STATISTIKA V EXCELU

Po přehledu statistického software je nutné zmínit, že pro většinu statistických výpočtů, které provádíme v sociologii statistický software ani nepotřebujeme. Většinu toho, co jsme probírali v této učebnici, zvládne i Excel, který je dnes nainstalován téměř na všech počítacích[1]. V Excelu existuje dvojí možnost práce se statistickými procedurami. První možností je využití speciálních statistických funkcí, druhou pak modulu pro analýzu dat, který je ovšem nutno doinstalovat. Těmto dvěma možnostem se budeme detailněji věnovat.

A) Statistické funkce Excelu

Pro práci s Excesem je charakteristické, že píšeme do buněk (angl. cells, tak se nazývají jednotlivá políčka, která mají označení dle řádku číslem a dle sloupce písmenem, např. první buňka je A1) různé vzorečky a odkazy na jiné buňky. Takto lze sčítat, násobit, dělit apod. Kromě těchto jednoduchých operací umí Excel i složitější matematické funkce (logaritmy, goniometrické funkce atd.), různé finanční funkce (třeba vypočítat splátku hypotéky, naučte se to sami!), textové funkce (sloučit dva texty, oříznout určitý počet znaků zleva či zprava), logické funkce (jako je logické ano, logické nebo, když apod.), dále umí Excel základy práce s databázemi a operace s datem a časem. V neposlední řadě má Excel implementovány i statistické funkce a na ně zaměříme svou pozornost nyní[2].

Statistické funkce implementované v Excelu lze rozdělit do několika kategorií:

a) funkce počítající popisnou statistiku – průměr, směrodatnou odchylku, rozptyl, špičatost atd.,
b) funkce pro jednotlivé statistické testy – t-testy, chí-kvadrát test,
c) funkce počítající kvantily různých statistických rozdělení - např. normálního, t-rozdělení, F-rozdělení, chí-kvadrát apod. a
d) funkce pro regresní a korelační analýzu.

Postupně se dle těchto skupin budeme věnovat nejdůležitějším funkcím, vysvětlíme si jejich zadání a jejich případná úskalí. Dopředu varujeme, že uživatelský komfort v Excelu při statistických výpočtech je nižší než v SPSS. Důvodem je skutečnost, že Excel není primárně zamřen na statistiku, ale statistika je pouze jeho jednou a spíše okrajovou částí.

Obecně o funkcích

Každá funkce v Excelu má své klíčové slovo, například pro funkci počítající průměr je to PRŮMĚR a poté je třeba zadat do závorky z čeho má příslušná funkce počítat (s jakými daty pracovat). Někdy je třeba zadat jednu oblast dat (v řeči Excelu i dále tzv. pole), někdy je třeba více takových oblastí případně určitých parametrů (jako je pravděpodobnost, počet stupňů volnosti apod.

Máme-li tedy například sto údajů v prvním sloupci (tedy sloupci A a v řádku 1-100) o příjmech a chceme spočítat jejich průměr, stačí do řádku určeného pro vzorce v Excelu zapsat tento výraz:

---

[1] Tato skutečnost je také nespornou výhodou Excelu a tento exkurz o Excelu činíme zejména proto, aby jste byli schopni se obejít bez SPSS a používat statistiku v Excelu.

[2] V naší učebnici nemáme prostor na výklad základů užívání Excelu. Pevně věříme, že je tomu věnována patřičná pozornost na středních školách, případně že si student/ka VŠ snadno doplní znalost samostudiem návodu k Excelu.

= PRŮMĚR(A1:A100[3])

Samozřejmě, že tento zápis si není třeba pamatovat, v Excelu můžeme využít tzv. průvodce funkcí (tlačítko f_x na začátku stavového řádku) a poté naklikat, z jakých dat se má počítat (buď za pomoci klávesnice nebo ještě snadněji označit myší) a Excel vytvoří vzorec resp. výpočet za nás a do buňky se provede výpočet.

Pojďme si teď stručně dle výše uvedeného popsat jednotlivé užitečné statistické funkce.

Aa) Funkce pro popisnou statistiku

Excel umí vypočítat základní charakteristiky střední hodnoty, tedy průměr, medián a modus.

PRŮMĚR(číslo1;číslo2;...) – Excel při výpočtu ignoruje prázdné buňky

MEDIAN(číslo1;číslo2;...)

MODE(číslo1;číslo2;...) - pozor pokud je rozdělení vícemodální vypočte Excel jen jeden modus a to ten s nejnižší hodnotou, vyplatí se tedy zobrazit si četnostní tabulky a v ní nalézt modus opticky (viz popis funkce ČETNOSTI dále)

Kromě těchto běžných počítá Excel i „useknutý“ (trimmed) průměr za pomoci funkce:

TRIMMEAN(pole;**procenta**), kde parametr procenta, znamená jaké procento nejvyšších a nejnižších hodnot má být při výpočtu průměru vynecháno (v SPSS standardně 5 %).

Excel samozřejmě počítá i základní charakteristiky rozptýlenosti (variability), maximum, minimum, rozptyl, směrodatnou odchylku.

MAX(číslo1;číslo2;...)

MIN(číslo1;číslo2;...)

VAR[4]**(číslo1**;číslo2;...)

SMODCH(číslo1;číslo2;...)

Pro charakterizování šikmosti a špičatosti lze užít též samostatných funkcí:

SKEW(číslo1;číslo2;...) – šikmost

[3] Povšimněte si, že spojitá oblast buněk mezi A1 a A100 se zapíše zjednodušeně za pomoci dvojtečky. Jiný způsob by byl vypsat všechny buňky a oddělit je čárkou (nikdo ovšem nepochybuje, že takový úmorný zápis by vyžadoval mnoho zbytečně promarněného času).

[4] Povšimněte si zvláštnosti českého Excelu, že některé funkce jsou česky (PRŮMĚR) a některé jsou ponechané v anglickém originále (VAR). Důvod této skutečnosti zřejmě zná jen české zastoupení firmy Microsoft.

KURT (číslo1;číslo2;...) – špičatost

Další charakteristiky popisné statistiky obsažené v Excelu jsou obecné kvantily, kvartily a funkce stanovující pořadí hodnot.

PERCENTIL(pole;k)[5] – parametr k označuje kolika procentní kvantil chceme spočítat, udává se jako pravděpodobnost (nikoliv jako %), tedy zadáváme hodnotu v intervalu 0 a 1.

QUARTIL(pole;kvartil) –vypočte kvartil, nutno zadat namísto parametru kvartil hodnotu mezi 0 a 4 (0-minimum 1 =1 . (25%) kvartil …3 = 3. (75 %) kvartil a 4 značí maximum)

RANK(číslo;odkaz;pořadí) – umožňuje určit pořadí hodnot. Číslo je konkrétní hodnota, pro níž chceme stanovit pořadí, odkaz znamená soubor hodnot, v rámci nichž se má stanovit pořadí. Pořadí je nepovinný prvek zadání[6], pokud není nic uvedeno, stanovuje se pořadí v seznamu seřazeném sestupně, pokud se uvede jakékoliv číslo různé od nuly, stanovuje se pořadí vzestupné.

V závěru se zaměříme na funkci, která umožní vygenerovat četnostní tabulku. Obecný formát této funkce je následující:

ČETNOSTI(data;hodnoty).

Práce s touto funkcí je ovšem poměrně složitá, proto si ji vysvětleme krok za krokem na příkladu.

1. Mějme 100 údajů o známkách (tzv data pro funkci ČETNOSTI) z prvního ročníku sociologie z předmětu A v buňkách A1:A100 (známky mohou být logicky 1-4).
2. Do sloupce B si musíme vepsat, jakých hodnot může naše proměnná nabývat (tedy vepíšeme hodnoty 1, 2, 3, 4 do buněk B1, B2, B3 a B4 –viz tabulka níže), Jde o tzv. hodnoty pro funkci ČETNOSTI
3. Do buňky C1 vepíšeme vzoreček (nebo za pomoci průvodce funkcí vyvoláme funkci ČETNOSTI): Za data vybereme A1-A100 za hodnoty B1-B4 (pro Excel lze tedy užít zápis A1:A100 resp. B1:B4). Výsledná funkce bude mít v buňce C1 tvar:
   = **ČETNOSTI**(A1:A100;(B1:B4)
4. Poté co stiskneme Enter napočítá se počet jedniček v našem souboru.
5. Ti, co z Vás co umí s Excelem, mají možná nápad, že uvedený vzorec v C1 stačí rozkopírovat do buněk C2-C4, ale tak jednoduché to bohužel není. Další kroky jsou mírně nelogické, ale je třeba je dodržet.
6. Nejdříve musíme myší najet na prostředek buňky C1 (kde je vypočtena první četnost) a objeví se bílý široký křížek. Teď stiskneme levé tlačítko myši a táhneme dolů až do buňky C4. Pustíme tlačítko myši a měly by zůstat vyznačené buňky C1-C4.
7. Nyní je třeba stisknout tlačítko F2 a zobrazí se vzorec pro ČETNOSTI s barevnými odkazy[7].

[5] Jednotlivé parametry funkce se oddělují středníkem.
[6] Nepovinné prvky v zadání fukcí budou zde i nadále označeny obyčejným písmem, povinné tučným.
[7] Tato funkce Excelu má obecné použití. Pokud máte složitější vzorec, můžete se za pomoci barevných odkazů snadno přesvědčit, zda se Váš vzorec odkazuje do správných buněk a velice snadno zkontrolujete správnost vašeho výpočtu.

8. Další trojkombinace kláves je podivná, ale musíte si ji pamatovat. Nejdříve stiskneme Ctrl+Shift a poté (obě klávesy stále držíme) ještě Enter. V políčkách C2-C4 se zobrazí vypočtené četnosti.

	A	B	C
1	1	1	= ČETNOSTI(A1:A100;(B1:B4)
2	2	2	
3	2	3	
4	1	4	
5	3		
6	4		
7	2		
8	3		

Není sporu o tom, že výše uvedený postup je poměrně složitý, ale jednodušeji to opravdu nejde (možná lze ještě oklikou pro výpočet četností použít okrajové četnosti u kontingenční tabulky –postup viz dále). Dodejme, že jsme vypočítali absolutní četnosti, z nich už není složité za pomoci jednoduchých vzorců dopočítat relativní četnosti a kumulativní četnosti.

Tip: Pokud chcete počítat četnostní tabulky pro intervalová data, např. pro příjmy Chcete-li zjistit kolik osob má příjmy do 10 tisíc, mezi 10 a 20 tisíci a vyšší, Excel umí i toto. Postup je obdobný výše uvedenému, jen do sloupce B zadáme horní meze příslušných intervalů (tedy 10000 a 20000).

Kontingenční tabulka a její generování

Další procedurou zasahující do popisné statistiky, ale zároveň již umožňující přesahy do statistického testování jsou kontingenční tabulky. Jejich zařazení mezi funkcemi není logicky správné, protože kontingenční tabulky se negenerují přes funkce ale přes nabídku, nicméně z hlediska logiky statistiky musí být vyložen způsob jejich generování dříve, než si ukážeme funkce, které s nimi pracují. Samotné nalezení procedury na generování kontingenčních tabulek nečiní problém, ale práce s procedurou je již poměrně složitá, proto si ji opět postupně vysvětleme.

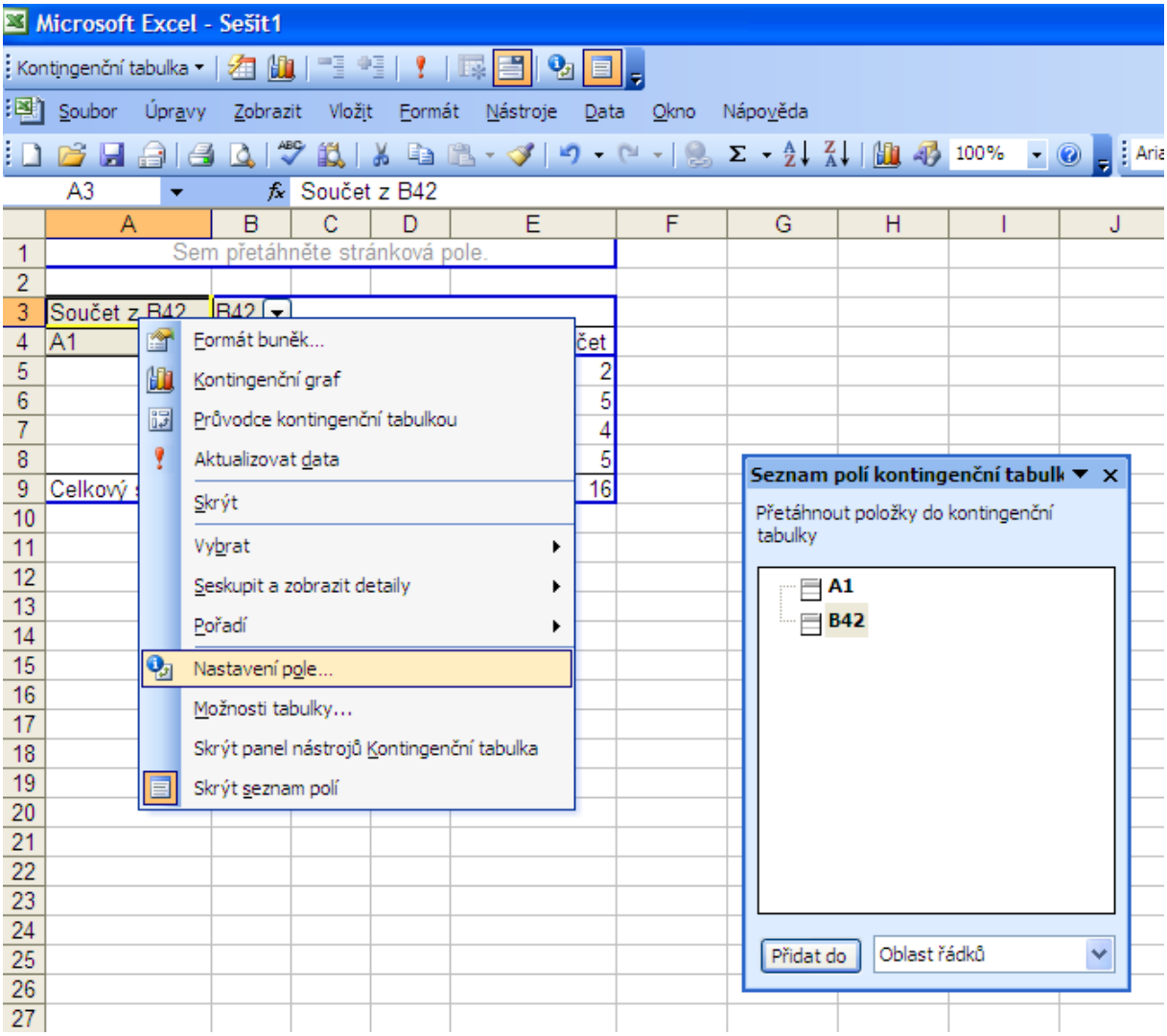
1. V menu najdeme Data-Kontingenční tabulka a graf[8]
2. Zde necháme přednastavenou volbu Seznam Microsoft Excel a Kontingenční tabulka a klikneme na Další.
3. Nyní je nutné vybrat oblast dat, z nichž se má generovat kontingenční tabulka. S ohledem na běžná data jde většinou o dva sloupce (ale Excelu nevadí, když budou dvě analyzované proměnné tvořit dva řádky). Pro jednodušší následnou práci lze doporučit, aby první řádek (nebo v případě proměnných v řádku první sloupec) tvořil popis proměnné alespoň technickým označením jako je A1 či B42. Po vybrání oblasti dat opět klikneme na další.
4. Poslední volbu, kterou musíme učinit, je zde se tabulka vytvoří na nový list (přednastaveno), nebo do stávajícího listu (pak je třeba zadat souřadnice pravého

[8] Průvodce kontingenční tabulkou je obdobný průvodci pro tvorbu grafů, ten v našem textu vynecháváme a očekáváme, že jej čtenář umí používat, nebo se to velice snadno sám naučí.

horního rohu). Vřele doporučujeme použít nový list, aby nedošlo k přepisu zdrojových dat. Poté stiskneme Dokončit.

5. Tímto překvapivě zadání kontingenční tabulky v Excelu (na rozdíl např. od SPSS) nekončí, ale spíše začíná. Pokud jste udělali vše dobře, máte před sebou dvě proměnné (v našem případě A1 a B42) a pak volná pole, kam je třeba tyto proměnné přetáhnout myší. Poměrně intuitivně přetáhneme proměnnou, kterou chceme mít v řádku do políčka kde je nápis Sem přetáhněte řádkové pole, obdobně pro sloupce. Poté stačí vybrat jednu z proměnných (je opravdu jedno kterou z nich) a přetáhnout jí do největšího pole s nápisem Sem přetáhněte datové položky. Poslední operace abychom mohli dále pracovat s napozorovanými (absolutními četnostmi) je tato: Klikneme pravým tlačítkem na šedé políčko, kde je napsáno Součet z B42 (v našem případě, u Vás tam může být jiná proměnná), zvolíme Nastavení Pole (viz i níže na obrázku) a zde namísto Součet vybereme Počet. Poté již získáváme kontingenční tabulku, kterou je možné jako celek Ctrl+C Ctrl+V kamkoliv kopírovat[9]. Tabulka se již velice blíží tomu, co nám poskytne běžný statistický software jen s tím rozdílem, že nemá popsané hodnoty a to musíme udělat sami.
6. Dodejme, že z absolutních četností již není obtížné za pomoci vzorců vypočítat relativní četnosti resp. procenta (ať už celková či řádková nebo sloupcová), to ponecháme na čtenářích.

[9] Tip: Chcete-li tabulku zkopírovat a zbavit se jejích rámečků a barev pak místo Vložit (Ctrl+V) zvolte z Menu *Úpravy-Vložit jinak* a nechte vložit jen hodnoty.



Ab) Funkce pro jednotlivé statistické testy

Po přehledu popisné statistiky se zaměřme na statistické testy, které nabízí funkce v Excelu. Prvním testem bude chí-kvadrát test o nezávislosti v kontingenční tabulce. Pro jeho použití potřebujeme kromě absolutních četností (v terminologii Excelu *aktuální*), které již umíme napočítat, ještě dopočítat očekávané četnosti. Samozřejmě by bylo možné odkázat na příslušnou kapitolu učebnice, ale uvedeme postup i zde. Očekávanou četnost získáme vynásobením součtu příslušného řádku a sloupce kontingenční tabulky poděleného celkovým počtem respondentů. Pro zjednodušení při kopírování vzorce lze využít absolutních adres se symbolem $. Ukažme si to opět na jednoduchém příkladu. V buňkách A3-C5 máme absolutní četnosti, za pomoci funkce SUMA (nebo tlačítka na liště **Σ**) uděláme součty řádků a sloupců a celkový součet (počet respondentů)[10]. Nyní si můžeme vytvořit první očekávanou četnost (např. pro muže, kteří odpověděli ano do buňky B9. Vzorec je patrný z následujícího obrázku:

	A	B	C	D
1		B42		
2	A1	muž	žena	součet

[10] Součty jsou v řádku 6 a sloupci D, ještě jsme je popsali nadpiskem součet.

3	ano	10	30	40
4	nerozhodnutý	30	20	50
5	nevím	20	50	70
6	součet	60	100	160
7				
8				
9		=$D3*B$6/D6		
10				
11				
12				
13				

Dolary ve vzorci zajistí, že když budeme vzorec kopírovat[11] (a počítat další očekávané četnosti) bude se vždy vzorec odkazovat do součtového sloupce D (viz dolar před písmenem ve výrazu $D3), nebo součtového řádku (viz dolar ve výrazu =B$6), nebo do vždy do buňky D6, kde je počet respondentů (dva dolary ve výrazu D6, kterým se dělí). Po nakopírování vzorce již máme vše potřebné a do některé z prázdných buněk lze napsat funkci pro chí-kvadrát test. Její zadání je velice jednoduché:

CHITEST(aktuální;očekávané)

Po zadání oblasti, kde jsou aktuální (napozorované) a očekávané četnosti se nám po stisknutí Enter objeví výsledek. Je jím vypočtená signifikace, tedy vypočtená chyba prvního druhu našeho testu. Tedy bude-li tato menší než 0,05 můžeme se domnívat, že proměnné v kontingenční tabulce jsou statisticky významně závislé.

Tip: Výše uvedená funkce CHITEST se dá využít i k testování rozložení četností (např. reprezentativity výběru apod.), jak o tom byla zmínka již v závěru kapitoly věnované dvourozměrné analýze. Stačí do Excelu vepsat (případně samozřejmě vypočítat za pomoci funkce ČETNOSTI) napozorované četnosti zkoumané proměnné. Dále je nutno dopočítat, jak by měly vypadat, kdyby např. respektovala známé rozdělení ze základního souboru a pak opět použít funkci CHITEST[12].

Pokračujme dalšími statistickými testy, které nabízí Excel konkrétně funkcemi pro t-testy. Ještě před těmito testy však zmíníme F test o shodě rozptylů, který je pro t-testy či analýzu rozptylu[13] logicky předcházející (testuje předpoklad pro t-testy a analýzu rozptylu[14]).

Namísto Leveneho testu o shodě rozptylu ve dvou skupinách je v Excelu implementován F test založený na Bartlettově odvození testového kritéria. Použití je poměrně jednoduché. Dvě

[11] Kopírovat vzorce lze buď klasicky Ctrl+C a Ctrl+V, pro kopie do větších oblastí je však praktičtější najet myší na pravý dolní roh buňky, kde je obsah ke kopírování (udělá se úzký křížek) a pak podržet levé tlačítko a pohnout myší kam chci obsah zkopírovat. Pozor je možné kopírovat jen po sloupcích či řádcích, pokud tedy potřebuji kopírovat po řádcích i sloupcích je to nutno učinit ve dvou po sobě následujících krocích.
[12] Tato skutečnost naznačuje formální statistickou podobnost obou těchto testů.
[13] Funkce pro analýzu rozptylu Excel nezná, ale je implementována v rámci modulu pro analýzu dat (viz dále).
[14] V detailech lze odkázat na výklad v kapitole Srovnávání středních hodnot. F test v Excelu ale bohužel umí jen testovat shodu rozptylů ve dvou skupinách, proto ho nelze užít pro analýzu rozptylu, ale jen pro rozhodnutí, jaký typ dvouvýběrového t-testu volit.

skupiny, o nichž chceme zjistit, zda mají obdobné rozptyly, zadáme jako pole do příslušné funkce:

FTEST(pole1;pole2).

Výsledkem je opět vypočtená signifikace (malá naznačuje neshodu rozptylů).

Nyní se přesuneme k t-testu, resp. spíše t-testům. V rámci funkce TTEST se nabízí tři typy t-testu, párový, dvouvýběrový pro shodné (equal variances v SPSS) a neshodné rozptyly (not equal variances v SPSS)[15]. Pro použití t-testů (ale totéž platí i pro již dříve popsaný F test, je třeba v Excelu data nejdříve upravit. Nelze tedy mít vedle sebe dvě proměnné např. příjem a pohlaví, ale musíme mít v jedné souvislé oblasti (v poli) příjmy pro muže a v jiné souvislé oblasti příjmy pro ženy. Nejjednodušeji toho dosáhneme tak, že seřadíme (viz nabídka Data-Seřadit) hodnoty v obou proměnných dle třídící proměnné (tedy v našem příkladu dle pohlaví). Kromě dvou polí musíme zadat, zda chceme jednostranný či dvoustranný test a poslední číslo ve funkci TTEST udává typ testu:
1 – párový
2 – dva nezávislé výběry se shodnými rozptyly (tedy pokud bude signifikace F testu větší než 0,05)
3 – dva nezávislé výběry s neshodnými rozptyly (tedy pokud bude signifikace F testu menší nebo rovna 0,05). Funkce TTEST tedy vypadá následovně:

TTEST(pole1;pole2;strany;typ).

Výsledkem je opět signifikace a nízké hodnoty indikují statisticky významné rozdíly v průměrech mezi skupinami.

Poznámka: V případě dostatečně velkých výběrů (cca 30-50 jednotek) je možné pro jednovýběrový t-test možné použít funkce ZTEST.

Ac) Funkce počítající kvantily různých statistických rozdělení

Excel lze využít místo statistických tabulek pro hledání kvantilů různých statistických rozdělení. Ještě před ukázkou hledání kvantilů si ukažme funkci pro standardizaci (výpočet z skórů) proměnné. Jde o funkci:

STANDARDIZE(x;střed_hodn;sm_odch).

Tato funkce tedy vyžaduje odkaz na hodnotu, ze které chceme počítat z skór (x), dále průměr (střed_hod) příslušné proměnné a její směrodatnou odchylku (sm_odch).

Nyní již věnujme pozornost funkcím pro výpočet kvantilů nejčastěji užívaných statistických rozdělení. Konkrétně se zaměříme jen na normované normální rozdělení, t-rozdělení, chí-kvadrát rozdělení a F (Fisherovo) rozdělení. Excel umí hledat kvantily i dalších rozdělení, ale vzhledem k tomu, že jejich využití v sociálněvědní statistice je okrajové, nevěnujeme jim pozornost.

[15] Ještě by bylo možné říct, že díky tomu, že Excel nabízí dvoustranné i jednostranné testy je vlastně t-testů 6 typů (to zda chceme jednostranný test či oboustranný zadáváme pomocí třetí hodnoty ve funkci TTEST, 1= jednostranný a 2 = oboustranný). Nicméně uvažujeme zde pro jednoduchost jen oboustranné testy.

První a nejjednodušší je funkce pro generování kvantilů normovaného normálního rozdělení. Funkce má tvar:

NORMSINV(P),
kde P udává, kolika procentní kvantil chceme získat (jde ale o pravděpodobnost, tedy číslo mezi 0 a 1). Například nejběžněji užívaný 97,5 % kvantil získáme po napsání výrazu:

= NORMSINV(0.975)

V Excelu se pak v příslušné buňce objeví nám již známá hodnota 1,96[16].

U funkce pro hledání kvantilů t-rozdělení je kromě pravděpodobnosti zadat i stupně volnosti.

TINV(prst;volnost)

Funkce pro hledání kvantilů t-rozdělení (stejné platí i pro F rozdělení-srovnej dále) má ale jednu zradu. Při hledání 97,5 % kvantilu nezadáme do pole pravděpodobnost hodnotu 0,975 jak by bylo možné čekat analogicky k funkci NORMSINV, ale doplněk do jedničky (1-0,975) navíc ještě znásobený dvěmi (tedy hodnotu 0,05). Důvod této zapeklitosti asi tuší jen programátoři Excelu[17]. Pro jistotu ještě uveďme příklad. Pokud hledáme 95 % kvantil t-rozdělení pro 20 stupňů volnosti bude mít zápis v Excelu podobu:

=TINV(0,1;20)

Další funkcí pro hledání kvantilů, která vyžaduje kromě pravděpodobnosti též zadání stupňů volnosti je funkce generující kvantity chí-kvadrát rozdělení.

CHIINV(prst;volnost)

Opět zadáváme počet stupňů volnosti a i zde nám tvůrci Excelu nachystali zapeklitost při zadávání pravděpodobnosti. Chceme-li 95 % kvantil, musíme do pravděpodobnosti zapsat doplněk do jedničky, tedy 0,05 apod.

Poslední funkce určená k hledání kvantilů, kterou zmiňujeme je funkce FINV pro hledání kvantilů F rozdělení., Toto rozdělení je založené na podílu dvou veličin s chí-kvadrát rozdělením a není proto divu, že kromě pravděpodobnosti je třeba zadat dvojí stupně volnosti (veličin v čitateli a jmenovateli z nichž testové kritérium s F rozdělením vzniká).

FINV(prst;volnost1;volnost2)

I zde je situace při zadávání složitější. Platí totéž co u funkce chí-kvadrát, pro hledání 95 % kvantilu píšeme do pole pravděpodobnost 0,05, pro hledání 99 % kvantilu 0,01 apod.

---

[16] Samozřejmě po zaokrouhlení na dvě desetinná místa. Excel umožňuje provést výpočet mnohem přesněji, ale dvě místa pro naše účely bohatě postačí.

[17] Násobení dvěmi lze pochopit s ohledem na to, že kvantily Excelem vypočtené jsou pro oboustranné testy (to naznačuje i nápověda Excelu k této funkci), ale hledání doplňku do jedničky takto jednouše vysvětlit nelze.

Ad) Funkce pro regresní a korelační analýzu

Pro výpočet Pearsonova korelačního koeficientu (tedy koeficientu pro spojité proměnné) nabízí Excel dvě výpočetně zcela stejné funkce:

PEARSON(pole1;pole2)

a

CORREL (pole1, pole2)

Stačí tedy zadat dva sloupce nebo dva řádky, kde máme proměnné, pro které chceme vypočítat korelaci a výstupem je vypočtený korelační koeficient. Narozdíl od SPSS tedy Excel nenabídne test u nulovosti koeficientu v celé populaci ani žádné hvězdičky pro ocenění hladiny statistické významnosti. Pokud někdo chce test o nulovosti korelačního koeficientu provést, nezbývá než si najít příslušný vzorec, dosadit do něj a porovnat testové kritérium s hodnotou příslušného kvantilu (vše již umíte, tak si to sami zkuste!).

Poslední statistickou funkcí, které věnujeme pozornost, je funkce pro jednoduchou lineární regresní analýzu, resp. jedná se o dvě funkce, první počítá regresní konstantu (INTERCEPT) a druhá pak směrnici (SLOPE). Máme-li tedy dva sloupce či řádky, kde máme hodnoty závislé a nezávislé proměnné lze použít uvedené funkce (do dvou odlišných buněk!). Zadání je poměrně jednoduché:

INTERCEPT(pole_y;pole_x)

a

SLOPE(pole_y;pole_x).

Výsledkem je odhad konstanty a směrnice regresní přímky proložené našimi daty. Nezískáme žádné testy, ani odhady indexu determinace, či jiné složitější regresní diagnostiky, o nichž bylo pojednáno v kapitole o regresní analýze. S ohledem na tyto nedostatky lze doporučit spíše výpočet regrese obsažený v modulu na analýzu dat (viz další část textu).

B) Modul analýza dat

Modul analýza dat výrazně rozšiřuje statistické možnosti Excelu a přibližuje také uživatelskou snadností Excel standardnímu statistickému software. Modul samotný nalezneme v nabídce *Nástroje-Analýza dat.* Pokud jste otevřeli Excel a nenašli nabídku Analýza dat, je vše v pořádku, protože nejde o součást standardní instalace. Je tedy nejdříve doinstalovat tento modul (slovy Excelu doplněk). K této operaci zpravidla potřebujete instalační CD sady MS Office[18].

Provedení instalace Analýzy dat je jednoduché (návod pro Excel 2003):

1) V menu najdeme *Nástoje-Doplňky*.
2) Zde vybereme hned první možnost nahoře Analytické nástroje a klikneme na OK.
3) Excel chvilku pracuje a poté by se již měla objevit nabídce nástroje Analýza dat (standardně úplně dole v menu *Nástroje*, pokud nemáte instalovány jiné doplňky)[19].

Nyní již můžeme navštívit nabídku *Nástroje-Analýza dat* a vidíme 19 statistických procedur, jejichž přehled dále uvádíme.

Přehled jednotlivých procedur v analytickém modulu

Analýza rozptylu jednofaktorová – vypočítá analýzu rozptylu s jednou nezávislou proměnnou (faktorem).

Než přikročíme k výčtu dalších procedur (věnovat se jim detailně by znamenalo napsat samostatnou mnohasetstránkovou učebnici) detailněji si ukážeme zadání a výstupy analýzy rozptylu v modulu Analýza dat právě pro jednofaktorovou analýzu rozptylu. Ovládání ostatních procedur je obdobné, proto předpokládáme, že si čtenář po pochopení ovládání jedné procedury sám zkusí ostatní a bez problémů je bude schopen používat.

Nejdříve si tedy uveďme, jak si musíme připravit data pro použití analýzy rozptylu. Příklad, který budeme řešit je následující: Zjišťujeme, zda se odlišuje názor na ospravedlnění euthanasie[20] dle vzdělání (pro jednoduchost uvažujeme rozdělení na osoby se základním (1), středním (2) a vysokoškolským vzděláním (3)). Část dat tedy může mít tuto podobu:

	A	B	C	D
1	vzdělání	Euthanasie		
2	1	10		
3	3	8		
4	2	6		
5	2	5		
6	1	3		
7	2	5		
8	3	3		
9	1	7		

[18] Pokud provádíte instalaci v síťovém prostředí, např. v počítačové učebně na fakultě, je zpravidla instalace obsažena na některém ze síťových disků a instalační CD nepotřebujete. Postupujete dle dále uvedených kroků.

[19] V případě, že se Vám to nepodaří, zkuste požádat někoho, kdo má více zkušeností s instalováním software případně s Excelem.

[20] Odpovědi uvažujeme na škále 1- zcela ospravedlnitelná až 10 – zcela neospravedlnitelná.

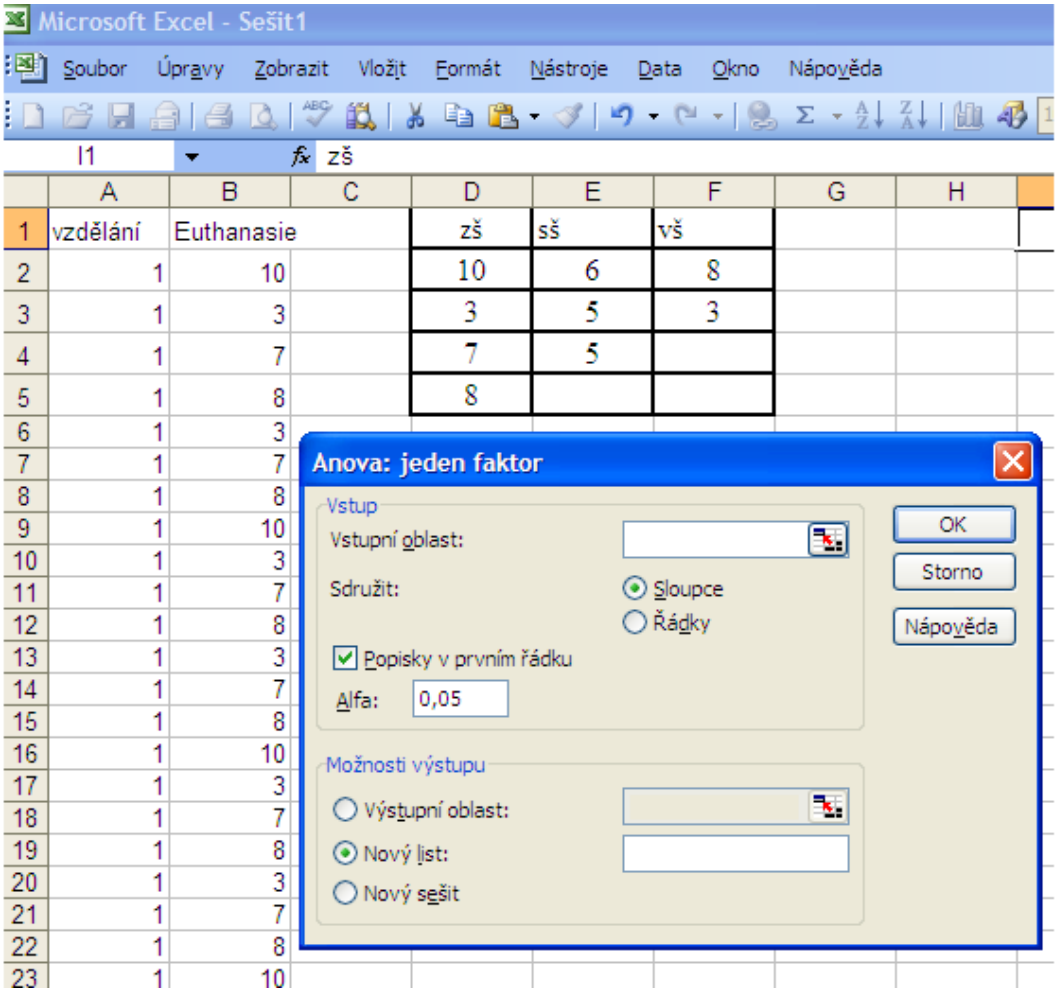
10	1	8
11	2	5
12	2	5
13		

Pro použití analýzy rozptylu (ale obdobně i pro t-testy, regresi apod.) je třeba data uspořádat tak, aby vždy data za jednotlivé skupiny (dle vzdělání) utvořili jednolitou oblast (nejlépe asi v jednom sloupci). Zopakujme, že nejjednodušeji toho dosáhneme tak, že naše data setřídíme podle proměnné pohlaví a pak optimálně zkopírujeme data pro jednotlivé vzdělanostní skupiny do jednotlivých sloupců (v našem příkladě volíme sloupce D,E,F). Výsek našich dat připravených pro analýzu má tedy podobu[21]:

	D	E	F	D
1	zš	sš	vš	
2	10	6	8	
3	3	5	3	
4	7	5		
5	8			
6				
7				
8				
9				
10				
11				
12				
13				

Takto připravená data již můžeme použít v analýze rozptylu. Vybereme *Nástroje-Analýza dat- Analýza rozptylu jednofaktorová* a získáme zadávací okno:

[21] Dodejme, že popisky v prvním řádku musíme vepsat sami, Excel to za nás neudělá. Popisky nejsou nutné, ale jsou velice vhodné, pokud chceme mít pěkné výstupy a případně se v nich chceme vyznat i po určitém čase.



Postup zadávání je tento:

1. Do vstupní oblasti zadáme oblast, kde máme naše data včetně popisků skupin a zaškrtneme volbu Popisky v prvním řádku (užitečnost tohoto kroku uvidíme ve výstupech procedury). Ponecháme sdružování sloupců, pokud bychom data pro jednotlivé skupiny měli v řádcích, změníme na sdružování dle řádků.
2. Dále nám Excel umožní měnit hladinu statistické významnosti (standard 0,05) a určit kam chceme zobrazit výstupy. Vřele doporučujeme využít přednastavenou volbu vygenerování výstupu na nový list Excelu, protože výstupy jsou zpravidla delší a při vložení výstupu do stávajícího listu hrozí, že si přepíšeme nějaká data.
3. Stiskneme OK a získáme výstupy:

Anova: jeden faktor

Faktor

Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
zš	42	276	6,571429	6,10453
sš	54	276	5,111111	4,100629
vš	18	84	4,666667	5,882353

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*

Mezi výběry	68,17043	2	34,08521	10,63908	5,93E-05	3,078057
Všechny výběry	355,619	111	3,203775			
Celkem	423,7895	113				

Výstup z analýzy rozptylu je rozdělen do dvou tabulek (např. u regrese jsou dokonce tři). V první tabulce je popisná statistika, počet jednotek ve skupinách (v našem výstupu máme 42 osob se ZŠ, 54 se SŠ a 18 s VŠ), dále můžeme vidět průměry ospravedlnitelnosti euthanasie v jednotlivých vzdělanostních skupinách a také rozptyly. Připomeňme, že bychom měli otestovat shodu rozptylů (ale to v případě více skupin Excel neumí), proto postačí se podívat na vypočtené rozptyly a konstatovat, že podíl největšího a nejmenšího je zhruba 1,5 a jsou si tedy podobné a analýzu rozptylu lze použít.
Druhá tabulka již uvádí výsledky samotné analýzy rozptylu. Pro nás je asi nejdůležitější Hodnota P, vypočtená chyba prvního druhu (tedy totéž co Sig. v SPSS). Tato hodnota je 5,93E-05[22]. Můžeme konstatovat, že rozdíly v průměrném hodnocení ospravedlnitelnosti euthanasie alespoň mezi dvěmi vzdělanostními skupinami existují. Z popisné statistiky máme náznak, jaké tyto rozdíly jsou a jak je věcně interpretovat. Problémem je, že Excel na rozdíl od statistického software (jako je SPSS, SAS, STATA) nenabízí následné testy (Post hoc tests v SPSS). Proto pokud budeme chtít zjistit statisticky významné rozdíly mezi dvojicemi (ZŠ a SŠ, ZŠ a VŠ a SŠ a VŠ) budeme muset třikrát spočítat dvouvýběrový t-test[23] (viz dále) a dle Bonferonniho korekce neužívat pro hodnocení statistické významnosti hodnotu 0,05, ale tuto podělit počtem provedených t-testů (tedy třemi), tj. užívat zhruba hladinu významnosti 0,02.

Analýza rozptylu dvoufaktorová bez opakování - pro výzkumné designy bez opakování příslušné kombinace hodnot faktorů. V sociologii nemají tyto experimentální designy příliš velké použití.

Analýza rozptylu dvoufaktorová s opakováním - pro výzkumné designy s opakováním příslušné kombinace hodnot faktorů. Tato procedura může být v případě, že chceme použít v sociologii dvoufaktorovou analýzu rozptylu použít, užitečnější než předchozí.

Korelace – počítá Pearsonův korelační koeficient mezi dvěmi či více proměnnými

Kovariance - počítá kovariance mezi dvěmi či více proměnnými (tedy korelaci násobenou součinem směrodatných odchylek dvou proměnných, z nichž je počítána korelace)

Popisná statistika – výpočet charakteristik polohy (průměr, medián a modus), charakteristik rozptýlenosti (minimum, maximum, variační rozpětí, směrodatná odchylka, rozptyl, špičatost, šikmost, počet analyzovaných jednotek, součet všech hodnot, chybu střední hodnoty (pokud odečteme její dvounásobek resp. tento přičteme k průměru, získáme interval spolehlivosti pro střední hodnotu v celé populaci)

[22] Pro ty z Vás co tento formát čísla neznají dodejme, že jde o vědecký zápis čísel, výraz E-05 nahrazuje skutečnost, že abychom číslo zobrazili v běžném desetinném zápisu musíme jej vynásobit 10^{-5}. Jde tedy o hodnotu 0,0000593 a Excel si zápis zjednodušil, aby číslo nezabíralo zbytečně moc místa a nebylo třeba rozšiřovat sloupec.
[23] Návod jak na to, naleznete dále.

Dvouvýběrový F test pro rozptyl – test shody rozptylů pro dvě skupiny (obdoba funkce FTEST)

Histogram – umožňuje kreslit různé histogramy četností

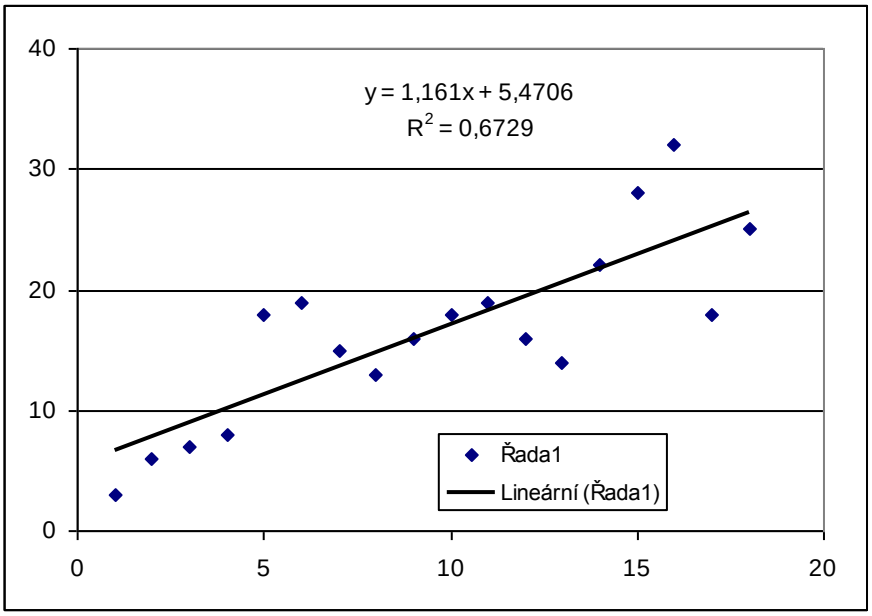
Pořadová statistika a percentily – počítá pořadí jednotlivých hodnot (od největší po nejmenší hodnotu) a pro každou hodnotu v souboru určí kolika procentním percentilem je.

Regrese – umožňuje výpočet lineární regresní analýzy jak jednoduché (s jednou nezávislou proměnnou) tak vícenásobné (s více nezávislými proměnnými). Výstupy z procedury Regrese jsou velice obdobné výstupům z SPSS. Nejdříve je tabulka, která udává index determinace a vícenásobný korelační koeficient (v Excelu ovšem nazvané chybně jako Hodnota spolehlivosti R a Násobné R), a také upravený (adjusted) index determinace (v Excelu tzv. Nastavená hodnota spolehlivosti R). V další tabulce je celkový F test pro regresní analýzu a poslední regresní výstup udává odhady parametrů[24], otestování jejich statistické významnosti pomoci dílčích t-testů a také intervaly spolehlivosti odhadnutých regresních parametrů[25].
Pro práci s regresí v Excelu (ale i s jinými metodami) by bylo vhodné ukázat tvorbu umělých (dummy) proměnných. Napovíme jen, že snadno to lze za pomoci logických funkcí (zejména funkce KDYŽ), nicméně detailnější ukázka již přesahuje možnosti tohoto textu a zvídavý čtenář si na potřebné postupy jistě přijde sám.

Exkurz[26] (Regrese v grafu): Neodpustíme si s ohledem na proceduru regrese v Excelu jeden exkurz, který může být užitečný pro jednoduchou regresi a částečně nahrazuje výše uvedenou proceduru. Pokud chcete řešit jednouchou regresní úlohu je vhodné si nakreslit graf (optimálně XY bodový). Pokud máme hotový graf je možné z Excelu získat regresní rovnici a index determinace. Postup je tento:

1. Klikneme na body, které zobrazují závislost mezi proměnnou X a Y a ony se vyžlutí.
2. Stiskneme pravé tlačítko myši a vybereme Přidat spojnici trendu.
3. Zde máme na výběr, zda chceme modelovat závislost našich proměnných za pomoci přímky (lineární), logaritmu, mocninné funkce, exponenciely a polynomu[27]. Pokud nahoře klikneme na záložku Možnosti, můžeme za pomoci voleb v dolní části nabídky nechat za pomoci volby Zobrazit rovnici regrese a Zobrazit hodnotu spolehlivosti R vepsat do grafu regresní rovnici a index determinace. Po stisknutí OK může vypadat výstup jako na následujícím obrázku.

[24] Konstanta je v Excelu nazvaná také podivně jako hranice.
[25] Vynecháváme nabídku Excelu kreslit nejrůznější grafy založené na reziduích, v kapitole o regresní analýze jsou tyto popsány pro SPSS a pro Excel platí totéž.
[26] Jsme si vědomi staré pravdy, že vědecké dílo se vyznačuje tím, že obsahuje exkurzy, ale exkurzy samotné z díla vědecké neudělají.
[27] Tento výběr křivek je také výhodou grafického řešení regrese, které umí i jiné než lineární typy regrese. Regrese v rámci modulu Analýza dat umí pouze lineární regresi.

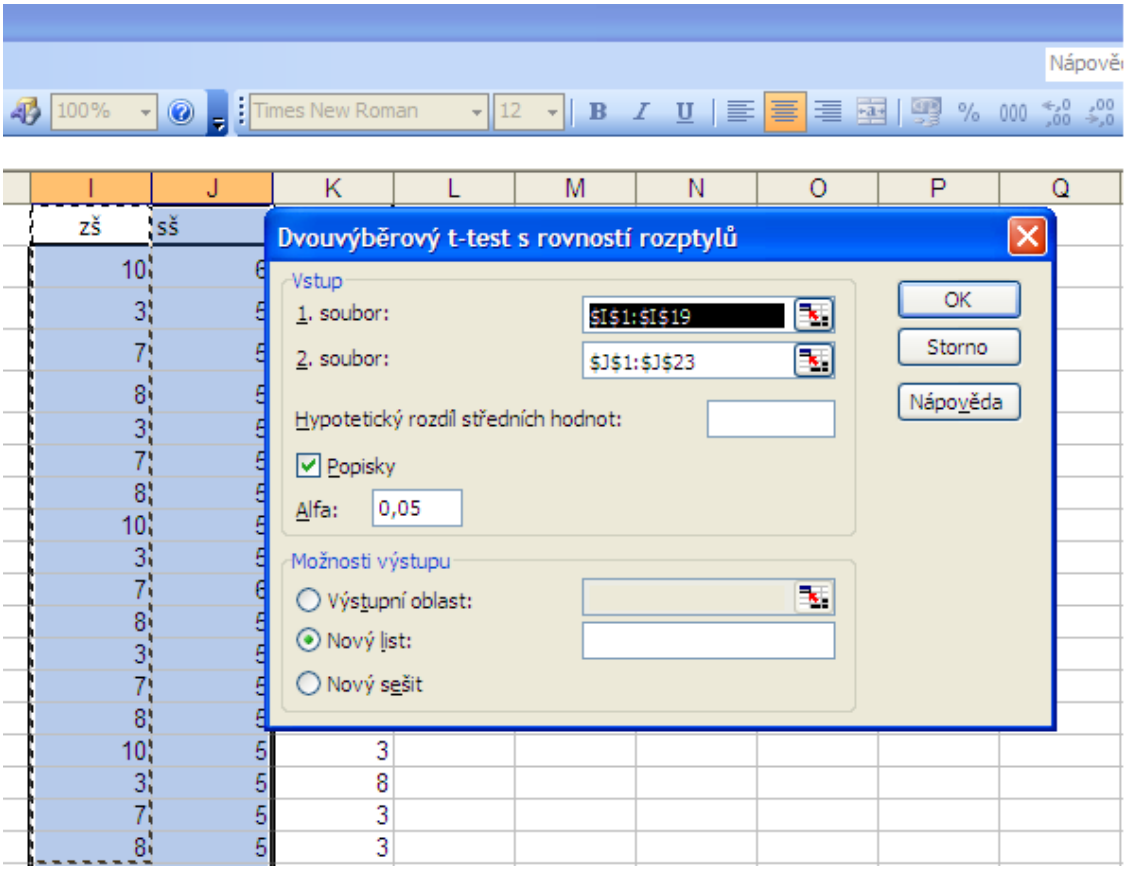


Kromě jednotlivých bodů, které zobrazují kombinace hodnot námi sledovaných proměnných je zobrazena regresní přímka (lineární křivka), která vystihuje závislost mezi proměnnými. Nahoře tedy vidíme rovnici regrese y = 1,161x + 5,4706 a také hodnotu indexu determinace R^2 = 0,6729. Rovnici regrese můžeme použít pro odhad hodnot za pomoci regresního modelu a pro výpočty reziduí apod.

Vzorkování – vybere z našeho souboru (buď náhodně, nebo systematicky) určitý počet jednotek. Tato volba má význam pro simulační úlohy, ale i pokud vybíráme z menších populací a máme jejich seznam (oporu výběru).

Dvouvýběrový t-test s rovností rozptylů – t-test pro dva nezávislé výběry při shodě rozptylů **(**viz **Dvouvýběrový F test pro rozptyl** nebo funkce **FTEST).**

Protože zadávání t-testů je mírně odlišné od výše uvedené analýzy rozptylu a také výstupy jsou poměrně specifické, zmíníme zde důležité rozdíly. Při zadávání t-testu se samostatně zadávají data pro jednu a druhou skupinu:



Výstupy jsou také odlišné, proto se na ně zaměřme:

Dvouvýběrový t-test s rovností rozptylů

	zš	*sš*
Stř. hodnota	6,666667	5,136364
Rozptyl	6,470588	4,123377
Pozorování	18	22
Společný rozptyl	2,962919	
Hyp. rozdíl stř. hodnot	0	
Rozdíl	38	
t stat	2,797275	
P(T<=t) (1)	0,004022	
t krit (1)	1,685954	
P(T<=t) (2)	0,008045	
t krit (2)	2,024394	

Uvedená tabulka navazuje na příklad z analýzy rozptylu, závislá proměnná je tedy opět ospravedlnitelnost euthanasie a sledujeme, zda je tato v průměru obdobná u osob se ZŠ a SŠ vzděláním. Na počátku tabulky je popisná statistika, tedy průměry a rozptyly v obou skupinách. Poté je uveden počet jednotek v každé ze skupin a dále již údaje potřebné pro výpočet t kritéria a statistické významnosti.
Pro naše účely je zřejmě nejdůležitější řádek nadepsaný P(T<=t) (2), který udává statistickou významnost pro oboustranný t-test. S ohledem na to, že hodnota je menší než konvenční 5 % hladina významnosti (tedy 0,05), lze usoudit, že rozdíl v míře ospravedlnitelnosti euthanasie mezi osobami s ukončeným základním středním vzděláním se statisticky významně odlišuje. Pokud bychom chtěli provádět jednostranný test (tedy měli směrovanou alternativní

hypotézu) museli bychom nahlédnout o dva řádky výše do řádku nadepsaného P(T<=t) (1). Dodejme, že vyhodnocení i zadání dalších typů t-testu i z testu je obdobné, proto je již takto detailně nerozebíráme.
Pozornost si zaslouží poslední parametr, který je možné v t-testech zadat a to je Hypotetický rozdíl středních hodnot. Za pomoci t-testu můžeme také testovat nulovou hypotézu, že rozdíl mezi středními hodnotami je roven nějaké námi předpokládané hodnotě.

Dvouvýběrový t-test s nerovností rozptylů - t-test pro dva nezávislé výběry při neshodě rozptylů (pro zájemce Excel nabízí v nápovědě k této funkci i vzorec, podívejte se na něj).

Dvouvýběrový párový t-test na střední hodnotu - test pro dva vzájemně závislé výběry.

Dvouvýběrový z-test na střední hodnotu – test pro dva nezávislé výběry v případě, že jejich velikost je minimálně 30 jednotek a rozptyly v obou výběrech jsou shodné **(**opět viz **Dvouvýběrový F test pro rozptyl** nebo funkce **FTEST).**

V přehledu byly vynechány některé speciální funkce pro analýzu časových řad, konkrétně **Exponenciální vyrovnání, Fourierova analýza a Klouzavé průměry**. Tyto procedury by vyžadovaly hlubší výklad časových řad v rozsahu cca 2 semestrálního kurzu, lůze odkázat na výklady podávané v učebnicích statistiky pro ekonomické obory či v ekonometrických učebnicích.