Regresní a korelační analýza

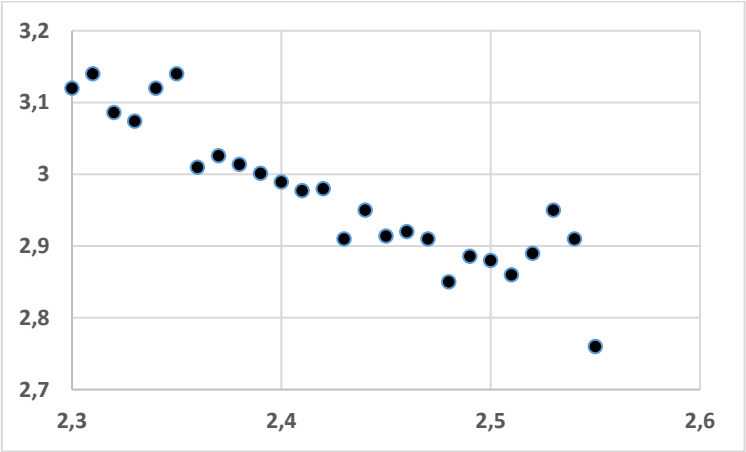
1. Hodnocení závislostí

Při statistickém hodnocení analyzovaných dat můžeme kromě klasického testování rozdílů mezi skupinami objektů také monitorovat potenciální vztah dvou a více kvantitativních proměnných. Příkladem hodnocení závislosti kvantitativních znaků může být sledování, zda existuje vztah mezi délkou křídla ptáků a jejich věkem. Je předpoklad, že délka křídla bude záviset na věku. Tuto hypotézu je ovšem nutné podložit vhodnými statistickými postupy, které nám umožní nalézt druh sledované závislosti a její sílu.

2. Regresní a korelační analýza

S využitím regresní analýzy jsme schopni nalézt vhodný matematický model charakterizující závislost hodnocených proměnných. Regresní analýza matematicky popisuje daný vztah s využitím vhodné funkce. Rozlišujeme dva základní typy regresní analýzy – jednoduchou a vícerozměrnou. Jednoduchá (jednorozměrná) regrese má za cíl nalézt relevantní model funkční závislosti veličiny Y na jedné veličině X. Typ vhodné funkce často jednoduše odhadneme z bodového diagramu, ve kterém je každá dvojice sledovaných údajů (tzv. korelační dvojice) graficky znázorněna jedním bodem v rovině (obrázek 1). Vícerozměrná regrese nám popisuje vztah závislosti veličiny Y na více veličinách.

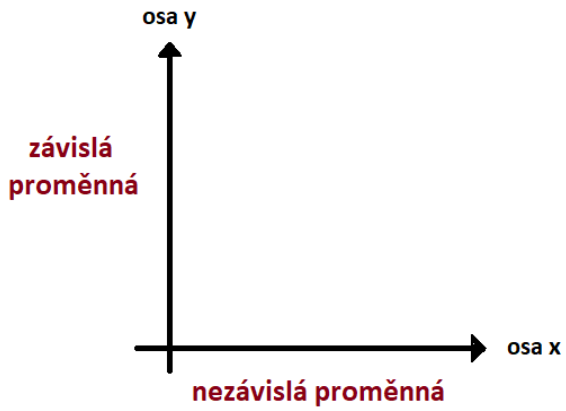
Obrázek 1: Ukázka bodového diagramu



V praxi se relativně často setkáváme s lineárním modelem, který předpokládá lineární závislost mezi dvěma veličinami. Jednoduchá lineární regrese nám umožní předpovědět s určitou pravděpodobností hodnotu jedné proměnné díky znalosti hodnoty druhé proměnné.

Základem jednoduché lineární regrese je rovnice přímky, která má obecný tvar: **y = kx + q.** Hodnoty x a y jsou sledované proměnné, u kterých hodnotíme potenciální vztah. Hodnota x odpovídá nezávislé (vysvětlující) proměnné a hodnota y je závislá (vysvětlovaná, predikovaná, očekávaná) proměnná (obrázek 2). Parametry *k* a *q* jsou konstanty, které charakterizují danou přímku (obrázek 3). Konstanta *k* je definována jako směrnice přímky, která určuje její sklon. Je-li směrnice přímky kladná, jedná se o přímku vzrůstajícího charakteru. Naopak pokud je směrnice záporná, jedná se o přímku klesajícího charakteru. Konstanta *q* udává posunutí přímky na ose y. Kvalitu regresního modelu udává tzv. koeficient determinace R^2. Tento parametr si lze podobně jako rovnici přímky nechat vyjádřit v tabulkovém procesoru Excel. Koeficient (index) determinace udává kolik procent rozptylu nezávislé proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno. Maximální hodnota tohoto indexu je rovna 1. Chceme-li tedy zjistit, zda pro naše data nebude vhodnější jiná matematická funkce (např. logaritmická, exponenciální), provedeme porovnání jednotlivých funkcí právě pomocí koeficientu (indexu) determinace. Čím je hodnota tohoto indexu blíž k hodnotě 1, tím je daná matematická funkce vhodnější pro popis našich dat, případně je třeba při výběru vhodného modelu diskutovat charakter závislosti s odborníkem v dané oblasti studované problematiky.

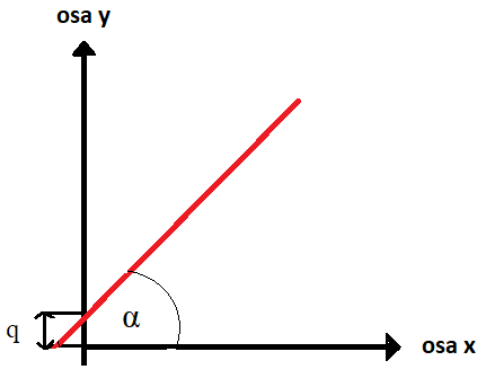
Obrázek 2: Grafické znázornění závislé a nezávislé proměnné.



Úkolem korelační analýzy je zhodnotit a změřit sílu (těsnost) dané závislosti. Korelace nám označuje stupeň asociace dvou proměnných. Uvádíme, že dané dvě proměnné jsou korelované, pokud určité hodnoty jedné proměnné mají tendenci se vyskytovat společně s určitými hodnotami druhé proměnné. Mírou těsnosti vztahu hodnocených proměnných je korelační koeficient. Korelační koeficient může dosahovat hodnot od 1 do -1. Kladná hodnota korelačního koeficientu značí pozitivní závislost, naopak záporná hodnota korelačního koeficientu indikuje negativní závislost. Pokud je korelační koeficient roven 0, neexistuje zde závislost. Hodnota korelačního koeficientu je bezrozměrné číslo vyjadřující těsnost vztahu

dvou proměnných a není závislá na použitých jednotkách. Pro finální posouzení závislosti proměnných se ověřuje významnost daného korelačního koeficientu.

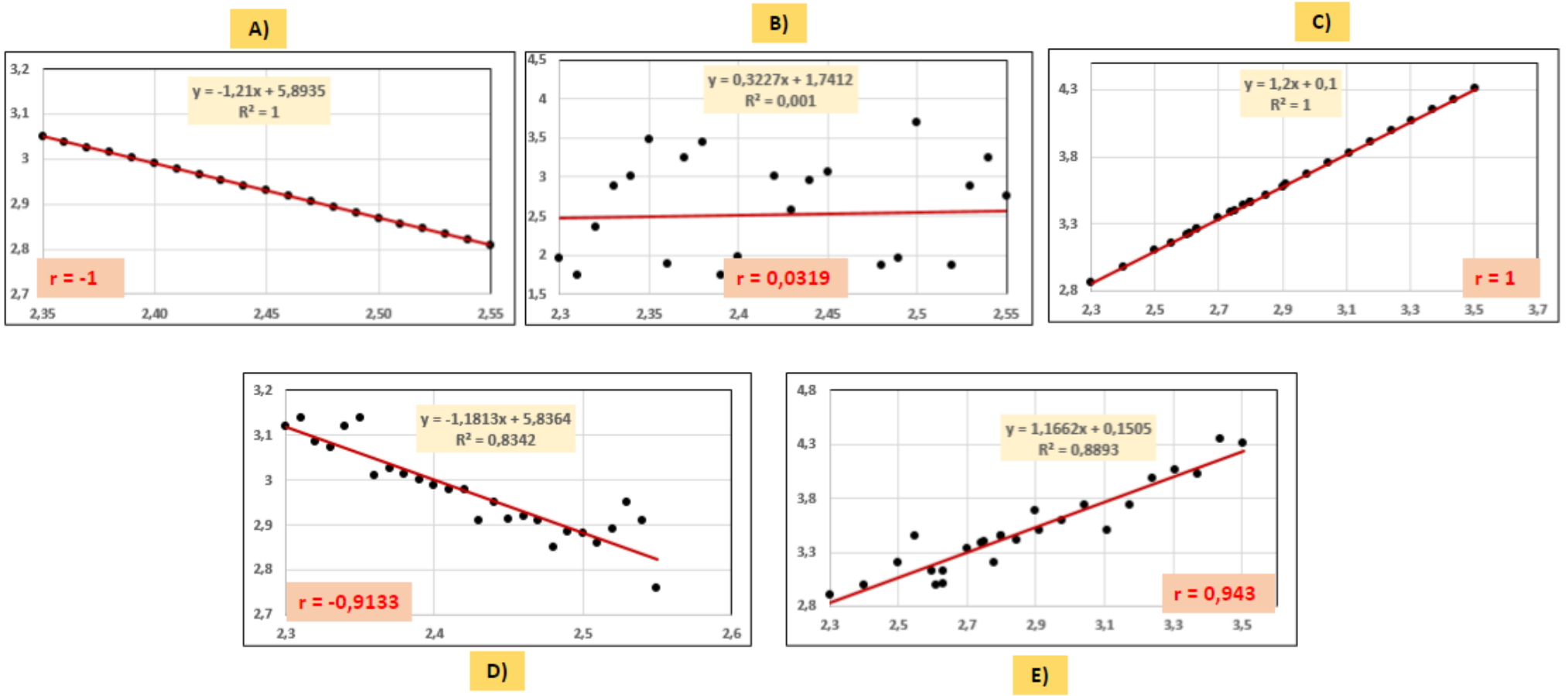
Obrázek 3: Definice konstant *k* (směrnice přímky) a *q* (posun na ose y).



Je-li u testovaných proměnných splněna podmínka normality, lze pro hodnocení těsnosti závislosti využít Pearsonův korelační koeficient. Pearsonův korelační koeficient patří mezi nejdůležitější míry síly vztahu dvou náhodných spojitých proměnných a vyjadřuje pouze sílu lineárního vztahu. Obdobně jako průměr či směrodatná odchylka je tento parametr ovlivněn odlehlými hodnotami. Na obrázku 4 jsou uvedeny příklady různých závislostí včetně výsledků Pearsonova korelačního koeficientu (r). Pokud není u obou výběrových souborů podmínka normality splněna nebo nelze předpokládat linearitu, využívá se vhodná neparametrická metoda. Nejčastěji používanou neparametrickou metodou pro hodnocení korelace je výpočet Spearmanova pořadového (korelačního) koeficientu (r_s). Výpočet tohoto koeficientu je založen podobně jako ostatní neparametrické testy na pořadí jednotlivých hodnot a v porovnání s Pearsonovým korelačním koeficientem je rezistentní vůči odlehlým hodnotám.

Výpočet Pearsonova korelačního koeficientu lze provést v tabulkovém procesoru Excel (funkce „CORREL“). Bohužel ovšem Excel není schopen provést zhodnocení významnosti tohoto korelačního koeficientu. Pro tyto účely je nezbytné využít vhodný statistický software. Ukázka výpočtu Pearsonova korelačního koeficientu v tabulkovém procesoru Excel a statistickém programu Unistat for Excel 6.5 je uvedena na obrázku 5 a 6. Jak již bylo uvedeno, není-li splněna podmínka normálního rozdělení, je třeba pro posouzení těsnosti závislosti kvantitativních znaků využít Spearmanův pořadový koeficient. Výpočet tohoto koeficientu ovšem nelze provést v tabulkovém procesoru Excel, musíme využít některý ze statistických programů, které nám zároveň ověří i významnost uvedeného korelačního koeficientu. Ukázka výpočtu Spearmanova pořadového koeficientu v programu Unistat for Excel 6.5 je uvedena na obrázku 7.

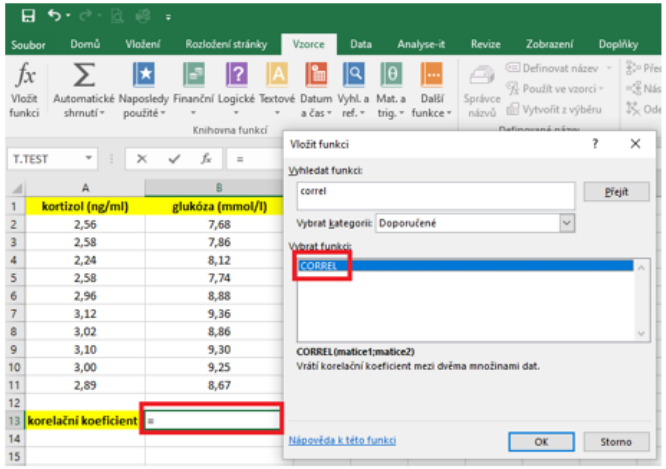
Obrázek 4: Hodnoty Pearsonova korelačního koeficientu (r) pro různé typy závislostí.



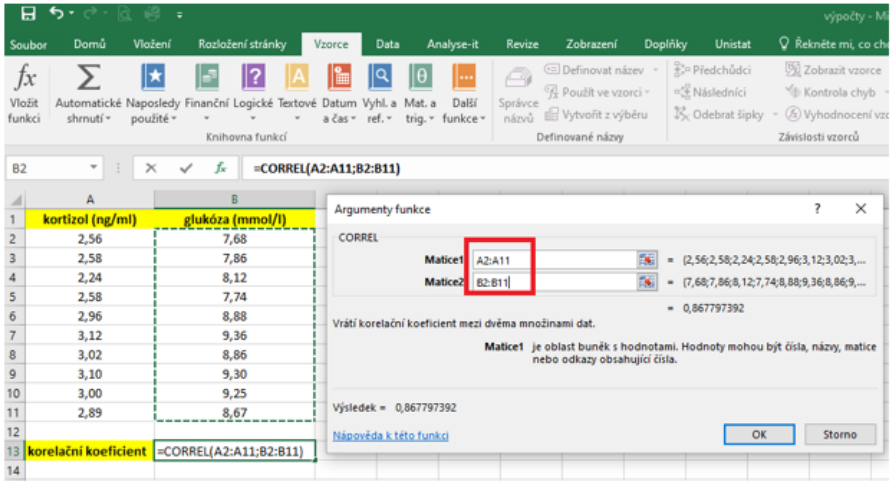
Poznámka: U všech výběrových souborů bylo s využitím Shapiro-Wilkova testu potvrzeno normální rozdělení ($p > 0,05$). Testování bylo provedeno s využitím programu Unistat for Excel 6.5.

Obrázek 5: Výpočet Pearsonova korelačního koeficientu v tabulkovém procesoru Excel.

A. V tabulkovém procesoru Excel lze vypočítat pouze Pearsonův korelační koeficient. Nelze ovšem ověřit jeho významnost. V nabídce „Vzorce“ si zvolíme funkci „Correl“.



B. Do kolonky Matice 1 a Matice 2 označíme data porovnávaných výběrových souborů.



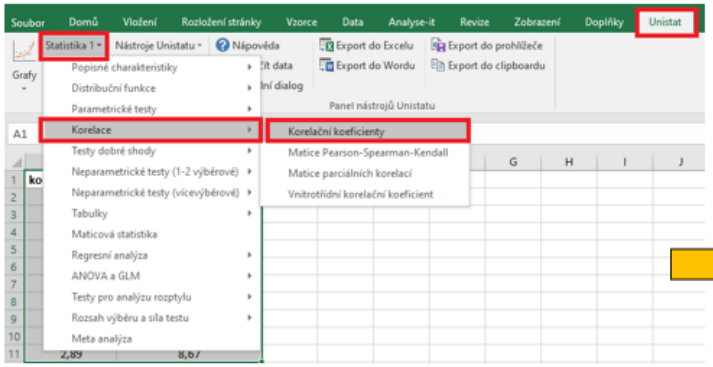
	A	B
1	kortizol (ng/ml)	glukóza (mmol/l)
2	2,56	7,68
3	2,58	7,86
4	2,24	8,12
5	2,58	7,74
6	2,96	8,88
7	3,12	9,36
8	3,02	8,86
9	3,10	9,30
10	3,00	9,25
11	2,89	8,67
12		
13	korelační koeficient	0,867797392
14		

VYHODNOCENÍ

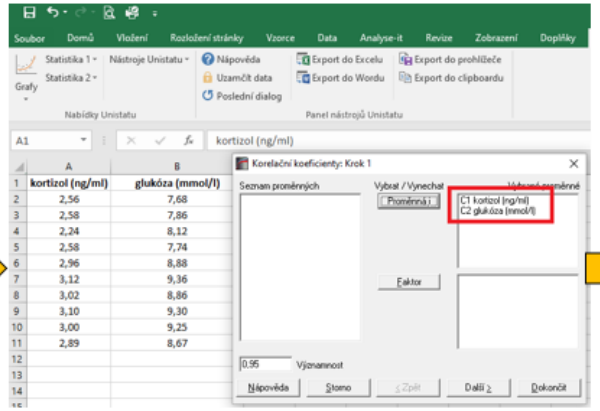
C. Hodnota korelačního koeficientu je 0,868, což znamená, že se jedná o pozitivní korelaci. Vzhledem k tomu, že tato hodnota je blízká číslu 1, lze usuzovat, že mezi testovanými znaky bude signifikantní závislost. Tento fakt, by ovšem bylo ale třeba ověřit pomocí vhodného statistického softwaru.

Obrázek 6: Postup provádění testu korelační analýzy s využitím Pearsonova koeficientu v programu v programu Unistat for Excel 6.5.

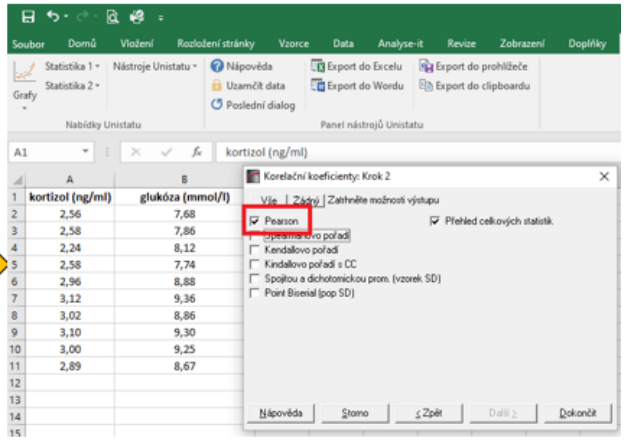
A. Označíme si zdrojová data a v hlavním menu si v nabídce Unistat zvolíme: **Statistika 1 → Korelace → Korelační koeficienty.**



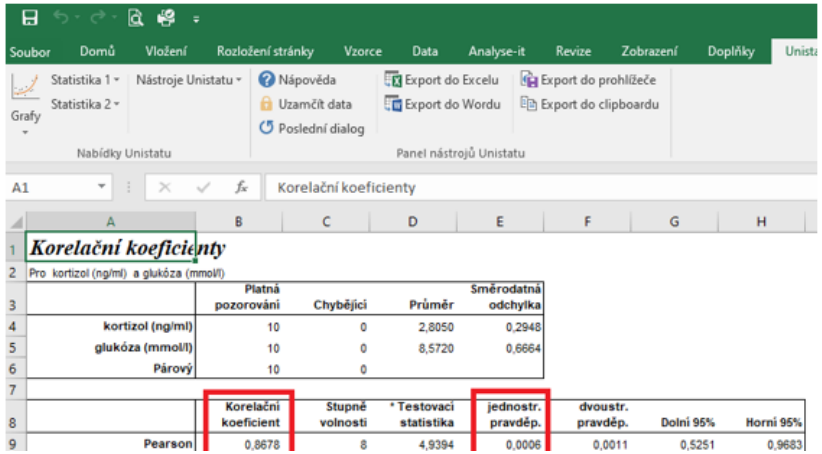
B. V dalším okně si vybereme do položky proměnné oba porovnávané soubory a zvolíme „Další“.



C. Při testování normality bylo u obou VS potvrzeno normální rozdělení, budeme proto používat využívající Pearsonův koeficient. Zvolíme si „Pearson“ a potvrdíme tlačítkem „Dokončit“.



VYHODNOCENÍ



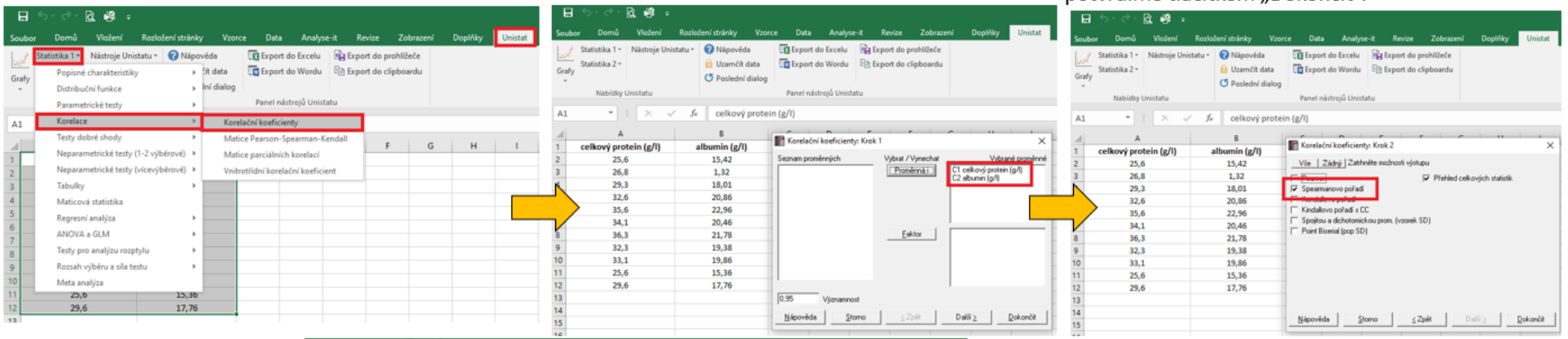
D. Dojde k vytvoření nového listu, kde v druhé tabulce vidíme hodnotu korelačního koeficientu (r = 0,8678) a výsledek jednostranné pravděpodobnosti p (p = 0,0006). Na základě získaných výsledků můžeme tvrdit, že mezi sledovanými výběrovými soubory byla potvrzena statisticky vysoce významná pozitivní korelace.

Obrázek 7: Postup provádění testu korelační analýzy s využitím Spearmanova pořadového koeficientu v programu v programu Unistat for Excel 6.5.

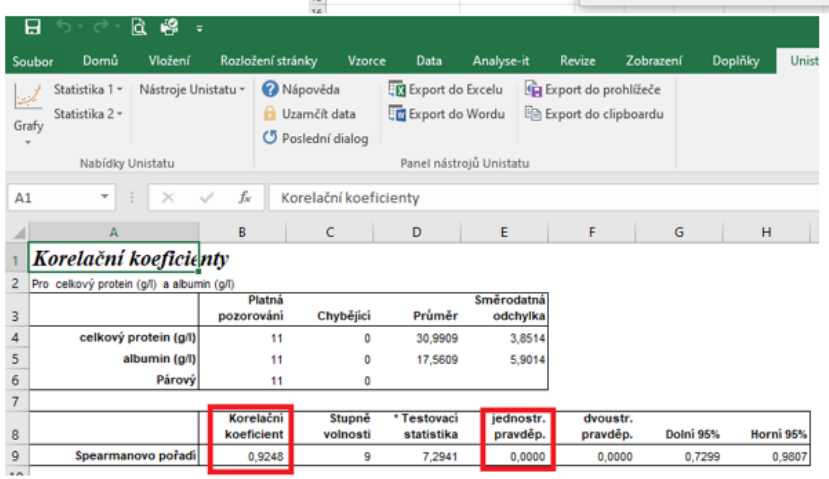
A. Označíme si zdrojová data a v hlavním menu si v nabídce Unistat zvolíme: **Statistika 1 → Korelace → Korelační koeficienty.**

B. V dalším okně si vybereme do položky proměnné oba porovnávané soubory a zvolíme „Další“.

C. Při testování normality nebylo u obou VS potvrzeno normální rozdělení, budeme proto používat využívající Spearmanův pořadový koeficient. Zvolíme si „Spearmanovo pořadí“ a potvrdíme tlačítkem „Dokončit“.



VYHODNOCENÍ



Korelační koeficienty

Pro celkový protein (g/l) a albumin (g/l)

	Platná pozorování	Chybějící	Průměr	Směrodatná odchylka
celkový protein (g/l)	11	0	30,9909	3,8514
albumin (g/l)	11	0	17,5609	5,9014
Párový	11	0		

	Korelační koeficient	Stupně volnosti	*Testovací statistika	jednostr. pravděp.	dvoustr. pravděp.	Dolní 95%	Horní 95%
Spearmanovo pořadí	0,9248	9	7,2941	0,0000	0,0000	0,7299	0,9807

D. Dojde k vytvoření nového listu, kde v druhé tabulce vidíme hodnotu korelačního koeficientu (r_s= 0,9248) a výsledek jednostranné pravděpodobnosti p (p = 0,000). Na základě získaných výsledků můžeme tvrdit, že mezi sledovanými výběrovými soubory byla potvrzena statisticky vysoce významná pozitivní korelace.

Testováním normality bylo potvrzeno, že jeden výběrový soubor (albumin) nesplňuje podmínku normálního rozdělení (p = 0,0013).

Zdroje:

Hendl, J. Přehled statistických metod – analýza a metaanalýza dat. 2015. Vydavatelství Portál Praha, 736 s.

Lepš, J. Biostatistika. 1996. Jihočeská univerzita v Českých Budějovicích, 166 s.

Meloun, M., Militký, J. 2012. Kompendium statistického zpracování dat. Nakladatelství Karolinum, Univerzita Karlova v Praze. 982 s.