



evropský
sociální
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



UNIVERSITAS
OSTRAVIENSIS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

TENTO PROJEKT JE SPOLUFINANCOVÁN EVROPSKÝM SOCIÁLNÍM FONDEM A STÁTNÍM ROZPOČTEM ČESKÉ REPUBLIKY.

KVANTITATIVNÍ METODY V PEDAGOGICKÉM VÝZKUMU

RADEK KRPEC



PODPORA TERCIÁRNÍHO VZDĚLÁVÁNÍ
STUDENTŮ SE SPECIFICKÝMI
VZDĚLÁVACÍMI POTŘEBAMI
NA OSTRAVSKÉ UNIVERZITĚ V OSTRAVĚ

CZ.1.07/2.2.00/29.0006

OSTRAVA, ČERVEN 2013

Studijní opora je jedním z výstupu projektu ESF OP VK.

Číslo Prioritní osy:	7.2
Oblast podpory:	7.2.2 – Vysokoškolské vzdělávání
Příjemce:	Ostravská univerzita v Ostravě
Název projektu:	Podpora terciárního vzdělávání studentů se specifickými vzdělávacími potřebami na Ostravské univerzitě v Ostravě
Registrační číslo projektu:	CZ.1.07/2.2.00/29.0006
Délka realizace:	6. 2. 2012 – 31. 1. 2015
Řešitel:	<u>PhDr. Mgr. Martin Kaleja, Ph.D.</u>

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

Název: Kvantitativní metody v pedagogickém výzkumu
Autor: Radek Krpec

Studijní opora k inovovanému předmětu: *Metody a techniky pedagogického výzkumu (KMD/2MTPD, KMD/2MTPE)*

Jazyková korektura nebyla provedena, za jazykovou stránku odpovídá autor.

Recenzent: *RNDr. Renáta Zemanová, Ph.D.*
Katedra matematiky s didaktikou PdF OU

© Radek Krpec
© Ostravská univerzita v Ostravě
ISBN 978-80-7464-445-0

POUŽITÉ GRAFICKÉ SYMBOLY



Průvodce studiem



Cíl kapitoly



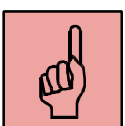
Klíčová slova



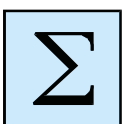
Čas na prostudování kapitoly



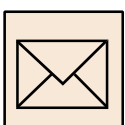
Kontrolní otázky



Pojmy k zapamatování



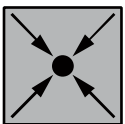
Shrnutí



Korespondenční úkol



Doporučená literatura



Řešený příklad



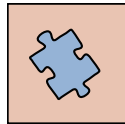
Otázky k zamyšlení



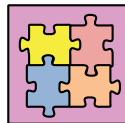
Část pro zájemce



Úlohy k textu



Testy a otázky



Řešení a odpovědi

Obsah

Slovo úvodem	7
1 Úvod do statistiky, základní pojmy a metody statistiky	8
1.1 Statistika a statistické šetření	8
1.2 Základní pojmy statistiky	10
1.3 Metody statistiky	11
<i>Shrnutí kapitoly</i>	15
2 Elementární statistické zpracování	17
2.1 Tabulky četností	17
2.2 Grafy	19
<i>Shrnutí kapitoly</i>	23
3 Empirické charakteristiky	25
3.1 Charakteristiky polohy	26
3.2 Charakteristiky variability	30
3.3 Charakteristiky šikmosti a špičatosti	32
3.4 Výpočet základních charakteristik v programu OpenOffice	34
<i>Shrnutí kapitoly</i>	39
4 Ověřování hypotéz	40
4.1 Statistické metody používané při ověřování hypotéz	41
4.2 Statistické testy	47
4.3 Parametrické testy	49
4.4 Neparametrické testy	67
<i>Shrnutí kapitoly</i>	81
5 Regrese a korelace	83
5.1 Regresní analýza	85

5.2	Korelační analýza	89
	<i>Shrnutí kapitoly</i>	92
6	Přehled použité literatury	94

Slovo úvodem

Tento učební text je určen především pro studenty oboru učitelství pedagogiky pro střední školy k předmětu Metody a techniky pedagogického výzkumu. Vzhledem k tomu, že se jedná o studenty nematematického zaměření, je tak pojato i zpracování tohoto textu. Studenti nematematických oborů potřebují ke svému výzkumu znát metody zpracování, aniž by museli umět matematické pozadí a matematickou teorii. Pedagogický výzkum můžeme rozdělit na dvě části, kde vedle kvantitativního výzkumu vystupuje ještě výzkum kvalitativní. V tomto učebním textu se zabýváme pouze kvantitativními metodami v pedagogickém výzkumu.

Seznámíme se zde se základními metodami statistického zpracování získaných dat. Na množství ukázek a příkladů mohou sledovat jednotlivé metody bez nutnosti znát matematické postupy. Veškeré postupy jsou uvedené s využitím základních programů jako OpenOffice, MS EXCEL, apod. Výhodou je, že se nemusíme učit žádné vzorečky a algoritmy zpracování. Nevýhodou je, že na rozdíl od matematických postupů, které jsou neměnné, se musí čas od času postupy v tomto textu upgradovat tak, jak vznikají nové verze výpočetních produktů.

V tomto učebním textu se seznámíme s elementárním statistickým zpracováním, základními popisnými charakteristikami. Nejdůležitější částí je ověřování statistických hypotéz. Seznámíme se zde s postupy pro ověřování hypotéz týkající se nominálních, ordinálních a metrických dat.

Věřím, že Vám tato opora usnadní studium tohoto náročného předmětu

Autor

1 Úvod do statistiky, základní pojmy a metody statistiky



Cíl kapitoly

Po nastudování této kapitoly byste měli být schopni:

- formulovat, čím se zabývá statistika,
- popsat statistické šetření a jeho etapy,
- definovat základní statistické pojmy,
- rozlišovat metody statistiky,
- tvořit výběrový soubor,
- tvořit škálu měření.



Klíčová slova: statistika, statistické šetření, statistická jednotka, statistický soubor, základní soubor, výběrový soubor, statistický znak (proměnná), metody statistiky, statistická indukce, náhodný výběr, škálování, měření



Čas na prostudování kapitoly

Průvodce studiem



Tato kapitola je teoretického charakteru, musíte se v něm seznámit se základními pojmy, se kterými se budete setkávat po celé studium statistických metod užívaných v pedagogickém výzkumu.

1.1 Statistika a statistické šetření

Statistika hraje v pedagogickém výzkumu nezastupitelnou roli. Jde o kvantitativní pedagogický výzkum, který je celý postaven na statistických šetřeních.

Co to vlastně ta statistika je?

Statistika je část matematiky, která se zabývá, podobně jako teorie pravděpodobnosti, analýzou zákonitostí hromadných náhodných jevů. Náhodným jevem nazveme jakýkoli výsledek náhodného pokusu, tedy pokusu, jehož výsledek není možné na základě předem daných podmínek stanovit. Jev nazveme hromadným, jestliže jej můžeme pozorovat na dostatečném množství objektů nebo jej můžeme dostatečně krát opakovat. Z teoretického hlediska je hromadný jev takový, který se může pozorovat na nekonečně mnoho objektech nebo jej můžeme nekonečně krát opakovat.



Na rozdíl od teorie pravděpodobnosti pracuje statistika s reálnými experimentálními údaji, tzv. *empirickými daty*. Výsledky získané statistickými ověřeními na těchto empirických datech nazýváme empirické.



Tyto empirické výsledky získáme na základě tzv. *statistického šetření*. Statistické šetření má celkem 4 etapy:



1. Sestavení plánu šetření;
2. Sběr statistických dat;
3. Statistické zpracování dat;
4. Rozbor získaných výsledků.

1. etapa – *sestavení plánu šetření* je jednou z nejdůležitějších etap statistického šetření. Do plánu šetření patří stanovení velikosti vzorku, použitých metod výběru vzorku, použitých metod sběru dat, použitých metod zpracování dat, apod. Jak vidíme, nesprávné stanovení kteréhokoliv parametru může vést k tomu, že celý výzkum, ač většina ostatních parametrů bude provedena správně, povede ke špatným závěrům.

2. etapa – na základě plánu šetření budeme mít stanoveno, jakým způsobem provedeme *sběr statistických dat*. Ke sběru dat slouží různé techniky jako např. pozorování, experiment, dotazník, rozhovor, apod.

3. etapa – na základě získaných empirických dat provedeme *statistické zpracování dat*, zde patří nejenom provedení příslušných statistických výpočtů, ale taktéž předcházející příprava získaných dat ke zpracování, a také shrnutí získaných statistických výsledků.

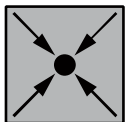
4. etapa – získané výsledky pak srovnáváme s dalšími výsledky podobných šetření, vyvozujeme nové závěry, hypotézy, apod.

1.2 Základní pojmy statistiky

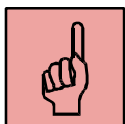
V této části se seznámíme se základními pojmy statistiky, se kterými se budeme setkávat na každém kroku tohoto textu.



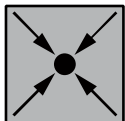
Objekty, se kterými budeme v rámci statistických šetření pracovat, nazýváme statistické jednotky. *Statistické jednotky* tvoří rozsáhlou množinu prvků, která má některé vlastnosti zcela shodné, a u které jsou zkoumány některé další vlastnosti, které jsou proměnlivé.



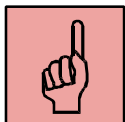
Příklady statistických jednotek mohou být žáci, třídy, školy, události, územní jednotky, občané, občanská seskupení, apod.



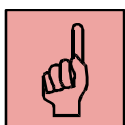
Seskupíme-li statistické jednotky stejného druhu, získáme soubor jednotek, který nazýváme *statistickým souborem*.



Statistické soubory mohou tvořit např. všichni žáci jedné třídy, všichni žáci dané školy, všichni žáci v určitém kraji, všichni učitelé českého jazyka v určitém kraji, úředníci daného útvaru, účastníci dané aktivity, apod.



Počet všech jednotek určitého statistického souboru nazýváme jeho *rozsahem*.



Statistické soubory dělíme do dvou skupin podle toho, jaké jednotky obsahují. Pokud daný statistický soubor obsahuje všechny jednotky, které zkoumáme, pak tento statistický soubor nazýváme *základním*. V případě, že nemáme k dispozici všechny jednotky základního souboru, ale pouze jeho určitou část, pak tuto část základního souboru nazýváme *výběrovým souborem*. Jak uvidíme v dalším textu, z hlediska výzkumu budeme mít zpravidla k dispozici soubor výběrový.

Jak jsme již dříve uvedli, statistické jednotky tvoří skupinu objektů, které mají některé vlastnosti stejné a v některých se liší. Těmto vlastnostem říkáme *statistické znaky* nebo také *statistické proměnné*.

Statistické proměnné (znaky) dělíme podle toho, jak jsou vyjádřeny, na



- *kvalitativní* - varianty statistické proměnné jsou vyjádřeny zpravidla slovně, výjimkou jsou číselná vyjádření, kdy se jedná ale pouze o identifikační charakter jako např. číslo domu,
 - *dichotomické* – nabývají pouze dvou variant, např. pohlaví, otázka s možnostmi ano-ne, apod.
 - *polytomické* – nabývají více než dvou variant, např. jméno, bydliště, okres, apod.
- *kvantitativní* – jsou vyjádřeny číslly, charakterizují zpravidla četnost nebo úroveň sledované vlastnosti, tyto dále dělíme na
 - *pořadové* – jednotky dle tohoto znaku seřadit, ale nemůžeme mezi nimi určit rozdíly (vzdálenosti), např. pořadí v cíli, stupeň vzdělání
 - *metrické* – můžeme určovat rozdíly mezi úrovněmi, např. věk, výška, IQ
 - diskrétní (věk v letech)
 - spojité (čas, rozměr měnící se spojitě)

1.3 Metody statistiky

N úvod této kapitoly se zkusme podívat, čím se liší následující dvě statistická vyhodnocení.



a) V 8. A psali žáci test z českého jazyka. Po ohodnocení určila paní učitelka aritmetický průměr a směrodatnou odchylku ze všech výsledků.

b) Ve sto 8. třídách náhodně vybraných ze všech osmých tříd České republiky psali žáci test z českého jazyka za účelem zjištění současného stavu znalostí u všech žáků v České republice. Na základě statistického zpracování byl určen průměrný výsledek a směrodatná odchylka žáků testovaných tříd a tyto byly vzaty jako průměrný výsledek a směrodatná odchylka výsledků všech žáků 8 tříd.

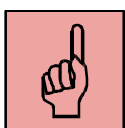
Jaký je rozdíl mezi oběma statistickými vyhodnoceními?

V případě a) pracujeme s výsledky žáků, které máme jako s celkem. Mám k dispozici všechny údaje potřebné k vyhodnocení.

V případě b) máme výsledky pouze od určité skupiny žáků náhodně vybraných ze všech žáků ČR. Znamená to, že nemáme k dispozici výsledky od všech žáků České republiky, ale snažíme se je odhadnout pomocí výsledků určité vybrané skupiny.



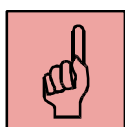
Na těchto dvou příkladech jsme si ukázali dvě základní metody statistického vyhodnocování naměřených výsledků. První metodě se říká *popisná (deskriptivní) statistika*. V tomto případě máme k dispozici všechny potřebné informace o všech jednotkách zkoumaného souboru.



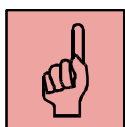
Druhou metodou je *statistická indukce (výběrová statistika)*. Zabývá se možnostmi zobecnění výsledků získaných zkoumáním výběrového souboru na soubor, z něhož výběr pochází. Umožňuje dělat závěry o charakteristikách, struktuře a vlastnostech základního souboru na základě předem určených charakteristik, struktury a vlastností výběru. Toto je silnější část statistiky, umožňuje nám dělat mnohem širší závěry, aniž bychom museli mít k dispozici všechny vyšetřované objekty, což většinou není možné.

Jedním z kroků vyhodnocování v rámci statistické indukce je určit ten pravý výběrový soubor. Metod, jak vybrat takovýto soubor je celá řada a my se zde seznámíme s několika z nich.

Základní typy výběrových souborů:



- *Náhodný výběr* – každá jednotka základního souboru má stejnou možnost být vybrána. Patří mezi nejjednodušší metody výběru. Mezi nejčastější procedury, jak vytvořit náhodný výběr, patří:
 - losování,
 - generátory pseudonáhodných čísel (najdeme je v jakémkoli statistickém softwaru, tabulkových procesorech jako např. MS EXCEL, OPENOFFICE, apod.)
 - tabulky náhodných čísel.

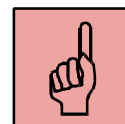


- *Záměrný výběr* – znalec vybírá takové jednotky, které umožňují nejlepší realizaci zamýšleného statistického šetření.
- *Stratifikovaný výběr* – jde o zvláštní typ náhodného výběru. Jednotky

základního souboru nejdříve rozdělíme podle vhodného statistického znaku do skupin, a z těchto skupin pak náhodně vybereme stanovený počet těchto jednotek. Zpravidla z každé skupiny vybereme stejný počet jednotek, ale nemusí tomu tak být vždy.

Příklad: Školy české republiky rozdělíme do skupin podle krajů a z každého kraje pak vybereme určitý počet škol.

- *Kontrolovaný výběr* – jde o zvláštní typ stratifikovaného výběru. Počet jednotek v jednotlivých skupinách výběru je úměrný počtu jednotek v odpovídajících skupinách základního souboru.



Příklad: Jestliže v základním souboru žáků základních škol bude např. dvakrát více dívek než chlapců, pak náhodně vybíráme určitý počet chlapců a ze všech dívek vybíráme dvojnásobek počtu chlapců rovněž náhodně.

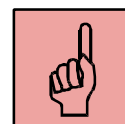
Nyní již víme, jak výběrový soubor můžeme vytvořit. Na co dalšího se budeme v této chvíli ptát? Když jsme probírali statistické šetření, jednou z otázek bylo, jak velký soubor budeme ke zpracování potřebovat. Jak velký výběrový soubor budeme tedy ke zpracování potřebovat, kolik by měl tento výběrový soubor obsahovat statistických jednotek? Na základě čeho se budeme rozhodovat?

Základním parametrem pro určení rozsahu výběrového souboru je rozsah základního souboru. Existuje několik vzorců pro určení rozsahu výběrového souboru. Často se používají meze:

$$n_{\min} = 0,1 \cdot \sqrt{n},$$

$$n_{\max} = \sqrt{n}.$$

Kde n je rozsah základního souboru, který také nazýváme také pojmem *populace*.



Musíme mít však na paměti, že jsou to pouze pomocné meze, vždy záleží na typu výzkumu. Stále bude platit, že čím více budeme mít ve výběru jednotek, tím se budou výsledky získané měřením na výběrovém souboru od výsledků získaných měřením potenciálního celého základního souboru při dodržení pravidel výběru lišit méně. Musíme mít na paměti, že pro některá podrobnější vyšetření se nesmíme řídit výše uvedenými mezemi, záleží na zvolených škálách a metodách.

Než se začneme zabývat vlastním statistickým vyhodnocováním, musíme se ještě seznámit s dvěma pojmy týkajícími se statistického šetření, jsou to škálování a měření. Prvním z nich je škálování. Všichni už jsme něco v životě měřili. Nejčastěji nějaký rozměr a víme, že podle toho, co měříme, volíme měřidlo. Když budeme měřit rozměr nábytku, budeme měřit s přesností na milimetry. Když měříme vzdálenosti měst, tak zpravidla s přesností na kilometry (popř. míle, apod.). Když počítáme dlažbu na dvůr, tak budeme měřit s přesností na centimetry. Tím, jakým měřidlem měříme, si vlastně volíme určitou „škálu“, podle které měření provádíme.

Jiný příklad může vypadat takto. Budeme sledovat např. závislost vztahu ke škole na věku. Předpokládejme, že naši populaci tvoří žáci ve věku od 11 do 19 let. Při měření můžeme pracovat se škálou, kde jednotlivé varianty jsou dané věkové hladiny 11, 12, 13, ... , 18, 19 let. Ale pro naše vyhodnocení může být obtížné zpracování z tak velkým množstvím variant. Pak často seskupujeme jednotlivé varianty do skupin a tvoříme např. intervaly 11-13, 14-16, 17-19 let, tzn. máme celkem tři varianty proměnné „věk“.

Tuto etapu, kdy volíme, jak budou jednotlivé varianty sledované proměnné vypadat, nazýváme *škálování*. Množina variant sledované proměnné se pak nazývá *škálou*.



Typy škál:

- *Nominální škála* – o každých dvou jednotkách lze pouze rozhodnout, zda jsou co do sledovaného znaku stejné či rozdílné. Mezi proměnné, které mají nominální škálu, patří např. pohlaví, adresa bydliště, navštěvovaná škola, apod.
- *Ordinální škála* – v tomto případě navíc lze oproti nominální škále určit pořadí jednotlivých variant statistické proměnné. Zde patří např. školní prospěch, dosažený stupeň vzdělání, umístění v soutěži apod.
- *Intervalová škála* – umožňuje jednoznačně určit vzdálenost (rozdíl) mezi variantami sledovaného znaku. Můžeme určit, o kolik se dvě varianty od sebe liší. Např. mezi 20 a 10 cm je rozdíl 10 cm. Patří sem např. teplota ve stupních Celsia, hodnocení didaktických testů, zkoumané rozměry, apod.
- *Absolutní škála* – zvláštní případ intervalové škály, navíc zde můžeme určit nulu škály jakožto počátek, tzn. varianty proměnné nemohou nabývat hodnot pod

nulou škály. Patří sem např. teplota ve stupních Kelvina, věk, počet chyb v didaktickém testu, apod.

Když už máme vytvořenou škálu, můžeme se zabývat měřením. Měření má ve statistice mnohem širší význam, než jak jej známe z běžného života. *Měřením* rozumíme přiřazení jednotlivým statistickým jednotkám příslušné hodnoty škály. Pod pojmem měření rozumíme tedy i



- určení vystudované základní školy respondenta,
- zařazení respondenta do příslušné věkové skupiny,
- přiřazení žákům jejich výsledky z testu,
- atd.

Shrnutí kapitoly

- V této kapitole jsme si popsali základní principy statistického šetření.
- Seznámili jsme se se základními pojmy, které budeme v průběhu celého studia používat.
- Ukázali jsme si, s jakými soubory při statistickém šetření pracujeme.
- Metody statistiky jsme si rozdělili na dva základní typy.
- Ukázali jsme si, s jakými typy dat budeme pracovat a čím se tato data odlišují.



Kontrolní otázky a úkoly:

1. Jaké jsou etapy statistického šetření?
2. Co je to popisná statistika a statistická indukce?
3. Jaké znáte typy statistických proměnných a jak souvisí s typy škál?
4. Co je to výběrový soubor a jaké metody výběru znáte?



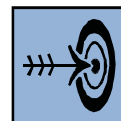


Korespondenční úkoly

- 1.** Definujte pět různých statistických proměnných. Ke každé statistické proměnné vytvořte alespoň tři různé škály.

2 Elementární statistické zpracování

Cíl kapitoly



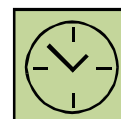
Po nastudování této kapitoly byste měli být schopni:

- přehledně zpracovat naměřené výsledky do tabulek a grafů,
- určit relativní a absolutní četnosti výskytu dané varianty sledované proměnné,
- vybrat vhodný typ grafu a výsledky do něj zaznamenat.

Klíčová slova: absolutní četnost, relativní četnost, četnostní tabulka, graf, sloupcový graf, histogram, výsečový graf, koláč, bodový graf, spojnicový graf.



Čas na prostudování kapitoly



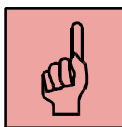
Průvodce studiem

V této kapitole seznámíte se základním shrnutím naměřených výsledků do jednoduchých výstupů jako jsou tabulky četností a grafy.



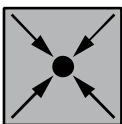
2.1 Tabulky četností

Nejjednodušší zpracování naměřených výsledků je jejich zpřehlednění do jednoduchých tabulek četností. Do těchto tabulek uvádíme k jednotlivým variantám měřené statistické proměnné (veličiny, statistického znaku) absolutní resp. relativní četnosti výskytu těchto variant v měřeném souboru. Tyto tabulky mohou být dále tvořeny dle určitých parametrů, např. podle pohlaví můžeme určovat zvlášť tabulku pro muže a zvlášť pro ženy. Než přejdeme k ukázkám, musíme si uvést základní pojmy.



Absolutní četnost výskytu dané varianty je počet všech jednotek, u kterých měřená proměnná nabývá dané varianty sledované proměnné.

Relativní četnost výskytu dané varianty je pak podíl absolutní četnosti a celkového rozsahu souboru. Relativní četnost se zpravidla udává v procentech.



Příklad

Ve třídě 30 žáků byly zjišťovány výsledky testu z dějepisu. Škála výsledků testu byla zvolena klasicky hodnotami známek od 1 do 5. Do tabulky 2.1 pak byly zaznamenány absolutní a relativní četnosti výskytu jednotlivých známek y testu z dějepisu.

Známka z testu z dějepisu	Počet (absolutní četnost)	Procentuální zastoupení (relativní četnost)
1	3	10,0 %
2	8	26,7 %
3	11	36,7 %
4	6	20,0 %
5	2	6,7 %
Celkem	30	100 %

Tabulka 2.1

Prohlédneme-li si podrobně tabulku, zjistíme, že součet relativních četností není 100 %, ale s tímto problémem se setkáme často. Nastává zaokrouhlováním a s tímto se musí počítat.

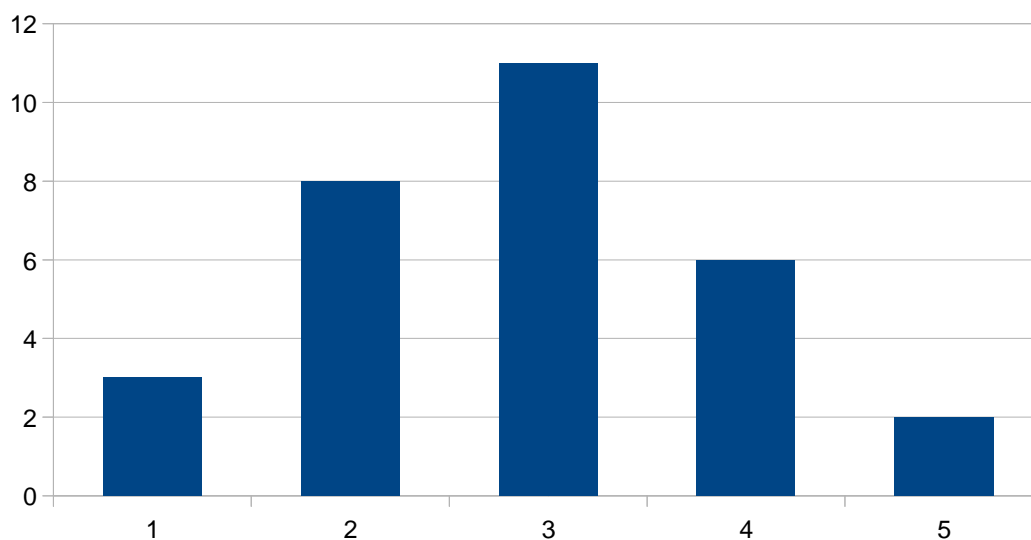
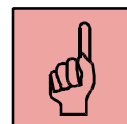
Tabulky můžeme tvořit s variantami sledovaného znaku seřazeného jak do sloupce, tak i do řádku, tzn., že tabulku četností můžeme mít vytvořenou i tak, že známky budou v prvním řádku a pod jednotlivými známkami pak budou absolutní a relativní četnosti.

2.2 Grafy

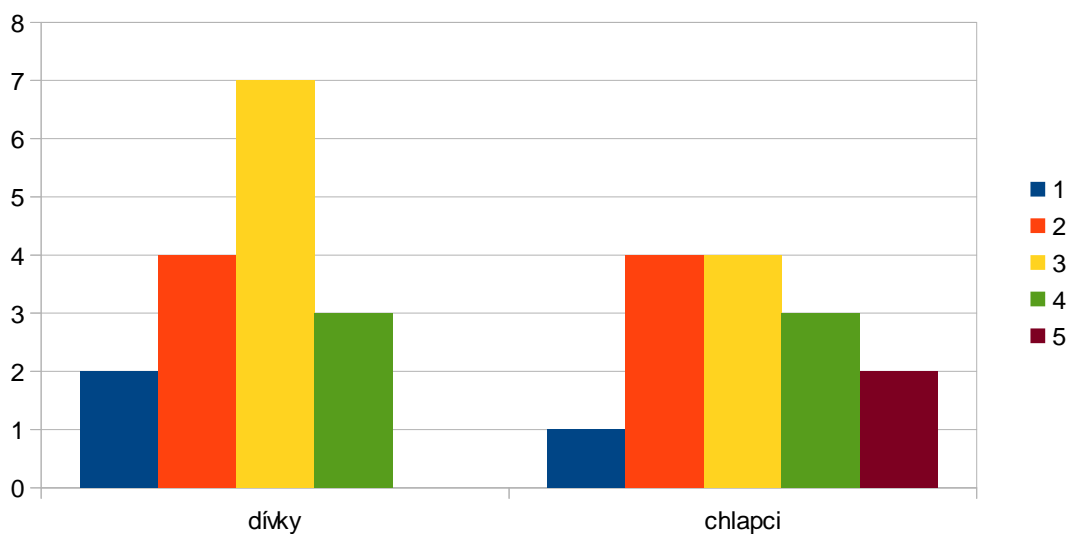
Dalším typem jednoduchého zpřehlednění naměřených údajů je vytvoření příslušného grafu. Těchto existuje nepřeberné množství, my si ukážeme pouze základní typy, jako jsou sloupcový graf, histogram, výsečový graf, spojnicový graf a bodový graf.

2.2.1 Sloupcový graf

Prvním typem je *graf sloupcový*, který používáme zpravidla k zaznamenání absolutních četností výskytů jednotlivých variant měřené statistické proměnné, viz obr. 2.1, na kterém nám výšky sloupců udávají absolutní četnosti jednotlivých známek z tabulky 2.1. Tento typ grafu je vhodný také v případech, pokud chceme srovnávat rozložení absolutních četností mezi různými skupinami, např. mezi chlapci a dívkami, viz obr. 2.2.



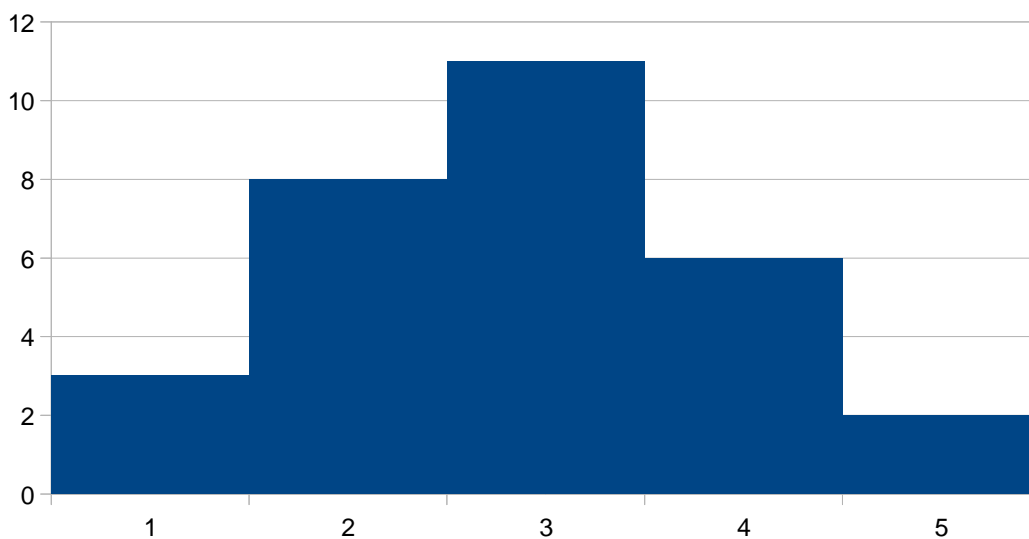
Obrázek 2.1



Obrázek 2.2



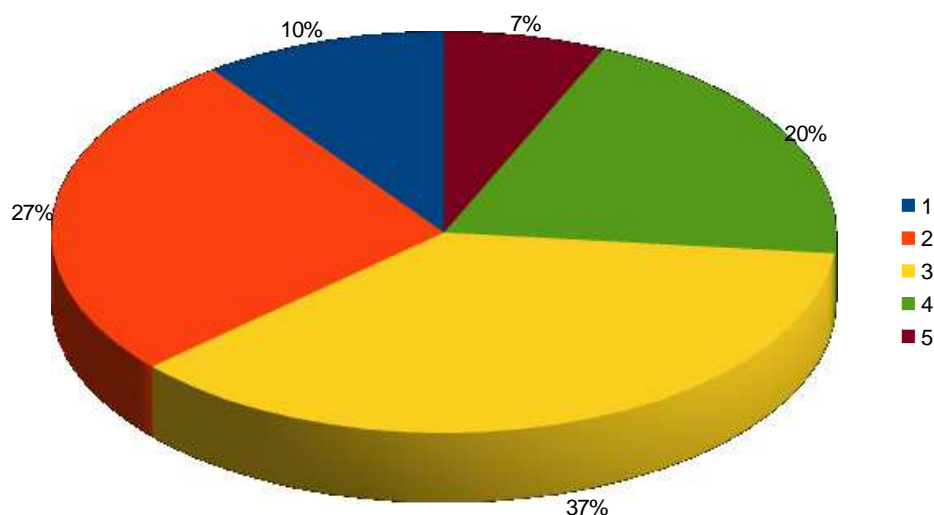
Zvláštním případem sloupcového grafu je *histogram*. Histogram je takový sloupcový graf, ve kterém se sloupce navzájem dotýkají, tzn., že šířka každého sloupce je přes celou jednu variantu sledované proměnné, jako to vidíme na obr. 2.3.



Obrázek 2.3

2.2.2 Výsečový graf - koláč

Dalším hojně používaným typem grafu je *graf výsečový*, který známe zpravidla pod pojmem „*koláč*“. Tento typ grafu se používá zejména pro zobrazení relativních četností výskytu jednotlivých variant sledované proměnné. Pokud vyjdeme z výsledků známek z testu z dějepisu zaznamenaných v tabulce 2.1, pak zaznamenání do výsečového grafu by mohlo vypadat jako na obrázku 2.4.

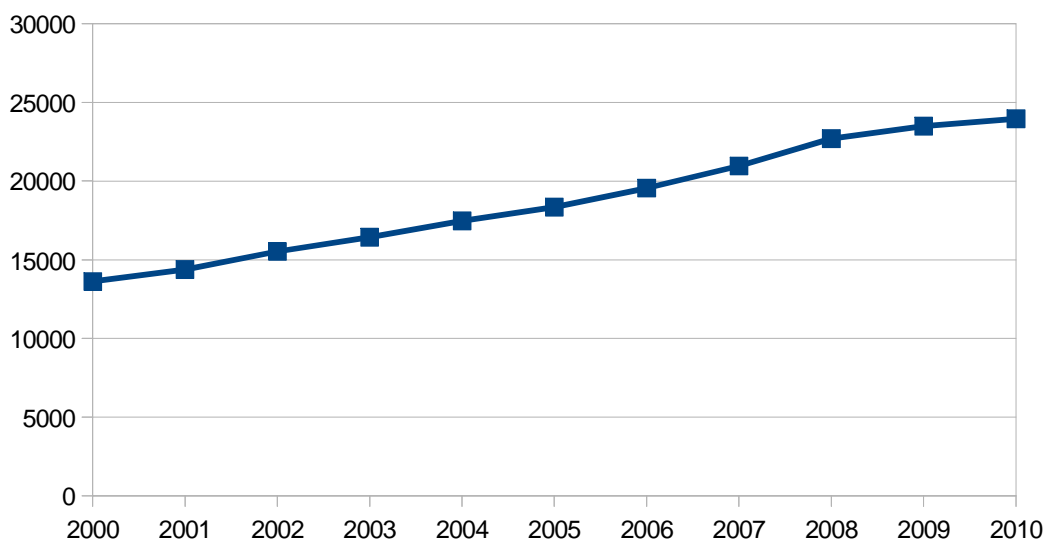


Obrázek 2.4

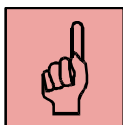
2.2.3 Spojnicový graf

Zatímco předchozí typy grafů byly soustředěny hlavně na četnosti výskytů jednotlivých variant sledované proměnné, další graf se může zaměřovat jak na četnosti, tak na úroveň sledované statistické proměnné. Tento typ grafu se zpravidla používá, chceme-li zobrazit vývoj nějaké proměnné v čase. Na obrázku 2.5 máme příklad grafu znázorňujícího vývoj průměrné mzdy v ČR v letech 2000—2010. Graf je tvořen tak, že na vodorovné ose zobrazíme jednotlivé časové etapy a ke každé etapě pak zobrazíme bod dle svislé osy a tyto body postupně spojíme úsečkami. V případě vývoje mezd nanese na vodorovnou osu jednotlivá léta 2000 až 2010 a ke každému roku pak zobrazíme průměrnou mzdu příslušné úrovně, která odpovídá v tom roce této mzdě, viz obr. 2.5.





Obrázek 2.5

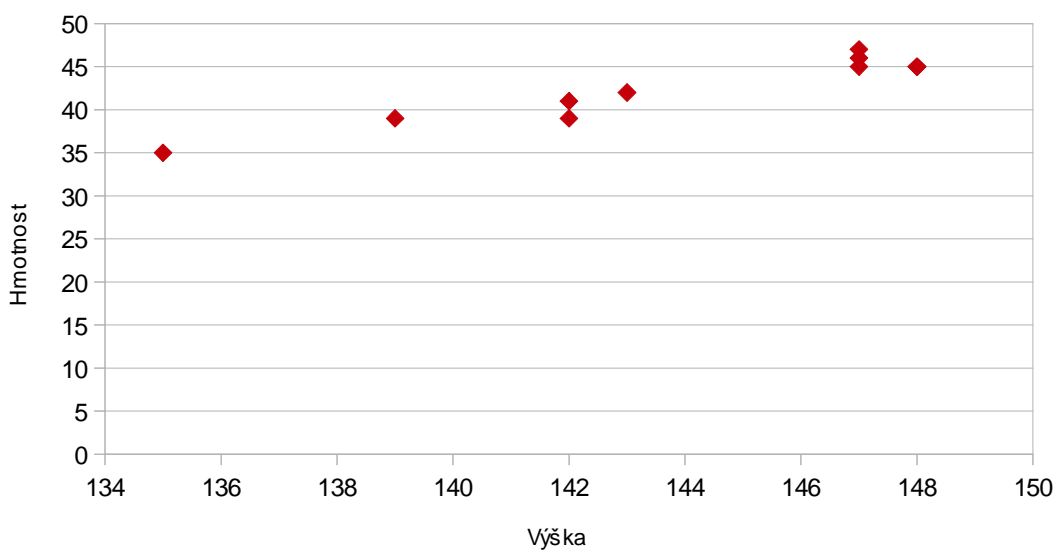


2.2.4 Bodový graf

Posledním grafem, který si v této kapitole uvedeme, je graf, ve kterém zpravidla zobrazujeme závislost dvou statistických proměnných. Tento typ grafu často používáme v souvislosti s regresí a korelací. Graf sestavujeme tak, že na jednu osu nanese hodnoty jedné proměnné a na druhou osu hodnoty druhé proměnné. Do grafu pak zaznamenáváme body, který na příslušných osách odpovídají naměřeným hodnotám proměnných. Jako ukázkou můžeme zvolit sledování závislosti výšky deseti dětí v určitém věku na jejich hmotnosti, viz obr. 2.6.

výška (cm)	142	148	135	147	148	143	139	142	147	147
hmotnost (kg)	39	45	35	46	45	42	39	41	47	45

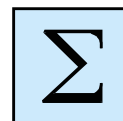
Tabulka 2.2



Obrázek 2.6

Shrnutí kapitoly

- Elementární statistické zpracování je jednoduché shrnutí do tabulek četností a grafů.
- Elementární zpracování je nejjednodušší zpřehlednění údajů.
- V kapitole jsme si uvedli nejpoužívanější typy grafů jako sloupcový, výsečový, spojnicový a bodový.



Kontrolní otázky a úkoly:

1. Co jsou to absolutní a relativní četnosti?
2. V jakém případě používáme sloupcový graf?
3. V jakém případě používáme výsečový graf?
4. V jakém případě používáme spojnicový graf?
5. V jakém případě používáme bodový graf?



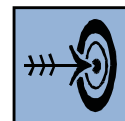


Korespondenční úkoly

1. V rámci výzkumu získejte statistická data a proveďte jejich elementární zpracování.

3 Empirické charakteristiky

Cíl kapitoly



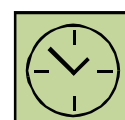
Po nastudování této kapitoly byste měli být schopni:

- rozdělit charakteristiky podle toho, jakou stránku souboru popisují,
- vyjmenovat základní charakteristiky polohy,
- rozeznat podmínky, kdy můžeme použít aritmetický průměr,
- určit, kdy je vhodné použít medián a modus,
- popsat, k čemu slouží charakteristiky variability,
- formulovat šikmost a špičatost.

Klíčová slova: empirické charakteristiky, charakteristiky polohy, charakteristiky variability, charakteristiky šikmosti a špičatosti, aritmetický průměr, medián, modus, rozptyl, směrodatná odchylka, mezikvartilové rozpětí, variační rozpětí, šikmost, špičatost.



Čas na prostudování kapitoly



Průvodce studiem

V této kapitole se seznámíme se základními empirickými charakteristikami a vhodnosti jejich použití při jednoduchém popisu vlastností statistického souboru.



Empirické charakteristiky jsou charakteristiky (míry), které stručně vystihují povahu zkoumaného statistického souboru. Jsou počítané z naměřených dat. Charakteristiky dělíme dvojím způsobem, jednak podle způsobu jejich konstrukce, jednak podle toho, jakou stránku statistického znaku postihují.



Dělení charakteristik:

Dělení charakteristik podle způsobu jejich konstrukce:

- momentové charakteristiky
- kvantilové charakteristiky
- jiné charakteristiky

Dělení charakteristik podle toho, jakou stránku statistického znaku postihují:

- charakteristiky polohy
- charakteristiky variability
- charakteristiky šikmosti a špičatosti

Pro naši další práci je důležitější druhé dělení, tedy podle toho, co popisují.

3.1 Charakteristiky polohy

Charakteristiky polohy nám udávají význačnou polohu v souboru naměřených dat.

Mezi nejvýznamnější charakteristiky polohy řadíme:

- Aritmetický průměr
- Medián
- Modus
- Minimum, maximum
- Decily, percentily

3.1.1 Aritmetický průměr



Mezi nejčastěji používanou charakteristiku polohy patří *aritmetický průměr*, taky někdy nazývaný jako střední poloha statistického souboru. Aritmetický průměr vypočteme tak, že sečteme všechny hodnoty a vydělíme počtem hodnot. Co to vlastně aritmetický

průměr je? Je to hodnota, vzhledem ke které jsou ostatní hodnoty přibližně rovnoměrně rozloženy nalevo a napravo od něj. Nemůžeme říct, že symetricky, protože většinou data nejsou symetricky rozložena.

Aritmetický průměr má smysl počítat pouze za určitých podmínek, což často nebývá dodržováno a pak bývá tento údaj zkreslující a nic nevyovídající. Základními podmínkami pro užití aritmetického průměru jako údaje udávající střední hodnotu souboru jsou:

- data jsou metrická,
- data mají přibližně normální rozdělení,
- data jsou přibližně symetricky rozložena,
- soubor neobsahuje extrémní hodnoty.

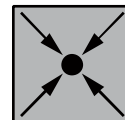
Příklad

Typickým příkladem nevhodného užití aritmetického průměru jsou mzdy. Rozložení mezd má málokdy symetrické rozdělení a zpravidla obsahuje extrémní hodnoty. Uvedme si jako příklad malou dílnu, která obsahuje 20 zaměstnanců a majitele (vedoucího). Zaměstnanci berou plat přibližně 19 000 Kč měsíčně a majitel má plat 100 000 Kč měsíčně. Jaký bude průměrný plat všech lidí v tomto malém podniku? Po jednoduchém výpočtu zjistíme, že zaokrouhleně 22 860 Kč. Rozdíl mezi skutečnou mzdou a průměrnou mzdou je téměř 4 000 Kč.

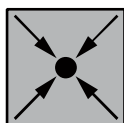
Častěji je vhodnější charakteristikou pro odhad mezd medián, kterému se budeme věnovat v následující kapitole.

3.1.2 Medián

Medián je takzvaná prostřední hodnota. Proč prostřední? Protože leží přesně uprostřed seřazené řady naměřených hodnot v souboru od nejmenší po největší. Definice nám tedy udává, jak medián jednoduše určíme. Vezmeme naše naměřené hodnoty, seřadíme je od nejmenší po největší a vybereme tu uprostřed. V případě, že máme lichý počet naměřených hodnot, tak je to hodnota přesně uprostřed. Pokud ovšem máme sudý počet hodnot, pak nemůžeme určit, která hodnota leží přesně



uprostřed, a musíme vzít dvě prostřední hodnoty, z těchto pak určíme průměr a to je medián daného souboru.



Příklad

Vzmemme-li si příklad se mzdami z kapitoly 3.1.1 Aritmetický průměr, pak medián určíme tak, že nejdříve seřadíme mzdy a pak určíme prostřední hodnotu.

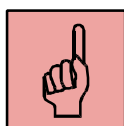
1	2	3	4	5	6	7	8	9
19 000	19 000	19 000	19 000	19 000	19 000	19 000	19 000	19 000
10	11	12	13	14	15	16	17	18
19 000	19 000	19 000	19 000	19 000	19 000	19 000	19 000	19 000
19	20	21						
19 000	19 000	100 000						

Medián určíme z tabulky tak, že si najdeme hodnotu uprostřed, což je hodnota na 11. místě, hodnota, která je v tabulce vyznačena červeně. Medián souboru mezd je tedy 19 000 Kč.

Výhodou mediánu je, že jej můžeme počítat i v následujících případech:

- Data jsou ordinální či metrická.
- Data obsahují odlehlé hodnoty.
- Rozdělení je silně zešikmené.

3.1.3 Modus



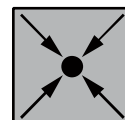
Modus je nejčastěji se vyskytující hodnota v souboru dat. Výhodou modu je, že je to jediná charakteristika, kterou můžeme použít i pro kvalitativní proměnné. Nevýhodou modu je, že nemusí být určen jednoznačně. Nejčastěji se vyskytujících hodnot se může vyskytovat v souboru více. V následujícím příkladu si ukážeme výpočet modu.

Příklad

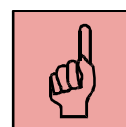
V souboru 30 žáků 8. třídy byly zjištěny následující výsledky testu z občanské výchovy.

Známka	1	2	3	4	5
Počet	3	7	12	4	4

V souboru se nejčastěji vyskytuje známka 3, celkem 12 krát, tedy známka 3 je *modem* souboru známek z testu z občanské výchovy.

**3.1.4 Minimum, maximum**

Dalšími dvěma důležitými charakteristikami jsou *minimální* a *maximální hodnota* souboru. Tyto dvě hodnoty nám uzavírají soubor zdola a shora. Můžeme si tak udělat obrázek, v jakém intervalu se nacházejí naše naměřené hodnoty.

**3.1.5 Decily, percentily**

Decily a percentily patří mezi tzv. kvantilové charakteristiky. S jednou kvantilovou charakteristikou jsme se již setkali. Byl to medián. Kvantilové charakteristiky jsou dány svou polohou v seřazeném souboru. Seřadíme-li tedy hodnoty vzestupně a bude nás zajímat, která hodnota se nachází na konci první desetiny souboru, pak se jedná o 1. decil, na konci druhé desetiny souboru se nachází 2. decil, atd. Pokud se hodnota na daném místě nenachází, pak vezmeme dvě hodnoty, které se kolem daného místa nacházejí, a určíme jejich průměr. *Decily* nám rozdělují soubor na desetiny.



Určování *percentilů* se provádí na stejném principu jako u decilů, akorát že v tomto případě se nám soubor rozdělí na setiny. Pokud tedy určíme v našem souboru 23. percentil, znamená to, že 23 % hodnot je menších než tento percentil a zbytek je větších. Příkladem použití percentilů v praxi jsou testy SCIO.



Medián jakožto prostřední hodnota je tedy 5. decil nebo také 50. percentil.

3.2 Charakteristiky variability

Charakteristikami polohy jsme si určili nějakou význačnou polohu v souboru, ale pro dobrý popis souboru to nestačí. Zatím nevíme další vlastnosti, především, jak jsou data rozložena vzhledem k této charakteristice polohy. Toto rozložení nám pomáhají popsat *charakteristiky variability*.

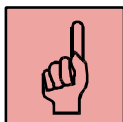


Mezi nejvýznamnější charakteristiky variability patří:

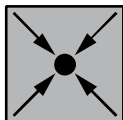
- Rozptyl
- Směrodatná odchylka
- Kvartilová odchylka
- Variační rozpětí (šíře)

3.2.1 Rozptyl, směrodatná odchylka

Nejčastěji používanými charakteristikami variability jsou rozptyl a směrodatná odchylka. *Rozptyl* je průměrná kvadratická odchylka měření od aritmetického průměru. *Směrodatná odchylka* je odmocninou z rozptylu. Obě charakteristiky nám popisují, jak jsou data rozložena (vzdálena) vzhledem k aritmetickému průměru. Čím větší je rozptyl (směrodatná odchylka) je, tím více jsou data rozložena do šíře vzhledem k aritmetickému průměru. Čím je rozptyl (směrodatná odchylka) menší, tím jsou data seskupena blíže k aritmetickému průměru, viz následující obrázek.

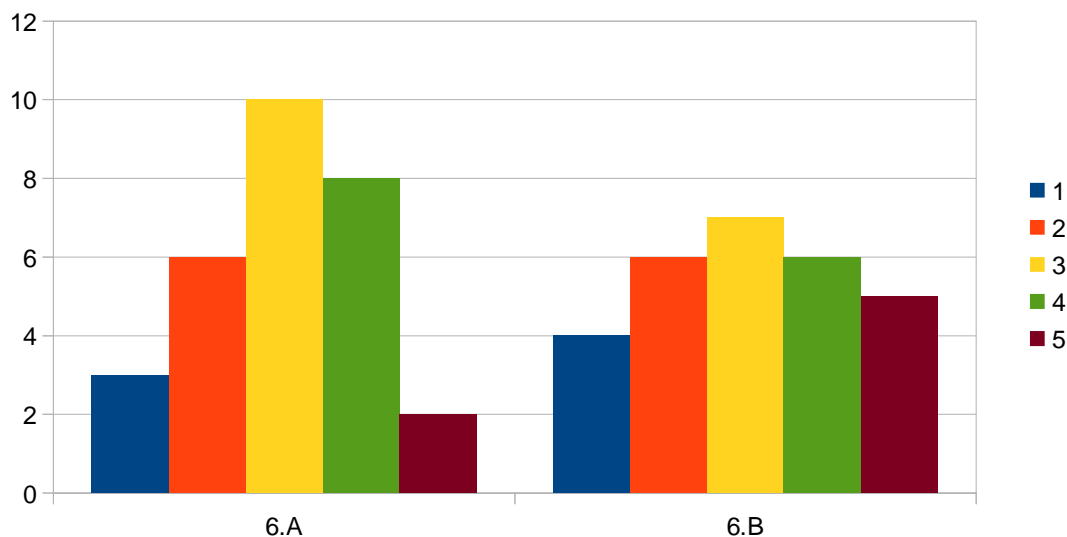


Příklad:



Z následujícího grafu vidíme, že v 6.A jsou data více soustředěna kolem průměrné hodnoty 3 než je tomu v případě 6.B, kde jsou data více roztažena. Tomu odpovídají i hodnoty rozptylu, u 6.A je rozptyl 1,21, kdežto u 6.B je rozptyl známek 1,84. Z výše uvedeného vidíme, že rozptyl má pro nás především srovnávací význam, chceme-li

srovnávat více souborů mezi sebou. Absolutní význam má pro nás směrodatná odchylka, ale tomu se budeme více věnovat až v kapitole o normálním rozdělení.



Rozptyl a směrodatná odchylka mají význam počítat pouze v případě, že data splňují podobně jako u výpočtu aritmetického průměru následující podmínky:

- data jsou metrická,
- data mají přibližně normální rozdělení,
- data jsou přibližně symetricky rozložena,
- soubor neobsahuje extrémní hodnoty.

3.2.2 Mezikvartilové rozpětí

Mezikvartilové rozpětí, jak už sám název napovídá, bude souviset s výpočtem kvartilů. Kvartily, podobně jako medián, decily a percentily, patří mezi kvantilové charakteristiky, tedy jejich hodnoty závisí na pozici v seřazeném souboru. Kvartily, jak už sám název napovídá, nám rozdělují soubor na čtvrtiny. Celkem máme 5 kvartilů, 0. kvartil je minimum, 1. kvartil (také se nazývá dolní kvartil) je hodnota nacházející se v jedné čtvrtině souboru dat seřazených vzestupně, 2. kvartil je medián, 3. kvartil (také horní kvartil) se nachází ve třech čtvrtinách seřazeného souboru a 4. kvartil je maximální hodnota.

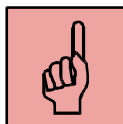


Pro výpočet mezikvartilového rozpětí nás budou zajímat 1. a 3. kvartil. Mezikvartilovým rozpětím nazveme rozdíl 3. a 1. kvartilu, je to tedy velikost intervalu od první do třetí čtvrtiny souboru. Tato velikost nám dobře popisuje, jak jsou data rozložena do šíře vzhledem k mediánu.

Podobně jako tomu bylo u mediánu, mezikvartilové rozpětí můžeme určovat v následujících případech:

- Data jsou ordinální či metrická.
- Data obsahují odlehlé hodnoty.
- Rozdělení je silně zešikmené.

3.2.3 Variační rozpětí

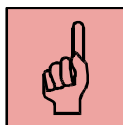


Poslední z uvedených charakteristik variability je *variační rozpětí*. Jeho hodnotu určíme jako rozdíl maximální a minimální hodnoty. Jeho hodnota je tedy velikost intervalu všech dat nacházejících se v našem naměřeném souboru.

3.3 Charakteristiky šikmosti a špičatosti

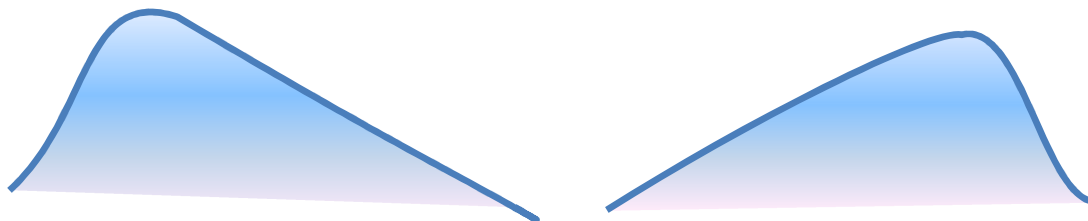
Zatím jsme se zabývali určením význačné polohy v souboru a rozložením dat do šíře vzhledem k dané význačné poloze. Nezajímali jsme se zatím tvarem rozložení dat v souboru. První důležitou podmínkou je, že data by se měla alespoň trochu blížit normálnímu rozdělení. V tomto případě pak můžeme sledovat jednak zešikmení rozložení dat v souboru a jednak špičatost dat.

3.3.1 Šikmost



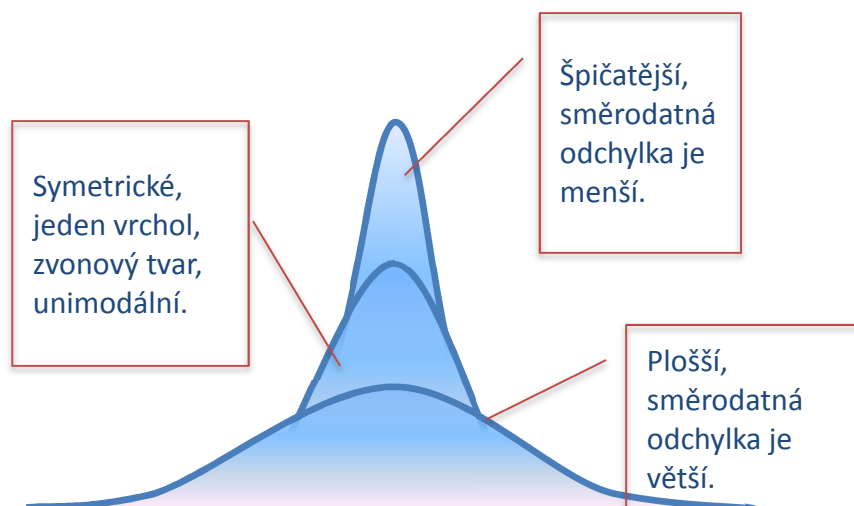
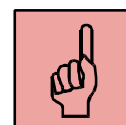
Šikmost nám popisuje zešikmení rozložení dat vzhledem symetrickému rozložení. Znaménko šikmosti nám udává směr zešikmení a velikost pak sílu zešikmení. Pokud je zešikmení zprava, pak je znaménko kladné a znamená to, že aritmetický průměr leží

vpravo od mediánu, viz obrázek vlevo. Pokud je zešikmení zleva, pak je znaménko záporné a znamená to, že aritmetický průměr leží vlevo od mediánu, viz obrázek vpravo.

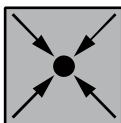


3.3.2 Špičatost

Výše uvedená šikmost nám popisuje tvar rozložení vzhledem k průměru. *Špičatost* nám udává, jak se data chovají v nejbližším okolí kolem průměru. Čím jsou data více soustředěna kolem průměru, pak je špičatost vysoká a směrodatná odchylka bývá malá, naopak pokud jsou data v nejbližším okolí průměru je málo naměřených dat, pak je špičatost nízká a směrodatná odchylka větší. Podobně jako je tomu na obrázku.



3.4 Výpočet základních charakteristik v programu OpenOffice



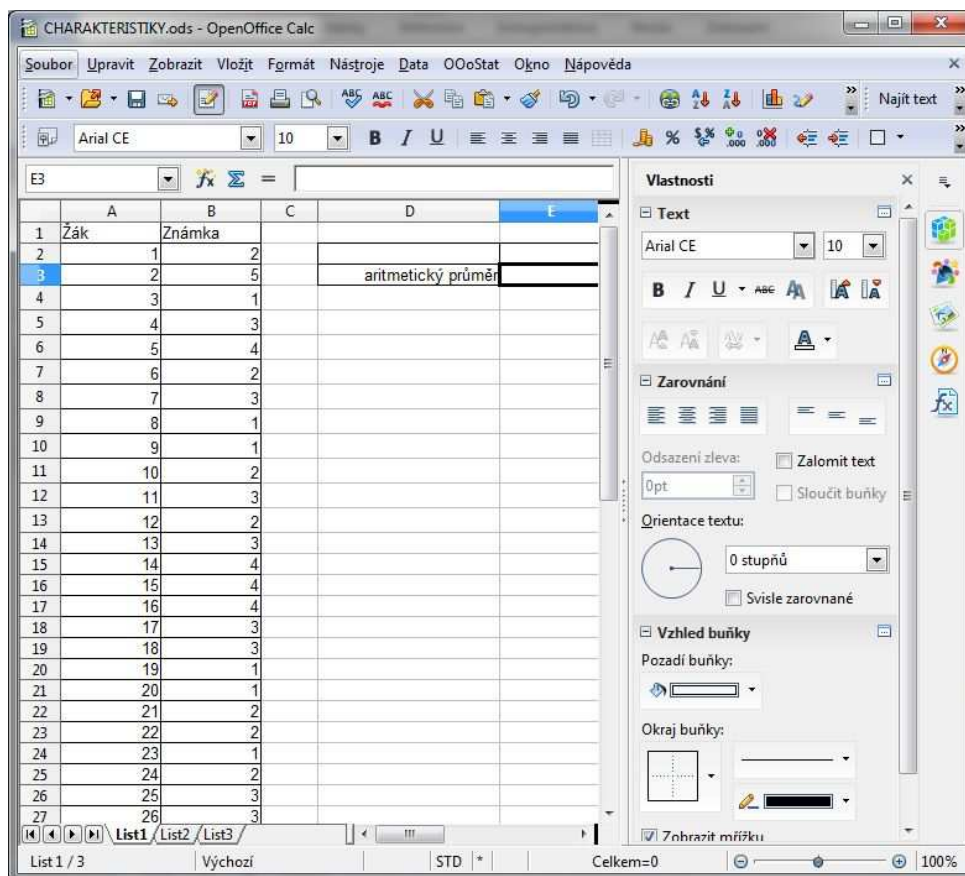
Máme v openoffice data (známky z testu v jedné třídě).

	A	B	C	D	E
1	Žák	Známka			
2		1	2		
3		2	5	aritmetický průměr	
4		3	1	rozptyl	
5		4	3	směrodatná odchylka	
6		5	4	medián	
7		6	2	modus	
8		7	3	dolní kvartil	
9		8	1	horní kvartil	
10		9	1	min	
11		10	2	max	
12		11	3	šikmost	
13		12	2	špičatost	
14		13	3		
15		14	4		
16		15	4		
17		16	4		
18		17	3		
19		18	3		
20		19	1		
21		20	1		
22		21	2		
23		22	2		
24		23	1		
25		24	2		
26		25	3		
27		26	3		

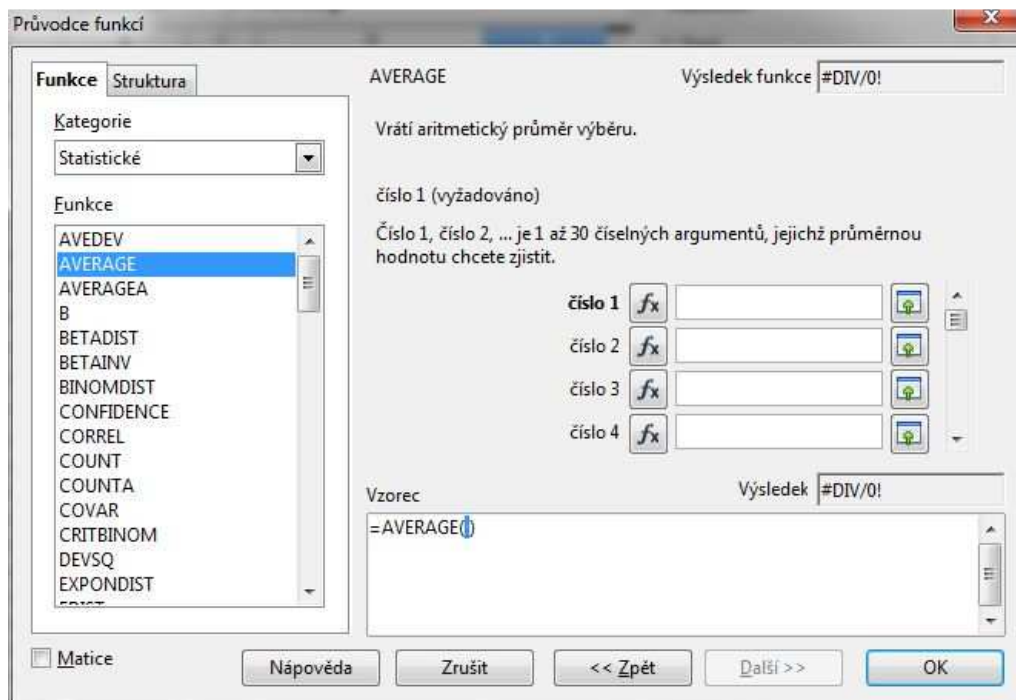
Aritmetický průměr vypočteme pomocí funkce AVERAGE(DATA).

Označíme si buňku, kam chceme průměr vypočít

Elementární statistické zpracování

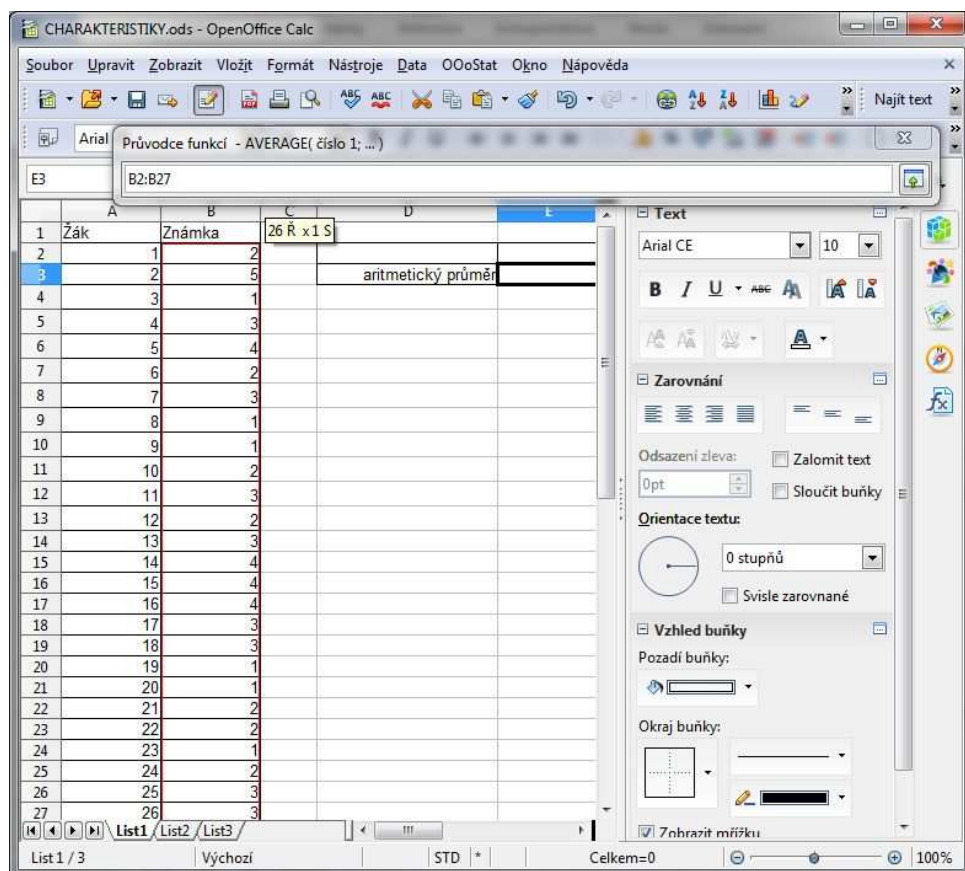


Vybereme z nabídky Vložit -> Funkce. Vybereme kategorii STATISTICKÉ a z funkcí pak vybereme funkci AVERAGE.

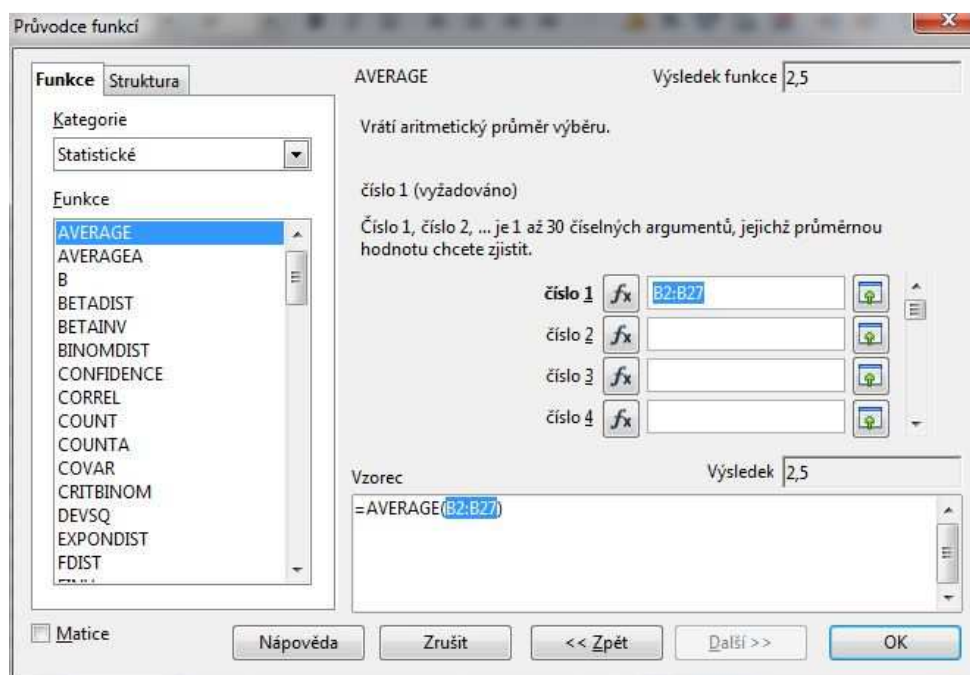


Elementární statistické zpracování

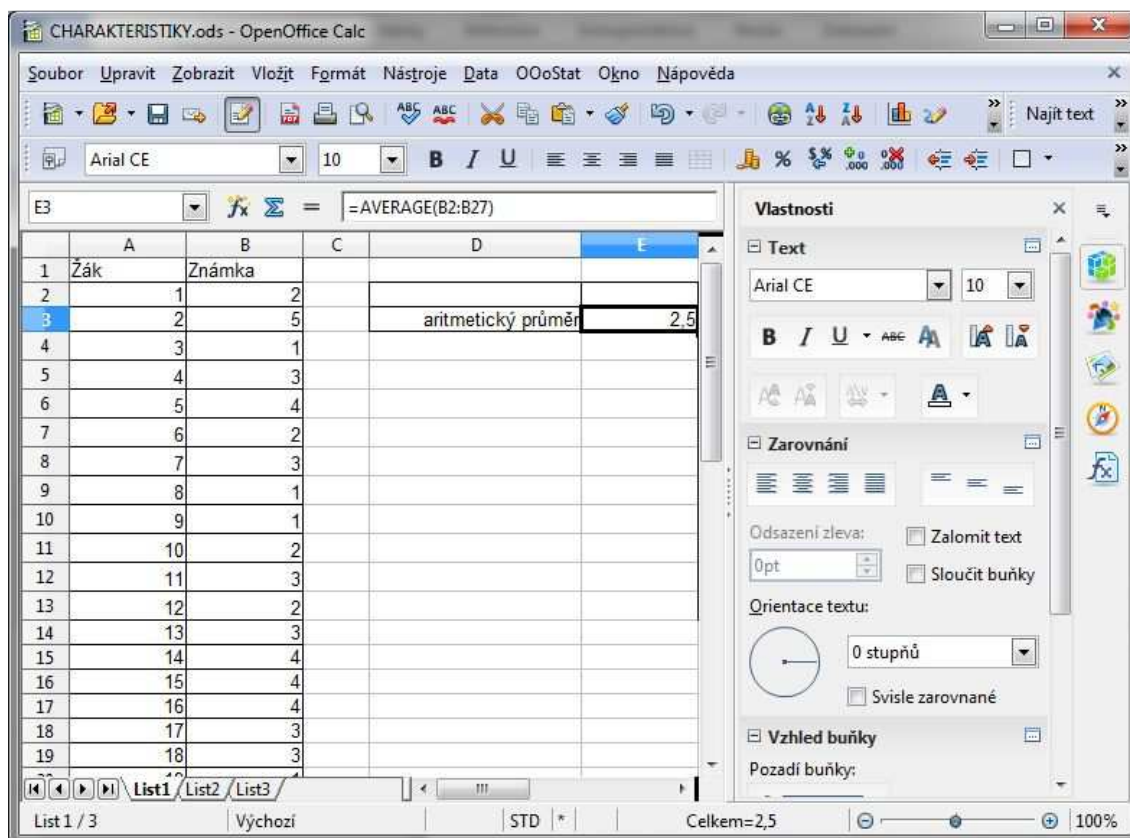
V řádku „číslo 1“ klikneme na zelenou šipku na konci řádku a vybereme data, ze kterých počítáme průměr.



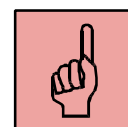
Klikneme na zelenou šipku na konci řádku a rozbálí se nám zpět výpočet průměru.



Pak už stačí jen dát OK a máme hodnotu aritmetického průměru.



Obdobně budeme postupovat při výpočtu dalších charakteristik.



Počítaná charakteristika	Název funkce v programu OpenOffice	Název funkce v programu EXCEL
aritmetický průměr	AVERAGE(data)	PRŮMĚR(data)
rozptyl	VAR(data)	VAR(data)
směrodatná odchylka	STDEV(data)	SMODCH(data)
medián	MEDIAN(data)	MEDIAN(data)
modus	MODE(data)	MODE(data)

dolní kvartil	QUARTILE(data;1)	QUARTIL(data;1)
horní kvartil	QUARTILE(data;3)	QUARTIL(data;3)
min	MIN(data)	MIN(data)
max	MAX(data)	MAX(data)
šikmost	SKEW(data)	SKEW(data)
špičatost	KURT(data)	KURT(data)

U kvartilů ještě vystupuje parametr, podle kterého se rozhodujeme, zda počítáme první neboli dolní kvartil a třetí neboli horní kvartil. V případě dolního kvartilu píšeme při zadávání dat do funkce do druhého řádku 1, v případě horního 3.

Výsledná tabulka by měla vypadat následovně.

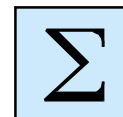
The screenshot shows the OpenOffice Calc interface with a spreadsheet titled 'CHARAKTERISTIKY.ods'. The spreadsheet contains a list of students and their grades, followed by a summary of statistical measures. The formula bar shows the formula for kurtosis: `=KURT(B2:B27)`. The Properties panel on the right is open, showing text formatting options.

Žák	Známka		
1	1		
2	2		
3	5	aritmetický průměr	2,5
4	1	rozptyl	1,3
5	3	směrodatná odchylka	1,140175425
6	4	medián	2,5
7	2	modus	3
8	3	dolní kvartil	2
9	1	horní kvartil	3
10	1	min	1
11	2	max	5
12	3	šikmost	0,263117406
13	2	špičatost	-0,668896321
14	3		
15	4		
16	4		
17	4		
18	3		
19	3		
20	1		
21	1		

Ještě bychom se měli vrátit ke špičatosti. V programech jako je OpenOffice, EXCEL a jiné není počítána přímo špičatost, ale pomocí funkce KURT vypočteme špičatost zmenšenou o špičatost normálního rozdělení se stejnými parametry. Tedy hodnota nám udává, zda špičatost našeho rozložení dat je menší (záporná hodnota KURT) nebo větší (kladná hodnota KURT) než u normálního rozdělení.

Shrnutí kapitoly

- Seznámili jsme se se základními popisnými metodami statistického souboru.
- Známe základní charakteristiky, metody jejich výpočtu a možnosti jejich použití při popisu statistického souboru.



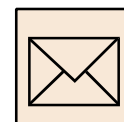
Kontrolní otázky a úkoly:

1. Co jsou to charakteristiky polohy, variability a šikmosti a špičatosti?
2. Kdy můžeme použít aritmetický průměr a směrodatnou odchylku?
3. Kdy můžeme použít mediána mezikvartilové rozpětí?
4. Která charakteristika polohy slouží k popisu nominálních dat?



Korespondenční úkoly

1. Na vybraném statistickém souboru proveďte měření tak, abyste získali různé typy dat (metrická, pořadová) a proveďte jejich popisnou statistiku.



4 Ověřování hypotéz



Cíl kapitoly

Po nastudování této kapitoly byste měli být schopni:

- popsat postup při ověřování statistických hypotéz,
- rozumět základním pojmům v ověřování hypotéz,
- vybrat správný typ testu pro ověřování dané hypotézy,
- vyvodit správné závěry získané na základě statistického ověření.



Klíčová slova: hypotéza, věcná hypotéza, statistické hypotézy, nulová hypotéza, alternativní hypotéza, hladina významnosti, statistické kritérium, p -hodnota, parametrické testy, neparametrické testy, jednostranný a oboustranný test, dvouvýběrový t-test, párový t-test, F-test, ANOVA, chí-kvadrát test dobré shody, chí-kvadrát test nezávislosti v kontingenčních tabulkách.



Čas na prostudování kapitoly

Průvodce studiem



V této kapitole se seznámíte se základními postupy při ověřování hypotéz. V první části se budeme věnovat správné formulaci hypotéz. V dalších částech obecnému postupu při ověřování hypotéz a posléze konkrétní typy testů pro ověřování určitých hypotéz.

4.1 Statistické metody používané při ověřování hypotéz

4.1.1 Věcné hypotézy

V kvantitativním výzkumu především ověřujeme hypotézy o vztazích mezi proměnnými. Ve vlastním výzkumu formulujeme tzv. *věcné hypotézy*. Jsou to hypotézy vyslovené v našem přirozeném jazyce. Musí být ovšem splněny požadavky na formulaci hypotéz:

1. Hypotéza je vždy ve tvaru oznamovací věty.
2. Proměnné vystupující v hypotéze musí být měřitelné. Hypotézu musí být možno empiricky ověřit.
3. Hypotéza musí vyjadřovat vztah dvou a více proměnných (tvrzení o rozdílech, vztazích nebo následcích).



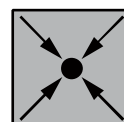
Hypotéza je předpoklad, který chceme buďto potvrdit nebo vyvrátit. Proto musí být ve tvaru oznamovací věty. Proměnné musí být měřitelné ve smyslu měření ve statistice. To znamená, že musíme mít dobře zvolenou škálu, na základě které jsme schopni získat věrohodná data pro naše zpracování.



Uvedme si příklady věcných hypotéz:

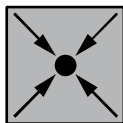
Příklady:

1. Řidiči muži ve věku 18 až 25 let způsobují více dopravních nehod než řidičky ženy.
2. Žáci mají po procvičení látky lepší znalosti než před procvičením.



Při ověřování hypotéz jde o prokazování její pravdivosti či nepravdivosti. Musíme si však uvědomit, že máme k dispozici nějaká data, která nikdy nejsou úplná, neznáme kompletní data, ale vše vyvozujeme na základě zpracování určitého vzorku (výběrového souboru). **Tudíž nikdy nemůžeme stoprocentně tvrdit na základě našeho ověření, že námi vyslovená hypotéza je pravdivá či nepravdivá. Všechna vyhodnocení jsou prováděna s určitým rizikem (s určitou pravděpodobností). Proto na závěr ověřování hypotézu pouze přijmeme nebo zamítneme.**

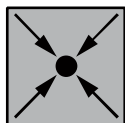
4.1.2 Chyby při formulaci a ověřování hypotéz



Při formulaci hypotéz se dopouštíme několika chyb, kterých je třeba se vyvarovat, proto si uvedeme nejčastější z nich:

- nevyjadřují vztah mezi proměnnými:
„Chlapci mají většinou rádi fyziku.“, „Základní školy jsou dobře vybaveny laboratorními pomůckami.“,
- nemají formu oznamovací věty;
„Jsou chlapci 4. tříd rychlejší v běhu na 60 m než děvčata?“
- někdy složité souvětí, z kterého nevyplývá jednoznačné tvrzení,
„Zaměstnanci s nižším vzděláním mají větší náchyllost k syndromu vyhoření, zvláště pak muži.“
- neurčité formulace:
„Chlapci jsou **někdy** méně ukáznění než děvčata.“,
- na konci ověření se často objeví chybný výsledek typu „Hypotéza byla částečně potvrzena“; hypotézu můžeme pouze přijmout nebo zamítnout, nikdy nemůže být částečně potvrzena, pak byla špatně formulována hypotéza nebo proměnné.

Musíme si také dobře uvědomit, zda díky našim prostředkům měříme ty proměnné, které jsme si v rámci výzkumu nadefinovali.



Příklad:

V tomto příkladu si ukážeme **špatně formulovanou hypotézu** na základě chybně definovaných proměnných.

V našem měření se zaměřujeme na zjišťování našich znalostí, schopností a dovedností před a po absolvování určitého školení.

Nadefinovali jsme si mimo jiné proměnné:

- jazyková dovednost,
- absolvování školení.

V dotazníku, kterým chceme ověřovat hypotézu:

Školení mělo vliv na zvýšení jazykových znalostí.

Na základě formulované hypotézy zvolíme otázky:

x. Jaké byly vaše jazykové znalosti před absolvováním školení?

nízké 1 2 3 4 5 vysoké

y. Jaké byly vaše jazykové znalosti po absolvování školení?

nízké 1 2 3 4 5 vysoké

Naměříme výsledky a provedeme vyhodnocení. Použijeme správný test na ověření hypotézy a zjistíme, že mezi jazykovými znalostmi před a po školení je významný rozdíl.

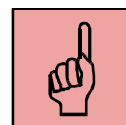
Můžeme nyní prohlásit, že „Školení mělo významný vliv na zlepšení jazykových znalostí“?



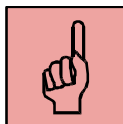
Zamysleme se nad tím, co vlastně měříme otázkami x. a y. v dotazníku.

Abychom mohli měřit jazykové znalosti, potřebovali bychom na to objektivní nástroj, tzn. nějaký standardizovaný didaktický test měřící úroveň jazykových znalostí. Odpovědi na otázky jsou pouze subjektivní pocity dotazovaných, a tudíž pomocí nich nemůžeme ověřovat výše uvedenou hypotézu.

Proměnné vystupující ve věcných hypotézách musíme zpravidla vyjádřit tak, abychom je mohli našimi prostředky měřit. Tomuto vyjádření říkáme *operacionalizace*. Budeme-li chtít měřit znalosti žáků, pravděpodobně k tomu použijeme nějaký standardizovaný didaktický test a úroveň znalostí pak bude odpovídat rozložení výsledků v uvedeném testu. Máme-li formulovanou věcnou hypotézu a operacionalizované proměnné, musíme přejít k formulaci statistických hypotéz.



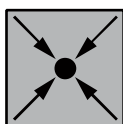
4.1.3 Statistické hypotézy



Statistické hypotézy jsou hypotézy, které vzniknou z věcné hypotézy převedením do statistického jazyka. Statistické hypotézy jsou dvě: nulová a alternativní.

Nulová hypotéza, značíme ji symbolem H_0 , je vždy formulována tak, že mezi proměnnými není závislost, v případě porovnávání úrovně proměnné mezi skupinami, formulujeme nulovou hypotézu tak, že mezi skupinami není v úrovni rozdíl.

Alternativní hypotéza je pak formulována proti nulové, značí se H_1 , ale nemusí být nutně její negací.



Příklady:

a) Věcná hypotéza: Řidiči muži ve věku 18 až 25 let způsobují více dopravních nehod než řidičky ženy.

Nejdříve si definujme proměnné:

1. proměnná: pohlaví,
2. proměnná: množství dopravních nehod v daném časovém období.

V tomto případě zkoumáme závislost mezi množstvím dopravních nehod a skutečností, zda řídil muž či žena. Hypotézy proto formulujeme následovně.

H_0 : Množství dopravních nehod v daném časovém období závisí na pohlaví.

H_1 : Množství dopravních nehod v daném časovém období nezávisí na pohlaví.

V tomto případě je alternativní hypotéza opravdu negací hypotézy nulové, protože žádná jiná varianta nemůže nastat.

b) Věcná hypotéza: Zaměstnanci s vyšším než základním vzděláním jsou platově lépe hodnoceni než zaměstnanci se základním vzděláním.

Proměnné:

1. proměnná: vzdělání,

2. proměnná: platové ohodnocení zaměstnance.

V tomto případě bychom mohli uvažovat, že opět zkoumáme závislost platového ohodnocení na vzdělání zaměstnance, což i ve skutečnosti ověřujeme. V tomto případě však musíme použít jiný typ testu, což pochopíme, až budeme jednotlivé testy ověřování hypotéz umět, a dle toho musíme zformulovat nulovou a alternativní hypotézy. V tomto případě porovnáváme průměrné platy mezi dvěma skupinami dle vzdělání a zjišťujeme, zda je mezi nimi rozdíl.

H_0 : Průměrné platové ohodnocení zaměstnanců se základním vzděláním se neliší od průměrného platového ohodnocení zaměstnanců s vyšším než základním vzděláním.

H_1 : Průměrné platové ohodnocení zaměstnanců se základním vzděláním je nižší než u zaměstnanců s vyšším než základním vzděláním.

V tomto případě vidíme, že alternativní hypotéza není negací nulové hypotézy, protože existuje ke dvěma uvedeným tvrzením ještě třetí možné tvrzení: Průměrné platové ohodnocení zaměstnanců se základním vzděláním je vyšší než u zaměstnanců s vyšším než základním vzděláním. Toto tvrzení nás ale v ověřování nebude zajímat. Formulovat budeme vždy pouze nulovou a alternativní hypotézy.

V případě, že bychom měli věcnou hypotézu formulovanou takto:

Zaměstnanci s vyšším než základním vzděláním jsou platově jinak hodnoceni než zaměstnanci se základním vzděláním.

Pak v tomto případě nás nebude zajímat, jestli jsou lépe či hůře ohodnoceni a nulová a alternativní hypotézy budou následující:

H_0 : Průměrné platové ohodnocení zaměstnanců se základním vzděláním se neliší od průměrného platového ohodnocení zaměstnanců s vyšším než základním vzděláním.

H_1 : Průměrné platové ohodnocení zaměstnanců se základním vzděláním je jiné než u zaměstnanců s vyšším než základním vzděláním.

V tomto případě je alternativní hypotéza negací nulové, protože žádná další možnost neexistuje.



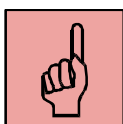
Důležité:

Nulová hypotéza je vždy formulována tak, že mezi proměnnými **není závislost** nebo mezi skupinami v úrovni proměnné **není rozdíl**.

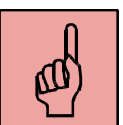
4.1.4 Hladina významnosti a p -hodnota

Dalším krokem při ověřování hypotéz je určení hladiny významnosti.

Co je hladina významnosti?



Jak jsme si již uvedli, veškeré závěry týkající se pravdivosti či nepravdivosti hypotéz se udávají s jistým rizikem. Jedním z těchto rizik je právě *hladina významnosti*, značíme α . Je to pravděpodobnost neboli riziko, že zamítneme nulovou hypotézu, která je ve skutečnosti pravdivá. Přestože bychom chtěli, aby toto riziko bylo nulové, statistickými metodami toho nemůžeme docílit. Se snižujícím se rizikem, že zamítneme pravdivou nulovou hypotézu, totiž roste riziko, že přijmeme nepravdivou nulovou hypotézu, tzv. riziko chyby druhého druhu. Proto bylo dlouholetým výzkumem zjištěno, že v pedagogických vědách je optimální nastavení hladiny významnosti na 0,05, výjimečně 0,01. Hladinu významnosti také občas nazýváme pravděpodobností chyby prvního druhu. Zatímco hladinu významnosti si volíme, pro rozhodnutí o přijetí či zamítnutí nulové hypotézy potřebujeme ještě experimentální hodnotu určenou ze získaných dat (měření, dotazník, test, apod.). Tuto hodnotu nazýváme p -hodnotou nebo také signifikancí.



p -hodnota (značíme malým p) znamená, zjednodušeně řečeno, jaká je pravděpodobnost, že by testovací kritérium dosáhlo své hodnoty, případně hodnot ještě více svědčících proti nulové hypotéze, pokud by nulová hypotéza opravdu platila.

Pro nás je důležité:

- a) pokud je p -hodnota menší nebo rovna hladině významnosti, pak nulovou hypotézu zamítáme na zvolené hladině významnosti;

- b) pokud je p -hodnota větší než hladina významnosti, pak nemůžeme nulovou hypotézu zamítnout a přijímáme ji na zvolené hladině významnosti.

4.1.5 Postup při ověřování statistických hypotéz

Nyní si shrneme celý postup při ověřování platnosti statistických hypotéz:

1. Vyslovení věcné hypotézy.
2. Stanovení nulové hypotézy a hypotézy alternativní.
3. Stanovení hladiny významnosti a použitého kritéria (výběr testu pro ověření).
4. Výpočet p -hodnoty (signifikance).
5. Přijetí nebo zamítnutí nulové hypotézy.
Je-li $p \leq \alpha$, pak zamítáme nulovou hypotézu na zvolené hladině významnosti.
Je-li $p > \alpha$, pak nulovou hypotézu přijímáme na zvolené hladině významnosti.
6. Správná interpretace výsledku testování.



4.2 Statistické testy

4.2.1 Jednostranné a oboustranné testy

V předchozí kapitole jsme se seznámili s metodami ověřování hypotéz. K ověření používáme statistické testy, které mohou být na základě formulovaných statistických hypotéz buď jednostranné nebo oboustranné.

V případě, že alternativní hypotéza není negací nulové hypotézy, pak se jedná o jednostrannou hypotézu.

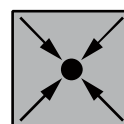


Příklad:

H_0 : Průměrné výsledky ze znalostního testu se mezi muži a ženami neliší.

H_1 : Průměrné výsledky ze znalostního testu jsou u žen lepší než u mužů.

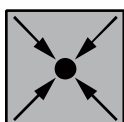
Symbolicky bychom mohli tyto hypotézy zapsat:



H_0 : průměr_ženy = průměr_muži

H_1 : průměr_ženy > průměr_muži

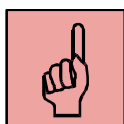
V tomto případě používáme jednostranný test (jednostranné kritérium), oboustranný test bychom použili v případě, že bychom se neptali na konkrétní nerovnost mezi výsledky u mužů a u žen, ale na obecnou nerovnost, jako v následujícím příkladu.



Příklad:

H_0 : Průměrné výsledky v běhu na 100 m se mezi muži a ženami neliší.

H_1 : Průměrné výsledky v běhu na 100 m jsou u žen jiné než u mužů.

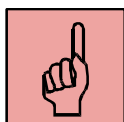


Jak vidíme v této alternativní hypotéze, netvrdíme konkrétně, že jsou lepší muži, nebo že jsou lepší ženy, ale tvrdíme obecně, že je mezi muži a ženami rozdíl. V takovém případě se jedná o oboustranný test (oboustranné kritérium).

Na správném výběru mezi jednostranným a oboustranným kritériem často závisí výpočet p -hodnoty, což si ukážeme u použití konkrétních testů.

4.2.2 Parametrické a neparametrické testy

Jak jsme si uvedli, hypotézy ověřujeme pomocí vybraného kritéria, které nazýváme také statistickým testem. Statistické testy dělíme především do dvou skupin podle toho, zda je či není známé rozdělení dat v měřeném souboru. V případě, že naše data odpovídají normálnímu rozdělení, pak používáme parametrické testy. *Parametrické* se nazývají proto, že se týkají některého z parametrů známého rozdělení. V případě, že naše data známému rozdělení neodpovídají, pak používáme testy, které nazýváme *neparametrické*.



V případě parametrických testů nejčastěji pracujeme s normálním rozdělením. Důvodem je, jak jsme již dříve uvedli, že většina jevů v reálném životě a v pedagogických a sociálních vědách se řídí právě normálním rozdělením. Z parametrických testů si uvedeme:

- dvouvýběrový t-test (test o rovnosti středních hodnot),
- párový t-test,
- F-test (test o rovnosti rozptylů),
- ANOVA (analýza rozptylu).

Z neparametrických testů si uvedeme především:

- chí-kvadrát test dobré shody (test rovnoměrnosti),
- chí-kvadrát test nezávislosti v kontingenčních tabulkách,
- test normality.

Kromě těchto neparametrických testů si u každého parametrického testu uvedeme jeho neparametrickou alternativu pro případ, že nemáme zaručeno dané rozdělení dat.

U základních testů si celý postup ověření platnosti hypotézy ukážeme na příkladu výpočtu v programu OpenOffice (popř. MS EXCEL). Některé testy – především neparametrické – v těchto programech nenajdeme a budeme muset použít jiný program. Pro jednoduchost si ukážeme výpočty v programu PAST, který má širokou oblast statistických výpočtů, je volně šířitelný, jeho předností je, že se nemusí instalovat a celý program zabírá pár MB, což znamená, že je lehce přenositelný na přenosných médiích jako USB apod.

4.3 Parametrické testy

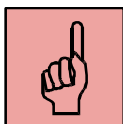
4.3.1 Dvouvýběrový t-test

První z testů – *dvouvýběrový t-test* – jak už sám název napovídá, se týká srovnávání dvou výběrů. Pomocí dvouvýběrového t-testu porovnáváme výsledky dvou různých souborů. Výsledky dvou souborů v tomto případě je založeno na srovnání průměrných výsledků těchto souborů.

Nulovou a alternativní hypotézu v případě dvouvýběrového t-testu formulujeme následovně.

V případě jednostranné alternativní hypotézy:





H_0 : průměr 1. souboru = průměr 2. souboru,

H_1 : průměr 1. souboru < průměr 2. souboru nebo průměr 1. souboru > průměr 2. souboru.

V případě oboustranné alternativní hypotézy:

H_0 : průměr 1. souboru = průměr 2. souboru,

H_1 : průměr 1. souboru \neq průměr 2. souboru.

Funkce v OpenOffice:

TTEST(data_1,data_2,režim, Typ)

Funkce v EXCELU:

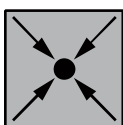
TTEST(Pole1,Pole2,Strany,Typ) EXCEL2010: (T.TEST(Matrice1,Matice2,Chvosty,Typ))

data_1 (Pole1; Matice1): oblast dat prvního souboru,

data_2 (Pole2; Matice2): oblast dat druhého souboru,

režim (Strany; Chvosty): vybíráme, zda chceme použít jednostrannou (hodnota 1) či oboustrannou (hodnota 2) alternativu,

Typ: vybíráme, zda jde o dva výběry se stejným rozptylem (hodnota 2) nebo s různým rozptylem (hodnota 3).



Příklad:

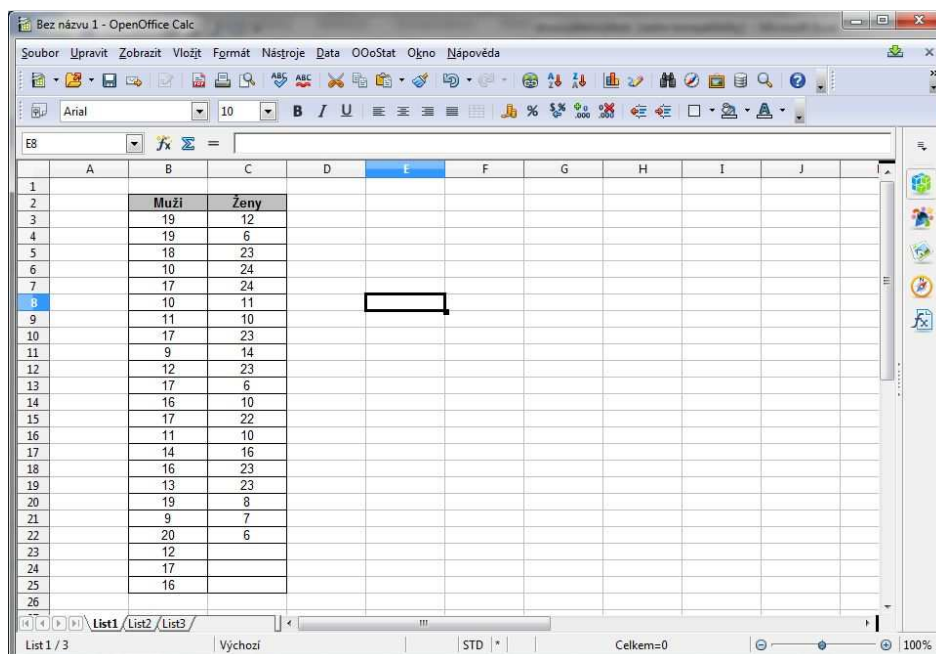
U skupiny mužů a skupiny žen byl proveden test znalostí z historie.

H: Mezi muži a ženami je rozdíl ve znalostech z historie.

1. Věcná hypotéza: Mezi znalostmi z historie mezi muži a ženami je rozdíl.
2. H_0 : průměr 1 je rovna průměr 2.
 H_1 : průměr 1 není roven průměr 2.
3. Hladina významnosti alfa se volí 0,05, kritériem je dvouvýběrový t-test.
4. Pomocí funkcí v programu OpenOffice (EXCEL, apod.) vypočteme signifikanci.

Ověřování hypotéz

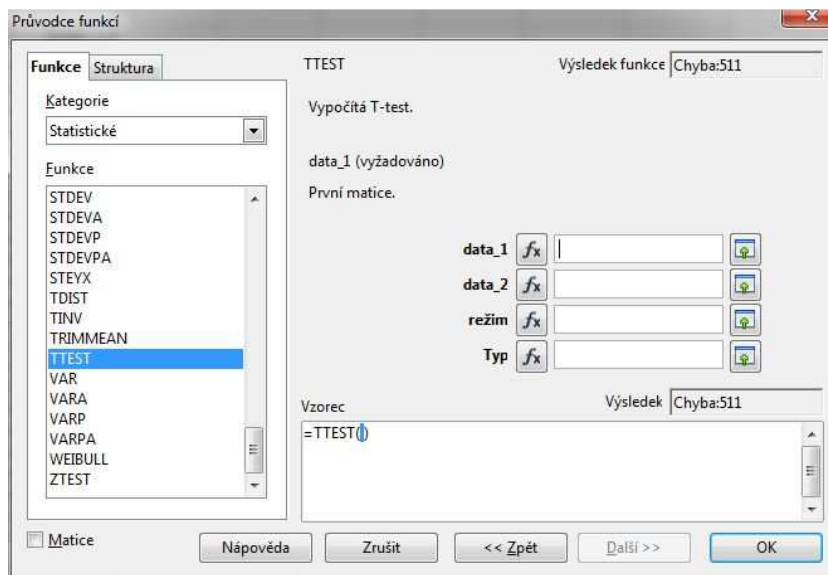
Máme připravenou tabulku dat (výsledků):



The screenshot shows an OpenOffice Calc spreadsheet with a table of data. The table has two columns: 'Muži' (Men) and 'Ženy' (Women). The data is as follows:

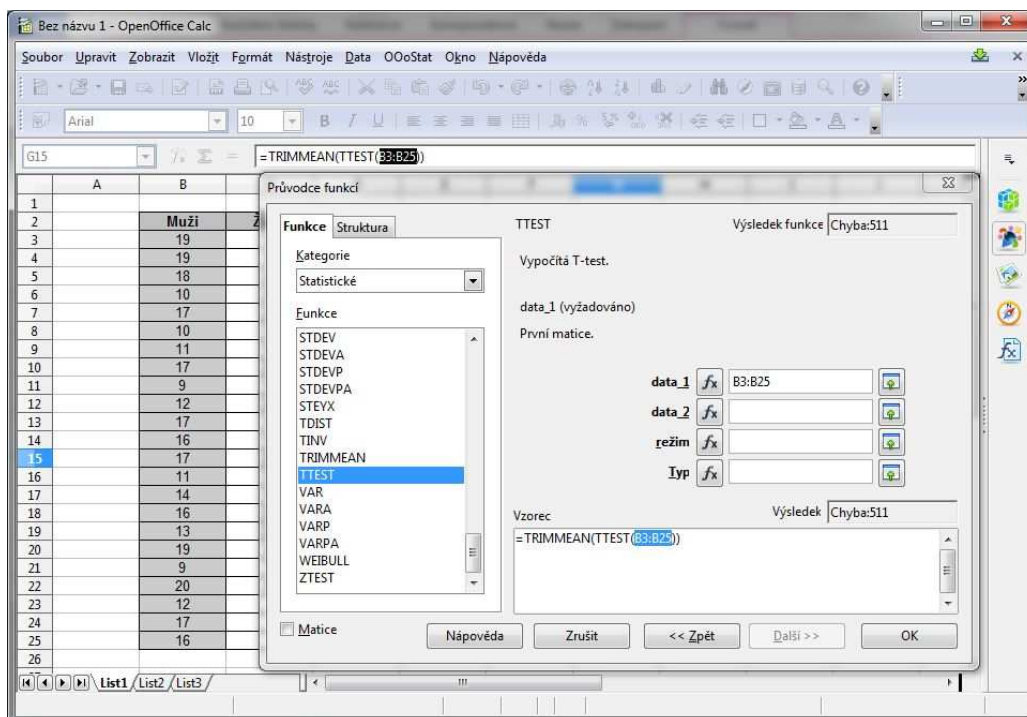
	Muži	Ženy
1		
2	19	12
3	19	6
4	18	23
5	10	24
6	17	24
7	10	11
8	11	10
9	17	23
10	9	14
11	12	23
12	17	6
13	16	10
14	17	22
15	11	10
16	14	16
17	16	23
18	13	23
19	19	8
20	9	7
21	20	6
22	12	
23	17	
24	16	
25		
26		

Označíme si buňku, kam chceme p -hodnotu vypočítat a vybereme z nabídky v OpenOffice příkaz „Průvodce funkcí“. Ve výběru kategorie zvolíme „Statistické“ a vybereme funkci TTEST.

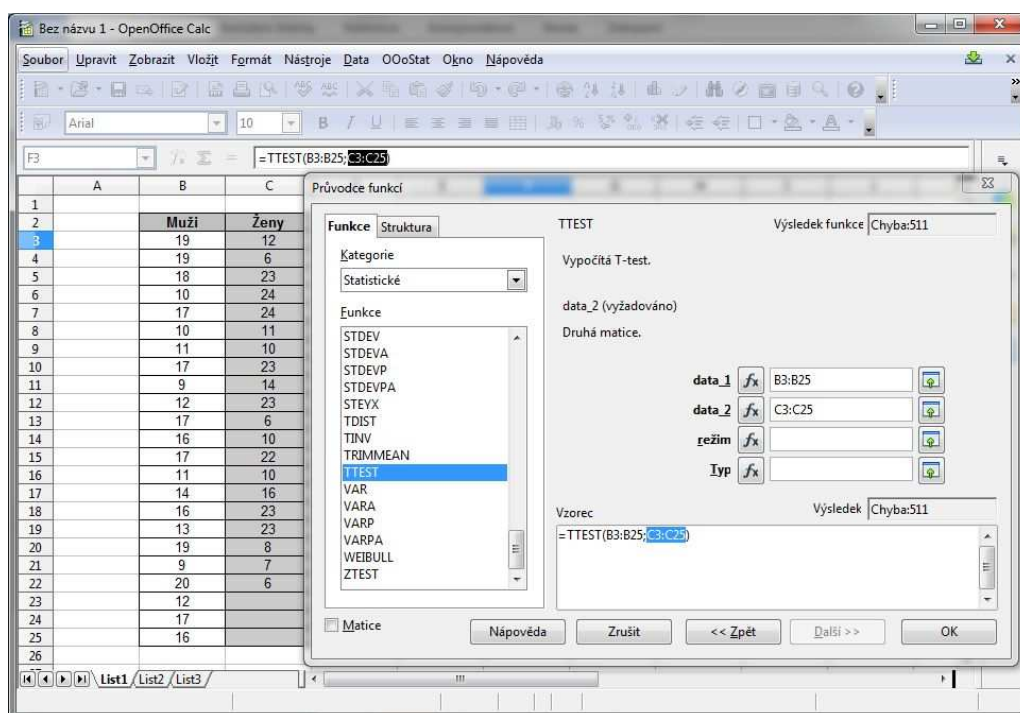


Klikneme do řádku „data_1“ a označíme oblast dat prvního souboru, tj. souboru „muži“, tedy označíme hodnoty v prvním sloupci.

Ověřování hypotéz



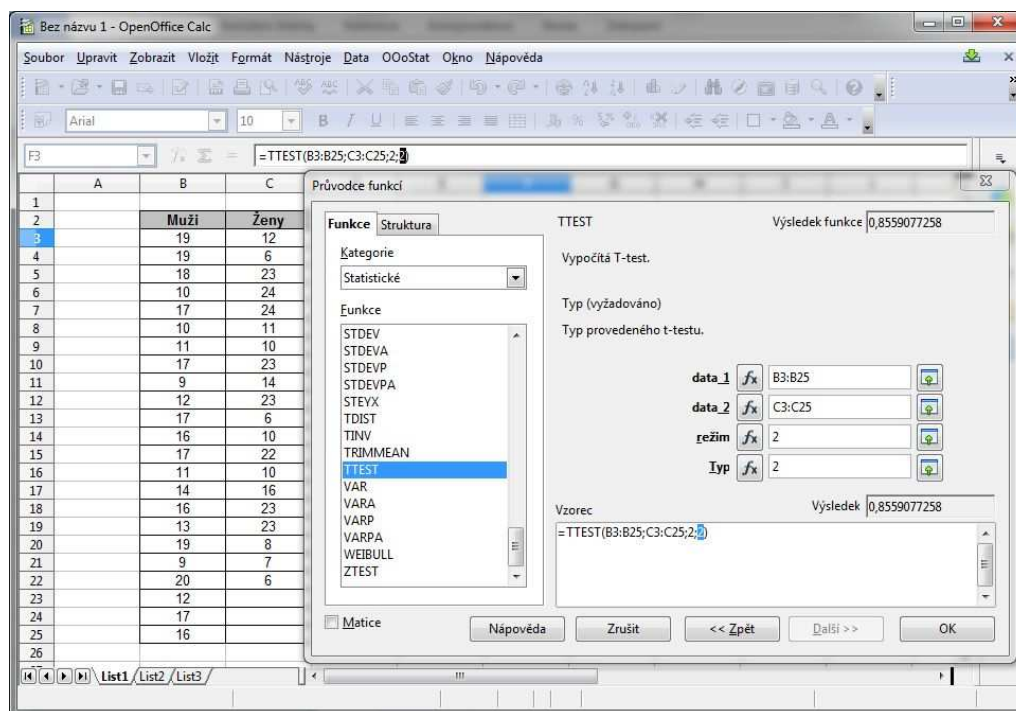
Klikneme do řádku „data_2“ a označíme oblast dat druhého souboru, tj. souboru „ženy“, tedy označíme hodnoty ve druhém sloupci.



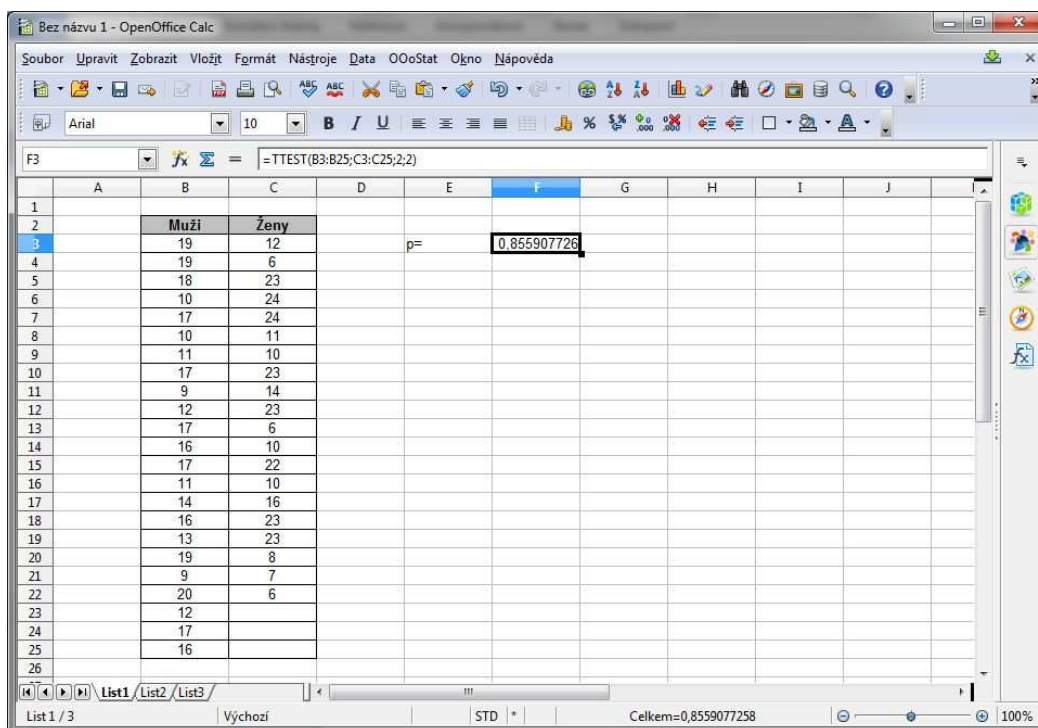
Vybereme režim, v tomto případě máme oboustrannou alternativní hypotézu, tedy do řádku režim napíšeme 2. A typ zvolíme podle toho, zda se rozptýly obou souboru

Ověřování hypotéz

rovnají či nikoli. Jak zjistíme, zda se rozptyly rovnají, si ukážeme pomocí až v kapitole F-test. Pomocí F-testu zjistíme, že rozptyly se neliší a tedy píšeme Typ 2.



Klikneme na OK nebo na klávesnici na Enter a v příslušné buňce dostaneme hodnotu p .



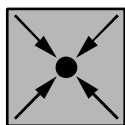
Obdrželi jsme p -hodnotu 0,856 po zaokrouhlení a nyní můžeme přejít ke kroku 5.

5. Signifikance $p = 0,856$ je větší než hladina významnosti $\alpha = 0,05$, tedy nulová hypotéza se přijímá na zvolené hladině významnosti 0,05.
6. Jelikož přijímáme nulovou hypotézu a věcná hypotéza odpovídá hypotéze alternativní, pak věcnou hypotézu (stejně jako alternativní) zamítáme na hladině významnosti 0,05.

Závěr tedy zní: ***Mezi muži a ženami není ve znalostech z historie statisticky významný rozdíl na hladině významnosti 0,05.***

V našem příkladu jsme měli oboustrannou alternativní hypotézu, což jednodušší, protože při jednostranné alternativní hypotéze musíme před výpočtem p -hodnoty ověřit, zda průměry obou souborů jsou v nerovnosti, jakou máme v naší alternativní hypotéze. Vše si ukážeme v následujících příkladech.

Příklady:



H: Chlapci ve věku 10 let jsou rychlejší v běhu na 100 m než děvčata v tomto věku.

2. Na základě této věcné hypotézy formulujeme hypotézy statistické.

H_0 : průměr chlapci = průměr děvčata,

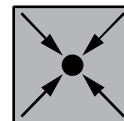
H_1 : průměr chlapci < průměr děvčata.

Chlapci mají být rychlejší, tzn. že čas by měl být u chlapců nižší než u děvčat.

3. Hladina významnosti $\alpha = 2$, volíme dvouvýběrový t-test.
4. Jednostranná alternativní hypotéza – před výpočtem p -hodnoty ověříme, že průměr chlapci je menší než průměr děvčata. Vypočteme pomocí příslušné funkce v OpenOffice oba průměry a porovnáme. Zjistili jsme, že průměrný čas u chlapců je 11,2 s a u děvčat je 12,3 s, tedy u chlapců je nižší než u děvčat a můžeme pokračovat v ověřování výpočtem p -hodnoty, zda tento rozdíl je ze statistického hlediska významný. Musíme ovšem dát ještě pozor při výpočtu p -hodnoty, že v OpenOffice „režim“ je roven 1, protože máme jednostrannou alternativní hypotézu. Dále již pokračujeme standardním postupem.

Nyní se podívejme na jiný příklad, kdy průměry obou souborů jsou v opačné nerovnosti, než jakou máme v alternativní hypotéze. Tento případ bývá zřídka, ale může se stát.

H: Absolventi gymnázií mají horší znalosti z fyziky než absolventi strojních průmyslových škol.



Znalosti budeme určovat na základě nějakého didaktického testu z fyziky a porovnávat průměrné výsledky obou skupin.

2. Statistické hypotézy:

H_0 : průměr gymnázia = průměr průmyslovka,

H_1 : průměr gymnázia < průměr průmyslovka.

Výsledky testu ovšem dopadly následovně:

Průměrný počet bodů z testu u absolventů z gymnázia je 25,8.

Průměrný počet bodů z testu u absolventů ze strojní průmyslové školy je 22,9.

V tomto případě již ve 4. kroku nepokračujeme ve výpočtu p -hodnoty a výsledek ověřování zní, že **Absolventi gymnázií nemají významně horší znalosti z fyziky než absolventi strojních průmyslových škol.**

V žádném případě nevyvozujeme závěr, že **Absolventi gymnázií mají významně lepší znalosti z fyziky než absolventi strojních průmyslových škol**, protože tuto hypotézu jsme nevyslovili a tudíž ani neověřovali.

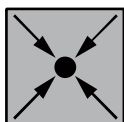
Na těchto příkladech jsme si ukázali, jak postupovat v případě, že věcná hypotéza odpovídá jednostranné alternativní hypotéze, tedy jednostrannému testu. Nestačí tedy určit p -hodnotu, ale musíme před tím porovnat průměry obou skupin.

4.3.2 Párový t-test

V předchozím testování jsme porovnávali dva různé soubory vzhledem k jejich výsledkům, tedy vzhledem k jejich průměrům. Často se ovšem můžeme setkat s ověřováním hypotézy, kdy máme jeden soubor objektů a srovnáváme u těchto



objektů dvojí výsledky. Tyto výsledky mohou být dvojího typu. V prvním případě budeme porovnávat výsledky u zkoumaných objektů před a po provedení nějaké akce. Ve druhém případě může jít o porovnávání výsledků použitím dvou různých metod na stejném souboru objektů.



Příklad:

U žáků 9. tříd byl proveden test z angličtiny před a po procvičení dané látky.

H: Procvičení látky mělo vliv na zlepšení jazykových dovedností žáků.

U tohoto párového t-testu zjišťujeme v konečném důsledku, zda mezi 1. a 2. výsledkem je rozdíl, tedy, zda rozdíl 1. a 2. výsledku je či není nulový. To jsme u dvou různých souborů zjišťovat nemohli, protože tam nelze výsledky spárovat. Na základě toho je postaven i test a formulace nulové a alternativní hypotézy.

H_0 : průměrný rozdíl 1. a 2. výsledku je nulový

H_1 : průměrný rozdíl 1. a 2. výsledku je nenulový – v případě, že zjišťujeme, zda obecně došlo ke změně, ale nezjišťujeme konkrétně zda došlo ke zlepšení či zhoršení.

V případě, že ověřujeme, že došlo *ke zvětšení hodnot*, pak jednostranná alternativní hypotéza bude mít tvar:

H_1 : průměrný rozdíl 1. a 2. výsledku je záporný,

protože vždy odčítáme *1. výsledek minus 2. výsledek*.

V opačném případě budeme formulovat rozdíl jako kladný.

V OpenOffice použijeme opět funkci TTEST(data_1,data_2,režim, Typ)

Funkce v EXCELU uijeme funkce:

TTEST(Pole1,Pole2,Strany,Typ) EXCEL2010: T.TEST(Matrice1,Matrice2,Chvosty,Typ)

data_1 (Pole1, Matrice1): oblast dat souboru před zásahem nebo použití první metody

data_2 (Pole2, Matrice2): oblast dat souboru po zásahu nebo použití druhé metody

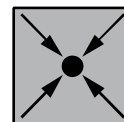
Ověřování hypotéz

režim (Strany, chvosty): výběr, zda používáme jednostrannou (hodnota 1) hypotézu či oboustrannou (hodnota 2)

Typ: vybíráme spárované výběry (hodnota 1)

Příklad:

U skupiny studentů byl proveden test před procvičením a po procvičení dané látky.



H: Po procvičení látky došlo ke zlepšení znalostí.

1. Věcná hypotéza: Po procvičení látky došlo ke zlepšení znalostí.

2. H_0 : Průměr rozdílu výsledků je nulový.

(výsledky 1. testu minus výsledky 2. testu)

H_1 : Průměr rozdílu výsledku je záporný.

(výsledky 2. testu by měly být lepší, tudíž by měl být rozdíl záporný)

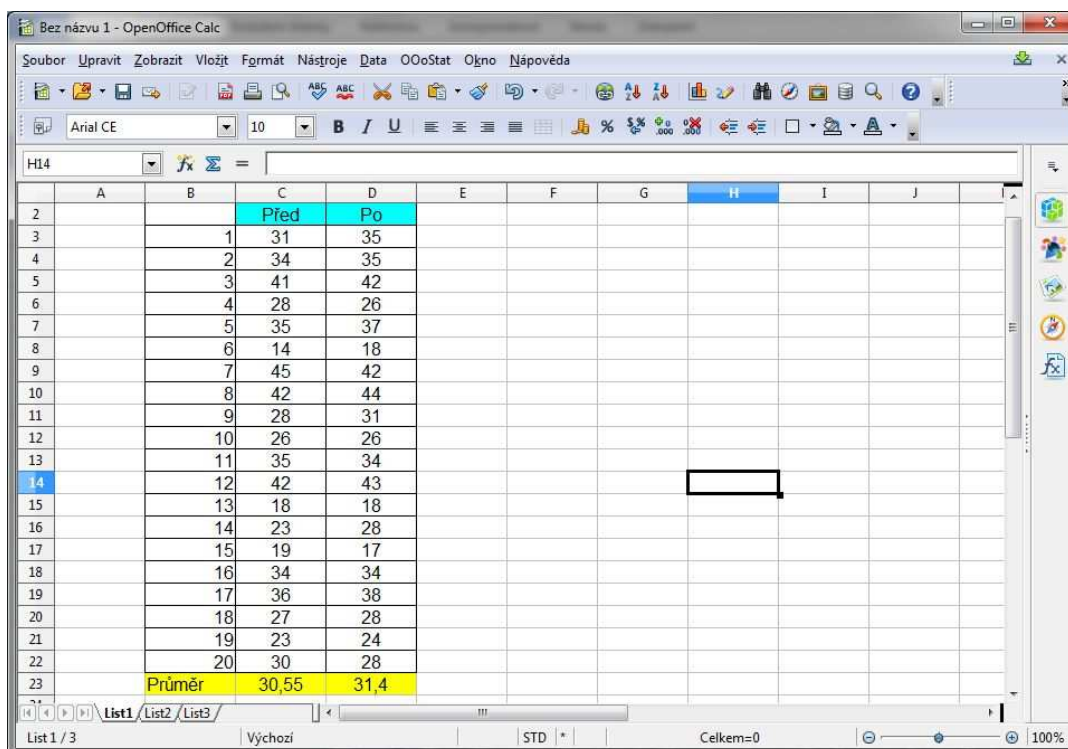
3. Hladina významnosti $\alpha = 0,05$, použijeme párový T-test.

4. Výpočet p -hodnoty v programu OpenOffice:

	A	B	C	D	E	F	G	H	I	J
2			Před	Po						
3		1	31	35						
4		2	34	35						
5		3	41	42						
6		4	28	26						
7		5	35	37						
8		6	14	18						
9		7	45	42						
10		8	42	44						
11		9	28	31						
12		10	26	26						
13		11	35	34						
14		12	42	43						
15		13	18	18						
16		14	23	28						
17		15	19	17						
18		16	34	34						
19		17	36	38						
20		18	27	28						
21		19	23	24						
22		20	30	28						
23										
24										

Ověřování hypotéz

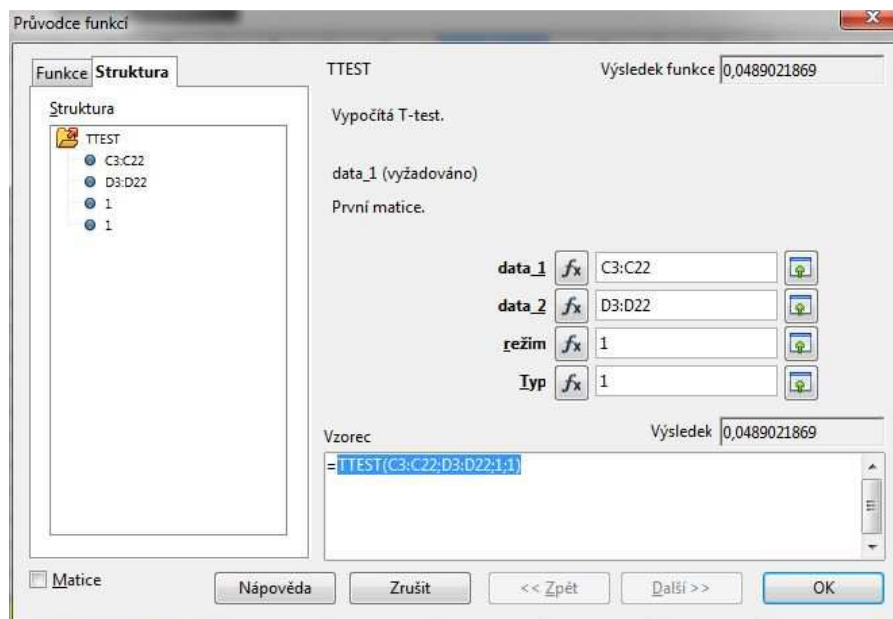
Než začneme výpočet p -hodnoty, vypočteme si průměrné hodnoty před a po procvičení. Pokud bude průměrná hodnota výsledků po procvičení nižší než před procvičením (průměrná hodnota rozdílu je kladná), což neodpovídá alternativní hypotéze, pak závěr zní: **„Po procvičení nedošlo k významnému zlepšení znalostí“**.



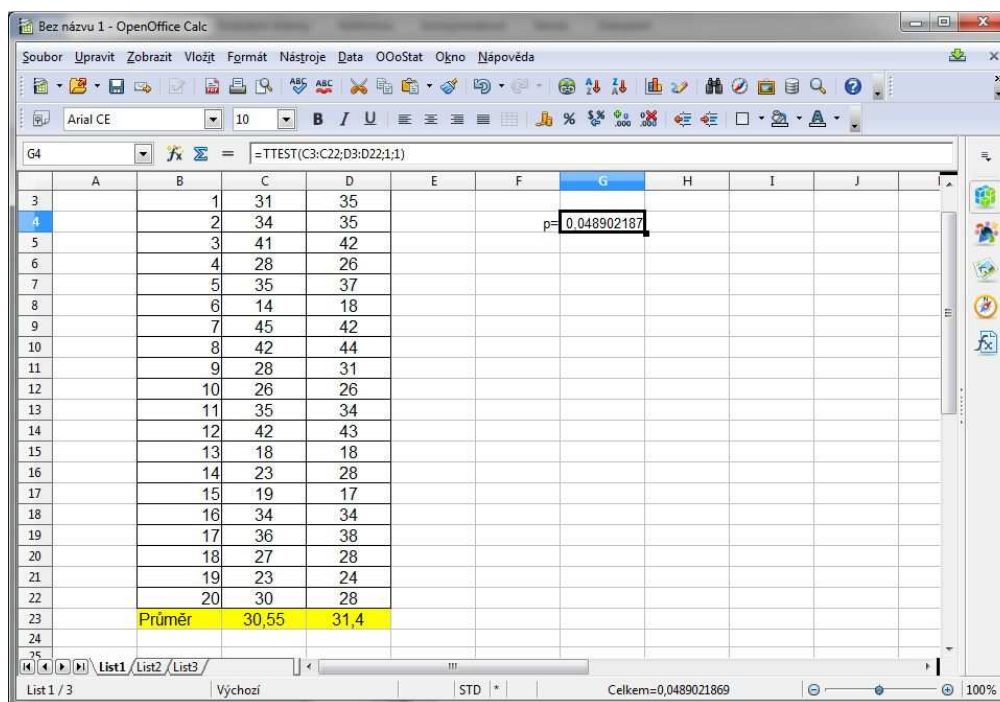
	A	B	C	D	E	F	G	H	I	J
2			Před	Po						
3		1	31	35						
4		2	34	35						
5		3	41	42						
6		4	28	26						
7		5	35	37						
8		6	14	18						
9		7	45	42						
10		8	42	44						
11		9	28	31						
12		10	26	26						
13		11	35	34						
14		12	42	43						
15		13	18	18						
16		14	23	28						
17		15	19	17						
18		16	34	34						
19		17	36	38						
20		18	27	28						
21		19	23	24						
22		20	30	28						
23			Průměr							
			30,55	31,4						

Z výsledků pod tabulkou vyplývá, že průměr po procvičení je větší než před procvičením a tedy můžeme pokračovat v ověřování, zda zlepšení je statisticky významné. Vypočteme si pomocí Průvodce funkcí hodnotu p . Podobně jako u dvouvýběrového t -testu, data_1 jsou ve sloupci „Před“ a data_2 jsou ve sloupci „Po“. Máme jednostrannou alternativní hypotézu, režim tedy volíme 1 a Typ je u párového t -testu hodnota 1. Zadání dat v průvodci funkcí tedy vypadá následovně:

Ověřování hypotéz



Klikneme pro dokončení a získáme hodnotu p .



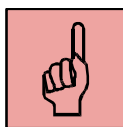
Obdrželi jsme hodnotu p po zaokrouhlení rovnou hodnotě 0,049.

- Hodnota p nám vyšla 0,049, což je menší než hladina významnosti $\alpha = 0,05$, a tedy nulovou hypotézu zamítáme ve prospěch alternativní hypotézy.

6. Jelikož věcná hypotéza odpovídala hypotéze alternativní, pak můžeme vyslovit závěr: „**Po procvičení došlo k významnému upevnění znalostí na hladině významnosti 0,05**“.

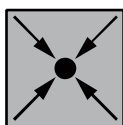
V případě, že ověřujeme hypotézu ve druhém tvaru, tedy srovnáváme výsledky dvou metod na stejném souboru objektů.

4.3.3 F-test



F-test také nazýváme testem o rovnosti rozptýlů. Tímto testem nesrovnáváme velikost (úroveň) hodnot u dvou souborů, ale porovnáváme rozptýlenost dat mezi dvěma soubory. Uvedme si příklad použití F-testu.

Příklad:



V jisté společnosti si měli vybrat mezi dvěma pracovními týmy. Jejich požadavek byl, aby tým byl kompaktní, tzn. aby se členové týmu co nejvíce shodovali ve svých názorech. K tomu měli psychologický test, díky kterému určovali výsledné skóre. Čím je tým kompaktnější, tím je rozptýlenost výsledných skór menší a obráceně.

H: Pracovní tým A je kompaktnější než pracovní tým B.

Z této věcné hypotézy formulujeme statistické hypotézy.

H_0 : Rozptyl A je roven rozptylu B.

H_1 : Rozptyl A je nižší než rozptyl B.

F-testem bychom nyní ověřili platnost věcné hypotézy.

Z příkladu je již zřejmé, jak budou formulovány nulová a alternativní hypotézy. Obecně budeme statistické hypotézy formulovat:

H_0 : rozptyl 1 = rozptyl 2,

H_1 : rozptyl 1 \neq rozptyl 2.

Funkce v OpenOffice: `FTEST(data_1,data_2)`.

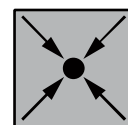
Funkce v EXCELU: `FTEST(Pole1,Pole2)` EXCEL2010: `F.TEST(Matice1,Matice2)`.

data_1 (Pole1, Matice1): oblast dat prvního souboru,

data_2 (Pole2, Matice2): oblast dat druhého souboru,

Příklad:

U dvou skupin zaměstnanců výchovných ústavů byl dotazníkem zjištěn statisticky významný rozdíl v názoru na vedení ústavu.



H: Mezi těmito skupinami zaměstnanců je rozdíl v názorové jednotnosti na vedení ústavu.

1. Věcná hypotéza: Mezi těmito skupinami zaměstnanců je rozdíl v názorové jednotnosti na vedení ústavu.
2. H_0 : Rozptyl 1 je roven rozptylu 2
 H_1 : Rozptyl 1 se liší od rozptylu 2
(ptáme se obecně, zda je mezi skupinami rozdíl v rozptylu)
3. Hladina významnosti $\alpha = 0,05$, použijeme F-test.
4. Přikročíme k výpočtu p -hodnoty v programu OpenOffice.

V OpenOffice máme naměřená data:

Ověřování hypotéz

	A	B	C	D	E	F	G	H	I	J
1										
2			skupina A	Skupina B						
3		1	28	15						
4		2	33	31						
5		3	11	26						
6		4	18	30						
7		5	28	29						
8		6	23	19						
9		7	25	36						
10		8	18	33						
11		9	10	25						
12		10	26							
13		11	14							
14		12	16							
15		13	19							
16		14	26							
17		15	22							
18										
19										
20										
21										

Vybereme funkci FTEST v programu OpenOffice a zadáme do obou řádků oblast dat.

Průvodce funkcí

Funkce: Statistické

Funkce: FTEST

Výsledek funkce: 1,0426931583

Vypočítá F-test.

data_2 (vyžadováno)

Druhá matice.

data_1: C3:C17

data_2: D3:D11

Vzorec: =FTEST(C3:C17;D3:D11)

Výsledek: 1,0426931583

Maticice

Nápověda Zrušit << Zpět Další >> OK

A kliknutím určíme hodnotu p .

	A	B	C	D	E	F	G	H	I	J
1										
2			skupina A	Skupina B						
3		1	28	15						
4		2	33	31		p=	0.9573068			
5		3	11	26						
6		4	18	30						
7		5	28	29						
8		6	23	19						
9		7	25	36						
10		8	18	33						
11		9	10	25						
12		10	26							
13		11	14							
14		12	16							
15		13	19							
16		14	26							
17		15	22							
18										
19										
20										
21										

Hodnota p rovna 0,957.

- p -hodnota je větší než hladina významnosti $\alpha = 0,05$. Nemůžeme tedy zamítnout nulovou hypotézu.
- Jelikož jsme přijali nulovou hypotézu a věcná hypotéza odpovídala alternativní hypotéze, pak závěr zní: „**V názorové rozptýlenosti není mezi skupinami významný rozdíl.**“

4.3.4 ANOVA

Test ANOVA neboli *analýza rozptylu* by nám dle názvu napovídala, že se jedná o srovnávání rozptylu. Analýza rozptylu však slouží k porovnávání průměrných hodnot souborů podobně jako dvouvýběrový t-test. V analýze rozptylu však můžeme porovnávat průměrné výsledky více než dvou souborů. Statistické hypotézy jsou formulovány:

H_0 : průměry všech souborů jsou stejné,

H_1 : průměr alespoň jednoho souboru se liší od ostatních.



Symbolicky:

H_0 : průměr 1 = průměr 2 = ... = průměr n ,

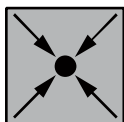
H_1 : není pravda, že průměr 1 = průměr 2 = ... = průměr n .

V OpenOffice pro určení p -hodnoty je třeba stáhnout statistický modul OOoStat, jehož součástí ANOVA je.

V EXCELU je třeba instalovat modul Analýza Dat.

Na příkladu si ukážeme ověření hypotézy Analýzou rozptylu.

Příklad:



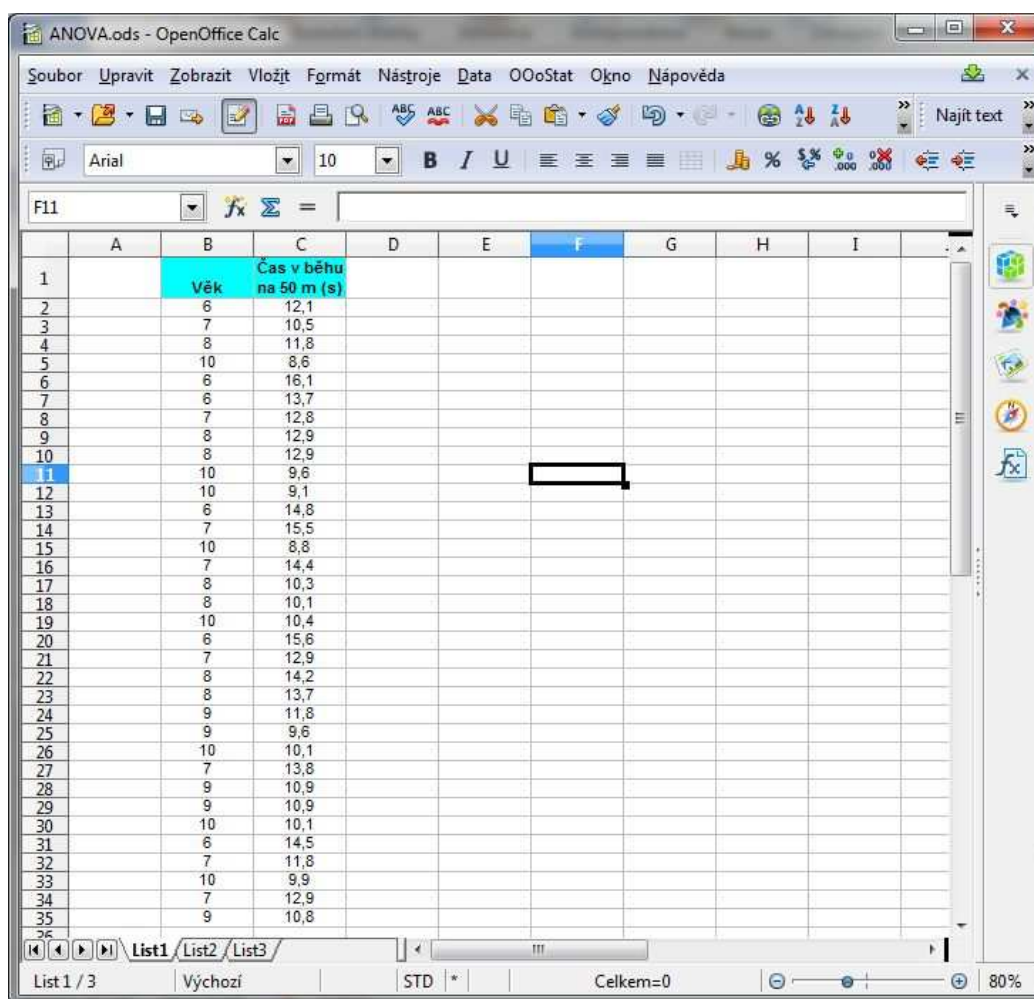
Na základní škole jsme zkoumali, jestli se liší výsledky v běhu na 50 m u chlapců ve věku 6 až 10 let dle jejich věku. Chlapce jsme rozdělili dle jejich věku do 5 skupin.

H: Výsledky v běhu na 50 m u dětí ve věku 6 až 10 let závisí na jejich věku.

1. Věcná hypotéza: Výsledky v běhu na 50 m u dětí ve věku 6 až 10 let závisí na jejich věku.
2. H_0 : průměr 6 let = průměr 7 let = průměr 8 let = průměr 9 let = průměr 10 let.
 H_1 : není pravda, že průměr 6 let = průměr 7 let = průměr 8 let = průměr 9 let = průměr 10 let.
3. Hladina významnosti $\alpha = 0,05$, použijeme test ANOVA.
4. Vypočteme p -hodnotu. Do OpenOffice připravíme data.

Data si připravíme tak, že do jednoho sloupce uvádíme proměnnou, dle které dělíme výsledky do skupin, v našem případě se jedná o věk. Ve druhém sloupci pak máme naměřené hodnoty proměnné.

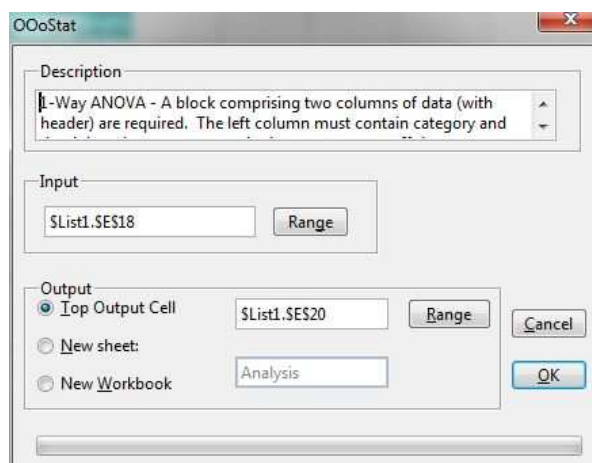
Ověřování hypotéz



The screenshot shows an OpenOffice Calc spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I
1		Věk	Čas v běhu na 50 m (s)						
2		6	12,1						
3		7	10,5						
4		8	11,8						
5		10	8,6						
6		6	16,1						
7		6	13,7						
8		7	12,8						
9		8	12,9						
10		8	12,9						
11		10	9,6						
12		10	9,1						
13		6	14,8						
14		7	15,5						
15		10	8,8						
16		7	14,4						
17		8	10,3						
18		8	10,1						
19		10	10,4						
20		6	15,6						
21		7	12,9						
22		8	14,2						
23		8	13,7						
24		9	11,8						
25		9	9,6						
26		10	10,1						
27		7	13,8						
28		9	10,9						
29		9	10,9						
30		10	10,1						
31		6	14,5						
32		7	11,8						
33		10	9,9						
34		7	12,9						
35		9	10,8						

Z nabídky vybereme OoStat -> Basic Stats -> 1-Way ANOVA. Otevře se okno pro výpočet ANOVy včetně hodnoty p .



The screenshot shows the OoStat dialog box with the following settings:

- Description: 1-Way ANOVA - A block comprising two columns of data (with header) are required. The left column must contain category and
- Input: \$List1.\$E\$18
- Output: Top Output Cell, \$List1.\$E\$20
- New sheet
- New Workbook, Analysis

Do pole Input zadáváme oblast dat, tzn. klikneme na Range, okno se sbalí a pak vybereme data ve sloupcích „Věk“ a „Čas v běhu na 100 m“ včetně záhlaví.

Ověřování hypotéz

ANOVA.ods - OpenOffice Calc

Soubor Upravit Zobrazit Vložit Formát Nástroje Data OOoStat Okno Nápověda

Najít text

Arial 10 B / U

	A	B	C	D	E	F	G	H	I
1		Věk	Čas v běhu na 50 m (s)						
2		6	12,1						
3		7	10,5						
4		8	11,8						
5		10	8,6						
6		6	16,1						
7		6	13,7						
8		7	12,8						
9		8	12,9						
10		8	12,9						
11		10	9,6						
12		10	9,1						
13		6	14,8						
14		7	15,5						
15		10	8,8						
16		7	14,4						
17		8	10,3						
18		8	10,1						
19		10	10,4						
20		6	15,6						
21		7	12,9						
22		8	14,2						
23		8	13,7						
24		9	11,8						
25		9	9,6						
26		10	10,1						
27		7	13,8						
28		9	10,9						
29		9	10,9						
30		10	10,1						
31		6	14,5						
32		7	11,8						
33		10	9,9						
34		7	12,9						
35		9	10,8						

OOoStat InputRange: Oblast

\$List1.\$B\$1:\$C\$35

35 Ř x 2 S

A nyní musíme nastavit, kam se provede výpočet. Zatrhneme New Sheet pro přehlednost.

OOoStat

Description

1-Way ANOVA - A block comprising two columns of data (with header) are required. The left column must contain category and

Input

\$List1.\$B\$1:\$C\$35 Range

Output

Top Output Cell \$List1.\$B\$37 Range

New sheet: Analysis

New Workbook

Cancel OK

Po provedení akce obdržíme výsledný list ve tvaru:

	A	B	C	D	E	F	G	H
1	Věk	N	Mean	S.E.M.	Sum square	Rank Sum	RS²/N	
2		6	14,46666667	0,584617634	1265,96	172,5	4959,375	
3		7	13,07500000	0,543714079	1384,2	182	4140,5	
4		8	12,27142857	0,605024315	1069,49	133,5	2546,035714	
5		10	9,575000000	0,235849528	736,56	43,5	236,53125	
6		9	10,80000000	0,350713558	585,66	63,5	806,45	
7	Total	34	11,99705882	0,363521235	5041,87	595	12688,89196	
8	Tukey HSD	6,5830721	2,05065162					
9	Tukey HSD	6,5830721	2,525802605					
10								
11	ANOVA Tab	Effect of Věk						
12	Source	SS	df	MS	F	p>F	Kruskall-Wa	p>KW
13	Between	100,5120868	4	25,12802171	15,2585628	0,0000	22,95521309	0,0001
14	Within	47,75761905	29	1,64681445				
15	Total	148,2697059	33	4,49302139				
16								
17								
18								
19								
20								
21								
22								
23								

Hodnota p je hodnota pod buňkou s nápisem $p>F$, na obrázku silně orámovaná buňka.

Hodnota p je rovna 0,000.

- p -hodnota je menší než hladina významnosti $\alpha = 0,05$, nulovou hypotézu tedy zamítáme ve prospěch hypotézy alternativní.
- Závěr bude znít: „**Výsledky v běhu na 50 m u dětí ve věku 6 až 10 let závisí na jejich věku**“.

4.4 Neparametrické testy

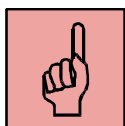
Ve všech předchozích případech jsme pro použití daného testu (kritéria) potřebovali mít splněnou důležitou podmínku, a to, že data mají normální rozdělení nebo jsou vybrána ze souboru s normálním rozdělením. V této kapitole si ukážeme testy, které takový požadavek na rozdělení dat nemají. Patří sem především chí-kvadrát testy dobré shody, chí-kvadrát test nezávislosti v kontingenčních tabulkách a pak alternativní neparametrické testy k testům parametrickým, kdy nemáme zaručeno normální

rozdělení. Než se začneme věnovat konkrétním testům, je důležité si ještě uvést, že chí-kvadrát testy lze použít pouze za předpokladu, že máme splněny podmínky:

- minimální počet respondentů je 40,
- všechny očekávané četnosti jsou větší než 1,
- maximálně třetina všech očekávaných četností je menších než 5.

V případě, že nejsou splněny tyto podmínky, chí-kvadrát testy nelze použít. Pak musíme buď doplnit počet měření, v některých případech stačí přeškálovat proměnné tak, že některé máločetné proměnné sloučíme do jedné, pokud je to z logického hlediska možné.

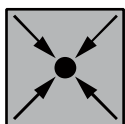
4.4.1 Chí-kvadrát test dobré shody – test rovnoměrnosti



Jako první z neparametrických testů si uvedeme test rovnoměrnosti, který patří mezi testy dobré shody. Pomocí těchto testů zjišťujeme, zda naše data mají určitý typ rozdělení. Mezi tyto testy patří i test normality, kterým si můžeme ověřit, zda naše data mají normální rozdělení. Tento test je ovšem pomocí programů jako OpenOffice a MS EXCEL na výpočet složitější a je vhodné pro ověření použít jiný program nebo metodu.

Vraťme se k testu rovnoměrnosti. Principem testu je předpoklad, že rozložení výsledků je rovnoměrné, tedy, že všechny varianty proměnné jsou zastoupeny stejným výskytem.

Příklad:



Při hodu klasickou hrací kostkou předpokládáme, že všech šest možností má stejnou pravděpodobnost, že padnou. Při 30 hodech by tedy každá hodnota měla padnout přibližně 5 krát. Samozřejmě, že při tak malém počtu se budou počty padnutí jednotlivých hodnot od hodnoty 5 více či méně lišit. Testováním můžeme ověřit, zda tyto odchylky jsou pouze dílem náhody, nebo zda se skutečně výrazně liší od předpokladu a tedy, že kostka není v pořádku.

Statistické hypotézy:

H_0 : Data mají rovnoměrné rozdělení.

H_1 : Data nemají rovnoměrné rozdělení.

V OpenOffice použijeme funkci: `CHITEST(Data_B, data_E)`.

V EXCELU použijeme funkci: `CHITEST(aktuální, očekávané)`.

V EXCELU 2010: `CHSQ.TEST(aktuální, očekávané)`.

Data_B (aktuální): oblast dat naměřených hodnot,

data_E (očekávané): oblast dat očekávaných četností.

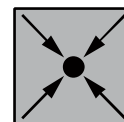
Nyní si ukážeme příklad z praxe:

Příklad:

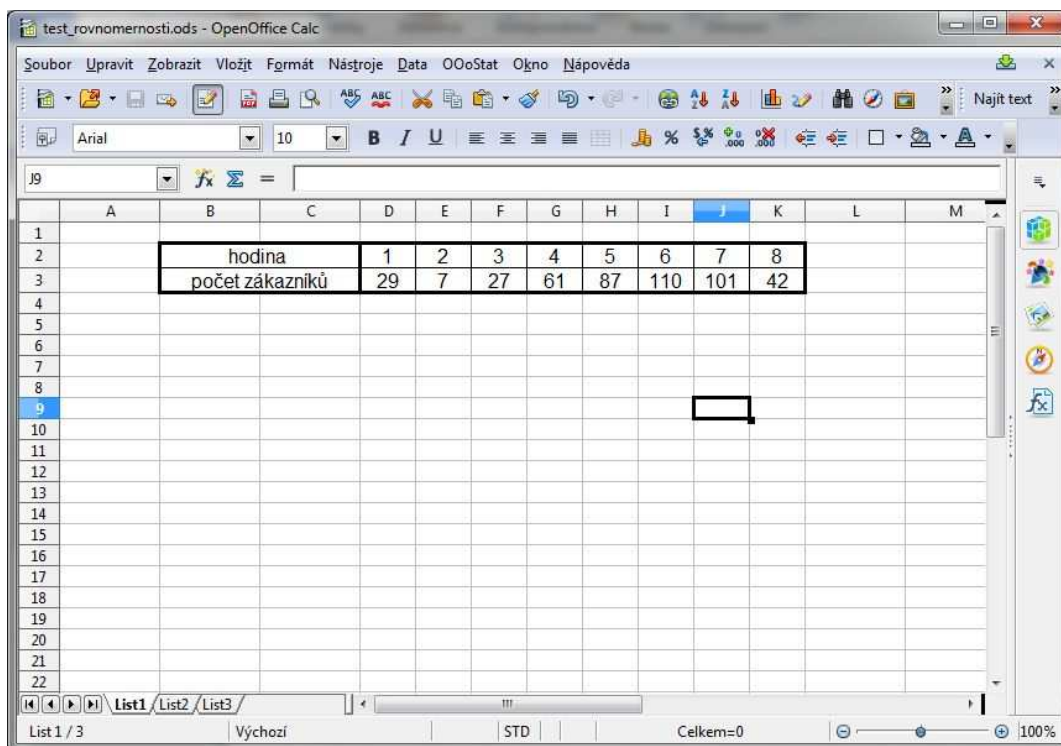
Při směně úřednice za přepážkou bylo zjišťováno, kolik má zákazník během směny v každé hodině.

H: Výskyt zákazníků během směny je nevyrovnaný.

1. Věcná hypotéza: Výskyt zákazníků během směny je nevyrovnaný.
2. H_0 : Rozložení počtu zákazníku během osmi hodin je rovnoměrné.
 H_1 : Rozložení počtu zákazníku během osmi hodin je nerovnoměrné.
3. Hladina významnosti $\alpha = 0,05$, kritériem je chí-kvadrát dobré shody – test rovnoměrnosti.
4. Výpočet p -hodnoty provedeme opět v OpenOffice. Do tabulky si zaneseme data:



Ověřování hypotéz



The screenshot shows an OpenOffice Calc spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2		hodina		1	2	3	4	5	6	7	8		
3		počet zákazníků		29	7	27	61	87	110	101	42		
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													
21													
22													

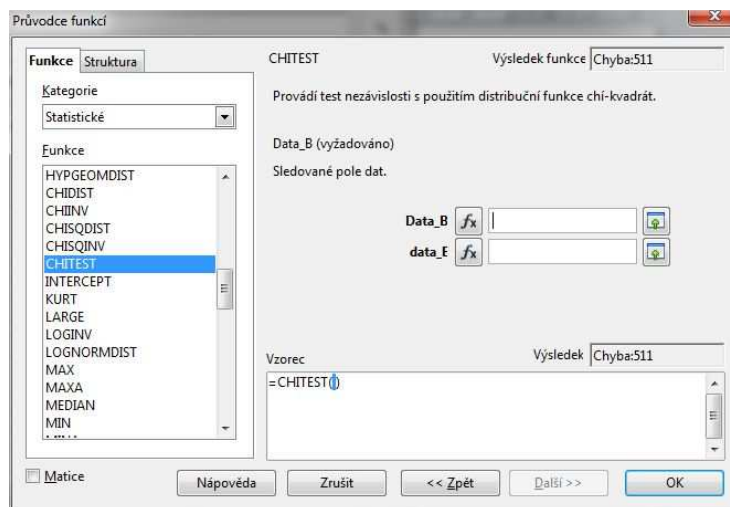
Při výpočtu p -hodnoty si musíme dopočíst k naměřeným hodnotám ještě očekávané hodnoty kritéria. V případě, že naším kritériem je test rovnoměrnosti, znamená to, že celkový počet výsledků musíme rovnoměrně rozpočít mezi všech osm hodin, tedy celkový počet zákazníků vydělíme osmi a dostaneme očekávaný počet zákazníků v každé hodině.

Celkem během směny úřednice obsloužila 464 zákazníků, tzn. že 464 vydělíme 8 a obdržíme, že během každé hodiny by měla obsloužit 58 zákazníků. Do tabulky tedy přidáme do dalšího řádku očekávané hodnoty.

Ověřování hypotéz

1	A	B	C	D	E	F	G	H	I	J	K	L	M
2		hodina		1	2	3	4	5	6	7	8	celkem	
3		počet zákazníků		29	7	27	61	87	110	101	42	464	
4		očekávaný počet		58	58	58	58	58	58	58	58		
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													
21													
22													

Nyní již můžeme vypočítat p -hodnotu pomocí funkce CHITEST. Vybereme si buňku, kam chceme p -hodnotu vypočítat.



Do řádku Data_B zadáme oblast skutečných naměřených hodnot. Do řádku data_E zadáme oblast očekávaných hodnot.

The screenshot shows a spreadsheet with the following data:

hodina	1	2	3	4	5	6	7	8	celkem
počet zákazníků	29	7	27	61	87	110	101	42	464
očekávaný počet	58	58	58	58	58	58	58	58	

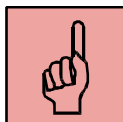
Below the table, the p-value is calculated as $p = 0,000$.

Hodnota p je rovna po zaokrouhlení 0,000.

5. Hodnota p je menší než hladina významnosti $\alpha = 0,05$. Nulovou hypotézu tedy zamítáme ve prospěch hypotézy alternativní.
6. Závěr: „**Výskyt zákazníků během směny je nevyrovnaný.**“

Na příkladu jsme si ukázali využití testu dobré shody při ověřování rovnoměrnosti rozložení počtu výskytů variant sledované proměnné. Testu dobré shody můžeme také využít, pokud chceme ověřit, zda data mají výsledky v určitém poměru. V tom případě se nejedná o test rovnoměrnosti, což si ukážeme v další části.

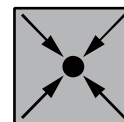
4.4.2 Chí-kvadrát test dobré shody – ověření určitého poměru rozložení četností



V této kapitole navážeme na předchozí kapitolu. Používat budeme stejného kritéria, jen rozložení počtu výskytů nebude v tomto případě rovnoměrné, ale bude odpovídat konkrétnímu poměru.

Příklad:

V určitém kraji byla v rámci dotazníku mimo jiné testována oblíbenost fyziky u žáků 7. tříd ZŠ. Do testování bylo vybráno 150 žáků základních škol onoho kraje.



H: Více než 60 % žáků 7. tříd v daném kraji nemá rádo fyziku.

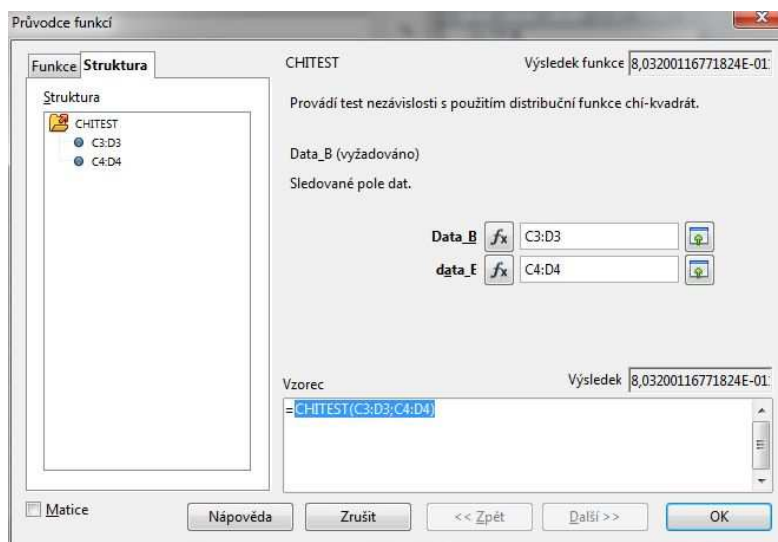
1. Věcná hypotéza: Více než 60 % žáků 7. tříd v daném kraji nemá rádo fyziku.
2. H_0 : Rozložení počtu žáků neoblíbená „fyzika : oblíbená fyzika“ je 60 : 40.
 H_1 : Rozložení počtu žáků je pro neoblíbenost fyziky více než 60 %.
3. Hladina významnosti $\alpha = 0,05$, jako kritérium použijeme chí-kvadrát test dobré shody.
4. Vypočteme p -hodnotu. Do jednoho řádku uvedeme skutečné počty, do druhého řádku uvedeme očekávané počty, tj. odpovídající poměru 60 : 40. Rozložíme-li počet 150 v poměru 60 : 40, obdržíme pro neoblíbenost 90 a oblíbenost 60, tedy 90 : 60.

	A	B	C	D	E	F	G	H
1								
2		Fyzika	oblíbená	neoblíbená				
3		skutečný počet	21	129				
4		očekávaný počet	60	90				
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								

V první řadě ale ještě zkontrolujeme, že alternativní hypotéze odpovídá skutečné rozložení četností, tzn. že fyzika je neoblíbená u více než 60 % žáků. 60 % ze 150 je 90,

Ověřování hypotéz

a skutečný počet žáků, kteří nemají rádi fyziku je 129, což je opravdu více než očekávaných 90. Kdyby rozložení četností neodpovídalo alternativní hypotéze, museli bychom přijmout závěr, že „Není pravda, že více než 60 % žáků 7. tříd v daném kraji nemá rádo fyziku“. V našem případě ale můžeme pokračovat ve výpočtu p -hodnoty. Vybereme buňku pro výslednou hodnotu p . Zadáme v OpenOffice funkci CHITEST, zadáme oblasti dat (viz chí-kvadrát test dobré shody – test rovnoměrnosti).



A vypočteme výslednou hodnotu p .

	A	B	C	D	E	F	G	H
1								
2		Fyzika	oblíbená	neoblíbená				
3		skutečný počet	21	129				
4		očekávaný počet	60	90				
5								
6								
7								
8								
9			p=	0,000				
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								

Výsledná hodnota p je rovna 0,000 po zaokrouhlení.

- Hodnota p je menší než hladina významnosti $\alpha = 0,05$. Tedy nulovou hypotézu zamítáme ve prospěch hypotézy alternativní na dané hladině významnosti.
- Nyní můžeme vyslovit závěr, že „**Více než 60 % žáků 7. tříd v daném kraji nemá rádo fyziku.**“

4.4.3 Chí-kvadrát test nezávislosti v kontingenčních tabulkách

Jedním z nejpoužívanějších testů je *test nezávislosti v kontingenčních tabulkách*. Testem nezávislosti ověřujeme závislost dvou kategorizovaných proměnných zpravidla nominálních, občas ordinálních. Kategorizovaná proměnná je proměnná, jejíž škálu tvoří pár hodnot, tzv. kategorií. V případě testu nezávislosti formulujeme statistické hypotézy následovně:



H_0 : Proměnná 1 a proměnná 2 jsou nezávislé.

H_1 : Proměnná 1 a proměnná 2 jsou závislé.

V OpenOffice použijeme stejnou funkci jako u testu dobré shody:

`CHITEST(Data_B, data_E)`.



V EXCELU použijeme funkci: `CHITEST(aktuální, očekávané)`.

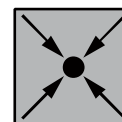
V EXCELU 2010: `CHSQ.TEST(aktuální, očekávané)`.

Data_B (aktuální): oblast dat naměřených hodnot,

data_E (očekávané): oblast dat očekávaných četností.

Příklad:

V jednom výzkumu v určitém roce byla zjišťována závislost mezi stupněm vzdělání a průměrným měsíčním příjmem. Průměrná měsíční mzda byla rozdělena do několika kategorií, stupeň vzdělání byl definován jako základní, střední a vysokoškolské. Výsledky jsme zanesli do tabulky v OpenOffice.



Ověřování hypotéz

Naměřené hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	293	156	3	452
7000 - 12999	876	609	34	1519
13000 - 18999	712	667	68	1447
19000 - 24000	173	112	18	303
nad 24000	67	25	2	94
součet	2121	1569	125	3815

Očekávané hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000				
7000 - 12999				
13000 - 18999				
19000 - 24000				
nad 24000				
součet				

Podobně jako u testu dobré shody musíme dopočítat očekávané hodnoty četností. Je to takové rozložení četností odpovídající nezávislosti uvedených proměnných. Očekávané četnosti dopočteme tak, že do tabulky stejné jako původní doplníme pouze celkové součty v řádku 24 a ve sloupci F. Dostaneme tak následující tabulku.

Očekávané hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000				452
7000 - 12999				1519
13000 - 18999				1447
19000 - 24000				303
nad 24000				94
součet	2121	1569	125	3815

Ověřování hypotéz

Nyní dopočteme očekávané četnosti uvnitř tabulky tak, že vynásobíme příslušný řádkový součet příslušným sloupcovým součtem a vydělíme celkovým počtem hodnot, tedy rozsahem souboru. Chceme-li tedy určit první očekávanou četnost v buňce C19, vynásobíme řádkový součet v řádku 19 (hodnota 452) sloupcovým součtem ve sloupci C (hodnota 2121) a vydělíme celkovým počtem (buňka F24 – hodnota 3815). První očekávaná četnost v buňce C19 je rovna $(452 \cdot 2121) / 3815 = 251$. Podobně určíme další očekávanou četnost v buňce C20. Obdobně jako v předchozím případě vypočteme očekávanou četnost vynásobením řádkové četnosti sloupcovou četností a vydělením celkovým rozsahem souboru, tedy zapsáno pomocí buněk: $F20 \cdot C24 / F24$, což je $1519 \cdot 2121 / 3815 = 845$. Tak dopočteme všechny hodnoty očekávaných četností.

The screenshot shows an OpenOffice Calc spreadsheet with two tables. The first table, 'Naměřené hodnoty', shows observed frequencies for different income levels and education levels. The second table, 'Očekávané hodnoty', shows the expected frequencies calculated for the same categories. The expected frequency for the first cell (C19) is 251.

Naměřené hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	293	156	3	452
7000 - 12999	876	609	34	1519
13000 - 18999	712	667	68	1447
19000 - 24000	173	112	18	303
nad 24000	67	25	2	94
součet	2121	1569	125	3815

Očekávané hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	251	186	15	452
7000 - 12999	845	625	50	1519
13000 - 18999	804	595	47	1447
19000 - 24000	168	125	10	303
nad 24000	52	39	3	94
součet	2121	1569	125	3815

Nyní můžeme přejít k výpočtu p -hodnoty. Určíme buňku, kam chceme vypočíst hodnotu p . Vybereme funkci CHITEST a zadáme oblasti dat. Do řádku Data_B zadáme oblast četností v tabulce Naměřené hodnoty bez řádkových a sloupcových četností v řádku 11 a sloupci F. V následujícím obrázku jde o červeně orámovanou vybranou oblast.

Ověřování hypotéz

Průvodce funkcí - CHITEST(Data_B; ...)

C6:E10

=CHITEST()

Naměřené hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	293	156	3	452
7000 - 12999	876	609	34	1519
13000 - 18999	712	667	68	1447
19000 - 24000	173	112	18	303
nad 24000	67	25	2	94
součet	2121	1569	125	3815

Očekávané hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	251	186	15	452
7000 - 12999	845	625	50	1519
13000 - 18999	804	595	47	1447
19000 - 24000	168	125	10	303
nad 24000	52	39	3	94
součet	2121	1569	125	3815

Nyní vybereme oblast do řádku data_E vybereme obdobnou oblast v tabulce Očekávané hodnoty (viz následující obrázek).

Průvodce funkcí - CHITEST(Data_B; ...)

C19:E23

=CHITEST(C6:E10;)

Naměřené hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	293	156	3	452
7000 - 12999	876	609	34	1519
13000 - 18999	712	667	68	1447
19000 - 24000	173	112	18	303
nad 24000	67	25	2	94
součet	2121	1569	125	3815

Očekávané hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	251	186	15	452
7000 - 12999	845	625	50	1519
13000 - 18999	804	595	47	1447
19000 - 24000	168	125	10	303
nad 24000	52	39	3	94
součet	2121	1569	125	3815

A nyní dokončíme výpočet hodnoty p .

The screenshot shows a spreadsheet with the following data:

Naměřené hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	293	156	3	452
7000 - 12999	876	609	34	1519
13000 - 18999	712	667	68	1447
19000 - 24000	173	112	18	303
nad 24000	67	25	2	94
součet	2121	1569	125	3815

Očekávané hodnoty				
Průměrný měsíční příjem	Stupeň vzdělání			součet
	základní	střední	vysokošk.	
nižší než 7000	251	186	15	452
7000 - 12999	845	625	50	1519
13000 - 18999	804	595	47	1447
19000 - 24000	168	125	10	303
nad 24000	52	39	3	94
součet	2121	1569	125	3815

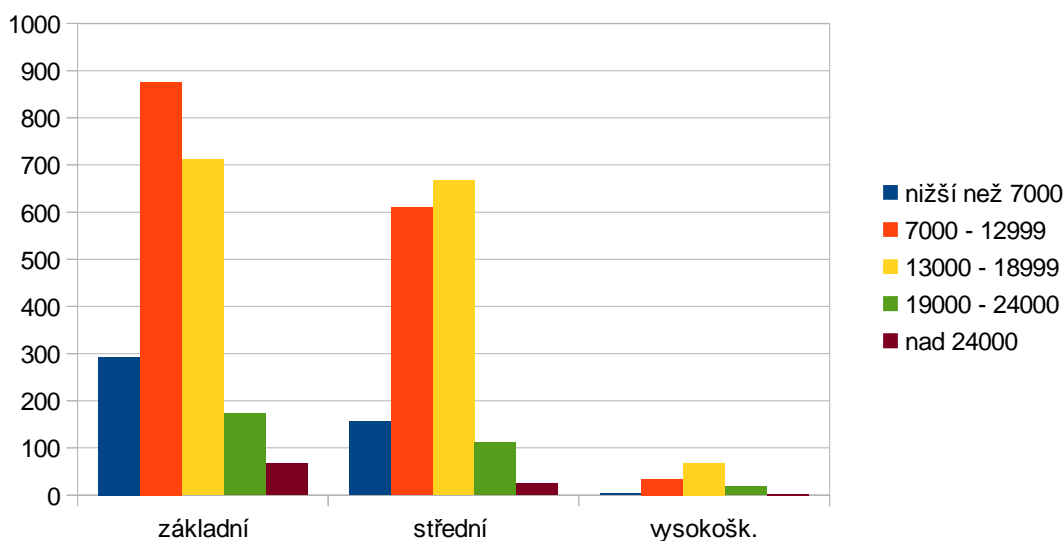
The formula bar shows: $=\text{CHITEST}(C6:E10;C19:E23)$

Cell H6 contains: $p=$ 0,000

Hodnota p je po zaokrouhlení rovna 0,000.

- p -hodnota je menší než zvolená hladina významnosti, nulovou hypotézu zamítáme ve prospěch hypotézy alternativní na hladině významnosti 0,05.
- Vyslovíme závěr: „**Stupeň vzdělání ovlivňuje průměrný měsíční plat.**“

Nyní bychom se ještě zaměřili, jakým způsobem stupeň vzdělání průměrný měsíční příjem ovlivňuje. Musíme analyzovat rozložení četností v tabulce naměřených hodnot četností. Z tabulky naměřených hodnot můžeme vysledovat, že s rostoucím vzděláním roste i průměrný měsíční příjem. U základního vzdělání je nejvíce respondentů s průměrným měsíčním příjmem 7000-12999, u středního vzdělání je jich nejvíce v kategorii 13000-18999, u vysokoškoláků je jich nejvíce opět v kategorii 13000-18999, ale je jich v poměru ke kategorii 7000-12999 několika násobně více než u středoškoláků. Ještě lépe můžeme rozložení četností u jednotlivých stupňů vzdělání pozorovat v následujícím grafu.



4.4.4 Další neparametrické testy jako alternativy k testům parametrickým

V poslední části této kapitoly si pouze uvedeme, jak postupovat v případě, že nemáme splněny podmínky pro použití parametrických testů, tedy naše data neodpovídají určitému typu rozdělení, v našem případě rozdělení normálnímu.

Ke každému parametrickému testu existuje 1 či více alternativních neparametrických testů. Tyto neparametrické testy bohužel nejsou součástí ani OpenOffice, ani MS EXCEL a je třeba k jejich výpočtu použít nějaký jiný program. Většina statistických programů jsou drahé a pro občasnou potřebu zbytečné. Existuje jednoduchý program PAST, což je zkratka PAleontologic STatistics, který je jednoduchý na zadávání dat, čtení výstupů je taktéž jednoduché a obsahuje širokou škálu statistických prostředků včetně všech potřebných neparametrických testů.

Nyní si pouze shrneme, které neparametrické testy jsou alternativou k testům parametrickým.

V případě, že nemůžeme použít dvouvýběrový t-test, pak nám jej může nahradit neparametrický *Mann-Whitney U-test*. V případě Mann-Whitney U-testu formulujeme statistické hypotézy:

H_0 : Mezi dosaženými výsledky obou skupin žáků nejsou rozdíly.

H_1 : Výsledky chlapců a dívek v testu z chemie jsou rozdíly.

Jelikož nemůžeme počítat průměry, pak v těchto testem nepoužíváme jednostranné hypotézy.

Namísto párového t-testu můžeme použít *znaménkový test* nebo *Wilcoxonův test*.

Statistické hypotézy:

H_0 : Mezi výsledky v testu před a po procvičení není rozdíl.

H_1 : Po procvičení jsou výsledky lepší než před procvičením.

V případě F-testu můžeme místo něj použít neparametrický Kolmogorovův-Smirnovův test.

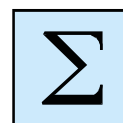
Statistické hypotézy:

H_0 : Složení obou skupin respondentů je stejné.

H_1 : Složení obou skupin respondentů je různé.

Shrnutí kapitoly

- Při ověřování hypotéz nejprve formulujeme věcnou hypotézu, na jejím základě pak formulujeme statistické hypotézy.
- Ověřování hypotéz provádíme vždy s určitým rizikem zamítnutí pravdivé nulové hypotézy, toto riziko se nazývá hladina významnosti a volíme jej zpravidla 0,05.
- Rozhodnutí o přijetí či zamítnutí nulové provádíme na základě srovnání hodnoty p s hladinou významnosti.
- Pro výpočet hodnoty p používáme kritérium, které vybíráme na základě toho, jaký typ hypotézy ověřujeme.
- Testy dělíme na parametrické a neparametrické.



- Podle tvaru alternativní hypotézy dělíme testy na jednostranné a oboustranné.



Kontrolní otázky a úkoly:

1. Co jsou to statistické hypotézy?
2. Jaký je rozdíl mezi jednostranným a oboustranným testem?
3. Které znáte parametrické a které neparametrické testy?
4. Jaký je postup při ověřování platnosti hypotéz?
5. Kdy používáme párový t-test?
6. Kdy používáme analýzu rozptylu?
7. Jaké typy dat musíme mít při chí-kvadrát testu nezávislosti v kontingenčních tabulkách?

Korespondenční úkoly



1. Rozhodněte, které z následujících formulací jsou hypotézy a formulujte k nim nulovou a alternativní hypotézu:
 - Učitelé, kteří používají ve třídě humor, dosahují lepší učební výsledky než učitelé, kteří nepoužívají humor.
 - Český denní tisk věnuje málo pozornosti problematice školy.
 - Žáci druhého ročníku mají školu raději než žáci pátého ročníku.
 - Absolventů gymnázií je na vysokých školách více než jiných absolventů.
 - Studium ve skupině přispívá ke zvýšení výkonu studentů.
 - Mezi vzděláním rodičů a inteligencí dítěte je pozitivní vztah.
 - Jak učitel zvýší počet pochval žáků, tak se žáci více naučí.
 - Předpokládáme, že se ne vždy používají adekvátní formy spolupráce školy s rodinou.
 - Čím větší je kohezivnost skupiny, tím větší je její vliv na jednotlivé členy.
 - Dívky dosahují v testech jazykových schopností lepší výkony než chlapci.
2. Zformulujte věcnou hypotézu, naměřte data a ověřte platnost této hypotézy.

5 Regrese a korelace

Cíl kapitoly

Po nastudování této kapitoly byste měli být schopni:

- určit regresní parametry jednoduché lineární regresní závislosti,
- odhadovat hodnoty závislé proměnné pro jakékoliv nezávislé proměnné,
- určit Pearsonův korelační koeficient,
- rozhodnout o těsnosti závislosti na základě korelačního koeficientu.

Klíčová slova: statistická závislost, regresní analýza, jednoduchá lineární regrese, regresní parametry, regresní přímka, korelační analýza, korelační koeficient, Pearsonův korelační koeficient, těsnost korelační závislosti.

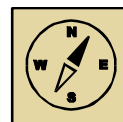


Čas na prostudování kapitoly

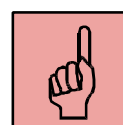


Průvodce studiem

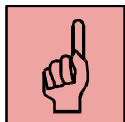
V této kapitole se seznámíte s analýzou závislosti dvou statistických metrických proměnných. Budete zkoumat tvar závislosti a dále těsnost zjištěné závislosti



Úkolem regrese a korelace je hledání závislostí mezi dvěma a více metrickými proměnnými u jednoho statistického souboru. Závislost mezi metrickými proměnnými může být buďto funkční nebo statistická. Proměnné vystupující v závislosti dělíme na nezávislé a závislé. Které volíme jako nezávislé a které jako závislé je z hlediska statistického výzkumu irelevantní, vyplývá to především z praktické podstaty výzkumu. Budeme-li zkoumat závislost výšky na hmotnosti žáků 9. tříd, pak si můžeme vybrat, kterou proměnnou si zvolíme jako nezávislou a kterou jako závislou. Jestliže budeme



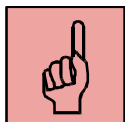
zkoumat závislost délky skoku do dálky na rychlosti skokana, pak jako nezávislou proměnnou budeme volit rychlost skokana, protože ta je příčinou délky skoku, tzn. že délka skoku závisí na rychlosti skokana.



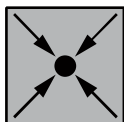
Funkční a statistickou závislost si ukážeme na případu dvou proměnných, tedy jedné nezávislé a jedné závislé proměnné. *Funkční závislost* je taková, která je dána předpisem

$$y = f(x).$$

V tomto případě jedné nezávislé proměnné odpovídá vždy pouze jedna závislá proměnná. Ze statistického hlediska je pro nás funkční závislost nezajímavá.

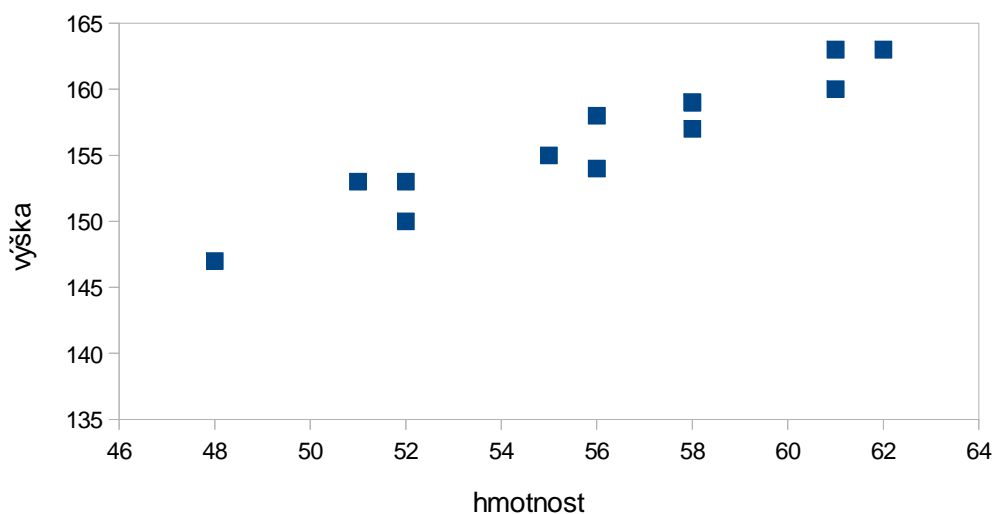


Statistická závislost je taková, kdy jedné nezávislé proměnné odpovídá rozložení závislé proměnné, zpravidla jde o normální rozložení výsledků. Ukážeme si statistickou závislost na příkladu.



Příklad:

Budeme zkoumat závislost výšky na hmotnosti žáků devátých tříd ZŠ. Nezávislá proměnná pro nás bude hmotnost v kg, závislá proměnná bude výška v cm. Při měření zjistíme, že jedné hmotnostní kategorii např. 56 kg bude odpovídat celá řada žáků s různými výškami, např. od 148 cm do 166 cm. Rozložení výšek pro tuto hmotnost bude s největší pravděpodobností normální s průměrnou hodnotou 158 cm. Nejvíce hodnot se bude pohybovat v okolí průměrné hodnoty a se vzdáleností od průměru budou počty ostatních hodnot klesat. Obdobně tomu bude pro kategorii 57 kg, kde hodnoty výšky budou mít opět normální rozložení s průměrnou hodnotou např. 158,5 cm apod. Tak jak se mění hmotnost, mění se i rozložení výšek žáků, ale pro každou hmotnost je velké množství žáků s různými výškami, které se řídí normálním rozložením.



Na jednoduchém grafu můžeme pozorovat statistickou závislost výšky na hmotnosti. Pro tak málo respondentů nemůžeme vypočítat, že výšky pro danou hmotnost mají normální rozdělení, ale vidíme, že jedné hmotnosti může odpovídat více výšek žáků.

Cílem regresní a korelační analýzy je zkoumání těchto statistických závislostí.

5.1 Regresní analýza

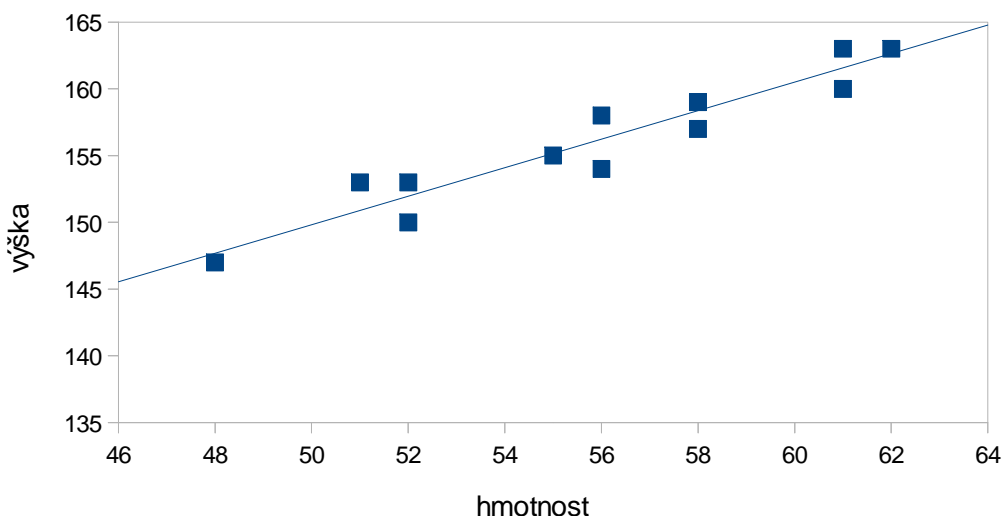
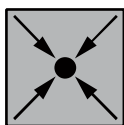
Regresní analýza se zabývá hledáním funkční závislosti, která by co nejlépe vystihovala danou statistickou závislost. V případě dvou proměnných se jedná o jednoduchou regresi, v opačném případě jde o vícenásobnou regresi. V případě jednoduché regrese je cílem proložit bodovým grafem závislosti křivku tak, aby odchylky od skutečných hodnot byly co nejmenší.

Dle toho, jakou křivku prokládáme, dělíme jednoduchou regresi na

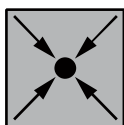
- lineární,
- nelineární.

V případě *jednoduché lineární regrese* prokládáme grafem závislosti přímkou, v opačném případě jde o obecnou křivku. V další části textu se budeme zabývat pouze jednoduchou lineární regresí. Přímka proložená naším grafem by byla takto.

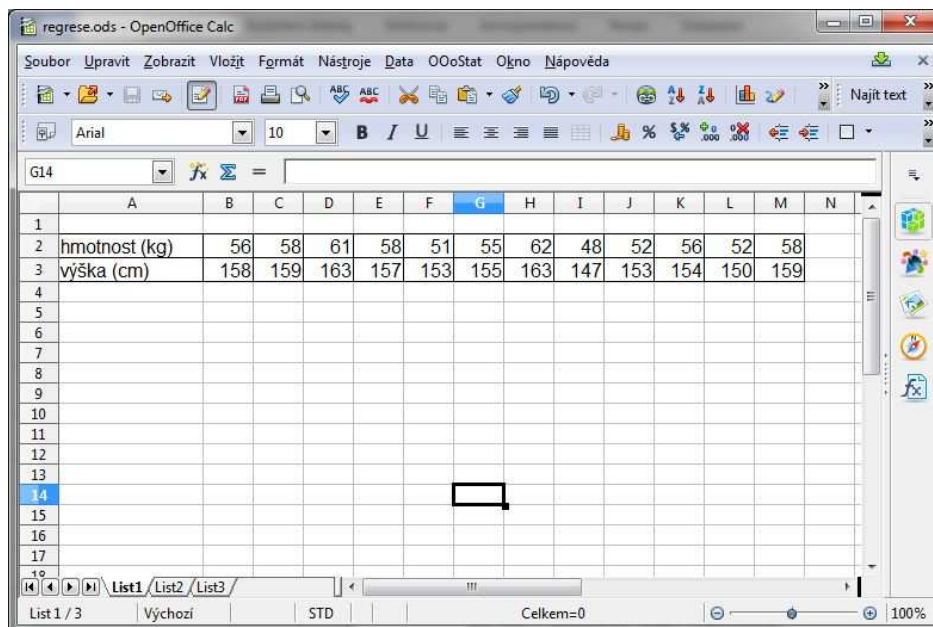




K určení proložené přímky nám poslouží funkce Proložit regresní přímku v OpenOffice. V případě, že bychom chtěli určit rovnici regresní přímky, použijeme opět funkce v OpenOffice, která se nachází v kategorii Matice a nazývá se LINEST. V případě použití MS EXCEL použijeme funkci LINREGRESE.



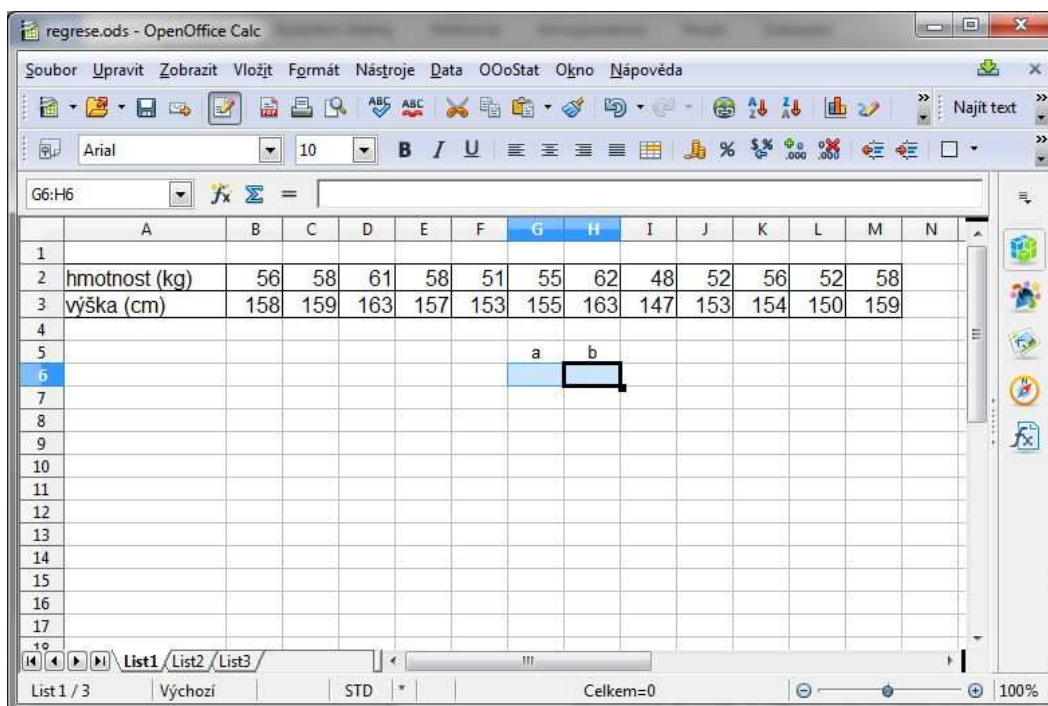
Příklad:



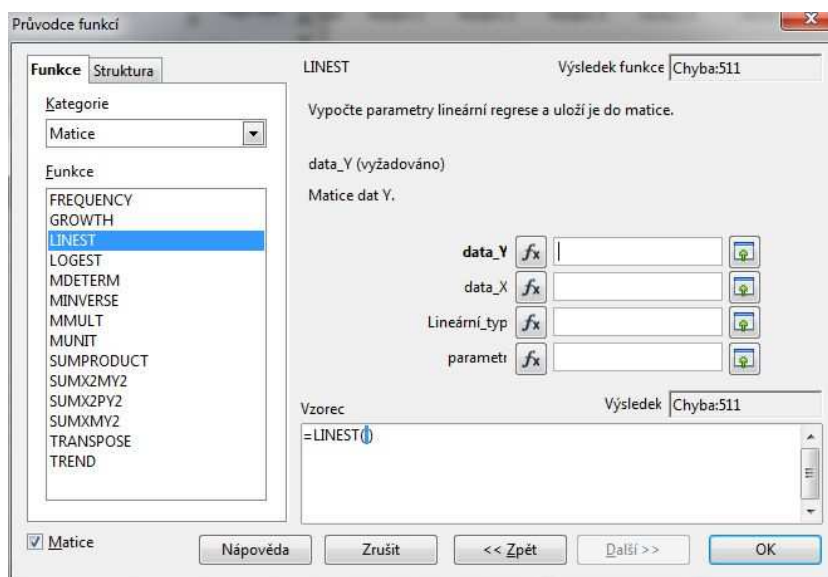
V OpenOffice máme data odpovídající závislosti výšky žáků na hmotnosti zobrazené v grafu na předchozí straně. Pro určení regresní rovnice musíme určit regresní

Regrese a korelace

parametry přímky. V OpenOffice si vybereme místo, kam chceme regresní parametry určit. Regresní parametry přímky jsou dva, tudíž si označíme dvě buňky vedle sebe.



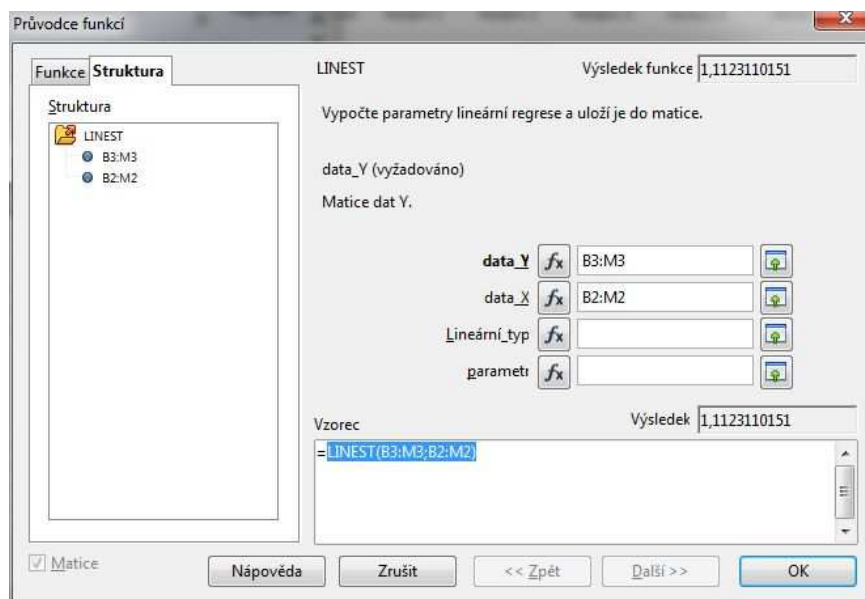
Nyní můžeme přejít k výpočtu regresních parametrů pomocí funkce LINEST. V kategoriích vybereme Matice ve funkcích najdeme LINEST.



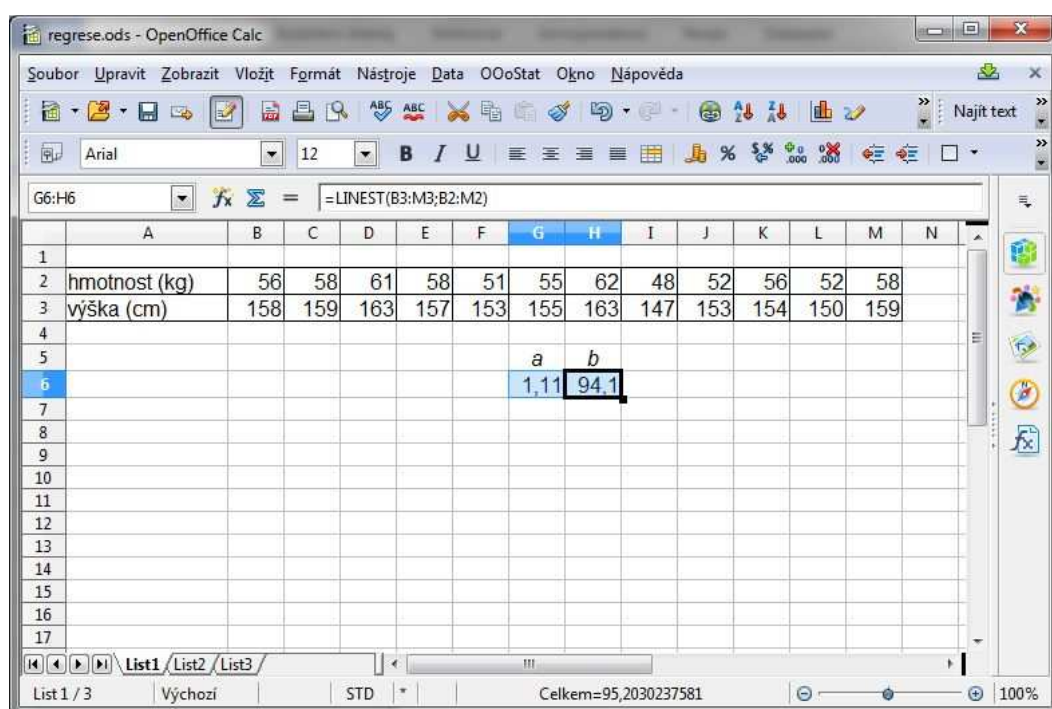
Objeví se nám čtyři řádky, pro nás jsou důležité v případě jednoduché lineární regrese pouze první dva, dalších dvou si nebudeme všimnout. Do prvního řádku označeného data_Y si vybereme oblast závislé proměnné, v našem případě řádek „výška“. Do

Regrese a korelace

druhého řádku označeného data_X vybereme oblast nezávislé proměnné, v našem případě řádek „hmotnost“.



Další dva řádky necháme prázdné a klikneme na OK pro provedení výpočtu regresních parametrů.



Obdrželi jsme regresní parametry $a = 1,11$ a $b = 94,1$. To znamená, že regresní rovnice má tvar

$$\text{výška} = 1,11 * \text{hmotnost} + 94,1.$$

Na základě regresní rovnice nyní můžeme odhadovat, kolem jaké hodnoty se budou pohybovat výšky žáků pro určitou vybranou hmotnost. Z výše uvedeného tedy můžeme vyslovit závěr, že *regresní rovnice* jednoduché lineární regrese bude mít vždy tvar

$$y = a * x + b,$$

kde a , b jsou regresní parametry určené na základě naměřených experimentálních hodnot, y je závislá proměnná a x je nezávislá proměnná.

Nyní již můžeme odhadovat, jak budou vypadat hodnoty závislé proměnné y pro dané hodnoty nezávislé proměnné x . Co ovšem nevíme, zda tyto odhady mají vůbec smysl počítat. Doposud nic nevíme o tom, jak dobře výše uvedená regresní rovnice dovede odhad provést. O tom jsme se v rámci regresní analýzy nic nedozvěděli, protože regresní analýza nám k tomu žádné prostředky neposkytuje. Pokud chceme vědět, jak dobrá závislost mezi závislou proměnnou a nezávislou proměnnou je, musíme použít jiný prostředek a tím je korelační analýza.

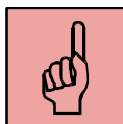
5.2 Korelační analýza

V rámci korelační analýzy se budeme zabývat pouze závislostí vyjádřené jednoduchou lineární regresí. Tedy budeme mít pouze jednu nezávislou a jednu závislou proměnnou a budeme zkoumat závislost proloženou přímkou.

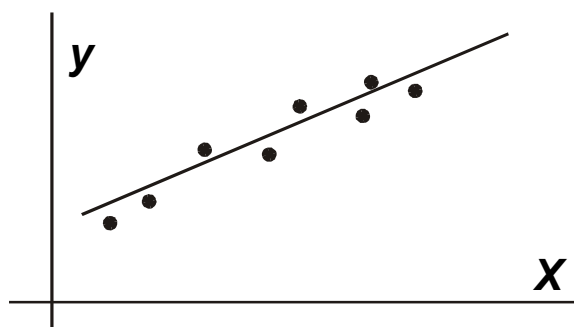
Lineární korelace zjišťuje, jak těsná je lineární závislost mezi dvěma proměnnými. Těsnost této závislosti se určuje na základě koeficientu, kterému se říká korelační koeficient.

Korelačních koeficientů je několik, nejčastěji používaný je *Pearsonův korelační koeficient*. Jeho hodnota je od -1 do 1, a čím se blíží více k 1 v absolutní hodnotě, tím je závislost těsnější, tím jsou body těsněji seskupeny kolem regresní přímky a tím nám regresní přímka lépe danou závislost popisuje. Čím se koeficient blíží více k nule, tím je závislost slabší, a tím horší odhady můžeme na základě regresní rovnice provádět.

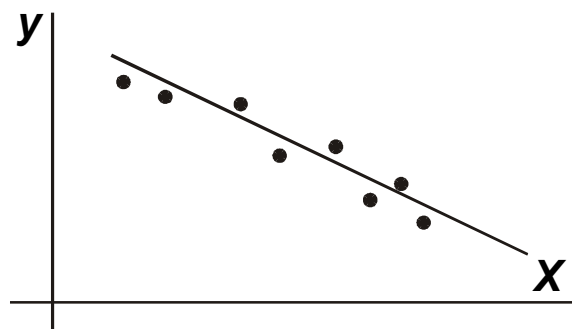




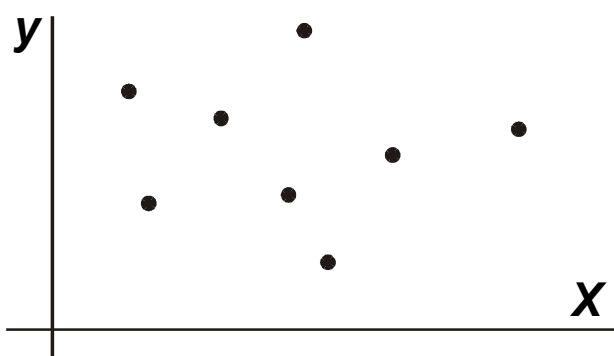
V případě, že koeficient je kladný, pak regresní funkce je rostoucí. V případě, že koeficient je záporný, pak regresní funkce je klesající. Na základě znaménka a velikosti korelačního koeficientu však nemůžeme říct nic o sklonu regresní přímky.



Statistická závislost s vysokým korelačním koeficientem blížícím se k jedné.



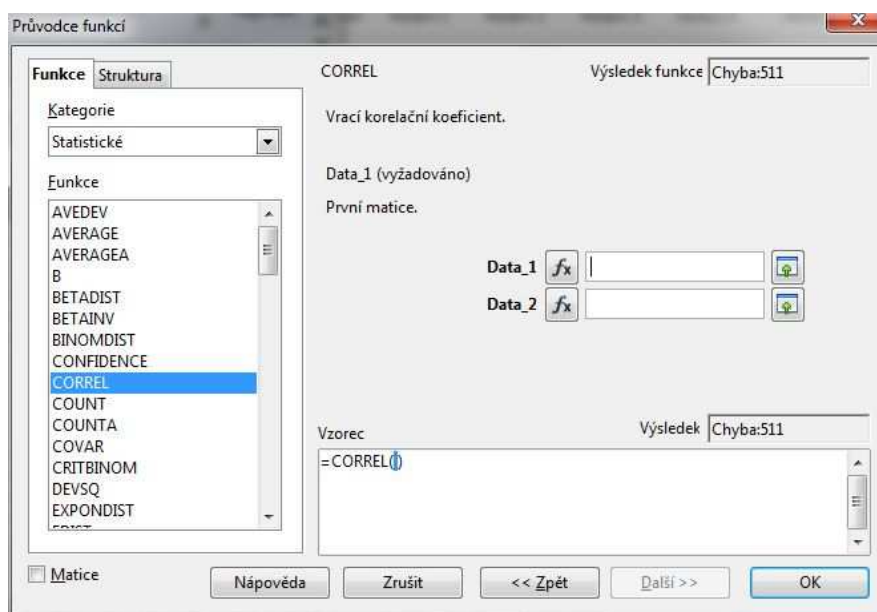
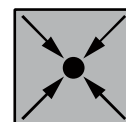
Statistická závislost s vysokým korelačním koeficientem blížícím se k minus jedné.



Statistická závislost s nízkým korelačním koeficientem blížícím se k nule.

Příklad:

V kapitole 5.1 Regresní analýza jsme se v příkladu věnovali závislosti výšky žáků na jejich hmotnosti. Určili jsme regresní závislost resp. regresní parametry a rovnici regresní přímky. Zbývá nám však určit, jak těsná je závislost mezi výškou a hmotností daná naměřenými experimentálními údaji. Pro výpočet v OpenOffice použijeme funkci CORREL se stejnými daty jako při výpočtu regresních parametrů, pomocí níž vypočteme Pearsonův korelační koeficient.



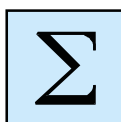
V případě výpočtu korelačního koeficientu máme řádky označeny Data_1 a Data_2. To je z toho důvodu, že při zjišťování těsnosti závislosti nezáleží která proměnná je závislá a která nezávislá, ale jde o vzájemný vztah dvou proměnných. Je proto jedno, do kterého řádku zadáme hmotnost, a do kterého výšku. Do řádků zadáme tedy naměřené hodnoty hmotnosti a výšky, a provedeme výpočet korelačního koeficientu.

The screenshot shows a spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2	hmotnost (kg)	56	58	61	58	51	55	62	48	52	56	52	58	
3	výška (cm)	158	159	163	157	153	155	163	147	153	154	150	159	
4														
5														
6														
7														
8	Pearsonův korelační koeficient				0,956									
9														
10														
11														
12														
13														
14														
15														
16														
17														

Hodnota Pearsonova korelačního koeficientu je rovna 0,956, což je hodnota velmi blízká jedné. Můžeme říct, že korelační závislost je velmi těsná, téměř lineární a vzhledem k tomu, že hodnota korelačního koeficientu je kladná, můžeme nazvat vztah obou proměnných za pozitivní.

Shrnutí kapitoly



- Mezi metrickými proměnnými může existovat závislost, která může být funkční nebo statistická.
- Závislost mezi metrickými proměnnými nazýváme statistickou, jestliže jedné hodnotě nezávislé proměnné odpovídá celá množina hodnot závislé proměnné, která má určité rozdělení.
- Zkoumáním tvaru rozdělení se zabývá regresní analýza, hledá tvar křivky, která by co nejlépe popsala danou statistickou závislost.
- Korelační analýza zkoumá, jak dobře regresní křivka danou statistickou závislost popisuje.

Kontrolní otázky a úkoly:

1. Jaký je rozdíl mezi funkční a statistickou závislostí.
2. Co jsou to regresní parametry a jaký má předpis jednoduchá lineární regresní závislost?
3. Jakých hodnot může nabývat Pearsonův korelační koeficient?
4. Můžeme na základě korelačního koeficientu určit sklon regresní přímky?



6 Přehled použité literatury



CYHELSKÝ, L. *Úvod do teorie popisné statistiky*. Praha : SNTL, 1974.

GAVORA, P. *Úvod do pedagogického výzkumu*. Brno : Paido, 2000. ISBN 80-85931-79-6.

HENDL, J. *Přehled statistických metod zpracování dat : analýza a metaanalýza dat*.

Praha : Portál, 2004. ISBN 80-7178-820-1.

CHRÁSKA, M. *Metody pedagogického výzkumu. Základy kvantitativního výzkumu*.

Praha : Grada, 2007. ISBN 978-80-247-1369-4.

KOMENDA, S. *Základy pravděpodobnostních a statistických metod v psychologickém a pedagogickém výzkumu*. Praha : SPN, 1967.

PELIKÁN, J. *Základy empirického výzkumu pedagogických jevů*. Praha : Karolinum,

1998. ISBN 80-7184-569-6.

ŘEHÁK, J.; ŘEHÁKOVÁ, B. *Analýza kategorizovaných dat v sociologii*. Praha : Academia, 1986.