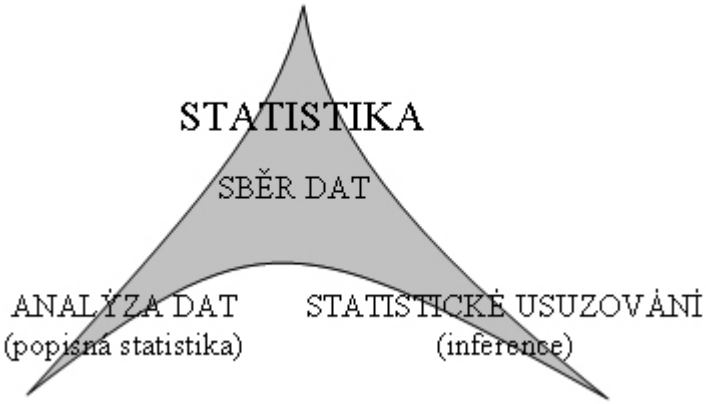
ZÁKLADY STATISTIKY

1 ÚVOD

Struktura

Jako každá vědecká disciplína, taktéž statistika má svou strukturu, členění. V našem případě bude mít toto členění především didaktický význam, neboť Vám umožní lépe pochopit smysl jednotlivých nástrojů statistiky.



Obrázek 1.1 Struktura statistiky

Popisná statistika (rovněž označovaná deskriptivní statistika) v sobě zahrnuje takové postupy, které umožňují dle naměřených skóre vyslovit závěry o vlastnostech souboru v určité "koncentrované podobě". Získané charakteristiky sledovaný jev pouze popisují, neumožňují přesnější srovnání mezi soubory dat. Často těchto charakteristik ve snaze podat zprávu o stavu jevu. Mezi základní charakteristiky lze zařadit střední hodnoty, hodnoty variability, rozdělení četností, aj.

Inferenční statistika. Jedná se o část statistiky, která umožňuje hlouběji zkoumat jevy, především ve smyslu ověřování předpokladů. Účel sběru dat zde souvisí s původním záměrem ověřit předpoklad. Na začátku libovolného výzkumu vždy stojí určitý názor na sledovaný jev, myšlenka (zvolená tréninková metoda měla pozitivní efekt na úroveň silových předpokladů sportovce). Postavení "Statistiky" je zde jako prostředku ověření pravdivosti tohoto názoru, myšlenky. Vzhledem k povaze sledovaného jevu a jiným okolnostem je zvolen odpovídající design sběru dat. Jejich dalším zpracováním lze poté vyslovit úsudek o původnímu názoru.

V inferenční statistice se již nelze spokojit s pouhým popisem jevu pomocí statistických charakteristik. Nyní se dostáváme k dalšímu kroku, kdy již s vybranými charakteristikami operujeme (snažíme se odpovídat na otázky typu "Liší se průměry dvou sledovaných souborů?", "Existuje vztah mezi dvěmi jevy?" aj.). Většinou ověřujeme tzv. *statistickou hypotézu* - tvrzení, které je možné jednoznačně ověřit. Její nulová forma vždy předpokládá jistý "*nulový rozdíl* " či "*nezávislost jevů*". Např. předpoklad, že dva výběrové průměry jsou stejné, tj. rozdíl mezi těmito průměry je roven nule. Dodržením přesných postupů a užitím odpovídajících metod lze tyto tvrzení buďto vyvracet a nebo výsledek uzavřít výrokem, že hypotézu o nulovém rozdílu nelze vyvrátit (s jistou mírou nepřesnosti znamená "existence rozdílu").

2 ZÁKLADNÍ STATISTICKÉ POJMY

Na začátku jakékoliv práce s daty je bezpochyby nutné přesně vědět, „Kdo a co“ je těmito daty. Není to nutné pouze pro vlastní potřebu při dalším zpracovávání dat (neboť ne na „všechny šroubky je stejný šroubovák“), ale také pro osoby, které mohou následně s našimi daty a výsledky pracovat a dále operovat. Je potřeba vědět, čeho a koho se čísla týkají, kdy byla data měřena, jaký je zdroj dat apod. Tak jako každá vědní disciplíně i statistika vládne vlastním slovním aparátem, který umožňuje data vymezit, pojmenovat, interpretovat, … V této se kapitole se seznámíte s výkladem základních pojmů, které budou v dalším textu běžně užívány.

Statistický soubor, výběr

Základní soubor (populace). Je souborem všech prvků, které na základě vymezené vlastnosti mohou teoreticky být předmětem sledování.

Příklad: Mějme množinu osob, které budou mít společnou tu vlastnost, že studují na Univerzitě Palackého v Olomouci. Tuto množinu pak nazveme statistickým souborem, přičemž je vytvořen na podkladě několika selekčních vlastností: (1) studující, (2) na Univerzitě Palackého, (3) v Olomouci.

Statistická jednotka. Prvky statistického souboru, které mají alespoň jednu společnou vlastnost, nazýváme statistickými jednotkami. Lze se setkat také s rovnocenným pojmem *případ*, především pak v prostředí statistického softwaru a v anglické literatuře (*case*). Statistické jednotky jsou často potenciální povahy a jejich existence je dána pouze jejich teoretickým vymezením. V důsledku toho má základní soubor zpravidla značný rozsah (značíme N) a zjištění zkoumaných vlastností všech jednotek je buďto nemožné nebo je příliš časově a ekonomicky náročné (např. zjištění výšky a váhy všech studujících tělesné výchovy v ČR).

Rozsah souboru (N). Rozsahem souboru chápeme počet případů (statistických jednotek; prvků), které vybraný soubor obsahuje. V závislosti na tom, zda se jedná o základní či výběrový soubor, jej značíme N či n.

Výběrový soubor, náhodný výběr. Z důvodů časových, ekonomických, prostorových nebo jiných nároků na sledování celého základního souboru snižujeme přesně vymezeným postupem počet jeho případů (rozsah). Tímto postupem (popsaným dále a dále pak v přiloženém příkladu) dostáváme svým rozsahem soubor menší a nazýváme jej *výběrový soubor* (jeho rozsah značíme n). Nejčastějším postupem je tzv. *náhodný výběr*. Je to takový výběr případů ze základního souboru, že každý z nich má stejnou možnost být vybrán. Znamená to, že pravděpodobnost, že bude případ vybrán, je pro všechny stejná. Můžeme také říct, že o tom, zda bude nebo nebude prvek vybrán vybrání, rozhoduje pouze náhoda.

Jmenujme tři praktické možnosti tvorby výběrového souboru metodou náhodného výběru:

- losováním statistických jednotek s jejich vracením do osudí (u malých souborů);
- losováním statistických jednotek bez vracení do osudí (u velkých souborů);
- pomocí tabulky náhodných čísel nebo pomocí určené funkce v prostředí statistického softwaru (generuje náhodné čísla).

Stratifikovaný výběr. Často se můžeme setkat s případy, kdy má statistický soubor určité vnitřní uspořádání, členění do několika typických skupin. Například populaci studentů Univerzity Palackého v Olomouci by bylo možné rozdělit na 8 částí podle toho, na které fakultě studují (Cyrilometodějská teologická fakulta, Lékařská fakulta, Filozofická fakulta,

Přírodovědecká fakulta, Pedagogická fakulta, Fakulta tělesné kultury, Právnická fakulta, Fakulta zdravotnických věd). Celá populace je takto rozdělena v určitém poměru (řekněme 5:18:5:25:15:25:6 %). Budeme-li chtít, aby byl výběrový soubor dostatečně reprezentativní, musíme dodržet při výběru jeho uspořádání. Je tedy nutné postupovat tak, aby byl poměr zachován i ve výběrovém souboru – z každé fakulty vybereme pouze počet studentů odpovídající poměrovému zastoupení. Tento typ výběru nazýváme stratifikovaným výběrem.

Typy statistických znaků

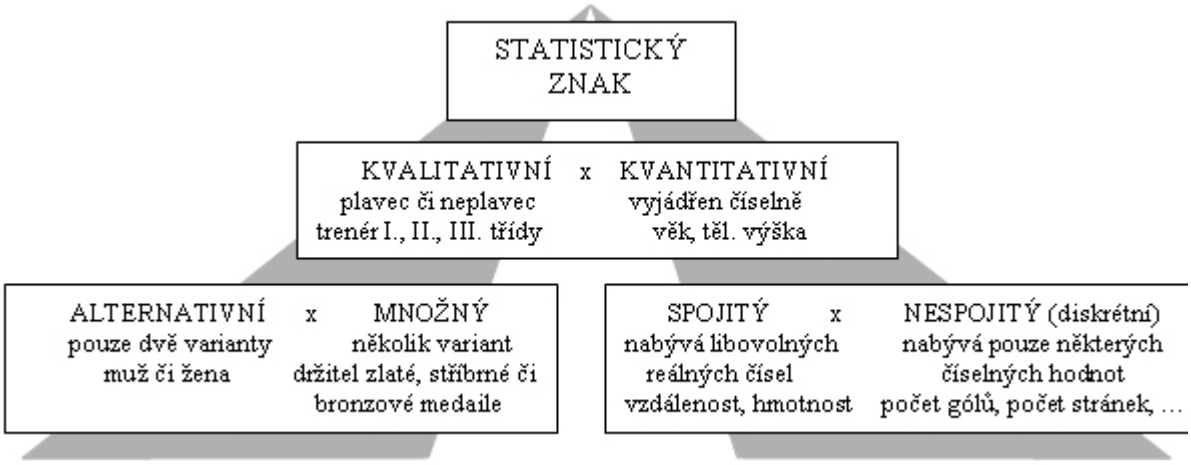
Statistický znak, proměnná. Statistické znaky nebo proměnné jsou charakteristiky prvků základního souboru, jež mohou nabývat více hodnot a existuje pro ně předpis, jak tyto hodnoty zjistíme (Hendl, 2004). Například student řešící problematiku vlivu tréninkové zátěže na kvalitu spánku v rámci své diplomové práce bude pracovat se dvěmi proměnnými. První proměnná bude souviset s tréninkovou zátěží (uběhnuté kilometry apod.) a druhá proměnná bude souviset s kvalitou spánku (délka non-REM spánku či jiná kvalita). Pojmy proměnná a znak jsou často užívány souběžně. Pojem statistický znak však chápeme jako určitou společnou vlastnost jednotek statistického souboru, která se může stát předmětem sledování. Při vlastním měření tento znak u jednotek nabývá různých hodnot, a proto již hovoříme o proměnné.

V závislosti na tom, kolik znaků u statistické jednotky určujeme, resp. sledujeme, lze členit statistický soubor na (1) jednorozměrný (sledujeme jeden znak) a (2) dvou- či vícerozměrný (sledujeme dva či více znaků).

Před vlastním šetřením by měla osoba zpracovávající data přesně vědět, s jakým typem statistického znaku pracuje, neboť to značně ovlivňuje možnosti zpracování a analýzy dat. V případě, že by chybně stanovila typ statistického znaku nebo použila statistický postup, který danému typu znaku nenáleží, dopustila by se samozřejmě chyby. To by s největší pravděpodobností vedlo k chybným výsledkům a následně i závěrům. Přinejmenším by nebylo možné nic říct o pravdivostní hodnotě zjištěných výsledků.

Jak je naznačeno na obr. 2.1, lze členění statistického znaku rozdělit do dvou úrovní. V první úrovni dělíme statistický znak na *kvalitativní* nebo *kvantitativní*. *Kvalitativní znak* je vesměs vyjádřen slovně. Lze se ovšem setkat s vyjádřením pomocí písmene nebo i číslice, které však nevyjadřují množství (resp. míru), ale jsou pouze zástupci (vyjádřením) sledované kvality jevu. Např. v tabulce dat, kde popisovaným znakem je plavecká dovednost, se lze setkat s hodnotami „1“ a „0“ u jednotlivých statistických jednotek, ovšem tyto číslice nevyjadřují míru této plavecké dovednosti, ale jsou pouze kódy pro úroveň znaku. Řekněme, že „0“ znamenalo v zápise „neplavec“ a „1“ znamenalo „plavec“. Význam však může být rovněž opačný. Při takovémto kódování je tedy zapotřebí souběžně s daty uvádět způsob jejich šifrování. Toto kódování se provádí čistě z praktických důvodů (rychlost zápisu, přehlednost, …). Při tomto vyjádření úrovně znaku je zapotřebí mít neustále na paměti, že se nejedná o číslo v pravém slova smyslu, ale pouze o znak. Především v prostředí statistického softwaru, který nedokáže rozeznat, zda jsou vaše hodnoty skutečné číslice nebo jen znaky, což může vézt k nesmyslným postupům a závěrům. Vezmeme-li si opět do úst předchozí příklad o plavecké dovednosti – bez uvědomění si souvislostí s určením typu znaku, budeme chybně (a nesmyslně!) určovat aritmetický průměr. Statistický software bude sčítat nuly s jedničkami (tedy plavce s neplavci!) a dělit je jejich počtem – to samozřejmě nelze. V další úrovni členění se u kvalitativního znaku dostáváme k rozdělení znaků na *alternativní* a *množné*. *Alternativní znak* zahrnuje vždy pouze a jen dvě varianty. Jedná se tak vlastně o dichotomické členění. U *množného znaku* je vždy možnost více alternativ, více úrovní znaku.

Dalším typem znaku, jak je vidět na obrázku, v rámci první úrovně je *znak kvantitativní*. Ten je vyjádřen vždy číselně. Číslice zastupuje nějaké množství nebo míru sledovaného znaku. Jeho dalším členěním se dostáváme k znakům spojitým a diskrétním. Znak kvantitativní spojitý může nabývat libovolné reálné hodnoty a jeho hodnoty tak lze vnímat jako libovolný bod číselné osy. U těchto znaků bude platit, že mezi libovolnými dvěmi hodnotami teoreticky existuje přinejmenším jedna další hodnota. Znak kvantitativní diskrétní (též nespojitý) nabývá pouze některých hodnot (většinou v oboru celých čísel). U znaku tohoto typu vždy existují takové dvě hodnoty, které nazveme sousedními a mezi nimiž teoreticky neexistuje hodnota, která by ležela mezi nimi.



Obrázek 2.1 Přehled typů statistických znaků

Závislá a nezávislá proměnná (dependent and independent variable). Většina výzkumných studií rozlišuje mezi těmito dvěmi typy proměnných. Někdy označujeme závisle proměnnou jako odpověďová, kriteriální nebo cílová; nezávisle proměnnou predikátor nebo explanační proměnná. Příklady: (1) závislost prospěchu (závisle proměnná) na pohlaví žáka (nezávisle proměnná); (2) závislost průměrného prospěchu (závisle proměnná) na počtu žáků ve třídě (nezávisle proměnná); závislost síly (závisle proměnná) jedince na objemu jeho svalů (nezávisle proměnná).

Příklad: „Základní pojmy“

Popis: Na prvním obrázku je vidět zápis dat do tabulky v prostředí aplikace Microsoft Excel. Nechť těchto 350 sledovaných osob představuje základní soubor - 350 případů (statistických jednotek), což jsou zde osoby, které nesou jméno „jméno 1“ až „jméno 350“. Rozsah souboru je tedy $N = 350$. Sledovány byly dva statistické znaky (dvě proměnné) – první „B1500M“ (běh 1500 metrů) a druhá „P100M“ (plavání 100 m). Jak první, tak i druhý statistický znak jsou znaky kvantitativní spojité, neboť měrnou jednotkou, kterou sledujeme, je čas. Ten jak víme nabývá libovolné hodnoty z oboru reálných čísel.

V pravé části obrázku je připravena tabulka pro náhodný výběrový soubor, který bude rozsahu $n = 50$. Úkolem je vytvořit náhodným výběrem vytvořit výběrový soubor. K tomu využijeme funkci obsaženou v MS Excel, která generuje náhodná čísla. Postup bude následující:(a) ke každému případu přiřadíme náhodné číslo pomocí funkce "NÁHČÍSLO( )"; (b) seřadíme případy podle velikosti vzestupně dle hodnoty náhodného čísla; (c)

vybereme prvních 50 případů.

Popis tvorby výběrového souboru:

1. Na stranách 2-6 je zobrazen postup výběru funkce generující náhodné číslo. První náhodné číslo generujeme v buňce "F5". Kliknutím na tuto buňku ji aktivujeme. Nyní klikneme na tlačítko „vložit funkci“ (str. 2) a tím zobrazíme nabídku funkcí, které můžeme využít.
2. Zde do první příkazové lišty charakterizujeme jednoduše funkci, kterou chceme nalézt – v našem případě nechť je to heslo „Náhodné číslo“; klikneme na „Přejít“ (str.3). Tím získáme užší výběr funkcí, které odpovídají popisu – v našem případě vyhovuje pouze jedna funkce – „NÁHČÍSLO“. Její bližší popis najdeme v dolní části okna. Vybereme tuto funkci a pokračujeme přes tlačítko „OK“ (str.4).
3. Přecházíme do okna „Argumenty funkce“. Zde se jindy zadávají nezbytné údaje pro funkci (argumenty), v našem případě však zde není zapotřebí vyplňovat žádné hodnoty, neboť funkce NÁHČÍSLO není závislá na žádných argumentech (str.5). Pokračujeme „OK“ a v zadané buňce dostáváme číselnou hodnotu z intervalu mezi nulou a jedničkou (str.6). Toto číslo nemá ukončený desetinný rozvoj, o čemž by se bylo možné přesvědčit přidáváním desetinných míst prostřednictvím tlačítka „vložit desetinné místo“ v horním panelu nástrojů.
4. Podobně je potřeba vygenerovat náhodné čísla pro ostatních 349 případů. Postup si zjednodušíme kopírováním vzorce, což se nám bude hodit také dalších případech. Je řada způsobů, jak kopírování docílit. Postupujme následovně: přiblížíme kurzor do blízkosti pravého dolního okraje aktivované buňky, ve které je již vzorec (funkce) použit (buňka F5), a to do takové blízkosti, až se kurzor změní na křížek tak (viz str.7). Podržíme-li levé tlačítko myši a budeme posouvat kurzor ve sloupci směrem dolů (str.8), budeme vzorec postupně kopírovat (str.9) – takto nakopírujeme vzorec pro všech 350 případů.
5. Nyní se dostáváme k další části – seřadit případy podle velikosti náhodného čísla. Na str.10 je vytvořen blok buněk B5 až F359 tak, aby v něm byly zahrnuty jak všechny případy, tak i všechny náhodné čísla. Dále využijeme položku „Data“ (jak naznačuje zelená šipka) a získáme tak nabídku pro práci s daty. Zde přejdeme hned na první položku „Seřadit“ (str.11). Přejdeme tak do okna (str.12), které nám dává několik možností volby. My si všimneme položky „Seřadit podle“. Chceme řadit náhodné čísla a proto vybereme „sloupec F“, kde jsou náhodné čísla umístěny. Budeme-li pokračovat přes OK, dostaneme již zaměněné pořadí případů 1 – 350, a to podle velikosti náhodných čísel (str.13). Všimněte si, že ačkoliv jsme řadili náhodné čísla podle velikosti, jejich hodnoty nyní seřazeny podle velikosti nejsou. To je dáno vlastností funkce „NÁHČÍSLO“, která po jakékoliv úpravě v sešitu vygeneruje nové náhodné čísla.
6. Posledním krokem náhodného výběru je vybrat do bloku prvních (ale může být také jiných za sebou jdoucích) 50 případů (str.14), kopírovat je kliknutím na blok pravým tlačítkem myši a výběrem položky „kopírovat“ (str.15), kliknutím pravým tlačítkem myši na první buňce tabulky náhodného výběru se dostaneme do výběru pro tuto buňku, vybereme vložit (str.16) a tím dostáváme výběrový soubor o rozsahu 50 případů (str.17).

(blíže k této problematice např. Hendl, 2004; Cyhelský, Novák, 1976)

TIP: Vytvořte náhodný výběr *(n* = 26) ze základního souboru - české města, jejichž počet

obyvatel převyšuje 2000 (seznam lze najít na stránkách Českého statistického úřadu, Wikipedia, Ministerstva pro místní rozvoj,...).

3 TEORIE MĚŘENÍ, MĚŘÍCÍ STUPNICE

Teorie měření

Měření se vyvinulo v průběhu historického vývoje lidské společnosti v běžnou každodenní proceduru, kterou si ani neuvědomujeme (užití hodinek, tachometru automobilu, atd.). Měření je nejen součástí naší každodenní zkušenosti, nýbrž také základem věd - obzvláště přírodních.

Historické počátky měření spočívaly v porovnávání objektů s počtem prstů, délkou palce, délkou chodidla, paže atd. Tyto primitivní měřící způsoby se s rozvojem vědy a techniky vyvinuly v měřící procedury založené často na složitých měřících přístrojích.

Otázky spjaté s problematikou kvantifikace řeší obor nazývaný teorie měření. Problematika teorie měření je podrobně rozpracována řadou autorů, přičemž ranná historie měření začíná již před rokem 1900 a je neodlučitelně spjata zvláště se jmény Helmholtz (1887), Campbell (1920) a Stevens (1946).

Teorie měření zjišťuje podmínky popř. předpoklady měřitelnosti různých vlastností. Měřitelné jsou jak fyzikální vlastnosti (např. délka, čas, hmotnost), tak i psychické vlastnosti (např. inteligence, strach, postoje) apod. Podle Campbellovy všeobecně uznávané reprezentační teorie měření lze vymezit pojem měření jako přiřazování číslic k reprezentaci vlastností daného jevu. Tuto koncepci měření doplnil Stevens o formulaci podmínek, za nichž je měření uskutečnitelné. Podle něj lze za měření považovat každé přiřazování číslic k objektům nebo událostem…podle pravidel.

Klasická koncepce měření vychází z rozlišení na fundamentální a odvozené měření, někteří autoři zmiňují ještě třetí druh: měření asociativní (Berka, 1977) resp. asociační (Blahuš, 1996), označované rovněž jako měření *per fiat*, *per Definition*, *by fiat* či měření *na základě konvence*. Tento třetí typ měření má svůj význam ve sportu, například při měření úrovně motorických schopností, které nejsou přímo měřitelné, a proto je měříme na základě asociace s jinou měrnou veličinou (zde specifickým motorickým výkonem).

Jednotlivé druhy měření lze charakterizovat takto:

- Fundamentální měření se vztahuje na bezprostřední měření extenzivních veličin“ a je to „každé měření, které nezahrnuje žádná předcházející měření“.
- Odvozené měření předpokládá jiná, již dříve provedená měření, z nichž je odvozeno na základě vztahů, a tedy závisí na předcházejících měřeních.
- Asociativní měření (asociační) je takové měření, kdy je přímo měřená veličina asociována s nepřímo měřitelnou veličinou, např. při měření teploty vycházíme ze závislosti změny objemu kapaliny na teplotě. Toto měření se často vyskytuje v oblasti sociálních věd, kde jde často o předpokládaný (hypotetický) vztah mezi pozorovanými vlastnostmi, které měřitelné nejsou a měřitelnými veličinami. Tento třetí typ měření má svůj význam ve sportu. Například při měření úrovně motorických schopností, které nejsou přímo měřitelné. Měříme je proto na základě asociace s jinou měrnou veličinou (specifickým motorickým výkonem). Příkladem lze uvést měření vytrvalostních schopností měřením výkonu v 12-minutovém běhu.

Z hlediska měřících procedur je možno měření dále diferencovat na

- *přímé měření*, které je založeno na bezprostředním srovnávání měřeného objektu se standardním měřidlem nebo se stupnicí měřícího přístroje;
- *nepřímé měření*, které zahrnuje přímé měření něčeho jiného a následně prováděné

výpočty.

Měřenou veličinu lze dle povahy rozdělit na

- *Extenzivní veličinu* – veličina charakterizující množství či objem určitého jevu;
- *Intenzivní veličinu* – veličina charakterizující intenzitu, resp. úroveň určitého jevu.

Měřící stupnice (škály)

Měřící stupnici (škálu) můžeme chápat jako určité měřítko, na kterém jsou hodnoty sledovaného jevu měřeny. Tato stupnice není pro všechny jevy svými vlastnostmi stejná. Na základě odlišností jednotlivých stupnic je možné rozlišit čtyři typy skupnic. Tyto odlišnosti jsou vždy dány povahou sledovaného jevu, jeho podstatou. Stupnice je odrazem toho, nakolik je zobrazení vlastnosti jevu do množiny reálných čísel (způsob, jakým přiřazujeme různým hodnotám proměnné čísla) plnohodnotné vzhledem k operacím mezi čísly.

Nominální (jmenná, klasifikační) stupnice. Číselné hodnoty této stupnice vznikají na základě přiřazování číslic jednotlivým úrovním jevu a tyto úrovně tak pouze pojmenovávají. Jde vlastně o pojmenování vlastností osob, věcí a jevů čísly. Vlastnosti jsou uspořádány do tříd, které se navzájem vylučují (např. čísla hráčů, pohlaví, kuřák a nekuřák, národnost, alternativní třídění). Na této stupnici jsou měřeny všechny kvalitativní veličiny. Základní empirickou operací je „určení rovnosti“, tj. relace =, "nerovnost". Při zpracování dat této stupnic lze používat pouze neparametrické statistické metody.

Ordinální (pořadová) **stupnice.** U této stupnice větší číselná hodnota představuje rovněž větší množství (úroveň) samotné proměnné a jednotlivé objekty tak mají přirozené uspořádání dle velikosti. Mezi hodnotami lze rozlišit vztah rovnosti či nerovnosti a vztah větší nebo menší (relace =, "nerovnost", >, <). Odstupy mezi hodnotami však nejsou známy (nelze například s jistotou říct, že mezi hodnotou 2 a 3 je stejná vzdálenost jako mezi 3 a 4). Jako příklad můžeme uvést body hodnocení v gymnastice, školní známky, stupnice tvrdosti nerostů. Tak jako u předchozí nominální stupnice i zde lze při zpracování kalkulovat pouze s neparametrickými statistickými metodami.

Metrická(intervalová a poměrová) **stupnice.** U měřené veličiny je zavedena jednotka měření a odstupy mezi hodnotami (intervaly) tak jsou na rozdíl od předchozí stupnice známy. To umožňuje užití parametrických statistických metod. Mezi metrické stupnice se řadí stupnice intervalová a poměrová.

- **Intervalová stupnice.** Vyžaduje stanovení měrové jednotky (měřítka; intervalu mezi hodnotami) a počátku měření (nulového bodu). Nula - nulový bod - je zvolená dohodou, většinou ve vztahu k nějakému jevu (u měření teploty ve °C je tímto bodem bod tání ledu. Pro tuto stupnici jsou přípustné všechny aritmetické operace, kromě násobení a dělení. Např. teplota ve °C, letopočet.
- **Poměrová stupnice.** Od intervalové stupnice se odlišuje pouze existencí přirozeného počátku, tzn. že nula je u této stupnice absolutní a znamená neexistenci jevu (je-li počet zásahů na koš v basketbalu roven nule, pak skutečně jev "zásah koše" nenastal). To již umožňuje uplatňovat všechny aritmetické operace včetně násobení a dělení. Např. věk, váha, počet (množství), výška, teplotní stupnice dle Kelvina, v podstatě všechny fyzikální jednotky.

Tabulka 3.1 Přehled typů škál (stupnic)

	Nemetrické škály		Metrické škály	
	NOMINÁLNÍ	ORDINÁLNÍ	INTERVALOVÁ	POMĚROV
Příklady	Číselné označení barev, psychologického typu, pohlaví, atd.	Školní známky, stupnice tvrdosti, služební pořadí, Richterova stupnice	Teplota ve °C, Fahrenheita, letopočet, inteligenční kvocient	Teplota °Ke věk, váha, vý velikost úhlu
Operace	=, ≠	=, ≠, >, <	Navíc: intervaly, nula zvolená	Navíc: nula absolutní
Statistické charakteristiky	Modus, Absolutní a relativní četnosti	Navíc: medián, procentily, kvantily a kvantilové odchylky,	Navíc: arit. průměr, směrodat.odchylka, šikmost, špičatost	Navíc: koeficient variability, geometr. prů
Testy Významnosti	χ 2 - test, McNemar test, Cochran test, ...	Znaménkový test, Mann-Whitney U-test, Friedmanova pořadová analýza variance, aj.	Parametrické metody: F-test t-test (pro závislé či nezávislé soubory)	Parametrické metody: F-test t-test (pro zá nezávislé sou
Míry závislosti	Kontingenční a čtyřpolní koeficient	Navíc: pořadová korelace	Navíc: Pearsonova součinová korelace	Navíc: Pearsonova součinová ko
Statistické metody	Některé neparametrické metody	Všechny neparametrické metody	Všechny neparametrické a parametrické metody	Všechny neparametric parametrické metody

4 DESKRIPTIVNÍ STATISTIKA

Data analyzujeme s cílem porozumět. Je zapotřebí si uvědomit, že porozumění vzniká až kombinací (a) znalostí kontextu, jak vznikly data a (b) schopností využít statistické grafy a numerické výpočty. Například samotné údaje z výzkumné studie o pohybové aktivitě běžné populace znamenají jen velmi málo, pokud nevíme, s jakým cílem se studie provedla a jak měření různých motorických parametrů přispívají k tomuto cíli. Na druhé straně měření mnoha stovek jedinců (statistických jednotek) má malou výpovědní hodnotu i pro experta oboru kinantropologie, pokud data nejsou pomocí statistických prostředků upravena, zobrazena transformována do několika vhodně zvolených numerických charakteristik – popisných charakteristik.

Každá množina naměřených dat obsahuje informace o skupině objektů v ní obsažených. Účelem analýzy dat je přehledně zpřístupnit data tak, aby byla dobře „čitelná“. Jde tedy o jistou kumulaci, zhuštění informací do grafů, tabulek a vypočtených statistických charakteristik. Až takto upravené informace popisující určitý soubor dokážeme porovnávat (muži x ženy, plavci x neplavci, apod.)

V následných kapitolách se seznámíme s metodami a postupy, pomoci kterých lze datům „porozumět“.

Statistické třídění dat

Výsledkem statistického šetření jsou neuspořádané, neroztříděné a nepřehledné statistické data. Chceme-li získat podrobnější informace, je třeba údaje uspořádat. Tato činnost se nazývá statistické zpracování (třídění) dat.

Nejjednodušší způsob zpracování dat je tzv. tabulka rozdělení četností.

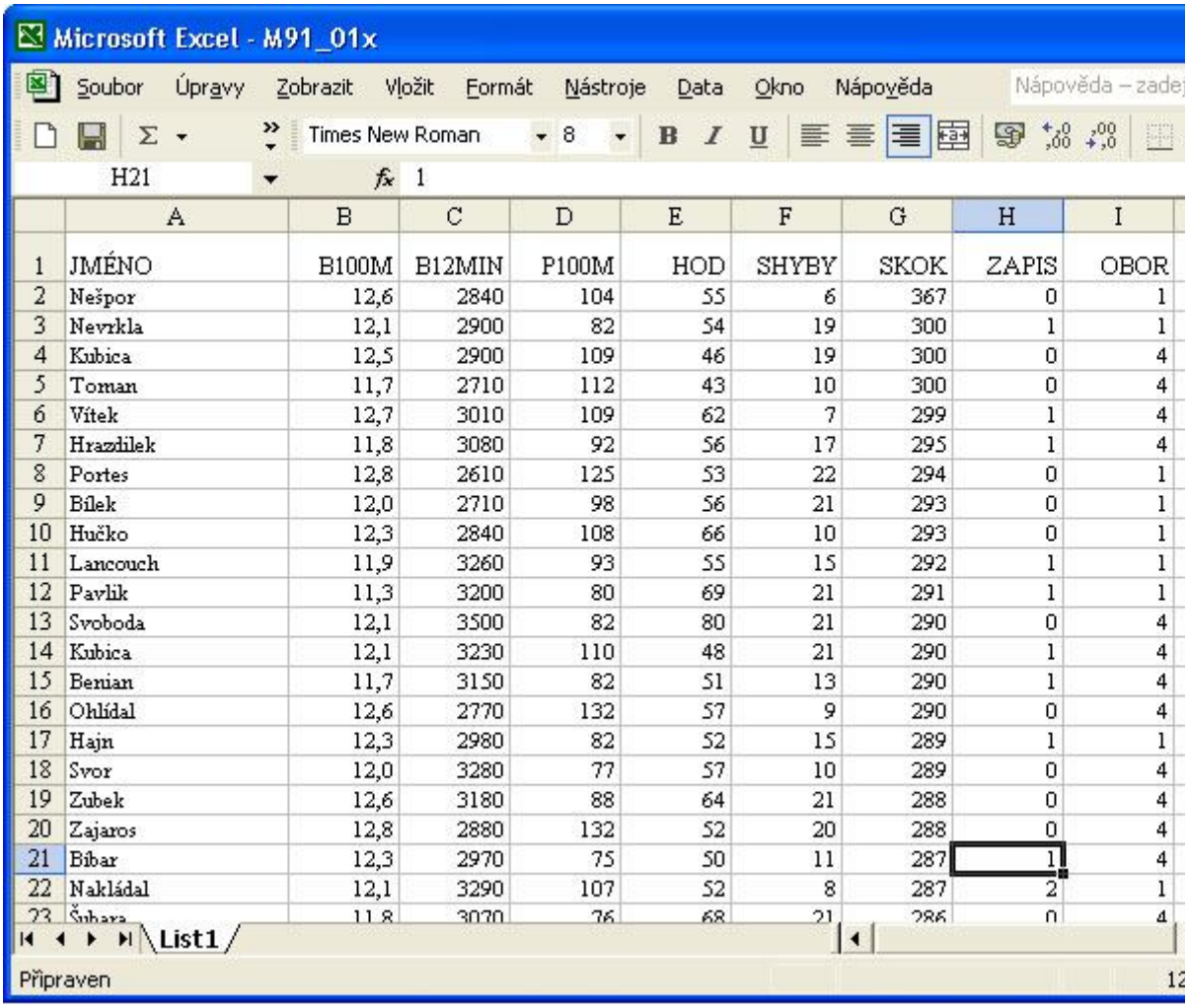
Organizace dat

Data shromažďujeme v různých kontextech a zároveň různými způsoby. Pro potřebu dalšího zpracování a práce s daty je zapotřebí datům dát nejprve takovou podobu, která umožňuje například počítačové zpracování. Jen stěží si lze představit přímé počítačové zpracování dat z papírových formulářů, kterých jsme při výzkumu v terénu použili. Nashromážděná data proto převádíme do tabulky dat neboli datové matice*. Ta má takovou podobu, že jednotlivé sloupce tabulky představují proměnné (variables) – znak, a jednotlivé řádky představují údaje u sledovaných objektů, jednotlivých měření, případů atd. (cases). Obrázek 4.1 zobrazuje příklad tabulky dat (matice dat) v prostředí programu MS Excel, který je možné pro základní statistické výpočty využívat. Zde jsou názvy proměnných označeny v prvním řádku - „Jméno“, „B100M“, „B12MIN“ atd., každé proměnné zde odpovídá právě jeden sloupec – například pro proměnnou „SKOK“ je to sloupec označený jako „G“. V řádcích 2 a dále jsou uvedeny příslušné „případy (cases)“, které speciálně zde představují dosažené hodnoty výkonů v uvedených disciplínách. Například řádek 6 uvádí výkony osoby nazvané „Vítek“.

Údaje jsou v tabulkách dat často kódovány. To znamená, že jednotlivým úrovním jsou u proměnných dle stanovených pravidel přiřazeny číselné kódy. Např. muž=0, žena=1; nebo červená=1, bílá=2, černá=2 apod.

* V praxi často analýzy dat využívají různé formy databází a jejich kombinace. Specifickými postupy jsou tabulky dat z těchto databází generovány.

Obrázek 4.1 Příklad tabulky dat v prostředí programu MS Excel



Microsoft Excel - M91_01x

Soubor Úpravy Zobrazit Vložit Formát Nástroje Data Okno Nápověda

Times New Roman 8

H21 1

	A	B	C	D	E	F	G	H	I
1	JMÉNO	B100M	B12MIN	P100M	HOD	SHYBY	SKOK	ZAPIS	OBOR
2	Nešpor	12,6	2840	104	55	6	367	0	1
3	Nevrkla	12,1	2900	82	54	19	300	1	1
4	Kubica	12,5	2900	109	46	19	300	0	4
5	Toman	11,7	2710	112	43	10	300	0	4
6	Vítek	12,7	3010	109	62	7	299	1	4
7	Hrazdilek	11,8	3080	92	56	17	295	1	4
8	Portes	12,8	2610	125	53	22	294	0	1
9	Bílek	12,0	2710	98	56	21	293	0	1
10	Hučko	12,3	2840	108	66	10	293	0	1
11	Lancouch	11,9	3260	93	55	15	292	1	1
12	Pavlik	11,3	3200	80	69	21	291	1	1
13	Svoboda	12,1	3500	82	80	21	290	0	4
14	Kubica	12,1	3230	110	48	21	290	1	4
15	Benian	11,7	3150	82	51	13	290	1	4
16	Ohlídal	12,6	2770	132	57	9	290	0	4
17	Hajn	12,3	2980	82	52	15	289	1	1
18	Svor	12,0	3280	77	57	10	289	0	4
19	Zubek	12,6	3180	88	64	21	288	0	4
20	Zajaros	12,8	2880	132	52	20	288	0	4
21	Bíbar	12,3	2970	75	50	11	287	1	4
22	Nakládal	12,1	3290	107	52	8	287	2	1
23	Šubara	11,8	3070	76	68	21	286	0	4

List1

Připraven

Jednorozměrné rozdělení četností

Zajímá-li nás pouze jedna vlastnost statistického souboru charakterizovaná jedním statistickým znakem, hovoříme o *jednorozměrném statistickém souboru* a o jednorozměrném rozdělení (rozložení) četností.

Konstrukce tabulky jednorozměrného bodového rozdělení četností je postup vhodný pro nespojité (diskrétní) kvantitativní statistické znaky (např.počet dětí v rodině, počet zásahů do koše, počet kliků atd.), případně pro spojité statistické znaky s malým počtem výskytu tzn. pro statistické soubory s malým rozsahem. Pro spojité kvantitativní statistické znaky (např.výsledky měření běhu na 100 m, tělesné výšky, skoku dalekého), popřípadě pro nespojité (diskrétní) statistické znaky s velkým počtem výskytů užíváme *jednorozměrné intervalové* (skupinové) *rozdělení četností*.

Princip konstrukce tabulky rozdělení četnosti spočívá v záznamu počtu výskytů jevů určité třídy. Ta je u diskrétních statistických znaků představována samotnými znaky, u spojitých znaků intervaly o stanovené šířce. Důležité je mít na paměti, že tabulka musí zahrnovat (pokrýt) celé spektrum znaků, jejichž výskyt je teoreticky možný mezi minimem a maximem a nikoliv pouze znaky (resp. intervaly), které se ve sledovaném souboru vyskytují. Při sestavování tabulky u spojitých znaků se řídíme několika pravidly, které nám napomáhají stanovit počet a šíři těchto intervalů (viz tabulka X). Vypočtené hodnoty však nejsou zavazující, ale pouze orientační a pomocné. Konečný počet intervalů nakonec stanovujeme

vlastní úvahou dle okolností. Při konstrukci zjišťujeme *absolutní, relativní, absolutní kumulativní a relativní kumulativní četnosti znaku* (blíže viz příklad).

Tabulka 4.1 Doporučená pravidla pro určení šířky a počtu intervalů u spojitých statistických znaků

Variační rozpětí	*R*	$R = x_{max} - x_{min}$
Šířka intervalu	*h*	$h = 0,08 . R$
Počet intervalů	*k*	$k = \sqrt{n}$ $k \leq 5 . \log n$ $k \approx 1 + 3.3 \log n$ (Sturgessovo pravidlo) Je-li $n < 30 \Rightarrow$ doporučuje se vytvořit ne více než 6 intervalů Je-li $30 < n < 100 \Rightarrow$ doporučuje se vytvořit 7 až 10 intervalů
Pozor! Intervaly musí být vytvořeny tak, aby jeden statistický znak nemohl být současně zařazen do dvou různých intervalů!!! Intervaly na sebe musejí navazovat!!!		

Příklad 4.1

Při dvakrát opakovaném testování střelby na koš byly u deseti osob (n=10) zjištěny následující výsledky (je zaznamenán počet úspěchů z deseti pokusů při 1. resp. 2. testování).

Úkol: Pro znaky xi sestavte tabulku rozdělení četností (frekvenční).

Tabulka 4.2 Tabulka dat

	x_i	y_i
Osoba 1	9	4
Osoba 2	6	8
Osoba 3	6	6
Osoba 4	8	8
Osoba 5	9	7
Osoba 6	8	8
Osoba 7	8	7
Osoba 8	8	4
Osoba 9	9	8
Osoba 10	10	10

Jedná se o znaky nespojité (diskrétní) kvantitativní, a proto budeme v tomto případě sestavovat Tabulku jednorozměrného bodového rozdělení četností pro proměnnou xi. Minimem v tomto souboru je znak o hodnotě 6 a maximum je 10. V tabulce tedy budou uvedeny všechny úrovně znaku od 6 do 10 (všimněte si, že ačkoliv se hodnota 7 v hrubých skóre vůbec nevyskytuje, přesto je v tabulce rozložení četností uvedena!)

Tabulka 4.3 Tabulka jednorozměrného rozdělení četností (frekvenční)

znak	Čárkovací metoda	n_i	f_i	N_i	F_i
6	//	2	0,2	2	0,2
7		0	0	2	0,2
8	////	4	0,4	6	0,6
9	///	3	0,3	9	0,9
10	/	1	0,1	10	1,0
Σ		10	1,0	-	-

Vysvětlivky:

n ... rozsah souboru

xi ... hodnota znaku

ni ... absolutní četnost

fi ... relativní četnost $f_i = \frac{n_i}{n}$

Ni ... absolutní kumulativní četnost (je vypočtena jako součet příslušné absolutní četnosti a

četností jí předcházejících)
Fi ... relativní kumulativní četnost

Všimněte si, že v tabulce je v posledním řádku suma (součet) u absolutní četnosti rovna celkovému počtu výskytů jevu (zde počet osob), relativní četnost je rovna 1, a že v předposledním řádku je absolutní kumulativní četnost taktéž rovna počtu osob a relativní kumulativní četnost rovna 1. Takto si lze ověřit, zda zjištěné četnosti u jednotlivých tříd jsou dobře stanoveny.

Příklad 4.2

Pro znaky yi sestavte tabulku skupinového (intervalového) rozdělení četností.

Pomocné výpočty pro určení šířky (h) a počtu intervalů (k):
R = 6 h = 0.08 x 6 = 0,48 ??1 (pokus)
k = 3.16 k ? 5 k ? 4.3 (log 10 = 1)

Tabulka 4.4 Určení šířky a počtu intervalů u spojitých statistických znaků

??Volíme šířku intervalu: 1
??Volíme počet intervalů: 4

Tabulka 4.5 Tabulka intervalového rozložení četností

Třída	Interval	Střed	n_i	f_i	N_i	F_i
1	4 – 5	4,5	2	0,2	2	0,2
2	6 – 7	6,5	3	0,3	5	0,5
3	8 – 9	8,5	4	0,4	9	0,9
4	10 – 11	10,5	1	0,1	**10**	**1,0**
Σ			**10**	**1,0**	-	-

Grafické znázornění rozdělení četností

Dalším možným způsobem rozdělení četností je grafické zobrazení. Jeho výhodou je značná přehlednost a umožňuje rychlou souhrnnou představu o četnosti výskytu sledovaného jevu. Prezentace dat číselně (tabulkou) či graficky má své pozitiva i negativa. To se týká především odlišné přehlednosti, exaktnosti, možnosti „čitelnosti údajů o jevu“ apod. Rozhodnutí, zda zobrazit data, resp. jejich četnost, graficky nebo číselně je proto vždy věcí citu. Nelze proto dopředu stanovit pravidla pro to, kdy kterou formu použít. V případě grafického znázornění rozdělení četnosti užíváme zpravidla 5 odlišných forem grafů.

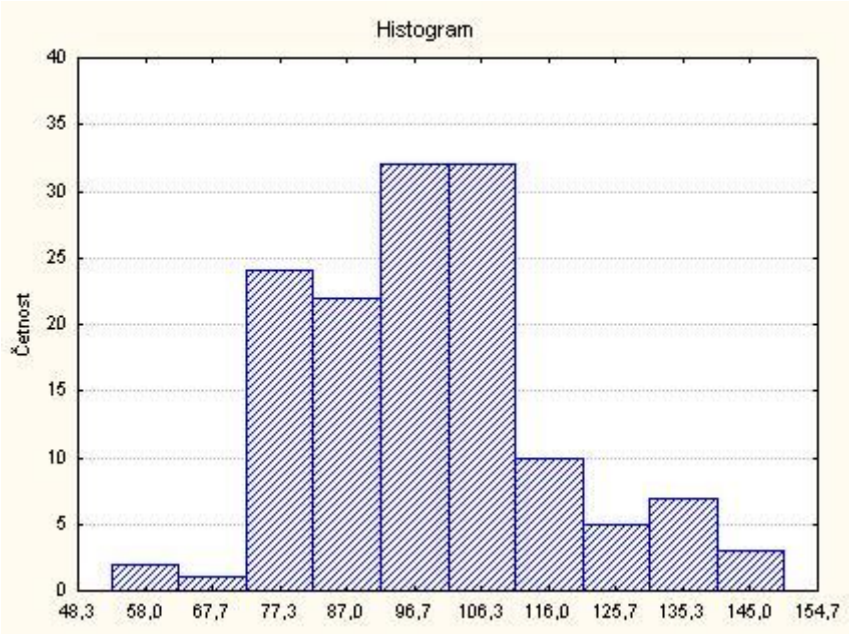
a. **Histogram četností** (sloupkový diagram, sloupcový graf). Histogram je jedna z nejčastěji užívaných forem grafického znázornění rozdělení četností. Je tvořen sloupci;jejichž šířka odpovídá šířce třídního intervalu; výška sloupce odpovídá absolutní četnosti sledovaného statistického znaku. Nutno připomenout, že tyto intervaly opět pokrývají celou šíři číselných hodnot mezi minimem a maximem proměnné (znaku). Na číselné ose „x“ jsou zobrazeny intervaly a na ose „y“ četnosti jednotlivých znaků.
b. **Polygon** (frekvenční graf**).** Jedná se o takovou formu grafického znázornění

rozdělení četností, kde jsou četnosti znaků příslušné třídy spojeny lomenou čárou – jde vlastně o spojnici středů intervalů a příslušných četností histogramu. Tato lomená čára je v prostředí statistických programů nahrazována vyhlazenou (plynulou) křivkou.

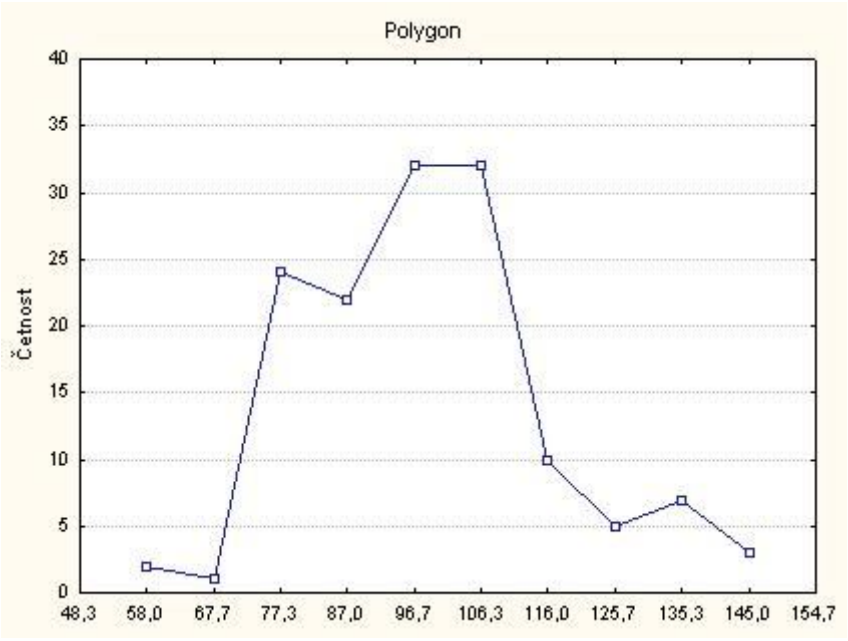
c. **Ogiva** (galtonova**).** Pojem ogival je používán pro lomený oblouk v architektuře, ve statistice tento pojem charakterizuje esovitě lomenou křivku znázorňující kumulativní četnosti (absolutní nebo relativní). Můžeme říct, že jde polygon s tím rozdílem, že namísto absolutní četnosti znaků je na ose „y“ vynášena kumulativní četnost.
d. **Výsečový** (sektorový) **graf.** Jedná se o kruhový graf, vyjadřující relativní četnosti jako charakteristiku struktury daného souboru (nejčastěji v %). Velikost jednotlivých výsečí kruhu znázorňuje příslušné relativní četnosti – procentuální zastoupení ploch jednotlivých tříd.
e. **Piktogram.** Je to obrázkový graf užívaný spíše pro laickou veřejnost.Vyjadřuje absolutní četnosti bez nároků na přesnost, má spíše informativní charakter a používá obrazových symbolů (např.lokomotiva, váček s penězi, postava vojáka). Úroveň absolutní četnosti je vyjádřena velikostí zvoleného symbolu.

(Blíže viz obrázek).

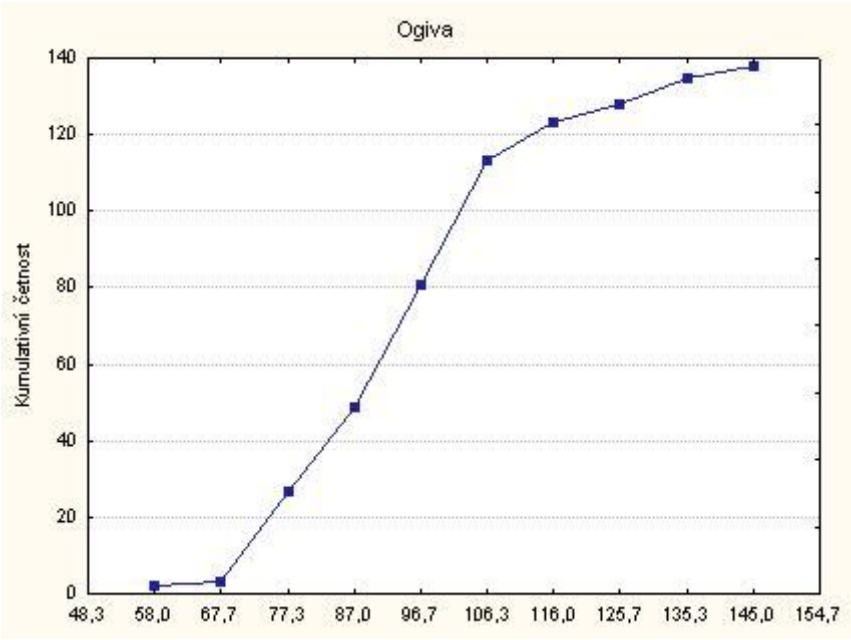
Obrázek 4.2 Histogram



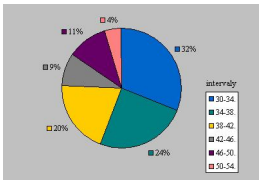
Obrázek 4.3 Polygon



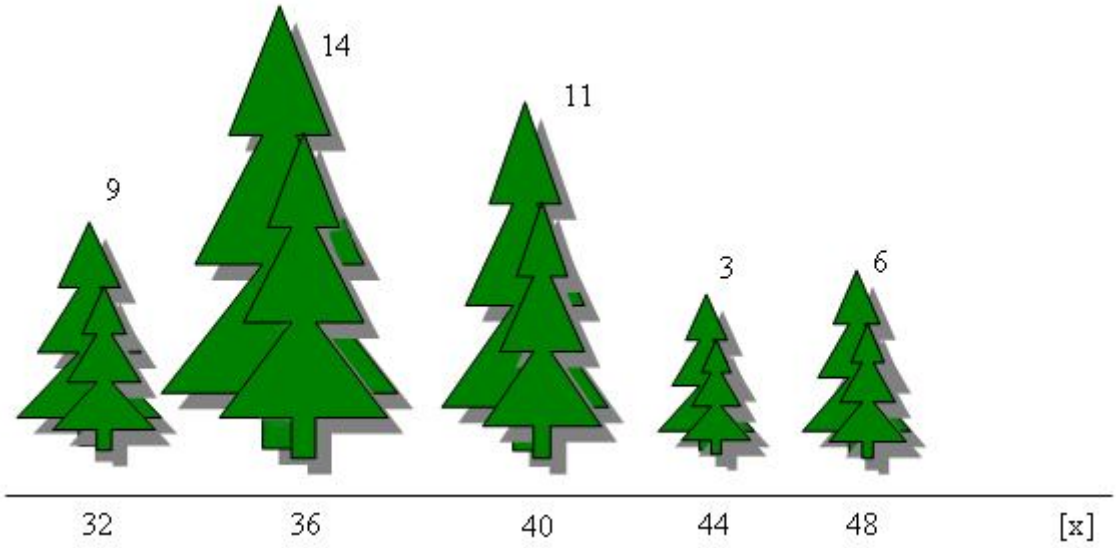
Obrázek 4.4 Ogiva



Obrázek 4.5 Výsečový graf



Obrázek 4.6 Piktogram



Příklad 4.3

Na podkladě tabulky intervalového rozložení četností sestavte v prostředí programu MS Excel histogram.

PDF2_Graf četností

Popis:

Na str.1 je v prostředí MS Excel zobrazen základní soubor o rozsahu N = 350. V následných několika krocích si ukážeme sestavení histogramu četností.

Chceme-li sestavit histogram, je nezbytné znát hodnotu absolutních četností v jednotlivých intervalech. Za tímto účelem je na str.1 rovněž sestavena tabulka intervalového rozdělení četností.

Jak víme, histogram je sloupcovým grafem, kde výška sloupce zobrazuje absolutní četnost znaků v daném intervalu. Naším prvním krokem tedy bude vložení grafu. Na str.2 vidíme, že graf se vytváří prostřednictvím položky „Vložit“ v základní nabídce a zde výběrem položky „Graf“. Potvrzením této položky se zobrazí okno s názvem „Průvodce grafem“ (str.3). Zde si volíme typ grafu – v našem případě volíme „Sloupcový“, chtěli-li bychom však sestavovat polygon nebo ogivu, zde bychom volili „Spojnicový“. Přejdeme-li v průvodci dále, nově zobrazené okno od nás vyžaduje zadat data, na jejichž podkladě má být graf sestaven (str.4). Pomocí myši označíme do bloku oblast E5 až E13, kde se nacházejí data s absolutními četnostmi znaku (str.5). Následně si všimneme ve stejném okně druhé záložky nazvané „Řada“ (str.6). Takto přejdeme do prostředí, kde lze volit další důležité charakteristiky grafu (str.7). My si všimneme položek „Název“ řady a „Popisky osy X (kategorie)“ – umístíme do řádku kurzor a myší označíme oblast, ve které jsou názvy intervalů (D5:D13) – tím se nám již na ose X nebudou zobrazovat kategorie jako 1, 2, 3, … , ale budou zde zobrazeny jednotlivé intervaly, ve kterých budou zobrazeny jednotlivé četnosti. Přejdeme na další okno (str.8). Zde pouze doplníme názvy „Grafu“ (změníme na Histogram), „Osy X“ a „Osy Y“. Přechodem dále (str.9) se dostaneme k poslednímu oknu průvodce grafem, kde volíme, zda se graf umístí přímo do zobrazeného sešitu, nebo bude vytvořen samostatný list, na kterém bude graf vyobrazen. My zvolíme druhou variantu – tedy „Jako oblast do listu“, abychom měli neustále možnost zároveň nahlížet do tabulky dat bez nutnosti přecházet mezi listy. Na str.10 vidíme, že vytvořený graf sice je histogramem, ovšem ne příliš přehledným. Je proto nutné jej dále poupravit. Chceme-li jakoukoliv část upravit, je to možné tak, že na tuto oblast poklepeme a dostaneme tak nabídku, kde lze měnit vlastnosti této části (např. velikost písma kategorií osy, velikost zobrazované oblasti grafu, vzdálenosti mezi sloupci, barva podkladu

aj.). Postupnou úpravou lze např. získat graf zobrazený na poslední str.11.

Budeme-li sestavovat polygon, ogivu nebo výsečový graf, postupujeme podobně, jen volíme jiný typ grafu a rovněž je nutné si uvědomit, co budou zdrojová data – kdy to bude absolutní četnost, kdy relativní a kdy kumulativní četnost.

PŘÍKLAD UKÁZKY PŘÍKLADU V MS EXCEL

Míry polohy

Jsou také označovány jako míry centrální tendence; charakterizují úroveň statistického souboru z hlediska polohy jeho střední hodnoty. Míry polohy svým způsobem zevšeobecňují, zastupují, reprezentují jednotlivé hodnoty sledovaného statistického znaku. Umožňují jednoduché srovnání polohy dvou či více rozdělení četností, srovnání úrovně dvou či více souborů.

Nejčastěji používané míry polohy jsou:

MODUS (Mo)…takto se označuje nejčastěji se vyskytující hodnota statistického souboru (hodnota s největší četností). Je to nejsnáze zjistitelná míra polohy. Soubor může mít jeden či více modů (bimodální, trimodální soubor). Modus je použitelný pro nominální stupnice (a všechny vyšší).

MEDIÁN (Me). Označujeme jím prostřední člen variační řady. Rozděluje řadu na dvě stejně veliké části co do počtu prvků tak, že hodnoty znaku v jedné části jsou menší než medián (nebo jsou mu nanejvýše rovny), ve druhé jsou pak větší (nebo jsou mu nanejvýše rovny). V případě lichého počtu hodnot je mediánem přímo hodnota prostředního členu řady – tedy vždy konkrétní hodnota znaku. Skládá- li se ovšem řada ze sudého počtu, pak nelze stanovit konkrétní prvek, který by rozděloval řadu na dvě stejné části co do počtu prvků. V tomto případě považujeme za medián libovolnou hodnotu ležící mezi dvou prostředních členů řady. Zpravidla se tu za medián pokládá poloviční součet hodnot dvou prostředních členů řady. (Viz obrázek 3.1)

ARITMETICKÝ PRŮMĚR ($\bar{x}$). Lze říct, že aritmetický je nejpoužívanější mírou polohy u metrických dat (není možné jej stanovit u nominálních a ordinálních dat!). Je definován jako součet všech hodnot statistického souboru, resp. proměnné, dělený rozsahem souboru (n). Jeho vlastností je, že je dobrým odhadem hodnot souboru, což znamená, že součet odchylek (rozdílů) každé z hodnot od aritmetického průměru je minimální. V případě, že bychom dosadili jakoukoliv jinou hodnotu a vypočetli součet odchylek, pak vždy obdržíme větší číslo. Tuto vlastnost je možné rovněž interpretovat tak, že nahrazujeme-li hodnoty souboru aritmetickým průměrem, pak se dopouštíme nejmenší chyby.

Aritmetický průměr prostý (jednoduchý)

$\bar{x} = \frac{\sum x_i}{n}$ xi - statistický znak

n = rozsah souboru

Vážený aritmetický průměr

Jeho výpočet vychází z rozdělení četností, užívá se u početnějších souborů; vážený se nazývá proto, že jednotlivým hodnotám znaku je přisuzována váha odpovídající počtu výskytů.

$$\bar{x} = \frac{\sum x_i n_i}{\sum n_i}$$ xi - statistický znak

ni= rozsah jednotlivých podsouborů souboru

Poznámky

V tzv.Gaussově normálním (symetrickém) rozdělením četností jsou hodnoty aritmetického průměru, modu a mediánu stejně velké.

Čím více se od sebe tyto hodnoty liší, tím více je rozdělení četností asymetrické (mluvíme o porušené normalitě rozložení četností).

Další míry polohy jsou např. aritmetický střed, harmonický průměr, geometrický průměr, kvadratický průměr.

Příklad 4.4

Vypočítejte modus, medián a aritmetický (a vážený aritmetický) průměr pro níže zadanou matici dat.

	proměnná x_i
případ 1	7
případ 2	6
případ 3	7
případ 4	8
případ 5	9
případ 6	8
případ 7	8
případ 8	8
případ 9	9
případ 10	10

Variační řada znaku xi: 6, 7, 7, 8, 8, 8, 8, 9, 9, 10

Modus Mo = 8

Median Me = 8

Aritmetický průměr

$$\bar{x} = \frac{6+7+7+\ldots+9+10}{10} = \frac{80}{10} = 8$$

Vážený AP

$$\bar{x} = \frac{6+2\cdot 7+4\cdot 8+2\cdot 9+10}{10} = \frac{80}{10} = 8$$

Příklad 4.5

Vypočítejte míry polohy (minimum, maximum, modus, medián, aritmetický průměr) v prostředí programu MS EXCEL

Míry polohy

Popis:

V přiloženém příkladu si ukážeme výpočet základních popisných charakteristik – měr polohy. Jako příklad zde byly použity výsledky z talentové přijímací zkoušky na FTK UP, které se zúčastnilo 332 osob (rozsah souboru N = 332), uvedeny jsou pouze testové disciplíny běh 100 m, běh 1500 m, plavání 100m a hod kriketovým míčkem.

Na první straně je zobrazen zápis zjištěných hodnot ve všech čtyřech disciplínách v prostředí programu MS Excel a připravena tabulka pro stanovené hodnoty základních popisných charakteristik – měr polohy. Jedná se tedy o stanovení charakteristik Maximum, Minimum, Modus, Medián a Aritmetický průměr.

Při výpočtu těchto hodnot využijeme funkcí, jež prostředí MS Excel nabízí. Volbu požadované funkce lze vyvolat kliknutím na tlačítko fx (Vložit funkci), jak je znázorněno na str. 2. Tím vyvoláme okno, které obsahuje seznam všech dostupných funkcí. Tyto funkce jsou různého typu – matematické, statistické, finanční, databáze, vyhledávací aj. Zde je možné popisem funkce v horní části (str.3) – např. heslem „minimum“ – vyhledat požadovanou funkci. Pomocí tlačítka „Přejít“ získáváme nabídku funkcí, které svou povahou odpovídají zadanému heslu (str. 4). Je nutné vybrat pouze jednu správnou. Klikneme-li jednou na jakoukoliv funkci, v dolní části okna se nám zobrazí její popis. Pro naší potřebu stanovení minima ze základního souboru je nejvíce vyhovující funkce „MIN“. Přejdeme dále přes OK. Nyní se dostáváme do okna nazvaného „Argumenty funkce“ – jedná informace, které musejí být zadány, aby funkce mohla stanovit požadovanou hodnotu (str. 5). Funkce MIN si vyžaduje zadat pouze jeden argument a tím jsou čísla, nebo oblast buněk, které čísla obsahují, ze kterých má být minimum vypočítáno resp. nalezena nejmenší hodnota. V zobrazeném okně nám je nám automaticky nabídnuta oblast dat, kterou MS Excel očekává. Ovšem ne vždy tato oblast odpovídá naší představě – jako je tomu i v tomto případě. Musíme proto zadat oblast nově. To je možné provézt tažením myši a označením tak zamýšlených dat do bloku (str. 6 a 7). Tím jsme do kolonky „Číslo 1“ zadali oblast dat B2:B333. Kolonku „Číslo 2“ necháváme volnou, neboť nechceme zadávat žádné další data (str.8). Jdeme dále přes OK.Takto jsme vypočetli minimální hodnotu proměnné Běh100M, tj. min = 13,2 (str.9).

Obdobně je možné postupovat také pro další míry polohy. Známe-li však názvy funkcí, které nám tyto míry polohy určují, je možné je přímo zadat do příkazového řádku (řádek vedle tlačítka „Vložit funkci“, které jsme v předchozím využívali. Naše funkce se jmenuje MAX a jako každá funkce má tvar NÁZEVFUNCE(argumenty funkce). Z toho plyne zápis znázorněný na str.10-12. Píšeme „=MAX(“ ,dále tažením myši označíme oblast dat a zápis uzavřeme závorkou. Potvrdíme ENTER a dostáváme hodnotu max = 18,7 (str.13). Podobně postupujeme pro Modus, Medián a Aritmetický průměr (str.13-17).

Dalším krokem bude vypočítat hodnoty všech pěti popisných charakteristik také pro testy Běh1500M, Plavání100M a Hod. Bylo by samozřejmě možné postupně tyto hodnoty dopočítat postupem, který jsme doposud užili. Využijeme však možnosti, kterou nám prostředí MS Excel nabízí, a to je „kopírování vzorce tažením“.

Jak je vidět na str.18, označíme všech pět dosazených vzorců pro proměnnou Běh 100M do bloku, kurzor myši přesuneme do těsné blízkosti pravého dolního rohu označeného bloku, až se kurzor myši promění v zobrazený křížek (jak ukazuje zelená šipka). Přidržením a tahem levého tlačítka myši označujeme oblast, kam chceme vzorec doplnit (str.19). Pustíme-li tlačítko myši, všech 15 hodnot se automaticky dopočítá (str.20 a 21).

Zastavme se ještě na chvíli u kopírování vzorce. Tak jako jsme nyní použili kopírování vzorce, jak program zjistí, se kterými hodnotami počítat (které hodnoty mají být argumenty funkce)? Uvažujme, že budeme mít zadánu funkci následovně „FUNKCE(A1:A20)“. Překopírujeme-li vzorec o jednu buňku doprava, bude do argumentu automaticky dosazena oblast dat o jeden sloupec napravo od původního. Tedy nový vzorec bude mít podobu „FUNKCE(B1:B20). Budeme-li kopírovat z původní pozice o jednu dolů, bude mít vzorec tvar „FUNKCE(A2:A21)“. Je tedy nutné si vždy ujasnit, zda skutečně kopírováním vzorce počítám s daty (s oblastí dat), se kterými chci počítat.

KE STAŽENÍ PŘÍKLAD UKÁZKY PŘÍKLADU V MS EXCEL

Míry variability

Popis statistického souboru pomocí měr polohy (středních hodnot) není dostačující. Především z toho důvodu, že nevypovídají o tom, do jaké míry jsou „dobrou náhradou“

všech hodnot souboru. Můžeme například uvažovat dva soubory dat, které mají stejný řekněme aritmetický průměr a jsou jím tedy „zastoupeny“, ovšem vzdálenosti dat prvního souboru od aritmetického průměru mohou být znatelně větší než u druhého souboru. Pro lepší, hlubší informovanost o vlastnostech souboru používáme tzv. *míry rozptýlení*. V literatuře jsou též označovány jako *míry variace, míry měnlivosti*.

Míry variability charakterizují vyrovnanost jednotek souboru, ukazují, jak jsou hodnoty souboru rozptýleny, jak se jednotlivé hodnoty znaku vzájemně liší.

Umožňují posoudit, do jaké míry je sledovaný soubor homogenní (stejnorodý) resp. heterogenní (nestejnorodý, různorodý).

Už pouhé seřazení dat podle velikosti, známé jako *variační řada*, je vlastně způsobem vyjádření rozptýlenosti dat. Rovněž výpočet *variačního rozpětí (R)* definovaného jako diference maximální a minimální hodnoty sledovaného znaku souboru; tedy R = xmax – xmin. Nevýhodou tohoto vyjádření je však velká citlivost vůči odlehlým (extrémním hodnotám), které nemusí dobře charakterizovat zbytek dat a mohou být pouze „výjimečné“.

Míry variability založené na kvantilech

V literatuře se lze rovněž setkat s názvem *Kvantilové míry variability*. **Kvantil** lze charakterizovat jako hodnotu variační řady, pod níž se nachází definované množství dat. Každý kvantil má vždy přiřazenu tzv. hladinu *q*, a je označován znakem *xq*. Číselný parametr q je hodnota intervalu 0 < q < 1, přičemž je jí vyjádřený relativní podíl údajů, které se nacházejí pod kvantilem. Výpočet kvantilu na dané hladině se děje tímto způsobem: Definujme $j = [qn]$, kde hranaté závorky značí zaokrouhlení čísla “*qn*” na nejbližší menší celé číslo a kde n udává počet dat souboru. V případě, že $qn = [qn]$, pak $xq = (xj + xj+1)/2$; v případě, že neplatí $qn = [qn]$, pak $xq = xj+1$, *kde* xj $(j = 1, ...,n)$ jsou data výběru seřazena podle velikosti (variační řada). Například máme-li vypočítat kvantil *x0,1* pro modelová data {5, 8, 7, 2, 10, 50, 8, 1, 0, 0, 4, 5, 9, 7, 6, 8, 4, 3, 1, 1}, pak $j = qn = 0{,}1.20 = 2$. Tedy platí $qn = [qn] \Rightarrow x0{,}1 = (x2 + x2+1)/2 = (0+1)/2 = 0{,}5$.

V praxi se setkáváme s celou řadou kvantilů, v závislosti na účelu jejich výpočtu. Někdy jsou hladiny q vyjádřeny v procentech – v tomto případě nalezené hodnoty nazýváme percentily. Často počítané a významné jsou percentily s hladinou 25%, 50% a 75%, označované jako **kvartily**. *Q1* je první kvartil neboli dolní kvartil ($q = 25\%$); *Q2* je druhý kvartil neboli medián ($q = 50\%$); *Q3* je třetí kvartil neboli horní kvartil ($q = 75\%$). Percentily s hodnotami 10%, 20%, …, 90% se nazývají **decily** a značí se *x10, x20 , ..., x90* .

Výše byla zmiňována míra variability variační rozpětí. Alternativou, která již není tak citlivá k odlehlým hodnotám, je výpočet **kvartilového** (*x75 - x25*), **decilového** (*x90 - x10*) a **percentilového rozpětí** (*x99 - x1*).

Momentové míry variability

Dosud uvedené míry variability udávají jen rozpětí, v němž se znaky pohybují. Nejvhodnější mírou variability je však taková charakteristika, v sobě zahrnuje současně všechny hodnoty (je ovlivněna všemi hodnotami) a která udává jak variaci ve smyslu vzájemné odlišnosti jednotlivých hodnot znaku, tak variaci ve smyslu odlišnosti jednotlivých hodnot znaku od průměru. Za tímto účelem jsou používány charakteristiky rozptyl a směrodatná odchylka, které mají mezi sebou velice úzký vztah, a variační koeficient.

Rozptyl. Je průměrnou kvadratickou odchylkou jednotlivých hodnot od aritmetického průměru, přičemž při průměrování této odchylky dělíme číslem (n-1). Pouze v případě, že počítáme rozptyl pro všechny prvky populace, pak dělíme přímo číslem n. Rozptyl „měří“

variaci ve smyslu odlišnosti jednotlivých hodnot znaku od průměru i ve smyslu vzájemné odlišnosti jednotlivých hodnot znaku.

Rozptyl pro výběrový soubor

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

Rozptyl pro základní soubor

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

Směrodatná (standardní) odchylka. Variance vyjádřená rozptylem je uvedena v druhé mocnině měrné jednotky, což je náročné pro interpretaci. Proto bývá výhodnější používat směrodatnou odchylku, která je počítána jako odmocnina z rozptylu.

$V_y = \frac{1,8}{7} = 0,26$**Variační koeficient**. Někdy je zapotřebí porovnat rozptyly, které jsou ovšem měřeny v rozlišných jednotkách, a proto vypočtené hodnoty rozptylů nelze stavět vedle sebe a porovnávat je. Za tímto účelem zavádíme charakteristiku Variační koeficient, který umožňuje provést srovnání variability dvou či více souborů nezávisle na jednotkách měření. Udává poměr směrodatné odchylky k aritmetickému průměru, což vyjadřuje % aritmetického průměru je tvořeno směrodatnou odchylkou. Vysoká hodnota variačního koeficientu poukazuje na fakt, že aritmetický průměr „špatně“ zastupuje data souboru, resp. není vhodným odhadem úrovně dat.

$v = \frac{s}{\bar{x}}$resp. $V = 100 \cdot \frac{s}{\bar{x}}$(vyjádřeno v %)

Příklad 4.6

Vypočítejte rozptyl, směrodatnou odchylku a variační koeficient pro níže uvedenou matici dat a porovnejte rozptyly proměnných xi a yi.

	proměnná x_i [jednotka cm]	proměnná y_i [jednotka kg]
případ1	7	4
případ2	6	8
případ3	7	6
případ4	8	8
případ5	9	7
případ6	8	8
případ7	8	7
případ8	8	4
případ9	9	8
případ10	10	10

Výpočet aritmetického průměru pro proměnnou xi jsme již provedli v příkladu 3.3 (viz výše). Pro výpočet rozptylu proměnné yi je nutné nejprve vypočítat její aritm. průměr.

aritmetický průměr proměnné yi:

$$\bar{y} = \frac{4+8+6+8+7+8+7+4+8+10}{10} = \frac{70}{10} = 7$$

rozptyl proměnné xi:

$s_x^2 = \frac{(7-8)^2 + (6-8)^2 + (7-8)^2 + \ldots + (10-8)^2}{10} = \frac{1+4+1+0+1+0+0+0+1+4}{10} = 1,2$

směrodatná odchylka proměnné xi:

$s_x = \sqrt{1,2} = 1,1$

variační koeficient proměnné xi:

$V_x = \frac{1,1}{8} = 0,14$

rozptyl proměnné yi:

$$s_y^2 = \frac{(4-7)^2 + (8-7)^2 + (6-7)^2 + \ldots + (10-7)^2}{10} = \frac{9+1+1+1+0+1+0+9+1+9}{10} = 3{,}2$$

směrodatná odchylka proměnné yi:

$$s_y = \sqrt{3{,}2} = 1{,}8$$

variační koeficient proměnné xi:

$Vx = 0{,}14$ (resp. 14 %) < $Vy = 0{,}26$ (resp. 26 %)

Z výpočtu vyplývá, že rozptyl proměnné *xi* s měrnou jednotkou centimetry je menší, než rozptyl proměnné *yi* s měrnou jednotkou kilogram.

Příklad 4.7

Výpočet měr variability (kvantily, rozptyl, směrodatná odchylka, variační koeficient) v prostředí programu MS EXCEL

PDF4_Míry variability

Popis:

Tak jako v případě měr polohy, i zde vycházíme z příkladu talentových přijímacích zkoušek na fakultu tělesné kultury, které se zúčastnilo 332 osob a u kterých byly zaznamenány čtyři testy – Běh100M, Běh1500M, Plavání100M a Hod kriketovým míčkem (str.1). V pravé části prvního obrázku je připravena tabulka s pěti mírami variability – Kvartil, Percentil, Rozptyl, Směrodatná odchylka a Variační koeficient. Pro výpočet popisné charakteristiky Kvartil budeme hledat příslušnou funkci. Už známým způsobem postupujeme přes tlačítko fx (Vložit funkci) (str.2). Takto dostáváme nabídku funkcí obsažených v tomto programu. Pro popis funkce použijeme heslo „kvartil“ a zmáčkneme tlačítko „Přejít“ (str.3). Vidíme, že MS Excel nám nabízí pouze jedinou možnost a tou je funkce QUARTIL (str.4). Přečteme-li si ve spodní části charakteristiku této funkce, zjistíme, že je pro naše potřeby vyhovující. Přejdeme dále přes OK. Vidíme, že pro tuto funkci je nutné dosadit dva argumenty (str.5) – „Pole“ a „Kvartil“. Argument „Pole“ již známe z předchozích funkcí – je jím oblast buněk obsahujících data, ze kterých chceme kvartil vypočítat (volíme tedy B2:B333). Jak víme z textu výše, známe tři kvartily – Q1, Q2 a Q3 . To, který kvartil stanovíme, nám umožní argument „Kvartil“, kde dosazujeme hodnoty 1, 2 nebo3. My stanovíme první kvartil, tedy Q1, čemuž odpovídá argument „1“ (str.6). Potvrzením funkce tlačítkem OK dostáváme hodnotu 14,9 (str.7). To znamená, že 1 hodnot je menší nebo rovna hodnotě 14,9 a můžeme tedy říct, že 25% kandidátů dosáhlo v testu běh na 100 metrů čas menší nebo roven 14,9 s.

Analogicky postupujeme pro výpočet Percentilu. Postupně vyhledáme požadovanou funkci (str.8 a 9) a obdržíme okno, kde opět máme dosadit argumenty této funkce (PERCENTIL) (str.10). Všimneme si již pouze argumentu „K“. Jak máme uvedeno v popisu tohoto argumentu, je tímto argumentem hodnota z intervalu mezi 0 a 1. Víme, že Percentil je kvantilem, který dělí variační řadu na 100 stejných částí. Budeme-li tedy hledat dvacátý percentil (tedy hodnotu, která bude dělit variační řadu v poměru 20 ku 80), zadáme u tohoto argumentu hodnotu 0,2. My vypočteme třicátý percentil – proto hodnota 0,3. Potvrdíme-li tuto funkci, obdrželi bychom hodnotu 15, což je již uvedeno ve spodní části okna.

Na straně 11 počítáme další míru variability, kterou je Rozptyl. Použijeme opět funkci (str.11). Při vyhledávání funkce se však dostáváme do situace, kdy není zcela jednoznačné, kterou funkci použít, neboť je zde mnohem bohatší nabídka funkcí, než tomu bylo doposud. Je tedy nutné si u každé z uvedených funkcí přečíst jejich charakteristiku a vybrat správnou

(str.12). V našem případě je touto funkcí funkce „VAR“. Tato má snadný argument a tím je pouze oblast buněk, ve kterých je obsažen základní soubor – zadáváme B2:B333 (str.13). Přejdeme dále a získáváme přibližnou hodnotu 0,95 (str.14).

KE STAŽENÍ PŘÍKLAD UKÁZKY PŘÍKLADU V MS EXCEL

Transformace dat, Standardní (odvozená) skóre

Ve snaze úplného popisu dat se můžeme dostat do situace, která vyžaduje interpretovat jednotlivé (individuální) výsledky. Budeme-li například uvažovat situaci, kdy jsme zjišťovali výkony dvaceti svěřenců v disciplíně plavání 100 m. K dispozici tedy máme 20 údajů. V prvé fázi nám míry polohy (modus, medián, aritm. průměr) a míry variability (směrodatná odchylka, variační koeficient, kvantily, …) umožnily popsat celkovou úroveň výkonnosti skupiny našich svěřenců. My se ovšem ptáme, co tyto hodnoty vypovídají ve vztahu k výkonům jednotlivců? Asi nejjednodušší a obvyklou možností je určit vzdálenost (odchylku) výkonu jedince od aritmetického průměru, resp. jiné míry polohy. Takto individuální výsledky všech 20 svěřenců *transformujeme na odchylky od průměru*. Tyto údaje je již snazší interpretovat – tento údaj lze chápat jako vzdálenost úrovně výkonu od průměrnosti.

S výpočtem odchylek se ovšem často nelze spokojit, neboť ty nám nic nevypovídají o významu velikosti odchylky. Můžeme se ptát, jak významná bude v našem případě plavců odchylka některého ze svěřenců 15 sekund, je-li aritmetický průměr souboru svěřenců roven 72 sekund. Tím se dostáváme k *standardizované transformaci* (standardizaci) *dat*. Tak jako v případě míry variability, kdy jsme počítali variační koeficient za účelem odstranění míry dat (jednotky měření), protože pouze takto byly porovnatelné variability dvou sledovaných proměnných měřených v odlišných jednotkách, bude naší snahou získat nový formát dat, které budou bez měrné jednotky a které bude ve vztahu k variabilitě sledovaného souboru. Je zřejmé, že čím menší bude hodnota směrodatná odchylka (menší variabilita souboru, data jsou méně rozptýlena), tím významnější (markantnější) bude odchylka výkonu jedince od průměru souboru a naopak. Tuto úvahu lze popsat transformačním vzorcem hrubých skóre (naměřených dat) na odvozená skóre, tzv. *z-transformace*:

$$z = \frac{x_i - \bar{x}}{s}$$

Takto odvozená skóre nazýváme **z-skóre** (z-stupnice, z-body), které mají vlastnost, že transformujeme-li takto všechny data souboru, získáme odvozená data, jejichž aritmetický průměr je roven nule a směrodatná odchylka je rovna jedné. Dále platí že většina transformovaných hodnot (96%) se nachází v rozmezí od -3 do 3. Ze z-skóre se vypočítávají dle potřeby další odvozená skóre, jako jsou **T-skóre, staniny, steny, MQ skóre** (viz tabulka 5.2).

Důležitou podmínkou těchto transformací je *normalita rozložení četností* užitých dat, v opačném případě je možné využít **percentilové stupnice**. Jedná se o stupnici založenou na percentilech, se kterými jsme se setkali v kapitole 3.4.1 pojednávající o variabilitě souboru. Takto jsou individuální skóre převedeny na hodnotu percentilu (též procentilu). (Výpočet percentilu lze nalézt v Měkota et al. (1988): Antropomotorika II. [Skriptum].Praha: SPN)

Standardní skóre tedy umožňují posoudit individuální výsledek ve vztahu k průměru a rozptylu uvažovaného souboru. Umožňují přímá inter- a intraindividuální srovnání.

Tabulka 7.8 Charakteristiky odvozených skóre

Příklad výpočtu odvozených skóre:

Výchozí charakteristiky:

$x_i = 6$

Z	=	(6-8)/1,1	=	-1,8	aritm.průměr 0
T	=	50 + 10.(-1,8)	=	32	aritm.průměr 50
Sta	=	5 + 2.(-1,8)	=	1,4	aritm.průměr 5
Ste	=	5,5 + 2.(-1,8)	=	1,9	aritm.průměr 5,5
MQ	=	100 + 15.(-1,8)	=	73	aritm.průměr 100

5 ANALÝZA ZÁVISLOSTI, MÍRY ZÁVISLOSTI

ZÁVISLOST PEVNÁ, VOLNÁ, STATISTICKÁ A KORELAČNÍ

Dosud jsme pracovali jen s jednorozměrnými soubory a předmětem úvah byl popis jeho stavu (podoba) ve smyslu matematické statistiky, k čemuž jsme užívali řadu technik. Ačkoliv to v mnoha případech má své opodstatnění - spokojit se jen s tímto popisem, přesto nutno říct, že jde o pouhou „izolovanou analýzu“. Obecně jsou jevy v přírodě a společnosti mnohem složitější struktury a proto také analýza jevů nekončí pouze jejich popisem. V tomto ohledu si tedy nevšímáme změn jednotlivé proměnné (znaku) ve spektru proměnných, ale zkoumáme rovněž vztahy mezi těmito proměnnými. Sledujeme, do jaké míry a jakým způsobem má vliv a do jaké míry může jedna proměnná ovlivňovat druhou.

V této kapitole se proto zaměříme na hledání, zkoumání a hodnocení *souvislostí* (závislostí) *mezi dvěma* (či více) *statistickými proměnnými* (znaky). Budeme se koncentrovat na metody, kterými můžeme závislosti analyzovat.

V přírodě i ve společnosti se lze setkat s jevy, které jsou zcela věcí náhody (přinejmenším doposud nikdo neprokázal, že by tomu bylo jinak) – tyto označujeme jako **náhodné jevy** neboli jevy nezávislé na libovolné jiné proměnné, resp. na jiném jevu. U těchto jevů není možné nalézt žádné vlivy (statisticky by jsme je chápali jako proměnnou, kterou by bylo možné sledovat, měřit a vyhodnocovat), na podkladě jejichž parametrů by bylo možno usuzovat na výsledek zamýšleného náhodného jevu. Jako klasický příklad lze uvézt hod kostkou – náhodným jevem v tomto případě bude výskyt jedné z šesti možností na horní ploše vržené kostky. Napadnout nahodilost jevu by znamenalo najít takovou proměnnou, dle které bychom dokázali předem usuzovat na výsledek hodu. V takovém případě by existoval mezi těmito dvěmi proměnnými vztah, který by šlo statisticky (resp. matematickým vyjádřením) popsat.

Jak to bude v opačném případě, v případě, že najdeme jev, kdy na podkladě jeho parametrů dokážeme usuzovat na výsledek původně uvažovaného náhodného jevu? Pak již samozřejmě není náš jev zcela nezávislý (tedy náhodný), ale lze zde o určité závislosti hovořit – mluvíme o tzv. **příčinné** (kauzální) **závislosti**. Uvažujme například opět hod kostkou, kdy sledujeme počet teček na horní ploše vržené kostky. Při hře se budeme udivovat nad tím, že se nejčastěji opakuje hodnota tři. Neustále to připisujeme Paní Náhodě, ale s postupem času nám tento výskyt nedá spát a začínáme stále více o Paní Náhodě pochybovat. Po chvilce zamyšlení a spekulací nám to nedá a začneme hledat důvod, proč právě číslo má tři má nejvyšší výskyt (nejvyšší četnost). Využijeme znalostí ze základní školy a ověříme si domněnku, že kostka není homogenní (tedy nemá stejnou hustotu ve všech částech stejnou) a že koncentrace její hmotnosti připadá na protější stěnu stěny s označením tři. Tato domněnka se nakonec potvrdila a tak bylo možné říct, že hod kostkou již nebyl jevem nahodilým, ale že existovala příčinná závislost mezi koncentrací hmotnosti v určité částí kostky a výskytu číselných hodnot na vržené kostce.

Definovat příčinnou (kauzální) závislost lze tedy tak, že se jedná o takovou závislost, kdy změny daného jevu či několika jevů (příčina) nutně vyvolávají za příslušných podmínek určité změny jevu jiného (následek, účinek).

Některé závislosti jsou zcela přirozené a nikoho z nás asi nenapadne nad nimi uvažovat – např. s nejjednodušší formou kauzální závislostí se můžeme setkat u přírodních jevů - např. zahříváním tělesa (elementární příčina) za konstantních podmínek dochází ke zvětšování jeho objemu (elementární účinek).

Příčinná závislost jevů má vzhledem k všeobecné souvislosti jevů všeobecný charakter. Každý jev je příčinou jiných jevů a současně účinkem (následkem) jiných jevů. Jevy, jenž jsou v jedněch souvislostech příčinami, vystupují v jiných souvislostech jako následky. Existuje tedy zřetězení příčin a následků.

Zkoumání těchto příčinných závislostí je jedním z nosných pilířů vědy. Objevení a poznání příčinných závislostí umožňuje člověku proniknout do zákonitostí přírody a společnosti a využít jejich znalostí pro praktickou činnost.

V oblasti tělesné kultury má zkoumání příčinných závislostí své nepostradatelné postavení. Ať už budeme existenci, podobu nebo míru závislosti hledat pouze intuitivním odhadem nebo exaktními statistickými metodami, vždy jej ve sportovní praxi použijeme. Těžko lze něco namítat proti tvrzení, že trenér má snahu využívat takové tréninkové metody, které povedou k nárůstu sportovní výkonnosti – to jinými slovy znamená, že bude užívat tréninkovou metodu, která má těsný vztah (je v příčinné závislosti) k úrovni sportovní výkonnosti. Z hlediska metody zkoumání rozlišujeme *závislosti pevné a volné*.

Pevnou závislostí označujeme případ, kdy výskytu jednoho jevu NUTNĚ ODPOVÍDÁ výskyt druhého jevu. V takovém případě lze říct, že změny hodnot jednoho jevu (jedné proměnné) jednoznačně vyvolávají odpovídající změny hodnot druhého jevu (proměnné). Takto každé hodnotě jedné proměnné odpovídá právě jedna hodnota jiné proměnné. Pro pevnou závislost je charakteristické, že se opakuje ve všech jednotlivých případech. Může být tedy charakterizována jediným pozorováním (větší počet pozorování slouží k ověření výsledků a vyloučení chyb). Setkáváme se s ní při formulování zákonitostí vztahů (zákonů) mezi proměnnými (např. Newtonův či Ohmův zákon atd.). Ačkoliv pevná závislost popisuje velké množství přírodních jevů, již méně si s ní vystačíme v případě společenských jevů. To je dáno skutečností, že u nich se většinou jedná o celé komplexy příčin a účinků. V rámci těchto komplexů působí vedle podstatných okolností také okolnosti, které jsou vzhledem k dané souvislosti náhodné a u nichž nelze v daném okamžiku stanovit ani směr, ani sílu jejich působení (lze do jisté míry hovořit o rušivých proměnných).

Závislost dvou jevů, při níž vznik nebo existence jednoho jevu není nerozlučně spjata se vznikem či existencí jevu druhého, nazýváme **volná závislost**. Volnou závislostí označujeme případ, kdy výskyt jednoho jevu OVLIVŇUJE výskyt druhého jevu. Lze říct, že takto hodnoty jednoho jevu ovlivňují hodnoty druhého jevu pouze částečně a nemusí nutně platit, že by změna první proměnné jednoznačně způsobila odpovídající (očekávanou) změnu druhé proměnné. Jde tedy vztah, kdy každé hodnotě jedné proměnné odpovídají různé hodnoty jiné proměnné.

Chceme-li nějakým způsobem popsat takovou závislost, je vždy potřeba (na rozdíl od pevné závislosti) provézt řadu pozorování. Na jejich podkladě následně popíšeme závislost tím, že definujeme jejich vztah, který ale bude mít pouze potenciální povahu. Důležitou roli při zkoumání volné závislosti hraje volba vhodných statistických proměnných - znaků (tedy takových, které postihují jevy co nejpřesněji) a dostatečný rozsah souboru (u malých souborů se může projevit vliv náhodných a vedlejších činitelů). Pokud se volná závislost týká kvantitativních statistických znaků, bývá označována za statistickou.

Statistická závislost (širší pojem) je pravděpodobnostním vyjádřením skutečnosti, že jev B nabude příslušné hodnoty v případě výskytu dané hodnoty jevu A.

Korelační závislost (užší pojem) udává těsnost vztahu mezi dvěma jevy A a B (viz dále).

V oblasti tělesné kultury se s pevnou závislostí příliš často nesetkáme. Převažují zde spíše jevy podmíněné celými komplexy faktorů – tedy proměnnými, které se na úrovni sledovaného jevu podílejí jen do jisté míry. Libovolný pohybový projev je důsledkem velkého množství vlivů/faktorů, jejichž existence je nám často zcela utajena, ačkoliv se je snažíme objevit a vytvořit tréninkový algoritmus, který by dokázal s těmito faktory operovat.

KORELAČNÍ POČET, REGRESNÍ A KORELAČNÍ ANALÝZA

K poznání a matematickému popisu statistických závislostí slouží metody regresní a korelační analýzy označované souhrnně jako korelační počet.

Hlavní úkoly korelačního počtu:

- Postižení povahy korelační závislosti, umožňující odhady neznámých hodnot závisle* proměnné y při známých hodnotách nezávisle proměnné x (hovoříme o regresi).
- Povaha korelační závislosti je nejčastěji vyjadřována matematickou funkcí - mluvíme o regresní funkci. Tento úkol korelačního počtu je označován jako regresní analýza.
- Měření těsnosti korelační závislosti, aby bylo možno posuzovat přesnost regresních odhadů a míru korelační závislosti (vlastní korelace).
- Je vyjadřována korelačním koeficientem. Tento úkol korelačního počtu je označován jako korelační analýza.

* pozn. … pojem závislé proměnné lze charakterizovat jako proměnnou, jejíž výsledek nám dopředu není znám a která je v experimentu postavena do pozice sledované (měřené) proměnné; naší snahou je vyvolat její změny prostřednictvím jiné proměnné (nezávislé proměnné), s jejíž hodnotami dokážeme manipulovat. Např. při otázce „Jak se změní u jedince čas potřebný k překonání vzdálenosti 100 m, bude-li jedinec chodit denně plavat?“ bude závislou proměnnou „čas potřebný k překonání vzdálenosti 100 m“ a nezávislou proměnnou řekněme „typ doplňkových tréninkových pohybových aktivit“.

V nejobecnějším smyslu slovo „korelace“ označuje míru stupně asociace dvou proměnných. Říká se, že dvě proměnné jsou korelované (resp. asociované), jestliže určité hodnoty jedné proměnné mají tendenci se vyskytovat s určitými hodnotami druhé proměnné. Míra této tendence může sahat od neexistence korelace (všechny hodnoty proměnné Y se vyskytují stejně pravděpodobně s každou hodnotou proměnné X) až po absolutní korelaci (s danou hodnotou proměnné X se vyskytuje právě jedna hodnota proměnné Y). Pro měření korelace byla navržena řada koeficientů. Liší se podle typů proměnných, pro které se používají. Statistické usuzování o korelačních koeficientech se opírá o teorii pravděpodobnosti pro společné rozdělení dvou nebo více náhodných proměnných.

Při zkoumání korelačních vztahů má rozhodující význam kvalitativní rozbor

příslušného materiálu. Nemá smysl měřit závislost tam, kde na základě logické úvahy nemůže existovat. Často je zbytečné měřit korelaci i z jiných důvodů. Je to zejména tehdy, když je korelace způsobena: (a) formálními vztahy mezi proměnnými; (b) nehomogenitou sledovaného studovaného základního materiálu; (c) působením společné příčiny.

Formální korelace vzniká například tehdy, když se zjišťuje korelace procentuálních charakteristik, jež se navzájem doplňují do 100 % (např. korelace procentuálního zastoupení bílkovin a tuku v potravinách).

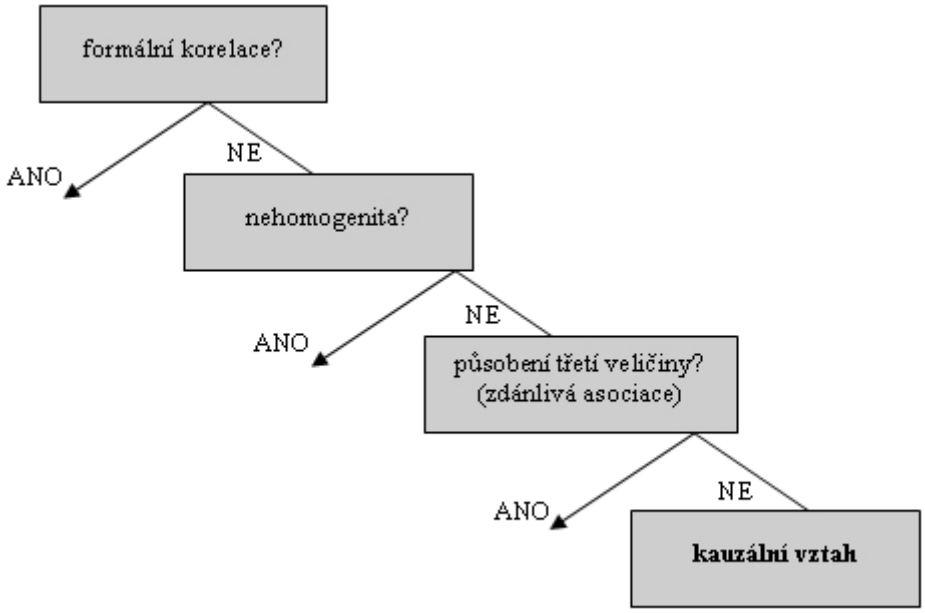
Jestliže populace, kterou sledujeme, obsahuje subpopulace, pro něž se průměrné hodnoty proměnných X a Y liší, vypočtené korelační vztahy jsou touto nehomogenitou silně ovlivněny a jejich hodnoty nepopisují skutečnou úroveň korelace, která ve skutečnosti mezi proměnnými reálně existuje. Nehomogenita materiálu se projeví na korelogramu (tento graf popsán níže) tak, že shluky bodů se budou nacházet v odlišných oblastech souřadnicového systému. Budou-li na korelogramu dva výrazné shluky jasně vypovídající o nehomogenitě souboru, bude výsledný koeficient korelace v sobě zahrnovat (a zároveň chybně zastupovat) dva jiné koeficienty korelace.

Příkladem korelací způsobených společnou příčinou jsou vztahy mezi některými mírami těla, např. mezi délkou levého a pravého femuru. Jiným známým příkladem jsou zdánlivé korelace způsobené časovým faktorem nebo faktorem modernizace u dvou řad údajů apod.

Mějme nyní příklad, kde byl vypočten silný korelační vztah mezi množstvím exportovaných automobilů z České republiky a počtem nově narozených dětí v rodině u osob, které pracují v automobilovém průmyslu. Můžeme si tedy položit otázku, zda zvýšením porodnosti u zaměstnanců v automobilovém průmyslu dosáhneme úspěšnějšího exportu automobilů.

Podobným korelacím se někdy říká „nesmyslné“ korelace. Hodnota korelace je vysoká. Nesmyslný by však byl závěr o přímém působení. Vztah zřejmě bude nutné vysvětlit působením přinejmenším třetí společné proměnné, která bude společnou příčinou.

Hendl (2004) popisuje postup pro ověřování kauzálního vztahu, jež je cílem našeho bádání – viz obr. 5.1.



Obrázek 5.1 Postup pro ověření kauzálního vztahu (Hendl, 2004, 243)

Lineární regrese

Regrese umožňuje postihnout povahu závislosti. Hledáme matematickou funkci, která by co nejlépe vyjadřovala charakter závislosti. Tato matematická funkce se nazývá regresní funkce a je vyjádřena regresní rovnicí. Regresní funkce může nabývat mnoha typů:

- - lineární regrese
- - kvadratická regrese
- - kubická regrese
- - polynomická regrese
- - hyperbolická regrese
- - logaritmická regrese

Nejjednodušší z nich je lineární regresní funkce, která má ve své empirické podobě tvar

$Y = a + b \,.\, x$

Pro konkrétní závislost (např. tělesné výšky a hmotnosti) je třeba určit tzv. regresní koeficienty a, b, přičemž vycházíme z empirických údajů (znaků) sledované závislosti.

Pro výpočet regresních koeficientů *a*, *b* je výhodné použít následující vzorce

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

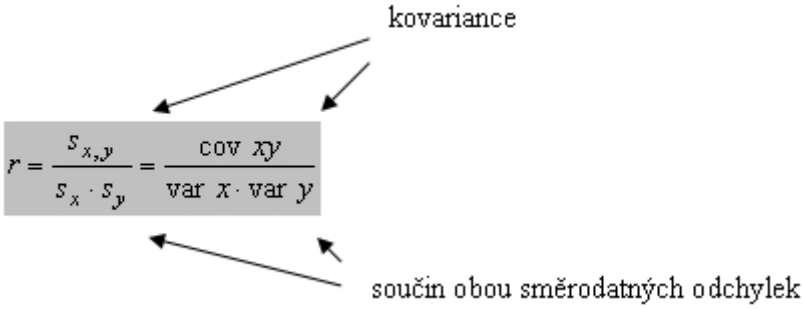
$$a = \frac{\sum y_i - b \sum x_i}{n}$$

Lineární korelace

Pro korelaci používáme symbol *r*, resp. r_{xy}. Lze ji definovat jako „volnou kvantitativní závislost dvou či více jevů“. Vyjadřuje stupeň neboli těsnost závislosti a jak již bylo popsáno v úvodu, je charakterizována číslem, tzv. korelačním koeficientem. Tento koeficient měří těsnost závislosti popsané lineární regresní

funkcí. Pro metrická data spojité náhodné veličiny užíváme k výpočtu míry těsnosti korelační závislosti tzv. Pearsonův koeficient součinové korelace a pro ordinální (pořadová) data tzv. Spearmanův koeficient pořadové korelace.

Výpočet korelačního koeficientu. Symbolická podoba vzorce pro výpočet korelačního koeficientu:



Korelace je tedy matematicky podíl kovariance a součinu dvou směrodatných odchylek příslušných dvou proměnných (viz dále). Kovarianci lze přitom chápat jako určitou „společnou směrodatnou odchylku obou proměnných".

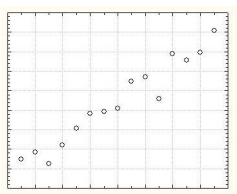
Pearsonův koeficient součinové korelace

Ačkoliv Pearsonův koeficient (značíme r_{xy}) má řadu nedostatků, přesto zůstává nejdůležitější mírou síly vztahu dvou náhodných spojitých proměnných X a Y. Počítáme jej z n párových hodnot $\{(x_i,y_i)\}$ změřených na n jednotkách náhodně vybraných z populace o rozsahu N. Matematické vyjádření Pearsonova korelačního koeficientu udává vzorec 5.2.3. Tento je však pro „manuální" výpočet značně nepraktický, proto se užívá odvozený výpočtový tvar (viz 5.2.4).

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

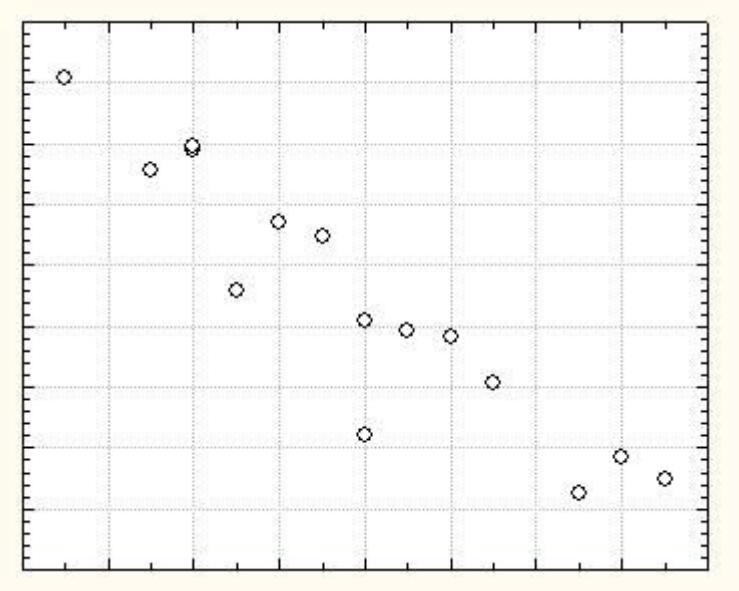
$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2] \cdot [n\sum y_i^2 - (\sum y_i)^2]}}$$

Korelační koeficient nabývá číselné hodnoty z intervalu <-1;1>, nebo-li r_{xy} je větší nebo rovno -1 a menší nebo rovno 1. Hodnoty -1 nabývá tehdy, pokud všechny body $[x_i,y_i]$ leží na přímce. Nule je roven v případě, že jsou veličiny nezávislé. Ovšem může nastat také situace, kdy je korelační koeficient roven nule, ačkoliv jsou proměnné funkčně závislé – a to tehdy, když se nejedná o lineární závislost. Proto je při použití Pearsonova korelačního koeficientu potřeba posoudit, kdy je jeho aplikace vhodná. Při měření lineární závislosti je znaménko korelačního koeficientu kladné, jestliže hodnoty obou proměnných X a Y rostou, znaménko záporné pak v případě, kdy hodnota jedné proměnné roste, zatímco druhé klesá. Na obrázcích 5.3 a 5.4 jsou zobrazeny typické příklady s kladným a záporným korelačním koeficientem.



Obrázek 5.3 Grafické znázornění korelační závislosti (korelogram); kladný

korelační koeficient



Obrázek 5.3 Grafické znázornění korelační závislosti (korelogram); záporný korelační koeficient

Při pohledu na hodnotu korelačního koeficientu by nás tedy mělo zajímat znaménko a velikost (absolutní hodnota). Znaménko udává směr, zatímco velikost udává, jak blízko jsou body nashromážděny kolem přímky, jež lze pomyslně oblastí zobrazených hodnot optimálně proložit.

Grafické znázornění korelační závislosti

Grafickým znázorněním odpovídajících si hodnot lze získat bodový graf, tzv. **korelogram**. Zde jsou hodnoty jednoho jevu (proměnná A) zaneseny na osu *X* a hodnoty druhého jevu (proměnná B) na osu *Y*. Tím dostáváme v rovině grafu *n* bodů (obr. 5.4), které postihují vlastní korelační závislost. Pro bližší posouzení závislosti se snažíme proložit získanými body „ideální přímku (křivku)“, která by nejlépe postihovala úroveň všech hodnot. Dle vzdáleností jednotlivých bodů od „ideální přímky“ usuzujeme na míru korelační závislosti (čím jsou body přímce blíž, tím vyšší je závislost; nejvyšší je závislost, pokud všechny body leží na přímce). Dále lze určit směr korelace – kladný (obr 5.3) nebo záporný (obr. 5.4). Tvar korelace (lineární, kvadratická, …) lze posoudit dle křivky, kterou lze „shlukem“ bodů proložit. V našem případě se omezíme pouze na případ lineární – body lze proložit přímku.

O nulové (žádné) korelační závislosti vypovídají jednoznačně následné tři případy: shlukem nelze proložit přímku, přímka je kolmá na osu *X*, přímka je kolmá na osu *Y*.

Pozn.: U statistických programů se lze setkat při výpočtech s uváděnou hodnotou *p*, tedy hladinou významnosti, na které je testována (ověřována) hypotéza, že mezi sledovanými proměnnými je korelace nulová. Což tedy znamená, že mezi proměnnými neexistuje žádný lineární vztah. Zde se je ale třeba zamyslet nad praktickým významem testování této hypotézy – pro rozsáhlé výběrové soubory se bude i velmi malá absolutní hodnota korelačního koeficientu statisticky významně lišit od nuly. Opět, jako pro jakoukoliv jinou statistiku, je nutné rozlišit statistickou

od její věcné významnosti. Dále je nutné si uvědomit, že *korelace neznamená příčinnost*!

Předešlé poznámky a úvahy shrneme do několika hlavních pravidel při práci s korelačním koeficientem:

1. Pearsonův korelační koeficient je užitečný pro určení míry lineární závislosti mezi dvěma metrickými spojitými proměnnými.
2. Tvar korelace. Vždy si nakreslete graf (korelogram), získáte rozumnou vizuální představu o korelaci. Korelace lineární x nelineární.
3. Směr korelace. Podívejte se na znaménko korelačního koeficientu. *Kladná* (pozitivní) korelace r náleží intervalu <0;1>; *záporná* (negativní) korelace r náleží intervalu <-1;0>.
4. Podívejte se na velikost absolutní hodnoty korelačního koeficientu. Korelace neznamená příčinnost. Významné jsou hodnoty (1) $r = 0$ (lineární nezávislost proměnných), (2) r = 1 (úplná-funkční-pozitivní lineární závislost), (3) r = -1 (úplná-funkční-negativní lineární závislost)
5. Hodnoty dosažené hladiny významnosti *p* uvedené spolu s hodnotami korelačních koeficientů (vypočtené pomocí statistických programů) berte kriticky – pro velký rozsah výběrových souborů může i *prakticky nevýznamná korelace být významná statisticky*.

Pokud nás zajímá pouze míra (síla) lineární závislosti, používáme místo korelačního koeficientu *r* spíše jeho druhou mocninu *r2*, které se říká **koeficient determinace**. Udává tu část rozptylu závisle proměnné *Y*, který lze vysvětlit variabilitou nezávislé proměnné *X*. Někdy bývá velice obtížné posoudit, která z proměnných *X* a *Y* je závislou a která nezávislou. Proto tuto determinaci s maximální mírou opatrnosti považujeme oboustranně. Např. získáme-li výpočtem koeficient determinace *r2* = 0,95 u proměnných běh na 100 m a běh na 60 m, znamená to, že proměnlivost proměnné Y (do jisté míry lze říct odchylky hodnot od průměru) lze z 95% vysvětlit proměnlivostí proměnné X. Nebo-li je pravděpodobné, že hodnoty proměnných jsou podmíněny 95 % společných faktorů a pouze 5 % faktorů je odlišných. U tohoto případů budou mezi společné faktory bezesporu patřit reakční rychlostní schopnost, výbušnost atd. Mezi odlišné pak pravděpodobně rychlostní vytrvalost event. jiné.

Čím více se bude r blížit hodnotě 1, tím považujeme danou závislost za silnější, čím více se bude r blížit hodnotě 0, tím považujeme danou závislost za slabší.

Poznámky ke korelacím

Matematicko-statistické předpoklady pro výpočet Pearsonova korelačního koeficientu jsou: linearita, normalita rozložení četností, dostatečný rozsah souboru. Je-li libovolná z těchto podmínek porušena, pak nelze použít Pearsonův korelační koeficient a je nutné volit jiný, jehož podmínky výpočtu nejsou tak silné.

Věcný a formální smysl znaménka korelačního koeficientu. Budeme-li uvažovat běh na 100 m a skok daleký kdy výpočtem získáme r = - 0,8, pak to znamená, jak bylo uvedeno v začátku, že s narůstajícími hodnotami v běhu na 100 m klesají hodnoty ve skoku dalekém a naopak. Bude tedy interpretace taková, že kdo dosahuje výborných výsledků v běhu na 100 m, bude podprůměrný ve skoku dalekém? Zřejmě nikoliv. Je zapotřebí si uvědomit, že nyní hovoříme o úrovni výkonu, zatímco výpočet korelačního koeficientu se odvíjel od výpočtu z naměřených hodnot. Tedy čím menších hodnot dosáhne jedinec v běhu na 100 m (zaběhne jej rychleji), tím

bude jeho úroveň výkonnosti větší (tedy nepřímá úměra). Ve skoku dalekém je tomu naopak – s rostoucí vzdáleností doskoku lze hovořit o vyšší výkonnosti (přímá úměra). Proto náš původní výsledek lze interpretovat tak, že zjištěná hodnota korelačního koeficientu hovoří o vysoké úrovni lineární závislosti mezi výkonností v běhu na 100 m a skoku dalekém.

Příklad 5.1 Postup při výpočtu Pearsonova korelačního koeficientu

PDF4_Pearson

Popis:

Ačkoliv rozsah souboru je malý, není ověřena podmínka normalita rozložení četností ani není ověřena linearita korelace – tedy podmínky, za kterých je možné použít Pearsonův korelační koeficient pro výpočet míry korelace mezi proměnnými *X* a *Y*, přesto data uvedené na první straně použijeme, a to z důvodu přehlednosti.

Při výpočtu Pearsonova korelačního koeficientu využíváme výpočtového tvaru vzorce, který je uveden na první straně. Po jeho shlédnutí vidíme, že pro jeho dosazení v podstatě potřebujeme znát pouze šest údajů – (1) rozsah výběrového souboru n; (2 a 3) součty všech hodnot proměnných *X* (Σx_i) a *Y* (Σy_i); (4 a 5) součty druhých mocnin hodnot obou proměnných *X* ($\Sigma x_i 2$) a *Y* ($\Sigma y_i 2$) a (6) součet součinů hodnot proměnných *X* a *Y* ($\Sigma(xi\ \Sigma yi)$). Za tímto účelem sestavujeme uvedenou tabulku (str.1), kde postupně do jednotlivých sloupců doplníme hodnoty. Mocniny hodnot lze vypočítat buď zadáním součinu v příkazovém řádku jako „=B7*B7“ (str.2) nebo pomocí funkce, kterou nalezneme v nabídce funkcí kliknutím na tlačítko fx vedle příkazového řádku. Takto otevřeme okno, kde v prvním podokně zadáme popis funkce, kterou chceme vyhledat (zde „mocnina“ - str.3). Kliknutí na Přejít dostáváme nabídku několika funkcí. Když si však projdeme popisy jednotlivých funkcí, umístěných pod výběrem, vybereme jedinou funkci, která vyhovuje našemu požadavku druhé mocniny, a to je funkce POWER (str.4). Přes OK se dostáváme k argumentům této funkce – ty jsou dva – zadání hodnoty, kterou chceme umocnit a druhým argumentem je mocnina, kterou chceme vypočítat (v našem případě „2“) (str.5).

Nyní si výpočet hodnot $x_i 2$ usnadníme „kopírováním vzorce“ – uchopíme levým tlačítkem myši pravý dolní roh buňky, jejíž vzorec chceme kopírovat (D8) a tažením označíme oblast, kam chceme vzorec kopírovat (D9:D21). Takovýmto kopírováním se kopíruje typ funkce a postupně se mění argument funkce (v našem případě číslo, jež chceme umocnit) dle kopírované buňky (str.7).

Obdobně aplikujeme kopírování vzorce na oblast pro výpočet $y_i 2$ (sloupec E). Označíme do bloku oblast D7:D21, jejíž vzorce budeme kopírovat uchopením levé myši za pravý dolní roh a tažením o jeden sloupec doprava, a nakopírujeme vzorec do oblasti E7:E21 (str.8).

Pro výpočet součinů $x_i * y_i$ použijeme přímé zadání vzorce, tedy v speciálně pro prví buňku bude platit vzorec „=B7*C7“ (str.9). Opět vzorec kopírujeme do zbývajících buněk (str. 9 a 10).

Pro výpočet součtů jednotlivých sloupců x_i , y_i , $x_i 2$, $y_i 2$, $x_i * y_i$ použijeme funkci součtu, kterou lze nalézt opět ve výběru funkcí a nebo najít na panelu tlačítek jako Σ (nazvaná „*AutoSum*“ viz str.11 a 12). Jsme-li na buňce, kde chceme zobrazit součet, kliknutím na tlačítko je nám nabídnuta oblast dat (str.13), kterou program předpokládá, že chceme sečíst. V našem případě tato nabídnutá oblast odpovídá požadované oblasti a proto pouze potvrdíme výpočet klávesou ENTER (str.14).

Tento vzorec opět kopírujeme do zbývajících buněk, kde chceme provézt výpočet

(C22:F22) (str.15).

Nyní již máme dostatečný podklad pro dosazení všech potřebných hodnot do výpočtového tvaru vzorce Pearsonova korelačního koeficientu.

Vzhledem k určité složitosti vzorce a s nutností, aby v něm matematické operace proběhly tak jak mají, je nutné dobře psát závorky, které určují postup výpočtu. Proto na str.16 vidíme výchozí tvar vzorce „=()/()“ (odděluje ve zlomku čitatel od jmenovatele), do kterého budeme postupně psát vždy nejprve závorky a teprve poté dosazovat hodnoty.

Nejprve se zaměříme na zadání pro výpočet čitatele. Ze vzorce pro výpočet korelačního koeficientu vidíme, že již není zapotřebí zadávat další závorky, které by upřednostnili určitou matematickou operaci, a proto již přímo odkazujeme na buňky, které obsahují výpočet hodnot patřících do vzorce (str.17).

Zadání pro výpočet jmenovatele již bude o něco náročnější, neboť je zde řada operací, které musí nastat v určitém sledu. Budeme postupovat z vnějšku a postupně upřesňovat vzorec.

Nejprve tedy zadáme odmocninu součinů dvou hodnot (tak jak je vidět ve vzorci korelačního vzorce – součin dvou hranatých závorek) (str.18). Jak vidíme, pro výpočet odmocniny použijeme funkce (lze ji opět nalézt ve výběru funkcí) „ODMOCNINA( )“, která obsahuje pouze jeden argument, a tím je číslo, které chceme odmocnit. Proto do argumentu zadáváme přímo součin „( )*( )“.

Můžeme postoupit dále směrem do vzorce, tedy do součinu dvou hodnot ohraničených závorkami. Odkazujeme se postupně na hodnoty, které mají být do vzorce dosazeny (postupně str. 19 a 20).

Je důležité mít vždy všechny závorky správně uzavřeny z obou stran, jak mají být uzavřeny, neboť i nepatrná změna může vézt ke zcela odlišnému výsledku nebo může dojít k situaci, že program bude hlásit chybu, že vzorec obsahuje chybu. Je také nutné mít na paměti, že vždy je lepší zadat více závorek a být si jistý postupem matematických operací, které ve vzorci proběhnou než naopak.

Potvrzením vzorce klávesou ENTER získáváme výsledek roven hodnotě 0,8835, což svědčí o středně vysoké úrovni korelace (str.21).

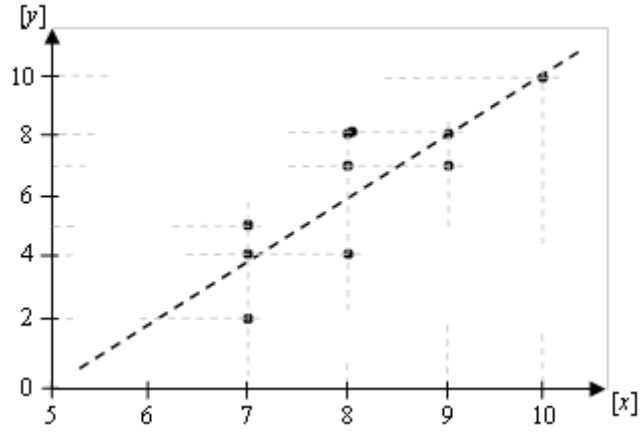
Náš výpočet si lze snadno ověřit pomocí funkce, kterou program nabízí (str.22). Tuto funkci vyhledáme v nabídce funkcí, kde popisem „korelace“ dostáváme nabídku tří funkcí (str.23-24). Po prostudování charakteristik těchto funkcí si logicky vybereme funkci „PEARSON( )“, která má dva argumenty – dvě oblasti dat (Pole1 a Pole2), mezi kterými má být stanovena míra korelace pomocí Pearsonova korelačního koeficientu. Tyto oblasti nejjednodušeji zadáme označením každého Pole jako blok dat pomocí myši (str.25-28). Potvrzením vzorce získáváme hodnotu 0,8835, která skutečně odpovídá původně vypočtené námi hodnotě.

KE STAŽENÍ PŘÍKLAD UKÁZKY PŘÍKLADU V MS EXCEL

Příklad 5.2

Konstrukce korelogramu. Sestrojte bodový graf vyjadřující korelační závislost proměnných X a Y. Proměnné X a Y lze zapsat jako deset uspořádaných dvojic $\{x_i; y_i\}$: {7;4},{7;2},{7;5},{8;8},{9;7},{8;8},{8;7},{8;4},{9;8},{10;10}

Postupujeme tak, že postupně vynášíme jednotlivé uspořádané dvojice hodnot na osy x a y. Tím získáme v souřadnicovém systému celkem 10 bodů, jejichž uspořádání vyjadřuje míru, tvar a směr korelace (viz obr. 5.4).



Příklad 5.3

Regresní rovnice, regresní přímka a korelační koeficient. Na podkladě dat z předchozího příkladu 5.2 sestavte rovnici regresní přímku $Y = a + b^* X$ a vypočtěte korelační koeficient r_{xy}.

	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
Osoba 1	7	4	49	16	28
Osoba 2	6	8	36	64	48
Osoba 3	7	6	49	36	42
Osoba 4	8	8	64	64	64
Osoba 5	9	7	81	49	63
Osoba 6	8	8	64	64	64
Osoba 7	8	7	64	49	56
Osoba 8	8	4	64	16	32
Osoba 9	9	8	81	64	72
Osoba 10	10	10	100	100	100
Σ	80	70	652	522	569

1. Výpočet koeficientů regresní přímky. Z teoretických poznatků uvedených výše víme, že pro výpočet koeficientů a a b regresní přímky $Y = a + b^*X$ používáme tyto vzorce :

$$b = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$a = \frac{\sum y_i - b \cdot \sum x_i}{n}$$

Dle tab. 5.1 tedy můžeme psát

$$b = \frac{10 \cdot 569 - 80 \cdot 70}{10 \cdot 652 - (80)^2} = \frac{90}{6520 - 6400} = \frac{90}{120} = 0,75 \cong 0,8$$

$$a = \frac{70 - 0,8 \cdot 80}{10} = \frac{6}{10} = 0,6$$

Vypočtené hodnoty pro $a = 0,6$ a $b = 0,8$ dosadíme do vzorce a dostáváme rovnici regresní přímky

$Y = 0,6 + 0,8 * x$.

2. Konstrukce regresní přímky za pomocí regresní rovnice. Jednou z možností (asi nejsnadnější), jak zakreslit regresní rovnici, známe-li regresní rovnici, je vypočítat

dvě uspořádané (vzájemně si odpovídající) dvojice hodnot, vynést jejich hodnoty do souřadného systému a jejich propojením dostaneme hledanou přímku.

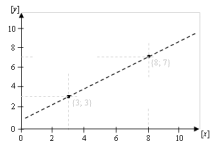
Vyjdeme tedy z výše vypočtené rovnice $Y = 0,6 + 0,8*x$, zvolíme si libovolné dvě hodnoty x_i nezávislé proměnné X (ovšem takové, aby se nám dobře vynášely v souřadném systému). Ty postupně doplníme do regresní rovnice a vypočteme jim odpovídající hodnoty y_i.

Mějme tedy například hodnoty $x_1 = 3$ a $x_2 = 8$. Dosadíme do rovnice.

$y_1 = 0,6 + 0,8 * 3 = 3$

$y_2 = 0,6 + 0,8 * 8 = 7$

Nyní máme hledané uspořádané dvojice {3;3} a {8;7}. Vynesením hodnot do souřadného systému a propojením získaných bodů dostáváme regresní přímku, která má rovnici $Y = 0,6 + 0,8 * x$ (viz obr. 5.5)



3. Výpočet Pearsonova korelačního koeficientu.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2] \cdot [n\sum y_i^2 - (\sum y_i)^2]}}$$

Za účelem snadného dosazení do výpočtového tvaru Pearsonova korelačního koeficientu (vzorec 5.2.7) použijeme již jednou vypočtené hodnoty z tab. 5.1, které jsme používali pro stanovení regresní rovnice. Jak je patrné z již uvedeného vzorce, budou pro nás užitečné hodnoty $\Sigma x_i = 80$; $\Sigma y_i = 70$; $\Sigma x_i 2 = 652$; $\Sigma y_i 2 = 522$ a $\Sigma x_i \Sigma y_i = 569$. Nyní můžeme tyto hodnoty dosadit do vzorce (viz 5.2.8):

$$r_{xy} = \frac{10 \cdot 569 - 80 \cdot 70}{\sqrt{(10 \cdot 652 - 6400) \cdot (10 \cdot 522 - 4900)}} = \frac{90}{\sqrt{120 \cdot 320}} = 0,46$$

Na základě sestaveného korelogramu a nyní vypočtené hodnotě korelačního $r_{xy} = 0,46$ můžeme říct, že korelace má v lineární tvar, pozitivní směr a její velikost je mírná.

Lze zároveň snadno dopočítat koeficient determinace $r2$, který bude roven hodnotě 0,2116. V souladu s výše uvedeným lze říct, že přibližně pouhých 21% variability závislé proměnné Y lze vysvětlit variabilitou nezávislé proměnné X.

Spearmanův koeficient pořadové korelace

V předchozí kapitole jsme se zabývali výpočtem korelačního koeficientu u spojitých metrických veličin, které navíc musely splňovat určité požadavky, jako je normalita rozložení četností, dostatečný rozsah souboru a linearita závislosti. Koeficientem korelace, který již nemá tak silné předpoklady, jako v případě Pearsonova korelačního koeficientu je Spearmanův koeficient pořadové korelace. Jak již plyne ze samotného názvu, je zřejmé, že pro jeho stanovení bude podstatné pořadí hodnot. Lze jej tedy použít pro výpočet těsnosti závislosti jak u metrických dat, které nesplňovaly výše uvedené požadavky, tak i u dat ordinálních. Užití Spearmanova korelačního koeficientu pro analýzu dat se řadí mezi neparametrické metody.

Princip Spearmanova korelačního koeficientu vychází z úvahy, že má-li mezi sebou vysoce korelovat řada hodnot proměnné X a řada hodnot proměnné Y, pak pořadí hodnot proměnné X by mělo odpovídat (přímo nebo nepřímo úměrně) pořadí hodnot

proměnné *Y*. Tedy seřadíme-li hodnoty proměnné *X* podle velikosti, pak jim odpovídající hodnoty proměnné *Y* by měly mít přibližně stejné pořadí. Z toho také vychází výpočet Spearmanova koeficientu pořadové korelace (vzorec 5.2.9).

$$r_{xy} = 1 - \frac{6 \cdot \sum (i_x - i_y)^2}{n \cdot (n^2 - 1)}$$

, kde *ix* resp. *ix* je index pořadí hodnot *x* resp. *y*.

Příklad 5.4

Postup při výpočtu Spearmanova korelační koeficient pořadí

PDF4_Spearman

Popis:

Pro výpočet Spearmanova koeficientu pořadové korelace použijeme stejných dat, které byly použity pro výpočet Pearsonova korelačního koeficientu. Takto bude možné porovnat vypočtené hodnoty.

Tak jako v případě Pearsonova korelačního koeficientu, také zde potřebujeme pro dosazení do vzorce pro výpočet Spearmanova korelačního koeficientu pouze několik hodnot, těmi jsou (1) rozsah výběrového souboru n a (2) součet mocnin rozdílů indexů i_x a i_y. Proto i v tomto případě si pro usnadnění výpočtu sestavíme výpočtovou tabulku (str.1).

Tato tabulka obsahuje samozřejmě sloupce se záhlavím x_i a y_i. Dále budou zapotřebí sloupce se stanovenými indexy i_x a i_y, označující pořadí hodnot ve variační řadě dané proměnné, aby bylo možné vypočítat druhou mocninu rozdílů (i_x - i_y) – tedy poslední sloupec tabulky.

Zaměřme se nyní na postup stanovení indexů i_x a i_y. Jak již bylo popsáno výše, znamenají tyto indexy pozici, jakou hodnota zaujímá ve variační řadě (hodnoty proměnné seřazené podle velikosti). Proto prvním krokem bude seřadit hodnoty podle velikosti. Při určení indexu i_x to bude podle proměnné x_i a u indexu i_y podle proměnné *yi*. Je ale nutné mít na vědomí, že jednotlivé dvojice hodnot x_i a y_i musejí být zachovány. Např. budeme-li sledovat proměnné školní známka a gymnastické body v akrobacii u osob určité základní školy. Postupně měříme hodnoty proměnných u osoby1, osoby2, atd. V případě, že bychom seřadili hodnoty podle velikosti samostatně pro jednu a poté pro druhou proměnnou, mohla by nastat situace, že hodnota školní známky a hodnota bodů v akrobacii by nebyly stejné pro stejnou osobu – jinými slovy uspořádanost dvojice hodnot by byla porušena. To za účelem výpočtu korelačního koeficientu nesmí nastat. Proto budeme postupovat následovně. Označíme oblast uspořádaných dvojic (blok B10:C24) a ty seřadíme nejprve podle proměnné *xi* (postupně na str.2-5). K tomu využijeme pod položkou „Data“ možnost „Seřadit …“. Kliknutím na tuto položku se otevře okno, kde zadáváme nezbytné údaje pro námi požadované seřazení hodnot. My si povšimneme dvou položek – „Seznam“ a „Seřadit podle“. Položka „Seznam“ obsahuje dvě možnosti zadání – „Se záhlavím“ a „Bez záhlaví“. Označujeme tím, zda se v námi označené oblasti nachází záhlaví nebo nikoliv. V případě znázorněném na str.3 nabídl program se záhlavím, což je chybně (to vidíme i na původně označené oblasti buněk – program zúžil označenou oblast – to my nechceme) a proto volíme druhou možnost. Po rozvinutí položky „Seřadit podle“ nám nyní program nabízí „SloupecB“ a „SloupecC“. Vzhledem k tomu, že potřebujeme seřadit hodnoty proměnné *X*, volíme „SloupecB“. Potvrdíme funkci přes OK. Dvojice hodnot jsou nyní seřazeny podle variační řady proměnné *X*.

Při stanovení indexů pořadí nyní mohou nastat dvě možnosti - (1) hodnota proměnné *X* se vyskytuje ve variační řadě pouze jednou a potom je jejím indexem číslo označující pozici v řadě (první místo – 1, páté místo – 5, …); (2) hodnota proměnné *X* se vyskytuje ve variační řadě vícekrát – hodnota indexu je rovna průměru ze všech pozic, které tato hodnota ve variační řadě zaujímá (jestliže hodnota 3 je v řadě 2krát a to na pozici 10 a 11, pak jejím indexem je číslo 10,5). Na str.6-8 jsou postupně doplňovány indexy proměnné *X*. Hodnota „1" proměnné *X* je ve variační řadě třikrát a to na pozici 1, 2 a 3 – průměr je roven číslu 2, proto index všech tří hodnot „1" je index „2". Hodnota „2" proměnné *X* se ve variační řadě nachází pouze jednou a to na pozici 4, proto je jejím indexem přímo číslo „4".

Nyní je potřeba doplnit indexy pro proměnnou *Y*. Budeme tedy postupovat analogicky jako v případě proměnné, ale budeme data řadit podle „SloupecC". Nezapomeňme ale, že při přeskupení pořadí dat již nemůžeme mít označeny pouze dva sloupce, jak tomu bylo v předchozím, ale že nám přibyl další sloupec, který musí s původními rovněž korespondovat. Tedy již nemáme pouze uspořádané dvojice, ale již uspořádané trojice hodnot, které se nám budou přeskupovat. Proto (jak je vidět na str.8) při seřazování dat označujeme oblast B10:D24 (tedy tři sloupce).

Po seřazení hodnot podle sloupce C (tedy podle hodnot yi) doplníme indexy iy (str.11-12).

Poslední sloupec snadno dopočítáme buď pomocí funkce „POWER((ix - iy);2)" nebo přímým zadáním vzorce do buňky jako „= (ix - iy)*(ix - iy)". Dále vzorec kopírujeme do zbývajících buněk již známým postupem – tedy uchopíme levou myší pravý dolní roh buňky, která obsahuje vzorec (u nás F10) a tažením dolů nakopírujeme vzorec do zbývajících buněk. Na závěr za pomocí funkce „SUMA( )" provedeme do buňky F25 součet všech druhých mocnin. (strany 13-15).

Vzorec pro výpočet Spearmanova korelačního koeficientu je na doplnění snadnější, než tomu bylo v případě Pearsonova korelačního koeficientu. Přesto nesmíme zapomenout na správné umístění závorek do vzorce tak, aby všechny matematické operace proběhly v pořadí tak jak mají (str.16-18).

Vidíme (str.19), že výsledkem je hodnota 0,8902, která se od výpočtu pomocí Pearsonova korelačního koeficientu (=0,8835) liší jen nepatrně.

V tomto případě bychom se přiklonili spíše k výsledku Spearmanova korelačního koeficientu, vzhledem k malému rozsahu souboru, neověřené normalitě rozložení četností proměnných *X* a *Y* a neověření linearity korelace.

KE STAŽENÍ PŘÍKLAD UKÁZKY PŘÍKLADU V MS EXCEL

Příklad 5.5

Pro hodnoty proměnných *X* a *Y* uvedených v příkladu 5.1 vypočtěte Spearmanův koeficient pořadové korelace.

Pořadí	x_i	y_i	i_x	i_y	$(i_x - i_y)^2$
1	7	4	2,5	1,5	1
2	6	8	1	7,5	42,25
3	7	6	2,5	3	0,25
4	8	8	5,5	7,5	4
5	9	7	8,5	4,5	16
6	8	8	5,5	7,5	4
7	8	7	5,5	4,5	1
8	8	4	5,5	1,5	16
9	9	8	8,5	7,5	1
10	10	10	10	10	0
Σ	-	-	-	-	87,5

$$r_{xy} = 1 - \frac{6\sum(i_x - i_y)^2}{n(n^2-1)} = 1 - \frac{6 \cdot 87{,}5}{10(100-1)} = 1 - \frac{525}{990} = 0{,}47$$

STATISTICKÁ A KORELAČNÍ ZÁVISLOST VÍCE ZNAKŮ

Dosud jsme zabývali závislostí dvou znaků, závislostí jedné závislé proměnné na jedné nezávisle proměnné, tedy jednoduchou (párovou) závislostí. Ve skutečnosti nejsou většinou závislosti jednoduché, ale vícenásobné (mnohonásobné), je tedy nutno zobecnit vztahy v části 7. na více proměnných. Budeme uvažovat závislost závisle proměnné y na více nezávisle proměnných $x_1, x_2, x_3, ..., x_i$.

Úkoly korelačního počtu jsou obdobné jako u jednoduché korelační závislosti, tedy:

1. Postižení povahy vícenásobné korelační závislosti (vícenásobná regrese)

2. Měření těsnosti vícenásobné korelační závislosti (vlastní vícenásobná korelace vyjádřená korelačním koeficientem)

Vícenásobná regrese a korelace

Chceme-li odhadovat individuální neznámé hodnoty závisle proměnné y ze známých individuálních hodnot nezávisle proměnných $x_1, x_2, x_3, ..., x_i$, musíme vystihnout průběh zkoumané závislosti analytickou (matematickou) funkcí η. Tuto funkci nazýváme vícenásobná regresní funkce.

Ve vícenásobné regresní funkci je proměnná η funkcí proměnných $x1, x2, x3, ..., xi$ a neznámých parametrů $\beta_1, \beta_2, ..., \beta_p$,

$$\eta = \eta(x_1, x_2, \ldots, x_i; \beta_1, \beta_2, \ldots, \beta_p)$$ a tuto funkci nazýváme teoretická regresní funkce.

Abychom odhadli neznáme hodnoty funkce η, je třeba odhadnout hodnoty parametrů $\beta_1, \beta_2, ..., \beta_p$. Označíme-li tyto odhady $b_1, b_2, ..., b_p$, dostaneme odhad regresní funkce Y.

$$Y = Y(x_1, x_2, \ldots, x_i; b_1, b_2, \ldots, b_p)$$ Hovoříme potom o empirické vícenásobné regresní funkci. Vícenásobné regresní funkce a korelace vyžadují poměrně složité matematické postupy, které jsou nad rámec učiva.

6 ČASOVÉ ŘADY

DEFINICE ČASOVÝCH ŘAD

V odborné literatuře se můžete setkat s řadou definic, které se však ve své podstatě shodují. Ve všech se hovoří o časové řadě (dále jako ČŘ) jako o sekvenci hodnot určitého ukazatele – proměnné, které jsou řazeny dle časové posloupnosti.

Pro naše potřeby definujme ČŘ následovně: „Časová řada je posloupnost věcně srovnatelných pozorování (dat, údajů), které jsou uspořádány na časové ose“.

S ČŘ se setkáváme snad ve všech oblastech lidské činnosti, ať se jedná o data “přírodního charakteru” nebo data ekonomická. Rovněž v oblasti sportu a pohybové aktivity vůbec mají své místo. Většinou sledujeme určité vývojové trendy některého ukazatele u sportovce (nebo i celé populace), jako jsou výsledky motorických testů, výsledky fyziologických testů, výkonnost, somatické parametry (váha, výška, obvodové charakteristiky) apod.

Obecně lze jmenovat dva hlavní cíle užití ČŘ: (1) identifikace podstaty jevu zastoupeného sekvencí pozorovaní, a (2) předpověď (predikce budoucích hodnot ČŘ).

Oba tyto cíle vyžadují, aby byl identifikován a alespoň v hrubé formě popsán „model ČŘ“. Jakmile je model sestaven, můžeme jej interpretovat a integrovat s dalšími daty.

Časové řady zapisujeme symbolem:

$$\{y_n\}_{t=1}^{n}$$

Vzorec (1.1)

To si můžeme představit takto: $y_1, y_2, \ldots, y_n$

(čteme „ypsilon-té pro t od 1 po n“)

KLASIFIKACE ČASOVÝCH ŘAD

ČŘ lze rozdělit podle několika hledisek. Při jejich dělení vždy záleží na druhu klasifikace a účelu klasifikace. Jejich členění přitom není samoúčelné. Zařazení sledované ČŘ do správné kategorie hned v úvodu vlastního výzkumu je nezbytným krokem, pokud se nechceme dopustit pozdějších chybných úsudků. Tak jako bylo například nutné rozlišovat typ dat (nominální, ordinální, …) ve smyslu možností užití statistických metod, které lze na jednotlivé typy dat aplikovat, i v tomto případě se jedná o aplikovatelnost statistických metod a případně nutnost úprav dat před dalším statistickým zpracováním – ne každý statistický postup je vhodný pro každý typ řady.

Obrázek (2.1)

Za základní členění ČŘ lze označit rozdělení na *intervalovou* a *okamžikovou ČŘ* (viz obrázek 2.1). Zatímco každá hodnota intervalová ČŘ vždy odráží jistou kumulaci jevu za určité období (období od předchozí hodnoty), je hodnota okamžikové ČŘ obrazem daného jevu pouze v daný okamžik bez vztahu k předchozí hodnotě.

Srovnejme následné dva měřitelné jevy, které je možné považovat za ČŘ. Představme si jedince, který běží maratón. Maratón se běží v parku nějakého města na okruhu nějaké vzdálenosti. Sportovec může svůj výkon hodnotit několika způsoby – uveďme například tyto: (1) každých 5 minut zaznamená vzdálenost, kterou v tomto časovém intervalu uběhl; (2) každou minutu si zaznamená tepovou frekvenci. V obou případech se jedná o určité charakteristiky, které vypovídají o jeho kvalitě výkonu a které bude (třeba) možné později vyhodnotit. Jak se tyto případy liší? Zatímco v prvním případě zaznamenaná hodnota vzniká sumací překonaných metrů za časový úsek 5 minut, v druhém případě hodnota odráží okamžitý stav (je možné, že kdyby měření proběhlo o 10 s dříve, hodnota by byla jiná). Samozřejmě oba typy záznamu výkonu mají své výhody a nevýhody z praktického/věcného hlediska, my se však přidržíme toho statistického hlediska. Uvědomění si typu časové řady bude mít vliv na její další zpracování, jak bude zřejmé později.

U intervalových ČŘ je někdy nutné provést tzv. *očištění dat od kalendářních variací*. Jedná se o případy, kdy data nejsou měřeny ve stejných časových odstupech – měříme-li hodnotu pravidelně na konci měsíce, bude výsledek zajisté ovlivněn faktem, že každý měsíc má jiný počet dní a tedy období pro kumulaci jevu je odlišné; rovněž měření jednou ročně by mohlo být zavádějící v důsledku přestupného roku apod. S touto skutečností se vyrovnáváme vcelku jednoduše – využíváme vzorce (2.1), který vytvoří novou ČŘ oproštěnou tohoto nedostatku.

$$y_t^{(0)} = \frac{y_t}{k_t}\bar{k}$$

Vzorec (2.1)

kde $y_t^{(0)}$ jsou hodnoty nové ČŘ, y_t hodnoty původní ČŘ, k_t délka časového období, kterého se příslušná hodnota týká a $\bar{k}$ je průměrná délka časového intervalu. Takto je tedy vytvořena ČŘ, která má hodnoty na časové ose ve stejné vzdálenosti

(všechny časové intervaly jsou shodné). Další možnosti členění jsou v přehledu tabulky:

Tabulka 2.1 Členění časových řad

hledisko	ČASOVÁ ŘADA		
délka časového intervalu	DLOUHODOBÁ		KRÁTKODOBÁ
původ hodnot ČŘ	ODVOZENÝCH UKAZATELŮ		ABSOLUTNÍCH UKAZATELŮ
	(průměry, součty, podíly aj.)		(počty, metry, sekundy, …)
povaha hodnot ČŘ	PENĚŽNÍCH UKAZATELŮ		NATURÁLNÍCH UKAZATELŮ
	(ekonomické hodnoty)		
homogenita časového intervalu	EKVIDISTANTNÍ		NEEKVIDISTANTNÍ
	(všechny časové intervaly jsou shodné)		

Úkol: Určete typ časové řady v případě sportovce běžícího maratón (obě sledované charakteristiky) dle všech hledisek.

ZÁKLADNÍ CHARAKTERISTIKY ČŘ

Tak jako u běžného souboru dat, bývá také v případě ČŘ vhodné charakterizovat povahu řady souhrnně, nahradit řadu vhodnou charakteristikou, která by ji vhodně postihovala. Asi za nejužívanější charakteristiky lze označit průměr, součty (pouze u intervalových řad!), tempo růstu, grafické znázornění apod.

Pro srovnání více ČŘ se nám nejvíce hodí již známá charakteristika průměr, charakterizující úroveň celé časové řady ve smyslu její polohy. Součty naměřených údajů však mají smysl pouze u ČŘ intervalové, a proto pouze u ní má smysl určovat průměr. Již z předchozích kapitolách jsme se seznámili s aritmetickým průměrem, zmíněn byl rovněž chronologický průměr, který má svůj význam obzvláště u časových řad.

Aritmetický průměr

$$\bar{y} = \frac{\sum y_i}{m}$$ Vzorec (3.1)

Kde y_i je hodnota i-tého členu (i = 1, 2, …, n), n je počet členů ČŘ, m je počet stejně dlouhých období v celém časovém úseku.

Chronologický průměr

$$\bar{y} = \frac{1}{k-1}\left(\frac{y_1 + y_2}{2} + \ldots + \frac{y_{k-1} + y_k}{2}\right)$$ Vzorec (3.2)

Zde je *k* počet stejně dlouhých časových úseků v průběhu celého pozorování.

Úkol 1: Užití aritmetického průměru.

Zadání: Úkolem je zjistit, jaká je průměrně uplavaná vzdálenost v plavecké jednotce. Data pocházejí z tréninkového záznamu plavce (smyšlený), který pravidelně střídal plavecké styly kraul a prsa (celkem desetkrát). Kraul byl vždy plaván po dobu 5 minut a plavecký styl prsa po dobu 3 minut. Záznam (je uveden v metrech): 650; 328; 645; 340; 658; 327; 658; 330; 652; 340

Řešení: Jak je ze zadání patrné, jedná se o kumulaci hodnot vždy ve dvou časových intervalech, které se pravidelně střídají. Aby ovšem bylo možné o oné „plavecké jednotce" hovořit souhrnně, je nutné nejprve data očistit od těchto rozdílů. Je nutné si uvědomit, že plavecká jednotka tak nebude ani 5 ani 3 minuty, ale bude průměrem těchto dvou hodnot, tj. (5+3)/2 = 4 minuty.

Pro očištění dat užijeme vzorec, jež byl již uveden dříve (vzorec 2.1):

$$y_t^{(0)} = \frac{y_t}{k_t}\bar{k}$$

V našem případě y_t jsou hodnoty záznamu, k_t délka jsou střídavě hodnoty 5 a 3 (délka příslušného intervalu), a $\bar{k}$ je 4 – průměrná plavecká jednotka.

K výpočtu a dosazení do vzorce pro výpočet aritmetického průměru sestavíme pro zjednodušení a přehlednost následnou tabulku

Tabulka (3.1)

Záznam	Původní hodnoty	Očištěné hodnoty
1.	650	520
2.	328	437
3.	645	516
4.	340	453
5.	658	526
6.	327	436
7.	658	526
8.	330	440
9.	652	521
10.	340	453
Součet	-	4830
Průměr	4830/10	**483**

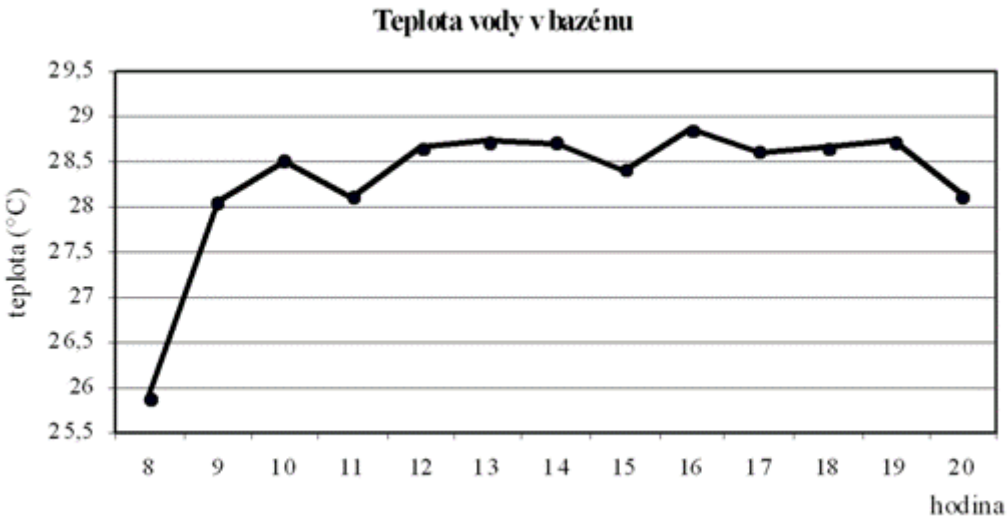
Výsledek: Průměrně uplavaná vzdálenost v plavecké jednotce je 483 metrů.

Grafické znázornění ČŘ

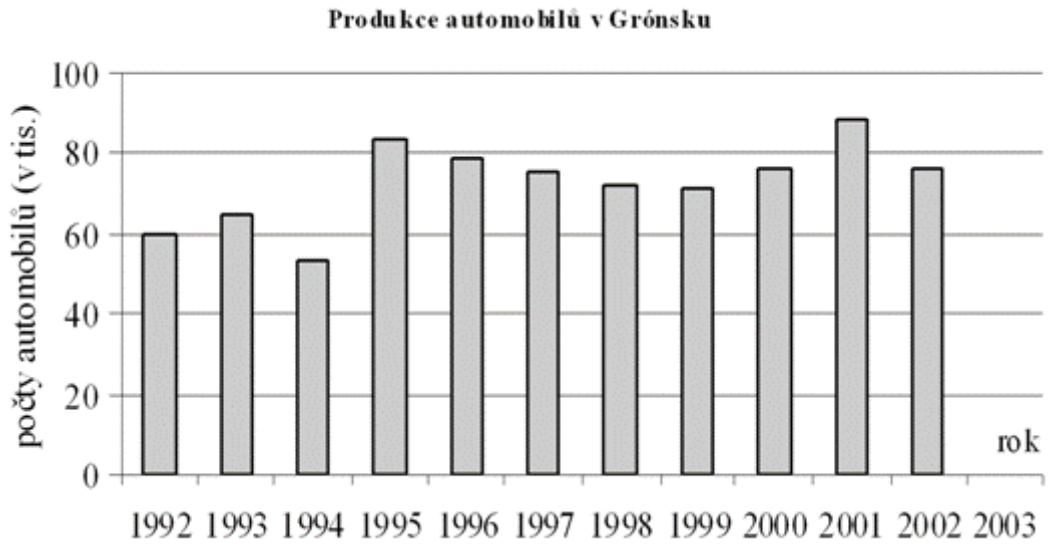
Grafické znázornění časové řady je jednou z nejpoužívanějších charakteristik pro

vyjádření ČŘ, a to především z důvodu přehlednosti. Při zkoumání ČŘ lze díky grafickému znázornění rychle získat hrubou představu o trendu řady, extrémních hodnotách, deviacím atd. Je vhodné jej rovněž užít pro prezentaci výsledků výzkumu, a to právě pro zmíněnou přehlednost.

Pro různé typy ČŘ jsou příhodné různé typy grafů – pro ČŘ okamžikovou je vhodné použít *spojnicový graf* v pravoúhlé soustavě souřadnic tak, že vodorovná osa je vyhrazena pro časové údaje (obrázek 3.1); pro ČŘ intervalovou je častěji použit *sloupcový graf*, kde šířka sloupce vždy odpovídá délce časového intervalu (obrázek 3.2). Do grafů jsou často zakreslovány další křivky, které vypovídají o dalších vlastnostech ČŘ (o popisovaném jevu). Nejčastěji se jedná o tzv. *spojnici trendu* a odhad dalšího průběhu ČŘ.



Obrázek (3.1)



Obrázek (3.2)

MODELOVÁNÍ ČASOVÝCH ŘAD

Při modelování časových řad se snažíme najít určitý model (často zastoupen matematickou funkcí), který by zpřehlednil naměřené hodnoty a tím umožnil snazší interpretaci. Tak jako jsme dříve hledali popisné charakteristiky, jako jsou průměr, směrodatná odchylka aj., abychom jimi výstižně postihli sledovaného hodnoty souboru (jednalo se tedy o určité zástupce souboru), podobně hledáme vhodný model ČŘ, který naměřená data zjednoduší a zpřehlední.

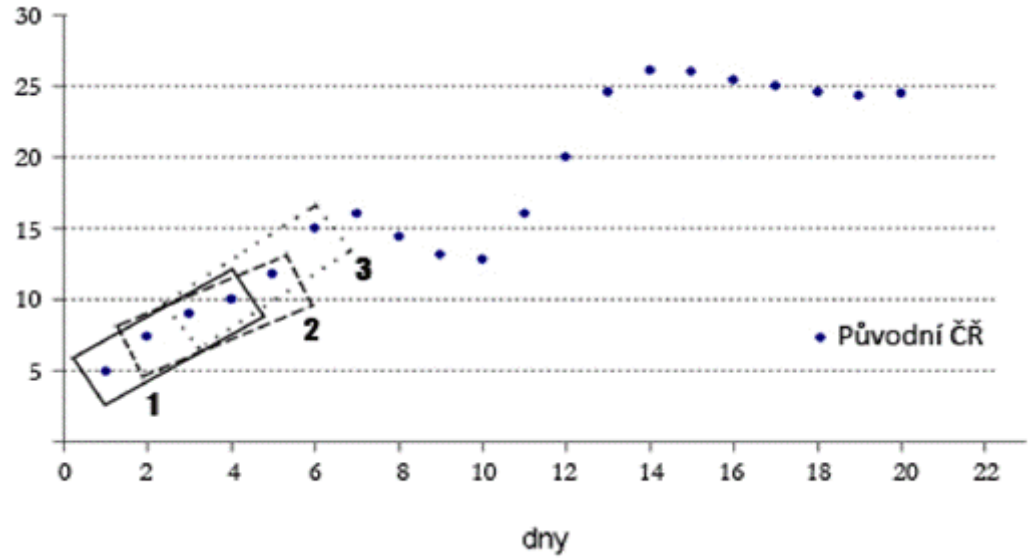
Většina modelů ČŘ může být popsána v termínech dvou základních složek: trend a sezónnost. Prvně jmenovaná representuje obecně lineární nebo (mnohem častěji) nelineární komponentu, která se mění v čase a neopakuje nebo jen minimálně opakuje v časovém rozpětí postiženém sledovanými daty. Lze také vysledovat obdobnou povahu jevu, která se systematicky opakuje v časových intervalech, což označujeme jako sezónnost. Tyto dvě základní složky ČŘ mohou koexistovat u dat povahy každodenního života.

V praxi se můžeme setkat s větším množstvím typů modelů – modelování. Při vlastním modelování jde zjednodušeně o to „dobře odhadnout“ jednotlivé složky tak, aby vytvořený model postihoval data co nejlépe a byl přitom vhodný pro interpretaci. Metody, které nám slouží k vytváření modelů, se liší jak do náročnosti, tak do přesnosti.

K nejjednodušším, ovšem málo spolehlivým, metodám tvorby modelu ČŘ (také vyrovnání ČŘ) patří *grafické* či *mechanické vyrovnání ČŘ*. S tím se však spokojíme pouze pro první hrubý odhad. Více informací nám dá teprve *analytické vyrovnání*, kdy model (jeho jednotlivé složky) popisujeme pomocí jednoduché teoretické analytické funkce. S takovouto funkcí se snáze dále pracuje – hledání periodicity, stanovení trendu, odhadu dalšího vývoje sledovaného jevu apod.

Grafické vyrovnání. Tato metoda spočívá v tom, že se lomená čára ve spojnicovém grafu naměřených hodnot „zkusmo nahradí“ vyrovnávající čarou. Tento způsob je jen velice přibližný a většinou nepřijatelný.

Mechanické vyrovnání (neboli vyrovnání časové řady **metodou klouzavých průměrů**). Touto metodou se každá empirická hodnota nahradí novou hodnotou, které se vypočítá jako aritmetický průměr z určitého počtu hodnot, které ji bezprostředně předcházejí a stejného počtu hodnot, které ji bezprostředně následují. Tyto nové hodnoty vytvářejí novou ČŘ, jejíž průběh je plynulejší a snáze interpretovatelný. Stanovení počtu hodnot, které mají být zahrnuty do vypočítávaného průměru, není jednoznačně definováno a záleží více na povaze dat – především pak na jejich celkovém počtu. Je nutné si uvědomit, že čím více původních hodnot zde bude zahrnuto, tím bude nová ČŘ plynulejší, ale zároveň „ořízнuta“ o krajní hodnoty (viz následný příklad).



Obrázek (4.1)

Obrázek (4.1) naznačuje postup při vytváření nové ČŘ metodou klouzavých průměrů. V grafu je zakresleno původních 20 empirických hodnot a dále obdélníky 1, 2 a 3 (pro přehlednost pouze tři; ve skutečnosti by měly být obdélníky zakresleny až do konce ČŘ), které naznačují, ze kterých empirických hodnot budou hodnoty nové ČŘ sestrojeny, resp. ze kterých bude počítán aritmetický průměr. Následné vzorce naznačují, jakým způsobem vypočítat hodnoty nové ČŘ (y'_1, y'_2, y'_3, …, y'_{20}).

$$y'_1 = \frac{\sum y_i}{4} \qquad y'_2 = \frac{\sum y_{i+1}}{4} \qquad y'_3 = \dots \qquad i = 1, 2, 3, 4$$

V tomto případě bylo pro výpočet každé nové hodnoty použity vždy čtyři empirické hodnoty. Lze se takto ptát, kolik je oněch několik zmiňovaných hodnot před a po původní hodnotě, když se jedná o sudý počet (4). V tomto případě nelze hovořit o hodnotách před a po původní empirické hodnotě, neboť zde není žádná „prostřední hodnota“. Tento způsob je rovněž možný. Kterou empirickou hodnotu původní časové řady v našem případě nová hodnota y'1 nahrazuje? Jedná se o y1? Ve skutečnosti by bylo možné novou hodnotu „zaznačit“ mezi hodnoty y'2 a y'3.

Proč zde bylo stanoven počet právě 4 původních empirických hodnot? Vzhledem k počtu hodnot v souboru (n = 20) lze toto číslo považovat za odpovídající (maximální) s přihlédnutím k faktu, že nová ČŘ takto mít o 4 hodnoty méně - počet hodnot se sníží na 4/5. Tip: Zkuste se zamyslet, kolik bude mít nová ČŘ hodnot v případě užití tří hodnot pro výpočet průměru.

Úkol 2 Vyrovnání ČŘ metodou klouzavých průměrů

Zadání: Následná tabulka zahrnuje údaje vyjadřující počet dětí, které v průběhu jednoho roku a tří měsíců umřely do tří týdnů po porodu v Indické provincii Džamkhed. Určete typ ČŘ a metodou klouzavých průměrů sestavte novou ČŘ. Obě řady zakreslete do jednoho grafu. Pokuste se charakterizovat, popsat vlastnosti ČŘ.

Tabulka (4.1)

rok	měsíc (t)	počet (y_t)
2001	květen	14

	červen	2
	červenec	1
	srpen	3
	září	1
	říjen	2
	listopad	6
	prosinec	7
2002	leden	8
	únor	9
	březen	8
	duben	11
	květen	16
	červen	3
	červenec	4

Řešení: **A)** Prvním úkolem bylo rozeznat typ časové řady. Jak víme, existuje několik typů členění. Uvedenou ČŘ lze považovat za intervalovou, absolutních ukazatelů, naturálních ukazatelů a ekvidistantní (i přes to, že každý měsíc má zde jiný počet dní, budeme s řadou pro jednoduchost pracovat jako s ekvidistantní; jako časovou jednotku budeme uvažovat pouze měsíc).

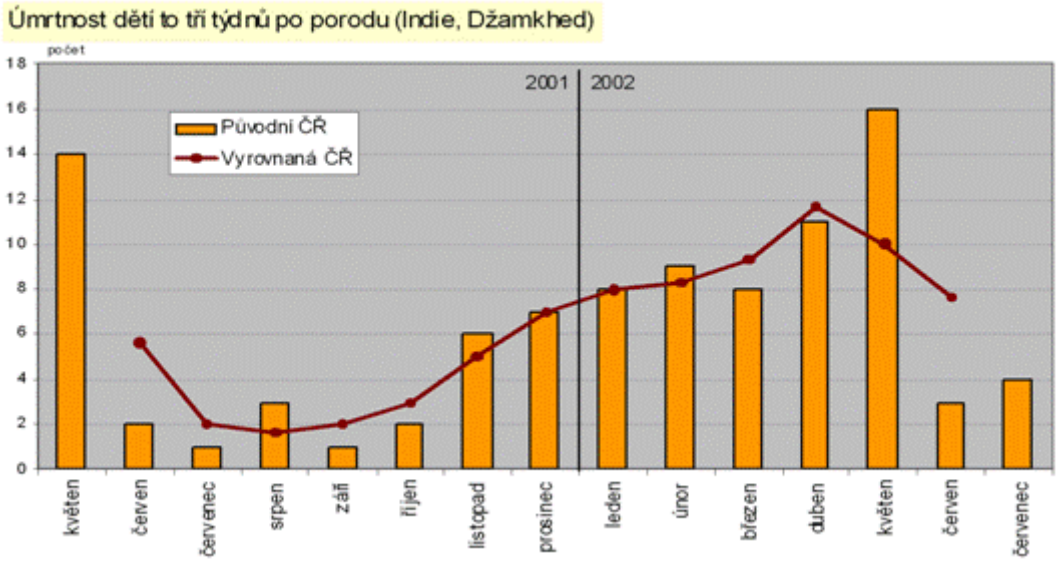
B) Vyrovnání ČŘ metodou klouzavých průměrů. Vzhledem k rozsahu řady (t = 15) si „nelze dovolit“ zahrnout vysoký počet hodnot pro výpočet průměru, neboť by došlo relativně k vysoké ztrátě dat. Budeme tedy volit počet k = 3. Výpočet lze přehledně zaznačit následnou tabulkou (tabulka XY).

Tabulka (4.2)

Původní řada empirických hodnot			**Nová ČŘ**		
měsíc (t)	**počet (y_t)**		**t'**	**výpočet**	**y'_t**
květen	14		1	(14+2+1)/3	17/3
červen	2		2	(2+1+3)/3	2
červenec	1		3	(1+3+1)/3	5/3
srpen	3		4	(3+1+2)/3	2
září	1		5	(1+2+6)/3	3
říjen	2		6	(2+6+7)/3	5
listopad	6		7	(6+7+8)/3	7
prosinec	7		8	(7+8+9)/3	8
leden	8		9	(8+9+8)/3	25/3
únor	9		10	(9+8+11)/3	7
březen	8		11	(8+11+16)/3	15/3
duben	11		12	(11+16+3)/3	10
květen	16		13	(16+3+4)/3	23/3
červen	3				

červenec	4				

C) Grafické znázornění původní a vyrovnané ČŘ. Vzhledem k povaze ČŘ je vhodným typem sloupcový graf (viz intervalová ČŘ).



Obrázek (4.2)

Jak je z obrázku XY patrné, je nová ČŘ plynulejší a výchylky v jednotlivých měsících tedy nejsou již tak výrazné. Z průběhu křivky lze vyčíst zřetelný pokles úmrtnosti v letních měsících a maximální úmrtnost v jarních měsících. To svým charakterem odpovídá sezónní složce ČŘ.

Při hlubším studiu ČR si většinou s předešlým modelem nepostačíme a použijeme jej pouze k prvotní informaci. Mnohem více informací a také možností (především pro predikci dalších hodnot) skýtá model definovaný analytickou funkcí. Její stanovení je již náročnější, nicméně řada programů stanovení modelu umožňuje.

Povahou analytické funkce vymezujeme model ČŘ (typ trendu). Mezi nejužívanější můžeme jmenovat následující:

- Lineární trend

$$\left(y_t = a_0 + a_1 \cdot t\right)$$

Vzorec (4.4)

- Kvadratický trend

$$\left(y_t = a_0 + a_1 \cdot t + a_2 \cdot t^2\right)$$

Vzorec (4.5)

- Polynomický trend $\left(y_t = a_0 + a_1 \cdot t + \ldots + a_k \cdot t^k\right)$ Vzorec (4.6)

  - hodnota *k* označuje tzv. *řád polynomického trendu*
- Exponencionální trend; Logistický trend, …

KORELACE ČŘ

Zatím jsme se zabývali jednotlivými řadami jako takovými, tj. nechali jsme stranou vztah dané časové řady k jiné časové řadě. Praxe však často vznáší otázky týkající se vztahu mezi jevy. Jedná-li se o jevy, které mají vazbu na časovou osu (jev se mění v čase) a tedy jedná se tedy o ČŘ, pak vzniká otázka, zda lze tak jako v dřívější kapitole věnující se korelaci a regresi využít podobných metod k postižení závislostí mezi jevy.

Závislostí ČŘ se věnuje korelace časových řad. Tato se od korelace běžných jevů (které byly diskutovány v dřívějších kapitolách) liší povahou statistických jednotek – zatímco u běžných prostorových řad jsou statistickými jednotkami osoby, zvířata, předměty apod., u ČŘ plní funkci statistických jednotek jednotlivé časové úseky a jejich uspořádání je přirozeně stanoveno. Rozdíl od prostorových řad je třeba chápat v tom, že zde nejsou změny způsobeny inter- či intra-individuálním srovnáním, ale vývojem jevu v čase.

Obecně je korelace definována jako obecné označení pro statistickou vazbu (závislost) mezi dvěma či více znaky ve statistickém souboru nebo mezi dvěma či více náhodnými veličinami. Cílem je objasnit závislost dvou jevů probíhajících v čase, resp. určit vliv latentních činitelů. Pro zjištění korelace mezi proměnnými x a y (jako zástupci dvou jevů X a Y) se používá nejčastěji korelační koeficient r_{xy}. Jeho vlastnosti lze shrnout do několika bodů:

- Korelační koeficient r_{xy} nabývá kladných, záporných hodnot nebo je roven nule.
- Korelační koeficient r_{xy} je roven nule právě tehdy, jsou-li jsou proměnné zcela nezávislé.
- Korelační koeficient r_{xy} je kladný právě tehdy, když je mezi proměnnými přímá závislost.
- Korelační koeficient r_{xy} je záporný právě tehdy, když je závislost proměnných nepřímá.
- Korelační koeficient $r_{xy} = -1$ (nebo 1) právě tehdy, když závislost proměnných lze vyjádřit lineární funkcí.

Aplikace časových řad ve sportu

Asi by bylo obtížné hledat obor v oblasti sportu, kde by ČŘ neměly své postavení a modelování ČŘ využití. Popisovat dynamiku „sportovního“ jevu je pro zainteresované osoby mnohem významnější, než pouhý statický popis souboru (jako jsou střední hodnoty, odchylky apod.). Analyzovat průběh sportovní výkonnosti, efektivity (něčeho), spotřeby (něčeho), … s přesahem k predikci budoucích hodnot je robustním nástrojem v analýze dat a jejich statistickém zpracování. Pro představu užití viz následný přehled užití ČŘ ve sportu.

- sportovní výkonnost v průběhu roku (sledování periodicity; načasování výkonnosti na závod; zpětný rozbor užitých metod, …)
- analýza fyziologických ukazatelů ve vztahu k souběžným faktorům tréninku
- sledování úrovně variability srdeční frekvence v průběhu celoročního tréninkového cyklu a posuzování ve vztahu k aktuálnímu zatížení; kontrola kratších tréninkových cyklů.
- predikce budoucích hodnot výkonů (celé populace) na podkladě dlouhodobého sledování
- sledování dlouhodobých (longitudinálních) trendů motorické výkonnosti a zdatnosti obecné populace

PŘÍKLADY K PROCVIČENÍ:

KE STAŽENÍ PŘÍKLADY K PROCVIČENÍ V MS EXCEL

Příklad 1 Určete typ ČŘ, vypočtěte aritm. průměr, zakreslete graf a odhadněte trend.

V letech 1991-1998 byly v běhu na 100m u TO zjištěny tyto hodnoty výkonů: 14,5; 14,1; 13,9; 13,8; 13,2; 13,0; 13,0; 12,9.

Příklad 1 Určete typ ČŘ, vypočtěte aritm. průměr, zakreslete graf a odhadněte trend.

V průběhu roku byla zjištěna u TO tato účast na tréninku (údaje jsou v jednotlivých měsících): 20; 18; 21; 14; 19; 18; 20; 21; 3; 6; 12; 19.

Příklad 3 Určete typ ČŘ, spočtěte chronologický průměr, znázorněte obě časové řady graficky a proveďte vyrovnání časových řad metodou klouzavých průměrů.

Individuální výsledky longitudinálního sledování síly levé ruky dvou hráček tenisu v letech 1996 až 1999 (xi: 18,6; 24,9; 28,0; 27,4; 28,5; 28,1 / yi: 15,2; 19,2; 20,8; 23,5; 26,7; 28,1)