovaných metod pedagogického výzkumu. Studenti se kromě teoretických poznatků vztahující se ke konkrétním metodám naučí i konkrétní postupy jejich výpočtu v programech MS Excel a Statistica. Učební text je doplněn celou řadou obrazových a tabelárních prvků, které umožňují čtenáři lépe pochopit obsah textu.

Klíčová slova: statistická analýza, kvantitativní metody, aplikace statistických metod, pedagogický výzkum, Statistica, Excel.

Text vznikl za podpory projektu s názvem „**Škola doktorských studií – aplikovaná a behaviorální studia**“ s registračním číslem CZ.02.2.69/0.0/0.0/16_018/0002727.

© Univerzita J. E. Purkyně v Ústí nad Labem

Autoři: Mgr. et Mgr. Petr Trahorsch (editor), PhDr. Vlastimil Chytrý, Ph.D., Mgr. Alena Nováková, Mgr. Věra Pavlátová

Obsah

Obsah	3
Předmluva.....	6
1. Deskriptivní statistika (analýza)	7
1.1. Grafický a číselný popis rozložení dat (základní tabulky a grafy)	7
Sloupcový graf (Bar chart)	9
Histogram.....	10
Spojnicový/čárový graf (link/line chart)	11
Bodový graf (scatter plot, korelační diagram)	11
Kruhový diagram (pie chart).....	14
Krabicový graf (boxplot).....	14
Třírozměrný graf.....	14
1.2. Míry centrální tendence.....	15
Střední hodnota.....	15
Vážený aritmetický průměr (weighted average).....	17
Geometrický průměr	18
Harmonický průměr	19
Medián.....	21
Modus	22
Používání měr centrální tendence	23
1.3. Míry rozptýlenosti (variability)	23
Variační rozpětí (šíře, R)	24
Rozptyl a směrodatná odchylka.....	24
Míry rozptýlenosti založené na empirických kvantilech	27
1.4. Intervaly spolehlivosti.....	29
1.5. Normální rozdělení.....	31
Testování normality.....	33
Testy normality.....	33
2. Typy proměnných.....	37

Závisle a nezávisle proměnné, rušivá proměnná.....	37
Diskrétní a spojité proměnné	38
Proměnné podle typu měřítka.....	38
3. Organizace dat (kódování).....	39
4. Zpracování Likertových škál	40
4.1. Transformace dat (spojitost s normálním rozdělením).....	41
Box-Coxova transformace	42
5. Detekce/odstraňování odlehlých hodnot a rezistentní odhady	45
5.1. Srovnání průměru a mediánu hodnot.....	46
5.2. Krabicový graf.....	48
5.3. Dean-Dixonův test.....	51
5.4. Grubbsův test	54
5.5. Rezistentní odhady	58
Useknutý průměr	58
Winsorizovaný průměr	58
6. Analýza závislostí.....	60
6.1. Zobrazování dvojrozměrných dat.....	60
6.2. Korelační analýza	64
Pearsonův korelační koeficient.....	67
Mnohonásobný koeficient korelace.....	69
Testování rozdílu mezi dvěma koeficienty korelace	70
Spearmanův koeficient korelace	71
Kendallův koeficient shody.....	75
Závislost mezi jevy zachycené nominálním měřením.....	77
Shrnutí korelační analýzy	78
6.3. Regresní analýza	78
Postup regresní analýzy v MS Excel	79
Shrnutí regresní analýzy	83
6.4. Vícenásobná regresní analýza	83

Základní postup vícenásobné regresní analýzy v MS Excel	84
Shrnutí vícenásobné regresní analýzy	93
Závěrem	94
Seznam tabulek.....	95
Seznam obrázků	96
Zdroje.....	98
Přílohy	101

Předmluva

Statistická analýza dat patří mezi základní metody kvantitativně orientovaného výzkumu. Při aplikaci těchto metod v pedagogickém výzkumu je nutné detailně znát výhody a úskalí jejich použití. Nežřídko se lze v závěrečných pracích studentů setkat s nevhodně použitými statistickými metodami (abychom studentům nekřivdili, podobné nedostatky vykazují i některé odborné články a publikace), což má za následek nevyhovující interpretaci získaných dat. Tento učební text tak charakterizuje a hodnotí vybrané metody statistické analýzy empirických dat; na rozdíl od jiných odborných textů se tato skripta zaměřují na konkrétní aplikaci vybraných metod kvantitativní analýzy; učební text uvádí i konkrétní příklady jejich aplikace. Studenti (nejen) pedagogických fakult, pro které je tento text určen, tak mohou dle příkladů uvedených v těchto materiálech lépe pochopit principy uvedených metod a potenciální interpretace získaných výsledků. V textu jsou dále uvedeny postupy výpočtů vybraných statistických testů ve dvou pravděpodobně nejpoužívanějších „statistických“ programech: MS Excel a Statistica. Studentům se tak dostává do rukou praktický návod k výpočtům vybraných koeficientů a testových kritérií pomocí informačních technologií, který jistě ocení při tvorbě svých závěrečných prací.

Za kolektiv autorů,

Petr Trahorsch

Ústí nad Labem, 15.3.2019

1. Deskriptivní statistika (analýza)

Statistika, jakožto „svět zisku a analýzy čísel“, se zabývá sběrem, prezentací, analýzou a interpretací dat. Můžeme ji rozdělit na deskriptivní (popisnou) a induktivní (inferenční). Deskriptivní statistika pracuje s daty, které jste ve vašem výzkumu získali. Tato data vám vhodně shrne a sumarizuje, aby bylo na první pohled jasné i ostatním, co vámi získaná data znamenají.

Metody induktivní statistiky vám umožní posoudit, zda to, co jste zjistili ve vašem vzorku probandů, lze zobecnit na celou skupinu, z níž byl vzorek vybrán.

Vraťme se tedy k deskriptivní statistice, která té induktivní předchází. Nejprve musíte získaná data vhodně numericky nebo graficky popsat. Není na tom nic složitého, určitě jste již s tabulkami a grafy pracovali. V deskriptivní analýze nás zajímá poloha a rozptýlení získaných dat na číselné ose, což podrobně popisuje následující text.

1.1. Grafický a číselný popis rozložení dat (základní tabulky a grafy)

Jednou z dovedností výzkumníka je vědět, jak nejlépe prezentovat získaná data (Walker, 2013, s. 90). Pro popis dat můžeme zvolit tabulky či grafy. Není ovšem vhodné prezentovat stejná data více nástroji; buď zvolíme tabulku, nebo data uvedeme jen v textu, nebo zvolíme graf. Volba grafu je ovšem klíčová; neřídíme se tím, který graf se nám nejvíce zamlouvá vizuálně, nýbrž tím, jaký typ dat chceme prezentovat. Pro data (proměnné) spojitá používáme jiné typy grafů než pro kategorické.

Spojité proměnné lze uspořádat, porovnávat, mění se spojitě, mohou mít jakoukoliv hodnotu – je to například výška, hmotnost, čas atd. **Kategorické proměnné** popisují omezený počet kategorií, které nemůžeme měnit, musíme jednotlivá data do určité kategorie zařadit – jedná se například o pohlaví, rodinný stav, dosažené nejvyšší vzdělání, formu studia atd.

Dalším důležitým bodem při prezentaci výsledků výzkumu je správný popis grafu či tabulky. Popis by měl být jasný a výstižný, srozumitelný i pro toho, kdo nečetl text vašeho výzkumu, ale podíval se jen na výsledky. Pro rozlišení tabulek a obrázků (sem patří i grafy) je popisujeme arabskými číslicemi (výjimečně lze použít římská čísla). Na konci své práce nezapomeňte uvést seznam obrázků a tabulek.

Tab. 1 Popis os používaných v 2D grafech

Obyčejné označení	Název souřadnice	Alternativní název
vodorovná osa	osa x	abscisa
svislá osa	osa y	ordináta

Obyčejné označení	Název souřadnice	Alternativní název
vodorovná osa	osa x	abscisa
svislá osa	osa y	ordináta

Obyčejné označení	Název souřadnice	Alternativní název
vodorovná osa	osa x	abscisa
svislá osa	osa y	ordináta

Obyčejné označení	Název souřadnice	Alternativní název
vodorovná osa	osa x	abscisa
svislá osa	osa y	ordináta

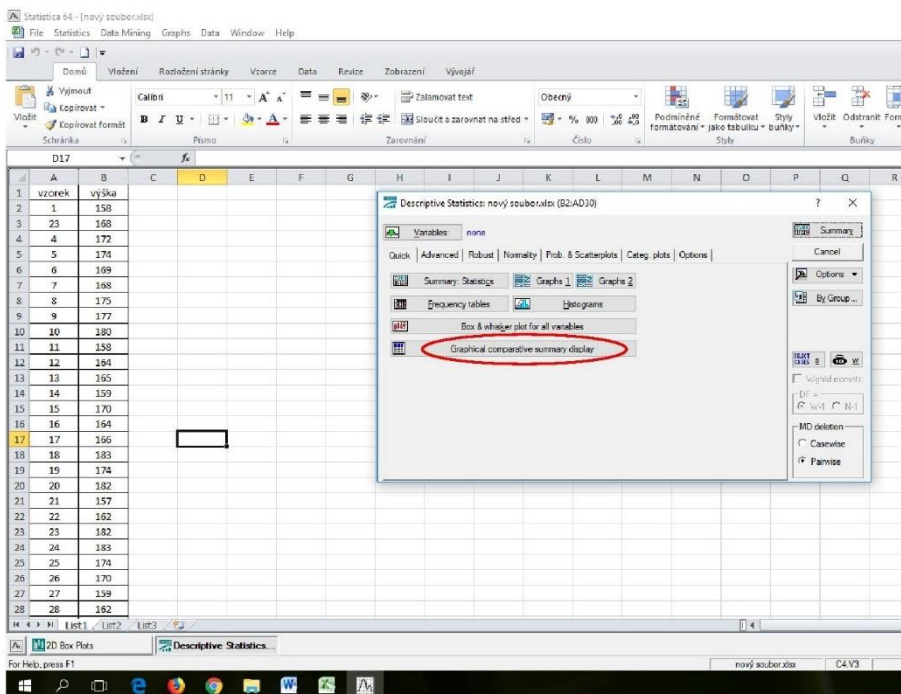
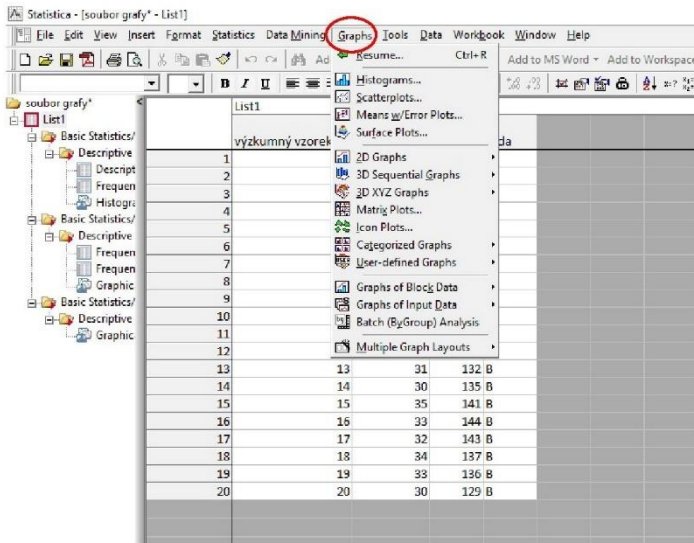
Zdroj: Walker (2013, s. 84)

Grafy nám usnadní práci s daty – pomohou prověřit jejich polohu, časový průběh, míru kolísání, vztah mezi dvěma či více proměnnými.

Kde najdete znázornění grafu na PC?

MS Excel: zadat data do tabulky → označit data → vložení → grafy

Statistica: načíst data → Graphs (je umístěno přímo na hlavní liště, viz obr. 1) nebo základní statistiky (basic statistics) → popisné statistiky (descriptive statistics) → Graphical comparative summary display (Graphical Summary, viz obr. 1)

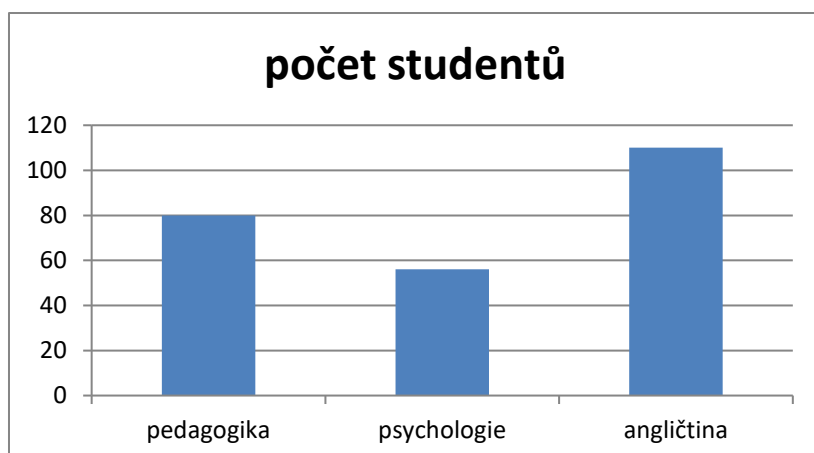


Obr. 1 Umístění grafů v programu Statistika

Zdroj: autoři

Sloupcový graf (Bar chart)

Tento typ grafu má na ose x vždy znázorněnu kategoričnou proměnnou, zatímco na ose y spojitou proměnnou. Sloupce se v tomto grafu nikdy nedotýkají. Porovnáváme jen výšky sloupců u jednotlivých kategorií, přičemž výška může reprezentovat četnost výskytu, nebo také vypočtený průměr, směrodatnou odchylku apod. Šířka sloupce nehraje žádnou roli.



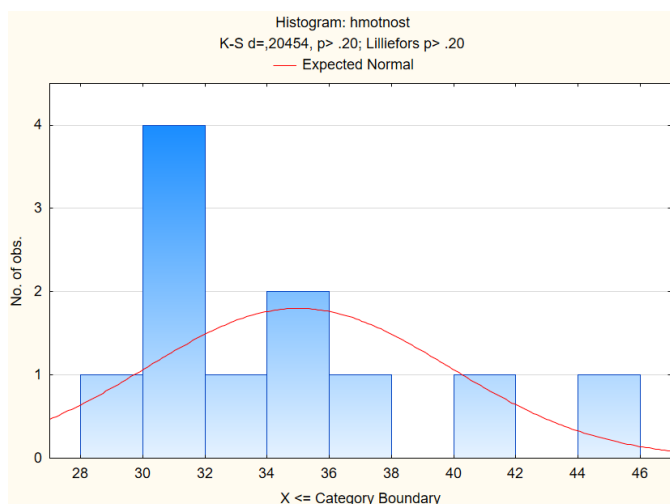
Obr. 2 Počet studentů jednotlivých oborů na fakultě

Zdroj: autoři

Histogram

Histogram poprvé představil na své přednášce v roce 1891 Pearson. Jedná se o grafickou verzi souboru spojitých dat (zachycujících čas, délku, teplotu...), která ukazuje počet případů spadajících do jednotlivých pravoúhlých sloupců, jež se nacházejí těsně u sebe (Magnello, 2010. s. 83). Slouží ke grafickému vyjádření velkých souborů spojitých dat (> 50).

Histogram má na obou osách spojitou proměnnou. Na osu x vyneseme intervaly proměnné, kterou jsme zjišťovali (hmotnost, výška, počet odpovědí, body v kognitivním testu) a osa y nám znázorňuje, jak často jsme určitý výsledek zjistili – četnost. Získáme tak rozdělení četnosti a jednotlivé sloupce se zde, na rozdíl od předchozího grafu, budou dotýkat. Výška sloupců je dána četností pozorování na daném intervalu vymezeným šířkou (viz obr. 3).

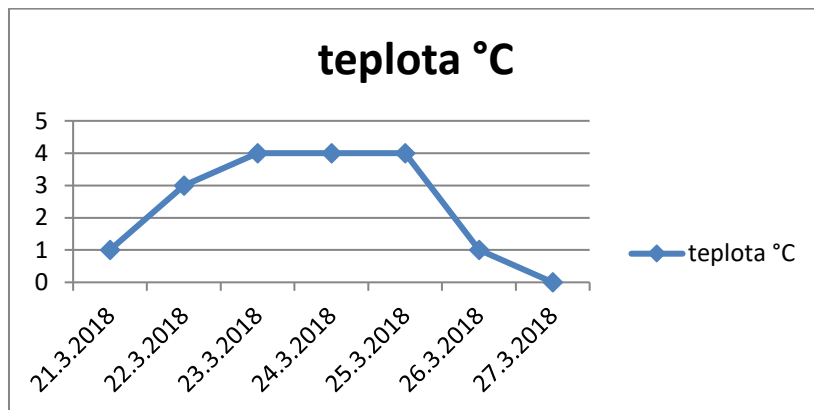


Obr. 3 Četnost hmotnosti u dvanáctiletých chlapců v rámci jedné třídy

Zdroj: autoři

Spojnicový/čárový graf (link/line chart)

Používá se ke znázornění toho, jak se data mění napříč kategoriemi nebo podél nějaké řady. Svislá osa *y* vždy vyjadřuje nějaký druh výsledku (teplota, tlak, výsledné body) nebo počet lidí jako ve sloupcovém grafu, ale osa *x* znázorňuje logickou posloupnost (obr. 4).

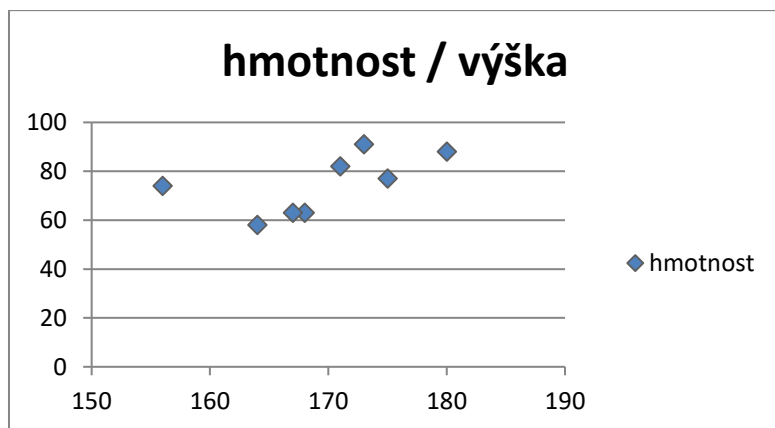


Obr. 4 Znáznornění polední teploty za sedm dní v Krupce

Zdroj: autoři

Bodový graf (scatter plot, korelační diagram)

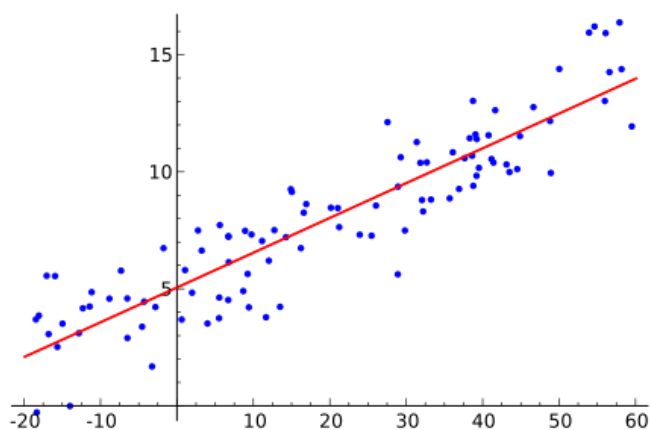
Tento graf se od ostatních liší v tom, že má na obou osách spojité proměnné. Používá se například k tomu, aby znázornil, jakého výsledku dosáhli probandi při dvou současných měřeních (viz kap. 6). Jednotlivého probanda pak představuje určitý bod v grafu (obr. 5).



Obr. 5 Znáznornění výšky a hmotnosti probandů

Zdroj: autoři

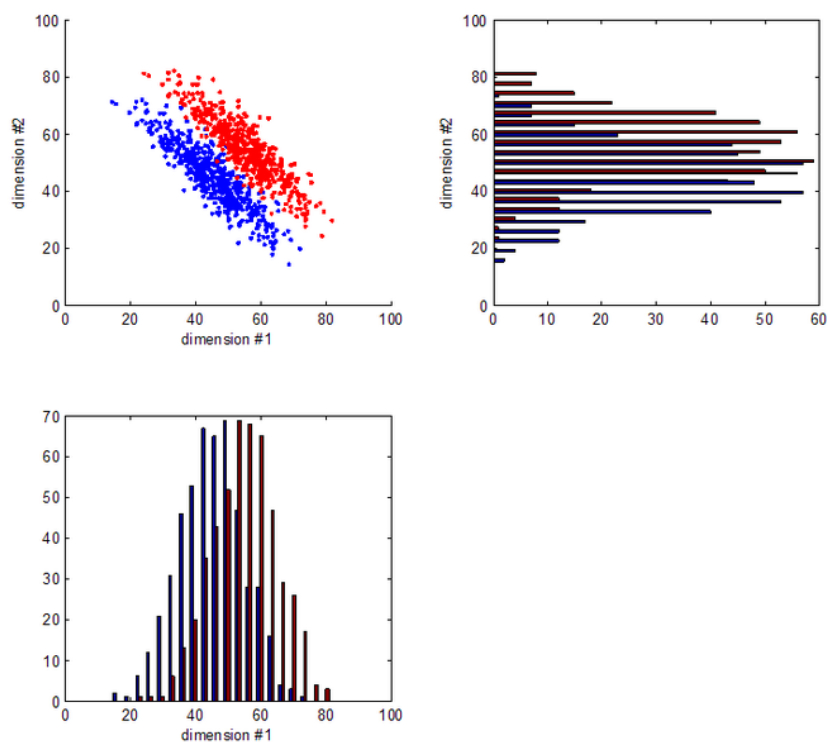
Bodové grafy (též korelační diagramy) se také využívají například při korelacích k tomu, abychom mohli sledovat vztah mezi dvěma odlišnými znaky v souboru (Walker, 2013, s. 87). Můžeme tak zjistit vzájemný vztah mezi oběma proměnnými, případně tuto závislost interpolovat (vylepšit vložením, vyjádřit) přímkou, křivkou, nebo jiným typem závislosti (viz obr. 6).



Obr. 6 Korelační diagram

Zdroj: Sewaqu (2010)

Bodový graf (Scatter plot) můžeme použít i pro kombinaci dat z více zdrojů. Interpretace a prezentace dat z tohoto typu grafu je pak přesnější a efektivnější, než kdybychom použili jednotlivé grafy (viz obr. 7).

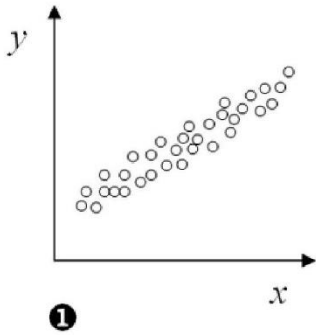


Obr. 7 Interpretace dat z více zdrojů

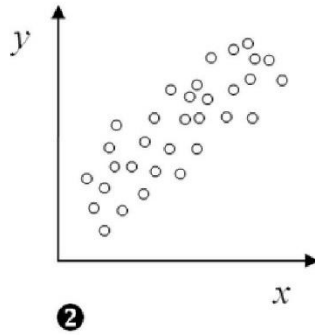
Zdroj: Jarekt (2010)

Při práci s korelacemi pozorujeme vztah mezi dvěma hodnotami na bodových grafech (korelačních diagramech). Tyto grafy nám pak usnadní naši interpretaci (viz obr. 8 a 9).

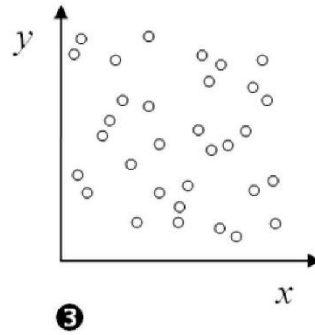
1. Silná korelace



2. Slabá korelace



3. Žádná korelace mezi proměnnými x a y

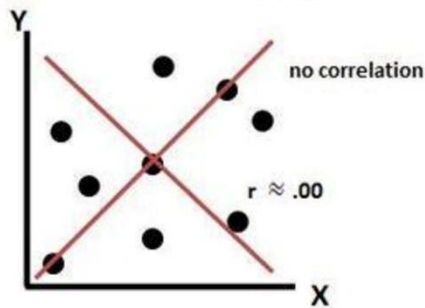


Obr. 8 Ukázka korelací

Zdroj: Atanassova (2010)

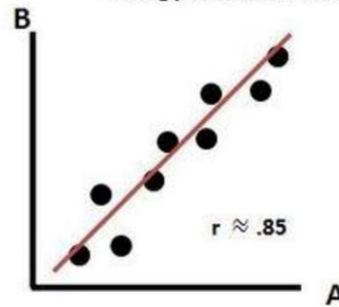
Mezi hodnotami neexistuje vztah.

Correlation – Linear $-1 \leq r \leq 1$



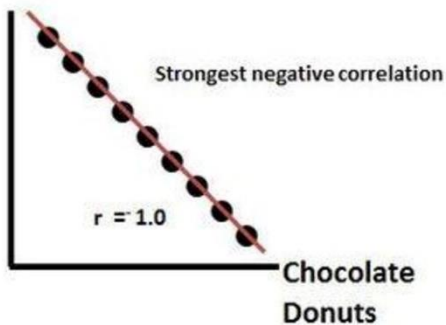
Body vyskytující se blízko přímky ukazují na silnou korelaci. Pokud tato přímka stoupá zleva doprava, jde o kladný vztah – pozitivní.

Strong positive correlation



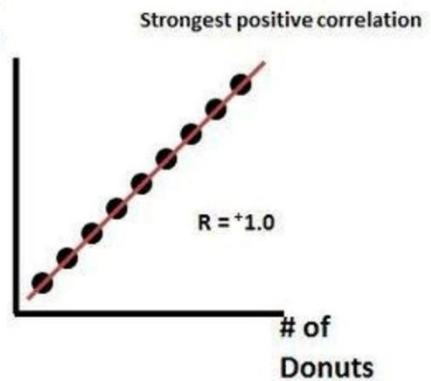
Zde body klesají zleva doprava, jde o nejsilnější záporný vztah (negativní korelaci). Jak se jedna naměřená hodnota zvedá, má ta druhá tendenci klesat.

Powdered Donuts



Zde je nejsilnější pozitivní korelace.

Cost of Donuts

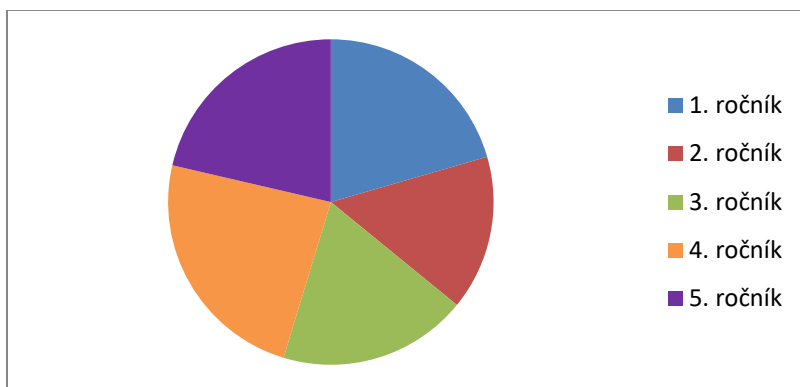


Obr. 9 Ukázka korelací včetně negativní korelace

Zdroj: Dwendland (2013)

Kruhový diagram (pie chart)

Kruhový diagram¹ nám zobrazuje podíl zastoupení jednotlivých kategorií vzhledem k celku. Například k výzkumu si vybereme žáky pěti různých ročníků a na kruhovém diagramu znázorníme rozložení jednotlivých ročníků ve výzkumném souboru a jejich četnost (šířky výseků). Těchto diagramů nepoužíváme k zobrazení výsledků jednotlivých skupin.



Obr. 10 Rozložení jednotlivých ročníků ve výzkumném souboru

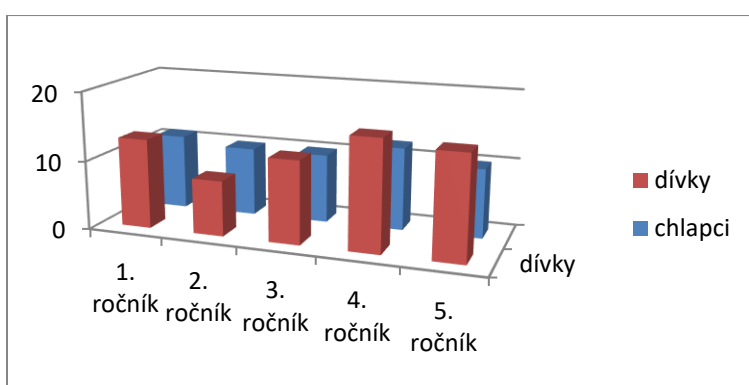
Zdroj: autoři

Krabicový graf (boxplot)

Na tomto typu grafu nejlépe rozpoznáme kvartily, interkvartilové rozpětí, medián, odlehlé hodnoty, symetrii. Podrobně se tomuto typu grafu budeme věnovat v kapitole 1.2 a 1.3

Třírozměrný graf

Pro zobrazení dat používejte spíše dvourozměrné grafy, třírozměrné použijte v případě zakreslení tří proměnných naráz.



Obr. 11 Genderové rozložení jednotlivých ročníků ve výzkumném souboru

Zdroj: autoři

¹ Velmi často se pro tento typ vizuálie užívá pojem koláčový graf, což je nepřesné označení. Vhodnější označení pro tento typ vizuálie je diagram, protože na rozdíl od jiných grafů, mu chybí osa x a osa y. Jelikož má tento diagram kruhový tvar, nazýváme ho kruhovým diagramem (v případě, že ho pojmenujeme koláčový diagram, lze ho s nadsázkou pojmenovat také jako pizzový diagram, lívancový diagram apod.).

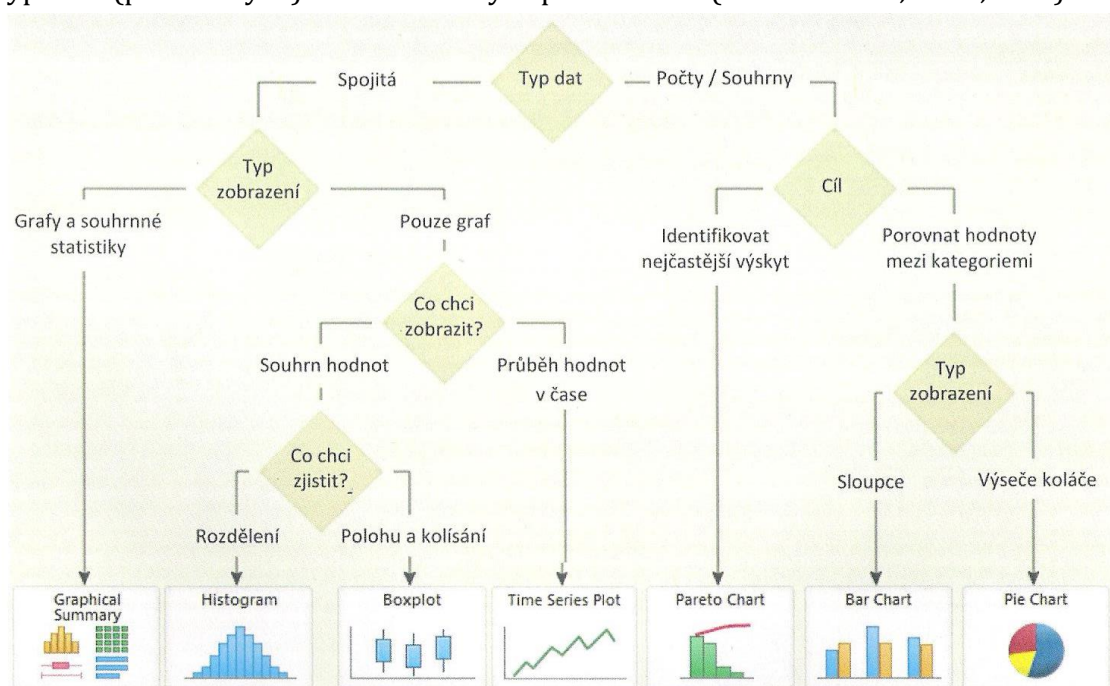
Další grafy (jejich užití bude uvedeno přímo v souvislosti s konkrétním testem)

Ke grafickému vyjádření menších souborů spojitých dat (do 50) používáme tečkový graf (dotplot), což je obdoba histogramu.

Pro jednotlivá měření spojitých dat v časové závislosti používáme časové řady (time series plot).

Pro kategorická data, kdy chceme porovnat četnost výskytu v sestupném pořadí (například počet propadajících žáků v jednotlivých ročnících), použijeme Paretův graf (pareto chart).

Další obrázek (obr. 12) nám shrnuje možnosti použití jednotlivých typů grafů v závislosti typu dat (proměnných) a na nastavených podmínkách (SC&C Partner, 2015, s. 20).



Obr. 12 Možnosti použití jednotlivých typů grafů

Zdroj: SC&C Partner (2015, s. 20)

1.2. Míry centrální tendence

Míry centrální tendence se zabývají polohou získaného data na číselné ose (vzhledem k ostatním naměřeným hodnotám), nejčastěji popisují *střední bod*, dle Hendla (2015) *typickou hodnotu dat*.

Střední hodnota

Nejnámější mírou centrální tendence je *střední hodnota*, o níž většina lidí nejčastěji mluví, když mluví o „průměru“ (Walker, 2013, s. 69). Zajisté i vy jste již ve své praxi aritmetický průměr (\bar{x} ; average, mean) počítali; sečetli jste dohromady získaná data a vydělili jejich počtem. Dospěli jste tedy k němu součtem všech hodnot v určitém souboru dat x a jeho vyděle-

ním celkovým počtem případů n (Magnello, 2010, s. 67). Můžeme si ho představit jako těžší-tě dat; součet podprůměrných hodnot je stejný jako součet nadprůměrných hodnot a oba součty jsou v rovnováze (Budíková et al., 2010, s. 42). Výhodou aritmetického průměru je především to, že jeho matematické vyjádření je jednoduché, je použitelný při odvozování dalších důležitých vztahů, jeho hodnota závisí na všech prvcích souboru dat. Nevýhodou je jeho citlivost k tzv. extrémním (odlehlym) hodnotám, tj. hodnotám, které se od ostatních značně odchyľují (Chráska, 2016, s. 42).

Uvedeme si příklad. Máme naměřené tyto hodnoty hmotností osmiletých chlapců (tab. 2).

Tab. 2 Hmotnost žáků

žák / prvek výběru	1	2	3	4	5	6	7	8
hmotnost / m [kg]	31	35	30	32	31	31	30	28

Zdroj: autoři

Poznámka: součet (Σ) všech hmotností : počet prvků výběru (tj. rozsah statistického souboru = n)

$$\bar{m} = \frac{\sum_{i=1}^n m_i}{n} \rightarrow \bar{m} = \frac{\sum_{i=1}^8 m_i}{8} \rightarrow \bar{m} = \frac{248}{8} \rightarrow \bar{m} = 31 \text{ kg}$$

Můžeme použít i tento ekvivalent vzorce:

$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

Průměrná hmotnost vybraného souboru (výběru) je tedy 31 kg. Co udělá s průměrem (střední hodnotou) přidání další jednotky do výběru, kterým bude obézní, 85 kg vážící osmiletý chlapec?

$$\bar{m} = \frac{333}{9} \rightarrow \bar{m} = 37 \text{ kg}$$

Tato odlehlá hodnota nám průměrnou hmotnost vybraného vzorku zvýší o 6 kg. Aritmetický průměr v tomto případě nemá příliš vypovídací hodnotu o centrální hodnotě, je lepší zvolit jinou míru centrální tendence (medián). Pro praxi je proto důležité detekovat v daném souboru odlehlé hodnoty.

Pokud máme statistický soubor určitých jednotek, u kterých je předem znám počet různých hodnot statistického znaku (r , v našem případě $r = 5$), využíváme k výpočtu aritmetického průměru četnosti (v , řecké ný; jde o počet, kolikrát se hodnota statistického znaku ve statistickém souboru vyskytuje). Například máme 20 žáků (statistický soubor o rozsahu $n = 20$), kdy analyzujeme jejich známky z testu (statistický znak) a víme, že známky se mohou pohy-

bovat v intervalu 1–5 (konkrétní hodnoty statistického znaku, označujeme x^*). Četnosti (absolutní) v znázorňuje tab. 3.

Tab. 3 Rozložení četností hodnot statistického znaku

x^*	1	2	3	4	5
v	5	5	4	3	3

Zdroj: autoři

Aritmetický průměr vypočítáme podle tohoto vzorce:

$$\bar{x} = \frac{\sum_{j=1}^r v_j \cdot x_j^*}{n} = \frac{\sum_{j=1}^5 v_j \cdot x_j^*}{20} = \frac{5 \cdot 1 + 5 \cdot 2 + 4 \cdot 3 + 3 \cdot 4 + 3 \cdot 5}{20} = \frac{5 + 10 + 12 + 12 + 15}{20} = \frac{54}{20} = 2,7$$

Můžeme říci, že tento zápis je jen jinou formou váženého průměru, kde jako váhy slouží četnosti.

Vážený aritmetický průměr (weighted average)

Tento typ průměru se používá pokud:

- hodnoty statistického znaku nemají stejnou váhu
- hodnoty statistického znaku jsou již určitým způsobem zatříděny (viz tab. 3)
- máme-li spočítat celkový průměr z několika podsouborů o různém počtu hodnot

$$\bar{x} = \frac{\sum_{i=1}^n x_i \cdot w_i}{\sum_{i=1}^n w_i}$$

Pokud budeme dále pracovat s naším příkladem (tab. 3), x_i je jednotka hodnoty (určitá známka z textu) a w_i jsou váhy (počet žáků, kteří dostali určitou známku – jedná se také o absolutní četnost). Hodnoty přepíšeme do tab. 4.

Tab. 4 Rozložení četností hodnot statistického znaku pro vážený průměr

známka = $x^* = x_i$	počet žáků = $w_i = v$ = absolutní četnost	váženo $x_i \cdot w_i$	vážený průměr
1	5	5	-
2	5	10	-
3	4	12	-
4	3	12	-
5	3	15	-
součet	20	54	54:20 = 2,7

Zdroj: autoři

Došli jsme tedy ke stejnému výsledku. Nyní do tab. 5 přidáme ještě relativní četnosti (relativní váhy), což je podíl z celkového počtu (počet žáků, kteří dostali určitou známku, vydělíme celkovým počtem žáků).

Tab. 5 Rozložení relativních četností hodnot statistického znaku pro vážený průměr

známka = $x^* = x_i$	počet žáků = $w_i = v$ = absolutní četnost	váha = relativní četnost = w_i	vážený průměr
1	5	$5:20 = 0,25$	$1.0,25 = 0,25$
2	5	$5:20 = 0,25$	$2.0,25 = 0,50$
3	4	$4:20 = 0,20$	$3.0,20 = 0,60$
4	3	$3:20 = 0,15$	$4.0,15 = 0,60$
5	3	$3:20 = 0,15$	$5.0,15 = 0,75$
součet	20	1	2,7

Zdroj: autoři

Použijeme-li relativní četnosti jako váhy, pak se jejich součet rovná jedné a vážený průměr je roven součtu $x_i \cdot w_i$.

Geometrický průměr

V některých situacích (např. tehdy, jestliže chceme postihnout tempo růstu v určité oblasti), se místo aritmetického průměru počítá tzv. geometrický průměr \bar{x}_G (Chráška, 2016, s. 42).

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Všechny hodnoty (jejich součet je n a hodnoty musí být nezáporné) se mezi sebou vynásobí a z jejich součinu vytvoříme n -tou odmocninu (je to totéž, pokud jejich součin umocníme na $\frac{1}{n}$). Uplatnění nachází u přírůstkových či růstových veličin. Například chceme zjistit průměrné tempo růstu bitcoinu za posledních 5 měsíců. Víme, že přírůstky, či poklesy jsou následující: 3%, 5%, -2% (pokles), 1%, 4%. Abychom se vyvarovali záporným hodnotám, vztáhneme vše k původním 100 % (situaci před 5 měsíci) a všechny změny vynásobíme:

$$1,03 * 1,05 * 0,98 * 1,01 * 1,04 = 1,113$$

$$\sqrt[5]{1,113} = 1,113^{\frac{1}{5}} = 1,021$$

Odečteme původních 100 % (= 1) a průměrné tempo růstu Bitcoinu za posledních 5 měsíců je tedy **2,1 %**. Geometrický průměr je vždy menší nebo roven aritmetickému průměru.

Harmonický průměr

Jedná se o převrácenou hodnotu aritmetického průměru převrácených hodnot zadaných členů, tedy o podíl rozsahu souboru (počtu členů) a součtu převrácených hodnot znaků. Používá se tam, kde porovnáváme určité znaky na stejných úsecích, například ke zjištění průměrné délky času nutné k provedení nějakého úkonu, kdy jsou dané úkoly prováděny současně několika osobami či stroji.

$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Například v továrně pracují současně dva stroje, starší vyrobí součástku za 4 minuty, novější stroj za 2 minuty. Jak dlouho trvá v průměru příprava jedné součástky?

$$\bar{x}_H = \frac{2}{\frac{1}{4} + \frac{1}{2}} = \frac{2}{\frac{3}{4}} = 2,67 \text{ minut}$$

Harmonický průměr je vždy menší nebo roven geometrickému průměru.

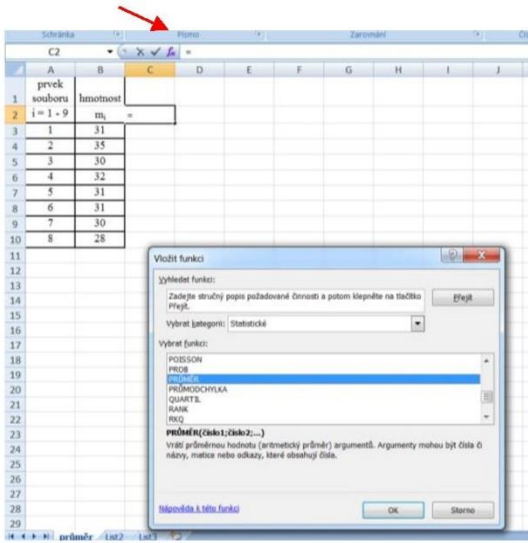
Kde najdete výpočet střední hodnoty (aritmetického průměru) na PC?

MS Excel: vložit funkci → statistické → průměr

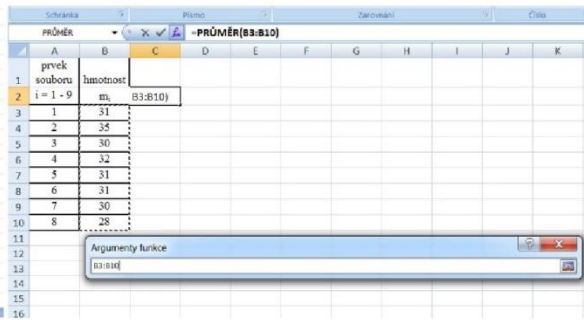
Statistica: statistics → základní statistiky (basic statistics) → popisné statistiky (descriptive statistics) → summary statistics → vybrat proměnnou → ok

Obr. 13 nám zachycuje postup výpočtu průměru v MS Excelu u souboru z tab. 2. Stejným způsobem vypočítáte v MS Excelu i následující míry, jako je medián, modus a směrodatná odchylka. Výpočty ve Statistice.cz nám také znázorňuje obr. 13.

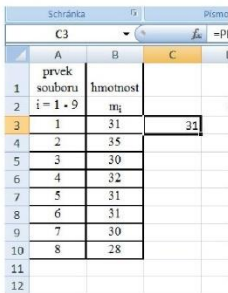
Excel: klikněte do libovolného políčka mimo tabulku, klikněte na: vybrat funkci f_x , vyberte kategorii statistické, průměr, ok



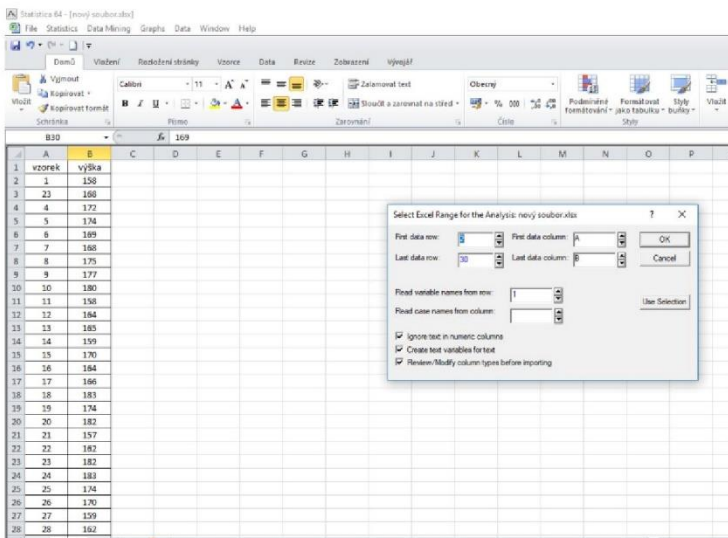
do políčka „argumenty funkce“ si levým tlačítkem myši označte (přetáhněte) celý sloupec čísel, u kterých chcete počítat průměr (zobrazí se vám souřadnice prvního a posledního čísla; B3: B10), dejte enter a potom ok



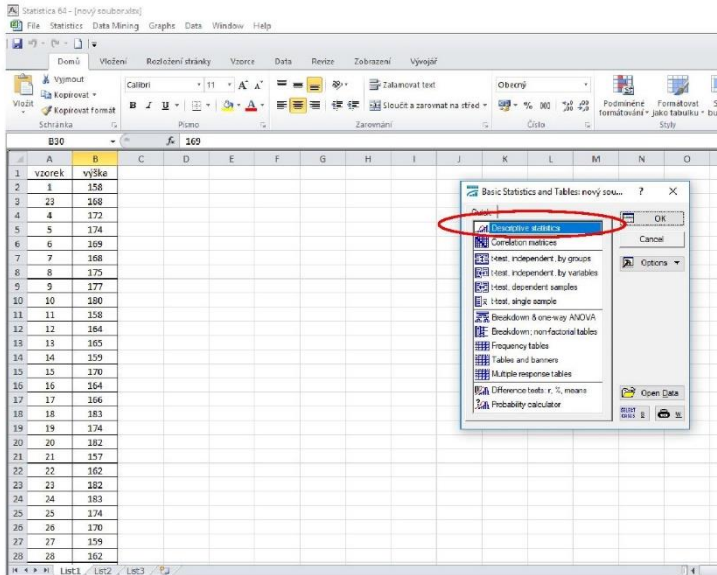
a objeví se vám v políčku, na které jste původně klikli, výpočet průměru (31)



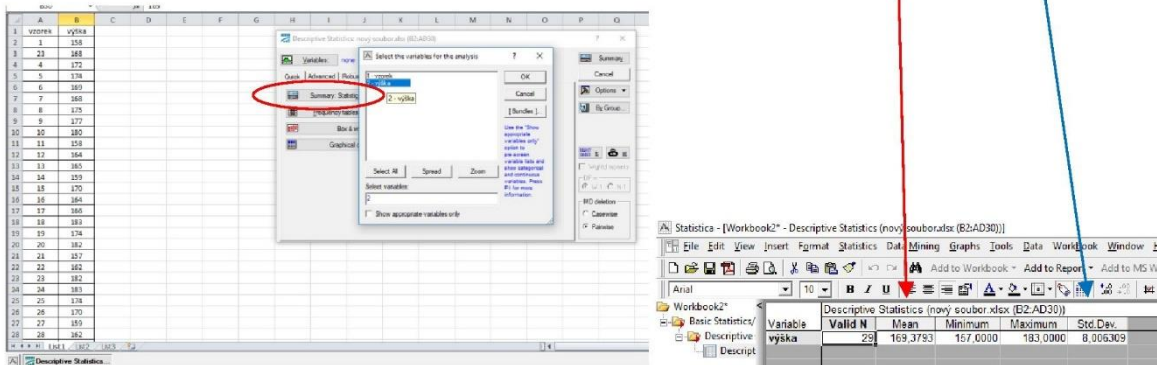
Statistica: načteme si data



vybereme proměnnou (výšku) → statistics → základní statistiky (basic statistics) → popisné statistiky (descriptive statistics)



summary statistics → vybrat proměnnou (výšku) → ok → ukáže se nám průměr i směrodatná odchyłka



Obr. 13 Postup výpočtu průměru v programu Excel a Statistica

Zdroj: autoři

Medián

Medián popisuje střední bod v řadě čísel ještě jednodušeji než aritmetický průměr; seřadíte svá naměřená data od nejnižšího po nejvyšší a medián naleznete přímo uprostřed. V případě sudého počtu naměřených dat je medián průměrem dvou prostředních hodnot (Walker, 2013, s. 71). Medián tedy udává prostřední hodnotu (respektive průměr dvou prostředních hodnot) v uspořádaném datovém souboru (Budíková et al., 2010, s. 41). Medián je bod, který dělí dané rozdělení na spodní polovinu a horní polovinu, a to tak, že 50 % hodnot se nalézá v

jedné polovině a 50 % ve druhé. Medián použil jako první v roce 1816 Gauss, do statistiky jej zavedl Galton (Magnello, 2010, s. 68).

Výhodou mediánu je, že není citlivý k extrémním hodnotám a jeho výpočet je možný i v případech, kdy o prvcích souboru dat nemáme úplné informace. Na rozdíl od aritmetického průměru, není nutné znát všechny hodnoty souboru (Chráska, 2016, s. 44). Podíváme se na hodnoty hmotností osmiletých chlapců, známé již z výpočtu průměru.

U prvního souboru dat seřadíme naměřené hodnoty: 28, 30, 30, 31, 31, 31, 32, 35 a zjistíme, že medián je 31 kg. Hodnota mediánu se v tomto případě shoduje s hodnotou průměru; $\tilde{m} = \bar{m} = 31 \text{ kg}$.

Seřadíme i naměřené hodnoty druhého souboru: 28, 30, 30, 31, 31, 31, 32, 35, 85 a zjistíme, že medián je opět 31 kg. Odlehlá hodnota neměla na hodnotu mediánu na rozdíl od hodnoty průměru žádný vliv.

Pokud máme data, která nejsou symetrická, tak potom platí, že:

- medián je odlišný od aritmetického průměru;
- medián má lepší vypovídací hodnotu o centrální tendenci.

Pokud jsou data symetrická, tak potom se průměr shoduje s mediánem. Symetrii či nesymetrii dat nejlépe zjistíme pomocí histogramu (kapitola 1.1).

Kde najdete výpočet mediánu na PC?

MS Excel: vložit funkci → statistické → median

Statistica: základní statistiky (basic statistics) → popisné statistiky (descriptive statistics)

Postup bude stejný, jako znázorňuje obr. 13.

Modus

Modus (vrchol nebo modální hodnota) je ta nejčastěji vyskytující se hodnota v souboru dat. Jedná se tedy o hodnotu, která se objevuje častěji než ostatní, je to bod největší (maximální) četnosti (Magnello, 2010, s. 71). Výhodou modu je jeho nezávislost na extrémních hodnotách měřené veličiny. Slouží jen jako provizorní charakteristika polohy a neumožňuje další statistickou analýzu. Modus je možno počítat u dat nominálních, ale i ordinálních či metrických (Chráska, 2016, s. 46).

U našeho vzorového souboru dat: 28, 30, 30, 31, 31, 31, 32, 35 je modus 31 kg.

Kde najdete výpočet modu na PC?

MS Excel: vložit funkci → statistické → mode

Statistica: základní statistiky (basic statistics) → popisné statistiky (descriptive statistics)

Postup bude stejný, jako znázorňuje obr. 13.

Používání měr centrální tendence

Existují tři hlavní míry centrální tendence; střední hodnota, medián a modus. Každá z nich vám poskytne odlišný způsob popisu řady dat. Střední hodnota a medián odvádějí prakticky stejnou práci; obě hledají v řadě dat střední bod. Pokud vaše data neobsahují odlehlé hodnoty (extrémně nízké nebo vysoké oproti ostatním), obvykle použijete střední hodnotu, protože ta bere v potaz všechna data, takže obsahuje více informací a je tedy vhodnější pro sumarizaci dat (Walker, 2013, s. 72).

Modus použijete, pokud zjistíte, jak často se odehrává či stává určitý jev a vy chcete najít ten nejběžnější, abyste s ním pak mohli dále pracovat. Například zjistíte, že nejčastějším kupujícím určitého produktu je žena ve věku 35–45 let. Další reklamní kampaně tedy můžete cílit přímo na tento soubor.

- Dle Hendla (2015) se *průměr* má použít:
 - pokud jsou data získána minimálně v intervalovém měřítku (nepoužíváme tedy pro kategoriální data);
 - pokud je rozdělení symetrické;
 - pokud chceme použít statistické testy.
- *Medián* se má použít:
 - pokud jsou data získána minimálně v ordinálním měřítku;
 - pokud chceme znát střed rozdělení dat;
 - pokud mohou data obsahovat odlehlé hodnoty;
 - pokud je rozdělení dat silně zešikmené.
- *Modus* se má použít:
 - pokud má rozdělení více vrcholů;
 - pokud chceme získat o rozdělení jenom základní přehled;
 - pokud se slovem „průměrně“ míní nejčastější hodnota.

1.3. Míry rozptýlenosti (variability)

Náhodně proměnlivé údaje nestačí charakterizovat jenom střední hodnotou, data se stejnou střední hodnotou mohou mít různou rozptýlenost (Hendl, 2015, s. 101). Informaci o tom, jak

dalece jsou jednotlivé hodnoty kolem střední hodnoty nakupeny (či naopak rozptýleny), vyjadřují tzv. míry variability (míry rozptýlenosti nebo charakteristiky rozptýlení) – (Chrásková, 2016, s. 46). Mezi ně patří variační rozpětí, rozptyl, směrodatná odchylka a míry rozptýlenosti založené na empirických kvantilech.

Máme například tyto dva soubory hmotností, jejichž střední hodnoty jsou totožné, ale liší se rozptylem dat.

$$S_1: 33, 33, 35, 36, 36 \quad \bar{m}_1 = 34,6 \text{ kg}$$

$$S_2: 17, 28, 35, 40, 53 \quad \bar{m}_2 = 34,6 \text{ kg}$$

To znamená, že pokud máme soubory dat, mohou být koncentrovány kolem nějaké hodnoty, nebo mohou být rozptýleny na širší škále. Způsobem, jak tento rozptyl zjistíme, se budeme zabývat v následujících kapitolách.

Variační rozpětí (šíře, R)

Pro přibližné posouzení rozptýlení hodnot (posouzení variability) můžeme vypočítat variační rozpětí neboli variační šíři R . Je to rozdíl mezi největším a nejmenším získaným datem (hodnotou). Jeho nevýhodou je velká citlivost k odlehlým hodnotám.

$$R = x_{max} - x_{min}$$

Pro náš modelový příklad máme dva soubory dat – jedná se o hmotnosti osmiletých chlapců ze třídy 3.A a 3.B. Z tab. 6 vidíme, že ačkoliv mají soubory stejný průměr, jejich variační rozpětí se velmi liší (můžete si zakrýt pravou část tabulky a zkusit si vypočítat míry centrální tendence a rozpětí sami).

Tab. 6 Hmotnosti osmiletých chlapců [kg] ze tříd 3.A a 3.B

soubor/ prvek výběru	1	2	3	4	5	6	7	8	9	M \bar{m}	Me \tilde{m}	Mo \hat{m}	R
1 (3.A)	31	35	30	32	31	31	30	28	85	37	31	31	57
2 (3.B)	38	36	37	35	38	38	37	36	38	37	37	38	3

Zdroj: autoři

Rozptyl a směrodatná odchylka

Směrodatná odchylka (standard deviation) je nejběžnější míra rozptýlení, se kterou budete ve statistice pracovat. Je to „typická míra, o níž se každé číslo v souboru odlišuje od střední hodnoty“ (Walker, 2013, s. 75). Popisuje nám, jakým způsobem jsou dané hodnoty rozptýlené. Soubory s velkým variačním rozpětím mají i větší hodnoty směrodatné odchylky. Smě-

rodatnou odchylku si můžeme vypočítat sami, nebo využít PC. Pokud počítáte sami za využití modelového příkladu, nejprve si od každého prvku souboru odečtete průměr, abyste zjistili, o kolik se liší hmotnost konkrétního chlapce od průměru, tedy od 37. Tento rozdíl umocníte na druhou a sečtete všechny výsledky. Pak vydělíte číslem, které získáte z počtu prvků výběru poníženého o 1 ($9-1=8$).

Tab. 7 Manuální výpočet směrodatné odchylky

prvek souboru $i = 1 - 9$	hmotnost m_i	odchylka od střední hodnoty ($\bar{m} = 37$) $m_i - \bar{m}$	odchylka na druhou $(m_i - \bar{m})^2$
1	38	1	1
2	36	-1	1
3	37	0	0
4	35	-2	4
5	38	1	1
6	38	1	1
7	37	0	0
8	36	-1	1
9	38	1	1
součet odchylek na druhou			10

Zdroj: autoři

Jak postupujete:

1. Vypočtete střední hodnotu analyzovaných čísel (pro nás 37).
2. Odečtete ji od každého čísla.
3. Získaný rozdíl umocníte na druhou.
4. Tyto druhé mocniny sečtete.
5. Součet vydělíte číslem o jedno nižším, než kolik je prvků v souboru ($9-1 = 8$).
6. Vypočtete druhou odmocninu a získáte směrodatnou odchylku, v našem případě tedy $10:8 = 1,25 \rightarrow \sqrt{1,25} \doteq 1,12$ **$s \doteq 1,12$** .

Většinou ve svých výzkumech pracujeme s výběrovým souborem (VS) a směrodatnou odchylku tedy počítáme tak, jak bylo naznačeno. Takto vypadá vzorec pro její výpočet:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Při zpracování na PC v programu MS Excel jde o funkci STDEVA, nebo SMODCH.VÝBĚR.

Pokud pracujeme s celým základním souborem (ZS), pak se směrodatná odchylka označuje σ a počítá dle tohoto vzorce:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

Při zpracování na PC v programu MS Excel jde o funkci SMODCH.

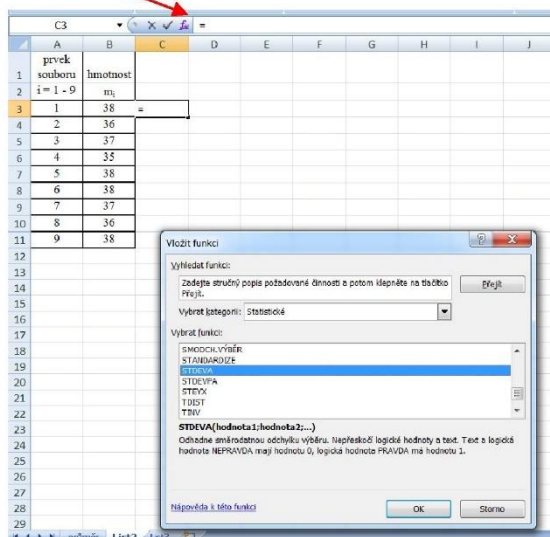
Rozptyl (variance) se používá především v inferenční statistice při výpočtu různých testovacích statistik (Hendl, 2015, s. 102). Jde vlastně o umocněnou směrodatnou odchylku, s^2 , nebo σ^2 .

Kde najdete výpočet směrodatné odchylky na PC?

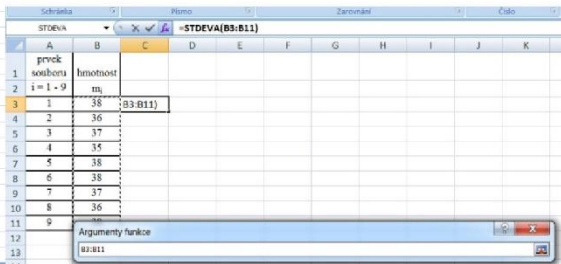
MS Excel: vložit funkci → statistické → STDEVA, nebo SMODCH.VÝBĚR

Statistica: statistics → základní statistiky (basic statistics) → popisné statistiky (descriptive statistics) → summary statistics → vybrat proměnnou → ok

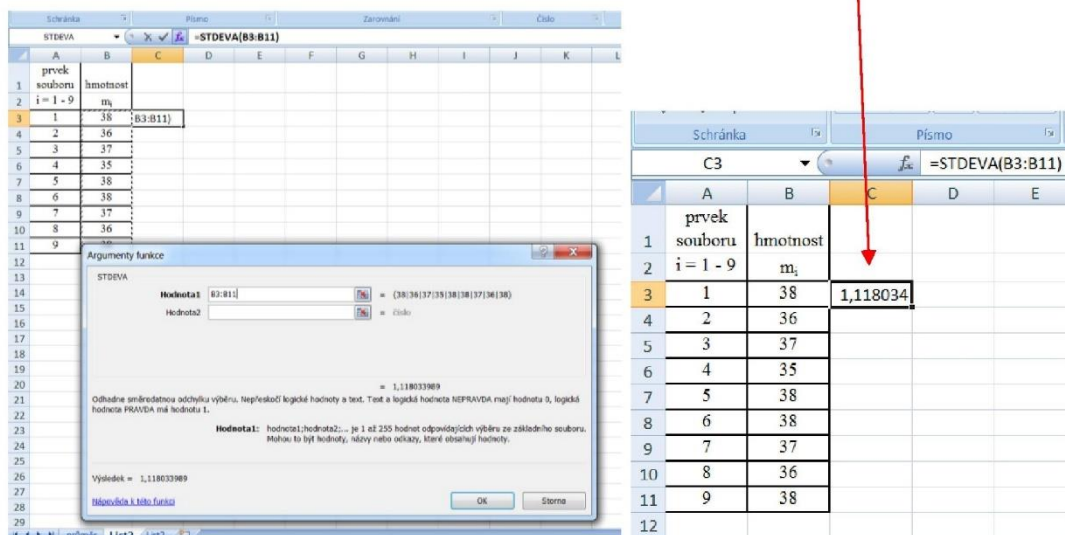
klikněte do nějakého políčka mimo tabulku, klikněte na: vybrat funkci f_x , vyberte kategorii statistické, STDEVA, ok



do políčka „argumenty funkce“ si levým tlačítkem myši označte (přetáhněte) celý sloupec čísel, u kterých chcete počítat průměr (zobrazí se vám souřadnice prvního a posledního čísla; B3: B11), dejte enter a potom ok



a objeví se vám v políčku, na které jste původně klikli, výpočet směrodatné odchylky



Obr. 14 Výpočet naší vzorové směrodatné odchylky v MS Excelu, výpočet ve Statistice znázorňuje obr. 13.

Zdroj: autoři

Míry rozptýlenosti založené na empirických kvantilech

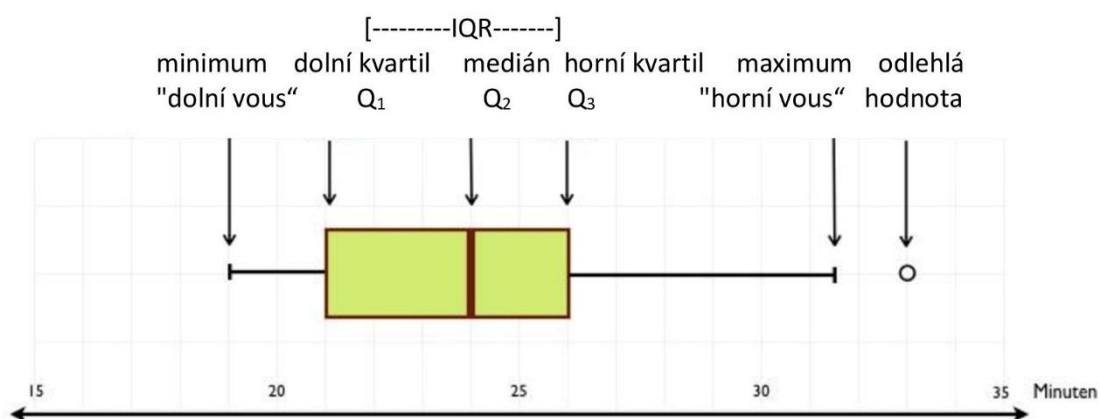
Empirický kvantil x je hodnota, pod níž leží definovaná část údajů. Udáváme jeho hladinu q a označujeme jej jako x_q . Hladiny q se někdy udávají v procentech, to je potom označujeme jako **percentily** = empirické percentily na dané úrovni (25% percentil je tedy rovný kvantilu o hladině 0,25). Percentily s hladinou 25 %, 50 % a 75 % nazýváme **kvartily** a takto je označujeme (Hendl, 2015, s. 104):

- Q_1 je první neboli dolní kvartil ($q = 25 \%$);
- Q_2 je druhý neboli **medián** ($q = 50 \%$) – již znáte z kapitoly 1.2 Míry centrální tendence;
- Q_3 je třetí neboli horní kvartil ($q = 75 \%$).

Někdy používáme k popisu tvaru dat **interkvartilové rozpětí** (R_Q , nebo *IQR*, interquartile range), které není tak citlivé k odlehlým hodnotám jako třeba směrodatná odchylka. V interkvartilovém rozpětí se nachází 50 % dat a spočítá se takto:

$$R_Q = Q_3 - Q_1$$

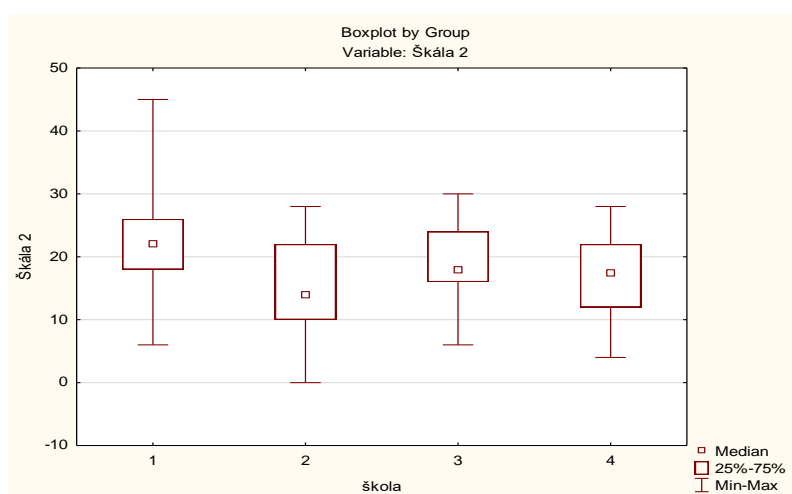
Kvartily i IQR nám nejlépe znázorní krabicové grafy (viz. obr. 15).



Obr. 15 Příklad krabicového grafu

Zdroj: Fachueber (2011); upraveno autory

Boxploty jsou velmi užitečné grafy při porovnávání více statistických souborů (obr. 16).



Obr. 16 Porovnání výsledků testů mezi čtyřmi školami

Zdroj: autoři

Pokud jsme střední hodnotu souboru dat charakterizovali pomocí mediánu, můžeme jako míru variability použít **kvartilovou odchylku** Q (Chráska, 2016, s. 49), kterou vypočítáme podle vzorce:

$$Q = \frac{Q_3 - Q_1}{2}$$

Například máme určit kvartilovou odchylku dat – hmotností žáků 3. třídy. Získali jsme tato data (v kg): 31, 32, 31, 35, 30, 42, 33, 32, 36, 45, 38. Nejprve si data seřadíme podle velikosti, určíme medián, dolní a horní kvartil a spočítáme kvartilovou odchylku Q .

30, 31, 31, 32, 32, 33, 35, 36, 38, 42, 45

$$Q = \frac{38 - 31}{2} = 3,5$$

1.4. Intervaly spolehlivosti

Interval spolehlivosti (konfidenční interval podle anglického confidence interval) je interval, ve kterém s určitou pravděpodobností (90, 95, nebo 99 %) leží skutečná hodnota veličiny odhadované na základě studia vzorku ze souboru. Pracujeme totiž jen s výběrovým souborem VS, což je náhodný výběr ze základního souboru ZS, jehož střední hodnotu chceme pomocí průměru VS odhadnout. Určíme si tedy interval, ve kterém se s určitou námi zvolenou pravděpodobností skutečná střední hodnota ZS nachází.

Konfidenční intervaly (intervaly spolehlivosti) jsou tak jedním ze způsobů, jak zjistit, nakolik můžeme generalizovat data naměřená na vzorku. Pokud totiž např. zjistíme u VS ($n = 20$), že dívky napsaly určený test o 20 % lépe, než chlapci; neznamena to, že tento výsledek analýzy platí pro celou populaci. Proto každou statistickou analýzu začínáme s tzv. „nulovou hypotézou“ H_0 , a to, že se v našem souboru nic neděje, neexistuje v ní žádný rozdíl mezi analyzovanými skupinami, žádná spojitost mezi zjištěnými daty. Jakmile nám ale míra pravděpodobnost (p -hodnota) poskytne dostatečnou jistotu, nulovou hypotézu opustíme a pracujeme s *alternativní hypotézou* H_1 (že nějaký rozdíl či spojitost existuje).

Může nám být např. známý jen jeden výběr ze souboru a jeho aritmetický průměr μ a nás zajímá, jak dobrý je to odhad střední hodnoty. Jinými slovy nás zajímá, v jakém pásmu kolem zjištěného aritmetického průměru μ se s předem stanovenou pravděpodobností nachází skutečná střední hodnota.

P se nazývá hladina (koeficient) spolehlivosti ($P = 1 - \alpha$, pro $0 < \alpha < 1$) a α je hladina významnosti neboli riziko. Jsou-li udány obě hranice intervalu, mluvíme o oboustranném intervalu, je-li dána pouze horní nebo dolní hranice, mluvíme o jednostranném intervalu spolehlivosti.

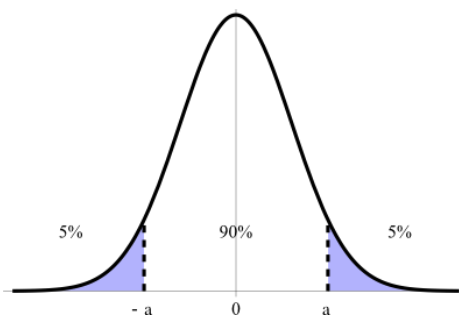
Hladina spolehlivosti 95 % (0,95) nebo 99 % (0,99) ovšem neznamená, že průměr μ leží uvnitř tohoto intervalu s touto pravděpodobností. Průměr μ je sice neznámý, ale pro daný soubor má určitou danou hodnotu; interval spolehlivosti je pak sestaven tak, aby pokryl tento parametr μ s danou spolehlivostí. 99% interval spolehlivosti tedy značí, že ve 100 náhodných výběrech se daná charakteristika objeví právě 99krát.

Pro zpracování na PC používáme nejčastěji oboustranný interval spolehlivosti 95 % (0,95), jemuž odpovídá hladina významnosti $\alpha = 0,05$. Co to znamená?

V praxi nám všechny testy inferenční statistiky poskytují hodnotu p , což je míra pravděpodobnosti, že uvidíme ve výzkumu nějaký účinek či vztah, byla-li nulová hypotéza pravdivá (nebo také míra pravděpodobnosti, že se budeme mýlit, začneme-li tvrdit, že jsme něco objevili). Také většina testů pracuje s hladinou významnosti α o hodnotě 0,05. Je to pro nás nejvyšší přijatelné riziko, že vyneseme mylné tvrzení. Použijeme-li hodnotu 0,05 jako hladinu významnosti pro posuzování p , říkáme tím, že přijmeme max. 5 % šanci, že přijdeme s mylným tvrzením. Víme, že musí existovat jisté riziko, že se mýlit budeme, a tak toto riziko kontrolujeme – staráme se o to, aby bylo pod 5 % (Walker, 2013, s. 104).

Pokud je **p -hodnota \geq než 0,05**, pak se dál držíme nulové hypotézy a docházíme k závěru, že se v populaci nic neděje, žádný účinek či spojitost jsme neodhalili. Pokud nám v testech vyjde **p -hodnota $<$ než 0,05**, odmítáme nulovou hypotézu a tvrdíme, že se v populaci opravdu něco děje, pracujeme s alternativní hypotézou a o výsledcích analýzy můžeme tvrdit, že jsou *signifikantní*. Alternativní hypotézu definujeme nejlépe jako dvoustrannou. Jednostranná hypotéza předpovídá zcela konkrétně, co se stane (např. pravidelné užívání guarany zvyšuje paměť); oboustranná hypotéza nechává možnost změny otevřenou ve více rovinách (např. pravidelné užívání guarany ovlivňuje paměť) a snižuje tak pravděpodobnost, že nějaký účinek či spojitost přehlédneme (na druhé straně ale bude muset být změna mnohem výraznější, aby byla signifikantní).

Proto také pracujeme s oboustranným intervalem spolehlivosti (viz obr. 17).



Obr. 17 Znárodnění oboustranného 90 % intervalu spolehlivosti na křivce standardizovaného normálního rozdělení

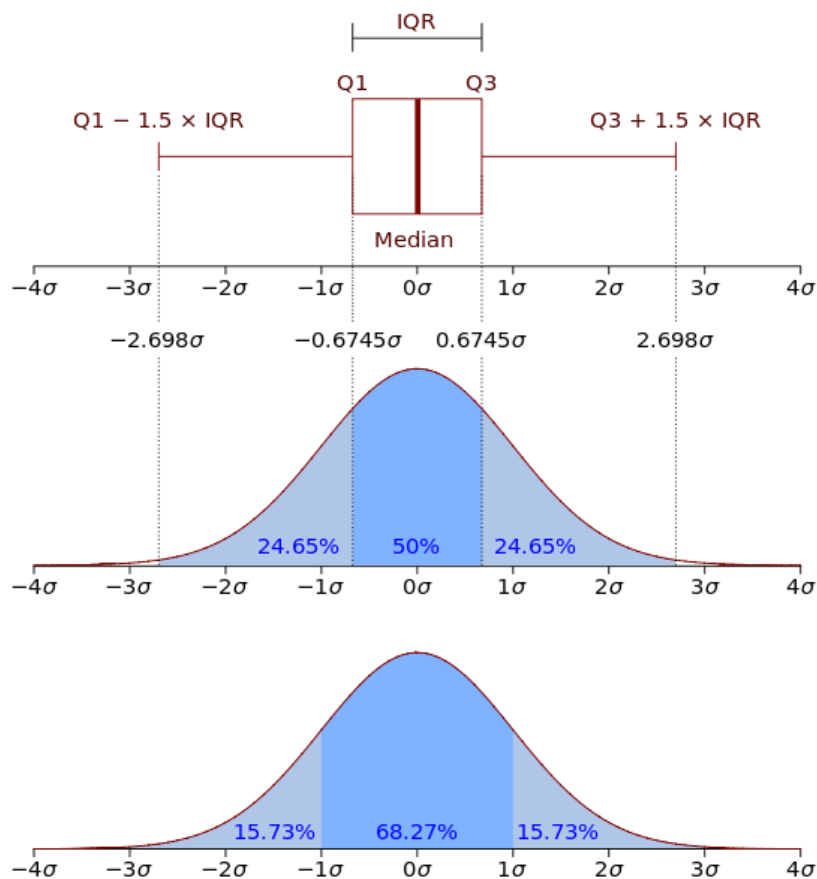
Zdroj: KendallVarent (2010)

1.5. Normální rozdělení

Normální rozdělení četností se vyznačuje tím, že značná část hodnot se soustřeďuje kolem průměrné hodnoty a na obě strany od ní jsou hodnoty stálé, méně časté, přičemž extrémní (odlehlé) hodnoty se vyskytují ojediněle. Graficky toto vyjadřujeme tzv. Gaussovou křivkou (Havel & Cihlář, 2011, s. 7).

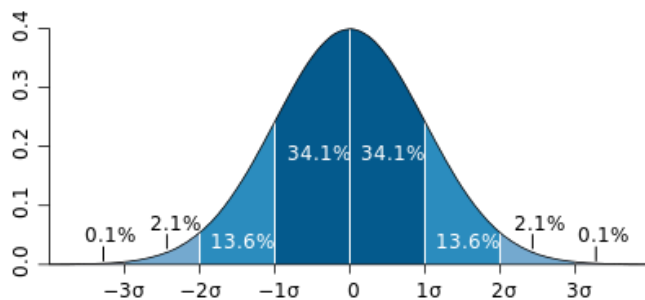
Znaky Gaussovy křivky (hustoty pravděpodobnosti), potažmo standardizovaného normálního rozdělení (Magnello, 2010, s. 60; Havel & Cihlář, 2011, s. 7):

- je symetrická kolem osy;
- má stejnosměrný zvonovitý tvar;
- její tvar je definovaný průměrem (který je roven 0) a směrodatnou odchylkou σ (která je rovna 1);
- vrchol křivky je totožný se střední hodnotou (M), modem (M_0) a mediánem (M_e);
- variační rozpětí $R \approx 6\sigma$;
- v intervalu $M \pm 1\sigma$ leží přibližně $\frac{2}{3}$ všech hodnot, tj. 68,27 % všech případů;
- v intervalu $M \pm 2\sigma$ leží přibližně $\frac{19}{20}$ všech hodnot, tj. 95,4 % všech případů;
- v intervalu $M \pm 3\sigma$ leží prakticky všechny hodnoty, tj. 99,73 % všech případů (pravidlo tří sigma);
- průměr ukazuje polohu rozdělení na ose x a rozptyl poukazuje na rozptýlení dat;
- šikmost křivky je nulová, protože je symetrická kolem průměru.



Obr. 18 Normální rozdělení četnosti včetně Boxplotu

Zdroj: Jhguch (2012)



Obr. 19 Graf normálního (Gaussova) rozdělení

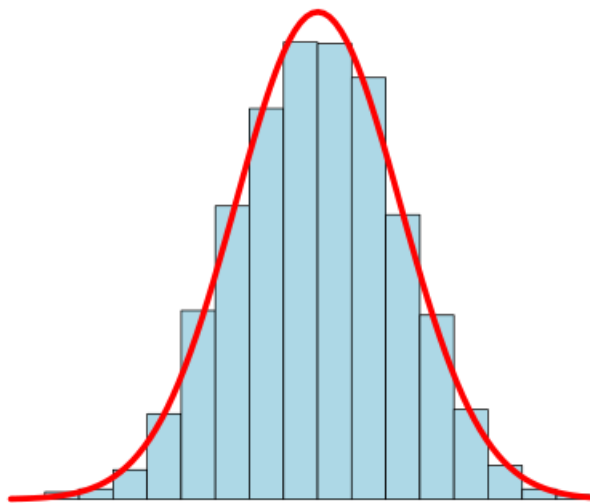
Zdroj: Toews (2007)

Normální rozdělení má mimořádný význam v teorii pravděpodobnosti i v matematické statistice. Je použitelné všude tam, kde je kolísání náhodné veličiny způsobeno součtem velkého počtu nepatrných a vzájemně nezávislých vlivů; je rozdělením limitním. Normální rozdělení mají i náhodné chyby při různých měřeních, můžeme na něj narazit v řadě technických, ekonomických, biologických, společenských a dalších situací (Neubauer et al., s. 125). Svět je prostě plný normálních rozdělení, ať měříte čas, který lidé tráví dojížděním do práce,

tloušťku šimpanzích chlupů, nebo množství pylu nasbírané včelami, často vám vyjdou podobné křivky (Walker, 2013, s. 126).

Testování normality

Pro zvolení správné statistické metody vyhodnocení našich dat je klíčové zjistit, zda jejich rozdělení je normální, či jiné než normální. Posoudit normální rozdělení dat můžeme testem špičatosti, šikmosti, posouzením histogramu či krabicového grafu, ale zejména testem normality.



Obr. 20 Histogram a normální rozdělení

Zdroj: Joxemai (2013)

Testy normality

Většina statistického softwaru má implementovány nějakou formu testů normality. Testů normality je několik: Shapirův-Wilkův, Andersonův-Darlingův, Kolmogorovův-Smirnovův, Lillieforsův, Test dobré shody a další.

Shapirův-Wilkův test je silný a velmi spolehlivý, dokonce i pro malé soubory. Je citlivý zvláště k nesymetrickým rozdělením, rozdělením s těžkými chvosty a též k odlehlým pozorováním (Malíková, 2014, s. 23).

Máme tedy soubor Y naměřených dat (hmotností) o četnosti 11 (31, 32, 31, 35, 30, 42, 33, 32, 36, 45, 48), střední hodnotě μ či \bar{m} (35) a směrodatné odchylce σ (4,878524). Ze směrodatné odchylky můžeme zjistit rozptyl σ^2 . Chceme zjistit, zda změřený náhodný výběr Y pochází z normálního rozdělení. Zformulujeme si hypotézy:

Nulová hypotéza H_0 : Změřený náhodný výběr pochází z normálního rozdělení s libovolnými parametry μ a σ^2 .

$$H_0: D(Y) \sim N(\mu, \sigma^2)$$

Alternativní hypotéza H_1 : Změřený náhodný výběr pochází z jiného, než normálního rozdělení s libovolnými parametry μ a σ^2 .

$$H_1: D(Y) / \sim N(\mu, \sigma^2)$$

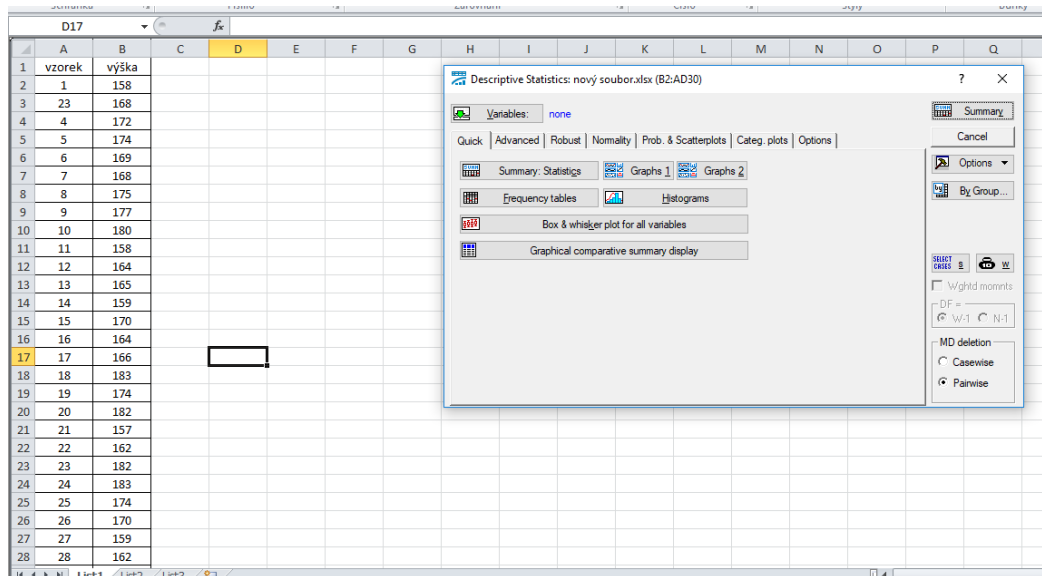
Použijeme Shapirův-Wilkův test a bude nás zajímat p -hodnota na hladině významnosti $\alpha = 0,05$.

Pokud bude naše získaná **p -hodnota > než 0,05**, platí nulová hypotéza a naše data pocházejí z normálního rozdělení (pro další statistické operace využíváme parametrické testy).

Pokud bude naše získaná **p -hodnota < než 0,05**, platí alternativní hypotéza a naše data pocházejí z jiného, než normálního rozdělení (pro další statistické operace využíváme neparametrické testy).

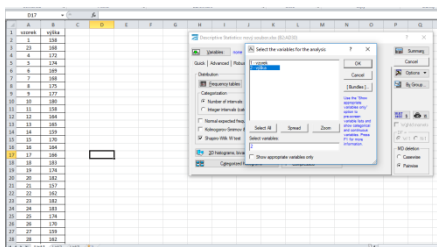
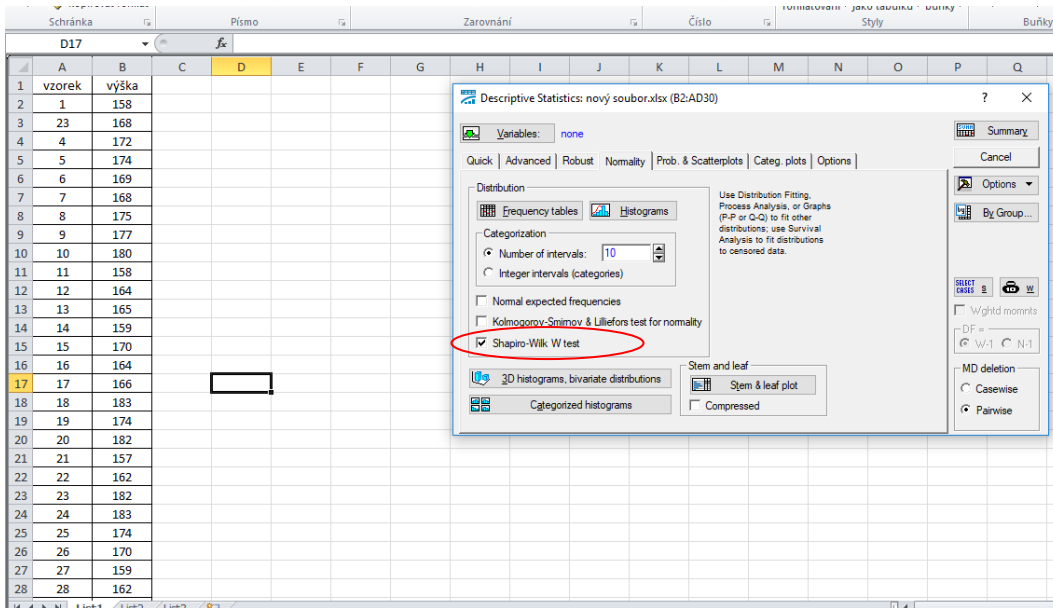
Kde najdete test normality na PC?

Statistica: statistics → základní statistiky (basic statistics) → popisné statistiky (descriptive statistics) → normality → Shapiro-Wilk test → vybrat proměnnou → frequency tables (nejlépe pak použít graphical summary, získáme kompletní přehled)



Obr. 21 Postup zjištění normality

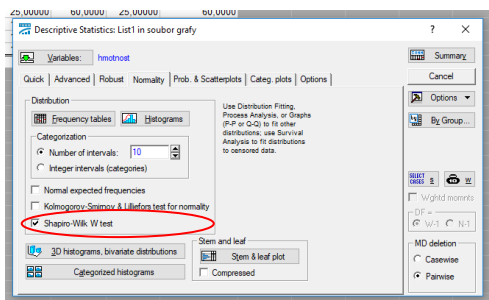
Zdroj: autoři



Statistics - [Workbook2] - Frequency table: výška (nový soubor.xlsx (B2:AD30))

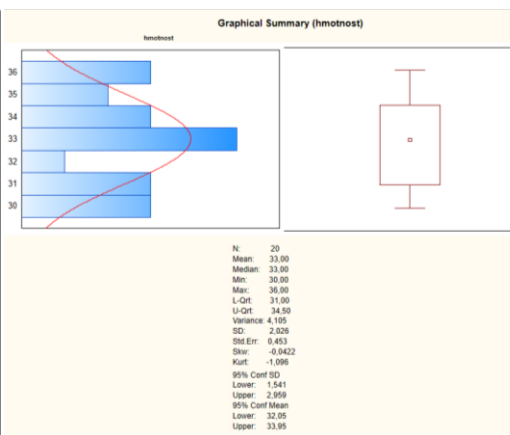
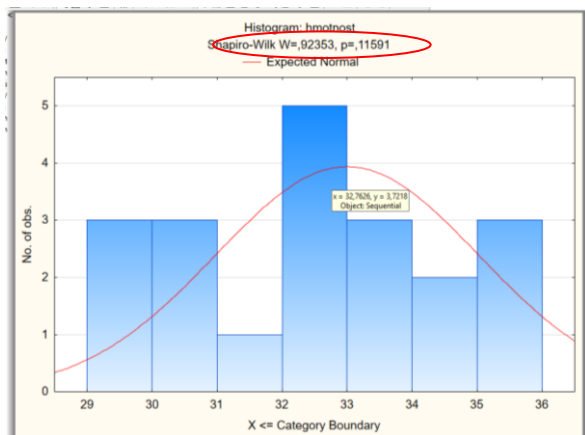
Frequency table: výška (nový soubor.xlsx (B2:AD30))
Shapiro-Wilk W = 948.6, p = 17095

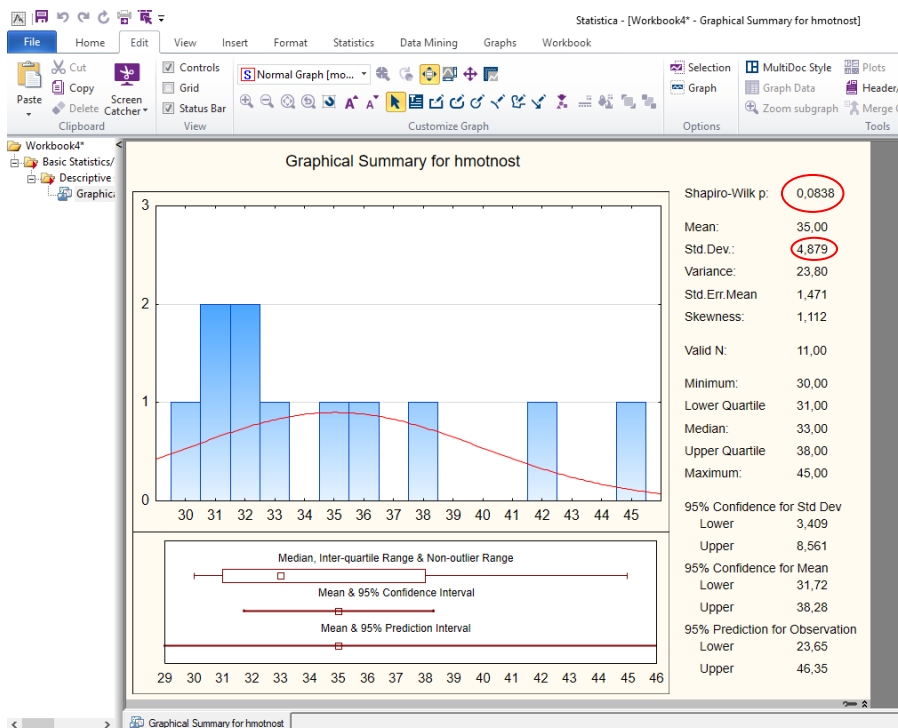
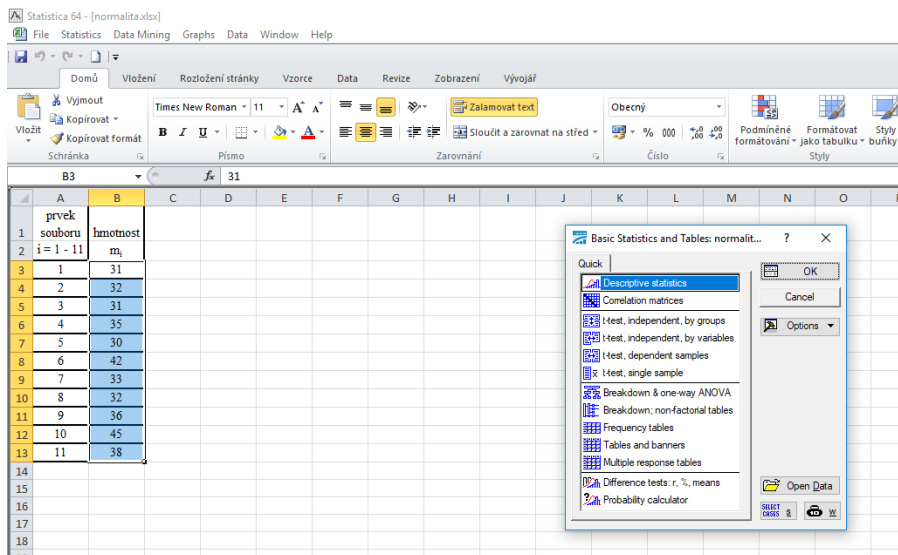
Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
150,0000<x<=155,0000	0	0	0,00000	0,0000	0,00000	0,0000
155,0000<x<=160,0000	5	5	17,24138	17,2414	17,24138	17,2414
160,0000<x<=165,0000	5	10	17,24138	34,4828	17,24138	34,4828
165,0000<x<=170,0000	8	18	27,58621	62,0690	27,58621	62,0690
170,0000<x<=175,0000	5	23	17,24138	79,3103	17,24138	79,3103
175,0000<x<=180,0000	2	25	6,89655	86,2069	6,89655	86,2069
180,0000<x<=185,0000	4	29	13,79310	100,0000	13,79310	100,0000
Missing	0	29	0,00000	0,00000	0,00000	100,0000



Frequency table: hmotnost (List1 in soubor grafy)
Shapiro-Wilk W = 92353, p = 11591

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
29,00000<x<=30,00000	3	3	15,00000	15,0000	15,00000	15,0000
30,00000<x<=31,00000	3	6	15,00000	30,0000	15,00000	30,0000
31,00000<x<=32,00000	1	7	5,00000	35,0000	5,00000	35,0000
32,00000<x<=33,00000	5	12	25,00000	60,0000	25,00000	60,0000
33,00000<x<=34,00000	3	15	15,00000	75,0000	15,00000	75,0000
34,00000<x<=35,00000	2	17	10,00000	85,0000	10,00000	85,0000
35,00000<x<=36,00000	3	20	15,00000	100,0000	15,00000	100,0000
Missing	0	20	0,00000	0,00000	0,00000	100,0000





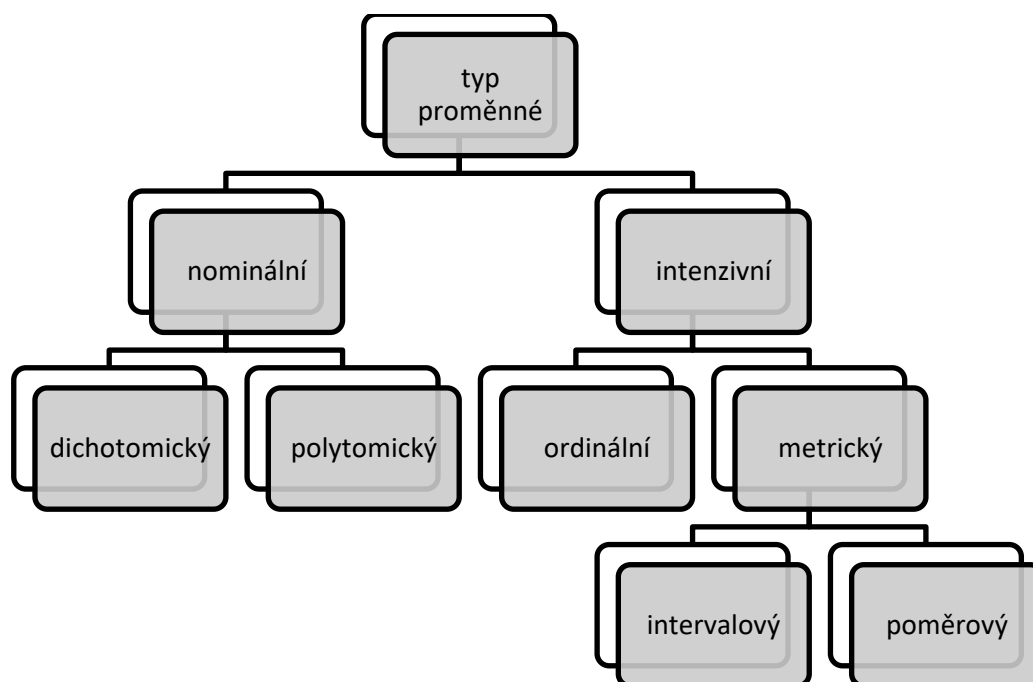
Obr. 22 Různé postupy zjišťování normality

Zdroj: autoři

Graphical summary je užitečný pracovní list, který sumarizuje důležité charakteristiky souboru, jako je průměr, směrodatná odchylka, normální rozdělení dat a další.

2. Typy proměnných

Při výzkumu sledujeme určité **znaky** nějaké osoby, skupiny osob či věci (např. inteligenci, kreativitu, dosažené vzdělání, věk, pohlaví, výšku). Tyto znaky nazýváme **proměnnými**. Aktuální hodnoty proměnných tvoří **data**. Data získáváme měřením, za použití různých testů, dotazníků. Při analýze dat těmto proměnným přiřazujeme číselnou hodnotu. Vždy však musíme mít na mysli, co daná hodnota představuje. Na základě použitého měřítka se totiž jednotlivé druhy proměnných mezi sebou významně liší (viz obr. 23 a tab. 8). Proměnná je tedy „*charakteristikou prvků základního souboru, jež mohou nabývat více hodnot a existuje pro ně předpis, jak tyto hodnoty zjistíme.*“ (Hendl, 2012)



Obr. 23 Přehled typů proměnných

Zdroj: dle Hendla (2012)

Závisle a nezávisle proměnné, rušivá proměnná

Na začátku výzkumu si musíme stanovit závisle a nezávisle proměnné. Mezi nimi předpokládáme nějaký příčinný vztah (pokud změním nezávisle proměnnou, změní se závisle proměnná).

Příkladem může být:

- druhu studované školy (nezávisle proměnná) a dosažený počet bodů v didaktickém testu (závisle proměnná);
- nejvyšší dosažené vzdělání rodičů žáka (nezávisle proměnná) a školní prospěch (závislá proměnná);

- pohlaví (nezávisle proměnná) a výsledky v testu manuální zručnosti (závisle proměnná).

Zjednodušeně lze říci, že závislá proměnná je to, co ve výzkumu měříme a nezávisle proměnná popisuje obecné charakteristiky jedince (pohlaví, věk, druh školy, dosažené vzdělání rodičů), podle kterých respondenty často třídíme do skupin (dívky/chlapci; žáci alternativní školy/“běžné“ školy). Pokud máme připravený test nebo dotazník, ptáme se většinou nejprve na nezávisle proměnné. (Hendl, 2012)

Rušivá proměnná je taková proměnná, která zkresluje výsledky našeho výzkumu.

Diskrétní a spojité proměnné

Spojité proměnné mohou nabývat libovolných hodnot reálných čísel (výška, váha – mohou například naměřit 28,64 kg), diskrétní proměnné nabývá konečného, spočetného množství variant (známky – máme pět známek) – (Litschmannová, 2012).

Proměnné podle typu měřítka

Podle Hendla (2012) rozlišujeme 4 různá měřítka. Liší se v tom, co přesně reprezentují čísla, která jsme sledovaným hodnotám přiřadili, zda lze tedy s nimi skutečně počítat jako s plnohodnotnými čísly, či zda jsou pouze zastupujícím symbolem. Než začneme s čísly provádět různé operace, musíme si uvědomit, co číselné symboly ve skutečnosti vyjadřují. Jednotlivé typy měřítek jsou zaznamenány i s popisem a příklady v tab. 8.

Tab. 8 Přehled měření s příklady

Měřítko	Popis	Příklad
Nominální	<ul style="list-style-type: none"> • rozlišuje kategorie („škatulkuje“) • kategorie však nemají určené pořadí 	<p><i>Dichotomický (na výběr ze 2 možností)</i></p> <ul style="list-style-type: none"> • pohlaví (muž/žena) <p><i>Polytomický (na výběr z více možností)</i></p> <ul style="list-style-type: none"> • nejoblíbenější předmět (matematika/prvouka/TV/...)
Ordinální	<ul style="list-style-type: none"> • podobné nominálním • ale hodnoty lze seřadit podle intenzity, lze je porovnávat 	<ul style="list-style-type: none"> • položka s Likertovou škálou (naprosto souhlasím/spíše souhlasím/nevím/spíše nesouhlasím/naprosto nesouhlasím)
Intervalové	<ul style="list-style-type: none"> • podobné ordinálním • vzdálenost mezi kategoriemi je dána jednotkou měření (můžeme je tedy např. sčítat) 	<ul style="list-style-type: none"> • teplota měřená ve stupních Celsia • inteligence
Poměrové	<ul style="list-style-type: none"> • podobné intervalovým • existuje pro ně ale i absolutní nulový bod 	<ul style="list-style-type: none"> • věk • počet bodů ze znalostního testu

Zdroj: dle Chytrý & Kroufek (2017)

Jak již bylo zmíněno výše, s ohledem na měřítko můžeme s daty provádět různé výpočty. Přehledně jsou možné výpočty pro jednotlivá měřítka uvedena v následující tabulce (tab. 9).

Tab. 9 Možnosti deskriptivní analýzy vzhledem k použitému měřítku

	Nominální	Ordinální	Intervalové
			Poměrové
Četnost	✓	✓	✓
Modus	✓	✓	✓
Medián		✓	✓
Průměr			✓
Směrodatná odchylka			✓
Lze kvantifikovat rozdíl mezi jednotlivými hodnotami (lze je sčítat, odčítat, násobit i dělit)			✓
Je možné přidat nebo ubrat hodnoty			✓

Zdroj: dle Chytrý & Kroufek (2017)

3. Organizace dat (kódování)

Naměřená data organizujeme do přehledných tabulek. Nejprve si shrneme základní pojmy a poté představíme několik zásad, kterých bychom se při kódování měli držet.

Základní pojmy (viz obr. 24):

- tabulka dat (datová matice) – v řádku jsou zaznamenány sledované charakteristiky jednoho objektu (respondenta), ve sloupci jsou data pro jednu proměnnou;
- buňka – průsečík řádku a sloupce, obsahuje právě jeden údaj/hodnotu;
- kód – jednoznačný předpis k přiřazování vhodných symbolů (čísel) sledovaným hodnotám (například, sleduji-li oblíbené předměty, přiřadím matematice vždy číslo 1, prvouce vždy číslo 2, ...) (Hendl, 2012).

Sloupce - jednotlivé proměnné

	A	B	C	D	E
1	Rspodent	Pohlaví	Známka z matematiky	Výsledek v testu	
2	1	1	1	12	
3	2	1	2	9	
4	3	0	2	6	
5	4	1	2	14	
6	5	0	3	5	
7	6	0	1	15	
8	7	0	1	11	
9	8	0	4	6	
10	9	0	1	17	
11	10	0	3	4	
12	11	1	2	15	
13	12	0	4	9	
14					

Řádek - charakteristiky 1 respondenta

Tabulka dat

Buňka

Obr. 24 Základní pojmy kódování (příklady)

Zdroj: dle Hendl (2012)

Dle Hendla (2012) je vhodné se při kódování řídit následujícím výčtem pravidel.

- Jasně vymezený způsob kódování (např. muž – 0, žena – 1). Je dobré si vymezený způsob kódování vypsát a za všech okolností se jím řídit.
- Každá proměnná zabírá jeden sloupec (u každého objektu ji nalezneme na stejné pozici)
- Kódy pro proměnnou jsou disjunktní – všechny různé pozorované hodnoty (skupiny) mají různé symboly. Hodnotě jedné proměnné u jedince nemůžou být přiřazeny dva symboly. Například u proměnná pohlaví máme pro každou ze skupin jiný symbol, muže označím např. číslicí 1, a ženy číslicí 0. Každému respondentovi následně přiřadíme jen jednu z uvedených hodnot.
- Kódování má zachovat maximum informací pro proměnnou (nezjednodušujeme, neredukujeme množství informací, to můžeme i později; například u věku nepíšeme pouze zařazení do dekad).
- Musí být definovány kódy pro všechny proměnné a všechny její specifické hodnoty (nutné je kódovat i scházející údaje např. prázdným znakem).
- Musíme se vyrovnat s kódováním nespecifických odpovědí (nevím, nejsem rozhodnut) a s rozlišením mezery a nuly.

4. Zpracování Likertových škál

Škálování je metoda výzkumu, s níž lze zachytit určitý kvalitativní jev v kvantitativní podobě (Rod, 2012). Při Likertově škálování označuje respondent míru souhlasu či nesouhlasu s daným tvrzením, velmi často se jedná o pěti stupňovou škálu. Blíže se tématu Likertových škál věnuje článek: Možnosti využití Likertovy škály – základní principy aplikace v pedago-

gickém výzkumu a demonstrace na příkladu zjišťování vztahu člověka k přírodě (Chytrý & Kroufek, 2017).

Postup při zpracování škál:

1. Uvědomit si, zda budeme počítat každou položku zvlášť, nebo nástroj jako celek.
2. Vymezit typ použitého měřítka (viz typy proměnných a měřítek).
3. Detekovat odlehlé hodnoty (viz odlehlé hodnoty).
4. Ověřit psychometrické vlastnosti nástroje
 - Reliabilita – ke zkoumání reliability se využívá Cronbachova alfa, která nabývá hodnot v intervalu $<0; 1>$; přijatelné jsou hodnoty mezi 0,7 a 0,95 (Tavakol & Dennick, 2011); Gavora (2010) i Chráska (2016) považují za dostatečně reliabilní výzkumný nástroj s hodnotou reliability alespoň 0,8.
 - Validita – Její hodnota poukazuje na to, do jaké míry zkoumáme to, co jsme chtěli zkoumat (Chráska, 2016). Na validitu lze nahlížet z různých hledisek, rozlišujeme tedy validitu obsahovou, kriteriální a konstruktovou (Parker & Lunney, 1998).
5. Vlastní analýza dat. Následující tabulka shrnuje možné analýzy dat různých statistických metod, a to jak pro jednotlivé hodnocení položek, tak pro hodnocení škály jako celku.

Tab. 10 Přehled možné analýzy dat pro hodnocení jednotlivých položek a hodnocení škály jako celku

Statistická metoda	Hodnocení položek	Hodnocení škály jako celku
Míra vnitřní konzistence	Ordinální alfa	Cronbachova alfa
Míra ústřední tendence	Medián nebo modus	Průměr
Míra variability	Frekvence (četnost)	Směrodatná odchylka
Míra asociace	Kendalovo tau <i>b</i> nebo <i>c</i> , Spearmanovo <i>ρ</i>	Pearsonovo <i>R</i>
Ostatní statistiky	χ^2 test, Mann-Whitney U-test	ANOVA, t-test

Zdroj: Chytrý & Kroufek (2017) podle Subedi (2016).

4.1. Transformace dat (spojitost s normálním rozdělením)

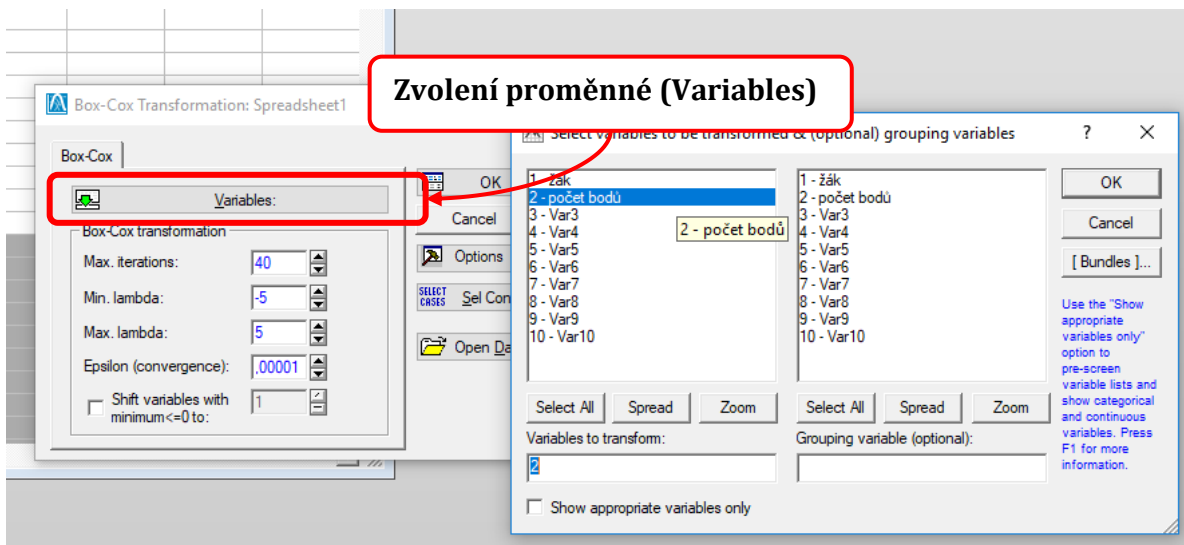
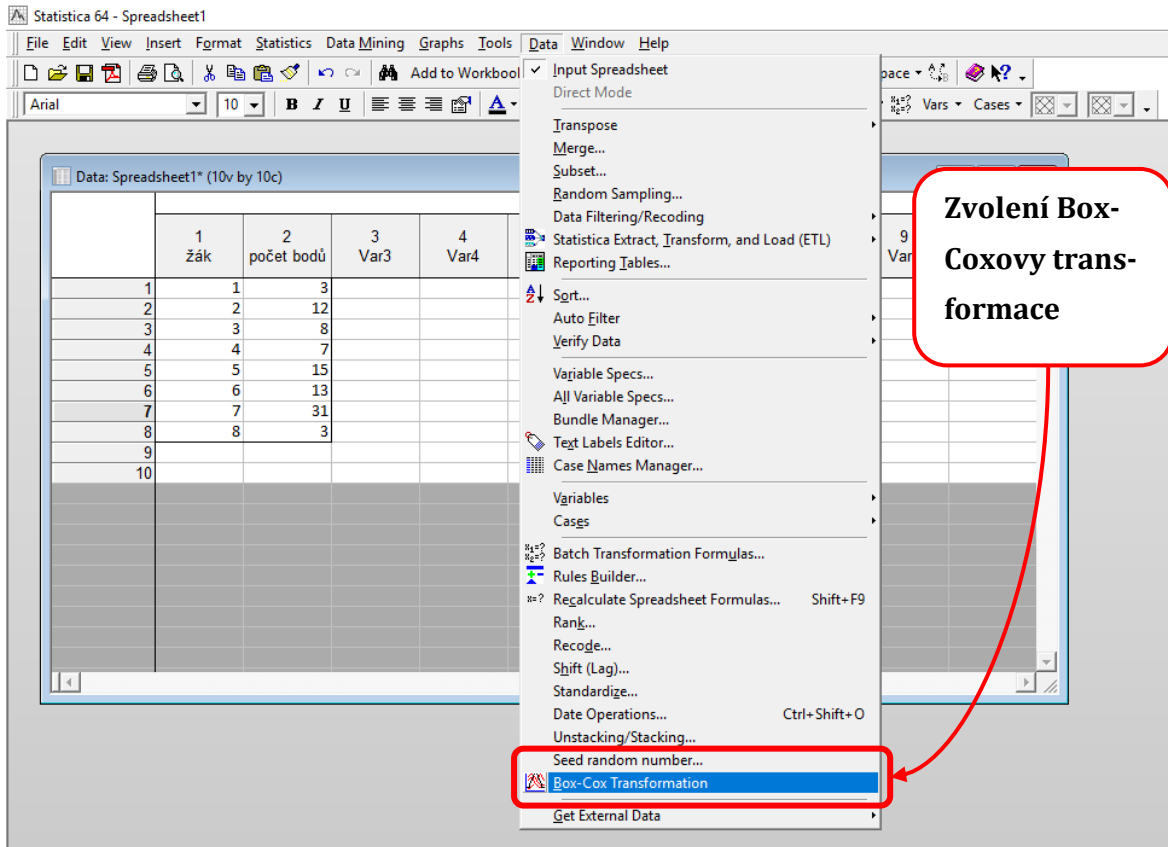
Lineární transformace (např. násobení hodnot proměnné konstantou) nemění výsledky analýzy v případech, že jde o analýzu vztahu proměnných (např. korelace); v případě, že je důležitá absolutní hodnota proměnné, však dochází k vážení jejího významu v analýze (Holčík & Komenda, 2015)

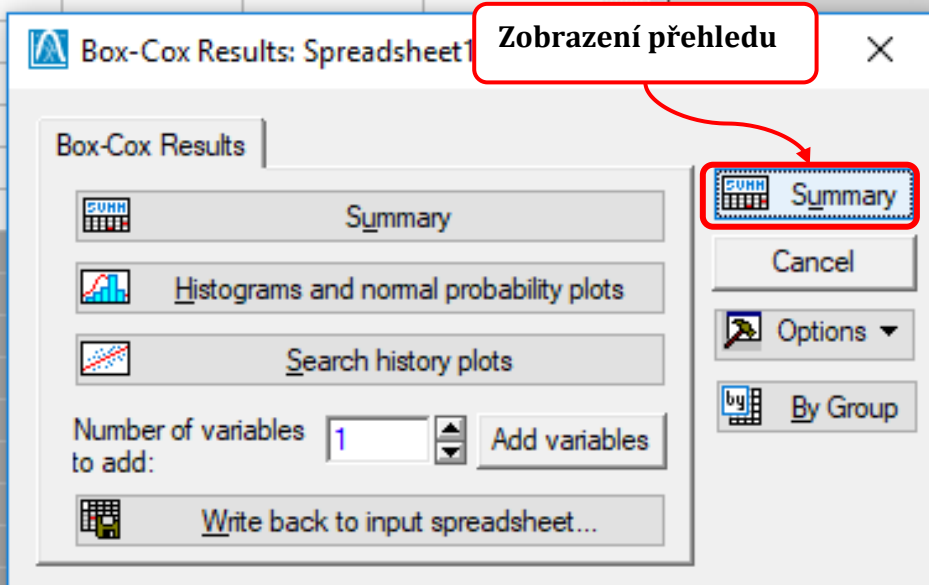
Box-Coxova transformace

Původní data by měla být transformována, pokud je výběrové rozdělení zešikmené, není homogenní a obsahuje-li odlehlé hodnoty (Meloun & Kupka, 2001). Box-Coxova transformace účinně přibližuje výběr normalitě jak z hlediska šikmosti, tak i z hlediska extrémních hodnot (Drápela, 2012).

Kde najdete výpočet Box-Coxovy transformace v programu Statistica?

Data → Box-Cox Transformation → Variables (proměnné) → zvolení příslušné proměnné
→ OK → Summary





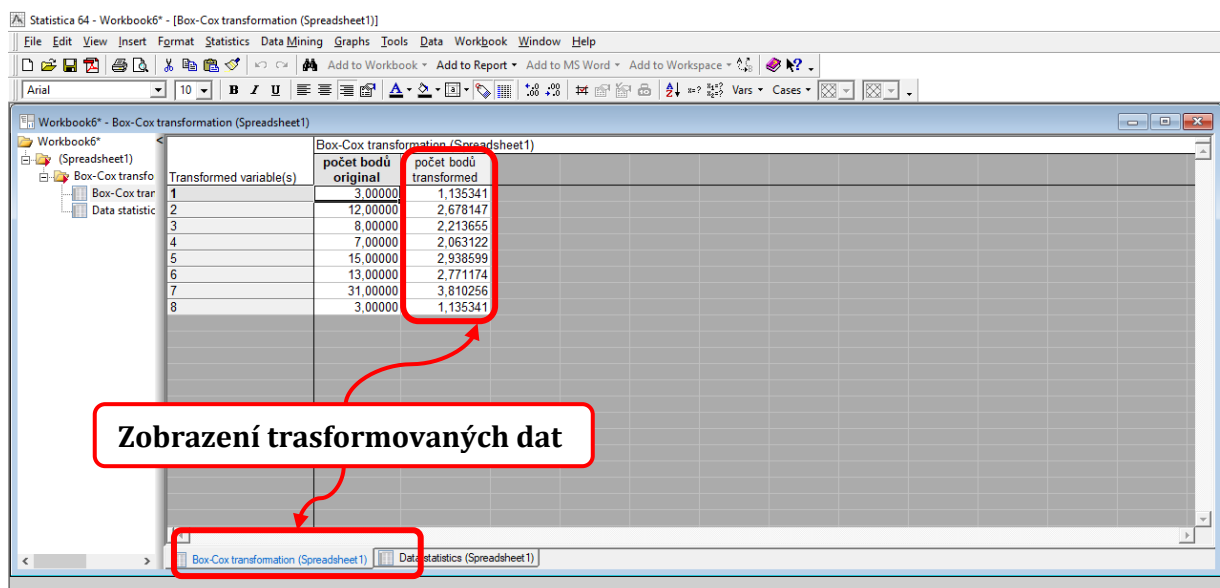
Statistica 64 - Workbook6* - [Data statistics (Spreadsheet1)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Workbook Window Help

Workbook6* - Data statistics (Spreadsheet1)

Transformed variable(s)	Lambda	Shift	Mean	Standard deviation	Lower Confidence Limit	Upper Confidence Limit	Formula for Box-Cox transformation
počet bodů	0.059542	0.00	2.343204	0.911449	-0.938937	0.985840	$((\lambda^2 + (0.059542)) - 1) / (0.059542)$

Workbook6 (Spreadsheet1) | Box-Cox transformation (Spreadsheet1) | Data statistics (Spreadsheet1)



Obr. 25 Box-Coxova transformace v programu Statistica

Zdroj: autoři

5. Detekce/odstraňování odlehlých hodnot a rezistentní odhady²

Při práci s daty je záhodno vypořádat se nejprve s podezřelými hodnotami. Extrémně vysoké nebo nízké hodnoty mohou být způsobeny chybou při měření či při práci s daty. Tyto chyby mohou zkreslit výsledky výpočtů. Ne vždy ale musí nutně jít o chyby, je tedy třeba se nad takovými hodnotami, jejich významem a možnými příčinami řádně zamyslet.

Pokud odlehlé hodnoty odhalíme, můžeme dle Hendla (2012) využít jeden z následujících postupů:

- vyřadit tato měření ze zpracování;
- provést výpočty s odlehlými hodnotami a bez nich, následně výsledky porovnat;
- použít rezistentní odhady, které nejsou tolik citlivé k odlehlým hodnotám (medián).

Tab. 11 Detekce odlehlých hodnot

Způsob zjištění	Použití
Srovnání průměru a mediánu hodnot	
Krabicový graf	
Dean-Dixonův test	Neparametrický/ neznámé rozložení
Grubbsův test	Parametrický

Zdroj: autoři

² Rezistentní odhad je takový odhad, který není ovlivněn odlehlými hodnotami (Budíková et al., 2005).

5.1. Srovnání průměru a mediánu hodnot

Při srovnání průměru a mediánu posuzujeme jejich rozdíl intuitivně. Pokud tedy například máme hodnoty v rozpětí 10 až 15 a hodnoty průměru a mediánu se liší o 1, pak je rozdíl zřejmě velký natolik, abychom se mohli domnívat, že se v souboru vyskytuje odlehlá hodnota. Pokud však srovnáváme průměr a medián hodnot mezi 10 a 800, a zjistíme, že se liší o 1, pak zřejmě nebudeme odlehlé hodnoty hledat.

Na příkladu (viz obr. 26) vidíme, že v prvním případě se aritmetickým průměr a medián liší o 1,375 ($11,375 - 10 = 1,375$), což nám poukazuje na možnost výskytu odlehlé hodnoty. Tu skutečně můžeme najít na 8. řádce, kde došlo k chybě při přepisu dat do tabulky.

Pokud by byl údaj zapsán správně (viz druhá tabulka vpravo), byl by rozdíl mezi aritmetickým průměrem a mediánem jen 0,5 ($8 - 7,5 = 0,5$).

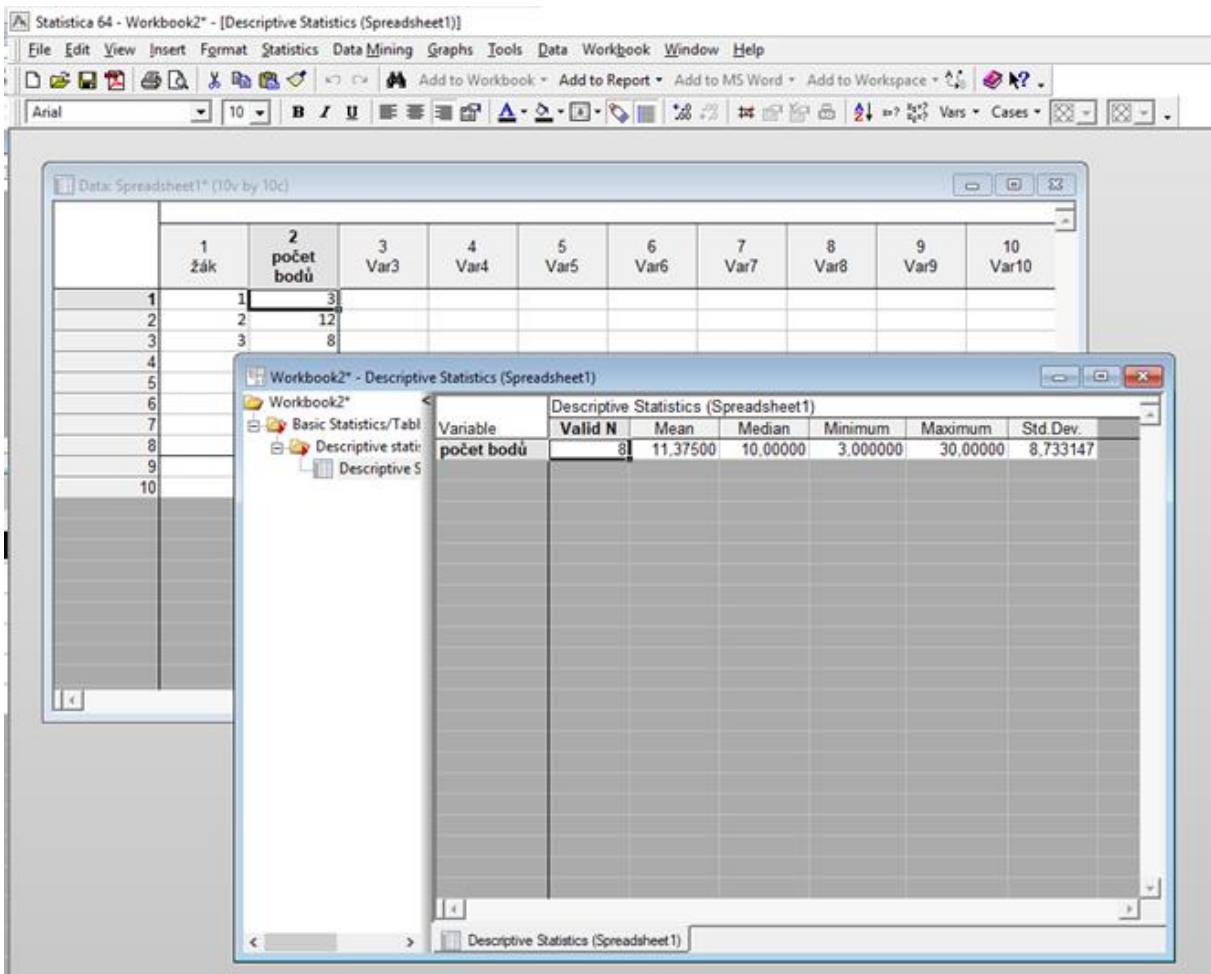
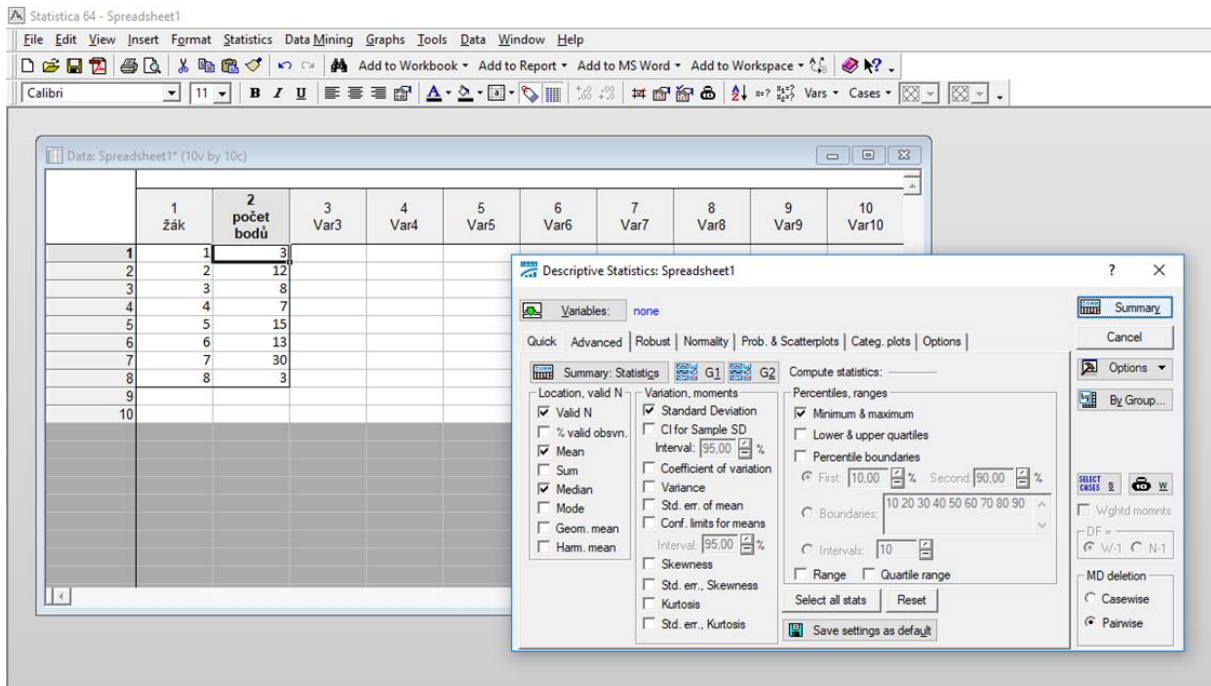
	A	B	C	D	E	F
1	žák	počet bodů		žák	počet bodů	
2	1	3		1	3	
3	2	12		2	12	
4	3	8		3	8	
5	4	7		4	7	
6	5	15		5	15	
7	6	13		6	13	
8	7	30		7	3	
9	8	3		8	3	
10						
11	Průměr	11,375		Průměr	8	
12	Medián	10		Medián	7,5	

Obr. 26 Rozdíl mezi aritmetickým průměrem a mediánem

Zdroj: autoři

Průměr a medián v programu Statistica

Statistics → Basic Statistics/Tables → Descriptive statistics → OK → Mean (průměr), Median → Variables → Mean (průměr), Median → zvolení příslušné proměnné → OK → Summary

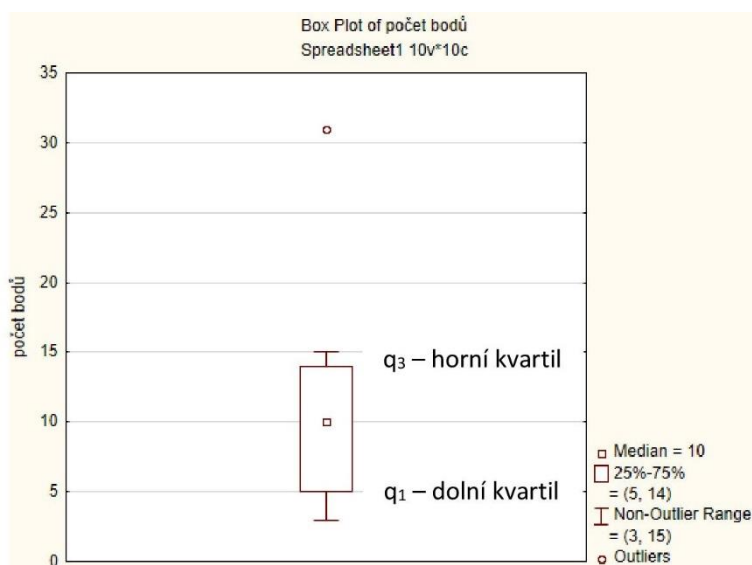


Obr. 27 Výpočet průměru a mediánu v programu Statistica

Zdroj: autoři

5.2. Krabicový graf

Odlehle hodnoty můžeme detekovat i prostřednictvím krabicového grafu (také nazývaný kvartilový graf). Jedná se o jednorozměrný graf, který je tvořen **krabicí**, jejíž výška je určena hodnotou interkvartilového rozpětí (Q), totiž **rozdílu horního (q_3) a dolního (q_1) kvartilu**. Uvnitř je např. čarou nebo puntíkem označen **medián**. Z krabice pokračují směrem nahoru i dolu tzv. vousy zakončené přilehlými body. **Přilehlé body** jsou ty body, které se nachází nejbližší vnitřním hradbám souboru. Horní hranici (BH) a dolní hranici (BD) hradeb spočítáme snadno.



Obr. 28 Ukázka krabicového grafu s vyznačenými kvartily

Zdroj: autoři

$$BH = q_3 + 1.5Q,$$

$$BD = q_1 - 1.5Q.$$

Vnitřní hradby v grafu znázorňovány nebývají. Koncové body úseček jsou tedy nejmenší a nejvyšší „bezproblémové“ hodnoty souboru. Body ležící mimo vnitřní hradby jsou považovány za „podezřelé“ (odlehle, vybočující) a jsou graficky znázorněny (křížky, kolečky apod.) v příslušných vzdálenostech (Drápela, 2012).

Popis postupu v MS Excel:

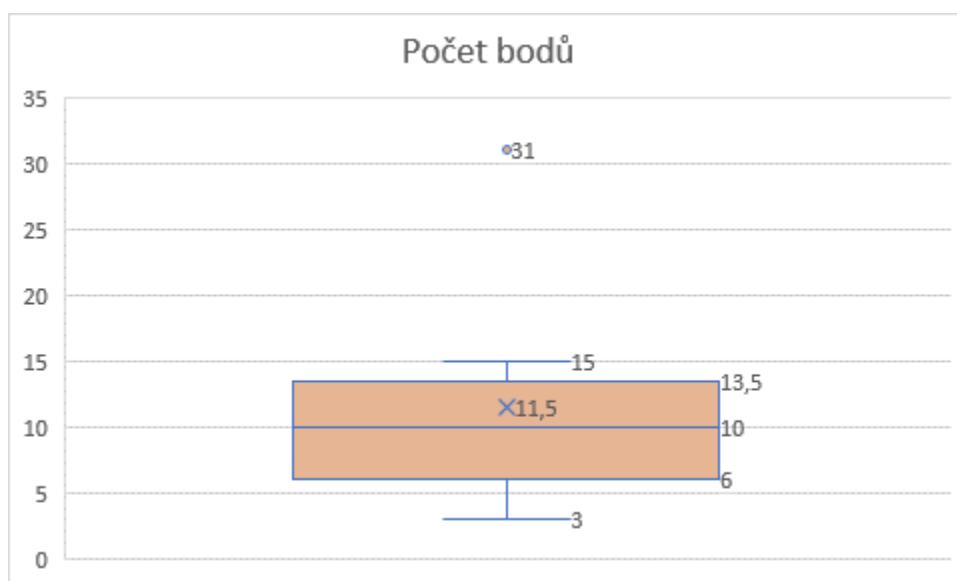
1. Označíme oblast dat, která chceme zaznamenat do krabicového grafu.
2. Na vrchní liště vybereme „Vložení“.
3. U grafů klikneme vpravo dole na šipku „Zobrazit všechny grafy“ (viz obr. 29).



Obr. 29 Nabídka „Zobrazit všechny grafy“ v MS Excel

Zdroj: autoři

4. Na horní liště nového okna zvolíme „Všechny grafy“.
5. V levém sloupci vybereme „Krabicový graf“.
6. Potvrdíme tlačítkem „OK“.
7. Vytvořený graf (obr. 30) pojmenujeme a vhodně dopravíme.



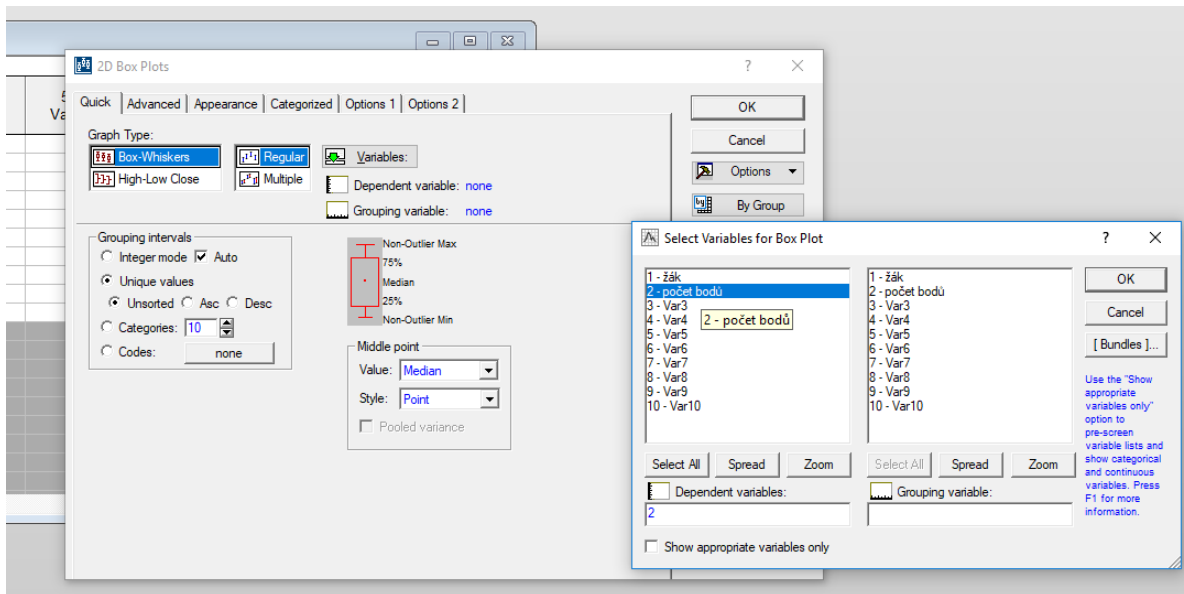
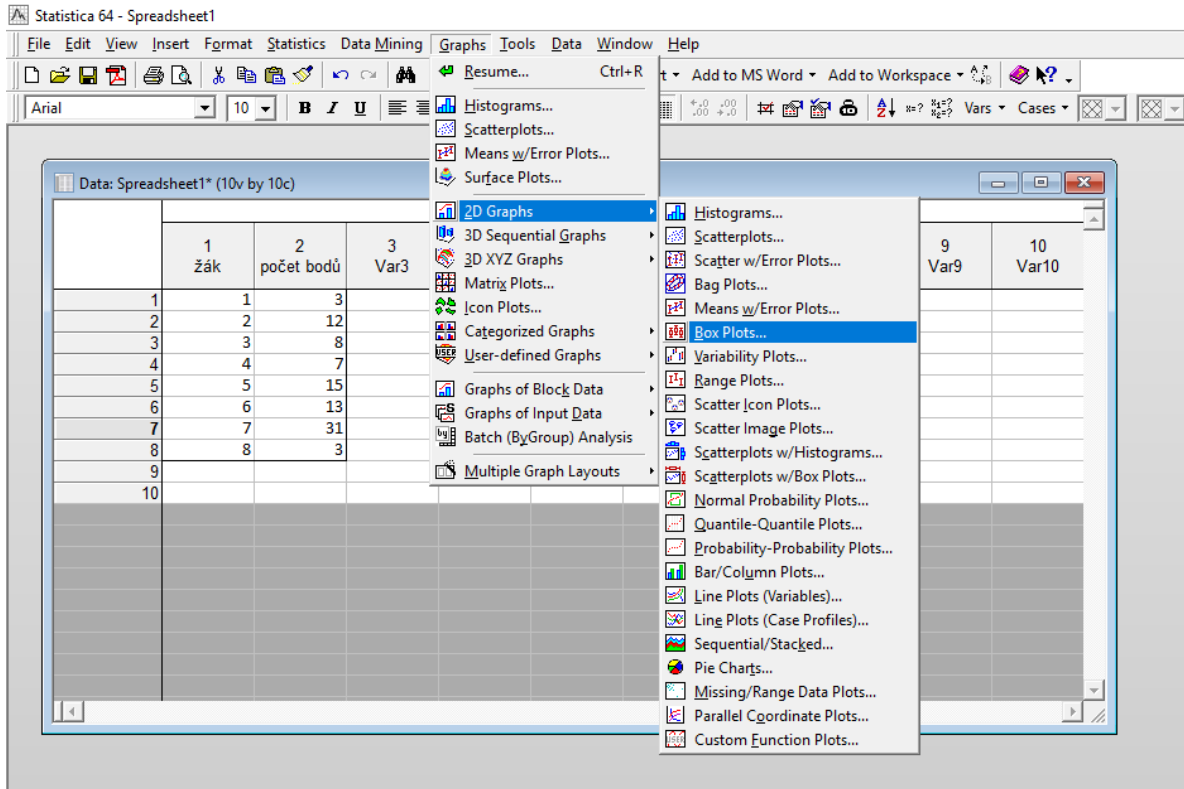
Obr. 30 Ukázka vytvořeného krabicového grafu v MS Excel

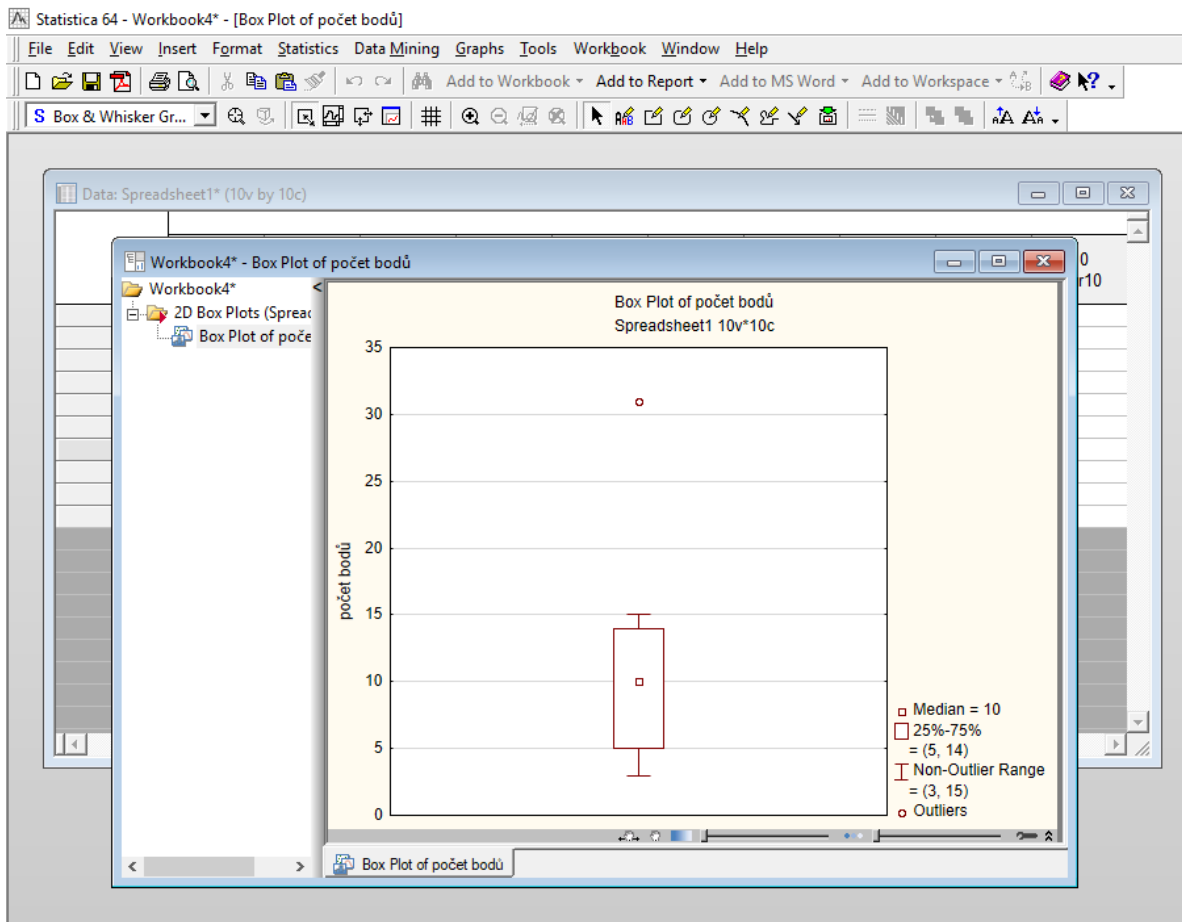
Zdroj: autoři

8. Puntík ukazuje odlehlou hodnotu. Křížek značí aritmetický průměr. Střední čára v obdélníku představuje medián (střední hodnotu). Hodnoty uprostřed obdélníku představují hodnoty mezi horním a dolním kvartilem.

Krabicový graf v programu Statistica

Graphs → 2D Graphs → Box Plots... → Variables → zvolení vybraných proměnných → OK





Obr. 31 Tvorba krabicového grafu v programu Statistica

Zdroj: autoři

5.3. Dean-Dixonův test

Grubbsovy testy se použijí k detekci odlehlých hodnot uvnitř skupin a k detekci odlehlých středních hodnot jednotlivých skupin.

Výpočet:

- R – variační rozpětí: $R = X_{\max} - X_{\min}$ (rozdíl nejvyšší a nejnižší hodnoty);
- Q_1 – testovací kritérium pro 1. hodnotu řady: $Q_1 = \frac{x_2 - x_1}{R}$;
- Q_n – testovací kritérium pro poslední hodnotu řady: $Q_n = \frac{x_n - x_{n-1}}{R}$;

1. Hodnoty se seřadí od nejmenší po největší.
2. Podle vzorce se vypočte testovací kritérium pro první a poslední hodnotu řady.

$$Q_1 = \frac{x_2 - x_1}{R} \quad Q_n = \frac{x_n - x_{n-1}}{R}$$


3. V tabulce vyhledáme kritickou hodnotu pro příslušný počet prvků a zvolenou hladinu α .

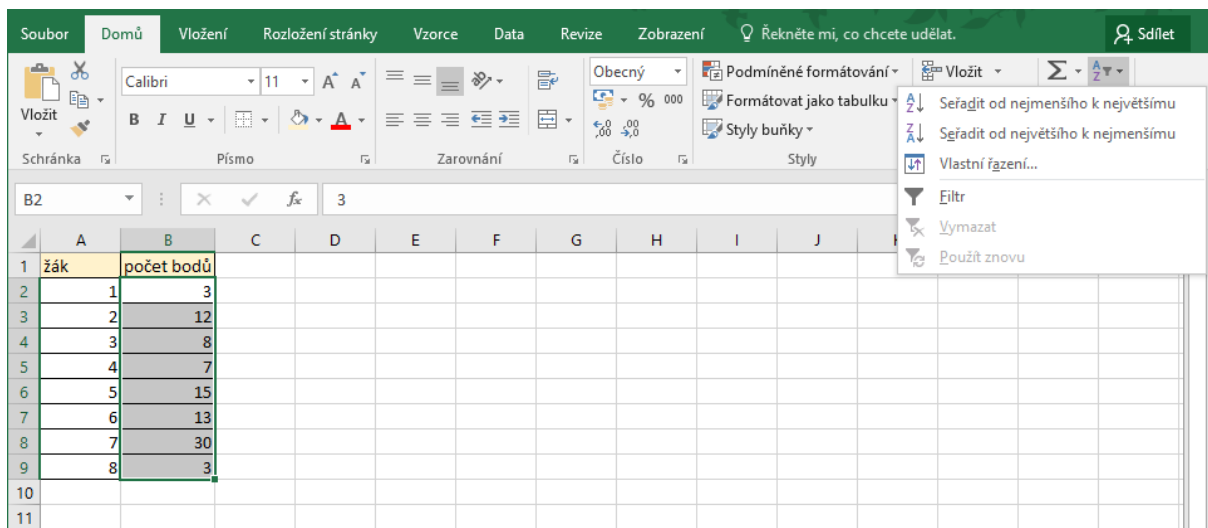
4. Pokud je námi vypočítaná hodnota testovacího kritéria vyšší, než je ta kritická, pak danou hodnotu ze souboru vyřadíme (Bednářová, 2019a).

Dean-Dixonův test v programu MS Excel

V příkladu jsou uvedeny výsledky testu (počet bodů) jednotlivých žáků. Počet žáků je 8 ($n = 8$).

1) Seřadíme hodnoty od nejmenší po největší.

- Označíme výsledky testu – počty bodů. V pravém horním rohu rozklikneme nabídku „Seřadit a filtrovat“ se symbolem  (viz obr. 32).
- Klikneme na „Seřadit od nejmenšího k největšímu“
- Potvrdíme, kliknutím na tlačítko „Seřadit“, že chceme „Rozlišit vybranou oblast“ (Zachová se tak spojení čísla žáka a jím dosaženého počtu bodů.)

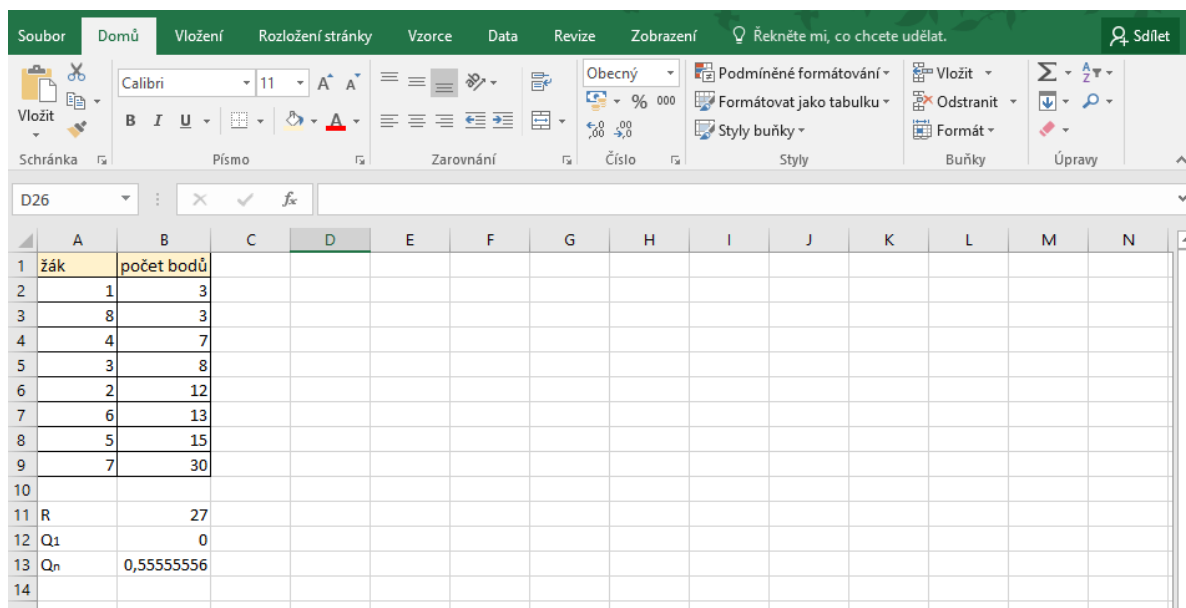


Obr. 32 Řazení hodnot v programu MS Excel

Zdroj: autoři

2) Podle vzorce se vypočte testovací kritérium pro první a poslední hodnotu řady.

- R – variační rozpětí: $R = x_{\max} - x_{\min}$;
- Q_1 – testovací kritérium pro 1. hodnotu řady: $Q_1 = \frac{x_2 - x_1}{R}$;
- Q_n – testovací kritérium pro poslední hodnotu řady: $Q_n = \frac{x_n - x_{n-1}}{R}$.



Obr. 33 Výpočet Dean-Dixonova vzorce v programu MS Excel

Zdroj: autoři

Tab. 12 Popis jednotlivých buněk z obr. 33

buňka	obsah	vzorec
B11	R	=B9-B2
B12	Q ₁	=(B3-B2)/B11
B13	Q _n	=(B9-B8)/B11

Zdroj: autoři

3) V tabulce (Bednářová, 2019b) vyhledáme kritickou hodnotu pro příslušný počet prvků a zvolenou hladinu α .

- V našem případě tedy pro $n = 8$, se zvolenou hladinou 0,05.

Tab. 13 Kritické hodnoty pro výpočet Dean-Dixonova vzorce

n	Q _{krit.} $\alpha = 0,05$	Q _{krit.} $\alpha = 0,01$
3	0,941	0,988
4	0,765	0,889
5	0,642	0,780
6	0,560	0,698
7	0,507	0,637
8	0,468	0,590
9	0,437	0,555

n	Q _{krit.} $\alpha = 0,05$	Q _{krit.} $\alpha = 0,01$
17	0,320	0,416
18	0,313	0,407
19	0,306	0,398
20	0,300	0,391
21	0,295	0,384
22	0,290	0,378
23	0,285	0,372

10	0,412	0,527
11	0,392	0,502
12	0,376	0,482
13	0,361	0,465
14	0,349	0,450
15	0,338	0,438
16	0,329	0,426

24	0,281	0,367
25	0,277	0,362
26	0,273	0,357
27	0,269	0,353
28	0,266	0,349
29	0,263	0,345
30	0,260	0,341

Zdroj: dle Bednářová (2019)

4) Pokud je námi vypočítaná hodnota testovacího kritéria vyšší, než je ta kritická, pak danou hodnotu ze souboru vyřadíme.

- Pokud $Q_{1(n)} \leq Q_{krit.}$ – hodnota není odlehlá a **ponecháme** ji v souboru.
- Pokud $Q_n > Q_{krit.}$ – hodnota je odlehlá a ze souboru ji **vyloučíme**.

Pro tento konkrétní příklad tedy platí:

- $0 \leq 0,468$ ($Q_1 \leq Q_{krit.}$) – první hodnota není odlehlá \Rightarrow **ponecháme** ji v souboru
- $0,556 > 0,468$ ($Q_n > Q_{krit.}$) – poslední hodnota je odlehlá \Rightarrow **vyloučíme** ze souboru.
 - Následně s novým souborem (bez poslední, deváté hodnoty) spočítáme znovu testovací kritérium pro novou poslední (nyní osmou) hodnotu.

Tab. 14 Popis jednotlivých buněk v MS Excel pro výpočet Dean-Dixonova vzorce

buňka	obsah	vzorec
B14	R ₂	=B8-B2
B15	Q _n	=(B8-B7)/B14

Zdroj: autoři

- Opět porovnáme s tabulkou (tentokrát s hodnotou pro $n = 7$)
 - $0,167 \leq 0,507$ ($Q_n \leq Q_{krit.}$) – nová poslední hodnota není odlehlá \Rightarrow **ponecháme** ji v souboru
- Poznámky: V případě, že je n vyšší než 30, srovnáváme vypočítanou hodnotu s kritickou hodnotou pro $n=30$.

5.4. Grubbsův test

Výpočet. Obdobně jako Dean-Dixonův test.

- S_n – směrodatná odchylka. Funkce SMODCH;
- \bar{x} – aritmetický průměr. Funkce PRŮMĚR;
- T_1 – testovací kritérium pro 1. hodnotu řady:

$$T_1 = \frac{\bar{x} - x_1}{S_n}$$

- T_n – testovací kritérium pro poslední hodnotu řady:

$$T_n = \frac{x_n - \bar{x}}{S_n}$$

Postup výpočtu Grubbssonova testu:

1. Hodnoty se seřadí od nejmenší po největší.
2. Podle vzorce se vypočte testovací kritérium pro první a poslední hodnotu řady.


$$T_1 = \frac{\bar{x} - x_1}{S_n} \quad T_n = \frac{x_n - \bar{x}}{S_n}$$

3. V tabulce vyhledáme kritickou hodnotu pro příslušný počet prvků a zvolenou hladinu α .
4. Pokud je námi vypočítaná hodnota testovacího kritéria vyšší, než je ta kritická, pak danou hodnotu ze souboru vyřadíme (Bednářová, 2019a).

Grubbsův test v programu MS Excel

V příkladu jsou uvedeny výsledky testu (počet bodů) jednotlivých žáků. Počet žáků je 8 ($n = 8$).

1) Seřadíme hodnoty od nejmenší po největší.

- Označíme výsledky testu – počty bodů. V pravém horním rohu rozklikneme nabídku „Seřadit a filtrovat“ se symbolem .
- Klikneme na „Seřadit od nejmenšího k největšímu.“
- Potvrdíme, kliknutím na tlačítko „Seřadit“, že chceme „Rozlišit vybranou oblast“ (zachová se tak spojení čísla žáka jeho a jím dosaženého počtu bodů).

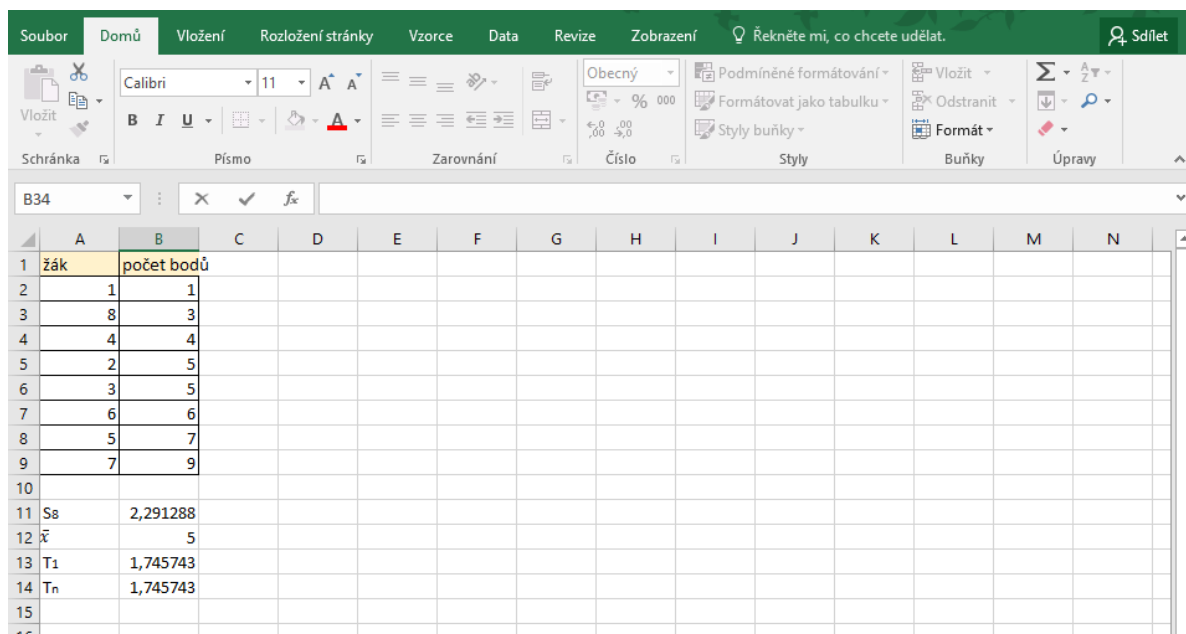
2) Podle vzorce se vypočte testovací kritérium pro první a poslední hodnotu řady.

- S_n – směrodatná odchylka. Funkce SMODCH.
- \bar{x} – aritmetický průměr. Funkce PRŮMĚR.
- T_1 – testovací kritérium pro 1. hodnotu řady.

$$T_1 = \frac{\bar{x} - x_1}{S_n}$$

- T_n – testovací kritérium pro poslední hodnotu řady.

$$T_n = \frac{x_n - \bar{x}}{S_n}$$



Obr. 34 Postup výpočtu Grubbssonova testu

Zdroj: autoři

Tab. 15 Popis jednotlivých buněk z obr. 34

buňka	obsah	vzorec
B11	S_n	=SMODCH(B2:B9)
B12	\bar{x}	=PRŮMĚR(B2:B9)
B13	T_1	=(B12-B2)/B11
B14	T_n	=(B9-B12)/B11

Zdroj: autoři

3) V tabulce (Epina e-Book Team, 2012) vyhledáme kritickou hodnotu pro příslušný počet prvků a zvolenou hladinu α .

Tab. 16 Kritické hodnoty pro výpočet Grubbssonova testu

n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$
3	1,1531	1.1546	15	2,4090	2.7049	80	3,1319	3.5208
4	1,4625	1.4925	16	2,4433	2.7470	90	3,1733	3.5632
5	1,6714	1.7489	17	2,4748	2.7854	100	3,2095	3.6002
6	1,8221	1.9442	18	2,5040	2.8208	120	3,2706	3.6619
7	1,9381	2.0973	19	2,5312	2.8535	140	3,3208	3.7121

8	2,0317	2.2208	20	2,5566	2.8838	160	3,3633	3.7542
9	2,1096	2.3231	25	2,6629	3.0086	180	3,4001	3.7904
10	2,1761	2.4097	30	2,7451	3.1029	200	3,4324	3.8220
11	2,2339	2.4843	40	2,8675	3.2395	300	3,5525	3.9385
12	2,2850	2.5494	50	2,9570	3.3366	400	3,6339	4.0166
13	2,3305	2.6070	60	3,0269	3.4111	500	3,6952	4.0749
14	2,3717	2.6585	70	3,0839	3.4710	600	3,7442	4.1214

Zdroj: autoři dle Epina e-Book Team (2012)

4) Pokud je námi vypočítaná hodnota testovacího kritéria vyšší, než je ta kritická, pak danou hodnotu ze souboru vyřadíme.

- Pokud $T_{1(n)} \leq T_{krit.}$ - hodnota není odlehlá a **ponecháme** ji v souboru.
- Pokud $T_n > T_{krit.}$ - hodnota je odlehlá a ze souboru ji **vyloučíme**.

V tomto konkrétním příkladě tedy dostaneme:

- V tomto případě nám vyšlo T_1 i T_n shodně.
- $1,7457 \leq 2,0317$ ($T_1 \leq T_{krit.}$) - první hodnota není odlehlá \Rightarrow **ponecháme** ji v souboru
- $1,7457 \leq 2,0317$ ($T_n \leq T_{krit.}$) - poslední hodnota není odlehlá \Rightarrow **ponecháme** ji v souboru
- Poznámka: V případě, že některá z vypočítaných hodnot bude větší, než je kritická hodnota, je třeba tuto hodnotu vyřadit ze souboru a znovu test spočítat s novými daty (bez vyřazené hodnoty). Změní se nám kromě poslední (nebo první) hodnoty, průměru a směrodatné odchylky i počet prvků a budeme tedy kritickou hodnotu hledat na jiném řádku.

Grubbsův test v programu Statistica

Statistics – Basic statistics / Tables – Descriptive statistics – Robust – Grubbs test of outliers – Variables (proměnné) – výběr proměnné – OK – Summary

5.5. Rezistentní odhady

Při měření či například přepisu dat mohou vzniknout chyby. Při zpracování dat je tedy výhodné využívat takové odhady, jež jsou vůči odlehlým hodnotám rezistentní.

Např. aritmetický průměr je velmi citlivý vůči odlehlým hodnotám (po jejich vyřazení se výrazně změní) na rozdíl od mediánu (Hendl, 2012). Tab. 17 shrnuje základní rezistentní odhady.

Tab. 17 Shrnutí rezistentních odhadů

Rezistentní odhady	
Střední hodnoty	medián
	percentilový průměr
	useknutý průměr
	winsorizovaný průměr
Rozptýlenosti	směrodatná odchylka

Zdroj: autoři

Mediánu a odchylce je podrobnější pozornost věnována v kapitole 1 (medián, odchylka, průměr).

Useknutý průměr

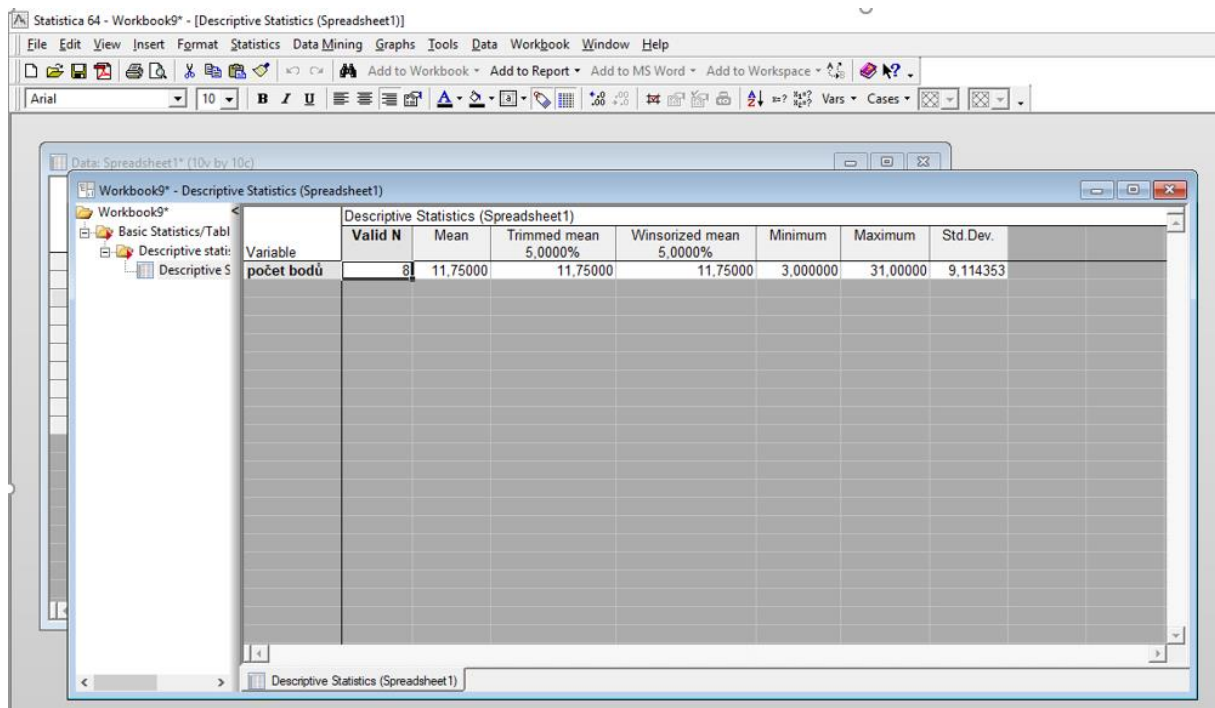
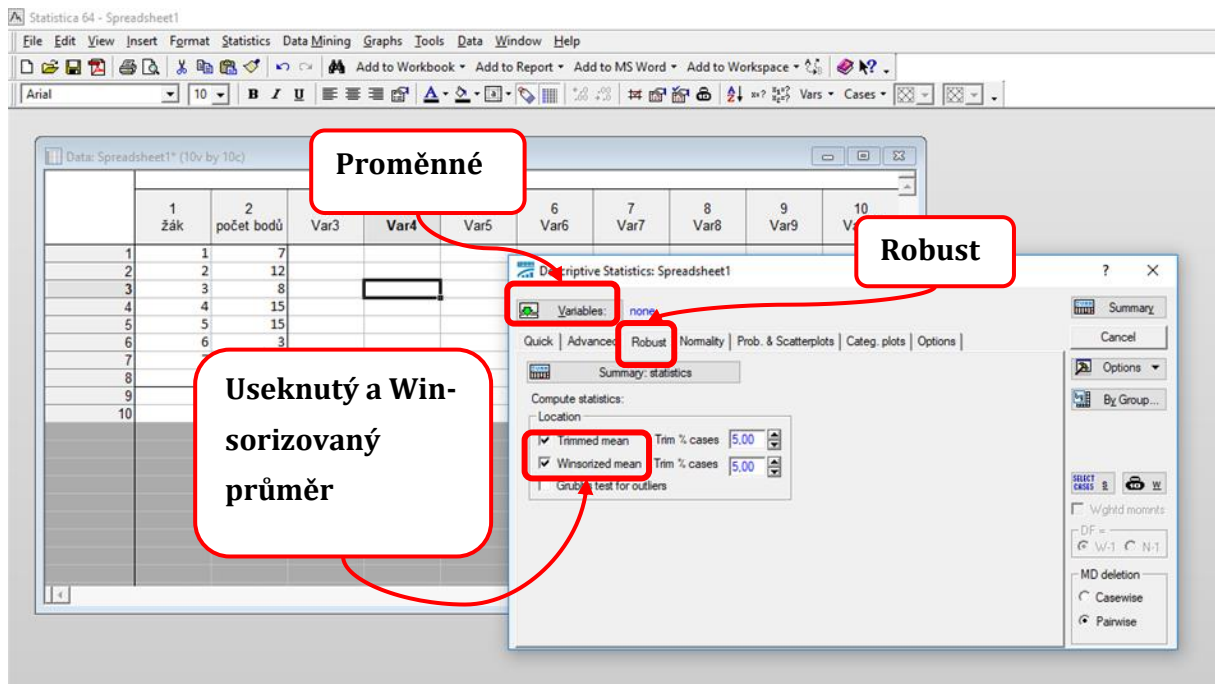
Při výpočtu aritmetického průměru nepočítáme s určitým procentem krajních hodnot. (Hendl, 2012).

Winsorizovaný průměr

Vychází se z toho, že určité části krajních hodnot z množiny dat se přiřadí jedna zvolená méně extrémní hodnota. Z takto upravených dat se následně počítá aritmetický průměr (Hendl, 2012).

Výpočet useknutého a winsorizovaného průměru v programu Statistica

Statistics → Basic Statistics/Tables → Descriptive statistics → Robust → Trimmed mean (useknutý průměr), Winsorized mean (Winsorizovaný průměr) → Variables (proměnné) → zvolení příslušné proměnné → OK → Summary



Obr. 35 Výpočet useknutého a winsorizovaného průměru v programu Statistica

Zdroj: autoři

6. Analýza závislostí

Při statistickém testování hypotéz pracujeme s určitými proměnnými, které mají povahu nominálních, ordinálních či metrických dat (Chytrý, Kroufek, 2017). V některých případech nás při výzkumu zajímá vztah sledovaných proměnných. Zjišťujeme, zda (popř. jak) jedna proměnná ovlivňuje druhou proměnnou, jaká je mezi proměnnými asociace. Zpravidla nás zajímá, zda zvýšení hodnoty jedné proměnné, zároveň zvýší (resp. sníží) hodnotu druhé proměnné.

Teze z předchozího odstavce ilustrujeme na příkladu z tělesné výchovy. Mezi dvě proměnné, mezi kterými budeme chtít zjišťovat vztah (asociaci), patří například hod medicinbalem a hod kriketovým míčkem. Po změření vzdálenosti hodu medicinbalem a kriketovým míčkem u určitého počtu probandů (žáků) lze zjistit, zda vyšší hozená vzdálenost medicinbalem indikuje i vyšší hodnotu hodu kriketovým míčkem, popř. zda vzdálenost hodu medicinbalem ovlivňuje vzdálenost hodu kriketovým míčkem. Dalším příkladem dvou proměnných, mezi kterými lze i logicky nalézt závislost je výška a hmotnost jedince. Čím je jedinec vyšší, velmi často je i jeho hmotnost vyšší v porovnání s menším jedincem.

Ke zjištění vztahu dvou proměnných a jejího ovlivňování jsou určeny různé koeficienty zvané korelační koeficienty. Ve statistické analýze zjišťujeme závislost (korelaci) dvou a více proměnných (např. výška a váha jedince). Korelační koeficienty zpravidla značíme r ; koeficienty zpravidla nabývají hodnot $\langle -1; 1 \rangle$. Pokud se korelační koeficient blíží hodnotě 1, mluvíme o pozitivní korelaci, v případě, že se blíží hodnotě -1, mluvíme o negativní korelaci. Pokud se vypočítaný korelační koeficient pohybuje kolem 0, mezi sledovanými proměnnými není vztah (závislost). Interpretace těchto hodnot bude vysvětlena dále v textu.

V následujících částech bude pojednáno o základních pravidlech užití nejpoužívanějších korelačních koeficientů a jejich výhod či nevýhod oproti jiným koeficientům. Autoři budou primárně odkazovat na konkrétní příklady užití jednotlivých koeficientů v praxi a jejich výpočet v programu MS Excel či programu Statistica. V těch případech, kdy uvedené koeficienty v těchto programech počítat nelze, uvede vztah výpočtu (vzorec), který ilustrují dosazením konkrétních hodnot.

6.1. Zobrazování dvojrozměrných dat

Před samotným užitím korelačních koeficientů je nutné data z tabulky graficky zobrazit. V podstatě se jedná o grafickou interpretaci dat pomocí vizuálí. Jedná se o první krok analýzy závislosti dvou proměnných. Právě volba vhodné vizuálie (možnosti zobrazení číselných dat) nám umožňuje vystihnout tendence dat. Nejčastěji se v statistice využívá tzv. dvojrozměrného bodového grafu, ale ve specifických případech lze užít i jiné typy vizuálí.

V současnosti máme pro zobrazení dat mnoho počítačových programů, a tak není nutné používat milimetrový papír. Zobrazování dvojrozměrných dat bude uvedeno na příkladu MS Excel. V tab. 18 jsou uvedeny hodnoty naměřené mezi 19 vysokoškolskými studenty (muži) v rámci Cooperova testu v plavání a v běhu. Tento test trvá 12 minut a zjišťuje vytrvalostní schopnosti jedince. Výsledkem testu je potom hodnota vzdálenosti (např. v metrech), kterou jedinec během 12 minut uplave, resp. uběhne. Je nutné upozornit, že data v každém řádku tabulky se vztahují pouze k jednomu probandovi; například z tabulky můžeme vyčíst, že proband 5 uplavala za 12 minut 430 metrů a uběhnul 2 900 metrů. Je nutné upozornit, že při měření je nutné pohlídat spárování hodnot (určitěmu probandovi odpovídá právě jedna hodnota uplavané vzdálenosti a právě jedna hodnota uběhnuté vzdálenosti).

Tab. 18 Uplavaná a uběhnutá vzdálenost v Cooperově testu u mužů

Proband	Uplavaná vzdálenost (v m)	Uběhnutá vzdálenost (v m)
Proband 1	350	2800
Proband 2	390	2000
Proband 3	420	2200
Proband 4	430	2400
Proband 5	430	2900
Proband 6	440	2900
Proband 7	450	2900
Proband 8	460	3350
Proband 9	470	3280
Proband 10	470	2900
Proband 11	470	2900
Proband 12	540	3090
Proband 13	550	3110
Proband 14	570	2960
Proband 15	590	3000
Proband 16	620	2880
Proband 17	700	2940
Proband 18	810	3350
Proband 19	912	3650

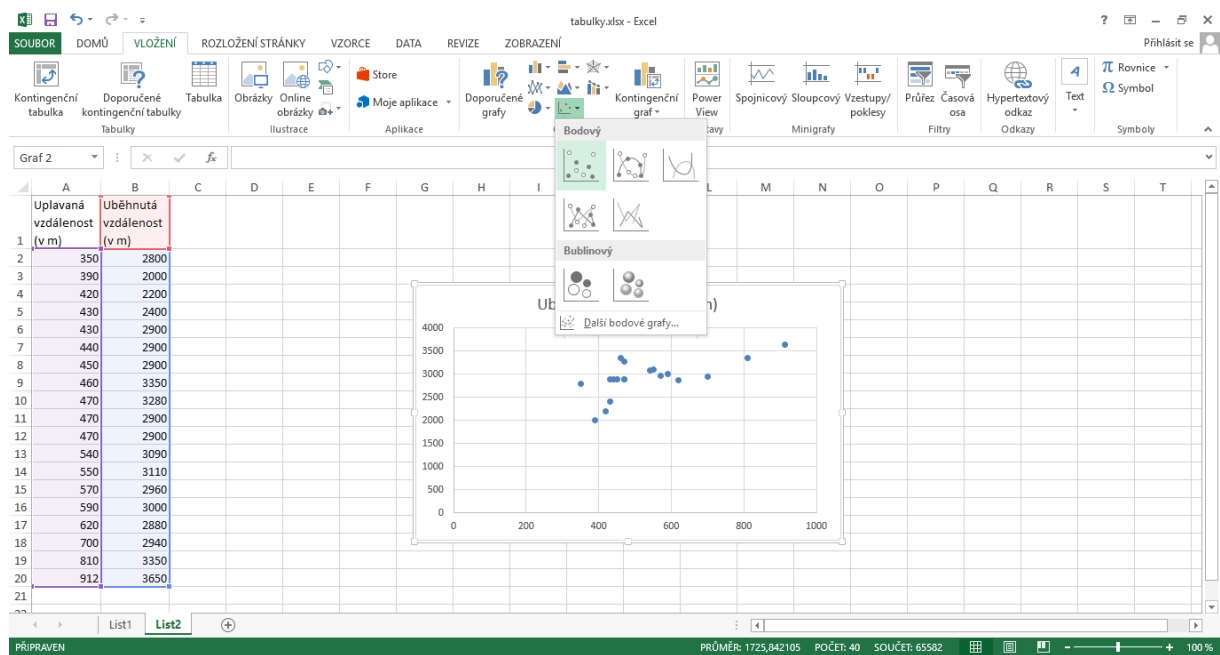
Zdroj: autoři

Pro zobrazení těchto dat ve dvojrozměrném (má pouze osu x a y) bodovém grafu zvolíme v MS Excel volbu *vložit* → *graf* → *XY bodový* (viz obr. 36). Nyní vidíme vynesení obou hodnot

každého jedince (plavání a běh) do dvojrozměrné soustavy souřadnic (viz obr. 37); jeden bod odpovídá jednomu páru měření, tedy jednomu probandovi (jednomu řádku v tabulce). Další vhodnou metodou, jak doplnit tento graf, je zobrazení lineární spojnice trendu. Tu lze přidat v záložce *návrh* → *přidat prvek grafu* → *spojnice trendu* → *lineární*. Lineární spojnice trendu ukazuje průměrný nejmenší rozdíl mezi naměřenými hodnotami dvou proměnných (uplavané a uběhnuté vzdálenosti). Dalšími výpočty (korelačními koeficienty) se potom zjišťuje, jak těsný je vztah mezi zobrazenými body a přímkou („jak blízko“ jsou body k přímce).

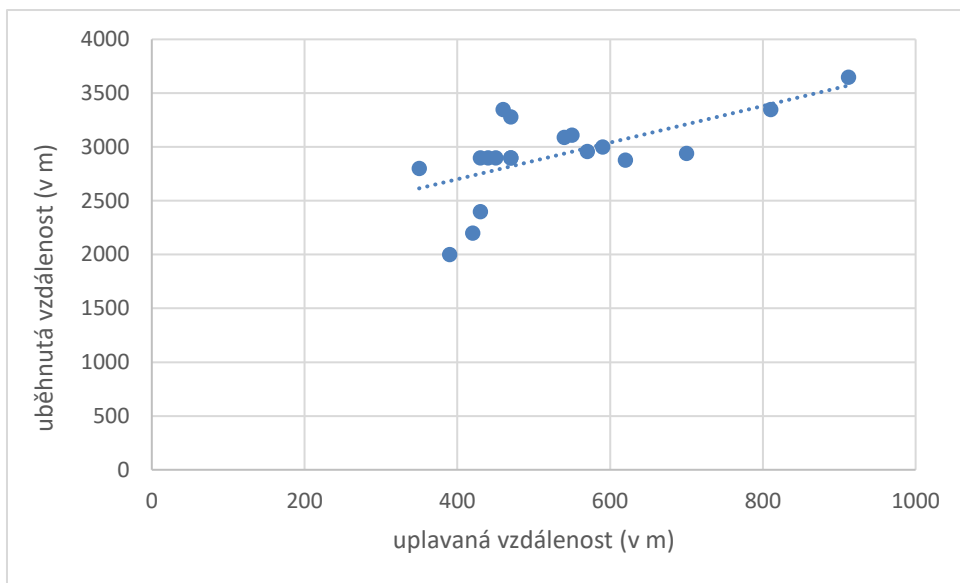
Již prvotní pohled na graf (obr. 37) nám ukazuje, že určitý vztah mezi uplavanou a uběhnutou vzdáleností by mohl existovat; pro srovnání jsou na obr. 38 uvedeny hodnoty, u kterých závislost není patrná. U obr. 37 je patrné, že čím více jedinec uplaval metrů, většinou i více metrů uběhl. Z obr. 38 tento vztah patrný není.

V případech, kdy bychom měli více informací o zobrazených datech (např. věk probandů, pohlaví atp.), lze body v grafu i odlišit různými vyjadřovacími prostředky, např. tvarem či barvou znaku (Kirk, 2016).



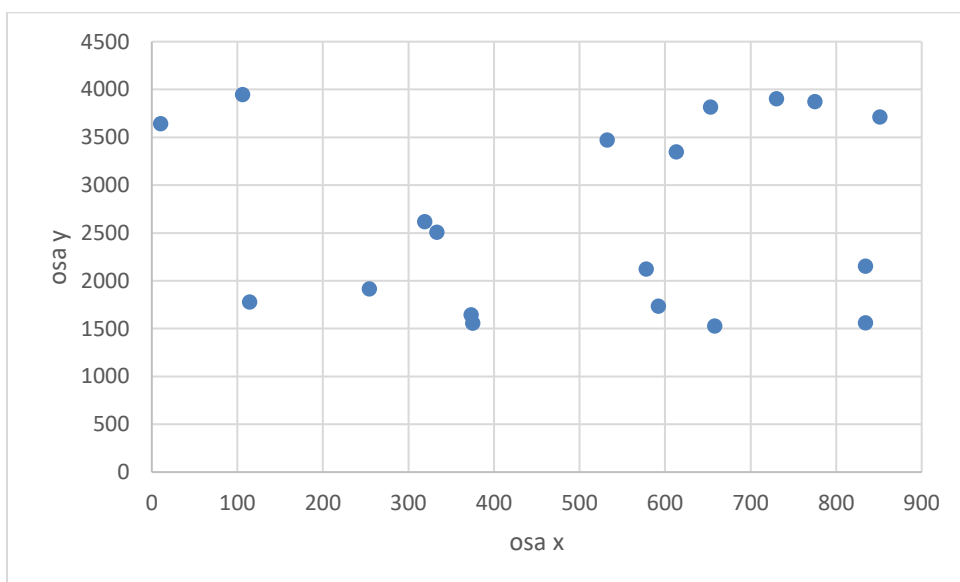
Obr. 36 Zobrazení dvojrozměrných dat v MS Excel

Zdroj: autoři



Obr. 37 Graf z dat z tab. 18 s lineární spojnicí trendu vytvořený v programu MS Excel

Zdroj: autoři



Obr. 38 Graf s daty, mezi kterými neexistuje závislost

Zdroj: autoři

Další variantou, jak „graficky“ zobrazit data je tzv. korelační tabulka. Do ní jsou naměřené hodnoty zaneseny ve formě počtu hodnot v určitém intervalu. Příklad korelační tabulky z naměřených hodnot z tab. 18 je uveden v tab. 19. Tabulka vznikla rozdělením datové řady (tj. sloupce dat) do intervalů (ideálně 3–5, dle počtu naměřených hodnot); dále byla zjištěna četnost probandů v určité dvojici intervalů (uplavané a uběhnuté vzdálenosti). Například v intervalech do 2 300 uběhlých metrů a zároveň uplavané vzdálenosti do 450 metrů, se nachází 2 probandi. Dále v intervalech v uběhlé vzdálenosti 2 301–3 000 metrů a zároveň uplavané vzdálenosti 451–599 metrů se nachází 4 probandi atd.

I z této tabulky je patrné, že určitý vztah mezi uplavanou a uběhnutou vzdáleností pravděpodobně existuje. Vztah mezi dvěma proměnnými (uplavané a uběhlé vzdálenosti) poznáme tak, že vyšší počet probandů v jednotlivých intervalech se koncentruje spíše zeleně označených okénkách tabulky. Naopak v intervalech, který je v tabulce označen červenou barvou, se nenachází žádný proband.

Tab. 19 Korelační tabulka z dat z tab. 18

Uběhnutá vzdálenost (v m)	Uplavaná vzdálenost (v m)			Celkem
	Do 450	451-599	600 a více	
do 2 300	2	0	0	2
2 301-3 000	5	4	2	11
3 001 a více	0	4	2	6
Celkem	7	8	4	19

Zdroj: autoři

6.2. Korelační analýza

Korelační analýza umožňuje zjistit vztah mezi dvěma a více proměnnými. Výzkumníci pomocí této analýzy zjišťují tendence jedné proměnné se vyskytovat společně s určitými hodnotami druhé proměnné.³ Tendence může mít charakter neexistence korelace (např. hodnoty proměnné X neodpovídají hodnotě druhé proměnné Y) až po tzv. absolutní korelaci, kdy s danou proměnou hodnoty X se vyskytuje určitá hodnota Y . Například neexistence korelace mohou mít proměnné IQ a běh na 100 metrů; pravděpodobně nebude znamenat, že vyšší míra inteligence znamená i lepší čas v běhu na 100 metrů. Naopak vysoké hodnoty korelace budou pravděpodobně mít již zmíněné proměnné výška a hmotnost jedince. Vyšší jedinec pravděpodobně bude mít i vyšší hmotnost.

Korelační koeficienty lze v základu počítat pro „všechna“ data, avšak jako v jiných statistických metodách dosáhneme zavádějících až chybných výsledků. Pro užití korelačních koeficientů existují obecná pravidla, která je nutno respektovat (viz např. Hendl, 2012).

1. **Má výpočet korelace smysl?** V některých případech nemá počítání závislosti dvou proměnných smysl. Například výpočtem korelačního koeficientu mezi délkou pažní kosti jedince a hodnotou IQ sice dosáhneme určitého výsledku, avšak nějaký vztah

³ Jak bylo zmíněno v úvodu: zjišťujeme, zda zvýšení hodnoty jedné proměnné může znamenat i zvýšení hodnoty druhé proměnné.

mezi těmito dvěma proměnnými nelze očekávat. Vždy je tedy nutné posoudit logickou souvislost proměnných.

2. **Formální korelace** vzniká u dvou proměnných, které se doplňují do 100 %. Jedná se například o procentuální zastoupení chlapců a dívek ve třídě. Pokud je podíl chlapců vyšší, logicky bude podíl dívek nižší na celkovém počtu žáků ve sledovaných třídách. Takové hodnoty nemá smysl mezi sebou počítat, protože závislost bude logicky negativní a vysoká (pokud se zvýší podíl chlapců, zároveň se sníží podíl dívek ve třídě).
3. **Vliv nehomogenity výzkumného souboru.** V některých případech může interpretaci korelace ztížit nehomogenita souboru, např. pokud se výzkumný soubor skládá z dílčích velmi nehomogenních celků. Například pokud budeme zjišťovat závislost běhu na 100 metrů a hmotnosti jedince u mužů a žen současně (v rámci jednoho výzkumného souboru). Výzkumný soubor se bude skládat z nehomogenních částí, protože muži a ženy mají rozdílné předpoklady pro běh na 100 metrů (např. úroveň silových schopností, podíl rychlých svalových vláken atd.). Výsledek korelační analýzy by byl zatížen chybou.
4. Korelace **způsobené společnou příčinou.** Jedná se například o zjišťování délky paže pravé a levé ruky, která je způsobena stejnou příčinou – tělesnými proporcemi a výškou postavy (primárně genetické dispozice). Závislost mezi těmito dvěma proměnnými bude logicky dosahovat vysokých hodnot, avšak její význam bude prakticky nulový.
5. **Vliv rušivé proměnné.** V některých případech mohou dvě sledované proměnné být ovlivněny třetí proměnnou, která s nimi koreluje. Interpretace korelace je potom zatížena určitou chybou. Typickým příkladem je věk, který ovlivňuje většinu sledovaných jevů v pedagogice i psychologii. Jak se dozvíme dále, tento faktor lze v některých případech alespoň částečně eliminovat.

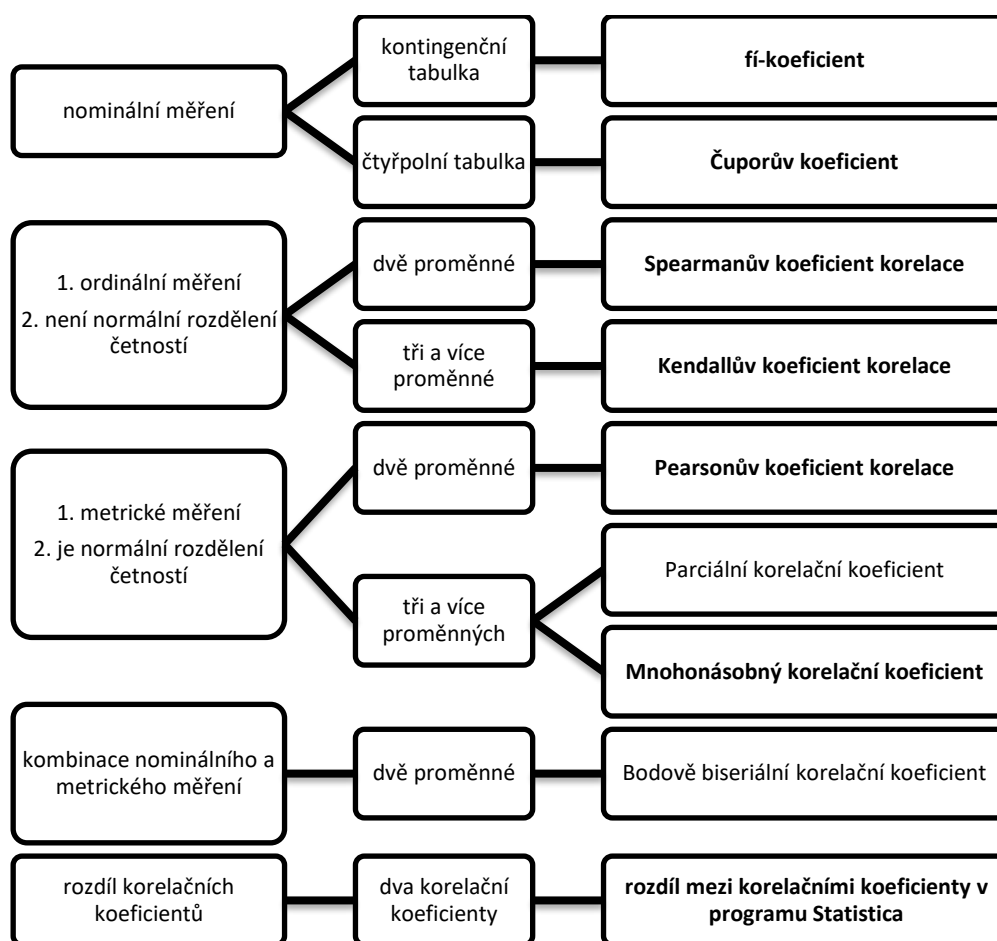
Koeficienty korelace většinou nabývají hodnot $<-1; 1>$. Čím se hodnota blíží 1 (resp. -1), tím je závislost proměnných vyšší, čím se hodnoty koeficientů blíží 0, tím je závislost proměnných menší. Pokud je koeficient roven 0, neexistuje mezi proměnnými žádná závislost. Prvotní interpretace hodnoty vypočítaného koeficientu (před testováním statistické významnosti) dává výzkumníkovi první informaci o stupni korelace a první podklad pro interpretaci výsledku. Jednotliví autoři rozdělují sílu asociace dvou proměnných od malé po velkou asociaci, avšak konkrétní hodnoty se u jednotlivých autorů liší. V tab. 20 je uveden přístup dvou vybraných autorů.

Tab. 20 Síla asociace proměnných dle různých autorů

Síla asociace	Hendl (2012)	Chráska (2016)
Nulová až velmi nízká	0,0	0,0
malá, nízká	0,1–0,3	0,2–0,4
střední	0,3–0,7	0,4–0,7
velká, vysoká	0,7–1,0	0,7–1,0

Zdroj: autoři dle Hendl (2012) a Chráska (2016)

V dalších částech budou uvedeny nejpoužívanější korelační koeficienty s příklady, ve kterých by bylo možno tyto koeficienty použít. Dále bude uveden alespoň stručný návod pro výpočet korelačních koeficientů v programu MS Excel a Statistica. Přehled popsanych korelačních koeficientů a kritéria výběru těchto koeficientů jsou uvedeny na obr. 39.



Obr. 39 Postup výběru vhodného korelačního koeficientu

Zdroj: autoři na základě Hendl (2012) a Chráska (2016)

Poznámka: o tučně zvýrazněných korelačních koeficientech je pojednáno dále v textu. Podstatu a výpočet ostatních koeficientů lze dohledat v odborné literatuře (např. Hendl, 2012; Chráska, 2016).

Pearsonův korelační koeficient

Pearsonův korelační koeficient (r) je jeden z nejpoužívanějších korelačních koeficientů.⁴ Při jeho užití je ale však nutné respektovat určitá pravidla jeho použití. V první řadě je nutné mít k dispozici metrická data. V případě nominálních či ordinálních dat se užívají jiné výpočty korelace. Dále je nutné mít k dispozici spojitě lineární proměnné. U nespojitých proměnných je výpočet zatížen chybou; s tímto kritériem použití souvisí i nevýhoda tohoto koeficientu – silné ovlivnění odlehlými hodnotami. Ty je nutné určitými matematickými operacemi odstranit. V odborné literatuře je i doporučováno tento koeficient užít pouze v případě náhodného výběru. Samozřejmostí je, že každé hodnotě odpovídá právě jedna hodnota druhé proměnné (tzn., že máme k dispozici vždy dvojice hodnot); ty musí mít normální rozdělení četností.

Pearsonův korelační koeficient, stejně jako ostatní korelační koeficienty, nabývá hodnot <1 ; 1 >. Hodnota menší než 0 značí negativní korelaci (čím vyšší hodnota jedné proměnné indikuje nižší hodnotu druhé proměnné) a naopak hodnota větší než 0 značí pozitivní korelaci (čím vyšší je hodnota jedné proměnné, tím vyšší je hodnota druhé proměnné). Vysoká hodnota korelačního koeficientu ale nemusí znamenat vysokou míru korelace. Například, pokud počítáme závislost mezi pěti dvojicemi hodnot, i vyšší hodnota korelačního koeficientu nemusí být statisticky významná.

Jak již bylo zmíněno v předchozím odstavci, klíčový význam pro interpretaci výsledků výzkumu je testování statistické významnosti. Jedná se o zjišťování, zda korelační koeficient je natolik vysoký, abychom mohli hovořit o statisticky významném vztahu. Tabulkové testové kritérium k tomuto koeficientu zjistíme z přílohy 1. Samozřejmě, že v případě užití programu Statistica se tyto tabulky nepoužívají, protože program nám tyto hodnoty porovná sám. Vypočítanou hodnotu korelačního koeficientu porovnáme s tabulkovou hodnotou, kterou získáme dle vztahu:

$$f = n - 2,$$

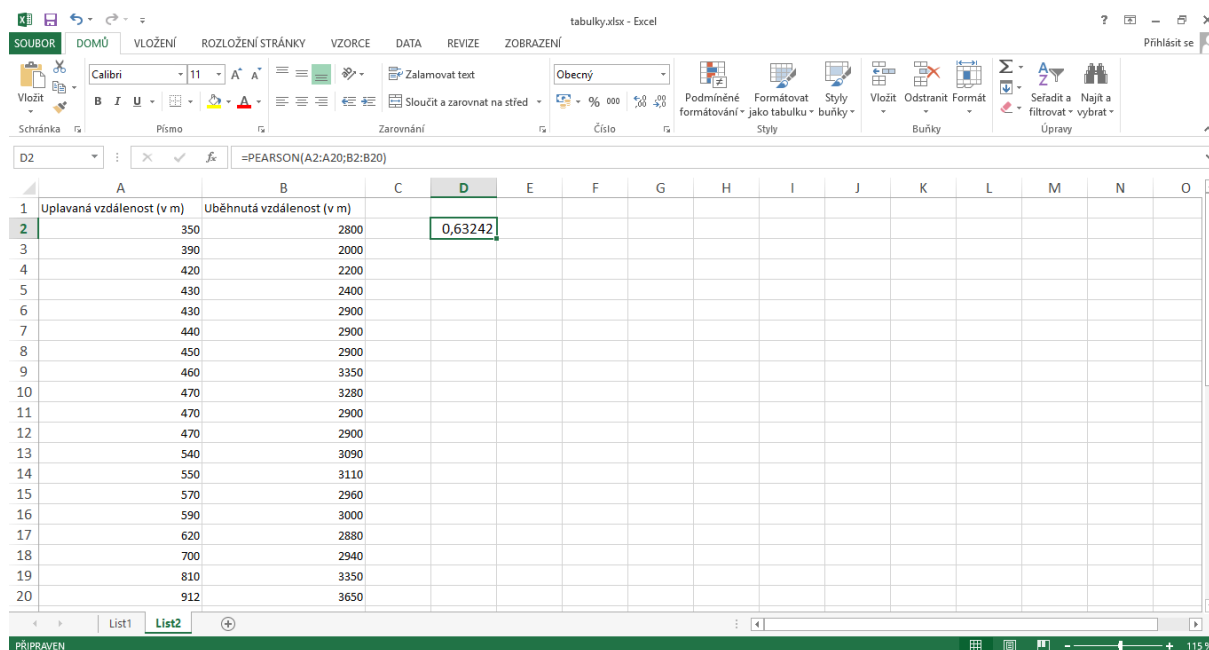
kde f je počet stupňů volnosti⁵ a n je počet porovnávaných dvojic.

⁴ Je označován také jako koeficient součinné korelace.

⁵ Podle hodnoty stupňů volnosti lze najít v tabulkách (např. Hendl, 2012, příloha 2) testové kritérium, které porovnáme s vypočtenou hodnotou. Pokud je vypočtená hodnota vyšší, nepotvrzujeme nulovou hypotézu o neexistenci závislosti mezi dvěma proměnnými. Pokud je vypočtená hodnota nižší než tabulková hodnota zjištěná podle stupňů volnosti, potvrzují nulovou hypotézu.

Použití Pearsonova koeficientu ilustrujeme na příkladu, který vychází z dat z tab. 18. Nebudeme užívat dosažení do vzorce, který je bez problémů dohledatelný v citované literatuře, nýbrž objasníme postup výpočtu v MS Excel a programu Statistica.

Postup v MS Excel: Vložíme funkci *Pearson*,⁶ v rámci které vybereme první proměnnou, oddělíme středníkem a vybereme druhou proměnnou (viz obr. 40). Vypočtenou hodnotu porovnáme s tabulkovou hodnotou. Jelikož máme celkem 19 probandů je počet stupňů volnosti 17 (tj. $f = 19 - 2$). Vypočtená hodnota ($r = 0,632$) je vyšší než tabulková (tj. 0,4556) – (viz příloha 1), a proto nelze potvrdit nulovou hypotézu o neexistujícím vztahu (závislosti) mezi uplavanou a uběhnutou u vysokoškolských studentů na hladině významnosti $\alpha = 0,05$.

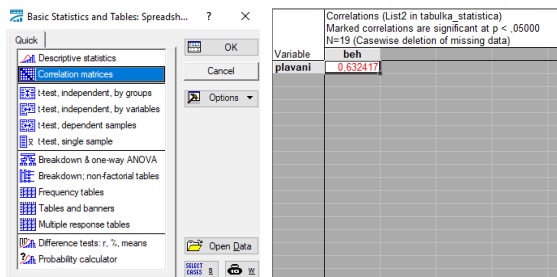


Obr. 40 Výpočet Pearsonova korelačního koeficientu v MS Excel

Zdroj: autoři

Postup v programu Statistica je velmi podobný. V tabulce máme hodnoty uplavané a uběhnuté vzdálenosti. Výpočet se provede zvolením volby *Statistics* → *basic statistics and tables* → *correlation matrices* (viz obr. 41) → *vybereme proměnné (plavání a běh)* a následně se nám zobrazí výsledek (viz obr. 41). Z výpočtu je patrné, že již nemusíme porovnávat vypočtenou hodnotu s tabulkami. Červeně zvýrazněná korelace (číslo) značí, že vypočtená hodnota je statisticky významná na hladině významnosti $\alpha = 0,05$, a proto nepotvrzujeme hypotézu o neexistenci vztahu.

⁶ Postup vložení funkce: do buňky v MS Excel napíšeme: *=pearson(hodnoty první proměnné; hodnoty druhé proměnné)*.



Obr. 41 Výpočet Pearsonova korelačního koeficientu v programu Statistica

Zdroj: autoři

Pokud nepotvrdíme nulovou hypotézu o neexistenci vztahu mezi dvěma proměnnými (tzn., potvrdíme alternativní hypotézu o existenci korelace), lze počítat i tzv. věcnou (praktickou) významnost neboli effect size (r^2). Ta nám udává, z kolika procent je ovlivněna závislost mezi sledovanými proměnnými. Věcná významnost je druhou mocninou koeficientu korelace (r). Tento výpočet ilustrujeme na předchozím příkladu. Druhá mocnina $r = 0,632$ je $r^2 = 0,399$. Lze tedy prohlásit, že závislost výkonnosti Cooperova testu v plavání a běhu je ovlivněna z 39,9 %. Výsledek lze interpretovat tak, že u sledované skupiny probandů je výkon ovlivněn z 39,9 % prostředím, ve kterém Cooperův test realizují a ve zbývajících 60,1 % připadá na další, nezjištěné činitele.

Mnohonásobný koeficient korelace

Mnohonásobný koeficient korelace se užívá v případech, kdy máme více jak dvě proměnné získané prostřednictvím metrického měření. Tímto koeficientem hodnotíme vliv několika proměnných na cílovou proměnnou; tzn., jaký mají vliv dvě a více proměnných na určitou proměnnou. Koeficient vypočítáme dle vztahu:

$$p_{x.yz} = \sqrt{\frac{p_{xy}^2 + p_{xz}^2 - 2p_{xz}p_{xy}p_{yz}}{1 - p_{yz}^2}}$$

Výpočet koeficientu ilustrujeme na příkladu. Z tělesné výchovy je známo, že motorické dovednosti jedince závisí na různých znacích jedince. Na závěr lyžařského kurzu se změřil čas ve slalomu u 36 dívek. Dále se u nich zjišťovaly další charakteristiky: test rovnováhy a test sociální úzkosti. Korelační koeficienty mezi sledovanými proměnnými vidíme v tab. 21.

Tab. 21 Korelační koeficienty mezi sledovanými jevy

	x	y	z
Čas ve slalomu (x)	1,00	-0,34	0,46
Test rovnováhy (y)	-0,34	1,00	0,45
Test sociální úzkosti (z)	0,46	0,45	1,00

Zdroj: dle Hendl (2012)

Mnohonásobný koeficient korelace má hodnotu:

$$r_{x,yz} = 0,342 + 0,462 - 2 \cdot (-0,34) \cdot 0,46 \cdot 0,45 = \sqrt{\frac{0,34^2 + 0,46^2 - 2(-0,34)(0,46)(0,45)}{1 - 0,45^2}} = 0,77$$

Statistickou významnost (testování nulové hypotézy o neexistenci vztahu) testujeme pomocí F-testu:

$$F = \frac{r_{x,yz}^2(n-3)}{2(1-r_{x,yz}^2)}$$

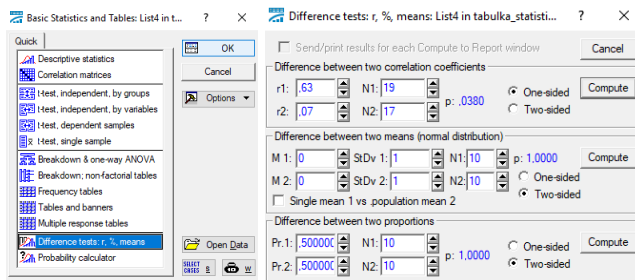
Pomocí tabulek F rozdělení se stupni volnosti 2 a $n - 3$ zjišťujeme tabulkovou hodnotu a porovnáváme ji s vypočtenou hodnotou. Vypočtená hodnota je vyšší než tabulková hodnota na hladině významnosti $\alpha = 0,05$, proto lze nepotvrdit nulovou hypotézu o neexistenci vztahu mezi sledovanými proměnnými. Čas ve slalomu je ovlivněn sociální úzkostí jedince a rovnováhovou schopností.

Testování rozdílu mezi dvěma koeficienty korelace

V některých případech máme k dispozici dva vypočítané koeficienty součinné korelace a chceme rozhodnout, zda mezi nimi je rozdíl. Sledujeme, zda míra závislosti jedné dvojice proměnných je stejná, resp. rozdílná jako druhá dvojice proměnných. Podmínkou samozřejmě je, že koeficient musí být vypočten pro stejné dvojice proměnných (např. výšku a hmotnost dvou skupin jedinců).

Využití výpočtu ilustrujeme na příkladu. Ve výzkumu jsme měřili Cooperovým testem uběhnutou a uplavanou vzdálenost u 19 mužů a 17 dívek. Pearsonův koeficient korelace mezi uplavanou a uběhnutou vzdáleností činil u mužů $r_1 = 0,632$ a u žen $r_2 = 0,073$. Výpočet lze provést v programu Statistica: *Statistics* → *basic statistics and tables* → *difference tests* → do prvního řádku s nadpisem *difference between two correlation coefficients* zadáme oba korelační koeficienty (r_1 a r_2) a zároveň počet testovaných probandů (N_1 a N_2) → označíme *one-sided* → *Compute* (viz obr. 42). Všimáme si hodnoty p , která je v našem případě menší než 0,05 (konkrétně $p = 0,038$); to znamená, že rozdíl mezi korelačními koeficienty uplavané

a uběhlé vzdálenosti u mužů a žen lze označit jako statisticky významný na hladině významnosti $\alpha = 0,05$.



Obr. 42 Výpočet rozdílu mezi korelačními koeficienty v programu Statistica

Zdroj: autoři

Spearmanův koeficient korelace

Tento korelační koeficient se nazývá také koeficient pořadové korelace. Je určen zejména pro ty data, která jsou zachycena pomocí ordinálního měření. Dále tento koeficient užíváme v případech, kdy nemůžeme předpokládat linearitu vztahu dvou proměnných, či nebylo zjištěno normální rozdělení hodnot. V těchto případech místo Pearsonova korelačního koeficientu užíváme Spearmanův.

Podstatou tohoto koeficientu je určení pořadí hodnot. Například vytváříme pořadí školních známek, bodového hodnocení písemných prací či schopnosti si zapamatovat určité věci atd. Na základě porovnání pořadí jedinců v rámci určitých dvou kritérií zjistíme, zda sledované proměnné mají mezi sebou nějaký vztah, tj. zda mezi sebou korelují.

Výpočet ilustrujeme na příkladu. Výzkumník hodnotil mapy povrchu Česka v učebnicích vlastivědy pomocí metody škálování dle předem stanovených kritérií. Výsledky škálování dvou vybraných kritérií (názornosti a odbornosti) jsou po přepočtu na jednotnou posuzovací stupnici uvedeny v tab. 22.

Tab. 22 Výsledky škálování dvou kritérií (názornosti a odbornosti) včetně uvedení jejich pořadí mezi různými nakladateli

	SPN	Nová škola	Nová škola-Duha	Taktik	Septima	Nakladatelství ČGS	Alter
Míra názornosti	119,85	68,49	128,41	68,49	77,05	94,17	85,61
Pořadí názornosti (pořadí)	2	6,5	1	6,5	5	3	4
Míra odbornosti	90,92	45,46	90,92	38,97	87,68	84,43	71,44
Pořadí odbornosti (pořadí)	1,5	6	1,5	7	3	4	5
Rozdíl pořadí (d)	0,5	0,5	-0,5	-0,5	2	-1	-1
Druhá mocnina rozdílu (d_i^2)	0,25	0,25	0,25	0,25	4	1	1

Zdroj: autoři

V prvním kroku je nutné určit pořadí každého nakladatelství v rámci každého kritéria. Pokud je míra plnění kritéria stejná u více nakladatelství, přiřadíme jim průměrnou hodnotu příslušných pořadí (pořadí sečteme a vydělíme počtem sečtených pořadí). V další fázi výpočtu pořadí odečteme a určíme druhou mocninu jejich rozdílu (viz tab. 22). Po sečtení druhých mocnin rozdílu (d^2) dostaneme hodnotu 7, kterou dosadíme do vzorce:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 7}{7 \cdot (49-1)} = 0,87$$

Po dosazení nám vyjde $r_s = 0,87$.

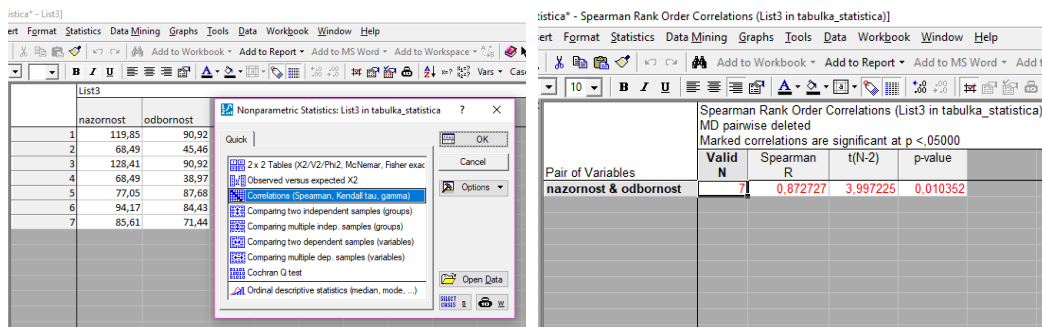
Pokud se jedná o náhodný výběr ze základního souboru, lze porovnávat výsledek pořadové korelace s kritickou hodnotou v tabulkách (viz příloha 2). Stupně volnosti určíme dle vztahu:

$$v = n - 2.$$

Hodnota $r_s = 0,87$ je vyšší než tabulková hodnota (0,714), proto nelze potvrdit nulovou hypotézu o neexistenci vztahu mezi mírou názornosti a odbornosti u schémat výškových stupňů zemského povrchu. Tyto dvě kritéria na sobě závisí.

Výpočet samozřejmě lze provést i v programu Statistica. Po vynesení hodnot do tabulky zvolíme *statistics* → *nonparametric statistics* → *correlations* (viz obr. 43) → poté zvolíme způsob

zobrazení dat a volbu proměnných → OK. Z obr. 43 vidíme, že hodnota je zvýrazněna červeně, což znamená, že musíme přijmout alternativní hypotézu o existenci vztahu (závislosti) mezi mírou odbornosti a názornosti u schémat výškových stupňů.



Obr. 43 Výpočet Spearmanova koeficientu korelace v programu Statistica

Zdroj: autoři

Stejně jako u Pearsonova korelačního koeficientu lze určit praktickou významnost prostřednictvím koeficientu determinace. To určíme jako druhou mocninu Spearmanova korelačního koeficientu, tedy v našem případě $0,87^2$. Výsledek je $r_s^2 = 0,76$. Míra odbornosti schémat výškových stupňů zemského povrchu je ovlivněna ze 76 % mírou názornosti vizuálie.

Výpočet Spearmanova korelačního koeficientu ilustrujeme i na intervalových proměnných. Tab. 23 udává obvod pasu a krku u dvaceti probandů. Zjišťujeme, zda je mezi těmito proměnnými nějaký vztah (korelace, závislost).

Tab. 23 Obvod pasu a krku u dvaceti probandů

	pas (v cm)	pořadí (pas)	krk (v cm)	pořadí (krk)	rozdíl pořadí (d)	druhá mocnina rozdílu pořadí (d_i^2)
proband 1	97	1	38	1	0	0
proband 2	73	7	31	17	-10	100
proband 3	68	13	34	7	6	36
proband 4	63	18	33	10	8	64
proband 5	60	20	31,5	15	5	25
proband 6	72	8	34	7	1	1
proband 7	65	16	33,5	9	7	49
proband 8	77	4	35	6	-2	4
proband 9	64	17	32	13	4	16
proband 10	62	19	31,8	14	5	25
proband 11	70	10	30	19	-9	81
proband 12	71	9	31,3	16	-7	49
proband 13	67	14	32,5	11	3	9
proband 14	75	6	32,3	12	-6	36
proband 15	76	5	35,5	4	1	1
proband 16	80	3	37	2	1	1
proband 17	85	2	36	3	-1	1
proband 18	70	10	35,1	5	5	25
proband 19	69	12	29	20	-8	64
proband 20	66	15	30,9	18	-3	9

$\Sigma = 596$

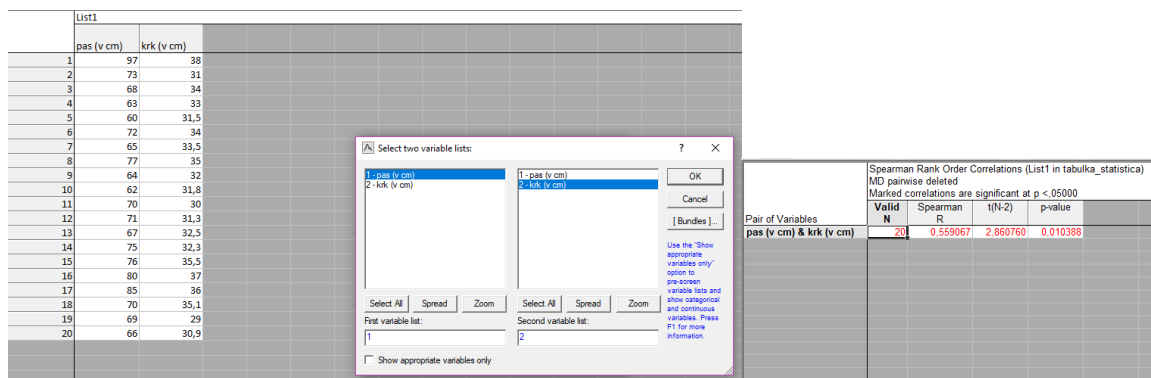
Zdroj: autoři

Stejně jako v předchozím příkladu, vypočtené hodnoty dosadíme do vzorce:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 596}{20 \cdot (400-1)} = 0,55$$

Výsledek ($r_s = 0,55$) srovnáme s tabulkovou hodnotou v příloze 2. Výsledek je vyšší než tabulková hodnota ($0,55 > 0,399$), proto nelze potvrdit nulovou hypotézu o neexistenci vztahu mezi obvodem pasu a krku u námi měřených dvaceti probandů. Tyto dvě proměnné na sobě závisí. Obvod pasu a krku na sobě závisí z 30 %; obvod pasu je z 30 % ovlivněn obvodem krku ($r_s^2 = 0,55^2 = 0,30$).

Výpočet příkladu je ilustrován i v programu Statistica (viz obr. 44).



Obr. 44 Výpočet Spearmanova korelačního koeficientu v programu Statistica (závislost obvodu pasu a krku)

Zdroj: autoři

Kendallův koeficient shody

Kendallův korelační koeficient měří vztah více než dvou proměnných. Data musí být zachycena minimálně ordinálním měřením. Podstatou tohoto testu je srovnání pořadí hodnot u každé kategorie. Využití tohoto koeficientu ilustrujeme na příkladu.

Výzkumník hodnotil vizuálie v učebnicích vlastivědy pomocí metody škálování (ordinální stupnice) na základě předem stanovených kritérií. Pro srovnatelnost vybral tři typy vizuálií: mapu, fotografii a schéma. Nyní chce zjistit, zda existuje vztah mezi kvalitou těchto tří vizuálií mezi nakladateli. Výsledky škálování podle nakladatelství jsou uvedeny v tab. 24.

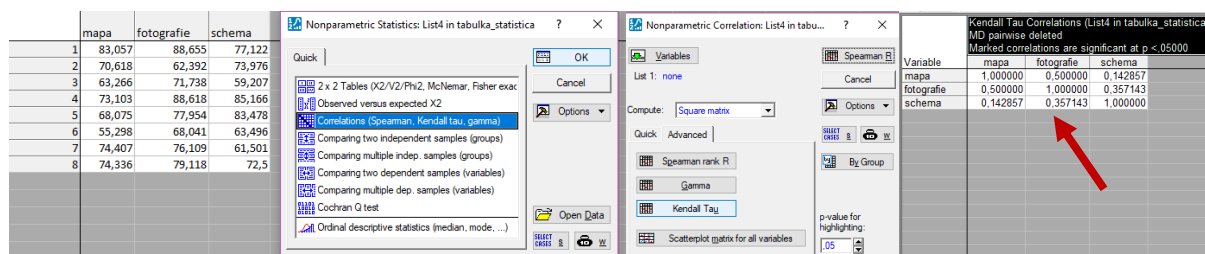
K dispozici máme celkem tři proměnné: hodnocení mapy, hodnocení fotografie, hodnocení schématu; dále jsou výsledky zachyceny ordinálním měřením. Lze tedy aplikovat Kendallův korelační koeficient na zjištění závislosti kvality těchto tří vizuálií.

Tab. 24 Výsledky škálování map, fotografií a schémat v učebnicích vlastivědy mezi různými nakladatelstvími (včetně uvedení pořadí)

Nakladatelství	SPN	Nová škola	Septima	Nová škola - Duha	Taktik	Dialog	ČGS	Alter
Hodnocení mapy	83,057	70,618	63,266	73,103	68,075	55,298	74,407	74,336
Pořadí	1	5	7	4	6	8	2	3
Hodnocení fotografie	88,655	62,392	71,738	88,618	77,954	68,041	76,109	79,118
Pořadí	1	8	6	2	4	7	5	3
Hodnocení schématu	77,122	73,976	59,207	85,166	83,478	63,496	61,501	72,500
Pořadí	3	4	8	1	2	6	7	5
Součet pořadí (X)	5	17	21	7	12	21	14	11

Zdroj: autoři

Výpočet provádíme v programu Statistica. Po vynesení hodnot do tabulky zvolíme *statistics* → *nonparametric statistics* → *correlations* (viz obr. 45) → *advanced* → volbu proměnných (*hodnocení mapy, hodnocení fotografie, hodnocení schématu*) → *Kendall tau* (viz obr. 45). Z obr. 45 je patrné (hodnoty nejsou zvýrazněny červeně), že na hladině významnosti $\alpha = 0,05$ potvrzujeme nulovou hypotézu o neexistenci závislosti mezi mírou kvality map, fotografií a schémat mezi nakladatelstvími v učebnicích vlastivědy.



Obr. 45 Výpočet Kendallova koeficientu shody v programu Statistica

Zdroj: autoři

Závislost mezi jevy zachycené nominálním měřením

I mezi jevy zachycené nominálním měřením lze zjišťovat sílu vztahu. V tomto případě vycházíme z výsledků čtyřpolní tabulky a výsledku testového kritéria chí-kvadrát. K výpočtu stupně závislosti v čtyřpolní tabulce vypočítáme tzv. r_ϕ -koeficient r_ϕ . Výpočet si ilustrujeme na příkladu.

Žáci základních škol a gymnázií měli za úkol identifikovat kartografický znak určený pro těžbu cínu. Jednalo se celkem o 189 respondentů a testové kritérium chí-kvadrát z čtyřpolní tabulky bylo vypočteno $X^2 = 78,96$. Nyní dosadíme do vzorce:

$$r_\phi = \sqrt{\frac{X^2}{n}} = \sqrt{\frac{78,96}{189}} = 0,646.$$

Míru asociace (vztahu) úspěšnosti identifikace kartografického znaku a stupně školy můžeme hodnotit jako střední až vysoký (viz tab. 20). Pokud nám ve výsledku vyjde záporné znaménko, nemá pro interpretaci výsledku význam.

Testové kritérium chí kvadrát nám udává rozdíl (souvislost) mezi jevy zachycenými nominálním měřením. Výsledek nám však nevypovídá o stupni závislosti mezi dvěma jevy. Ten lze určit např. pomocí koeficientu kontingence či Čuporova koeficientu K . Jeho výhodou oproti jiným testům (např. koeficientem kontingence) je zohlednění počtu proměnných díky zařazení informace o počtu řádků a sloupců kontingenční tabulky do výpočtu. Čuporův koeficient K se vypočte dle vztahu:

$$K = \sqrt{\frac{\sqrt{X^2}}{\sqrt{n * (r - 1) * (s - 1)}}}$$

kde X^2 je vypočítaná hodnota testového kritéria chí-kvadrát, n je celková četnost v kontingenční tabulce a r je počet řádků v tabulce a s je počet sloupců v tabulce. Výpočet ilustrujeme na příkladu.

Žáci základní a střední školy (celkem 185 žáků) se měli v dotazníku vyjádřit k vhodnosti kartografického znaku určeného pro těžbu fosfátů ve školních atlasech. Dotazníková položka se skládala z pěti vybraných kartografických znaků. Pomocí chí-kvadrát testu jsme zjišťovali, zda existuje souvislost mezi stupněm školy a volbou nejvhodnějšího znaku pro těžbu fosfátů. Byla vytvořena tabulka o dvou řádcích (stupeň školy) a pěti sloupcích (pět možností výběru znaku v dotazníkové položce) – viz tab. 25.

Tab. 25 Pozorované četnosti počtu respondentů s volbou určitého kartografického znaku

	znak „A“	znak „B“	znak „C“	znak „D“	znak „E“	součet
Žáci ZŠ	74	8	13	5	12	112
Studenti gymnázií	21	26	9	2	15	73
součet	95	34	22	7	27	185

Testové kritérium chí-kvadrát bylo vypočteno $X^2 = 34,77$. Po porovnání s tabulkovou hodnotou (9,45 – viz např. Chráska, 2016, s. 234) jsme zjistili, že stupeň školy má vliv na volbu nejvhodnějšího znaku určeného pro těžbu fosfátů. Nyní chceme zjistit, jak těsný je vztah mezi stupněm školy a volbou kartografického znaku. Proto hodnoty dosadíme do vzorce:

$$K = \sqrt{\frac{\sqrt{34,77}}{\sqrt{185 \cdot (2-1) \cdot (5-1)}}} = \sqrt{\frac{5,9}{27,2}} = 0,47.$$

Zjistili jsme, že těsnost vztahu mezi stupněm školy a volbou kartografického znaku je střední (viz tab. 20)

Pro zájemce je však nutné upozornit, že koeficientů zabývajících se zjišťováním těsnosti vztahu mezi nominálními proměnnými je celá řada; nevyznačující se však tak velkou přesností či vypovídací schopností, jako koeficienty počítající s metrickými či ordinálními daty.

Shrnutí korelační analýzy

Analýza závislostí je ve výzkumu cennou informací, která nám dává přehled o vztahu (asociaci) dvou a více proměnných. Vždy je ale nutné znát pravidla jejího užití, postup a možnosti užití jednotlivých korelačních koeficientů. Jak bylo zmíněno v textu, korelaci lze vypočítat prakticky ze všech dat, ale pouze logicky promyšlené užití jednotlivých koeficientů nám dává dobrý podklad pro interpretaci dat. Vždy je tedy nutné, aby si výzkumník rozmyslel, jaká data má k dispozici a co chce z dat „získat“. Je nutné brát v ohledu typy dat (nominální, ordinální, metrické), počet proměnných (dva, tři a více apod.), rozdělení dat (normální, jiné typy rozdělení) a na základě těchto kritérií vybírat vhodný statistický test. Navržený postup výběru vhodného statistického testu ukazuje obr. 39. V neposlední řadě je nutné upozornit na správnou interpretaci statistických dat (např. zobecnitelnost výsledků).

6.3. Regresní analýza

Regresní analýza umožňuje popsat tvar vztahu mezi proměnnými. Na rozdíl od korelační analýzy umožňuje předjímat trendy a predikovat hodnotu jedné proměnné v závislosti na druhé proměnné. Například na základě uplavané a uběhnuté vzdálenosti u probandů nám

umožňuje zjistit, kolik metrů uběhnou jedinci s určitou uplavanou vzdáleností. Požadavkem je, aby sledované veličiny byly spojité. Je vhodné ji využít především v těch případech, kde chceme použít prognózu a stanovit trend. Znovu se využívá grafického zobrazení dvojrozměrných dat, kde je využito proložení bodových znaků tzv. regresní přímkou.

V regresní analýze pracujeme se dvěma skupinami proměnných. Jednu z proměnných nazýváme cílová (též závislá) proměnná a druhou nebo několik dalších ovlivňujícími proměnnými, které nazýváme též nezávisle proměnné. Nezávisle proměnná má vliv (ovlivňuje) na cílovou proměnnou. V případě, že je mezi proměnnými určitý vztah (závislost), znamená, že pokud se změní hodnota nezávisle proměnné (např. uplavaná vzdálenost jedince), změní se i hodnota závisle proměnné (např. uběhlá vzdálenost). Například pokud budeme analyzovat vztah hmotnosti a výšky jedince, tak cílovou proměnnou může být hmotnost a nezávisle proměnnou výška jedince. Pokud se zvýší výška jedince, pravděpodobně se zvýší i jeho hmotnost. Regresní analýza nám umožňuje zjistit, o kolik se zvýší hmotnost jedince, pokud se zvýší výška jedince např. o 1 cm. Samozřejmě, že při predikci vždy pracujeme s chybou (viz dále).

Jak je uvedeno v prvním odstavci, regresní analýzou zjišťujeme tvar vztahu mezi proměnnými. Ten je popsán tzv. regresní rovnicí, kterou lze vygenerovat pomocí různých počítačových programů (MS Excel, Statistica – viz dále). Tato funkce obsahuje několik neznámých parametrů; po dosazení jedné proměnné do rovnice nám tato rovnice může s určitou pravděpodobností říci, jakých hodnot bude nabývat druhá proměnná.

Regresní analýzu používáme v těchto případech (Hendl, 2012):

- odhad neznámých parametrů regresní funkce,
- testování hypotéz o těchto parametrech,
- ověřování předpokladů regresního modelu.

Podle počtu proměnných rozlišujeme modely jednoduché regrese a vícenásobné regrese. Tato kapitola je zaměřena na jednoduchou lineární regresi; další typy regrese (exponenciální, logaritmická) je obsahem odborných publikací (např. Hendl, 2012). V další kapitole bude pojednáno o vícenásobné regresi.

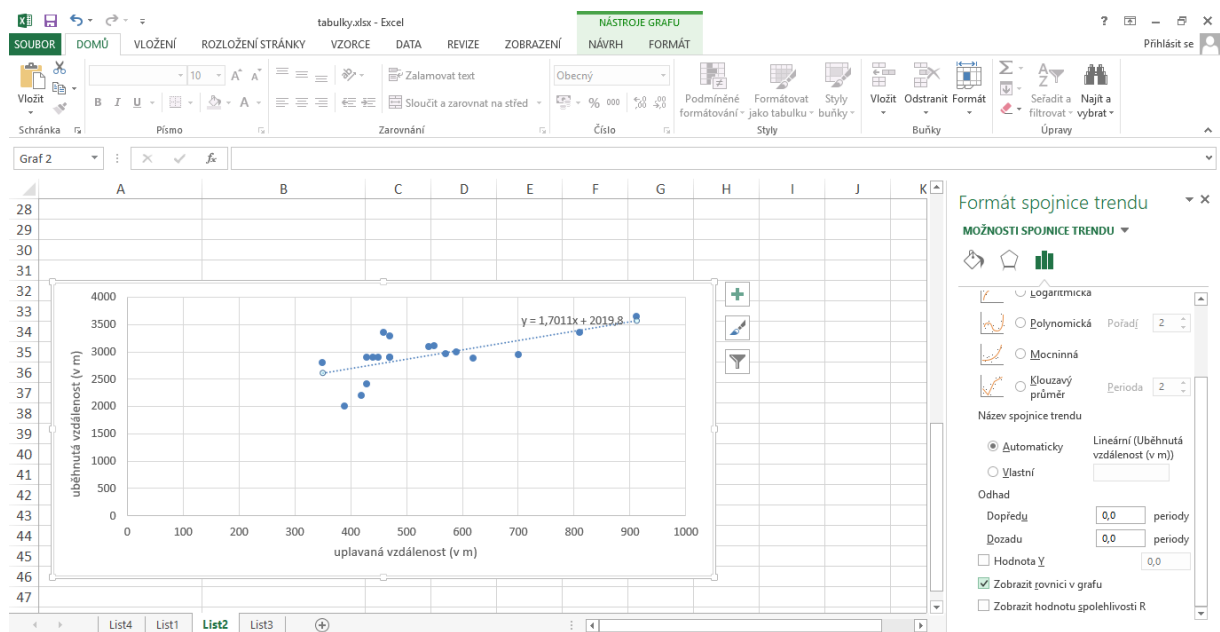
Postup regresní analýzy v MS Excel

Pokud používáme regresní analýzu, máme k dispozici podobný charakter dat, jako při korelační analýze, tzn., že vždy máme dvojice hodnot. Například od určitého počtu probandů máme k dispozici údaje o dvou proměnných (např. od každého jedince o výšce a hmotnosti).

Pro ilustraci budeme vycházet z dat z tab. 18 o uplavané a uběhnuté vzdálenosti jedinců v rámci Cooperova testu. V první fázi prokládáme bodový dvojrozměrný graf přímkou. Po-

stup programu MS Excel byl popsán v příslušné kapitole (viz zobrazování dvojrozměrných dat). Tato přímka má nejmenší součet druhých mocnin z rozdílu předpovědi (tj. hodnoty přímky) a skutečně naměřené hodnoty (tj. hodnoty uběhnuté vzdálenosti).

Jak bylo zmíněno v úvodu kapitoly, v regresní analýze nám jde o definici funkce, která nám umožňuje predikovat („předpovědět“) hodnotu uběhnuté vzdálenosti pomocí hodnot uplavané vzdálenosti. Tuto rovnici lze vygenerovat v MS Excel tak, že kliknutím na lineární spojnicí trendu pravým tlačítkem myši, zvolíme *formát spojnice trendu* a v nabídce zaškrtneme volbu *zobrazit rovnici v grafu* (viz obr. 46).



Obr. 46 Vložení regresní funkce v MS Excel

Zdroj: autoři

Nyní se nám v grafu zobrazila regresní funkce:

$$y = 1,7011x + 2\,019,8$$

Rovnice nám říká, že pokud budeme chtít predikovat uběhnutou vzdálenost u libovolného probanda na základě jeho uplavané vzdálenosti, musíme hodnotu naměřené uplavané vzdálenosti dosadit do rovnice za proměnnou x . Například změříme, že student uplave za 12 minut vzdálenost 650 metrů; na základě této hodnoty chceme zjistit, kolik metrů pravděpodobně uběhne v rámci Cooperova testu. Dosadíme tedy do rovnice:

$$\text{Uběhnutá vzdálenost v metrech} = 1,7011 * 650 + 2\,019,8 = 3\,126.$$

Zjistili jsme, že jedinec, který uplaval 650 metrů za 12 minut, pravděpodobně uběhne za stejný čas 3 126 metrů. Samozřejmě, že výsledek je zatížen určitou chybou. Jedinec při reálném měření uběhl za 12 minut 3 100 metrů. Při predikci jsme se dopustili chyby 26 metrů ($3126 - 3\,100 = 26$). Této chybě se říká chyba predikce (též reziduální hodnota).

Díky určité chybě predikce, vždy počítáme hodnotu v určitém intervalu spolehlivost, který vypočteme dle vztahu:

$$y' \pm u_1 - \frac{\alpha}{2} * \sqrt{\frac{s_r}{n-2}},$$

kde $u_1 - \frac{\alpha}{2}$ je při spolehlivosti predikce 95 % hodnota 1,96 a hodnotu s_r vypočteme z tab. 26, n je počet probandů.

Tab. 26 Postup výpočtu hodnoty s_r potřebné k určení intervalu spolehlivosti

Uplavaná vzdálenost (v m) x_i	Uběhnutá vzdálenost (v m) y_i	Predikční hodnota $y' = 1,7011 x_i + 2 019,8$	Odchylka $\Delta = y_i - y_i'$	Druhá mocnina Δ_i^2
350	2800	2615,185	184,815	34156,58
390	2000	2683,229	-683,229	466801,9
420	2200	2734,262	-534,262	285435,9
430	2400	2751,273	-351,273	123392,7
430	2900	2751,273	148,727	22119,72
440	2900	2768,284	131,716	17349,1
450	2900	2785,295	114,705	13157,24
460	3350	2802,306	547,694	299968,7
470	3280	2819,317	460,683	212228,8
470	2900	2819,317	80,683	6509,746
470	2900	2819,317	80,683	6509,746
540	3090	2938,394	151,606	22984,38
550	3110	2955,405	154,595	23899,61
570	2960	2989,427	-29,427	865,9483
590	3000	3023,449	-23,449	549,8556
620	2880	3074,482	-194,482	37823,25
700	2940	3210,57	-270,57	73208,12
810	3350	3397,691	-47,691	2274,431
912	3650	3571,2032	78,7968	6208,936

Σ 1655445

Zdroj: autoři

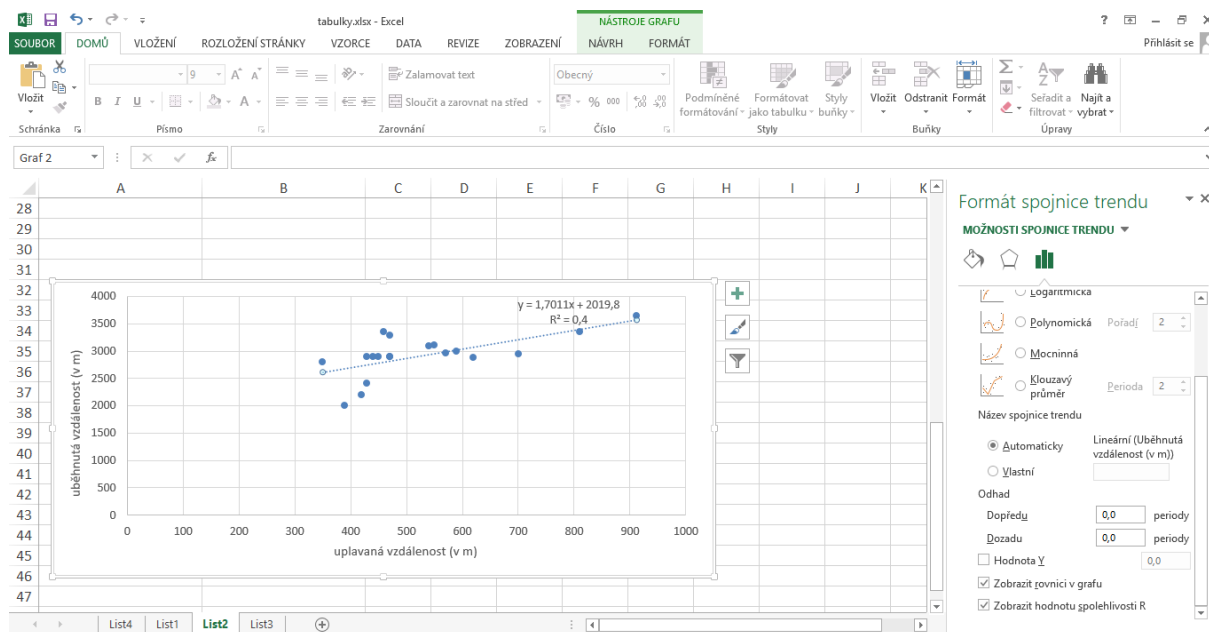
Součet v posledním sloupci je hodnota s_r . Dosazením do vzorce dostaneme:

$$y' \pm 1,96 * \sqrt{\frac{1 655 445}{17}} = 612$$

$$y' \pm 612 \text{ metrů}$$

Výsledek lze interpretovat tak, že predikci uběhnuté vzdálenosti z uplavané vzdálenosti v Cooperově testu lze určit při spolehlivosti 95 % s přesností ± 612 metrů.

Pomocí programu MS Excel lze určit i koeficient determinace R^2 . Tato hodnota nám říká, kolik procent variability v závisle proměnné se nám podařilo vysvětlit (z kolika procent je ovlivněna cílová proměnná nezávisle proměnnou). V MS Excel v možnostech lineární spojnice trendu (viz předchozí příklad – zobrazení regresní funkce) zaškrtneme *zobrazit hodnotu spolehlivosti R* (viz obr. 47).



Obr. 47 Vložení koeficientu determinace (R^2) v MS Excel

Zdroj: autoři

Koeficient determinace byl vypočítán $R^2 = 0,4$. Jedná se o druhou mocninu koeficientu korelace, která byla v předcházející kapitole vypočítána $r = 0,632$. Výsledek lze interpretovat tak, že uplavaná vzdálenost je schopna vysvětlit 40 % variability uběhnuté vzdálenosti. Zbytku (tj. 60 %) variability veličiny uběhnuté vzdálenosti je nutné hledat jinde (např. délka končetin, hmotnost jedince atd.). Výsledek v Cooperově testu v plavání ovlivňuje ze 40 % výsledek v běhu.

Pomocí doplňku analýzy dat v MS Excel lze zjistit další parametry vztahující se k regresní analýze. Tento doplněk budeme ilustrovat na stejném příkladu (z tab. 18). Postup je následující: *data* → *analýza dat* → *regrese* → *výběr vstupní oblasti X (tj. uplavaná vzdálenost) a vstupní oblasti Y (tj. uběhnutá vzdálenost)* → dále zvolíme, zda jsme v tabulce označili i *popisky* → zvolíme *hladinu významnosti* → *OK*. Výsledky regresní analýzy se nám zobrazí na novém listě (viz obr. 48). Nyní vysvětlíme údaje, které jsou pro interpretaci základní regresní analýzy důležité. Žluté pole nám značí výsledek korelační analýzy. Modré pole značí hodnotu koeficientu determinace. Červené pole ukazuje počet pozorování probandů. Zelená pole jsou hodnoty regresních koeficientů, které dosazujeme do regresní funkce (viz regresní

funkce); tyto hodnoty jsme získali i bez doplňku analýzy dat (viz výše).⁷ Hodnota v oranžovém poli značí hodnoty testového kritéria p , pomocí kterého testujeme nulovou hypotézu. Ta bude znít: neexistuje statisticky významný vliv (závislost) uběhnuté vzdálenosti na uplavané vzdálenosti během Cooperova testu. Hodnota p je menší než 0,05, proto nulovou hypotézu nepotvrzujeme ($p = 0,0036$) na hladině významnosti $\alpha = 0,05$. Uplavaná vzdálenost má vliv na uběhnutou vzdálenost v Cooperově testu.

Regresní statistika								
Násobné R	0,63241739							
Hodnota spolehlivosti R	0,39995175							
Nastavená hodnota spolehlivosti R	0,3646548							
Chyba stř. hodnoty	312,056244							
Pozorování	19							
ANOVA								
	Rozdíl	SS	MS	F	Významnost F			
Regrese	1	1103407,945	1103408	11,3310552	0,003666595			
Rezidua	17	1655444,687	97379,1					
Celkem	18	2758852,632						
Koefficienty								
	Koefficienty	Chyba stř. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	2019,83503	277,2859031	7,2843	1,2761E-06	1434,812917	2604,85715	1434,81292	2604,85715
Uplavaná vzdálenost (v m) (x)	1,70106576	0,505342618	3,36616	0,0036666	0,634886032	2,76724549	0,63488603	2,76724549

Obr. 48 Výsledek regresní analýzy v MS Excel s vyznačením důležitých údajů

Zdroj: autoři

Shrnutí regresní analýzy

Jednoduchá lineární regresní analýza umožňuje popsat tvar vztahu mezi dvěma proměnnými. Korelační analýza naopak popisuje sílu vztahu dvou (či více proměnných). Metody regresní analýzy nám umožňují predikovat („předpovědět“) hodnotu závisle proměnné (např. uběhnutá vzdálenost) z hodnoty nezávisle proměnné (např. uplavaná vzdálenost). Výslednou hodnotu závisle proměnné zjistíme na základě dosazení hodnoty závisle proměnné do tzv. regresní funkce (rovnice), kterou lze jednoduše vygenerovat v MS Excel. Mezi další základní údaje regresní analýzy, kterou lze pomocí MS Excel zjistit patří koeficient determinace a hodnota p , pomocí které je možné testovat nulovou hypotézu.

6.4. Vícenásobná regresní analýza

V předchozí kapitole bylo pojednáno o analýze vztahu mezi dvěma proměnnými, z nichž jedna byla závisle proměnná a druhá nezávisle proměnná. Pro hlubší analýzu dat a zjištění příčin změn závisle proměnné je v některých případech nutné analyzovat komplexnější data, tzn. vliv více nezávisle proměnných na jednu závisle proměnnou. Chceme například zjistit, které faktory ovlivňují uplavanou vzdálenost během 12 minut u vysokoškolských studentů

⁷ Hodnota 1,701 nám říká, že jedinec, který uplave o 1 metr více, zároveň uběhne v průměru o 1,701 metru více. V praxi se tyto hodnoty mohou využít v tréninkovém procesu například takto: pokud se jedinec zlepšil v Cooperově testu v plavání o 30 metrů, zároveň se zlepšil v Cooperově testu v běhu o 51 metrů ($30 \cdot 1,701 = 51$).

studující studijní obor tělesná výchova a sport. Dalším příkladem je zjišťování závislosti (vztahu, vlivu) výsledků žáka 5. třídy v testu přírodovědné gramotnosti na výsledcích v jiných testech (např. testu environmentální gramotnosti, testu logického myšlení, testu vizuální gramotnosti atd.).

I v rámci této metody se pracuje s již popsanými prostředky jednoduché analýzy – korelace, bodové dvojrozměrné grafy, regresní koeficienty atd. Základní příklad vícenásobné regresní analýzy bude ilustrován na příkladu z tab. 18. V předchozích kapitolách jsme zjišťovali vztah a sílu (korelační analýza) i tvar (regresní analýza) mezi uplavanou a uběhnutou vzdáleností v Cooperově testu. Nyní budeme zjišťovat vliv základních tělesných ukazatelů, tedy výšky a hmotnosti jedince společně s uplavanou vzdáleností, na výsledek Cooperova testu v běhu. Uběhlou vzdálenost tedy označujeme cílovou proměnnou, naopak hmotnost, výšku a uplavanou vzdálenost jako nezávisle proměnné. Postup vícenásobné regresní analýzy bude popsán v MS Excel (doplňek analýza dat).

Vícenásobnou regresní analýzu lze provádět různými způsoby (metodami). V další části textu bude první popsána metoda *Enter*, kdy bereme v úvahu všechny proměnné najednou a v druhé části kapitoly bude popsána metoda *Stepwise*, kdy do regresního modelu zahrnujeme proměnné postupně a sledujeme význam („důležitost“) jednotlivých modelů na výslednou regresní funkci (rovnici).

Základní postup vícenásobné regresní analýzy v MS Excel

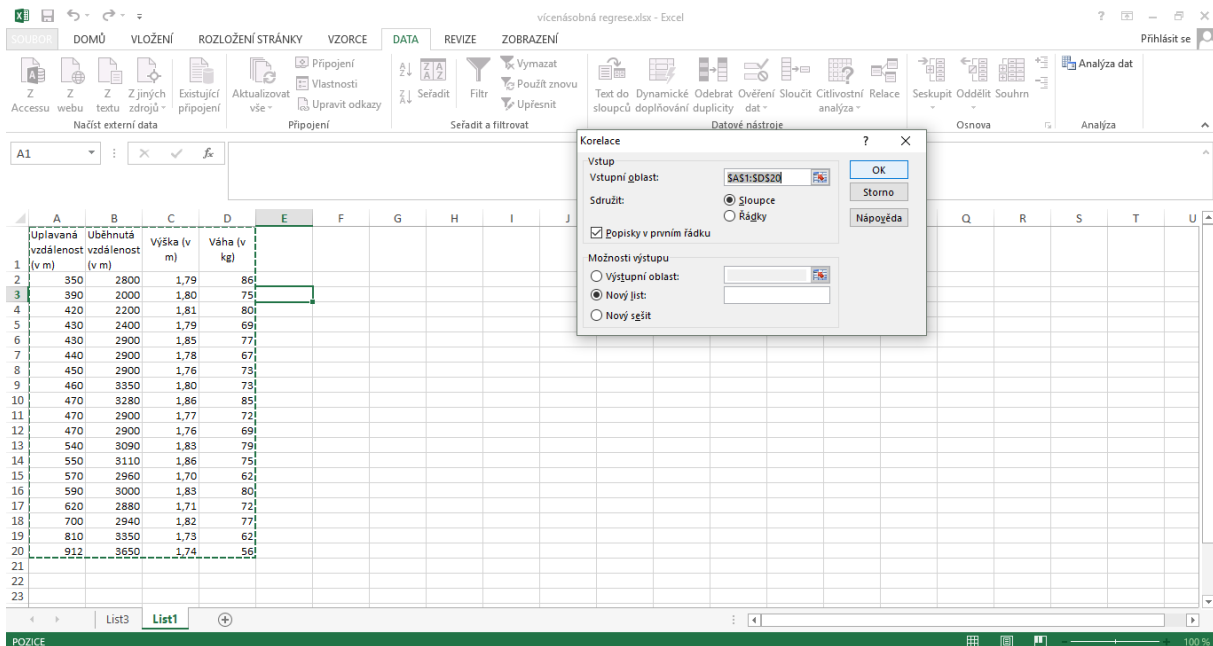
Budeme vycházet z příkladu uvedeném v tab. 18. Data z tabulky jsou rozšířena o údaje o hmotnosti a výšce jedince (viz tab. 27). Znovu platí, že je nutné dbát na správné spárování dat – každému probandovi odpovídá jeden řádek v tabulce. Například z tab. 27 lze zjistit, že proband 9 uplaval 470 metrů, uběhl 3 280 metrů, měří 1,86 metru a váží 85 kilogramů. Počet proměnných (značíme k), které budeme v tomto příkladu analyzovat je 4 (uplavaná vzdálenost během 12 minut, uběhnutá vzdálenost během 12 minut, výška, hmotnost: $k = 4$).

Tab. 27 Výsledky Cooperova testu v běhu a plavání u 19 probandů včetně jejich hmotnosti

Proband	Uplavaná vzdálenost (v m)	Uběhnutá vzdálenost (v m)	Výška (v m)	Hmotnost (v kg)
Proband 1	350	2800	1,79	86
Proband 2	390	2000	1,80	75
Proband 3	420	2200	1,81	80
Proband 4	430	2400	1,79	69
Proband 5	430	2900	1,85	77
Proband 6	440	2900	1,78	67
Proband 7	450	2900	1,76	73
Proband 8	460	3350	1,80	73
Proband 9	470	3280	1,86	85
Proband 10	470	2900	1,77	72
Proband 11	470	2900	1,76	69
Proband 12	540	3090	1,83	79
Proband 13	550	3110	1,86	75
Proband 14	570	2960	1,70	62
Proband 15	590	3000	1,83	80
Proband 16	620	2880	1,71	72
Proband 17	700	2940	1,82	77
Proband 18	810	3350	1,73	62
Proband 19	912	3650	1,74	56

Zdroj: autoři

Prvním krokem je provedení korelační analýzy mezi sledovanými proměnnými. Tu lze v MS Excel provést v záložce *data* → *analýza dat* → *korelace* (pokud vybíráme i první řádek tabulky je nutné zaškrtnout položku *Popisky* v prvním řádku) → *OK* (viz obr. 49).



Obr. 49 Korelační analýza v MS Excel

Zdroj: autoři

Výsledek je na novém listě vytvořená korelační matice. Ta nám udává korelační koeficienty mezi jednotlivými proměnnými (viz obr. 50). Některé údaje chybí, protože matice (tabulka) je symetrická; například hodnota korelace mezi výškou a hmotností jedince je stejná, jako mezi hmotností a výškou jedince (tj. $r = 0,719$) – viz červeně zbarvené pole. Hodnoty $r = 1$ na diagonále (modře zbarvené pole) znamenají, že každá z veličin je mezi sebou korelovaná (hodnotě o výšce jedince odpovídá ta samá hodnota výšky jedince).

	A	B	C	D	E
1		Uplavaná vzdálenost (v m)	Uběhnutá vzdálenost (v m)	Výška (v m)	Váha (v kg)
2	Uplavaná vzdálenost (v m)	1			
3	Uběhnutá vzdálenost (v m)	0,632417388	1		
4	Výška (v m)	-0,364768597	-0,095969463	1	
5	Váha (v kg)	-0,597752542	-0,317681104	0,71924831	1

Obr. 50 Výsledek korelační analýzy v MS Excel

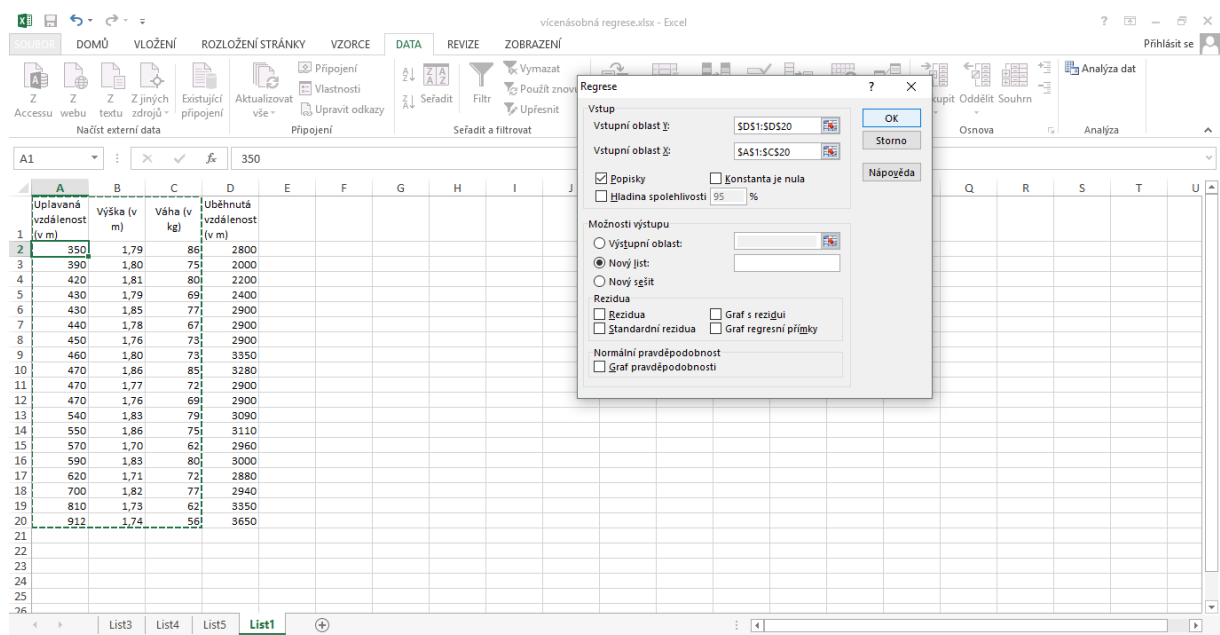
Zdroj: autoři

Korelační koeficienty nabývají hodnot $\langle -1; 1 \rangle$; interpretace těchto hodnot viz příslušná kapitola. V dalších krocích je nutné, aby byla určitým způsobem řešena silná korelace mezi faktory. Právě vysoká hodnota korelačního koeficientu může zkreslit výsledek regresní analýzy. Jednou z možností, jak řešit vysokou korelaci, je vypustit jednu z proměnných.⁸ Přesto lze

⁸ Sám autor analýzy vždy rozhoduje o tom, které z proměnných do regresní analýzy „pustí“ a které nikoliv. Na první pohled se může zdát, že čím více proměnných zahrne do vícenásobné regresní analýzy, tím bude model přesnější. Bohužel tomu tak není. Právě vhodným výběrem proměnných ovlivní kvalitu výsledné regresní funkce, která slouží k predikci (předpovědi) cílové, tedy závislé proměnné na základě nezávisle proměnných.

pro naše účely hodnotu korelačního koeficientu mezi hmotností a výškou jedince prozatím v regresním modelu ponechat. Pokud by však hodnota korelačního koeficientu byla vyšší než 0,8, již by bylo vhodné tak vysokou korelaci řešit právě odstraněním jednoho z faktorů z vícenásobné regresní analýzy.

Nyní přistupme k samotné vícenásobné regresní analýze metodou **Enter**. Cílem bude zjistit vliv nezávisle proměnných uplavané vzdálenosti, výšky a hmotnosti jedince na závisle proměnnou tedy uběhnutou vzdálenost. Vícenásobnou regresní analýzu provedeme: *data* → *analýza dat* → *regrese* → *volba cílové proměnné Y (uběhnutá vzdálenost) a volba nezávisle proměnných X (uplavaná vzdálenost, hmotnost a výška)* → OK. Vždy je nutné, aby sloupce nezávisle proměnných byly vedle sebe, tzn., aby označovaná oblast byla spojitá (viz obr. 51).



Obr. 51 Vícenásobná regresní analýza v MS Excel (metoda Enter)

Zdroj: autoři

Výsledek se nám vygeneruje na novém listě (viz obr. 52). Červeně označené pole nám udává hodnotu vícenásobného korelačního koeficientu. Ten vypovídá o síle vztahu všech faktorů dohromady. Modré pole označuje hodnotu indexu determinace $R^2 = 0,42$; hodnota říká, že uběhnutá vzdálenost je modelem vystižena ze 42 % (blíže viz příslušná kapitola – regresní analýza). Fialové pole označuje p-významnost celého modelu; hodnota je menší než 0,05, proto ji lze model označit jako významný. Sloupec tabulky s názvem Hodnota P nám dává informaci o významnosti jednotlivých koeficientů a jejich vlivu na cílovou proměnnou (tedy uběhnutou vzdálenost). Žlutě označené koeficienty jsou vyšší než 0,05, proto jsou na hladině významnosti $\alpha = 0,05$ nevýznamné.

Regresní statistika								
Násobné R	0,64942819							
Hodnota spolehlivosti R	0,42175698							
Nastavená hodnota spolehlivosti R	0,30610837							
Chyba stř. hodnoty	326,117288							
Pozorování	19							
ANOVA								
	Rozdíl	SS	MS	F	Významnost F			
Regrese	3	1163565,349	387855,116	3,64688342	0,037225454			
Rezidua	15	1595287,282	106352,485					
Celkem	18	2758852,632						
ANOVA - detailní								
	Koeficienty	Chyba stř. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	-553,221486	3511,561393	-0,15754288	0,87691816	-8037,937422	6931,49445	-8037,93742	6931,49
Uplavaná vzdálenost (v m)	1,80182546	0,663312793	2,7164039	0,01592592	0,388007707	3,21564321	0,38800771	3,21564
Výška (v m)	1510,70438	2334,833116	0,64702885	0,52739146	-3465,874601	6487,28337	-3465,8746	6487,28
Váha (v kg)	-2,50224606	16,64400721	-0,15033916	0,88250001	-37,97810766	32,9736156	-37,9781077	32,9736

Obr. 52 Výsledek vícenásobné regresní analýzy v MS Excel (metoda Enter)

Zdroj: autoři

Stejně jako v jednoduché lineární regresní analýze, stejně i u tohoto typu regresní analýzy zjistíme tvar regresní funkce, tedy rovnice, která nám může pomoci predikovat cílovou proměnnou (uběhnuté vzdálenosti) na základě nezávisle proměnných (výšky, hmotnosti a uplavané vzdálenosti). Hodnoty koeficientů zjistíme z prvního sloupce třetí tabulky (vyznačeno zeleně). Regresní funkce má tedy tvar:

$$\text{Uběhnutá vzdálenost v metrech} = -553,22 + 1,80 * x_1 + 1510,70 * x_2 - 2,50 * x_3,$$

kde x_1 je uplavaná vzdálenost v metrech, x_2 je výška v metrech, x_3 je hmotnost v kilogramech.

Pokud budeme chtít zjistit přibližnou uběhnutou vzdálenost jedince, který uplaval 530 metrů, váží 65 kilogramů a měří 1,75 metru, dosadíme do rovnice:

$$\begin{aligned} \text{Uběhnutá vzdálenost v metrech} &= -553,22 + 1,80 * 530 + 1510,7 * 1,75 - 2,5 * 65 = \\ &= 2882 \text{ metru.} \end{aligned}$$

Jedinec pravděpodobně uběhne 2 888 metrů.

Doposavad byla popisována jen jedna z metod vícenásobné regresní analýzy. Jednalo se o tzv. metodu Enter, kdy do analýzy jsme zahrnuli všechny sledované faktory. Pokud bychom chtěli regresní rovnici zpřesnit, bylo by nutné využít další metodu vícenásobné regresní analýzy zvanou **Stepwise**. V rámci této metody jsou do regresního modelu vkládány postupně nezávisle proměnné. K ilustraci využijeme předchozí příklad. Již bylo zmíněno, že výška i hmotnost jedince byly v rámci regresní analýzy metodou Enter shledány jako nevýznamné faktory, které mají vliv na uběhnutou vzdálenost. Pro detailnější analýzu dat provedeme lineární regresní analýzu (viz předchozí kapitola) mezi všemi nezávisle proměnnými se závisle proměnnou postupně; tj. regresní analýzu výška-uběhnutá vzdálenost, dále hmotnost-uběhnutá vzdálenost a dále uplavaná vzdálenost-uběhnutá vzdálenost. Ze všech regresních analýz si poznamenejme (viz tab. 28) Hodnotu P z poslední tabulky (na obr. 53, kte-

rý ukazuje příklad regresní analýzy vztahu výšky jedince a uběhnuté vzdálenosti, vyznačena červeně). Hodnoty větší než 0,05 nám značí nevýznamný vliv faktoru nezávisle proměnné (např. výšky jedince) na závisle proměnnou (na uběhnutou vzdálenost). Naopak hodnoty menší než 0,05 nám značí významný vliv nezávisle proměnné na závisle proměnnou. V našem případě se jedná jen o faktor uplavané vzdálenosti ($p = 0,004$). Jelikož je statisticky významný pouze tento faktor (nezávisle proměnná), je ze sledovaných nezávisle proměnných nejlepším prediktorem uběhnuté vzdálenosti právě uplavaná vzdálenost. Pokud by i jiná nezávisle proměnná měla hodnotu P nižší než 0,05, bylo by ji možné zařadit do vícenásobné regresní analýzy (viz dále).

Regresní statistika								
Násobné R	0,09596946							
Hodnota spolehlivosti R	0,00921014							
Nastavená hodnota spolehlivosti R	-0,04907162							
Chyba stří. hodnoty	400,987254							
Pozorování	19							
ANOVA								
	Rozdíl	SS	MS	F	Významnost F			
Regrese	1	25409,41298	25409,413	0,1580278	0,69592588			
Rezidua	17	2733443,219	160790,778					
Celkem	18	2758852,632						
	Koeficienty	Chyba stří. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	4330,26586	3544,818666	1,22157613	0,23853818	-3148,647781	11809,1795	-3148,64778	11809,1795
Výška (v m)	-787,438993	1980,843517	-0,39752711	0,69592588	-4966,653503	3391,77552	-4966,6535	3391,77552

Obr. 53 Vícenásobná regresní analýza v MS Excel (metoda Stepwise)

Zdroj: autoři

Tab. 28 Výsledky závislosti uběhnuté vzdálenosti na nezávisle proměnných

Nezávisle proměnná	Hodnota P (uběhnutá vzdálenost- nezávisle proměnná)
Uplavaná vzdálenost	0,004
Výška	0,700
Hmotnost	0,185

Zdroj: autoři

Vliv hmotnosti a výšky jedince na uběhnutou vzdálenost se ukázal jako nevýznamný; jedinou nezávisle proměnnou, která byla označena vícenásobnou regresní analýzou metodou Stepwise jako významná, je uplavaná vzdálenost.

Nyní si ukážeme vhodnější příklad pro ilustraci aplikace metody Stepwise ve vícenásobné regresní analýze. Budeme vycházet z dat z tab. 27. Nyní naší cílovou proměnnou nebude uběhnutá vzdálenost, nýbrž hmotnost jedince. Nezávisle proměnné budou výška, uplavaná vzdálenost a uběhnutá vzdálenost v Cooperově testu. V prvním kroku, stejně jako v předešlém příkladě, bychom měli provést korelační analýzu všech proměnných pro vyloučení proměnných, které mezi sebou mají vysokou hodnotu korelačního koeficientu. Jelikož

jsme tato data již korelační analýze podrobili (viz obr. 49), již nemusíme tuto analýzu provádět a rovnou zahájíme vícenásobnou regresní analýzu. V prvním kroku provedeme vícenásobnou regresní analýzu metodou Enter, tedy vložíme do modelu všechny proměnné: *data* → *analýza dat* → *regrese* → *vložíme cílovou proměnnou Y (hmotnost) a nezávisle proměnné (uplavaná a uběhnutá vzdálenost, výška)* → OK. Výsledek vidíme na obr. 54. Většinu hodnot jsme interpretovali v předchozím příkladě. Nyní se zaměříme na sloupeček Hodnota *P* v třetí tabulce. Žlutě jsou na obrázku vyznačeny nevýznamné koeficienty na hladině významnosti 0,05. Protože hodnoty jsou vyšší než 0,05, model tedy není plně dobrý. Nyní provedeme postupnou lineární regresní analýzu mezi cílovou proměnnou a postupně všemi nezávisle proměnnými. Provádíme tedy celkem tři regresní analýzy: hmotnost-výška, hmotnost-uběhnutá vzdálenost, hmotnost-uplavaná vzdálenost. Tuto regresní analýzu provádíme stejně jako v předchozí kapitole: *data* → *analýza dat* → *regrese* → *cílová proměnná Y (tj. vždy hmotnost) a nezávisle proměnná X (postupně výška, uplavaná vzdálenost, uběhnutá vzdálenost)*. Vygenerují se nám tedy tři nové listy. Pro ilustraci uvádíme jen list regresní analýzy, který analyzuje vztah hmotnost a uplavané vzdálenosti (viz obr. 54 dole).

Regresní statistika								
Násobné R	0,80473739							
Hodnota spolehlivosti R	0,64760227							
Nastavená hodnota spolehlivosti R	0,57712272							
Chyba stří. hodnoty	5,05525827							
Pozorování	19							
ANOVA								
	Rozdíl	SS	MS	F	Významnost F			
Regrese	3	704,4549317	234,818311	9,18851362	0,001077726			
Rezidua	15	383,3345419	25,5556361					
Celkem	18	1087,789474						
	Koeficienty	Chyba stří. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	-84,6529934	49,90205388	-1,69638295	0,1104614	-191,0167035	21,7107167	-191,016703	21,7107167
Uplavaná vzdálenost (v m)	-0,01954832	0,01150031	-1,69980803	0,10980301	-0,044060649	0,00496401	-0,04406065	0,00496401
Výška (v m)	94,9595176	27,30098993	3,47824448	0,0033707	36,76883498	153,1502	36,768835	153,1502
Uběhnutá vzdálenost (v m)	-0,00060127	0,003999419	-0,15033916	0,88250001	-0,00912583	0,00792329	-0,00912583	0,00792329

Regresní statistika								
Násobné R	0,59775254							
Hodnota spolehlivosti R	0,3573081							
Nastavená hodnota spolehlivosti R	0,3195027							
Chyba stří. hodnoty	6,41282484							
Pozorování	19							
ANOVA								
	Rozdíl	SS	MS	F	Významnost F			
Regrese	1	388,6759915	388,675991	9,45124364	0,006872744			
Rezidua	17	699,1134822	41,1243225					
Celkem	18	1087,789474						
	Koeficienty	Chyba stří. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	90,0294972	5,698286652	15,7993977	1,3546E-11	78,0071633	102,051831	78,0071633	102,051831
Uplavaná vzdálenost (v m)	-0,03192618	0,010384903	-3,0742875	0,00687274	-0,053836406	-0,01001595	-0,05383641	-0,01001595

Obr. 54 Výsledek vícenásobné regresní analýzy metodou Stepwise

Zdroj: autoři

Do tabulky si postupně zapíšeme hodnoty *P* (viz tab. 29). Ze zapsaných hodnot je zřejmé, že jako významné nezávisle proměnné mající vliv na hmotnost se jeví výška jedince a uplavaná vzdálenost. Hodnota *P* je menší než 0,05, proto je faktor na hladině významnosti $\alpha = 0,05$ významný. Do „finální“ vícenásobné regresní analýzy zahrneme pouze ty nezávisle proměnné

né, které byly regresní analýzou označeny jako významné (hodnota P menší než 0,05), což jsou výška a uplavaná vzdálenost.

Tab. 29 Výsledky závislosti hmotnosti na nezávisle proměnných

Nezávisle proměnná	Hodnota P (hmotnost-nezávisle proměnná)
Uplavaná vzdálenost	0,01
Uběhnutá vzdálenost	0,19
Výška	<0,01

Zdroj: autoři

Výsledek vícenásobné regresní analýzy je ukázán na obr. 55. Ten bere v úvahu vliv výšky a uplavané vzdálenosti jedince na jeho hmotnost.

Regresní statistika								
Násobné R	0,80440741							
Hodnota spolehlivosti R	0,64707128							
Nastavená hodnota spolehlivosti R	0,60295519							
Chyba stř. hodnoty	4,89841905							
Pozorování	19							
ANOVA								
	Rozdíl	SS	MS	F	Významnost F			
Regrese	2	703,8773268	351,938663	14,6674667	0,000240711			
Rezidua	16	383,9121469	23,9945092					
Celkem	18	1087,789474						
	Koeficienty	Chyba stř. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	-84,4474114	48,3356862	-1,74710277	0,0997827	-186,9144887	18,0196659	-186,914489	18,0196659
Uplavaná vzdálenost (v m)	-0,02066279	0,008519488	-2,4253557	0,02749416	-0,038723297	-0,00260228	-0,0387233	-0,00260228
Výška (v m)	94,1928928	25,98843012	3,62441642	0,00227859	39,09988206	149,285903	39,0998821	149,285903

Obr. 55 Výsledek vícenásobné regresní analýzy (vliv výšky a uplavané vzdálenosti jedince na jeho hmotnost)

Zdroj: autoři

V této fázi nás zajímají zeleně vyznačená pole, což jsou koeficienty regresní funkce. Výsledná rovnice má tedy tvar:

$$\text{Hmotnost v kilogramech} = -84,45 - 0,02 * x_1 + 94,19 * x_2,$$

kde x_1 je hodnota uplavané vzdálenost v metrech a x_2 je hodnota výšky jedince v metrech.

Nyní výsledek interpretujeme. Index determinace nám říká, že výsledný model vystihuje hodnotu proměnné hmotnost z 64,7 %. Zbytek, tedy 35,3 % připadá na jiné faktory. Hodnota významnost F je menší než 0,05, proto lze tento model označit jako významný na hladině významnosti $\alpha = 0,05$. Koeficienty nám říkají, že pokud jedinec bude měřit o 1 metr více,

bude o 94,19 kilogramu těžší⁹ a také pokud jedinec uplave o 0,02 metru méně během Cooperova testu, bude o 1 kilogram těžší.¹⁰

Samozřejmě, že znovu lze predikovat hmotnost jedince (resp. studenta tělesné výchovy, protože v příkladu byl analyzován soubor studentů tělesné výchovy) z uplavané vzdálenosti a z jeho výšky. Pokud bude jedinec (student tělesné výchovy) měřit 180 cm (tj. 1,8 m) a uplave vzdálenost 520 metrů za 12 minut, bude pravděpodobně vážit 74,69 kg:

$$\text{Hmotnost v kilogramech} = -84,45 - 0,02 * 520 + 94,19 * 1,8 = 74,69 \text{ kg.}$$

Jak bylo zmíněno výše, výsledný model vždy závisí na tom, které proměnné do modelu zahrneme a které nikoliv. Predikce cílové proměnné je potom tímto faktorem ovlivněna. Pro ilustraci tohoto tvrzení jsou v tab. 30 ukázány hodnoty predikce hmotnosti po řazení různých proměnných. Bereme v úvahu, že predikujeme hmotnost jedince, který měří 1,8 metru (označení v), uplave 520 metrů (označení p) a uběhne 2 800 metrů (označení b) v Cooperově testu. Hodnoty jsou vypočítány dosazením hodnot výše uvedených nezávisle proměnných do regresní funkce; koeficienty regresní funkce byly vygenerovány programem MS Excel v rámci regresní analýzy, do které jsme vložili hmotnost jedince jako cílovou proměnnou a nezávisle proměnné uvedené v prvním sloupci tab. 30.

Tab. 30 Různé výsledky regresního modelu v závislosti na zařazení nezávisle proměnných do regresní funkce

Zařazené nezávisle proměnné do regresního modelu	Tvar regresní funkce	Výsledná hodnota hmotnosti jedince (v kg)
Výška (v)	$-136,53 + 117,18 * v$	74,40
Uplavaná vzdálenost (p)	$90,029 - 0,032 * p$	73,39
Výška (v) a uplavaná vzdálenost (p)	$-84,45 - 0,02 * p + 94,19 * v,$	74,69
Výška (v), uplavaná vzdálenost (p), uběhnutá vzdálenost (b) ¹¹	$-84,65 - 0,02 * p + 94,96 * v - 0,0006 * b$	74,20

Zdroj: autoři

⁹ V příkladu jsme počítali výšku v metrech. Po převodech jednotek lze říci, že pokud bude jedinec o 1 cm vyšší, bude o 0,94 kg těžší (1 metr má 100 centimetrů $\rightarrow 94,19 / 100 = 0,94$).

¹⁰ Jedná se pouze o model, který jsme počítali na souboru studentů tělesné výchovy, proto tento model predikuje hodnoty pouze u studentů tělesné výchovy. Jistě nelze očekávat, že jedinec z obecné populace s výškou 1,75 m, který uplave 200 metrů, bude vážit 76,38 kg.

¹¹ Jedná se o vícenásobnou regresní analýzu provedenou metodou Enter.

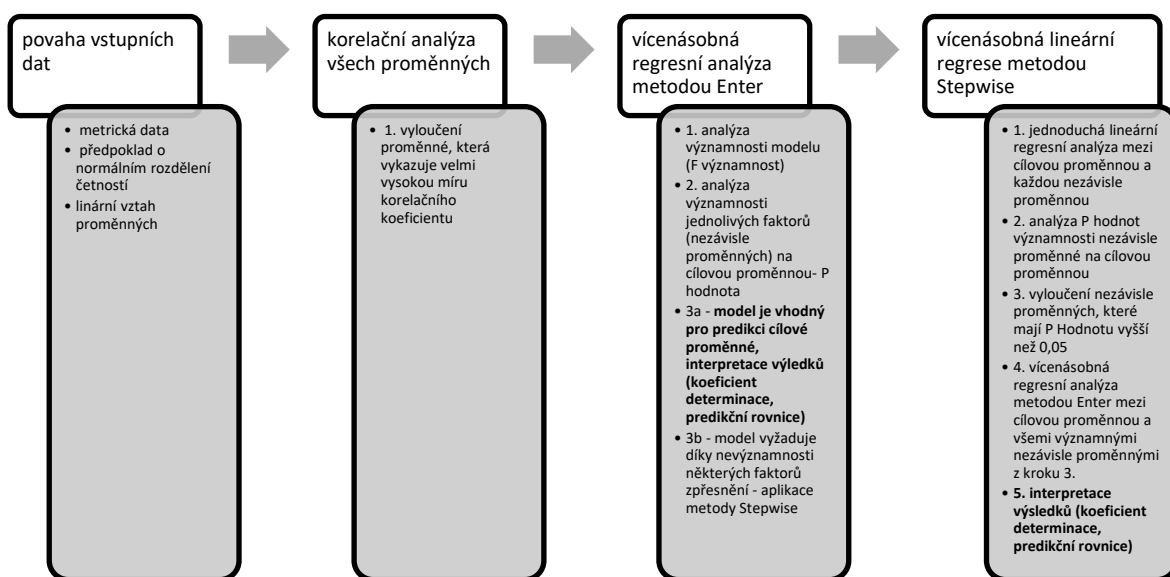
Z tabulky je zřejmé, že hodnoty predikce hmotnosti na třech nezávisle proměnných se nepatrně liší. Pro volbu nejvhodnějšího modelu metodou Stepwise je vždy nutné zohlednit významnost jednotlivých nezávisle proměnných na závisle (cílovou) proměnnou. Nemusí nutně znamenat, že čím více nezávisle proměnných je do modelu zahrnuto, tím je regresní model přesnější. Jde spíše o vhodnou volbu nezávisle proměnných, které budou do modelu zahrnuty.

Shrnutí vícenásobné regresní analýzy

Vícenásobná regresní analýza nám umožňuje zjistit vztah dvou a více nezávisle proměnných se závisle proměnnou. Umožňuje zjistit, který faktor má vliv na změnu cílové proměnné a u kterého faktoru je naopak vliv na cílovou proměnnou nevýznamný. Stejně jako u jednoduché regresní analýzy je jedním z výsledků definice tvaru regresní funkce (rovnice), do které po vložení hodnot nezávisle proměnných lze predikovat hodnotu závisle proměnné. Je však třeba mít na paměti, že výsledky se vztahují pouze k predikci cílové proměnné u jedince, který pochází z výzkumného souboru. Například pokud jsme na příkladech aplikovali vícenásobnou regresní analýzu na studenty (muže) tělesné výchovy, nelze tvar regresní rovnice zobecnit na běžnou populaci či na studentky (ženy) tělesné výchovy.

Vícenásobnou regresní analýzu lze provádět několika metodami. V textu byla popsána metoda Enter a metoda Stepwise. U každé z těchto metod dosáhneme poněkud odlišných výsledků.

V níže uvedeném obr. 56 je schematicky naznačen postup vícenásobné regresní analýzy.



Obr. 56 Schéma postupu vícenásobné regresní analýzy

Zdroj: autoři

Závěrem

Tato skripta vyplňují nedostatečnou publikační činnost v oblasti praktické aplikace vybraných kvantitativně orientovaných metod v pedagogicko-psychologickém výzkumu. Orientace obsahu této učební opory na praktické aplikace vybraných metod statistické analýzy dat odlišuje tento text od ostatních odborných publikací, čímž může oslovit především studenty a studentky pregraduálního studia koncipující svůj výzkum v rámci svých závěrečných prací.

Do budoucna je cílem tento učební text rozšířit o další často používané metody statistické analýzy. Bude se jednat například o témata testování rozdílů mezi parametrickými a neparametrickými daty, normování dat či aplikaci statistické analýzy na nominální data.

Seznam tabulek

Tab. 1 Popis os používaných v 2D grafech.....	8
Tab. 2 Hmotnost žáků	16
Tab. 3 Rozložení četností hodnot statistického znaku.....	17
Tab. 4 Rozložení četností hodnot statistického znaku pro vážený průměr	17
Tab. 5 Rozložení relativních četností hodnot statistického znaku pro vážený průměr	18
Tab. 6 Hmotnosti osmiletých chlapců [kg] ze tříd 3.A a 3.B	24
Tab. 7 Manuální výpočet směrodatné odchylky	25
Tab. 8 Přehled měření s příklady	38
Tab. 9 Možnosti deskriptivní analýzy vzhledem k použitému měřítku	39
Tab. 10 Přehled možné analýzy dat pro hodnocení jednotlivých položek a hodnocení škály jako celku	41
Tab. 11 Detekce odlehlých hodnot.....	45
Tab. 12 Popis jednotlivých buněk z obr. 33	53
Tab. 13 Kritické hodnoty pro výpočet Dean-Dixonova vzorce.....	53
Tab. 14 Popis jednotlivých buněk v MS Excel pro výpočet Dean-Dixonova vzorce	54
Tab. 15 Popis jednotlivých buněk z obr. 34	56
Tab. 16 Kritické hodnoty pro výpočet Grubbsonova testu	56
Tab. 17 Shrnutí rezistentních odhadů.....	58
Tab. 18 Uplavaná a uběhnutá vzdálenost v Cooperově testu u mužů	61
Tab. 19 Korelační tabulka z dat z tab. 18.....	64
Tab. 20 Síla asociace proměnných dle různých autorů	66
Tab. 21 Korelační koeficienty mezi sledovanými jevy	70
Tab. 22 Výsledky škálování dvou kritérií (náznornosti a odbornosti) včetně uvedení jejich pořadí mezi různými nakladateli.....	72
Tab. 23 Obvod pasu a krku u dvaceti probandů.....	74
Tab. 24 Výsledky škálování map, fotografií a schémat v učebnicích vlastivědy mezi různými nakladatelstvími (včetně uvedení pořadí).....	76
Tab. 25 Pozorované četnosti počtu respondentů s volbou určitého kartografického znaku..	78
Tab. 26 Postup výpočtu hodnoty sr potřebné k určení intervalu spolehlivosti.....	81
Tab. 27 Výsledky Cooperova testu v běhu a plavání u 19 probandů včetně jejich hmotnosti	85
Tab. 28 Výsledky závislosti uběhnuté vzdálenosti na nezávisle proměnných.....	89
Tab. 29 Výsledky závislosti hmotnosti na nezávisle proměnných.....	91
Tab. 30 Různé výsledky regresního modelu v závislosti na zařazení nezávisle proměnných do regresní funkce.....	92

Seznam obrázků

Obr. 1 Umístění grafů v programu Statistika.....	9
Obr. 2 Počet studentů jednotlivých oborů na fakultě.....	10
Obr. 3 Četnost hmotnosti u dvanáctiletých chlapců v rámci jedné třídy.....	10
Obr. 4 Znázornění polední teploty za sedm dní v Krupce	11
Obr. 5 Znázornění výšky a hmotnosti probandů.....	11
Obr. 6 Korelační diagram.....	12
Obr. 7 Interpretace dat z více zdrojů	12
Obr. 8 Ukázka korelací.....	13
Obr. 9 Ukázka korelací včetně negativní korelace.....	13
Obr. 10 Rozložení jednotlivých ročníků ve výzkumném souboru	14
Obr. 11 Genderové rozložení jednotlivých ročníků ve výzkumném souboru	14
Obr. 12 Možnosti použití jednotlivých typů grafů	15
Obr. 13 Postup výpočtu průměru v programu Excel a Statistica.....	21
Obr. 14 Výpočet naší vzorové směrodatné odchylky v MS Excelu, výpočet ve Statistice znázorňuje obr. 13.....	27
Obr. 15 Příklad krabicového grafu.....	28
Obr. 16 Porovnání výsledků testů mezi čtyřmi školami	28
Obr. 17 Znázornění oboustranného 90 % intervalu spolehlivosti na křivce standardizovaného normálního rozdělení	30
Obr. 18 Normální rozdělení četnosti včetně Boxplotu.....	32
Obr. 19 Graf normálního (Gaussova) rozdělení.....	32
Obr. 20 Histogram a normální rozdělení.....	33
Obr. 21 Postup zjištění normality.....	34
Obr. 22 Různé postupy zjišťování normality.....	36
Obr. 23 Přehled typů proměnných	37
Obr. 24 Základní pojmy kódování (příklady)	40
Obr. 25 Box-Coxova transformace v programu Statistica	45
Obr. 26 Rozdíl mezi aritmetickým průměrem a mediánem	46
Obr. 27 Výpočet průměru a mediánu v programu Statistica.....	47
Obr. 28 Ukázka krabicového grafu s vyznačenými kvartily.....	48
Obr. 29 Nabídka „Zobrazit všechny grafy“ v MS Excel	49
Obr. 30 Ukázka vytvořeného krabicového grafu v MS Excel.....	49
Obr. 31 Tvorba krabicového grafu v programu Statistica.....	51
Obr. 32 Řazení hodnot v programu MS Excel	52
Obr. 33 Výpočet Dean-Dixonova vzorce v programu MS Excel.....	53

Obr. 34 Postup výpočtu Grubbsonova testu	56
Obr. 35 Výpočet useknutého a winsorizovaného průměru v programu Statistica	59
Obr. 36 Zobrazení dvojrozměrných dat v MS Excel	62
Obr. 37 Graf z dat z tab. 18 s lineární spojnicí trendu vytvořený v programu MS Excel.....	63
Obr. 38 Graf s daty, mezi kterými neexistuje závislost.....	63
Obr. 39 Postup výběru vhodného korelačního koeficientu	66
Obr. 40 Výpočet Pearsonova korelačního koeficientu v MS Excel	68
Obr. 41 Výpočet Pearsonova korelačního koeficientu v programu Statistica	69
Obr. 42 Výpočet rozdílu mezi korelačními koeficienty v programu Statistica.....	71
Obr. 43 Výpočet Spearmanova koeficientu korelace v programu Statistica.....	73
Obr. 44 Výpočet Spearmanova korelačního koeficientu v programu Statistica (závislost obvodu pasu a krku).....	75
Obr. 45 Výpočet Kendallova koeficientu shody v programu Statistica	76
Obr. 46 Vložení regresní funkce v MS Excel.....	80
Obr. 47 Vložení koeficientu determinace (R^2) v MS Excel.....	82
Obr. 48 Výsledek regresní analýzy v MS Excel s vyznačením důležitých údajů	83
Obr. 49 Korelační analýza v MS Excel.....	86
Obr. 50 Výsledek korelační analýzy v MS Excel.....	86
Obr. 51 Vícenásobná regresní analýza v MS Excel (metoda Enter).....	87
Obr. 52 Výsledek vícenásobné regresní analýzy v MS Excel (metoda Enter).....	88
Obr. 53 Vícenásobná regresní analýza v MS Excel (metoda Stepwise).....	89
Obr. 54 Výsledek vícenásobné regresní analýzy metodou Stepwise	90
Obr. 55 Výsledek vícenásobné regresní analýzy (vliv výšky a uplavané vzdálenosti jedince na jeho hmotnost).....	91
Obr. 56 Schéma postupu vícenásobné regresní analýzy.....	93

Zdroje

- Atanassova, V. (2010). Spiritia [cit. 2018-03-31]. Dostupný pod licencí Creative Commons na [www: <https://commons.wikimedia.org/wiki/File:Strong--weak--no-correlation.png>](https://commons.wikimedia.org/wiki/File:Strong--weak--no-correlation.png)
- Bednářová, I. (2019a). *Vylučování extrémních hodnot souboru*. Dostupné z: <https://cit.vfu.cz/statpotr/POTR/Teorie/Predn2/extremy.htm>.
- Bednářová, I. (2019b). *Statistické tabulky*. Dostupné z: <https://cit.vfu.cz/statpotr/POTR/Teorie/tabulky.htm#Dixon>.
- Budíková, M., Králová, M. & Maroš, B. (2010). *Průvodce základními statistickými metodami*. Praha: Grada. Budíková, M., Lerch, T. & Mikoláš, Š. (2005). *Základní statistické metody*. Brno: Masarykova univerzita.
- Drápela, K. (2012). *Průzkumová analýza dat*. Brno: Masarykova univerzita. Dostupné z: http://user.mendelu.cz/drapela/Statisticke_metody/teorie%20text%20II.pdf.
- Dwedland (2013) [cit. 2018-03-31]. Dostupný pod licencí Creative Commons na [www: <https://commons.wikimedia.org/wiki/File:Correlation_image.JPG>](https://commons.wikimedia.org/wiki/File:Correlation_image.JPG)
- E-academia.cz (2014) [cit. 2018-03-25]. Dostupný z: http://www.e-academia.cz/online-video-kurzy/uvod-do-statistiky/course.php?lecture_id=24.
- Epina e-Book Team (2012). *Fundamentals of Statistics*. Dostupné z: http://www.statistics4u.com/fundstat_eng/ee_grubbs_outliertest.html.
- FachueberM28 (2011) [cit. 2018-02-21]. Dostupný pod licencí Public Domain na [www: <https://commons.wikimedia.org/wiki/File:Boxplot_BeispielM28.jpg>](https://commons.wikimedia.org/wiki/File:Boxplot_BeispielM28.jpg).
- Gavora, P. (1996). *Výzkumné metody v pedagogice: příručka pro studenty, učitele a výzkumné pracovníky*. Brno: Paido.
- Gavora, P. (1999). *Úvod do pedagogického výzkumu*. Bratislava: Univerzita Komenského.
- Gavora, P. (2010). *Úvod do pedagogického výzkumu*. Brno: Paido.
- Havel, Z. & Hnizdil, J. (2008). *Cvičení z antropomotoriky*. Ústí nad Labem: Univerzita J. E. Purkyně.
- Havel, Z. & Cihlář, D. (2011). *Vybrané neparametrické statistické postupy v antropomotorice*. Ústí nad Labem: Univerzita J. E. Purkyně.
- Hayes, N. (2003). *Základy sociální psychologie*. Praha: Portál.
- Hendl, J. (2012). *Přehled statistických metod*. Praha: Portál.
- Hendl, J. (2014). *Statistika v aplikacích*. Praha: Portál.

- Hendl, J. (2015). *Přehled statistických metod: analýza a metaanalýza dat*. Praha: Portál.
- Holčík, J. & Komenda, M. (2015). *Matematická biologie: e-learningová učebnice*. Brno: Masarykova univerzita.
- Chráska, M. (2000). *Základy výzkumu v pedagogice*. Olomouc: Univerzita Palackého.
- Chráska, M. (2016). *Metody pedagogického výzkumu: základy kvantitativního výzkumu*. Praha: Grada.
- Chytrý, V. & Kroufek, R. (2017). Možnosti využití Likertovy škály – základní principy aplikace v pedagogickém výzkumu a demonstrace na příkladu zjišťování vztahu člověka k přírodě. *Scientia in educatione*, 8(1), 2–17.
- Jarekt. (2010). [cit. 2018-03-30]. Dostupný pod licencí Creative Commons na [www: <http://commons.wikimedia.org/wiki/File:Data_Fusion_-_Scatter_plot.png>](http://commons.wikimedia.org/wiki/File:Data_Fusion_-_Scatter_plot.png)
- Jhguch (2012). [cit. 2018-02-21]. Dostupný pod licencí Creative Commons na [www: <http://commons.wikimedia.org/wiki/File:Boxplot_vs_PDF.svg>](http://commons.wikimedia.org/wiki/File:Boxplot_vs_PDF.svg)
- Joxemai (2013). [cit. 2018-03-25]. Dostupný pod licencí Creative Commons na [www: <http://commons.wikimedia.org/wiki/File:Banakuntza_normala_histograma_01.png>](http://commons.wikimedia.org/wiki/File:Banakuntza_normala_histograma_01.png)
- KendallVarent (2010).[cit. 2018-03-23]. Dostupný pod licencí Public Domain na [www: <http://commons.wikimedia.org/wiki/File:Confidence_Interval_90P.png>](http://commons.wikimedia.org/wiki/File:Confidence_Interval_90P.png)
- Kirk, A. (2016). *Data vizualization*. London: SAGE publication.
- Litschmannová, M. (2012). *Úvod do statistiky (interaktivní učební text)*. Ostrava: Vysoká škola báňská – Technická univerzita Ostrava. Dostupné z: <http://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/interaktivni_uvod_do_statistiky.pdf>.
- Malíková, K. (2014). *Shapirův-Wilkův test normality* (bakalářská práce). Praha: Univerzita Karlova.
- Magnello, E. (2010). *Statistika*. Praha: Portál.
- Maňák, J. (1994). *Kapitoly z metodologie pedagogiky*. Brno: Masarykova univerzita.
- Marek, L. (2015). *Statistika v příkladech*. Praha: Kamil Mařík – Professional Publishing.
- Meloun, M. & Kupka, K. (2001). The Box-Cox transformation for rigorous statistical analysis of metallurgical data. *Acta Metallurgica Slovaca*, 34(7), 1–8.
- Neubauer, J., Sedlačík, M. & Kříž, O. (2016). *Základy statistiky: aplikace v technických a ekonomických oborech*. 2., rozšířené vydání. Praha: Grada.

- Parker, L. & Lunney, M. (1998). Moving beyond content validation of nursing diagnoses. *Nursing Diagnosis*, 9(4), 144–150.
- Průcha, J., Walterová, E. & Mareš, J., (2009). *Pedagogický slovník*. Praha: Portál.
- Rod, A. (2012). Likertovo škálování. *E-Logos Electronic Journal for Philosophy*, 13(1), 2–14.
- SC&C Partner (2015). *Minitab 17 – příručka uživatele*. Brno: SC&C Partner
- Sewaqu (2010). [cit. 2018-03-30]. Dostupný pod licencí Public Domain na [www:
<https://commons.wikimedia.org/wiki/File:Linear_regression.svg>](http://www.commons.wikimedia.org/wiki/File:Linear_regression.svg)
- Subedi, B. P. (2016). Using Likert type data in social science research: Confusion, issues and challenges. *International Journal of Conterporary Applied Sciences*, 3(2), 36–49.
- Škoda, J. & Doulík, P. (2007). *Tvorba a hodnocení didaktických testů: cvičebnice pro studenty učitelství a účastníky kurzu DPS*. Ústí nad Labem: Univerzita J. E. Purkyně.
- Švaříček, R. & Šed'ová, K. (2014). *Kvalitativní výzkum v pedagogických vědách*. Praha: Portál.
- Tavakol, M. & Dennick, R. (2011). Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, 2(1), 53–55.
- Toews, M. W. (2007). [cit. 2018-03-24]. Dostupný pod licencí Creative Commons na [www:
<https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg>](http://www.commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg)
- Walker, I. (2013). *Výzkumné metody a statistika*. Praha: Grada.

Přílohy

Příloha 1: kritické hodnoty součinného (Pearsonova) korelačního koeficientu r

v / α	0,95	0,99	v / α	0,95	0,99
1	0,997	0,999	51	0,271	0,351
2	0,950	0,990	52	0,268	0,348
3	0,878	0,959	53	0,266	0,345
4	0,811	0,917	54	0,263	0,342
5	0,755	0,875	55	0,261	0,339
6	0,707	0,834	56	0,259	0,336
7	0,666	0,798	57	0,256	0,333
8	0,632	0,767	58	0,254	0,330
9	0,602	0,734	59	0,252	0,327
10	0,576	0,708	60	0,250	0,323
11	0,553	0,684	61	0,248	0,322
12	0,532	0,661	62	0,246	0,320
13	0,514	0,641	63	0,244	0,317
14	0,497	0,623	64	0,242	0,315
15	0,482	0,601	65	0,241	0,313
16	0,468	0,590	66	0,239	0,310
17	0,456	0,575	67	0,237	0,318
18	0,444	0,561	68	0,235	0,306
19	0,433	0,549	69	0,234	0,304
20	0,423	0,537	70	0,232	0,302
21	0,413	0,526	71	0,230	0,300
22	0,404	0,515	72	0,229	0,299
23	0,396	0,505	73	0,227	0,296
24	0,389	0,496	74	0,226	0,294
25	0,381	0,487	75	0,224	0,292
26	0,374	0,479	76	0,223	0,290
27	0,367	0,471	77	0,213	0,288
28	0,361	0,463	78	0,220	0,286
29	0,356	0,456	79	0,219	0,285
30	0,349	0,449	80	0,218	0,283
31	0,344	0,442	81	0,216	0,281
32	0,339	0,436	82	0,215	0,280
33	0,334	0,423	83	0,213	0,278
34	0,329	0,424	84	0,212	0,276
35	0,325	0,418	85	0,211	0,275
36	0,320	0,413	86	0,210	0,273
37	0,316	0,408	87	0,208	0,272
38	0,312	0,403	88	0,207	0,270
39	0,308	0,398	89	0,206	0,269
40	0,304	0,393	90	0,205	0,267
41	0,301	0,389	91	0,204	0,266
42	0,297	0,384	92	0,202	0,265
43	0,294	0,380	93	0,201	0,263
44	0,297	0,376	94	0,200	0,262
45	0,288	0,373	95	0,199	0,260
46	0,285	0,368	96	0,198	0,259
47	0,282	0,365	97	0,197	0,258
48	0,279	0,361	98	0,196	0,257
49	0,276	0,358	99	0,195	0,255
50	0,273	0,354	100	0,194	0,254

Příloha 2: kritické hodnoty pořadového (Spearmanova) korelačního koeficientu r

v / α	0,95	0,99
4	1,000	
5	0,900	1,000
6	0,829	0,943
7	0,714	0,893
8	0,643	0,833
9	0,600	0,783
10	0,564	0,764
12	0,506	0,712
14	0,456	0,645
16	0,425	0,601
18	0,399	0,564
20	0,377	0,534
22	0,359	0,508
24	0,343	0,485
26	0,329	0,465
28	0,317	0,448
30	0,306	0,432