UNIVERZITA J. E. PURKYNĚ V ÚSTÍ NAD LABEM
Ústav zdravotnických studií

ZÁKLADY BIOSTATISTIKY S VYUŽITÍM EXCELU

Karel Hrach

Základy biostatistiky s využitím Excelu

Karel Hrach

Tento projekt je součástí
IPRM Ústí n. L. – Centrum.

Tato publikace vznikla v rámci projektu Posilování kompetencí vysokoškolských pracovníků pro rozvoj konkurenceschopnosti vysokého školství v Ústeckém kraji, registrační číslo CZ.1.07/2.2.00/07.0117, realizovaného v rámci OP Vzdělávání pro konkurenceschopnost.

http://pokrok.ujep.cz

Obsah

Poznámky úvodem 4

Literatura 4

Statistické veličiny 5
- Základní pojmy 5
- Textové veličiny 5
- Číselné veličiny 6
- Poznámky k typům veličin 7

Deskriptivní charakteristiky kategoriálních veličin 8
- Absolutní a relativní četnosti 8
- Kumulativní četnosti 9
- Grafické znázornění četností a modus 10
- Problém „multiple responses“ 12
- Věrohodnostní poměr a podíl šancí 13

Deskriptivní charakteristiky číselných veličin 14
- Medián 14
- Aritmetický průměr 15
- Rozptyl a směrodatná odchylka 15
- Geometrický průměr 17

Metody statistické indukce – parametrické odhady 18
- Parametry pravděpodobnostních modelů 18
- Bodové odhady 18
- Intervalové odhady aneb intervaly spolehlivosti 20

Metody statistické indukce – parametrické testy 22
- Úvod do testování hypotéz – nulová a alternativní hypotéza 22
- Princip testování hypotéz – hladina významnosti a p-hodnota 23
- Párový t-test 24
- Dvou-výběrový F-test 26
- Dvou-výběrový t-test 28
- Jedno-faktorová analýza rozptylu (ANOVA) 29
- Statistická versus klinická významnost 32

Metody statistické indukce – regresní modely 33
- Jednoduchá regrese a korelace 33
- Vícenásobná regrese a korelace 37
- Regrese s využitím dummy proměnných 39

Metody statistické indukce – testy typu chí-kvadrát 41
- Test dobré shody 41
- Test nezávislosti 43

Příloha – ukázka postupu při získání dat z www databáze ÚZIS 47

Poznámky úvodem

Kurz realizovaný na jaře 2011 na ÚZS UJEP byl věnován statistickým analýzám, obvyklým v medicínské statistice, s důrazem na ukázku jejich provedení pomocí běžně dostupného SW MS-Office Excel (používána byla verze 2007). Tento studijní materiál si klade za cíl být užitečnou pomůckou všem, kteří se potřebují seznámit stručně se základy praktického statistického zpracování dat ve zdravotnicky zaměřených oborech. Obsah byl koncipován zejména podle série učebnic Biomedicínská statistika autorského kolektivu vedeného prof. Zvárovou, jehož jsem byl členem, a dále podle mých skript, především Sbírka úloh ze statistiky. Technické detaily, tedy vzorce či dokonce důkazy jejich platnosti, byly zcela vynechány, nebo alespoň omezeny na nejnutnější minimum. K podrobnému nastudování statistických metod je proto rozhodně nutno použít jinou odbornou literaturu, např. Statistické metody prof. Anděla.

Technická poznámka: Symbol □ označuje v textu konec příkladu.

Literatura

- ANDĚL, Jiří. *Statistické metody*. 4. vydání, 2007. Praha: Matfyzpress. ISBN 80-7378-003-8.
- BYSTROŇ, Marian – ČERVINKA, Pavel – ŠPAČEK, Radim, a kol. Randomized Comparison of Endothelial Progenitor Cells Capture Stent Versus Cobalt-Chromium Stent for Treatment of ST-Elevation Myocardial Infarction. *Catheterization and Cardiovascular Interventions*, November 2010, vol. 76, no. 5, s. 627-631.
- HRACH, Karel. *Sbírka úloh ze statistiky*. 1. vydání, 2006. Ústí nad Labem: FSE UJEP. ISBN 80-7044-845-8.
- NETER, John – WASSERMAN, William – KUTNER, Michael H. *Applied Linear Statistical Models*. 3rd edition, 1990. Boston: Irwin. ISBN 0-256-08338-X.
- ZVÁROVÁ, Jana. *Biomedicínská statistika I. - Základy statistiky pro biomedicínské obory*. 1. vydání, 2001. Praha: Karolinum. ISBN 80-7184-786-0.
- ZVÁROVÁ, Jana – MALÝ, Marek, a kol. *Biomedicínská statistika III. - Statistické metody v epidemiologii*. 1. vydání, 2003. Praha: Karolinum. ISBN 80-246-0765-4.

Statistické veličiny

Základní pojmy

Ve statistice jsou zpracovávána data, neboli údaje zaznamenávající pro danou statistickou jednotku z nějaké populace (též: ze základního souboru - tedy z množiny všech statistických jednotek daného typu) příslušný sledovaný údaj neboli hodnotu sledované statistické veličiny (též: znaku). Populace mohou být konečné (kraje; nemocnice; pacienti;…) nebo, aspoň hypoteticky, nekonečné (všechny krevní vzorky, které by bylo možno odebrat;…). Vždy je důležité jasně specifikovat věcné, časové a místní vymezení sběru dat – tedy co, kdy a kde bylo získáváno. Důležitou roli hraje také to, zda data představují chování celé populace (pak mluvíme o úplném šetření), nebo (častěji) zda jsou sledované statistické jednotky pouze výběrem z celé populace (pak jde o výběrové šetření, zkráceně: výběr). V případě výběru požadujeme, aby byl tzv. reprezentativní, tedy aby v něm zahrnuté statistické jednotky „kopírovaly" důležité charakteristiky celé zkoumané populace. Např. jsou-li v celé populaci zastoupeni muži a ženy zhruba rovnoměrně, pak i ve výběru by měla být tato proporce zachována; výjimkou jsou ty případy, kdy bychom sledovali nějaký jev, který je zcela nezávislý na pohlaví - pak nebude pohlaví důležitou charakteristikou a nemusíme příslušnou populační proporci ve výběru dodržet. Poměrně spolehlivou zárukou reprezentativnosti bývají situace, kdy statistické jednotky vybereme metodou náhodného výběru, tedy např. losováním. Počet statistických jednotek ve výběru nazýváme rozsah výběru.

Příklad 1 (statistické veličiny a jejich hodnoty):
U statistické jednotky „pacient" s vymezením např. pacienti hospitalizovaní po jakékoli operaci – během prvního čtvrtletí – za všechna oddělení sledované nemocnice, můžeme sledovat např. veličiny „pohlaví" (s hodnotami muž – žena), „pocit po zákroku" (s možností volným textem vypsat své pocity), nebo „věk" (s hodnotou kladnou celočíselnou udávající věk v letech), atd.
U statistické jednotky „kraj" s vymezením, o jaký kraj se jedná a k jakému okamžiku má být údaj zaznamenán (např. k poslednímu dni daného kalendářního roku), můžeme sledovat např. veličiny „existence fakultní nemocnice v daném kraji" (s hodnotami ano – ne), „celkový počet nemocničních lůžek", atd.

Cílem statistiky je pak tzv. hromadné zpracování, tedy (souhrnně za všechny vybrané statistické jednotky) statistická deskripce (popis chování získaných dat) a pokud možno i hlubší statistická analýza vztahů mezi veličinami či statistická analýza příčin jejich chování, neboli statistická indukce (čímž je vlastně míněno zobecnění chování výběru na celou populaci).

Textové veličiny

Statistické (též náhodné) veličiny lze rozlišovat do několika různých typů podle toho, jakých hodnot mohou u konkrétních statistických jednotek nabývat. Veličiny nazýváme textové - otevřené, pokud je odpovědí volný text. V příkladu 1 šlo např. o veličinu „pocit po zákroku". Z pohledu statistického zpracování lze použití tohoto typu doporučit jen v opravdu odůvodněných případech, kdy nelze použít jinou možnost získání dat, protože tento typ prakticky vylučuje možnost jakéhokoliv automatického hromadného zpracování. Pokud potřebujeme získávat informace textového charakteru, je mnohem vhodnější použít veličiny textové – uzavřené (kategoriální), u nichž je hodnota vybírána z předem připravených variant (kategorií). Jediným problémem někdy může být to, abychom již při přípravě statistického průzkumu nezapomněli připravit varianty tak, aby každé statistické jednotce mohla být přiřazena odpovídající kategorie. Často např. jako poslední možnost uvádíme variantu „ostatní" nebo „jiné" (např. u veličiny „důvod hospitalizace na chirurgickém oddělení" bychom mohli uvádět jako hodnoty varianty „úraz", „plánovaný operační zákrok", „jiný důvod"). Jsou-li rozlišovány pouze dvě možné varianty, hovoříme speciálně o veličině alternativní

(dichotomické). V příkladu 1 se jednalo o veličiny „pohlaví“ či „existence fakultní nemocnice v daném kraji“. V případě, kdy je možný výběr z více než dvou variant, můžeme takovéto textové veličiny ještě dál rozlišovat na neuspořádané (nominální) a uspořádané (ordinální). Nominální veličiny jsou ty, kde mezi jejich kategoriemi neexistuje přirozené uspořádání ve smyslu, která hodnota je nižší či vyšší, resp. ve smyslu, která kategorie je lepší či horší. Příkladem takové veličiny je o pár řádků dříve zmíněná veličina „důvod hospitalizace na chirurgickém oddělení“, protože nelze říci, že by kategorie „úraz“ byla „víc“ či „líp“ (nebo naopak méně či hůře) než kategorie „plánovaný operační zákrok“. Naopak ordinální veličina je každá taková, pro jejíž kategorie existuje jejich logické uspořádání. Příkladem by byla veličina „pocit po zákroku“, u níž bychom jako možné hodnoty povolili výběr z kategorií „došlo k výraznému zlepšení“, „došlo k částečnému zlepšení“, „nedošlo k žádné změně“, „došlo ke zhoršení“.

Číselné veličiny

Dalším typem statistických veličin jsou veličiny nabývající číselných hodnot. Statistická veličina se nazývá číselná diskrétního typu (diskrétní veličina, diskrétně rozdělená veličina), pokud nabývá tzv. spočetně mnoha možných číselných hodnot. Obvykle jde o hodnoty kladné, často včetně nuly. Příkladem by byla veličina „počet dětí“, nabývající hodnot 0, 1, 2, atd. Jedná se vlastně o jakousi analogii s veličinou ordinálního typu: jednotlivé možné hodnoty lze považovat za jednotlivé kategorie, jednoznačně uspořádané podle velikosti. Druhým typem je veličina číselná spojitého typu (spojitá veličina, veličina se spojitým rozdělením), která může nabývat tzv. nespočetně mnoha možných číselných hodnot. Obvykle jde o hodnoty reálné, pokud jsou celočíselné, pak jde vlastně pouze o jejich celočíselné zaokrouhlení. Příkladem je veličina „věk pacienta“, která je obvykle uváděna v rocích. Pokud dva různí jedinci uvedou shodně věk 27 let, neznamená to ještě, že slaví narozeniny v tentýž den a že by tedy ve skutečnosti byli zcela shodného stáří. Dalšími příklady jsou „tělesná teplota“, „koncentrace látky XY v krevním vzorku“ atd., obecně řečeno patří sem všechny číselné údaje, které jsou uváděny v nějakých měrných jednotkách (fyzikálních, nebo třeba procentuálních). Poznamenejme, že číselné veličiny obou typů lze převést do tvaru uspořádané kategoriální veličiny (ale ne naopak) jednoduše tak, že zavedeme tzv. intervalové rozdělení, tj. že jednotlivé kategorie vymezíme rozmezím od-do. Např. veličinu „počet dětí“ lze rozlišovat v kategoriích „bezdětní“ (tedy původně číselná hodnota 0), „jedno nebo dvě děti“ (tedy původně číselné hodnoty 1 nebo 2) a „tři či více dětí“ (původně hodnoty ostatní). Podobně veličinu „tělesná teplota“ lze zaznamenat např. jako „kategorii 1“ (do 37°C včetně), „kategorii 2“ (nad 37°C, nejvýše však 40°C včetně) a „kategorii 3“ (nad 40°C); povšimněme si, že hraniční hodnoty mezi jednotlivými kategoriemi (zde 37°C a 40°C) musí být jednoznačně vymezeny tak, aby bylo jasné, do které kategorie patří, resp. nepatří.

Příklad 2 (použití funkce =KDYŽ):
V tabulce 1 jsou do Excelu přepsaná data z učebnice Statistické metody v epidemiologii, s nimiž budeme i zde ještě později pracovat. Nyní si na veličině „hmotnost“ (data ze sloupce D) předvedeme, jak bychom pro tuto veličinu automaticky zavedli intervalové rozdělení např. s kategorií 1 (zahrnující děti s hmotností do 50 kg včetně), s kategorií 2 (hmotnost nad 50 kg, nejvýše však 60 kg včetně) a kategorií 3 (hmotnost nad 60 kg). Lze k tomu v Excelu využít logickou funkci =KDYŽ s touto strukturou

=KDYŽ(podmínka;hodnotaANO;hodnotaNE) (1)

Prvním argumentem příkazu (1) je „podmínka“, tedy obecně jakékoli tvrzení, o němž Excel může rozhodnout, zda je či není splněno. Druhým argumentem (označen jako „hodnotaANO“) je konkrétní hodnota, kterou chceme získat jako výsledek celého příkazu v případě, že vyhodnocovaná podmínka platí. Posledním, třetím argumentem (označen jako „hodnotaNE“), je ta konkrétní hodnota, kterou chceme získat jako výsledek v případě, že vyhodnocovaná podmínka naopak neplatí. Novou, intervalovou veličinu „hmotnostní kategorie“, bychom tedy mohli definovat ve sloupci E tak, že v buňce E3 (tedy na řádku u prvního zaznamenaného dítěte) zapíšeme vzorec:

Tabulka 1: Data o věku, výšce a hmotnosti 13 dětí přepsaná do Excelu (zdroj Zvárová a kol.)

	A	B	C	D
1	Tabulka Věk, výška a hmotnost u 13 dětí			
2	Dítě (i)	Věk ($X_{1,i}$)	Výška ($X_{2,i}$)	Hmotnost (Y)
3	1	10	146	51
4	2	10	152	59
5	3	10	138	34
6	4	11	150	40
7	5	11	160	67
8	6	11	154	49
9	7	11	155	47
10	8	12	164	68
11	9	12	158	65
12	10	12	162	65
13	11	13	165	57
14	12	13	169	60
15	13	13	158	61

=1+KDYŽ(D3>50;1;0)+KDYŽ(D3>60;1;0) (2)

Jelikož u prvního dítěte je ve vyhodnocované buňce D3 hodnota hmotnosti 51 (kg), je splněna podmínka D3>50, ale není splněna podmínka D3>60. Znamená to, že hodnotou pro KDYŽ(D3>50;1;0) bude 1, ale hodnotou pro KDYŽ(D3>60;1;0) bude 0. Výsledná hodnota vzorce (2) pro první zaznamenané dítě tedy bude v buňce E3 činit 1+1+0=2, což znamená, že toto dítě bylo zařazeno do druhé hmotnostní kategorie. Teď již stačí vzorec z buňky E3 překopírovat do buněk E4 až E15, čímž Excel automaticky zjistí hodnotu příslušné hmotnostní kategorie i pro zbylé děti.

Poznámky k typům veličin

V datech v tabulce 1 se ve sloupci A vyskytl příklad tzv. identifikátoru, jednalo se o pořadové číslo každé statistické jednotky. Nejde o statistickou veličinu v pravém slova smyslu, není to údaj, který bychom dál statisticky zpracovávali, identifikátory slouží pouze k odlišení jednotlivých konkrétních pozorování. Jiným příkladem identifikátoru může být jméno a příjmení či rodné číslo (u pacientů), nebo třeba název, adresa sídla či IČO (u zdravotnických zařízení). Zvláště v případě osobních údajů pacientů se často volí ten postup, že tyto údaje jsou nahrazeny právě pořadovým číslem (přičemž pouze ten zdravotník, který data získával, má k dispozici seznam, podle nějž může každému pořadovému číslu zpětně přiřadit skutečnou identitu příslušného pacienta) a k dalšímu statistickému zpracování jsou předána již pouze anonymizovaná data třeba v podobě jako v tabulce 1.

Na závěr této kapitoly si odpovězme na otázku, proč je vlastně tak důležité umět rozlišovat jednotlivé typy veličin? Protože každý typ má odpovídající způsob statistického zpracování; to, co se hodí pro jeden typ, může být u jiného typu zcela nepoužitelné, nevhodné. Jednoduchým a názorným příkladem budiž určení průměrné hodnoty. Průměr má samozřejmě smysl počítat u číselných veličin (zjistíme tak běžně i laiky správně chápaný a interpretovaný průměrný počet dětí nebo průměrnou teplotu, připadající na jednoho pacienta), ale určitě ne u veličiny textové, např. u veličiny nominálního typu „barva očí“. Pokud bychom si její jednotlivé kategorie číselně kódovali (1=modrá, 2=zelená, 3=hnědá, 4=jiná), mohli bychom technicky z napozorovaných hodnot 1-4 jejich průměr spočítat, ale pokud by tento dejme tomu vyšel 1,972, o čem by tato hodnota „průměrné barvy očí“ vlastně vypovídala, jak bychom ji vlastně interpretovali? V dalším textu bude proto u jednotlivých

statistických metod uváděno, pro jaký typ veličin je můžeme používat, přičemž veličiny textové-otevřené nebudou již dále zmiňovány (jak již bylo řečeno, pro ně není automatické statistické zpracování použitelné a pokud je statisticky zpracovat chceme, musíme je nejprve uměle převést do podoby veličiny kategoriální).

Deskriptivní charakteristiky kategoriálních veličin

Absolutní a relativní četnosti

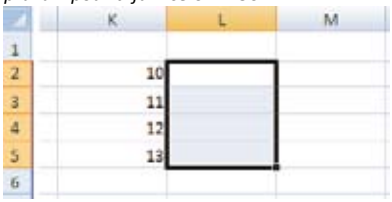
Četnosti lze obvykle smysluplně použít u všech veličin s výjimkou číselných-spojitých. Jinak řečeno, četnosti můžeme určovat vždy, kdy má smysl hovořit o jednotlivých kategoriích, ať už nečíselných či číselných. Např. u spojité veličiny „hmotnost“ z dat v tabulce 1 by nemělo smysl určovat četnosti jednoduše proto, že mezi pozorovanými 13 hodnotami byly (až na výjimky) všechny hodnoty vzájemně různé, takže nelze hovořit o opakovaném výskytu jednotlivých kategorií a nemá smysl pro tyto víceméně jedinečné hodnoty určovat jejich četnosti. Smysl by to ovšem mělo pro hmotnost zaznamenanou intervalově, kterážto veličina by vznikla úpravou podle příkladu 2. Pro ni už bychom mohli určovat četnosti jednotlivých tří hmotnostních kategorií.

Četnosti absolutní udávají skutečný počet výskytů dané kategorie v datech. Četnosti relativní udávají poměrný počet výskytů dané kategorie vzhledem k celkovému počtu pozorování, jde tedy vždy o hodnotu z rozmezí 0-1, přičemž obvykle ji vynásobením 100 převádíme na procenta.

Příklad 3 (absolutní četnosti a použití funkce =ČETNOSTI):
Na ukázku určíme absolutní četnosti veličiny „věk“ pro data z tabulky 1. Veličina věk patří sice mezi veličiny spojitého typu, v uvedených datech se však vyskytly pouze čtyři její různé hodnoty (věk 10, 11, 12 nebo 13 let), takže můžeme tyto čtyři hodnoty považovat za věkové kategorie a má smysl určit četnosti jejich výskytu. Vzhledem k tomu, že pozorování bylo celkem pouze 13 a navíc byla data uspořádána vzestupně právě podle veličiny věk, jsou z údajů ve sloupci B v tabulce 1 absolutní četnosti zřejmé na první pohled: popořadě mají pro jednotlivé věkové kategorie hodnoty 3, 4, 3 a 3. Interpretace je evidentní – např. do druhé věkové kategorie (tedy 11 let) patřily 4 děti ze všech sledovaných.

Nyní si předvedeme, jak bychom četnosti určili pomocí Excelu. (Užitečné to bude v případě, kdy by databáze byla mnohem rozsáhlejší a četnosti by nebyly tak snadno patrné jako z tabulky 1.) Excel toto umožňuje vždy, když jsou jednotlivé kategorie reprezentovány číselnými hodnotami (jako např. zde hodnotami 10 až 13). Pokud bychom potřebovali zpracovat veličiny kategoriální-textové, stačilo by mít jednotlivé slovní kategorie zaznamenány pod číselným kódem (čili např. u veličiny „pohlaví“ bychom každého muže kódovali třeba hodnotou 1 a ženu hodnotou 2). Jediné, co si ještě musíme připravit předem, je přehled jednotlivých kategoriálních hodnot, které si přejeme rozlišovat. Tyto hodnoty si vypíšeme kdekoli na stejném listu, v němž už máme připravena data ke zpracování, ale někde mimo oblast těchto dat (tedy třeba až do buněk K2 až K5 jako v tabulce 2).

Tabulka 2: Data o věku – příprava k použití funkce ČETNOSTI



Předpokládejme, že výsledné četnosti chceme mít v sousedních buňkách, tedy v buňkách L2 až L5. Tyto buňky si označíme (jak je naznačeno v tabulce 5) a zadáme příkaz s touto obecnou strukturou:

=ČETNOSTI(data;kategorie) (3)

Argument „data“ označuje oblast buněk, v nichž jsou zadána zpracovávaná data (v našem příkladě jde o veličinu věk, čili o data z oblasti B3 až B15). Poznamenejme, že není nutno mít tato data předem seřazená vzestupně jako ve sloupci B v tabulce 1, Excel by zpracoval i data nesetříděná. Argument „kategorie“ udává, jaké kategoriální hodnoty chceme rozlišovat; tuto informaci máme již připravenou v buňkách K2 až K5. Příkaz tedy bude mít v našem příkladě konkrétní podobu

=ČETNOSTI(B3:B15;K2:K5) (4)

Pozor – abychom skutečně získali četnosti pro všechny požadované kategorie, musí Excel poznat, že zadaný příkaz platí ne pro jednu buňku (jako obvykle), ale skutečně pro všechny předem vybrané buňky (L2 až L5). Říkáme tomu, že zadáváme „maticový vzorec“ a technicky to znamená, že místo stisknutí klávesy ENTER musíme příkaz odeslat tak, že držíme stisknuté současně klávesy CTRL a SHIFT a při tom stiskneme klávesu ENTER. Pokud opomineme použít tento „troj-hmat“, příkaz se neprovede správně. Po správném provedení se v buňkách L2 až L5 objeví požadované hodnoty příslušných absolutních četností (tedy popořadě čísla 3, 4, 3 a 3).

Obvykle ještě pod získané četnosti uvádíme jejich součet. Ten získáme v Excelu automaticky tak, že v příslušné buňce (zde v buňce L6) zadáme příkaz =SUMA(), přičemž v závorce uvedeme odkaz na buňky, jejichž hodnoty chceme sčítat. V našem příkladě bychom tedy zadali v buňce L6 příkaz

=SUMA(L2:L5) (5)

Tento příkaz lze vložit jednoduše výběrem ikony „Automatické shrnutí“ označené sumačním symbolem Σ. Dokonce i požadovaná oblast buněk se zadává automaticky – Excel do příkazu sám vkládá nejbližší buňky s číselným obsahem.

□

Příklad 4 (relativní četnosti a využití „zamykacího dolaru“):
Pokračujme v určování četností z příkladu 3. Relativní četnosti získáme vždy tak, že jednotlivé absolutní četnosti vydělíme celkovým počtem pozorování. V našem příkladě budou mít tedy relativní četnosti jednotlivých věkových kategorií hodnoty popořadě 3/13=0,23; 4/13=0,31; 3/13=0,23; 3/13=0,23. Znamená to, že např. do druhé věkové kategorie (11 let) patřilo 31 % ze všech 13 sledovaných dětí. Předpokládejme, že chceme tyto četnosti určit v Excelu (do dalšího sloupce, tedy sloupce M). Výpočet si připravíme v buňce M2 pro první kategorii (tedy pro kategorii desetiletých):

=L2/L$6 (6)

Tímto příkazem spočítáme hodnotu podílu 3 (hodnota z buňky L2) ku 13 (hodnota z buňky L6). Symbol „$“ zapsaný před číslicí 6 má tzv. „zamykací“ funkci. Jak by mělo být čtenáři známo, pokud v Excelu kopírujeme nějaký vzorec do jiné buňky, automaticky se v něm posunou odkazy na buňky, které byly v tom vzorci použity. Neplatí to však pro souřadnice „zamčené“ právě symbolem „$“ – takové odkazy se při kopírování nemění. Zkopírujme nyní vzorec (6) o jednu buňku níž, tedy do buňky M3 (např. pomocí klávesových zkratek CRTL C, CTRL V). Výsledkem v buňce M3 bude příkaz

=L3/L$6 (7)

tedy hodnota podílu 4 (hodnota z buňky L3) ku 13 (opět hodnota z buňky L6). Analogicky zkopírujeme tento výpočet i do buněk M4 a M5. Pokud na závěr pod takto zjištěné relativní četnosti (do buňky M6) vložíme hodnotu jejich součtu, tj. sumaci =SUMA(M2:M5), získáme hodnotu 1.

□

Kumulativní četnosti

V případě, že zpracováváme veličinu s uspořádanými kategoriemi (tedy veličinu textovou-ordinálního typu, nebo veličinu číselnou-diskrétního typu), má smysl zavádět i četnosti kumulativní (kumulované), jak absolutní, tak relativní. Vznikají jako součet četností (absolutních, resp. relativních)

za danou a všechny předešlé kategorie. (Proto je nutno dbát na to, aby se jednalo skutečně o veličiny s uspořádáním: kvůli následné interpretaci musí být jasné, které kategorie byly ty „předešlé“).

Příklad 5 (kumulativní četnosti):
Dokončeme určování četností z příkladů 3 a 4 pomocí Excelu. Předpokládejme, že chceme ve sloupci N získat hodnoty kumulativních absolutních četností. Konkr. v buňce N2 (u první věkové kategorie) chceme tedy mít přímo absolutní četnost této kategorie (protože žádná „předešlá“ kategorie, míněno kategorie jedinců mladších než 10 let, v těchto datech neexistuje). V buňce N2 tedy zapíšeme pouze odkaz

=L2 (8)

V další buňce N3 (u druhé věkové kategorie) už musíme mít součet absolutních četností této (druhé) kategorie (tedy údaj z buňky L3) a právě zjištěný kumulativní údaj za všechny kategorie předešlé (tedy údaj již připravený v buňce N2). Do buňky N3 tedy zapíšeme příkaz

=L3+N2 (9)

Nyní již stačí tento poslední příkaz zkopírovat do příslušných buněk u zbývajících kategorií (tedy do buněk N4 a N5). V našem příkladě by výsledkem měly být popořadě hodnoty 3, 7, 10, 13. Pokud bychom chtěli kumulovat četnosti relativní, je postup analogický (v našem příkladě bychom se pouze místo na buňky ze sloupce L odkazovali na buňky ze sloupce M).

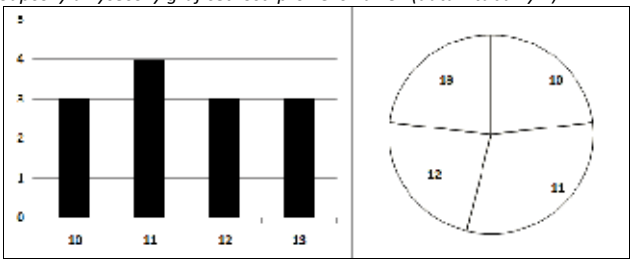
□

Jaká by byla interpretace zjištěných kumulativních absolutních četností z příkladu 5? Např. hodnota 10 (zjištěná pro třetí věkovou kategorii, tedy pro kategorii 12-letých) znamená, že mezi sledovanými dětmi jich 10 patřilo do této nebo předešlých kategorií; jinak řečeno: deset dětí bylo ve věku 12 let nebo mladších; nejelegantněji řečeno: 10 dětí bylo ve věku nejvýše 12 let. Ještě jednou zde připomeňme požadavek „uspořádatelnosti“ - slovo „nejvýše“ bychom u veličiny textové-neuspořádané nemohli při rozumné interpretaci použít. Pro úplnost dodejme, že u kumulativních četností (ani absolutních, ani relativních) nemá smysl určovat jejich součet.

Grafické znázornění četností a modus

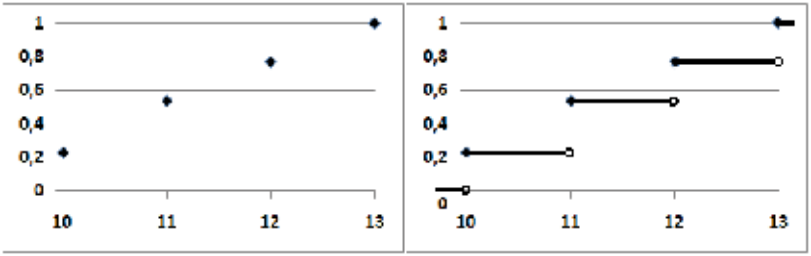
Pro znázornění četností (nejprve těch ne-kumulovaných) jsou nejvhodnější tyto dva typy grafů: sloupcový (histogram) a výsečový (též kruhový či koláčový, což je překlad anglického termínu pie-chart). V případě sloupcového grafu vyneseme na vodorovnou osu hodnoty jednotlivých kategorií (proto je tento typ vhodný především pro číselné veličiny, případně pro veličiny textové-ordinální, jejichž jednotlivé kategorie můžeme reprezentovat jejich pořadovým číslem: tedy 1 pro první kategorii, 2 pro druhou kategorii, atd.). Na svislou osu vynášíme hodnotu příslušné četnosti, ať už jde o graf četností absolutních nebo relativních. Graf výsečový odpovídá vlastně znázornění četností relativních – plocha celého kruhu představuje 100 % (tedy všechna pozorování), jednotlivým kategoriím odpovídají výseče, jejichž velikost (daná velikostí úhlu) reflektuje poměrné zastoupení výskytu dané kategorie. K označení toho, o jakou kategorii se jedná, lze použít popisky vložené přímo na jednotlivé výseče. Oba typy grafů jsou pro veličinu věk z dat v tabulce 1 uvedeny na obrázku 1. Vytvoření těchto grafů pomocí Excelu je i pro jen málo zkušeného uživatele velmi intuitivní: Stačí vybrat buňky s hodnotami, které si přejeme znázornit (v našem případě s četnostmi, tj. vybereme oblast buněk L2 až L5) a pak přes nabídku Vložení – Grafy pouze zvolíme požadovaný typ grafu. Jedinou nevýhodou je, že Excel automaticky označí jednotlivé kategorie jejich pořadovými čísly (tedy 1 až 4 a nikoli 10 až 13, jak by zde bylo vhodnější). Je proto ještě následně nutno kliknout na vytvořený graf pravým tlačítkem a z nabídky, která se objeví, aktivovat volbu „Vybrat data…“. Zde nalezneme „Popisky vodorovné osy (kategorie)“ a po kliknutí na příslušné tlačítko „Upravit“ můžeme zadat oblast buněk, v nichž se nacházejí požadované popisky – v našem případě oblast K2 až K5. Další možné úpravy grafů (barevnost, písmo atd.) si případně nezkušený čtenář může již osvojit sám postupným zkoušením jednotlivých nabídek.

Obrázek 1: Sloupcový a výsečový graf četností pro veličinu věk (data z tabulky 1)



S pojmem četnosti souvisí i charakteristika, kterou nazýváme modus. Jedná se o tu kategoriální hodnotu, která se v datech vyskytla nejčastěji. Je přitom jedno, zda modus určujeme dle četností absolutních či relativních – jako výsledek totiž neuvádíme hodnotu maximální četnosti, ale skutečně jen to, v jaké kategorii bylo této maximální četnosti dosaženo (což musí vyjít shodně podle obou typů četností). Pro veličinu věk (data z tabulky 1) je nejjednodušší určit modus s využitím již hotového grafu četností (viz obrázek 1) – nejčetnější výskyt evidentně zaznamenala kategorie 11-letých, takže modus má pro tuto veličinu (a tato data) hodnotu 11. Poznamenejme, že slovo modus skloňujeme: modus – modu -…- modem. Dále poznamenejme, že v případě, kdy je maximální četnosti dosaženo souhlasně ve dvou (nebo dokonce více) kategoriích, hovoříme o veličině (a datech) bi-modálních (případně více-modálních, ale v takových případech už spíš modus vůbec neurčujeme).

Obrázek 2: Kumulativní četnosti veličiny věk – bodový graf (vlevo) a graf distribuční funkce (vpravo)



Co se týče grafického znázornění kumulativních četností, znázorňujeme obvykle ty relativní. Z nabídky grafů v Excelu je k tomuto účelu asi nevhodnější volbou graf bodový (viz obrázek 2 vlevo), přičemž jako hodnoty pro vodorovnou osu musíme vybrat buňky obsahující kategoriální hodnoty a jako hodnoty pro svislou osu vybrat buňky s vypočítanými kumulativními relativními četnostmi. Ještě vhodnějším grafickým znázorněním by byl tzv. graf distribuční funkce (viz obrázek 2 vpravo), jenž sice Excel nemá ve své standardní nabídce grafů, ale lze jej získat z grafu bodového jeho drobnou úpravou v nějakém grafickém editoru: stačí z každého Excelem vyneseného bodu vést úsečku až k úrovni další kategorie, kde tuto úsečku zakončíme otevřeným kolečkem, naznačujícím, že v tuto chvíli přestává dosavadní hodnota platit a distribuční funkce „skáče“ na vyšší úroveň. Podle grafů na obrázku 2 (je jedno, zda použijeme graf bodový či graf distribuční funkce) např. vidíme, že pro x=11 má kumulativní četnost hodnotu cca 54 % (součástí obou grafů je bod o souřadnicích [11;0,54]). Znamená to, že 54 % dotázaných bylo ve věku nejvýše 11 let. Distribuční funkce (graf vpravo) ale umožňuje interpretovat i věkové hodnoty, které se v datech vůbec nevyskytly – např. x=11,5; jelikož jde o bod ležící na stejné úsečce jako bod [11 ; 0,54], musí být příslušná y-ová souřadnice stejná, tedy 0,54; znamená to doslovně, že 54 % dotázaných bylo věku nejvýše 11,5 roku (kterážto informace je zcela v souladu se

zjištěnými daty). Analogicky např. pro x=9,7 podle grafu distribuční funkce snadno zjistíme, že odpovídající y-ová hodnota je 0, protože graf distribuční funkce vždy od nejmenší kategorie doleva (tedy zde od 10 let) splývá s vodorovnou osou; toto znamená, že nikdo z dotázaných (0 %) nebyl ve věku nejvýše 9,7 roku (tj. ve věku 9,7 roku nebo mladší), což je opět zcela v souladu s danými daty. A konečně např. pro x=13,2 podle grafu distribuční funkce snadno zjistíme, že odpovídající y-ová hodnota je 1, protože graf distribuční funkce vždy od největší kategorie doprava (tedy zde od 13 let) zůstává na úrovni 1; toto znamená, že všichni dotázaní (100 %) byli ve věku nejvýše 13,2 roku, což je opět zcela v souladu s danými daty.

Problém „multiple responses“

Pokud sečteme absolutní četnosti přes všechny sledované kategorie, musíme jako hodnotu součtu získat celkový počet statistických jednotek, jako tomu bylo např. v závěru příkladu 3 po zadání příkazu (5). Toto pravidlo však nemusí platit v případě, kdy povolíme tzv. „multiple responses“, tedy to, že respondent jako odpověď na danou otázku smí vybrat více než jednu uvedenou kategorii.

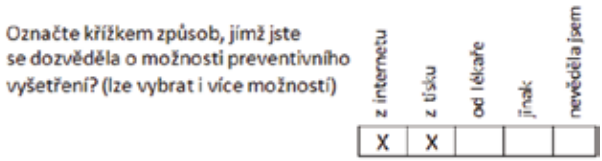
Příklad 6 (zpracování dat s „multiple responses“):
Dotázaných 20 respondentů v rámci rychlé ankety uvádělo, z jakých zdrojů se dozvěděli (pokud vůbec) o možnosti preventivního vyšetření na daný typ karcinomu. Možnosti byly: z internetu (kategorie 1), z tisku (2), od lékaře (3), jinak (4), nedozvěděl/a jsem se o tom (5). Zjištěné absolutní četnosti výskytu jednotlivých odpovědí shrnuje tabulka 3.

Tabulka 3: Ukázka absolutních četností u 20 statistických jednotek s „multiple responses“

Kategorie č.	1	2	3	4	5	celkem
Četnost výskytu	6	4	8	5	6	29

Součet absolutních četností v tabulce 3 není 20 (což byl celkový počet dotázaných), ale 29, a to právě proto, že někteří dotázaní vybrali jako odpověď na tuto otázku více než jednu z nabízených možností, jako např. jedinec z obrázku 3.

Obrázek 3: Ukázka vyplnění dotazníku s povolením „multiple responses“



V případě, že se v datech u některé otázky povolily multiple responses, není možno přepsat příslušnou veličinu do Excelu jako jeden sloupec (jako např. u veličin v tabulce 1), ale musíme zapsat každou kategorii zvlášť do samostatného sloupce (viz tabulka 4).

Každá kategorie se tak vlastně stane samostatnou alternativní veličinou, přičemž hodnota 1 u ní znamená, že dotyčný respondent tuto kategorii vyznačil, hodnota 0 (resp. prázdná buňka) znamená, že ji neoznačil. Např. pro respondenta č.1 (jehož dotazník byl uveden na obrázku 3) tak i z tabulky 4 vidíme, že jako odpověď vybral možnosti „z internetu“ a „z tisku“, tedy celkem volil dva různé způsoby – tento počet vidíme i z příslušné součtové hodnoty v buňce G2. Zdůrazněme, že součtové buňky ve sloupci G byly připraveny tak, aby sčítaly opravdu jen počty „způsobů, jimiž se respondent o prevenci dozvěděl“, tedy aby sčítaly kategorie 1-4 (neboli hodnoty ze sloupců B až E); pokud někdo uvedl, že o prevenci nevěděl (konkr. šlo o dotazníky č. 15 až 20), má příslušný součet ve sloupci G hodnotu 0 (dotyčný se o prevenci nedozvěděl ani jedním z uvedených způsobů). Hodnoty uvedené

v tabulce 3 získáme snadno jako součty ve sloupcích - např. hodnota 6 v buňce B25 je výsledkem příkazu =SUMA(B2:B21). Tato hodnota znamená, že 6 ze 20 (nikoli ze 29) dotázaných respondentů uvedlo jako zdroj své informovanosti internet (přičemž ovšem někteří z nich mohli k tomu navíc uvést ještě i jiný zdroj). Pokud bychom chtěli získat tabulku četností, v níž bude platit pravidlo, že součet četností je roven počtu statistických jednotek, můžeme zpracovat součtovou veličinu ze sloupce G, tedy zde veličinu s kategoriemi 0, 1, 2 a 3. Odpovídající hodnoty četností ve sloupci J byly získány aplikací příkazu =ČETNOSTI podobně jako v příkladu 3. Např. dvojice údajů v buňkách I5 a J5 pak znamená, že 2 respondenti (ze 20 dotázaných) uvedli 3 různé zdroje své informovanosti.

Tabulka 4: Zpracování dat s „multiple responses“

	A	B	C	D	E	F	G	H	I	J
1	odpověď / respondent č.	1-internet	2-tisk	3-lékař	4-jinak	5-NIJAK	počet způsobů (suma B až E)		počet způsobů	četnost
2	1	1	1				2		0	6
3	2			1	1		2		1	7
4	3	1					1		2	5
5	4		1	1	1		3		3	2
6	5			1			1		suma	20
7	6				1		1			
8	7	1		1			2			
9	8		1				1			
10	9			1			1			
11	10				1		1			
12	11			1			1			
13	12	1	1				2			
14	13	1		1			2			
15	14	1		1	1		3			
16	15					1	0			
17	16					1	0			
18	17					1	0			
19	18					1	0			
20	19					1	0			
21	20					1	0			
22										
23										
24	odpověď	1	2	3	4	5	suma			
25	počet	6	4	8	5	6	29			

Věrohodnostní poměr a podíl šancí

Častým úkolem při zpracování nejen biomedicínských dat je porovnání dvou či více pod-skupin. Např. v článku autorů Bystroň a kol. byly porovnány dvě skupiny pacientů, kteří prodělali infarkt myokardu s tím, že jim byl implementován buď stent typu EPC (skupina 1), nebo stent typu CoCr (skupina 2). Všichni pacienti byli sledováni po dobu 6 měsíců a bylo zaznamenáno, zda u nich během této doby došlo k nějaké závažné kardiovaskulární příhodě (MACE – major adverse cardiovascular event). Ve skupině 1 bylo celkem 50 pacientů, z nich 12 prodělalo MACE. Ve skupině 2 bylo také celkem 50 pacientů, z nich 5 prodělalo MACE. Označme relativní četnosti MACE v jednotlivých skupinách popořadě jako p_1, p_2, jejich hodnoty určíme podobně jako v příkladu 4 výpočtem:

$$p_1=12/50=0{,}24\ (24\ \%) \qquad p_2=5/50=0{,}10\ (10\ \%) \tag{10}$$

Obě skupiny lze nyní porovnat pomocí věrohodnostního poměru (likelihood ratio), který má obecně tvar

$$p_1/p_2 \tag{11}$$

takže v tomto případě má věrohodnostní poměr konkrétní hodnotu 0,24/0,10=2,4. Tento výsledek interpretujeme jednoduše tak, že sledovaný jev (zde MACE) měl 2,4-krát vyšší četnost výskytu v první skupině ve srovnání se skupinou druhou.

Jinou možností, běžnější v anglo-saské odborné literatuře, je určit místo relativní četnosti tzv. šanci. Zatímco relativní četnost je poměr absolutní četnosti výskytu sledovaného jevu k celkovému počtu uvažovaných statistických jednotek, šance (odds) je poměr absolutní četnosti výskytu sledovaného jevu k počtu zbývajících statistických jednotek. (Může se tedy někdy stát, že na rozdíl od relativní četnosti vyjde hodnota šance i vyšší než 1; hodnoty šance proto nepřevádíme na %.) Označíme-li šance na výskyt MACE v obou skupinách popořadě jako $š_1$, $š_2$, bude

$$š_1 = 12:(50–12) = 12:38 = 6:19 \quad (12)$$

$$š_2 = 5:(50–5) = 5:45 = 1:9 \quad (13)$$

Upřesněme, že ve vztahu (12) bylo poslední úpravou vykrácení dvěma, ve vztahu (13) bylo poslední úpravou vykrácení pěti. Můžeme sice výrazy (12), (13) ještě dopočítat jako hodnoty (symbol : značí běžné dělení, takže $š_1$=6/19=0,32, resp. $š_2$=1/9=0,11), ale při interpretaci je obvykle ponecháme přímo v uvedených tvarech a výsledek čteme takto: šance na výskyt MACE byla „6 ku 19“ ve skupině 1, resp. „1 ku 9“ ve skupině 2. Obě skupiny lze nyní porovnat pomocí podílu šancí (odds ratio), který má obecně tvar

$$š_1/š_2 \quad (14)$$

takže v tomto případě má podíl šancí konkrétní hodnotu 0,32/0,11=2,84. Tento výsledek interpretujeme jednoduše tak, že šance na výskyt sledovaného jevu (zde MACE) byla 2,84-krát vyšší v první skupině ve srovnání se skupinou druhou.

Všechny tyto výpočty lze provést jednoduše v Excelu. Zapišme si do buňky A1 hodnotu 12 (počet výskytů MACE v 1. skupině) a do sousední buňky B1 hodnotu 50 (celkový rozsah 1. skupiny). Podobně na druhý řádek si připravíme údaje za druhou skupinu pacientů (tedy do buňky A2 hodnotu 5, do buňky B2 opět hodnotu 50). Hodnoty (10) dostaneme popořadě v buňkách C1, resp. C2 příkazy =A1/B1, resp. (stačí jen překopírovat) =A2/B2. Hodnotu (11) získáme v buňce pod nimi (C3) příkazem =C1/C2. Výpočet (12), resp. (13) jakožto číselných hodnot (tedy 0,32, resp. 0,11) zadáme v dalším sloupci, konkr. pro 1. skupinu v buňce D1 příkazem =A1/(B1–A1), resp. pro 2. skupinu v buňce D2 (stačí jen překopírovat) =A2/(B2–A2). Hodnotu (14) pak získáme v buňce pod nimi (D3) příkazem =D1/D2 (stačí tam překopírovat příkaz z buňky C3).

Deskriptivní charakteristiky číselných veličin

Medián

Medián patří mezi tzv. kvantilové charakteristiky, nazýváme jej také 50% kvantil. Určujeme jej u číselných veličin (diskrétního i spojitého typu) a lze jej určit tak, že všechny napozorované hodnoty seřadíme vzestupně podle velikosti (pokud se některá hodnota opakuje, uvedeme ji tolikrát, kolikrát se vyskytla) a za medián označíme hodnotu na prostřední pozici; byl-li počet všech pozorování lichý, půjde přesně o prostřední hodnotu (tedy např. ze sedmi hodnot by se jednalo o čtvrtou); byl-li počet všech pozorování sudý, označuje se za medián obvykle průměr ze dvou prostředních hodnot (tedy např. z osmi hodnot by mediánem byl průměr ze čtvrté a páté hodnoty). Jiným způsobem určení mediánu (který ale může případně poskytnout co do hodnoty poněkud jiný výsledek než postup právě uvedený) je použít kumulativní relativní četnosti (pokud jsou k dispozici, obvykle to tedy lze u veličin diskrétních): za medián označíme tu číselnou kategorii, u níž poprvé je hodnota kumulativní relativní četnosti na úrovni 50 % (nebo kde je poprvé nad touto úrovní).

Příklad 7 (určení mediánu):
Určíme medián veličiny věk pro data v tabulce 1. Jde o 13 již seřazených (!) hodnot ze sloupce B. Prostřední z nich, tedy sedmou hodnotou, je hodnota 11. Kumulativní relativní četnosti byly znázorněny na obrázku 2. Z obou typů grafů by zde mělo být zřejmé, že na(d) úroveň 50 % se data dostala poprvé v kategorii 11 let. Oba postupy shodně zde tedy jako medián určily hodnotu 11. Nejjednodušším způsobem je ale použít příkaz v Excelu, přičemž data ani nemusíme předem řadit podle velikosti (!). Stačí v kterékoli buňce (samozřejmě ne v té, která je již nějakou hodnotou obsazena) zadat příkaz v obecném tvaru

=MEDIAN(data) (15)

kde argument „data“ může být odkazem na oblast buněk obsahující analyzované číselné údaje. V našem případě by tedy měl mít příkaz tvar

=MEDIAN(B3:B15) (16)

a výsledkem bude opět hodnota 11. Vzhledem k nemalé souvislosti mezi mediánem a kumulováním bude interpretace výsledku analogická jako u kumulativních relativních četností, a to: polovina (50 %) respondentů (statistických jednotek) byla ve věku nejvýše 11 let.

□

Aritmetický průměr

Aritmetický průměr je nejběžnějším typem průměrné hodnoty, takže pokud hovoříme jen o „průměru“, máme obvykle na mysli právě tento typ. Patří mezi tzv. momentové charakteristiky, nazýváme jej také první obecný moment. Určujeme jej u číselných veličin (diskrétního i spojitého typu) a lze jej určit tak, že všechny jednotlivé napozorované hodnoty sečteme a tento výsledek (tzv. úhrn) vydělíme rozsahem výběru. V Excelu je možno využít příkaz v obecném tvaru

=PRŮMĚR(data) (17)

kde argument „data“ může být odkazem na oblast buněk obsahující analyzované číselné údaje.

Příklad 8 (určení aritmetického průměru):
Určíme průměr veličiny věk pro data v tabulce 1, tedy pro hodnoty ze sloupce B. Při ručním výpočtu zjistíme nejprve hodnotu úhrnu (součet všech 13 napozorovaných hodnot činí 149) a tento výsledek vydělíme rozsahem výběru, tedy 13: Výsledkem je 149/13=11,46. Rychlejší je využít příkaz (17), který zde musí mít konkrétní podobu

=PRŮMĚR(B3:B15) (18)

Výsledkem je přímo (bez nutnosti určovat nejprve úhrn) hledaná hodnota 11,46 let. Pokud tento vzorec překopírujeme do dvou sousedních buněk směrem doprava, Excel v nich automaticky vypočte průměry i sousedních veličin, tedy průměrnou výšku (=156,23 cm) a průměrnou hmotnost (=55,62). Interpretace je zřejmá – pro sledované jedince činil průměr jejich věku (výšky, hmotnosti) tolik a tolik roků (cm, kg).

□

Rozptyl a směrodatná odchylka

I tyto údaje se týkají číselných veličin (diskrétního i spojitého typu) a patří mezi tzv. momentové charakteristiky. Konkr. rozptyl je tzv. druhým centrovaným momentem, směrodatná odchylka je jeho druhou odmocninou. Obě charakteristiky slouží jako míry variability neboli proměnlivosti dat: obecně platí, že čím různorodější jsou jednotlivé napozorované hodnoty, tím větší hodnotu má rozptyl a tudíž i směrodatná odchylka, a naopak, čím podobnější si jsou jednotlivá pozorování, tím menší hodnotu má rozptyl a směrodatná odchylka (extrémním případem je situace, kdy by byly všechny napozorované hodnoty zcela identické – pro taková data mají rozptyl i směrodatná odchylka nejmenší možnou, tj. nulovou hodnotu; v takových datech by byla nulová variabilita, nebyla by v nich žádná proměnlivost).

Nebudeme zde vysvětlovat vzorce a „ruční“ výpočty (ty lze nalézt v jiných učebnicích statistiky), pouze okomentujeme fakt, že při výpočtu rozptylu pracujeme s druhými mocninami, takže výsledkem je hodnota, jejíž příslušné měrné jednotky jsou vlastně „na druhou“: např. pro veličinu hmotnost v kg bude jejím rozptylem hodnota v kg^2. Právě proto se při interpretaci běžně používá směrodatná odchylka jakožto odmocnina rozptylu, neboť příslušné měrné jednotky se odmocněním „vrátí do původní podoby“ (např. z uvedených kg^2 zpět na kg). V Excelu je možno využít k výpočtu této užitečné směrodatné odchylky příkaz v obecném tvaru

=SMODCH(data) (19)

kde argument „data“ může být odkazem na oblast buněk obsahující analyzované číselné údaje. Pokud je potřeba uvést hodnotu i rozptylu, je možno jednoduše umocnit již zjištěnou hodnotu směrodatné odchylky, nebo lze hodnotu rozptylu v Excelu určit přímo z dat příkazem v obecném tvaru

=VAR(data) (20)

Příklad 9 (určení směrodatné odchylky):
Určíme směrodatnou odchylku veličiny věk pro data v tabulce 1, tedy pro hodnoty ze sloupce B. Příkaz (19) zde musí mít konkrétní podobu

=SMODCH(B3:B15) (21)

Výsledkem je hodnota 1,08. Nejvhodnější interpretace je spolu s již zjištěným průměrem ve smyslu, že „typické“ zjištěné hodnoty veličiny věk se pohybovaly v rozmezí 11,46±1,08 roku, tedy v rozmezí 10,38 až 12,54 roku. Analogicky bychom určili (argumentem „data“ by byla oblast buněk C3:C15, resp. D3:D15) a interpretovali směrodatné odchylky zbylých dvou veličin (výšky a hmotnosti).

□

Rozmezí dané výpočtem průměr ± směrodatná odchylka nemusí být vždy skutečně rozmezím „typických“ hodnot (proto byly v ukázkové interpretaci v příkladu 9 použity uvozovky). Předpokládejme, že sledujeme pacienty s jistým typem kožního onemocnění, které zasahuje nehty na rukou. Tato choroba zasáhne nejprve jeden z nehtů a hned tehdy část pacientů vyhledá lékaře. Velká část pacientů však lékaře vyhledá až ve chvíli, kdy je nemoc rozšířená již na všechny nehty. Dále předpokládejme, že u sledovaných pacientů byla zaznamenána veličina „počet zasažených prstů na rukou“, tedy veličina s hodnotami 1 až 10. Zjištěné četnosti pacientů zachycuje tabulka 5.

Tabulka 5: Absolutní četnosti u pacientů – veličina počet zasažených prstů na rukou (fiktivní data)

Počet prstů s nemocí	1	2	3	4	5	6	7	8	9	10	celkem
Četnost pacientů	35	14	5	2	1	1	3	7	21	41	130

Pro veličinu z tabulky 5 vychází průměrná hodnota 5,95 („průměrný pacient“ má tedy zasaženo 6 prstů, přičemž ve skutečnosti této průměrné hodnoty nabyl jediný ze všech 130 pacientů) a hodnota směrodatné odchylky činí 3,95. Poznamenejme, že pokud by čtenář chtěl určit průměr a směrodatnou odchylku z tabulky 5, musel by buď nastudovat tzv. vážené metody výpočtu, nebo zrekonstruovat původní data hodnotu po hodnotě tak, aby měl k dispozici sloupec 130 čísel - jedniček až desítek v počtech odpovídajících četnostem z tabulky 5; teprve pak by bylo možno aplikovat v Excelu vzorce (17), resp. (19). Rozmezí průměr ± směrodatná odchylka zde tedy vychází 2,00 až 9,90, přičemž „typičtí“ pacienti by byli ale v tomto příkladu evidentně naopak ti, kteří se nacházejí mimo toto rozmezí (tedy pacienti buď s jedním, nebo se všemi deseti zasaženými prsty). Poznamenejme, že uvedený případ rozdělení četností se někdy nazývá U-rozdělení: důvod bude zřejmý, jakmile si čtenář četnosti z tabulky 5 znázorní ve sloupcovém grafu.

Někdy (především při intervalových odhadech nebo některých statistických testech, obojí viz později) se jako míry variability používají poněkud pozměněné charakteristiky, které nazýváme výběrový rozptyl (oproti „obyčejnému“ rozptylu se při jeho výpočtu ve jmenovateli místo rozsahu výběru

objevuje tzv. počet stupňů volnosti) resp. jeho druhá odmocnina, tedy výběrová směrodatná odchylka. K jejich určení v Excelu stačí aplikovat příkazy v obecném tvaru:

=VAR.VÝBĚR(data) (22)

=SMODCH.VÝBĚR(data) (23)

Geometrický průměr

V některých případech je nutno při určování průměrné („typické“) hodnoty použít jiný typ průměru, než je průměr aritmetický. Připomeňme, že výpočet aritmetického průměru je technicky založen na použití úhrnu (součtu průměrovaných hodnot). Pokud určení úhrnu nemá matematické opodstatnění, nemá pak význam ani aritmetický průměr.

Tabulka 6: Vývoj počtu nemocí z povolání v Ústeckém kraji za roky 2000-2004, zdroj: ÚZIS

Rok	2000	2001	2002	2003	2004
Počet nemocí z povolání	104	98	77	80	59
Řetězový index	X	0,94	0,79	1,04	0,74

Příklad 10 (nevhodné použití úhrnu):
Tabulka 6 je ukázkou tzv. časové řady, tedy záznamu vývoje nějakého ukazatele (zde počtu nemoci z povolání, viz druhý řádek tabulky 6) v čase. Data byla získána z databáze ÚZIS (postup viz Příloha). Aby se jednalo o časovou řadu, musí být dodržen požadavek stále stejné věcné a místní definice (tedy musí být měřeno „stále totéž pro stále stejné území“). Zaznamenaný vývoj lze kromě jiného charakterizovat tzv. řetězovými indexy (viz poslední řádek tabulky 6), nazývanými též koeficienty růstu, které jsou vždy podílem hodnoty za dané období ku hodnotě za období předešlé (proto nelze nikdy určit hodnotu řetězového indexu pro první období – chybí zde vždy hodnota za období předešlé). Např. pro rok 2001 tedy činí hodnota tohoto indexu

98/104 = 0,94 (24)

a znamená to, že údaj za rok 2001 (tedy hodnota 98) představuje 94 % z hodnoty období předešlého (tedy z hodnoty 104 za rok 2000). Je zřejmé, že pokud řetězový index vyjde jako zde menší než 1, došlo v daném období k poklesu oproti období předešlému, a naopak pokud řetězový index vyjde větší než 1 (jako např. v tabulce 6 pro rok 2003), došlo v daném období k nárůstu oproti období předešlému. Pokud bychom ze čtyř řetězových indexů chtěli spočítat jejich aritmetický průměr, musel by mít nějaký matematický význam jejich úhrn (tedy součet). Problém je zde ale v tom, že nelze sčítat procentuální hodnoty počítané z různých základů, což právě zde nastává: pro každý index byla základem jiná hodnota – vždy ta z předešlého období. Zkrátka v tomto a podobných případech „není procento jako procento“ a tudíž by bylo zcestné řetězové indexy sčítat.

□

Ukazuje se, že místo sčítání je užitečné řetězové indexy mezi sebou násobit. Předveďme si to na prvních dvou vypočtených indexech (tedy na hodnotách 0,79 a 0,94). To, jaký význam má vlastně jejich součin, zjistíme snadno, když si za oba indexy dosadíme původní zlomky (tedy to, jak byly vlastně oba indexy vypočteny):

$$0{,}79 \cdot 0{,}94 = (77/98) \cdot (98/104) \quad (25)$$

Pokud ve výrazu (25) na pravé straně vykrátíme hodnotu 98, dostaneme vztah (25) ve tvaru

$$0{,}79 \cdot 0{,}94 = 77/104 \quad (26)$$

Pravá strana v rovnici (26) má matematicky jasnou interpretaci: udává, jak velkou změnu představuje hodnota za rok 2002 (hodnota 77) oproti hodnotě za první sledované období (tedy oproti hodnotě 104; již tedy nejde o srovnání s hodnotou za období předchozí).

Jako důsledek úvah uvedených u výpočtů (25), resp. (26) lze konstatovat, že u řetězových indexů budeme při určování jejich průměru používat průměr geometrický, neboť podstatou jeho výpočtu je násobení průměrovaných hodnot. V Excelu je možno geometrický průměr určit pomocí příkazu

=GEOMEAN(data) (27)

Příklad 11 (ukázka určení geometrického průměru v Excelu):
Předpokládejme, že tabulku 6 máme kompletně připravenou v Excelu tak, že hodnoty řetězových indexů jsou vypočteny v buňkách C3 (hodnota 0,94) až F3 (hodnota 0,74). Jejich geometrický průměr (pozor, že neprůměrujeme počty nemocí z povolání, ale odpovídající indexy) určíme výpočtem

=GEOMEAN(C3:F3) (28)

jehož výsledkem je hodnota 0,87 (tedy číslo o 0,13 menší než hodnota 1). Zjištěný výsledek interpretujeme tak, že v Ústeckém kraji docházelo v letech 2001-2004 každoročně oproti předešlému roku k poklesu počtu nemocných v průměru o 13 %.

Metody statistické indukce – parametrické odhady

Parametry pravděpodobnostních modelů

Předešlé kapitoly byly věnovány deskripci dat, tedy popisu jejich chování s využitím vhodných charakteristik (četností, kvantilů, momentů…). Významnou součástí matematické statistiky jsou ale vedle deskripce také pravděpodobnostní modely, tedy modely, kterými se snažíme data prokládat, a které jsou založené na principech teorie pravděpodobnosti. K detailnímu teoretickému nastudování této problematiky musí čtenář sáhnout po jiné odborné statistické literatuře, zde budou pouze ukázány možnosti, jak některé základní postupy provádět pomocí Excelu, včetně interpretace výsledků. Souhrnně lze metody, které na základě dat (tedy „konkrétna“) usuzují na obecně platné zákonitosti (tedy na „abstraktno“), označit jako metody statistické indukce.

Stejně jako ve statistické deskripci, i zde platí, že volba vhodné metody závisí na typu analyzované veličiny. Některé metody jsou vhodné pro kategoriální veličiny, jiné pro veličiny spojitého typu. Pokud hledáme pravděpodobnostní model, pak všechny mají společné to, že jsou jednoznačně určeny hodnotami svých tzv. parametrů. Při interpretaci nalezených modelů často používáme úvahu zhruba tohoto typu: „Na základě analýzy dat lze předpokládat, že nejvhodnějším modelem pro chování celé populace je model…“ S ohledem na tento způsob interpretace se parametry modelů někdy též nazývají „populační parametry“.

Základním parametrem modelů kategoriálních veličin je parametr označovaný řeckým písmenem π (pí), kterým označujeme pravděpodobnost sledovaného jevu. Tento parametr může nabývat hodnoty mezi 0 až 1 (0 % až 100 %).

Mezi modely pro spojité veličiny zaujímá významné postavení model Gaussova normálního rozdělení, jehož průběh charakterizuje známá Gaussova křivka (viz např. křivka na obrázku 4). Parametry, které jednoznačně určují tvar této křivky, jsou parametr označovaný řeckým písmenem μ (mí), kterým označujeme střední hodnotu daného modelu (jde o teoretický první obecný moment), a parametr označovaný σ^2 (sigma na druhou, sigma kvadrát), kterým označujeme (populační) rozptyl daného modelu (jde o teoretický druhý centrovaný moment). Analogicky jako u rozptylu deskriptivního, určovaného pro data, platí, že jeho odmocninu (zde σ) nazýváme směrodatná odchylka, přičemž jde o odchylku teoretickou (též „populační“), charakterizující příslušný model. Od Gaussova modelu je odvozeno mnoho dalších, s některými se zde později setkáme (model t-Studentova či F-rozdělení).

Bodové odhady

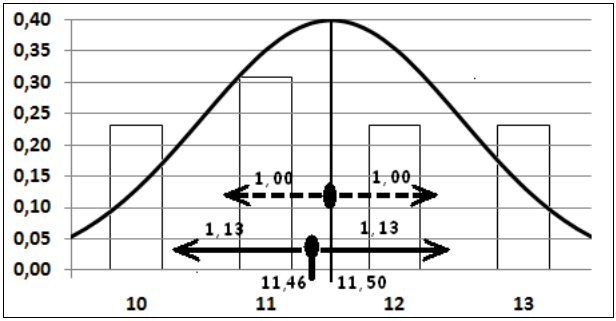
Obecně platí, že populační parametry jsou vlastně jakousi teoretickou obdobou výběrových charakteristik dat, jejichž nejdůležitější typy byly představeny v předešlých kapitolách. Pokud nějaká výběrová charakteristika splňuje jisté vlastnosti, nazýváme ji bodový odhad pro příslušný parametr.

Tabulka 7: Přehled nejdůležitějších populačních parametrů a jejich bodových odhadů

PARAMETR	JEHO BODOVÝ ODHAD
Pravděpodobnost sledovaného jevu (π)	Relativní četnost tohoto jevu v datech
Střední hodnota Gaussova modelu (μ)	Aritmetický průměr
(Populační) směrodatná odchylka (σ)	Výběrová směrodatná odchylka

V praxi nejčastěji využívané bodové odhady (pro parametry představené v předešlé kapitole) jsou uvedeny v tabulce 7. Vztah mezi parametry μ, resp. σ, a jejich odhady ilustruje obrázek 4. Čtyři sloupečky odpovídají relativním četnostem veličiny věk (data viz tabulka 1, znázornění odpovídajících absolutních četností viz obrázek 1). Pro tato data vyšel průměrný věk 11,46 (tato hodnota je na obrázku 4 tučně vyznačena na vodorovné ose). Pokud bychom určili hodnotu výběrové směrodatné odchylky pomocí příkazu (23), měla by hodnotu 1,13 (tato variabilita dat je na obrázku 4 naznačena šipkami směřujícími na obě strany od zjištěné průměrné hodnoty). Jako model (byť třeba ne ten nejvhodnější pro daná data, ale jde pouze o ukázku) byl zvolen model Gaussova normálního rozdělení s parametry μ=11,5 a σ^2=1 (po odmocnění σ=1). Odpovídající Gaussova křivka je na obrázku 4 znázorněna jako hladká tučná křivka, s maximální hodnotou právě v bodě 11,5, kolem něhož je tato křivka souměrná, přičemž její šířka (teoretická variabilita modelu) je dána hodnotou 1 (vyznačeno jako čárkované šipky). Spíše pro zajímavost zde zmiňme alespoň ten důležitý fakt, že celková velikost plochy pod Gaussovou křivkou musí být vždy rovna 1 (představuje 100 %).

Obrázek 4: Data versus model



Příklad 12 (ukázka interpretace bodových odhadů):
V tabulce 1 byla zaznamenána data o 13 dětech. Předpokládejme, že těchto 13 dětí představovalo reprezentativní výběr z nějaké populace dětských pacientů s jistým (zde nespecifikovaným) typem onemocnění. V příkladu 4 byla pro zaznamenaná data zjištěna relativní četnost druhé věkové kategorie (11 let), která činila 31 %. Naším (bodovým) odhadem tedy nyní bude, že kategorie 11-letých tvoří 31 % celé této populace. V příkladu 8 byl zjištěn průměrný věk pro zaznamenaná data, který činil 11,46 let. Naším (bodovým) odhadem tedy nyní bude, že střední věk celé zkoumané populace má hodnotu 11,46 let (takže „ideální" model Gaussovy křivky by měl být vycentrován právě kolem této hodnoty a ne kolem hodnoty 11,5, jako na obrázku 4).

Tento typ odhadů se nazývá „bodový" proto, že jako výsledek poskytnou vždy jedinou hodnotu – jeden bod na číselné ose. Výhodou bodových odhadů je jejich jednoduchost: především relativní četnost nebo aritmetický průměr dokáže z dat určit i laik. Naopak nevýhodou je to, že jejich výsledná hodnota závisí na náhodě: průměrný věk 13 dětí vyšel podle toho, jakých 13 dětí bylo do průzkumu náhodně zařazeno. Pokud bychom bývali vybrali jiné děti, vyšel by jejich průměrný věk (neboli bodový odhad hledaného parametru střední věk) jinak. Tuto nevýhodu lze odstranit tak, že místo

jedinou hodnotou bude hledaný parametr odhadnut celou množinou hodnot – nejčastěji nějakým intervalem. K takovýmto úvahám je však nutno pochopit teoretické vlastnosti bodových odhadů. Zde se zmíníme alespoň o jedné z obecných zákonitostí, kterou je tzv. zákon velkých čísel. Podstatou tohoto zákona je to, že čím více pozorování bylo učiněno, tím blíže by měla být hodnota bodového odhadu k odhadované hodnotě příslušného populačního parametru. Ilustraci tohoto zákona si může čtenář sám zrealizovat analogicky jako v následujícím příkladě.

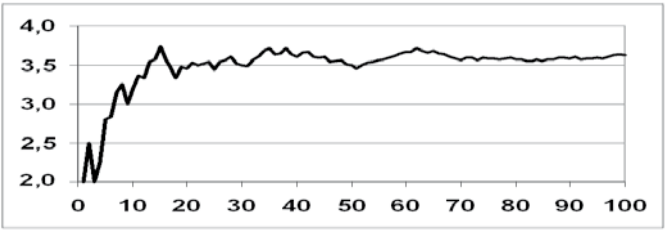
Tabulka 8: Ilustrace zákona velkých čísel pomocí 100 hodů hrací kostkou (první a poslední hody)

Pořadí hodu	1	2	3	4	5	…	98	99	100
Hozeno	2	3	1	3	5	…	5	5	3
Průměr	2,00	2,50	2,00	2,25	2,80	…	3,62	3,64	3,63

Příklad 13 (ilustrace zákona velkých čísel):
Provedeme 100 hodů běžnou hrací kostkou a v každém hodu si zaznamenáme hozenou hodnotu (tedy číslo 1-6). Zatímco při standardním statistickém zpracování bychom provedli deskripci dat až po posledním hodu (takže až po provedení 100. hodu bychom např. spočítali průměr ze všech hozených čísel), v rámci tohoto příkladu spočítáme po každém hodu průměr ze všech dosud hozených čísel. Po prvním hodu bude tímto průměrem přímo ona sama hozená hodnota; po druhém hodu určíme průměr z prvního a druhého hozeného čísla; hned po třetím hodu určíme průměr ze všech tří dosud hozených čísel; atd. Ukázka toho, jak vypadalo 100 konkrétních realizovaných hodů, včetně postupně zjištěných průměrů, je v tabulce 8, resp. na obrázku 5. Samozřejmě, že pokud si bude čtenář tento pokus sám opravdu realizovat, bude jeho posloupnost sta naházených čísel jiná, než byla zde, a tudíž i jím spočítané průměry budou odlišné, nicméně zákon velkých čísel by se měl projevit stejně – tj. hodnota průměru by se s rostoucím počtem hodů měla blížit zhruba k hodnotě 3,5, přičemž kolísání kolem této hodnoty by se mělo postupně zmenšovat (viz obrázek 5).

V příkladu 13 byla zmíněna hodnota 3,5, která zde byla známou teoretickou střední hodnotou (jde o průměr všech šesti hodnot na kostce, tedy průměr ze šesti čísel 1 až 6). U konkrétní hrací kostky se ale její střední hodnota bude od čísla 3,5 více či méně lišit (záleží na tom, nakolik se ona konkrétní kostka liší svým tvarem, polohou těžiště atd. od ideální krychle). V praxi navíc pracujeme skoro vždy s modely, jejichž teoretické parametry známe jen přibližně, nebo vůbec - právě proto je potřebujeme umět odhadnout, ať již bodově, nebo složitěji (viz následující kapitola).

Obrázek 5: Zákon velkých čísel – závislost průměru (svislá osa) na rozsahu výběru (vodorovná osa)



Intervalové odhady aneb intervaly spolehlivosti

V rámci předchozí kapitoly byly představeny bodové odhady včetně jejich nevýhody, kterou je právě ona „bodovost“, tedy to, že se jedná o jedinou hodnotu, aniž bychom věděli, nakolik se na ni můžeme jakožto na odhad neznámého parametru spolehnout. Zároveň bylo řečeno, že místo bodových odhadů budeme sestrojovat celou množinu „kandidátů“ na to být tím správným odhadem pro neznámou hodnotu parametru. Nechť je onou množinou interval (A; B), který nazveme intervalový

odhad pro daný parametr. Jediné, co budeme od nalezeného intervalu požadovat je, aby s dostatečně velkou pravděpodobností pokrýval nám neznámou hodnotu parametru, jinak řečeno, aby neznámá hodnota parametru ležela s dostatečně velkou pravděpodobností právě v tomto intervalu. Tuto pravděpodobnost („že jsme se trefili“) nazýváme spolehlivost intervalového odhadu a obvykle požadujeme, aby činila 95 %. Intervalový odhad (A; B), který splňuje tento požadavek, proto také často nazýváme 95% interval spolehlivosti. Vzorce na určení intervalů spolehlivosti existují pro každý parametr, zde si předvedeme pouze způsob, jak sestrojit pomocí Excelu intervalový odhad pro neznámou střední hodnotu (μ) Gaussova modelu.

Příklad 14 (interval spolehlivosti pro parametr μ):
K sestrojení intervalového odhadu pro neznámý střední věk na základě dat v tabulce 1 lze v Excelu využít jeden z připravených analytických nástrojů. Nejsou-li analytické nástroje aktivní, jejich aktivace se provádí následovně:

a) klepneme v Excelu na velké tlačítko Office (zcela vlevo nahoře);
b) v panelu nabídek, jenž se otevře, vybereme „Možnosti aplikace Excel“ (zcela dole, vpravo);
c) zde (v nabídce vlevo) otevřeme „Doplňky“;
d) v pravé části menu zcela dole vedle „Spravovat:“ klikneme na „Přejít…“ u položky rozbalovacího seznamu „Doplňky aplikace Excel“;
e) v menu, které se otevře, zaškrtneme políčko u možnosti „Analytické nástroje“;
f) aktivaci analytických nástrojů potvrdíme kliknutím na OK.

Byly-li analytické nástroje správně aktivovány, objeví se v Excelu po otevření karty „Data“ mezi nabídkami také „Analýza dat“. Máme-li již připravena data jako v tabulce 1, klikneme na nabídku „Analýza dat“ a postupně vybíráme, resp. zadáváme:

a) vybereme nástroj „Popisná statistika“ (OK);
b) jako vstupní oblast zadáme oblast buněk B2:B15 (tedy včetně popisky „Věk ($X_{1,i}$)“ v buňce B2); lze zde též zadat oblast buněk B3:B15 (tedy pouze hodnoty věku bez nadpisu);
c) v kolonce „Sdružit“ musí být vybrána možnost „Sloupce“ (jsou-li data ve sloupcích jako zde);
d) kolonku „Popisky v prvním řádku“ zaškrtneme pouze, pokud jsme vstupní oblast zadali i včetně buňky B2 (tedy včetně popisku); pokud jsme jako vstupní oblast zadali B3:B15, kolonku „Popisky v prvním řádku“ nezaškrtáváme;
e) jako „Možnost výstupu“ ponecháme přednastavenou volbu „Nový list“ (výstup se pak objeví na novém listu s automaticky přiděleným názvem);
f) na závěr zaškrtneme jako požadovaný obsah výstupu možnosti „Celkový přehled“ i „Hladina spolehlivosti pro stř. hodnotu“, u níž ponecháme přednastavených 95 %;
g) analýzu potvrdíme kliknutím na tlačítko OK.

Automaticky se otevře nový list obsahující číselné charakteristiky pro zadanou veličinu „věk“ ve zpracovávaných datech. V tuto chvíli jsou potřebné pouze zcela první a zcela poslední údaj: první údaj (v české verzi Excelu je pro něj použito poněkud zavádějící označení „stř. hodnota“, konkr. zde hodnota 11,46) je již známým bodovým odhadem neboli „obyčejným“ aritmetickým průměrem věku; poslední údaj (v české verzi Excelu poněkud nešťastně pojmenován „hladina spolehlivosti“) je údaj potřebný k dopočtení hledaných mezí intervalu spolehlivosti (označme jej jako H, zde vyjde H=0,68). Obecně pak určíme pomocí průměru a hodnoty H obě meze hledaného intervalového odhadu takto:

$$A=\text{průměr}-H \qquad B=\text{průměr}+H \tag{29}$$

V příkladu s věkem tedy po dosazení do (29) dostáváme konkrétně hodnoty

$$A = 11{,}46-0{,}68 = 10{,}78 \tag{30}$$

$$B = 11{,}46+0{,}68 = 12{,}14 \tag{31}$$

Na základě zaznamenaných 13 hodnot jsme zjistili, že pro celou populaci sledovaných dětských pacientů pro daný (zde blíže nespecifikovaný) typ onemocnění je (s 95% spolehlivostí) neznámým středním věkem hodnota z rozmezí 10,78 až 12,14 roku.

□

Poznamenejme závěrem, že intervalový odhad typu (A; B) se někdy nazývá oboustranný nebo též dvoustranný interval spolehlivosti, abychom jej odlišili od případu, kdy nás z obou mezí zajímá pouze jedna. Takovým případům pak říkáme, že konstruujeme tzv. jednostranné intervaly spolehlivosti. Neplatí při tom, že bychom sestrojili interval oboustranný a pak pouze „převzali" jednu z jeho mezí (A nebo B); na konstrukci jednostranných intervalových odhadů je potřeba použít speciální vzorce (které jsou těm oboustranným sice dosti podobné, ne však zcela).

Metody statistické indukce – parametrické testy

Úvod do testování hypotéz – nulová a alternativní hypotéza

Předchozí kapitola byla věnována problému, jak odhadnout na základě statistického výběru neznámou hodnotu populačního parametru „střední hodnota" v Gaussově modelu normálního rozdělení. Položme si nyní ale otázku, zda je vůbec Gaussův model pro naše data vhodný? Na rozdíl od problematiky odhadů, kdy jako odpověď očekáváme číselný údaj, jde v tomto případě o otázku zjišťovací, tedy otázku se dvěma možnými odpověďmi (obecně „ano" nebo „ne"). Jiná situace, byť stále v souvislosti s parametrem „střední hodnota": předpokládáme, že by střední koncentrace dané látky v populaci zdravých jedinců měla činit 12 mmol/l a chceme na základě statistického výběru toto tvrzení potvrdit nebo vyvrátit. Ani zde tedy již nejde o úkol odhadnout neznámou hodnotu parametru, ale opět se jedná o otázku zjišťovací, neboť očekáváme pouze jednu ze dvou odpovědí (zde: „platí" či „neplatí"). Ve statistice ale mohou být kladeny i zjišťovací otázky, které se parametrů netýkají – např. můžeme chtít ověřit, zda je (nebo není) závislost mezi použitým léčebným postupem a mírou jeho úspěšnosti.

Pro každý typ problému, vedoucího ve statistice ke zjišťovací otázce, je dána dvojice navzájem se doplňujících hypotéz, které obecně značíme H_0 (nulová hypotéza) a H_a, případně H_1 (alternativní hypotéza). Nulová hypotéza má pro každý typ problému pevně daný tvar a alternativní hypotéza je tvrzením k ní opačným. Při tom je jedno, zda si testované tvrzení přejeme prokázat nebo naopak zamítnout.

Příklad 15 (ukázka formulace statistických hypotéz):
Předpokládejme, že pomocí statistického výběru prokazujeme tvrzení o tom, jestli prevalence nějakého onemocnění v dané populaci činí či nečiní 20 %. Konkrétní zadání by mohlo znít takto: „Prokažte, že prevalence daného onemocnění skutečně činí 20 %". Jde o statistický test, týkající se pravděpodobnostního parametru π, přičemž chceme rozhodnout, jestli $\pi=0{,}2$, nebo naopak $\pi\neq0{,}2$. Dvojice statistických hypotéz má tvar

$$H_0: \pi=0{,}20 \qquad H_a: \pi\neq0{,}20 \tag{32}$$

přičemž s ohledem na výše uvedené konkrétní zadání bychom zřejmě rádi prokázali platnost H_0.

Ke zcela stejné zjišťovací otázce by ale vedlo i např. toto konkrétní zadání: „Prokažte, že prevalence daného onemocnění je jiná než 20 %". (Např. pokud bychom chtěli prokázat, že se zkoumaná populace v tomto ohledu liší od jinak běžně platných standardů.) I v tomto případě bude mít dvojice statistických hypotéz tvar přesně stejný jako (32); jedinou změnou je teď to, že nyní bychom zřejmě naopak rádi prokázali neplatnost H_0.

□

Jako poučení z příkladu 15 bychom si měli zapamatovat, že pro každý typ statistického testu je důležité znát odpovídající znění nulové hypotézy. Právě o jejím osudu vlastně test rozhoduje. Výsledkem statistického testu je tedy vždy pouze jedna ze dvou odpovědí a) či b):

a) na základě dat zamítáme H_0 (příp. nepodařilo se prokázat H_0);
b) na základě dat nelze H_0 zamítnout.

Uvedené možnosti a), resp. b) jsou tedy výsledkem statistického testu. Následné slovní interpretace však už samozřejmě můžeme (jako důsledek tohoto rozhodnutí o osudu H_0) formulovat (pokud je to pro nás zajímavější) i jako rozhodnutí o H_a. Takže pokud výsledkem bude situace a), můžeme pak v odpovědi prohlásit např., že „jsme nevyvrátili platnost alternativní hypotézy"; a pokud nastane naopak situace b), můžeme prohlásit např., že „se nepodařilo alternativní hypotézu prokázat".

Povšimněme si jakési opatrnosti ve výše uvedených ukázkách možných slovních odpovědí. Snažíme se tím naznačit toho, že jsme si vědomi možného rizika mylného rozhodnutí. Statistickým testem totiž rozhodujeme o chování celé populace, ovšem jen na základě výběru z ní. Takže vždy je tu (byť s malou pravděpodobností) nebezpečí, že byla náhodou vybrána natolik nereprezentativní data, že naše rozhodnutí na nich založené je chybné.

Příklad 16 (ukázka situace vedoucí k chybnému rozhodnutí kvůli nereprezentativnosti výběru):
Chceme pomocí statistického výběru prokázat např. tvrzení o tom, že střední výška mužů (tedy parametr μ) v dané populaci činí 180 cm. Předpokládejme, že navíc na rozdíl od praxe víme, že je to pravdivé tvrzení. Dvojice statistických hypotéz má zde tvar

$$H_0: \mu=180 \qquad H_a: \mu\neq180 \tag{33}$$

Rozhodování založíme na výběru 30 mužů. Bude-li tento výběr dostatečně reprezentativní (co do zkoumané veličiny tělesná výška), bude výsledkem odpověď b) - na základě dat nelze H_0 zamítnout, což by bylo správné rozhodnutí.

Co kdyby ale všech 30 mužů byli ligoví basketbalisté? Podle jejich tělesných výšek by to pak mohlo vypadat, že muži v dané populaci musí být výrazně vyšší než 180 cm, takže výsledkem na základě takovýchto dat by mohla být odpověď a) - na základě dat zamítáme H_0, tedy rozhodnutí chybné.

□

Princip testování hypotéz – hladina významnosti a p-hodnota

Zdrojem rizika chybného rozhodnutí při statistickém testování je tedy vlastně riziko, že náhodně vybereme nedostatečně reprezentativní data. Toto riziko má sice naštěstí skutečně malou pravděpodobnost, ale v praxi s ním musíme kalkulovat. Uvědomme si, že při jakémkoli statistickém testu může nastat jedna z těchto čtyř možností (viz též tabulka 9):

A. na základě dat zamítáme H_0 a je to správné rozhodnutí;
B. na základě dat nelze H_0 zamítnout a je to správné rozhodnutí;
C. na základě dat zamítáme H_0 a je to chybné rozhodnutí;
D. na základě dat nelze H_0 zamítnout a je to chybné rozhodnutí.

Sice v praxi nevíme, která možnost vlastně nastala (protože nemůžeme zjistit, zda rozhodnutí učiněné na základě dat skutečně je či není správné), ale můžeme alespoň technicky omezit velikosti pravděpodobností, že k chybnému rozhodnutí (tedy k některé z možností C. nebo D.) dojde. Situaci C., tedy chybné zamítnutí nulové hypotézy, nazýváme chyba prvního druhu. Situaci D., tedy chybné přijetí nulové hypotézy, nazýváme chyba druhého druhu. Ještě před provedením statistického testu stanovujeme, jaká pravděpodobnost chyby prvního druhu je pro nás přijatelná. Tato stanovená hranice pro pravděpodobnost prvního druhu se nazývá hladina významnosti (příp. „stanovená hladina významnosti", nebo jen „významnost") statistického testu. Značíme ji řeckým písmenem α (alfa). Není-li uvedeno jinak, pracujeme automaticky s hodnotou 5 % (α=0,05). Další používané volby

Tabulka 9: Možnosti při statistickém testování

		Ve skutečnosti (což ale v praxi nelze ověřit):	
		H_0 platí	H_0 neplatí (platí H_a)
Rozhodnutí dle dat:	zamítáme H_0	nastala chyba I. druhu	správné rozhodnutí
	nelze zamítnout H_0	správné rozhodnutí	nastala chyba II. druhu

jsou 10 % (α=0,10), 1 % (α=0,01), příp. 1‰ (α=0,001). V odpovědi bychom vždy měli hodnotu α uvést (např. použitím formulace „na 5% hladině významnosti jsme na základě dat nevyvrátili tvrzení…“).

Zbývá prozradit jediné – podle čeho vlastně při provádění statistického testu vybereme jednu z oněch dvou možných odpovědí a) či b)? Jednou z možností (při využití počítačového SW se zabudovanými statistickými metodami, tedy i v Excelu) je spustit pro zadaná data odpovídající test. Výsledným počítačovým výstupem je tzv. p-hodnota (též tzv. „dosažená hladina významnosti“; vždy jde o číslo z rozmezí 0 až 1). Rozhodnutí učiníme porovnáním s předem zvolenou hladinou významnosti (α) takto:

a) pokud vyjde p-hodnota $\leq \alpha$, zamítáme (na dané hladině významnosti) H_0;
b) pokud vyjde p-hodnota $> \alpha$, nelze (na dané hladině významnosti) H_0 zamítnout.

V případě, že testujeme tvrzení o střední hodnotě (μ) jako v příkladě 16, můžeme místo konceptu s p-hodnotou využít intervalový odhad (A; B), sestrojený pro tento parametr s nějakou spolehlivostí. Přitom hladina významnosti (v procentech) odpovídá hodnotě

100–spolehlivost (34)

takže nejčastějšímu případu, kdy volíme spolehlivost 95 %, odpovídá 5% hladina významnosti. Rozhodnutí pak učiníme takto:

a) pokud testovaná hodnota nenáleží do (A; B), zamítáme (na dané hladině významnosti) H_0;
b) pokud testovaná hodnota náleží do (A; B), nelze (na dané hladině významnosti) H_0 zamítnout.

Příklad 17 (dokončení příkladu 16 - provedení statistického testu pomocí intervalu spolehlivosti): Testujeme dvojici statistických hypotéz (33). K dispozici byla data o výškách 30 mužů (zde nebudeme uvádět), předpokládejme, že na jejichž základě již byl postupem jako v příkladu (19) sestrojen s 95% spolehlivostí tento intervalový odhad pro střední výšku: (179,64; 187,56). Protože testovaná hodnota 180 cm je číslo, které do tohoto intervalu spolehlivosti náleží, nelze na 5% hladině významnosti H_0 zamítnout. Na základě dat, která byla k dispozici, bychom tak potvrdili, že střední výška mužů v dané populaci může činit 180 cm.

□

V následujících kapitolách budou předvedeny některé nejdůležitější typy statistických testů, které lze pomocí Excelu (s využitím konceptu p-hodnoty) realizovat. Zdůrazněme, že k tomu, aby byl test skutečně správně použit, musí být v praxi ověřena i platnost některých důležitých předpokladů. Tyto předpoklady budou u jednotlivých testů také uvedeny, přičemž ale bohužel ne vždy je možno tyto předpoklady pomocí Excelu jednoduše ověřit.

Párový t-test

Předpokládejme, že máme k dispozici za každou statistickou jednotku dvojici (pár, odtud název párový test) číselných hodnot spojitého typu, např.:

- věk muže a ženy pro každý manželský pár (statistickou jednotkou je zde manželský pár);
- délku (v cm) pravé a levé paže (statistickou jednotkou je zde jedinec), apod.

Chceme ověřit, zda v zjišťovaných párech hodnot je či není statisticky významná odlišnost, tedy zda

- jsou či nejsou muži a ženy v manželských párech stejně staří?
- mají či nemají jedinci obě paže stejně dlouhé? apod.

To, co nás tedy vlastně za každou sledovanou statistickou jednotku skutečně zajímá, není ona dvojice hodnot, ale jejich vzájemný rozdíl. Můžeme tedy vždy spočítat např.

- pro každý manželský pár: o kolik roků je muž starší (příp. mladší) než jeho manželka;
- pro každého jedince: o kolik cm je pravá paže delší (příp. kratší) než levá.

Tabulka 10: Data v Excelu pro párový test (převzato ze Zvárová: Biomedicínská statistika)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Hypertonik č.	1	2	3	4	5	6	7	8	9	10	11
2	Placebo	211	210	210	203	196	190	191	177	173	170	163
3	Hydrochlorothiazid	181	172	196	191	167	161	178	160	149	119	156

O takto vytvořené „rozdílové“ veličině budeme předpokládat, že se řídí modelem Gaussova normálního rozdělení. Tento předpoklad normality je nezbytný k tomu, aby směl být následující test vůbec aplikován. Alespoň přibližně bychom si představu o splnění tohoto předpokladu udělali tak, že hodnoty jednotlivých rozdílů převedeme intervalovým rozdělením do podoby kategoriální veličiny (alespoň s pěti kategoriemi) a sestrojíme její histogram; pokud bude tento histogram přibližně kopírovat průběh Gaussovy křivky, můžeme konstatovat, že normalitě dat bylo zřejmě vyhověno.

Test o shodě, resp. rozdílnosti párových hodnot, se díky zavedení jejich rozdílů zjednodušil na test

$$H_0: \mu=0 \qquad H_a: \mu\neq 0 \tag{35}$$

kde μ značí střední hodnotu rozdílové veličiny (střední rozdíl). Slovně lze dvojici hypotéz (35) vyjádřit např. takto: nulová hypotéza předpokládá, že není statisticky významný (střední) rozdíl mezi analyzovanými párovými údaji, naopak alternativní hypotéza předpokládá, že (střední) rozdíl je statisticky významný. Takto formulované dvojici hypotéz říkáme oboustranný test (test s oboustrannou, příp. s dvoustrannou alternativou).

Pokud bychom předpokládali, že případný rozdíl v párových hodnotách musí být převážně ve prospěch jednoho z obou párových údajů (čili že např. v populaci převažují manželské páry, v nichž muž je starší než jeho manželka), lze test (35) přepsat do tvaru

$$H_0: \mu=0 \qquad H_a: \mu>0 \tag{36}$$

kterému říkáme jednostranný test (test s jednostrannou alternativou); znaménko nerovnosti v H_a přitom závisí na tom, v jakém pořadí byly jednotlivé rozdíly vypočítávány – pokud bychom např. v manželských párech počítali vždy rozdílovou hodnotu jako věk ženy mínus věk muže, bylo by zřejmě logické použít naopak situaci s H_a: $\mu<0$. Každopádně i v případě jednostranného testu předpokládá nulová hypotéza, že (střední) rozdíl mezi analyzovanými párovými údaji není statisticky významný, zatímco alternativní hypotéza tvrdí, že rozdíl významný je.

Příklad 18 (párový t-test - studie účinku hydrochlorothiazidu na krevní tlak):
Tento příklad je převzatý z učebnice Zvárová, J.: Biomedicínská statistika I. - Základy statistiky pro biomedicínské obory. Jedná se o studii Colton (Boston, 1974), která se zabývala studiem účinku hydrochlorothiazidu (HCT) na krevní tlak. Dvojicí hodnot pro každého z 11 hypertoniků bylo měření jeho systolického tlaku po podání placeba a (po měsíci) po podání HCT, viz tabulka 10. Otázka zní, zda tato data prokázala statisticky významný rozdíl mezi účinkem placeba a HCT. Není-li předem jasné, zda očekávaným účinkem HCT má být významné snížení či snad naopak zvýšení krevního tlaku, použijeme oboustranný test (35). Výhodou použití Excelu je to, že není nutno ze zadaných dat počítat rozdíly, stačí odkaz na párové hodnoty tak, jak byly napozorovány. Obecný vzorec na spuštění párového testu v oboustranné verzi je

$$=\text{TTEST(data1;data2;2;1)} \tag{37}$$

pokud bychom potřebovali aplikovat jednostrannou verzi, má příkaz tvar

$$=\text{TTEST(data1;data2;1;1)} \tag{38}$$

V našem příkladu musí „data1“ odkazovat např. na oblast obsahující údaje o tlaku po podání placeba (zde na buňky B2:L2) a „data2“ odkaz na druhé údaje, tedy na tlak po podání HCT (zde na buňky B3:L3). Příkaz (37) bude mít konkr. podobu =TTEST(B2:L2;B3:L3;2;1) a jeho výsledkem je odpovídající p-hodnota (=0,00012). Jelikož vyjde p-hodnota menší než α (platí: $0{,}00012\leq 0{,}05$), zamítáme na 5%

hladině významnosti nulovou hypotézu. Znamená to, že data prokázala významný rozdíl mezi účinkem placeba a účinkem HCT (neboli že byl prokázán významný vliv HCT na krevní tlak, bez bližšího upřesnění, zda HCT způsobuje snížení či naopak zvýšení).

Pokud bychom od začátku předpokládali, že HCT má buď účinek srovnatelný s placebem, nebo že HCT krevní tlak snižuje (ale rozhodně ho nemůže zvyšovat), použili bychom jednostranný přístup. Vzorec (38) by pak měl tvar =TTEST(B2:L2;B3:L3;1;1) a jeho výsledkem je odpovídající p-hodnota (=0,00006). Mimochodem: p-hodnota jednostranného testu je vždy poloviční oproti p-hodnotě testu oboustranného. I tímto přístupem bychom tedy prokázali významný rozdíl mezi účinkem placeba a účinkem HCT, ale mohli bychom také říci, že byl prokázán významný vliv HCT na snížení krevního tlaku (formulace se slovem „snížení" při oboustranném přístupu přípustná nebyla).

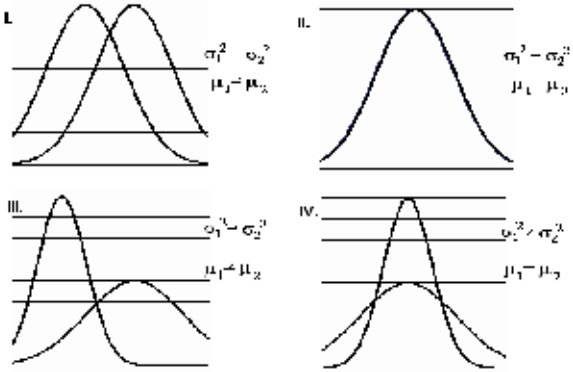
Poznamenejme, že předpoklad normality byl pro tato data prokázán speciálním testem normality (který je ale nad rámec jednoduchých aplikací v Excelu).

S dalším typem t-testu se setkáme u dvou-výběrových testů. Na závěr vysvětlení, proč je používán název t-test. Důvodem je to, že p-hodnota je počítána porovnáním s chováním tzv. Studentova t-rozdělení, což je jeden z modelů odvozených od modelu Gaussova normálního rozdělení.

Dvou-výběrový F-test

Předpokládejme, že máme k dispozici výběry ze dvou navzájem nezávislých populací. V obou populacích zaznamenáváme stejnou veličinu spojitého typu, např. u mužů a u žen jejich věk. Na rozdíl od párového přístupu ale zde nejsou jednotliví muži přiřazeni ke konkrétním ženám (dle požadavku vzájemné nezávislosti obou skupin), dokonce počet vybraných mužů může být obecně jiný než počet vybraných žen. Předpokládáme pouze, že u obou populací je vhodným modelem pro sledovanou veličinu (zde pro věk) model Gaussova normálního rozdělení (v praxi musí být tento předpoklad ověřován, což zde jako už u párového testu vynecháme). Znamená to, že každá z obou populací může být reprezentována vhodnou Gaussovou křivkou, jejíž tvar je jednoznačně určen hodnotami parametrů μ (střední hodnota) a σ^2 (teoretický rozptyl). Označme tyto parametry pro první populaci jako μ_1 a σ_1^2 a pro druhou populaci jako μ_2 a σ_2^2. Podle toho, zda platí $\sigma_1^2=\sigma_2^2$ (shodná variabilita, tj. křivky musí mít stejnou šířku) nebo $\sigma_1^2\neq\sigma_2^2$ (odlišná variabilita – křivky mají různé šířky), resp. $\mu_1=\mu_2$ (křivky mají vrchol v témže bodě) nebo $\mu_1\neq\mu_2$ (vrcholy křivek jsou v různých bodech), mohou nastat čtyři typově různé situace, označené na obrázku 6 jako možnosti I. až IV.

Obrázek 6: Možnosti pro dvojice populací charakterizovaných Gaussovými křivkami



Tabulka 11: Data pro dvou-výběrový test

	A	B
1	13,33	ALTE
2	32,5	Ostatní
3	11,67	ALTE
4	19	Ostatní
5	5	ALTE
6	22,67	ALTE
7	11	ALTE
8	29	Ostatní
9	23	ALTE
10	20,67	Ostatní
11	14,33	Ostatní
12	9,33	ALTE
13	14,17	ALTE
14	27,83	Ostatní
15	13,83	ALTE
16	9,17	ALTE
17	32	Ostatní
18	7	ALTE
19	8,83	ALTE
20	22,33	ALTE
21	8,17	ALTE
22	8,33	ALTE
23	22,33	Ostatní
24	21,17	ALTE
25	15,5	ALTE
26	11,33	Ostatní
27	9,33	ALTE
28	17,33	ALTE
29	15,17	ALTE
30	18	ALTE
31	7,67	ALTE
32	20,6	ALTE
33	35	Ostatní
34	13,67	Ostatní
35	31,33	Ostatní
36	17,33	Ostatní
37	31,17	Ostatní
38	9,33	ALTE
39	24,67	ALTE
40	9,67	ALTE
41	17,83	Ostatní

To, která ze situací I. až IV. nastala, pomáhají na základě dat rozhodnout dvou-výběrové testy (nejprve provádíme F-test, následně t-test). Konkrétně dvou-výběrový F-test rozhoduje o platnosti nulové hypotézy týkající se shody rozptylů:

$$H_0: \sigma_1^2=\sigma_2^2 \qquad H_a: \sigma_1^2\neq\sigma_2^2 \qquad (39)$$

Za platnosti nulové hypotézy víme, že zkoumané dvojici populací odpovídá na obrázku 6 situace I. nebo II. a naopak za platnosti alternativní hypotézy je odpovídající situací III. nebo IV. (K jednoznačnému rozhodnutí o konkrétním typu bude nutno následně provést ještě dvou-výběrový t-test, to ale viz až další kapitola.)

Příklad 19 (rozdíly mezi dvěma skupinami novorozenců – seřazení dat a provedení F-testu):
Také tento příklad je (s drobnými úpravami) převzatý z učebnice Zvárová, J.: Biomedicínská statistika I. - Základy statistiky pro biomedicínské obory. Jedná se o studii na posouzení, zda se chování veličiny LTV („long term variability“, definovanou jako rozdíl mezi maximální a minimální hodnotou tepové frekvence) významně liší u novorozenců, kteří prodělali ALTE („apparent life threatening event“, tedy událost očividně ohrožující jejich život), a u novorozenců ostatních.

Prvním úkolem v Excelu bývá příprava dat. Obvyklý způsob zápisu totiž v tomto případě bývá podobný jako v tabulce 11: ve sloupci A máme zaznamenán zjištěný údaj LTV hodnotu po hodnotě (tedy tak, jak byly hodnoty chronologicky získávány), ve vedlejším sloupci B je poznačeno, do které ze dvou skupin dotyčný novorozenec patřil (ALTE nebo Ostatní). Takto zapsaná data by bylo nutno seřadit: Nejprve vybereme oba sloupce A a B a v nabídce Data zvolíme možnost Seřadit. (Pokud byl vybrán jen jeden sloupec, potvrdíme, že chceme rozšířit vybranou oblast; v menu, které se pak otevře, zkontrolujeme, že pro tato data není zaškrtnuta možnost „Data obsahují záhlaví“.) Pak v nabídce „Seřadit podle“ vybereme možnost Sloupec B a odklikneme OK. Data se seřadí tak, že ve sloupci A již nebudou hodnoty LTV napřeskáčku jako původně, ale pohromadě pro skupinu ALTE (na 1. až 26. řádku) a pro skupinu Ostatní (na 27. až 41. řádku).

Máme-li všech 41 hodnot LTV již seřazeno pod sebou ve sloupci A (nejprve tedy 26 hodnot LTV pro skupinu ALTE, pod nimi 15 hodnot LTV pro skupinu Ostatní), je vše připraveno ke spuštění F-testu (39). Příkaz má obecný tvar:

=FTEST(data1;data2) (40)

kde argument „data1“, resp. „data2“ odkazuje na buňky s hodnotami pro první, resp. druhou skupinu. V našem případě musí mít tedy příkaz (40) konkrétní podobu

=FTEST(A1:A26;A27:A41) (41)

Výsledkem je p-hodnota p=0,17. Protože je tato p-hodnota větší než α=0,05, nelze na 5% hladině významnosti zamítnout H_0. Podle (39) to znamená, že rozptyly (variability) obou skupin lze považovat za srovnatelné, jinak řečeno, oběma populacím novorozenců (ALTE i Ostatní) odpovídá na obrázku 6 buď možnost I. (Gaussovy křivky pro LTV budou stejně široké, jen navzájem posunuté), nebo možnost II. (Gauusovy křivky pro LTV budou identické – splynou). Abychom rozhodli jednoznačně, která možnost odpovídá našim datům, musí ještě následovat t-test, viz další kapitola. (Podobně jako u příkladu 18 bylo i zde vynecháno poměrně komplikované ověření samotné normality.)

Na závěr vysvětlení, proč je tento test nazýván F-testem. Důvodem je to, že p-hodnota je určována porovnáním s chováním tzv. Fisherova F-rozdělení, což je další z modelů odvozených od modelu Gaussova normálního rozdělení.

Dvou-výběrový t-test

Předpokládejme stejně jako v předešlé kapitole, že máme k dispozici výběry ze dvou navzájem nezávislých populací a že u obou populací je vhodným modelem pro sledovanou veličinu model Gaussova normálního rozdělení, přičemž střední hodnotu pro první populaci značíme μ_1 a pro druhou populaci μ_2. Dále předpokládejme, že již byl aplikován dvou-výběrový F-test, takže je známo, zda lze u obou populací předpokládat shodnou či naopak rozdílnou variabilitu. To, o čem zbývá nyní rozhodnout, je dvojice statistických hypotéz

$$H_0: \mu_1=\mu_2 \qquad H_a: \mu_1\neq\mu_2 \tag{42}$$

Příklad 20 (dokončení příkladu 19 - rozdíly mezi dvěma skupinami novorozenců - provedení t-testu):
Mějme všech 41 hodnot LTV z tabulky 11 již seřazeno pod sebou ve sloupci A (nejprve tedy 26 hodnot LTV pro skupinu ALTE, pod nimi 15 hodnot LTV pro skupinu Ostatní) jako před spuštěním F-testu. Příkaz pro provedení dvou-výběrového t-testu má tento obecný tvar:

=TTEST(data1;data2;strany;typ) (43)

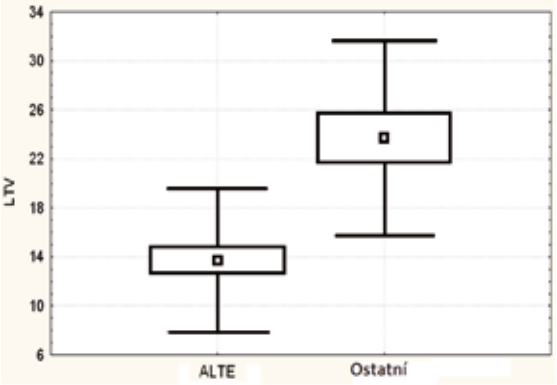
kde argument „data1“, resp. „data2“ odkazuje na buňky s hodnotami pro první, resp. druhou skupinu; za argument „strany“ dosazujeme hodnotu 1, chceme-li test s jednostrannou alternativou, nebo 2 při oboustranné alternativě; a za argument „typ“ můžeme dosadit hodnotu 2 v případě, kdy byla F-testem prokázána shoda rozptylů, nebo hodnotu 3 vždy, když předchozí F-test naopak prokáže rozdílnost rozptylů obou populací (hodnota 1 byla vyhrazena pro párový t-test, srovnej s (37), (38)). V našem případě, kdy chceme provést test s oboustrannou alternativou a kdy díky testu (41) již víme, že rozptyly lze považovat za srovnatelné, bude mít tedy příkaz (43) konkrétní podobu

=TTEST(A1:A26;A27:A41;2;2) (44)

Výsledkem je p-hodnota p=0,00004. Protože je tato p-hodnota menší než α=0,05, zamítáme na 5% hladině významnosti H_0 ve znění (42). Jinak řečeno, data prokázala statisticky významnou odlišnost středních hodnot LTV mezi populací novorozenců ALTE a populací novorozenců ostatních. Na základě výsledků obou dvou-výběrových testů (F-testu a t-testu) bylo zjištěno, že ze čtyř možností na obrázku 6 odpovídá našim datům situace I.

V případě, kdy porovnáváme chování spojité veličiny ve dvou (nebo i více) skupinách, je vhodné použít pro její grafické znázornění tzv. box-plot (do češtiny někdy překládáno jako krabicový graf). Přestože v moderní odborné literatuře jde o oblíbený typ grafu, v Excelu jej jednoduše přímo z nabídky vytvořit nelze (nejblíže mu je asi typ grafu „Burzovní“, ale pro jeho použití by bylo nutno nejprve data předzpracovat). Na obrázku 7 byl tento graf vytvořen pro data z příkladu 19 a 20 pomocí SW *STATISTICA*.

Obrázek 7: Box-plot s momentovými charakteristikami pro data LTV



V grafu typu box-plot jsou na vodorovné ose uvedeny jednotlivé kategorie, pro něž byly samostatně provedeny deskriptivní výpočty, a které tím pádem můžeme co do deskriptivních charakteristik vzájemně porovnat (na obrázku 7 můžeme porovnat kategorie ALTE a Ostatní). Číselné hodnoty vybraných deskriptivních charakteristik odečítáme na svislé ose. Obvykle značí poloha malého čtverečku průměrnou hodnotu sledované veličiny (na obrázku 7 jde tedy o průměrné hodnoty LTV) v jednotlivých kategoriích. Obdélník kolem tohoto čtverečku vymezuje (ve svislém směru) rozmezí intervalu spolehlivosti pro střední hodnotu sledované veličiny v rámci dané kategorie, určeného podobně jako v příkladu 14. Úsečky směrem nad, resp. pod obdélník pak vymezují (symetricky ve svislém směru) od polohy průměru vzdálenost plus, resp. mínus směrodatná odchylka, vymezují tedy „typické" rozmezí dat podobně jako v příkladu 9. Z obrázku 7 bychom tak měli být schopni vyčíst, že obě skupiny (ALTE a Ostatní) jsou zhruba srovnatelné co do variability - jim odpovídající grafy byly zhruba stejně vysoké (míněna je celá jejich výška, tedy od dolní k horní úsečce), ale v datech se projevil výrazný rozdíl mezi průměry - hodnoty LTV ve skupině ALTE byly zhruba poloviční oproti skupině Ostatní (krabičky mají své středy poblíž hodnot 14, resp. 24). Jinak řečeno, obrázek 7 je vlastně grafickým potvrzením toho, co jsme exaktně prokázali pro obě srovnávané populace díky testům provedeným v příkladech 19 a 20. Pro úplnost dodejme, že box-plot může jindy znázorňovat třeba kvantilové či jiné charakteristiky - čtvereček uprostřed pak odpovídá třeba poloze mediánu, úsečky vymezují rozmezí mezi maximální a minimální zjištěnou hodnotou, apod., takže aby nedošlo k dezinterpretaci, je vhodné u každého grafu tohoto typu uvést vysvětlivky.

Na závěr ještě alespoň jeden komentář: Co když budeme chtít porovnat dvě populace, jejichž výběry nebudou splňovat požadavek normality? Zkušený statistik má pro takový případ řešení (lze použít tzv. neparametrický test, viz např. Anděl, Statistické metody), ale to už je opět nad rámec standardních možností Excelu.

Jedno-faktorová analýza rozptylu (ANOVA)

Předpokládejme podobně jako v předešlých kapitolách, že máme k dispozici výběry z navzájem nezávislých populací a že u těchto populací je vhodným modelem pro sledovanou veličinu model Gaussova normálního rozdělení. Nyní však provedeme zobecnění a to v tom smyslu, že těchto populací může být libovolný počet (nejen dvě). Označme počet porovnávaných populací (a tudíž výběrů) jako K, kde tedy obecně $K \geq 2$. Kdybychom nyní chtěli (podobně jako na obrázku 6) rozlišovat situace s různými či stejnými variabilitami a středními hodnotami, došli bychom jen při třech populacích ($K=3$) k 25 různým možnostem a pro vyšší počty ($K>3$) by počet různých možností narůstal ještě dramatičtěji. Doplníme proto předpoklad o homogenitě variability (neboli o shodě rozptylů).

Jinak řečeno, budeme předpokládat, že všechny Gaussovy křivky odpovídající jednotlivým populacím mají shodnou šířku. Jediné, co zbývá posoudit, je chování jejich středních hodnot. Označme střední hodnotu v první, druhé atd. až poslední (*K*-té) kategorii jako μ_1, μ_2,…, μ_K. Pak nulová hypotéza předpokládá, že střední hodnoty budou shodné pro všechny kategorie:

$$H_0: \mu_1=\mu_2=\ldots=\mu_K \quad (45)$$

Alternativní hypotéza předpokládá opak, tedy to, že aspoň v jedné kategorii se střední hodnota významně liší od ostatních. To v sobě zahrnuje více různých možností, které zapisujeme jako:

$$H_a: \text{non } H_0 \quad (46)$$

kde latinské slovo „non“ značí obecně negaci, opak.

Tento test se často nazývá ANOVA, což je zkratka anglických slov „analysis of variance“, doslovně přeloženo jako „analýza rozptylu“. Důvodem k tomuto označení není znění hypotéz (ty se vůbec o rozptylu, tedy o parametru σ^2 nezmiňují), ale to, že číselné charakteristiky variability jsou podstatou výpočtů, na nichž je pak založeno rozhodnutí o nulové hypotéze. Tyto výpočty si však zde vysvětlovat nebudeme, opět se spokojíme s ukázkou praktického provedení tohoto testu v Excelu.

Příklad 21 (ANOVA – ilustrační příklad):
Sledujeme hmotnost pacientů, ty rozlišujeme podle vzdělání do tří (*K*=3) kategorií (ZŠ-SŠ-VŠ). Předpokládáme, že se veličina hmotnost ve všech třech vzdělanostních pod-populacích řídí Gaussovým normálním rozdělením se shodnou variabilitou. Testem ANOVA zde rozhodujeme o této dvojici hypotéz:

$$H_0: \mu_1=\mu_2=\mu_3, \qquad H_a: \text{non } H_0 \quad (47)$$

Nulová hypotéza tedy předpokládá, že střední hmotnost bude ve všech třech vzdělanostních kategoriích stejná. Alternativní hypotéza naopak předpokládá, že v aspoň jedné ze tří vzdělanostních kategorií je střední hmotnost významně odlišná od ostatních.

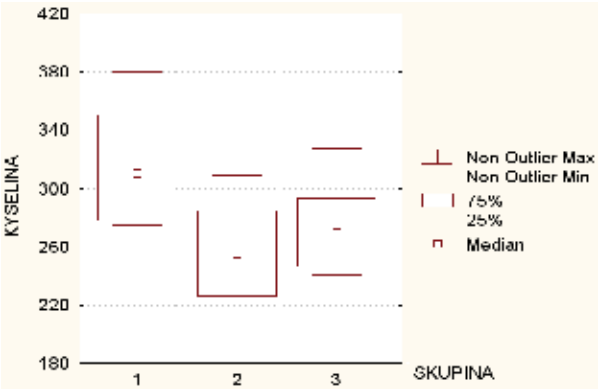
Uvědomme si, že alternativní hypotézu z příkladu 21 lze také chápat jako závislost sledované číselné veličiny (zde: hmotnosti) na tom, do jaké skupiny (zde: kategorie dle vzdělání) pacient patří: „Střední hmotnost závisí na tom, v jaké jsme vzdělanostní kategorii“. Tuto situaci též interpretujeme tak, že podle alternativní hypotézy by bylo vzdělání faktorem, který významně ovlivňuje hmotnost (vzdělání jako faktor, na němž hmotnost významně závisí). Naopak nulová hypotéza je vlastně hypotézou o nezávislosti sledované číselné veličiny (zde hmotnosti) na zkoumaném faktoru (zde na vzdělání). Z uvedeného důvodu se takto formulovaný test také někdy nazývá jedno-faktorová ANOVA. Existují i problémy vedoucí k více-faktorovému testu ANOVA - např. kdybychom zjišťovali, jak závisí hmotnost např. na dvou faktorech: na vzdělání a k tomu ještě na pohlaví.

Příklad 22 (provedení testu typu jedno-faktorová ANOVA v Excelu):
Dvacet dva pacientů, kteří podstoupili operaci srdce, bylo náhodně rozděleno do tří skupin. Skupina 1: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi nepřetržitě po dobu 24 hodin. Skupina 2: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi pouze během operace. Skupina 3: Pacienti nedostali žádný oxid dusný, ale dostali 35-50 % kyslíku po dobu 24 hodin. (Zdroj: Zvárová, Základy statistiky…). Sledovanou číselnou veličinou byla u všech pacientů hodnota koncentrace soli kyseliny listové v červených krvinkách. Předpokládáme, že se tato veličina ve všech třech populacích řídí Gaussovým normálním rozdělením se shodnou variabilitou. Data (již seřazena dle hodnot ve sloupci B) jsou k dispozici v tabulce 12. Jejich deskriptivní charakteristiky graficky nejlépe vystihne box-plot (podobně jako u porovnání dvou skupin dvou-výběrovým testem), viz obrázek 8 (na rozdíl od obrázku 7 je v kvantilové verzi). Díky grafu 8 lze konstatovat, že vzhledem k podobné výšce krabiček lze předpokládat splnění podmínky o shodě variability ve všech třech skupinách. Dále vidíme, že nejvyšších hodnot koncentrace soli kyseliny listové v krvi bylo dosaženo ve skupině první, naopak nejnižších ve skupině druhé, přičemž však rozdíl mezi druhou a třetí skupinou (alespoň tedy

Tabulka 12: Data ANOVA

	A	B
1	KYSELINA	SKUPINA
2	276	1
3	280	1
4	275	1
5	291	1
6	347	1
7	354	1
8	380	1
9	330	1
10	206	2
11	210	2
12	226	2
13	249	2
14	255	2
15	273	2
16	285	2
17	295	2
18	309	2
19	241	3
20	246	3
21	270	3
22	293	3
23	328	3

Obrázek 8: Box-plot pro srovnání tří kategorií



Tabulka 13: Výstup testu ANOVA v Excelu

	A	B	C	D	E	F	G
1	Anova: jeden faktor						
2							
3	Faktor						
4	*Výběr*	*Počet*	*Součet*	*Průměr*	*Rozptyl*		
5	Sloupec 1	8	2533	316,625	1699,411		
6	Sloupec 2	9	2308	256,444	1378,028		
7	Sloupec 3	5	1378	275,600	1288,300		
8							
9							
10	ANOVA						
11	*Zdroj variability*	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
12	Mezi výběry	15663,476	2	7831,738	5,301	0,015	3,522
13	Všechny výběry	28073,297	19	1477,542			
14							
15	Celkem	43736,773	21				

mezi jejich mediány) byl poměrně malý. Tyto úvahy byly ale pouze popisem dat a zatím nemůžeme nic prohlásit o chování populací, dokud neprovedeme odpovídající test. Testem ANOVA zde rozhodneme o stejné dvojici hypotéz, jako byla dvojice (47), neboť i zde máme K=3. Nulová hypotéza zde předpokládá, že mezi třemi porovnávanými populacemi pacientů není statisticky významný rozdíl ve střední koncentraci soli kyseliny listové, naopak alternativní hypotéza předpokládá, že se (co do střední hodnoty) alespoň jedna ze tří uvažovaných populací statisticky významně liší od ostatních.

Před samotnou realizací testu je ale nutno ještě data z tabulky 12 připravit do Excelem požadované podoby, podle níž musí být číselné údaje za každou skupinu v samostatných a sousedících sloupcích (případně řádcích). Hodnoty ze sloupce A musíme tedy překopírovat např. do sloupců C, D a E a to tak, aby v prvním z nich byly jen číselné údaje týkající se skupiny 1 (čili do sloupce C, konkr. do buněk C1 až C8, překopírujeme osm hodnot z buněk A2 až A9), ve druhém údaje pro skupinu 2 (čili do sloupce D, konkr. do buněk D1 až D9, překopírujeme devět hodnot z buněk A10 až A18) a ve třetím údaje pro skupinu 3 (do sloupce E, konkr. do buněk E1 až E5, překopírujeme pět hodnot z buněk A19 až A23). Celou tuto nově vytvořenou oblast dat vybereme pomocí myši (ve výběru tedy musíme mít obdélník C1 až E9) a aktivujeme (viz příklad 14) kliknutím nabídku Data – Analýza dat. V menu zvolíme analytický nástroj „Anova: jeden faktor“ a potvrdíme (OK). Zkontrolujeme, že jako vstupní

oblast dat máme vybrány buňky C1:E9, že je zaškrtnuto sdružení dat do sloupců, že naopak není zaškrtnuta možnost „Popisky v prvním řádku“ (první řádek ve sloupcích C až E obsahuje číselné údaje a ne vysvětlující popisek, jako tomu bylo ve sloupcích A a B) a že hladina významnosti alfa je automaticky nastavena na standardní hodnotu 0,05. Jako možnost výstupu ponecháme „Nový list“, takže po potvrzení (OK) se v Excelu automaticky vytvoří nový list s výstupy ANOVA jako v tabulce 13.

První část výstupu (řádky 3-7) obsahuje deskriptivní charakteristiky pro jednotlivé kategorie (zde „Sloupce“) zkoumaného faktoru. Povšimněme si, že hodnoty průměrů (hodnoty z buněk D5 až D7) zhruba odpovídají hodnotám mediánů na obrázku 8 a že hodnoty rozptylů (hodnoty z buněk E5 až E7) korespondují s již učiněným komentářem o tom, že lze předpokládat (nikoli však „prohlásit za ověřené“) splnění podmínky o shodné variabilitě všech populací. Samotné výpočty pro test ANOVA jsou v tabulce 13 dole (počínaje řádkem 10). Ze všech zde uvedených hodnot se soustředíme pouze na jedinou, a to na p-hodnotu=0,015 (údaj z buňky F12). Protože vyšla p-hodnota menší než hladina významnosti (0,015<0,05), zamítáme na 5% hladině významnosti nulovou hypotézu. Na základě dat tedy bylo prokázáno, že mezi třemi uvažovanými populacemi pacientů je statisticky významný rozdíl ve střední koncentraci soli kyseliny listové v krvi. Jinak řečeno, bylo prokázáno, že střední koncentrace soli kyseliny listové v krvi závisí na zkoumaném faktoru (tedy na tom, do které ze tří skupin pacient náleží).

Pomocí testu ANOVA bylo prokázáno, že mezi třemi uvažovanými populacemi pacientů je statisticky významný rozdíl ve střední koncentraci (viz příklad 22). Na co ale nebylo zodpovězeno, je to, která či které populace se od sebe významně odlišují. Všechny možnosti („ukryté“ v H_a) jsou tyto:

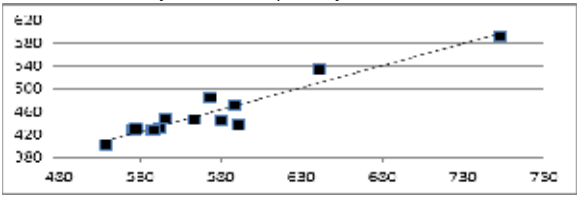
a) populace 1 se významně liší od populací 2 a 3, mezi nimiž vzájemně není významný rozdíl;
b) populace 2 se významně liší od populací 1 a 3, mezi nimiž vzájemně není významný rozdíl;
c) populace 3 se významně liší od populací 1 a 2, mezi nimiž vzájemně není významný rozdíl;
d) populace 1 se významně liší od populace 2, mezi zbylými dvojicemi není významný rozdíl;
e) populace 1 se významně liší od populace 3, mezi zbylými dvojicemi není významný rozdíl;
f) populace 2 se významně liší od populace 3, mezi zbylými dvojicemi není významný rozdíl;
g) každá populace se významně liší od každé ze dvou zbývajících.

Nezapomeňme, že znázornění box-plotem (viz obrázek 8) je deskripcí dat (čili pouhého výběru ze srovnávaných populací), nikoli dokladem chování celých populací. Můžeme podle něj tedy pouze tipovat, které z možností a) až g) připadají, resp. nepřipadají v úvahu: pro data z příkladu 22 by zřejmě připadala v úvahu některá z možností a), d) nebo g). Pokud bychom chtěli na základě dat opravdu rozhodnout, která možnost je pro tři zkoumané populace ta správná, museli bychom provést návazné testy, které již nejsou v Excelu zakomponovány. Pouze zde uveďme, že jde o takzvaná mnohonásobná porovnání (nejčastěji bývají aplikovány metody Bonferroniho, Tukeyova či Scheffého, podrobněji viz např. Anděl, Statistické metody). Pro zajímavost - aplikací Tukeyho metody v SW *STATISTICA* bylo pro data z příkladu 22 zjištěno, že odpovídající situací je (na 5% hladině významnosti) situace d), tedy že populace 1 se významně liší od populace 2, ale rozdíl ani mezi populací 1 a 3, ani mezi populací 2 a 3 nelze prohlásit za statisticky významný. Jinou možností je aplikace regresního přístupu (viz kapitola o regresi s využitím tzv. dummy proměnných).

Statistická versus klinická významnost

Pomocí testů představených v předešlých kapitolách byly zjišťovány statisticky významné rozdílnosti (zamítnutí H_0 zde vlastně znamenalo prokázání přítomnosti statisticky významného rozdílu). Je na místě alespoň stručně upozornit čtenáře, že zjištění statisticky významného rozdílu ještě nemusí nutně znamenat prokázání rozdílu významného z hlediska klinické praxe. Jak je uvedeno v knize Zvárová: Základy statistiky…: „Např. při porovnávání krevního tlaku na levé a pravé ruce byl zjištěn průměrný rozdíl 1 mm Hg. Tento rozdíl je vysoce statisticky významný…, ale není důležitý klinicky“. (Podrobněji viz zmíněná či jiná odborná publikace.)

Obrázek 9: Bodový graf s vloženou regresní přímkou (závislost standardizované incidence novotvarů na hodnotách hrubé incidence, zdroj: ÚZIS – data pro kraje ČR za rok 2008, viz Příloha



Metody statistické indukce – regresní modely

Jednoduchá regrese a korelace

Regresní a korelační metody slouží při zkoumání závislosti mezi dvěma či více číselnými veličinami. Nejde „pouze" o testy, jako v předešlých kapitolách, i když v rámci těchto metod se testy také využívají; základním úkolem regresní analýzy je najít model této závislosti, návazně úkolem korelační analýzy je změřit výstižnost (kvalitu) nalezeného regresního modelu. Výsledkem samotné regresní analýzy je tedy rovnice (vzorec modelu), zatímco výsledkem korelace je číselný údaj (míra výstižnosti), týkající se příslušného modelu.

Zkoumáme-li závislost mezi dvěma číselnými veličinami, nazýváme tento úkol jednoduchá regrese (simple regression). V takovém případě bývá zvykem závislou veličinu (tu řízenou, určovanou modelem) značit jako Y a nezávislou veličinu (tu řídící, určující, též: regresor) značit jako X. To, která veličina je závislá a která je regresorem, je zřejmé z kontextu řešeného problému. Můžeme hledat např. model závislosti porodní hmotnosti (Y) na porodní délce (X) u novorozenců; model závislosti hodnoty krevního tlaku (Y) na množství jisté podané látky (X), apod. Vhodným a užitečným grafickým znázorněním dat je tzv. bodový graf (scatter plot) jako na obrázku 9. Každé statistické jednotce přísluší dvojice konkrétních číselných hodnot (x;y), které určují body (na obrázku 9 čtverečky) s příslušnou x-ovou souřadnicí na vodorovné ose, resp. y-ovou souřadnicí na svislé ose. Na obrázku 9 byly statistickými jednotkami jednotlivé kraje ČR, na x-ové ose je pro ně znázorněna vždy hodnota hrubé incidence (skutečné nové počty) novotvarů na 100.000 obyvatel, na y-ové ose je vynesena vždy odpovídající hodnota incidence přepočtené vůči standardní evropské populaci.

Obecně lze hledaný model jednoduché regresní závislosti označit jako

$$Y=f(X) \qquad (48)$$

kde f značí odpovídající matematickou funkci. Obvykle již pohled na bodový graf naznačí, jaký funkční model by byl pro daná data nejvhodnější (může se však samozřejmě i stát, že data na grafu jsou natolik nahodile rozházená, že na první pohled žádnou funkční závislost odhalit nelze). Nejjednodušším modelem, kdy je vhodné proložit daty přímku (jako např. na obrázku 9), je model jednoduché lineární regrese. Obecně odpovídá tomuto modelu rovnice

$$Y=\beta_0+\beta_1 \cdot X \qquad (49)$$

kde β_0 a β_1 jsou neznámé parametry regresní přímky, konkr. β_0 určuje hodnotu, v níž graf přímky protíná svislou osu (musí jít o bod na přímce se souřadnicemi $(0;\beta_0)$) a β_1 je tzv. směrnice (též: jednotkový přírůstek), určující sklon regresní přímky; u rostoucí přímky musí být β_1 kladné, u klesající záporné. Z obrázku 9 není hodnota β_0 přímo patrná, protože vodorovná osa není znázorněna nad x-ovou souřadnicí rovnou nule, nýbrž byla posunuta do bodu x=480. Hodnotu směrnice β_1 lze přibližně odhadnout následující úvahou: vidíme, že např. pro bod na přímce s hodnotou x=680 činí jeho y-ová souřadnice přibližně 540. Znamená to, že podle modelu by pro kraj, v němž na 100.000 obyvatel připadá 680 nových novotvarů, měla mít standardizovaná incidence novotvarů hodnotu cca 540

nových případů na 100.000 obyvatel. Dále pro bod na přímce s hodnotou x=730 je jeho y-ová souřadnice přibližně 580. Pokud tedy v modelu x-ová souřadnice vzrostla o 50 jednotek (ze 680 na 730), odpovídal tomu nárůst y-ové souřadnice o zhruba 40 jednotek (z 540 na 580). Směrnici určíme tak, že zjistíme, jak velká je y-ová změna při změně x-ové souřadnice o 1 jednotku (proto název „jednotkový přírůstek“), což zde znamená, že stačí vydělit 40/50=0,80; směrnice přímky na obrázku 9 by tedy měla mít přibližně hodnotu 0,80, což lze interpretovat tak, že pokud dojde k nárůstu hrubé incidence o 1 (případ na 100.000 obyvatel), reaguje model zvýšením standardizované incidence o 0,8.

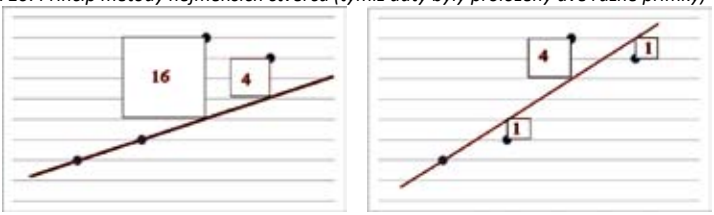
Zásadní problém v praxi ale je to, že nevíme, jaké jsou hodnoty parametrů oné „ideální“ regresní přímky. Tyto parametry na základě dat pouze odhadujeme. Konkrétně odhad pro parametr β_0 značíme b_0 a odhad pro parametr β_1 značíme b_1. Přímka na obrázku 9 byla také „pouze“ odhadem na základě zjištěných dat, takže správně bychom její rovnici měli zapsat ne ve tvaru (49), ale ve tvaru

$$Y=b_0+b_1 \cdot X \qquad (50)$$

Hodnoty b_0 a b_1 stanovíme ze všech zaznamenaných hodnot x a y tak, aby jim odpovídající regresní přímka co nejlépe vystihla celkový průběh dat. Pro data na obrázku 9 je tedy např. zřejmé, že hodnota směrnice b_1 musí být kladná a to přibližně 0,80; ale kolik má být přesně? A kolik činí hodnota b_0? Vzorce pro přesné vypočtení hodnot b_0 a b_1 jsou výsledkem tzv. metody nejmenších čtverců.

Příklad 23 (ilustrace principu metody nejmenších čtverců):
Princip této metody je ilustrován dvojicí grafů na obrázku 10, kde byly týmiž daty (pro jednoduchost byla jako ilustrace zvolena pouze čtveřice bodů) proloženy dvě různé, nahodile vybrané regresní přímky. Přímku na grafu vlevo budeme značit jako přímka I, přímku na grafu vpravo jako přímka II. Přímka I byla zvolena tak, aby přesně procházela dvěma body v levé části grafu (zbylé dva leží nad ní), přímka II byla zvolena tak, že přesně prochází pouze jedním ze čtyř bodů, zbylé dva leží pod ní a jeden nad ní. Úkolem je posoudit, která z těchto dvou přímek lépe vyhovuje daným datům, která je pro ně výstižnější. Kritériem této výstižnosti jsou velikosti tzv. reziduálních čtverců; za vhodnější prohlásíme ten z modelů, jehož celková plocha reziduálních čtverců bude menší (odtud název celé metody). Nejprve je ale nutno definovat pojem reziduum - na bodovém grafu jde o svislou (tedy nikoli kolmou) vzdálenost každého bodu od regresního modelu (zde od přímky), tedy vlastně o rozdíl y-ové souřadnice každého bodu (na grafech 10 každého „puntíku“) a jemu odpovídajícího svislého průmětu na regresní přímku. Např. na obrázku 10 vlevo mají dva body nulové reziduum (jde o dva body ležící přesně na přímce), jeden bod má hodnotu rezidua 2 (leží 2 měrné jednotky nad přímkou – na grafu odpovídá měrné jednotce vzdálenost „vodorovných linek“) a zbylý dokonce 4 (leží 4 měrné jednotky nad přímkou). Připomeňme si známý fakt, že plocha čtverce má velikost rovnou druhé mocnině délky jeho strany, takže na grafu 10 vlevo mají odpovídající reziduální čtverce plochy o velikostech popořadě 0, 0 (dva body vlastně žádné reziduální čtverce nevytvořily), 4 a 16, celkový součet těchto čtverců tedy činí 20. Analogicky na grafu 10 vpravo mají rezidua velikosti 0 (jde o bod ležící přesně na přímce), 1 (shodně pro oba body ležící pod přímkou), resp. 2 (pro bod ležící nad přímkou), takže odpovídající hodnoty „čtverců“ jsou 0, 1, 1 a 4 a jejich součet činí 6. Podle výše uvedeného kritéria je tedy vhodnějším modelem přímka II (na grafu 10 vpravo).

Obrázek 10: Princip metody nejmenších čtverců (týmiž daty byly proloženy dvě různé přímky)



Tabulka 14: Data pro jednoduchou lineární regresi v Excelu, zdroj: Zvárová, Základy statistiky…

	A	B	C
1	Nemocnice v USA	Spádová populace v tisících osob (X)	Celkem pracovních hodin (Y)
2	č. 1	25,5	304,37
3	č. 2	294,3	2616,32
4	č. 3	83,7	1139,12
5	č. 4	30,7	285,43
6	č. 5	129,8	1413,77
7	č. 6	180,8	1555,68
8	č. 7	43,4	383,78
9	č. 8	165,2	2174,27
10	č. 9	74,3	845,30
11	č. 10	60,8	1125,28
12	č. 11	319,2	3462,60
13	č. 12	376,2	3682,33

Ilustrační příklad 23 sice odpověděl na otázku, která z uvažovaných dvou regresních přímek byla pro daná data vhodnější, ne ale na otázku, která ze všech možných přímek by byla ta vůbec nejlepší – co kdybychom dokázali najít k uvažovaným datům jinou přímku, jejíž celková plocha reziduálních čtverců by byla ještě menší než 6? V praxi není samozřejmě možno postupovat tak, že bychom uvažovali všechny možné navzájem různé přímky (vždyť jich je nekonečně mnoho, na grafech 10 byly z nich vybrány pouze dvě), pro každou si sečetli velikosti ploch reziduálních čtverců a pak vybrali tu přímku s nejmenším součtem. Formálně postupujeme tak, že rovnou hledáme takovou přímku, aby byl odpovídající součet ploch reziduálních čtverců pro daná data minimální. Parametry odpovídající právě takové přímce značíme jako b_0 a b_1. Odvození příslušných vzorců (z matematického hlediska jde o úlohu hledání extrému) lze najít v odborné statistické literatuře, zde pouze uvedeme způsob, jak lze hodnoty těchto parametrů pro konkrétní data získat pomocí Excelu.

Příklad 24 (určení regresní přímky pomocí Excelu):
V tabulce 14 je pro 12 vybraných amerických nemocnic uvedena velikost jejich spádové populace (v tisících obyvatel) a celkový počet odpracovaných hodin. Úkolem je zjistit, jakým modelem by se mohla řídit závislost počtu odpracovaných hodin na velikosti spádové populace. Zřejmě lze očekávat, že u větších populací bude zapotřebí i více odpracovaných hodin, takže pokud je hledaná závislost lineárního typu, mělo by jít o model rostoucí přímky. Představu o tom, jaký model by mohl datům odpovídat, získáme z jejich bodového grafu. Ten sestrojíme v Excelu tak, že vybereme oba sloupce s daty (tedy sloupce B a C; Excel vždy předpokládá, že první sloupec obsahuje hodnoty pro vodorovnou osu a druhý pro svislou) a v nabídce Vložení - Grafy zvolíme typ Bodový (Bodový pouze se značkami). Značky budou na výsledném grafu skutečně uspořádány zhruba podél rostoucí přímky. Nyní je úkolem tuto přímku najít, což znamená určit hodnoty obou jejích parametrů b_0 a b_1. V Excelu jsou tyto hodnoty výsledkem příkazů (popořadě) v obecném tvaru

=INTERCEPT(dataY;dataX) =SLOPE(dataY;dataX) (51)

Pro data podle tabulky 14 je tedy nutno zadat příkazy (51) v konkrétním tvaru

=INTERCEPT(C2:C13;B2:B13) =SLOPE(C2:C13;B2:B13) (52)

Výsledné hodnoty činí (popořadě) 180,7, resp. 9,4. Znamená to, že hledaným lineárním modelem, který co nejlépe vystihuje zadaná data, je model

$$Y=180,7+9,4\cdot X \quad (53)$$

Dosazením za X-ové hodnoty zjistíme podle modelu očekávanou hodnotu Y-ovou. Např. při X=100 (tis. osob) lze očekávat tento potřebný počet odpracovaných hodin: Y=180,7+9,4·100=1120,7. Hodnotu směrnice 9,4 lze samostatně interpretovat takto: zvýšení velikosti spádové oblasti o 1 jednotku (o 1 tisíc osob) znamená v průměru zvýšení potřebného počtu odpracovaných hodin o 9,4.

Nalezenou regresní přímku je v Excelu možno jednoduše vložit přímo do bodového grafu podobně jako např. na obrázku 9. Máme-li v již hotovém bodovém grafu kurzor přesně na kterémkoli vyneseném bodu (značce), po kliknutí pravým tlačítkem myši se objeví menu s možností Přidat spojnici trendu. Po otevření této nabídky zvolíme jako Možnosti spojnice trendu typ Lineární. Pokud pak v dolní části nabídky zaškrtneme možnost Zobrazit rovnici regrese, objeví se po potvrzení v grafu nejen regresní přímka, ale i jí odpovídající rovnice (53), pouze s přehozeným pořadím sčítanců a případně jiným zaokrouhlením.

Již víme, jak najít nejvhodnější přímku, ale stále nevíme, jak kvalitním modelem tato přímka vlastně je, neboli jaká je korelace (korelovanost) mezi zkoumanými veličinami. Mírou kvality pro lineární modely je korelační koeficient (pro nelineární modely je nutno použít jiné míry). V Excelu lze jeho hodnotu získat příkazem v obecném tvaru

=CORREL(dataY;dataX) (54)

Korelační koeficient vždy nabývá hodnot z rozmezí -1 až 1. Kladné hodnoty odpovídají modelům rostoucích přímek (tedy tzv. přímé lineární závislosti – s rostoucím X také Y vzrůstá), zatímco záporné hodnoty odpovídají modelům klesajících přímek (tedy tzv. nepřímé lineární závislosti – s rostoucím X Y naopak klesá). Důležitější než samotné znaménko je ale vzdálenost hodnoty korelačního koeficientu od nuly: čím blíž k nule (ať je hodnota kladná nebo záporná), tím slabší je lineární závislost (tím méně výstižným modelem je nalezená přímka), hodnota nula odpovídá dokonalé (lineární) nezávislosti; a naopak čím blíž k plus nebo mínus jedné, tím silnější je lineární závislost (tím více výstižným modelem je nalezená přímka); hraniční hodnota plus jedna pak odpovídá dokonalé přímé lineární závislosti (všechny body by musely ležet přesně na nalezené rostoucí regresní přímce), hraniční hodnota mínus jedna odpovídá dokonalé nepřímé lineární závislosti (všechny body by musely ležet přesně na nalezené, ale klesající regresní přímce).

Další možnou charakteristikou kvality modelu je index determinace. Jeho hlavní předností je jeho snadná interpretace, nabývá totiž vždy hodnot mezi 0 až 1 a převádí se vynásobením sty na procenta. Lze potom (byť poněkud zjednodušeně) říci, z kolika procent daný model vysvětluje analyzovanou závislost mezi veličinami. Jeho další výhodou je jeho univerzálnost – je definován pro jakýkoliv (tedy nejen lineární) model. Speciálně u lineárního modelu platí, že jeho hodnota je rovna druhé mocnině korelačního koeficientu.

Příklad 25 (dokončení příkladu 24 - určení kvality nalezené regresní přímky pomocí Excelu):
Hodnotu korelačního koeficientu, odpovídajícího nalezené regresní přímce (53), získáme příkazem

=CORREL(C2:C13;B2:B13) (55)

Výsledkem je hodnota 0,97 (hodnota kladná a to velmi blízko hraniční hodnotě +1). Znamená to, že příslušná regresní přímka je rostoucí (což už jsme koneckonců viděli z toho, že kladná vyšla i hodnota směrnice 9,4) a že data vykazují velmi silnou lineární závislost. Kdybychom chtěli míru této závislosti vyjádřit v procentech, spočteme hodnotu indexu determinace $0{,}97^2=0{,}95$ (výpočet byl proveden s nezaokrouhlenou hodnotou korelačního koeficientu), což lze interpretovat tak, že model (53) vysvětluje závislost počtu pracovních hodin na velikosti spádové populace z 95 %.

Součástí regresní a korelační analýzy jsou i testy statistických hypotéz, o nich bude pojednáno alespoň stručně v rámci následující kapitoly.

Vícenásobná regrese a korelace

Zkoumáme-li závislost regresoru na dvou či více číselných veličinách, nazýváme tento úkol vícenásobná regrese (multiple regression). Opět obvykle závislou veličinu (tu řízenou, určovanou modelem) značíme jako Y, nezávislé veličiny (ty řídící, určující, též: regresory) značíme např. jako X a Z, nebo, je-li jich více, jako X_1, X_2, X_3, atd. Můžeme hledat např. pro novorozence model závislosti porodní hmotnosti (Y) na jejich porodní délce (X) a na délce trvání těhotenství (Z); u pacientů model závislosti hodnoty jejich krevního tlaku (Y) na podaném množství jisté účinné látky (X_1), na jejich věku (X_2) a hmotnosti (X_3), apod. Problém se ale někdy může komplikovat, protože vzájemné závislosti mohou fungovat i mezi jednotlivými regresory (třeba u příkladu s krevním tlakem může být podané množství účinné látky X_1 určováno v závislosti na hmotnosti pacienta X_3, apod.).

Vzhledem k více-dimenzionalitě problému již nelze využít ke grafickému znázornění jednoduchý bodový graf. Obecně lze hledaný model vícenásobné regresní závislosti označit jako

$$Y=g(X_1, X_2,\dots) \tag{56}$$

kde g značí odpovídající matematickou funkci. Nejjednodušším modelem je model vícenásobné lineární regrese, jemuž obecně odpovídá rovnice

$$Y=\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\dots \tag{57}$$

Jde vlastně o zobecnění modelu (49) pro více než jen jeden regresor. Speciálně při dvou regresorech by byl model (57) modelem roviny ve 3D prostoru (toto by ještě bylo poměrně snadno graficky znázornitelné i pomocí Excelu), u více než dvou regresorů se už jedná o modely tzv. nadrovin ve více než trojrozměrných prostorech.

Na základě dat získáme odhady regresních parametrů β_0, β_1, β_2 atd., které značíme b_0, b_1, b_2 atd. K jejich zjištění Excelem je nutno využít připravený analytický nástroj Regrese (ten by byl použitelný už i v případě regrese jednoduché jako alternativa k postupu uvedenému v příkladech 24 a 25). Použití si ilustrujeme na následujícím příkladu.

Příklad 26 (ukázka použití analytického nástroje Regrese):
Na základě dat z tabulky 1 hledáme model závislosti hmotnosti dítěte na jeho věku a výšce. Předpokládejme, že data jsou zadána v Excelu přesně podle tabulky 1 a že jsou aktivovány analytické

Tabulka 15: Výstup analytického nástroje Regrese pro data podle tabulky 1

Regresní statistika	
Násobné R	0,7498
Hodnota spolehlivosti R	0,5595
Nastavená hodnota spolehl.	0,9595
Chyba stř. hodnoty	7,0291
Pozorování	13

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	876,9979	438,4990	8,8751	0,0051
Rezidua	10	494,0790	49,4079		
Celkem	12	1371,0769			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	-117,6460	42,0732	-2,7962	0,0189	-211,3019	-23,9919
Věk (X1,l)	3,4078	3,2740	1,0408	0,3094	-3,8684	10,6838
Výška (X2,l)	1,3623	0,4305	3,1642	0,0101	0,4030	2,3216

nástroje (aktivace byla popsána na úvod příkladu 14). V nabídce Data – Analýza dat vybereme metodu Regrese (OK). Jako Vstupní oblast Y zadáme buňky D2:D15 a jako vstupní oblast X zadáme buňky B2:C15. Jelikož takto zadaná data obsahují i řádek s názvy (popisky) jednotlivých veličin, zaškrtneme možnost Popisky. Pokud jako Možnosti výstupu zvolíme Nový list, po potvrzení (OK) se otevře nový list (viz tabulka 15), obsahující souhrnné výsledky regresní a korelační analýzy (bohužel v české verzi s množstvím nevhodně či přímo chybně přeložených termínů). Důležité jsou zejména:

- „Hodnota spolehlivosti R“ (zde 0,6396) - jde o hodnotu indexu determinace;
- „Významnost F“ (zde 0,0061) - jde o p-hodnotu tzv. F-testu (jak i naznačuje název této části výstupů, jedná se o analogii testu ANOVA), podrobněji viz níže a pak příklad 27;
- Údaje ze sloupce „Koeficienty“ - konkr. u Hranice jde o hodnotu b_0 (zde -117,65), u Věk o hodnotu b_1 (zde -3,45) a u Výška o hodnotu b_2 (zde 1,36), takže odhad modelu (57) má zde konkrétní tvar

$$Y= -117{,}65–3{,}45\cdot X_1+1{,}36\cdot X_2 \qquad (58)$$

  kde proměnná Y značí hmotnost dítěte (kg), X_1 jeho věk (roky) a X_2 tělesnou výšku (cm);
- Údaje ze sloupce „Hodnota P“ (popořadě 0,0189, 0,3094 a 0,0101) odpovídají p-hodnotám tzv. t-testů, podrobněji viz níže a pak příklad 27.

Díky modelu (58) lze např. odhadnout, že u 10-letých dětí o tělesné výšce 1,5 metru lze v průměru očekávat hmotnost $-117{,}65–3{,}45\cdot 10+1{,}36\cdot 150=51{,}85$, tedy necelých 52 kg. Díky indexu determinace lze konstatovat, že vytvořený model (58) vysvětluje závislost hmotnosti dětí na jejich věku a výšce z téměř 64 % (63,96 %).

Součástí výstupů analytického nástroje regrese, komentovaných v příkladu 26, byly i p-hodnoty testů dvou typů, s nimiž jsme se již setkali: jde o F-test a t-testy. I v případě regrese musí být splněny předpoklady normality k tomu, abychom byli oprávněni tyto testy vůbec použít. Konkrétně v případě regrese se požadavek normality týká chování reziduí (podrobněji viz např. Anděl: Statistické metody), což je bohužel opět předpoklad, který se nedá pomocí Excelu jednoduše ověřit.

F-testem v regresi ověřujeme významnost regresního modelu jako celku. Nulová hypotéza předpokládá nulovost všech „násobících“ regresních parametrů:

$$H_0: 0=\beta_1=\beta_2=\ldots \qquad (59)$$

Jinak řečeno, za platnosti nulové hypotézy (59) by se model (57) zredukoval na tvar $Y=\beta_0$, takže ani jedna z uvažovaných vysvětlujících veličin X_1, X_2, atd. by do modelu nepatřila, ani jedna z nich by nebyla významným regresorem a model by tak neprokázal žádnou významnou závislost. Alternativní hypotéza naopak předpokládá, že alespoň jeden z „násobících“ parametrů je nenulový, tudíž že závislost byla prokázána alespoň na jednom ze všech uvažovaných regresorů.

Pomocí t-testů v regresi ověřujeme významnost jednotlivých regresních parametrů (včetně parametru β_0). Nulová hypotéza zvlášť pro každý parametr předpokládá jeho nulovost, alternativní hypotéza naopak jeho nenulovost:

$$H_0: \beta_i=0 \qquad H_a: \beta_i\neq 0 \qquad (60)$$

kde i=0,1,2,… Podle výsledků t-testu lze konstatovat, zda model (57) bude či nebude obsahovat příslušný parametr (a případně jemu odpovídající veličinu jakožto potenciální regresor).

Příklad 27 (dokončení příkladu 26 – interpretace výsledných p-hodnot):
Na základě dat v tabulce 1 byl sestrojen model závislosti hmotnosti (Y) dítěte na jeho věku (X_1) a výšce (X_2): $Y=-117{,}65–3{,}45\cdot X_1+1{,}36\cdot X_2$. Celkovému F-testu odpovídá p-hodnota 0,0061<0,05, takže na 5% hladině významnosti zamítáme odpovídající nulovou hypotézu o nevýznamnosti modelu. Pomocí t-testů je nyní možno zjistit, který (které) z parametrů je (jsou) v modelu významný (významné):

- t-test pro H_0: $\beta_0=0$ versus H_a: $\beta_0\neq 0$: p-hodnota=0,0189<0,05, takže na 5% hladině významnosti zamítáme odpovídající H_0; jinak řečeno, parametr β_0 je v modelu významně nenulový;

- t-test pro H_0: $\beta_1=0$ versus H_a: $\beta_1\neq0$: p-hodnota=0,3094>0,05, takže na 5% hladině významnosti nelze zamítnout odpovídající H_0; jinak řečeno, parametr β_1 je možno považovat za nulový, takže odpovídající proměnnou X_1 (věk) nelze označit za významný regresor;
- t-test pro H_0: $\beta_2=0$ versus H_a: $\beta_2\neq0$: p-hodnota=0,0101<0,05, takže na 5% hladině významnosti zamítáme odpovídající H_0; jinak řečeno, parametr β_2 není možno považovat za nulový, takže odpovídající proměnnou X_2 (výška) lze označit za významný regresor.

Díky výsledkům testů můžeme nyní prohlásit, že model (58) je statisticky významným regresním modelem pro vysvětlení závislosti hmotnosti dětí, ovšem z obou v něm zahrnutých vysvětlujících veličin (věk a výška) lze pouze jednu (výška) označit za významný regresor.

Na základě diskuse na konci příkladu 27 se vnucuje myšlenka – co kdybychom nevýznamnou proměnnou věk z modelu zcela vypustili? Problém by se tím pádem změnil na problém jednoduché regrese, neboť bychom u dětí hledali závislost jejich hmotnosti na výšce. Je ale nutno předem upozornit, že

- nelze očekávat, že výsledný model bude mít tvar (58), „pouze s vynecháním“ proměnné X_1;
- dosavadními výsledky není vůbec zaručeno, že tento nový model bude statisticky významný.

Zkrátka – vytváříme-li nový model (byť jde o „sub-model“ modelu již odhadnutého), musíme opět provést celou analýzu. Tento zdlouhavý proces je možno formálně zkrátit aplikací tzv. krokové regrese (stepwise regression), která ale není v Excelu implementována. Pouze pro informaci tedy alespoň uvedeme, že rozlišujeme dva typy krokové regrese:

- typ forward, kdy začneme s co nejjednodušším modelem a do něj postupně přidáváme vždy jeden regresor tak, aby model jako celek zůstal významný;
- typ backward, kdy naopak začneme s modelem obsahujícím všechny potenciální regresory, z nichž postupně ubíráme vždy jeden („nejhorší“) tak, aby model jako celek zůstal významný.

Metody krokové regrese jsou standardní součástí specializovaných statistických SW.

Vraťme se ještě k modelu (58) z příkladu 26. Odhad koeficientu b_1 (u veličiny věk) má zápornou hodnotu -3,45, což bychom mohli interpretovat tak, že u dětí o stejné tělesné výšce lze s nárůstem jejich věku o 1 rok očekávat v průměru pokles hmotnosti o 3,45 kg, jinak řečeno, čím starší děti, tím nižší bude jejich hmotnost. Tento zdánlivý paradox může mít z matematického hlediska různá vysvětlení (např. tzv. multikolinearitu či jiné problémy, viz např. Zvárová: Statistické metody v epidemiologii), jejichž společným důvodem může být závislost buď navzájem mezi potenciálními regresory (zde mezi věkem a výškou dětí), nebo nějaká specifická funkční závislost mezi zkoumaným Y a jednotlivými potenciálními regresory.

Regrese s využitím dummy proměnných

Regresní modely je možno konstruovat např. i pro nečíselné regresory. V případě jednoho kategoriálního regresoru půjde vlastně o jakousi vylepšenou jedno-faktorovou ANOVu. Připomeňme, že v rámci příkladu 22 byl metodou ANOVA na základě dat z tabulky 12 zjištěn statisticky významný rozdíl ve střední koncentraci soli kyseliny listové mezi třemi srovnávanými pod-populacemi pacientů. Samotný výstup metody ANOVA (tak, jak jej poskytuje Excel, viz tabulka 13) však již neumožnil rozhodnout konkrétně, které skupiny se navzájem významně liší. To nyní umožní poměrně snadno právě regresní přístup. Představme si, že chceme modelovat závislost veličiny „kyselina“ (v tabulce 12 šlo o číselné hodnoty ze sloupce A) na veličině „skupina“ (tamtéž ve sloupci B). Veličina „skupina“ sice nabývala hodnot 1, 2 nebo 3, nejde však o číselnou veličinu, nýbrž o veličinu kategoriální nominálního typu, kde hodnoty 1-3 představují pouze kódy rozlišující jednotlivé srovnávané kategorie (kategorie mohly být stejně tak rozlišovány třeba pomocí kódů A, B a C). Nemůžeme proto konstruovat přímo model jednoduché regrese obecného typu (48), protože nemáme k dispozici číselný regresor X (do nalezeného modelu bychom pak těžko dosazovali za X třeba hodnotu „B“). Každou nečíselnou kategoriální veličinu však můžeme převést do podoby několika regresorů alternativního typu (tedy

s možnými hodnotami 0-1), a to zavedením tzv. dummy proměnných (český termín není vžitý, mohli bychom hovořit o jakýchsi „pomocných“ či „fiktivních“ proměnných). Výhodou alternativních veličin (byť byly v úvodu této knihy formálně zařazeny mezi veličiny nečíselné) je to, že pokud je kódujeme hodnotami 0 a 1, lze je považovat za číselné veličiny diskrétního typu, označující „počet výskytů sledovaného jevu při jediném pokusu“. Oním „sledovaným jevem“ bude pro každou vytvářenou dummy proměnnou „náležení“ do vybrané kategorie. Obecně platí, že pokud měla sledovaná kategoriální veličina *K* různých kategorií, musíme definovat *K*-1 dummy proměnných.

Příklad 28 (zavedení dummy proměnných v datech z příkladu 22):
Předvedeme si zavedení dummy proměnných místo kategoriální veličiny „skupina“ z tabulky 12. Počet jejích kategorií *K*=3, takže musíme zavést dvě (*K*-1=2) dummy proměnné. Nechť první dummy proměnná bude značit náležení do kategorie 2 (označme ji SKUP2) takto: SKUP2=0, pokud dotyčný pacient nepatřil do kategorie 2 a SKUP2=1, pokud dotyčný do kategorie 2 patřil. Podobně nechť druhá dummy proměnná bude značit náležení do kategorie 3 (označme ji SKUP3): SKUP3=0, pokud dotyčný nepatřil do kategorie 3 a SKUP3=1, pokud dotyčný do kategorie 3 patřil. Tudíž pro každého

- z původní kategorie 1 (v datech z tabulky 12 všichni pacienti s údaji z řádků 2 až 9 včetně) budou mít jemu zavedené dummy proměnné tyto hodnoty: SKUP2=0, SKUP3=0;
- z původní kategorie 2 (v datech z tabulky 12 všichni pacienti s údaji z řádků 10 až 18 včetně) budou mít jemu zavedené dummy proměnné tyto hodnoty: SKUP2=1, SKUP3=0;
- z původní kategorie 3 (v datech z tabulky 12 všichni pacienti s údaji z řádků 19 až 23 včetně) budou mít jemu zavedené dummy proměnné tyto hodnoty: SKUP2=0, SKUP3=1.

Tabulka 16: Data z tabulky 12 se zavedenými dummy proměnnými

KYSELINA	SKUPINA	SKUP2	SKUP3
276	1	0	0
280	1	0	0
275	1	0	0
291	1	0	0
347	1	0	0
354	1	0	0
380	1	0	0
330	1	0	0
206	2	1	0
210	2	1	0
226	2	1	0
249	2	1	0
255	2	1	0
273	2	1	0
285	2	1	0
295	2	1	0
309	2	1	0
241	3	0	1
246	3	0	1
270	3	0	1
293	3	0	1
328	3	0	1

Vyplnit v Excelu hodnoty do sloupců SKUP2 a SKUP3 je možno ručně, nebo lépe automaticky s využitím příkazu =KDYŽ() analogicky jako v příkladu 2. Konkrétně hodnoty pro dummy proměnnou SKUP2 bychom pro prvního pacienta mohli definovat příkazem

=KDYŽ(B2=2;1;0)

a tento příkaz bychom překopírovali do celého sloupce SKUP2. Podobně hodnoty pro dummy proměnnou SKUP3 bychom pro prvního pacienta mohli definovat příkazem

=KDYŽ(B2=3;1;0)

a tento příkaz bychom překopírovali do celého sloupce SKUP3. Nová podoba dat (již připravených k aplikaci regresního přístupu) je v tabulce 16.

□

Nyní je již možno aplikovat model vícenásobné lineární regrese (57) analogicky jako v příkladu 26, přičemž nyní dvěma regresory budou právě zavedené dummy proměnné SKUP2 a SKUP3. Podstatná část výstupu je uvedena v tabulce 17. Z hodnoty „Významnost F“ (jde o p-hodnotu celkového F-testu) vidíme, že model s oběma uvažovanými regresory SKUP2 a SKUP3 je statisticky významný (0,0148<0,05). Zmiňme shodu s výsledkem testu ANOVA (viz tabulka 13) – jím zjištěná p-hodnota je (až na zaokrouhlení) stejná.

Tabulka 17: Výstup analytického nástroje Regrese pro data z tabulky 16 s dummy proměnnými

ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	2	15660,4755	7830,7378	5,3005	0,0148
Reziduá	19	28073,2972	1477,5420		
Celkem	21	43733,7727			

	Koeficienty	Chybová sl.	t Stat	Hodnota P	Dolní	Ho
Hranice	316,6250	13,5902	23,2981	0,0000	288,1804	345
SKUP2	-60,1805	18,6779	-3,2220	0,0045	-99,2730	-21
SKUP3	-41,0250	21,9135	-1,8721	0,0767	-86,8900	4

Stejně jako na základě testu ANOVA lze tedy konstatovat, že mezi třemi srovnávanými pod-populacemi pacientů byl zjištěn významný rozdíl ve střední koncentraci soli kyseliny listové v krvi. Navíc oproti testu ANOVA máme nyní k dispozici následující regresní model (viz zaokrouhlené hodnoty ze sloupce „Koeficient“ v tabulce 17):

$$Y=317-60\cdot SKUP2-41\cdot SKUP3 \qquad (61)$$

Pokud do tohoto modelu dosadíme příslušnou „kombinaci“ nul a jedniček, dostaneme odhady středních koncentrací v jednotlivých kategoriích, konkr.:

- kategorie 1 (dosadíme-li SKUP2=0, SKUP3=0): Y=317–60·0–41·0=317;
- kategorie 2 (dosadíme-li SKUP2=1, SKUP3=0): Y=317–60·1–41·0=317–60=257;
- kategorie 3 (dosadíme-li SKUP2=0, SKUP3=1): Y=317–60·0–41·1=317–41=276.

Kategorii, reprezentovanou dosazením samých nul, považujeme v modelu vždy za tzv. bazickou. V případě modelu (61) byla tedy bazickou kategorií kategorie 1, pro niž odhad střední koncentrace činí 317 (příslušných měrných jednotek). Pro kategorii 2 je pak odhad střední koncentrace oproti bazické hodnotě o 60 jednotek nižší, činí tedy 257. Podobně vidíme, že pro kategorii 3 je odhad střední koncentrace oproti bazické hodnotě o 41 jednotek nižší, činí tedy 276. I tyto tři odhady středních hodnot byly k dispozici již ve výstupu testu ANOVA: na obrázku 13 je najdeme (s přesností na 3 desetinná místa) ve sloupci „Průměr“, šlo o „obyčejné“ aritmetické průměry koncentrací v jednotlivých kategoriích, neboli o bodové odhady středních hodnot. Hlavní výhodou regresního modelu je však to, že nyní máme k dispozici p-hodnoty pro posouzení významnosti jednotlivých dummy proměnných (viz sloupec „Hodnota P“ v tabulce 17): zatímco proměnná SKUP2 je v modelu (61) statisticky významným regresorem (0,0045<0,05), proměnnou SKUP3 nelze na 5% hladině významnosti prohlásit za statisticky významný regresor (0,0767>0,05). Zjistili jsme tudíž na základě dat, že za statisticky významný rozdíl lze prohlásit rozdíl mezi bazickou kategorií (kategorií 1) a kategorií, reprezentovanou dummy proměnnou SKUP2 (tedy kategorií 2). Tento výsledek je v souladu s komentářem ze závěru kapitoly o jedno-faktorové analýze rozptylu, jenž tam byl učiněn na základě tzv. Tukeyho metody. Poznamenejme závěrem, že velmi podrobně je o regresních modelech, a to i v souvislosti s problematikou ANOVA, pojednáno např. v Neter-Wasserman-Kutner: Applied Linear Statistical Models.

Metody statistické indukce – testy typu chí-kvadrát

Test dobré shody

Teoretickým modelem kategoriální veličiny je pravděpodobnostní zastoupení jejích jednotlivých kategorií. Nechť u dané veličiny rozlišujeme K kategorií, přičemž pravděpodobnost výskytu první kategorie označíme jako π_1, pravděpodobnost výskytu druhé kategorie jako π_2, atd., až pravděpodobnost výskytu poslední kategorie označíme jako π_K. Pro jednotlivé pravděpodobnosti musí platit, že jejich součet přes všechny kategorie musí být roven jedné (100 %):

$$\pi_1+\pi_2+\ldots+\pi_K = 1 \quad (62)$$

V praxi máme k dispozici náhodný výběr, na jehož základě můžeme určit pozorované absolutní četnosti jednotlivých kategorií, označme je popořadě jako n_1, n_2, atd., až n_K. Z kapitoly o četnostech již víme, že (označíme-li rozsah výběru jako n) musí platit:

$$n_1+n_2+\ldots+n_K = n \quad (63)$$

Pomocí těchto absolutních četností lze nyní ověřit shodu dat s konkrétním teoretickým pravděpodobnostním modelem (odtud název „test dobré shody“, angl. goodness of fit). Nulová hypotéza předpokládá shodu, alternativní naopak statisticky významnou odlišnost mezi daty a modelem. Formálně provádíme test porovnáním pozorovaných absolutních četností (63) s tzv. očekávanými absolutními četnostmi (značenými popořadě jako $o_1,\ldots,o_K$), které pro jednotlivé kategorie určujeme takto:

$$o_1=n\cdot\pi_1,\ o_2=n\cdot\pi_2,\ \ldots\ ,\ o_K=n\cdot\pi_K \quad (64)$$

Očekávané četnosti (64) tedy odpovídají „ideálním“ absolutním četnostem, tedy případu, kdy by se všechny statistické jednotky přerozdělily do jednotlivých kategorií zcela přesně podle poměru určeného jednotlivými pravděpodobnostmi. Poznamenejme, že v praxi se často stává, že očekávané četnosti nevycházejí celočíselně, což není žádná chyba. Aby test správně fungoval, musí mít ale všechny vypočtené očekávané četnosti hodnotu alespoň 5; pokud se v řešeném problému vyskytne kategorie s menším očekávaným zastoupením, přistupuje se někdy v praxi ke sloučení této kategorie s nějakou jinou. Samotné porovnání pozorovaných četností (63) a očekávaných četností (64) si nyní předvedeme prakticky s využitím Excelu.

Příklad 29 (test chí-kvadrát dobré shody pomocí Excelu):
Předvedeme si snadné provedení testu dobré shody v Excelu na následující úloze převzaté z knihy Zvárová: Základy statistiky…: Vylučovatelství substancí ABH určuje dominantní alela Se. Heterozygotní rodiče-vylučovatelé, tj. rodiče s kombinací alel (Se,se), mohou mít potomka typu:

- kategorie 1: nevylučovatel, tj. potomek má kombinaci alel (se, se);
- kategorie 2: heterozygotní vylučovatel, tj. potomek má kombinaci alel (Se, se);
- kategorie 3: homozygotní vylučovatel, tj. potomek má kombinaci alel (Se, Se).

Z pravděpodobnostního hlediska by uvedené tři kategorie měly být zastoupeny popořadě dle poměru

$$\pi_1=0{,}25\,(25\ \%) \qquad \pi_2=0{,}50\ (50\ \%) \qquad \pi_3=0{,}25\ (25\ \%) \quad (65)$$

což představuje teoretický pravděpodobnostní model. Při statistickém průzkumu byly v uvažovaných třech kategoriích potomků zjištěny tyto absolutní četnosti:

$$n_1=159 \qquad n_2=321 \qquad n_3=159 \quad (66)$$

čili celkem bylo vyšetřeno 159+321+159=639 potomků. Podle (64) budou mít tedy očekávané četnosti pro model (65) hodnoty:

$$o_1=639\cdot0{,}25=159{,}75 \quad o_2=639\cdot0{,}50=319{,}50 \quad o_3=639\cdot0{,}25=159{,}75 \quad (67)$$

Pouhým pohledem na skutečné hodnoty (66) a jim odpovídající hodnoty očekávané (67) je zřejmá značná podobnost, nicméně k tomu, abychom mohli model (65) prohlásit za prokazatelně platný, je nutno provést test dobré shody. Předpokládejme, že hodnoty pozorovaných četností (66) máme v Excelu v buňkách B1, B2 a B3, a že odpovídající hodnoty očekávaných četností (67) byly vypočteny v buňkách C1, C2 a C3; připomeňme, že násobení je v Excelu zadáváno pomocí symbolu *, takže např. hodnotu v buňce C1 bychom získali příkazem =639*0,25. Obecný příkaz na provedení testu typu chí-kvadrát má tvar

```
=CHITEST(pozor.četnosti;oček.četnosti)   (68)
```

takže konkrétní zadání teď musí mít podobu

```
=CHITEST(B1:B3;C1:C3)   (69)
```

Výsledkem příkazu je vždy p-hodnota, která v případě (69) činí přibližně 0,99, tedy více než 0,05, takže na 5% hladině významnosti nelze zamítnout odpovídající H_0 o „dobré shodě“ mezi daty a modelem, jinak řečeno, data potvrdila vhodnost modelu s procentuálním zastoupením 25-50-25 (%).

□

Název „chí-kvadrát“ používáme u těch testů, které ke svému provedení potřebují využívat kvantily tzv. rozdělení chí-kvadrát (značeno χ^2), při postupu podle příkladu 29 však tento princip zůstane čtenáři skrytý za výslednou p-hodnotou. Závěrem poznamenejme, že test dobré shody existuje i ve verzi pro spojité veličiny, nejde však již o metodu přichystanou v Excelu k přímému použití.

Test nezávislosti

Představme si, že u každého pacienta zaznamenáváme použitý typ léčby (nechť možnosti byly čtyři, rozlišujme je jako A-B-C-D) a míru úspěšnosti (pacient byl vyléčen zcela – částečně – vůbec ne). Obě uvažované veličiny, tedy jak typ léčby, tak její úspěšnost, jsou kategoriálního typu. To, co budeme chtít prokázat, je případná závislost mezi typem a úspěšností léčby. Tato závislost je zde míněna tak, že se jednotlivé typy léčby liší ve své úspěšnosti. Naopak nezávislost by znamenala, že mezi jednotlivými typy léčby nejsou významné rozdíly v jejich úspěšnosti.

Obecně předpokládejme, že sledujeme dvojici kategoriálních veličin. Důležitým předpokladem je náhodnost při výběru statistických jednotek, která je poměrně spolehlivou zárukou jejich reprezentativního rozřazení do jednotlivých kategorií. (V uvedeném motivačním příkladu by samozřejmě bylo zásadní chybou při zahrnutí např. zcela vyléčených pacientů do průzkumu preferovat jen určitý typ léčby.) Testem nezávislosti vybíráme z dvojice hypotéz

H_0: nezávislost obou veličin H_a: závislost mezi oběma veličinami (70)

(Zdůrazněme, že název „test nezávislosti“ odpovídá znění nulové hypotézy, nikoli tedy tomu, co si obvykle „přejeme prokázat“, což v praxi často bývá právě naopak závislost.)

Podobně jako u testu dobré shody bude rozhodnutí založeno na porovnání skutečně pozorovaných a teoreticky očekávaných absolutních četností. Vzhledem k tomu, že se zabýváme vzájemným vztahem dvojice kategoriálních veličin, musí mít ale záznam absolutních četností specifický tvar, který nazýváme kontingenční tabulka. Jde vlastně o „dvourozměrnou“ tabulku četností, kdy v jednom směru (řádcích) rozlišujeme kategorie jedné veličiny a ve druhém směru (sloupcích) kategorie veličiny druhé. Četnost zaznamenaná v tabulce pak odpovídá výskytu pro příslušnou kombinaci řádkové a sloupcové kategorie. Např. v ukázkové kontingenční tabulce 18 je počet pacientů, kteří byli léčeni léčbou typu B a u nichž došlo k úplnému vyléčení, označen jako četnost n_{12} (dolní indexy 1 a 2 značí, že jde o údaj v prvním řádku a ve druhém sloupci tabulky; vždy první index odpovídá pořadí řádku a druhý pořadí sloupce). Celkový počet pacientů, u nichž došlo k úplnému vyléčení (bez ohledu na to, jakému typu léčby byli podrobeni) najdeme v tabulce 18 jako součet na prvním řádku, označený jako $n_{1\bullet}$ (každý řádkový součet má místo druhého, tedy „sloupcového“ indexu, tečku). Analogicky celkový počet pacientů léčených léčbou B (bez ohledu na výsledek) najdeme v tabulce 18 jako součet ve druhém sloupci, označený jako $n_{\bullet 2}$ (každý sloupcový součet má místo prvního, tedy „řádkového“ indexu, tečku). Celkový součet (ať již přes řádky či sloupce kontingenční tabulky) musí být roven rozsahu výběru, neboli v tabulce 18 musí platit:

$$n_{1\bullet}+n_{2\bullet}+n_{3\bullet} = n_{\bullet 1}+n_{\bullet 2}+n_{\bullet 3}+n_{\bullet 4} = n \quad (71)$$

Také očekávané četnosti budeme zapisovat do tabulky stejného vzhledu. Očekávané četnosti musí odpovídat testované nulové hypotéze o nezávislosti, tedy z matematického hlediska tomu, že se pozorované celkové počty na každém řádku (tedy popořadě hodnoty $n_{1\bullet}$, $n_{2\bullet}$ atd.) „přerozdělí“ do jednotlivých sloupců ve zcela stejném poměru, přičemž tímto společným poměrem bude poměr daný pozorovanými celkovými počty v jednotlivých sloupcích (takže konkr. pro data z tabulky 18 v poměru $n_{\bullet 1}$: $n_{\bullet 2}$: $n_{\bullet 3}$:$n_{\bullet 4}$). Vzorce na výpočet očekávaných četností odpovídajících tabulce 18 jsou v tabulce 19. Označíme-li obecně očekávanou četnost na pozici i,j (tedy na i-tém řádku a j-tém sloupci) jako o_{ij}, platí pro její výpočet obecný vzorec

$$o_{ij} = n_{i\bullet}\cdot n_{\bullet j}/n \quad (72)$$

Vzorec (72) se dá slovně interpretovat tak, že k výpočtu očekávané četnosti pro jakékoli políčko je nutno vynásobit mezi sebou příslušný řádkový a sloupcový součet („příslušný“ znamená z toho řádku,

Tabulka 18: Ukázka kontingenční tabulky s obecným označením pozorovaných absolutních četností

Počty pacientů		léčených typem léčby				celkem
		A	B	C	D	
vyléčených	zcela	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1\bullet}$
	částečně	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2\bullet}$
	vůbec ne	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3\bullet}$
celkem		$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	$n_{\bullet 4}$	n

resp. sloupce, pro nějž očekávanou četnost určujeme) a výsledek vydělit celkovým počtem pozorování. Konkr. očekávaný počet pacientů, kteří byli léčeni léčbou typu B a u nichž došlo k úplnému vyléčení, tedy očekávaná četnost odpovídající pozorované četnosti n_{12}, by se podle vzorce (72) vypočetla takto:

$$o_{12}=n_{1\bullet}\cdot n_{\bullet 2}/n \qquad (73)$$

I zde jako v testu dobré shody požadujeme, aby měly všechny vypočtené očekávané četnosti hodnotu alespoň 5, a není-li tomu tak, přistupuje se někdy v praxi ke slučování kategorií (zde dvou různých řádků, resp. sloupců kontingenční tabulky). Ani zde nemusí vyjít očekávané četnosti celočíselně; jsou-li však všechny očekávané četnosti správně vypočteny, musí být nakonec jejich součty v řádcích, resp. sloupcích úplně stejné (příp. až na zaokrouhlení), jako byly součty v tabulce četností skutečně pozorovaných. Jinak řečeno, číselně musí mít tabulky 18 a 19 stejné hodnoty v „celkovém" řádku i sloupci, ale hodnoty v jednotlivých políčkách uvnitř obou tabulek se budou více či méně odlišovat. Podstatou testu je potom (stejně jako u testu dobré shody) výpočet jakési míry vzájemné podobnosti mezi pozorovanými a očekávanými četnostmi, načež jsou opět využity kvantily rozdělení chí-kvadrát; i tento test si však ukážeme bez těchto podrobností, pouze s využitím Excelu.

Tabulka 19: Vzorce na výpočet očekávaných četností odpovídajících tabulce 18

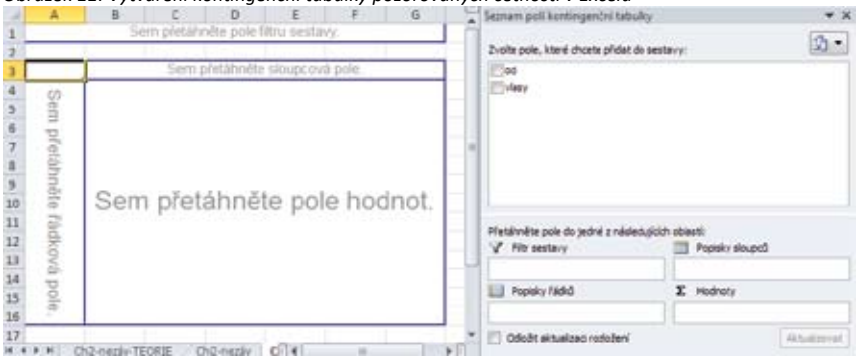
Počty pacientů		léčených typem léčby				celkem
		A	B	C	D	
vyléčených	zcela	$o_{11}=n_{1\bullet}\cdot n_{\bullet 1}/n$	$o_{12}=n_{1\bullet}\cdot n_{\bullet 2}/n$	$o_{13}=n_{1\bullet}\cdot n_{\bullet 3}/n$	$o_{14}=n_{1\bullet}\cdot n_{\bullet 4}/n$	$n_{1\bullet}$
	částečně	$o_{21}=n_{2\bullet}\cdot n_{\bullet 1}/n$	$o_{22}=n_{2\bullet}\cdot n_{\bullet 2}/n$	$o_{23}=n_{2\bullet}\cdot n_{\bullet 3}/n$	$o_{24}=n_{2\bullet}\cdot n_{\bullet 4}/n$	$n_{2\bullet}$
	vůbec ne	$o_{31}=n_{3\bullet}\cdot n_{\bullet 1}/n$	$o_{32}=n_{3\bullet}\cdot n_{\bullet 2}/n$	$o_{33}=n_{3\bullet}\cdot n_{\bullet 3}/n$	$o_{34}=n_{3\bullet}\cdot n_{\bullet 4}/n$	$n_{3\bullet}$
celkem		$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	$n_{\bullet 4}$	n

Příklad 30 (kontingenční tabulka a test chí-kvadrát nezávislosti pomocí Excelu):
Problematiku budeme ilustrovat úlohou převzatou z knihy Anděl: Statistické metody. Pro každého ze 6800 mužů byla zjišťována dvojice veličin „barva očí" (s použitými kódy: „1" kódující oči modré, „2" jako kód pro oči hnědé a „3" jako kód pro oči jiné barvy) a „barva vlasů" (s možnými hodnotami: „1" jako kód pro vlasy světlé, „2" pro kaštanové, „3" pro černé a „4" pro zrzavé). Zaznamenaná data by v Excelu tedy vypadala např. tak jako v tabulce 20. Konkr. muž č.1 měl oči hnědé barvy (kód 2) a černé vlasy (kód 3); muž č.2 byl modrooký a světlovlasý, atd. Poznamenejme, že Excel by uměl zpracovat i kódy nečíselné (pro barvu očí tedy např. slovní kódy „modrá", „hnědá", „jiná"; nevýhodou takovéhoto kódování by bylo to, že při konstrukci přehledných tabulek pak Excel automaticky seřadí

Tabulka 20: Část vstupních dat v úloze s testem nezávislosti pro dvojici kategoriálních veličin

	A	B	C
1	muž č.	oči	vlasy
2	1	2	3
3	2	1	1
4	3	3	4

Obrázek 11: Vytváření kontingenční tabulky pozorovaných četností v Excelu

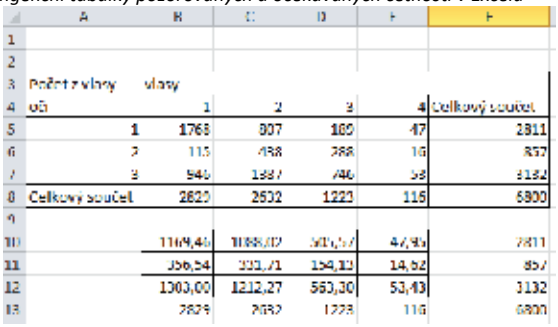


nalezené kategorie abecedně). Celkem by tato data v Excelu obsahovala 6800 řádků, pro každého muže jeden samostatný řádek; a jelikož první řádek v tabulce 20 obsahuje názvy sledovaných veličin, byl by zde posledním řádkem řádek s pořadovým číslem 6801. Pro další jednoduché zpracování je vhodné, aby obě analyzované kategoriální veličiny byly v sousedních sloupcích (jako v tabulce 20 ve sloupcích B a C).

Prvním úkolem při praktickém zpracování vlastních dat bývá sestrojit kontingenční tabulku skutečných, pozorovaných četností. Pokud bychom skutečně měli v Excelu k dispozici data jako v tabulce 20, ovšem kompletní, postupovali bychom následovně. Zadávací panel aktivujeme z nabídky Vložení – Kontingenční tabulka. V položce „Vybrat tabulku či oblast“ zadáme buňky obsahující všechny údaje o obou srovnávaných kategoriálních veličinách, pro data podle tabulky 20 tedy oblast B1:C6801. Zbytek zadání lze ponechat a potvrdit (OK); otevře se nový list jako na obrázku 11 s prozatím prázdnou tabulkou (vlevo) a nástroji pro její vyplnění (vpravo). V pravé části v sekci „Zvolte pole, které chcete přidat do sestavy“ najedeme kurzorem na název veličiny „oči“, podržíme levé tlačítko myši, najedeme do levé části do oblasti „Sem přetáhněte řádková pole“ a tlačítko myši uvolníme. Automaticky se vytvoří popisy jednotlivých řádků vznikající kontingenční tabulky (nadpis „oči“ a kategorie 1, 2 a 3); analogicky přetažením pole „vlasy“ zprava pomocí myši do oblasti „Sem přetáhněte sloupcová pole“ vytvoříme popisy jednotlivých sloupců kontingenční tabulky (nadpis „vlasy“ a kategorie 1, 2, 3 a 4); v obou směrech vzniknou zároveň automaticky i pole pro celkové součty (konkr. na řádku 8, resp. ve sloupci F). V dalším kroku vytváření tabulky je jedno, jakou veličinu zprava („oči“ nebo „vlasy“) přetáhneme myší doleva do oblasti „Sem přetáhněte pole hodnot“: vyplní se tak i vnitřek tabulky, bohužel obvykle automaticky nikoli požadovanými četnostmi, ale zde nepotřebnými součty. Toto opravíme následovně: na kterékoli políčko uvnitř vzniklé tabulky klikneme pravým tlačítkem myši a zvolíme možnost „Nastavení polí hodnot“. Jako kritérium je zvýrazněna volba „Součet“, místo níž klikneme na variantu „Počet“ a volbu potvrdíme (OK). Výsledkem je požadovaná kontingenční tabulka pozorovaných četností ve struktuře tabulky 18 (viz tabulka 21, konkr. řádky 3 až 8). Čtenáři lze doporučit, aby si uvedený postup vyzkoušel na vlastních datech menšího rozsahu, protože kompletní data v podobě podle tabulky 20 zde vzhledem k jejich obrovskému rozsahu nemůžeme uvádět.

Nyní jsme tedy ve fázi, že původní data byla automaticky převedena do podoby kontingenční tabulky. Abychom si nyní mohli vyzkoušet provedení testu nezávislosti, kontingenční tabulku si do Excelu sami přepišme přesně podle řádků 3 až 8 z tabulky 21 (podstatné jsou hodnoty pozorovaných četností včetně celkových součtů, tedy hodnoty v buňkách B5 až F8). Vidíme např., že mezi všemi 6800 muži jich 1768 bylo modrookých se světlými vlasy (četnost n_{11} z buňky B5).

Tabulka 21: Kontingenční tabulky pozorovaných a očekávaných četností v Excelu



	A	B	C	D	E	F
1						
2						
3	Počet z vlasy	vlasy				
4	oči		1	2	3	4 Celkový součet
5		1	1768	807	189	47 2811
6		2	118	436	268	16 857
7		3	946	1387	746	53 3132
8	Celkový součet		2832	2632	1223	116 6800
9						
10			1169,46	1086,82	504,57	47,95 7811
11			356,54	331,71	154,03	14,62 857
12			1303,00	1212,27	563,30	53,43 1132
13			2829	2632	1223	116 6800

Další fází je vypočtení odpovídajících očekávaných četností v Excelu podle obecného vzorce (72). Začneme očekávanou četností o_{11}, jejíž výpočet si připravíme do buňky B10 příkazem

=$F5*B$8/F8 (74)

V příkazu (74) se odkazujeme na hodnotu příslušného součtu pro první řádek (buňka F5), resp. pro první sloupec (buňka B8), resp. na celkový součet (buňka F8). Symboly dolaru v příkazu (74) slouží k zamknutí odkazu (viz příklad 4), takže jakmile příkaz (74) překopírujeme do jiné buňky, stále bude odkazovat do příslušného řádkového součtu (sloupec F), resp. do příslušného sloupcového součtu (řádek 8). Překopírujme tedy příkaz (74) do zbylých buněk z oblasti B10 až E12. Výsledky vidíme v dolní části tabulky 21. Např. očekávaná četnost o_{11} má hodnotu 1169,46 (viz buňka B10 z tabulky 21). Pouze pro kontrolu správnosti spočítáme v políčkách F10 až F12 hodnoty řádkových součtů (postupem analogickým závěru příkladu 3), resp. v políčkách B13 až F13 hodnoty sloupcových součtů; výsledky musí odpovídat celkovým součtům pro skutečné četnosti.

K provedení testu nezávislosti chí-kvadrát v Excelu jsou postačující pozorované, resp. očekávané četnosti i bez celkových součtů (v tabulce 21 tedy hodnoty z buněk B5 až E7, resp. B10 až E12). Obecná struktura příkazu je stejná jako u příkazu (68), pro kontingenční tabulky podle tabulky 21 bude tedy mít příkaz konkrétní podobu

=CHITEST(B5:E7;B10:E12) (75)

Výsledkem je p-hodnota, která zde činí 1,12E-228, což je v Excelu zápis pro číslo $1{,}12 \cdot 10^{-228}$ (tj. číslo, v němž se první nenulová číslice nachází až na 228. místě za desetinnou čárkou). Jde tedy o p-hodnotu odpovídající velmi malé kladné hodnotě a rozhodně splňující nerovnost $1{,}12 \cdot 10^{-228} < 0{,}05$. Na 5% hladině významnosti tak můžeme zamítnout nulovou hypotézu o nezávislosti. Jinak řečeno: na základě analyzovaných dat byla testem typu chí-kvadrát prokázána statisticky významná závislost mezi barvou očí a barvou vlasů.

Postupem podle příkladu 30 bychom mohli zpracovat např. i data z kapitoly o věrohodnostním poměru a podílu šancí. Odpovídající kontingenční tabulka (zde speciálně tzv. čtyřpolní) viz tabulka 22. Po spočtení očekávaných četností lze v Excelu získat p-hodnotu 0,062>0,05, takže nelze zamítnout H_0 o nezávislosti. Data tedy neprokázala závislost mezi aplikovaným typem stentu a výskytem MACE; jinak řečeno, rozdíl ve výskytu MACE mezi oběma skupinami pacientů nebyl shledán významným.

Tabulka 22: Čtyřpolní tabulky (vlevo pozorované, vpravo očekávané četnosti), data dle Bystroň a kol.

počty pacientů	s MACE	bez MACE	celkem
skupina 1	12	38	50
skupina 2	5	45	50
celkem	17	83	100

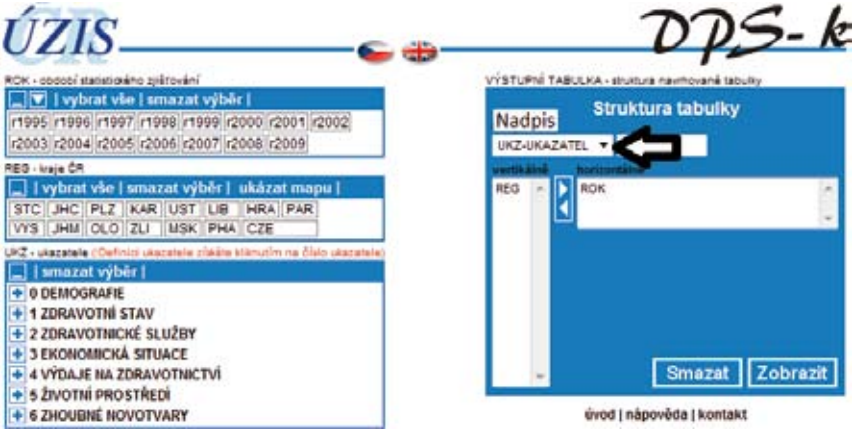
s MACE	bez MACE	celkem
8,5	41,5	50
8,5	41,5	50
17	83	100

Příloha – ukázka postupu při získání dat z www databáze ÚZIS

V kapitolách o geometrickém průměru či o regresních modelech byla na ukázku (viz tabulka 6 či obrázek 9) použita data, získaná z databáze Ústavu zdravotnických informací a statistiky ČR (ÚZIS), dostupné na www stránkách www.uzis.cz. Postup při získávání těchto dat je následující:

1) Na hlavní stránce ÚZIS (www.uzis.cz) otevřeme odkaz UKAZATELE DPS (DPS je zkratka slov Data Presentation System);

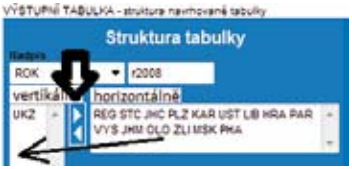
2) Otevře se úvodní stránka, kde zvolíme jazykovou verzi výběrem VSTUP pro českou mutaci (pro anglickou bychom volili ENTER);
3) Tím jsme vstoupili přímo na stránku DPS. V levé části volíme časové, místní a věcné hledisko požadovaných dat (sekce ROK, REG a UKZ), struktura požadované výstupní tabulky se vytváří v pravé části, kde ji lze případně ještě doupravit. Konkrétně data potřebná pro vytvoření obrázku 9 obsahovala dvě veličiny (ukazatele hrubé a standardizované míry incidence novotvarů) pro všechny kraje ČR v jediném vybraném roce (vybrán byl rok 2008, což byl v době tvorby tohoto textu rok s nejaktuálnějšími údaji o novotvarech). Jako nadpis celé tabulky proto zvolíme (kvůli jeho jednoznačnosti) rok 2008: nejprve v pravé části u sekce Nadpis rozklikneme rozbalovací šipku (naznačeno na obrázku)



a namísto přednastavené volby UKZ-UKAZATEL vybereme ROK; to, o jaký rok konkrétně se jedná, upřesníme nyní volbou v levé části u sekce ROK (výběrem r2008);

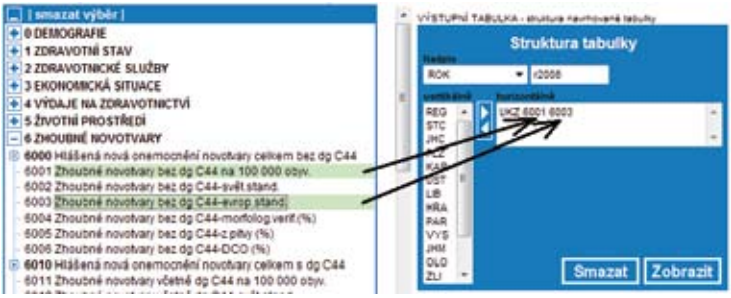
4) Vlevo v sekci REG provedeme jednoduše klikáním výběr všech krajů (kromě hodnoty CZE - ta by udávala souhrn pro celou ČR; rychlejší by zde bylo kliknout na „vybrat vše“ a přebytečnou možnost CZE kliknutím na ni odebrat). Chceme-li mít hodnoty za jednotlivé kraje seřazené pod

sebou v řádcích, musíme ale opět vpravo upravit strukturu výstupní tabulky – nyní stačí kliknout na symbol dvou bílých šipek (naznačeno na obrázku),



tím se vybrané regiony přesunou z horizontálního do vertikálního zobrazení;

5) Posledním krokem je zadání věcného hlediska (sekce UKZ); základní kategorie dostupných ukazatelů jsou DEMOGRAFIE, ZDRAVOTNÍ STAV atd.; konkrétní typy ukazatelů uvidíme až po rozkliknutí symbolu + (vlevo vedle každé kategorie ukazatelů). Konkr. při získání hodnot pro obrázek 9 byla rozevřena kategorie 6 ZHOUBNÉ NOVOTVARY a postupným kliknutím byly vybrány ukazatele 6001 Zhoubné novotvary bez dg C44 na 100 000 obyv. (tedy hrubá incidence) a 6003 Zhoubné novotvary bez dg C44-evrop.standard (naznačeno na obrázku).



6) Tím je tabulka vybrána a k jejímu zobrazení stačí kliknout vpravo na Zobrazit. Tím se otevře nové okno obsahující výslednou tabulku s vybranými číselnými údaji, bohužel ve formátu nevhodném ke kopírování. K překopírování dat např. do Excelu ale postačí z bleděmodrého menu vlevo nahoře (s možnostmi zavřít-tabulka-tisk) vybrat možnost „tisk“, čímž se tabulka otevře znovu ve formátu, z nějž je už možné kopírování pomocí známých klávesových zkratek (CTRL-C, CTRL-V).

2008

	Zhoubné novotvary bez dg C44 na 100 000 obyv.	Zhoubné novotvary bez dg C44-evrop.stand.
STC	526.4	427.3
JHC	588.4	471.7
PLZ	753.3	593.2
KAR	641	534.6
UST	573.7	484.9
LIB	545.9	448.8
HRA	580.7	444.9
PAR	542.3	431.6
VYS	537.5	428.5
JHM	563.5	446.6
OLO	538.8	427.4
ZLI	508.9	402.3
MSK	527	430.1
PHA	591.1	436.8

Ústí nad Labem 2011

Název	Základy biostatistiky s využitím Excelu
Autor	Karel Hrach
Nakladatel	Univerzita J. E. Purkyně v Ústí nad Labem
Edice	Studijní opory projektu Posilování kompetencí vysokoškolských pracovníků pro rozvoj konkurenceschopnosti vysokého školství v Ústeckém kraji, registrační číslo CZ.1.07/2.2.00/07.0117
Náklad	50 ks
Počet stran	48 stran
Tisk a vazba	PrintActive s.r.o., www.dum-tisku.cz

ISBN 978-80-7414-398-4