Cvičení 2: Vícerozměrné t-testy

Příklad na vícerozměrný jednovýběrový t-test

Podle údajů na obalu má rybí konzerva obsahovat 55 g masa, 30 g zeleniny a 15 g oleje. Náhodně bylo vybráno 10 konzerv a v každé z nich byla zjištěna hmotnost masa (proměnná X1), hmotnost zeleniny (proměnná X2) a hmotnost oleje (proměnná X3). Získané údaje jsou uloženy v souboru rybi_konzervy.sta.

Úkol 1.: Vypočtěte vektor výběrových průměrů **M** a výběrovou varianční matici **S**.
Řešení:
Výpočet vektoru **M**: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza - Proměnné X1, X2, X3 – OK – OK – záložka Popisné statistiky - Shrnutí popisných statistik

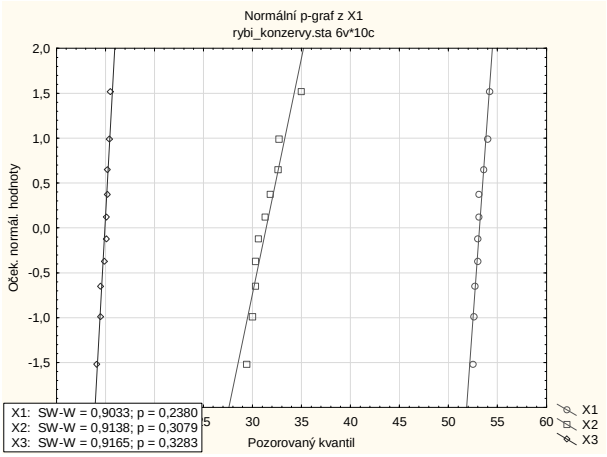
	Souhrn. statistiky (rybi_konzervy.sta)	
Proměnná	Průměr	Sm. Odch.
X1	53,18000	0,576965
X2	31,40000	1,675974
X3	14,95000	0,447834

Výpočet matice **S**: Návrat do výsledky hlavních komponent – Kovarianční matice

	Kovariance (rybi_konzervy.sta)		
Proměnná	X1	X2	X3
X1	0,332889	-0,408889	-0,032222
X2	-0,408889	2,808889	0,307778
X3	-0,032222	0,307778	0,200556

Komentář: Ve zkoumaných 10 konzervách je v průměru o něco méně masa a oleje než jsou deklarované hodnoty, zato více zeleniny. Dále vidíme, že s klesajícím podílem masa roste podíl zeleniny a podíl oleje. S rostoucím podílem zeleniny roste i podíl oleje. Největší variabilitu vykazuje zelenina, menší maso a nejmenší olej.

Úkol 2.: Na hladině významnosti 0,05 testujte hypotézu, že proměnné X1, X2, X3 se řídí normálním rozložením. Vytvořte normální pravděpodobnostní grafy.
Řešení:
Grafy – 2D grafy – Normální pravděpodobnostní grafy – Proměnné X1, X2, X3 – OK - zaškrtneme S-W test a Více grafů v jednom obrázku – OK



Komentář: S-W test ani v jednom případě nezamítá hypotézu o normalitě dat na hladině významnosti 0,05. Rovněž tečky v N-P grafech leží v těsné blízkosti ideální přímky. Data budeme tedy považovat z realizace výběru z třírozměrného normálního rozložení.

Úkol 3.: Na hladině významnosti 0,05 testujte hypotézu H_0: $\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\end{pmatrix} = \begin{pmatrix}55\\ 30\\ 15\end{pmatrix}$ proti alternativě H_1: $\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\end{pmatrix} \neq \begin{pmatrix}55\\ 30\\ 15\end{pmatrix}$.

Řešení:

Statistiky – Základní statistiky a tabulky – t-test. samost. Vzorek – OK – Proměnné X1, X2, X3 – OK – záložka Možnosti – zvolíme Test průměrů vůči různým volitelným konstantám Specif. X1: 55, X2: 30, X3: 15 – OK – zaškrtneme Vícerozměrný test (Hotellingovo T^2) – Výpočet

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (rybi_konzervy.sta) T2(celé případy ChD)=103,532 F(3,7)=26,842 p<,00033							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
X1	53,18000	0,576965	10	0,182452	55,00000	-9,97520	9	0,000004
X2	31,40000	1,675974	10	0,529990	30,00000	2,64156	9	0,026845
X3	14,95000	0,447834	10	0,141618	15,00000	-0,35306	9	0,732169

Komentář: Testová statistika vícerozměrného jednovýběrového t-testu se realizuje hodnotou 26,842, odpovídající p-hodnota je 0,00033, tedy na hladině významnosti 0,05 považujeme za prokázané, že složení konzerv neodpovídá údajům na obalu.

Úkol 4.: Zjistěte, vzhledem ke kterým složkám vektoru **μ** byla nulová hypotéza zamítnuta, tj. simultánně testujte H_{01}: $\mu_1 = 55$, H_{02}: $\mu_2 = 30$, H_{03}: $\mu_3 = 15$ proti H_{11}: $\mu_1 \neq 55$, H_{12}: $\mu_2 \neq 30$, H_{13}: $\mu_3 \neq 15$.

Řešení:

Použijeme 3 jednovýběrové t-testy, kde hladinu významnosti $\alpha = 0{,}05$ upravíme pomocí Bonferroniho korekce. H_{0j} zamítneme na hladině významnosti $\alpha = 0{,}05$, když vypočtená p-hodnota bude menší nebo rovna $\frac{\alpha}{\text{počet testů}} = \frac{0{,}05}{3} = 0{,}017$.
Podíváme-li se na tabulku uvedenou u úkolu 3, vidíme, že vícerozměrná hypotéza byla zamítnuta kvůli první složce, tj. kvůli podílu masa. U zeleniny a oleje se neprokázala odlišnost od deklarovaných hodnot.

Příklad na vícerozměrný dvouvýběrový t-test

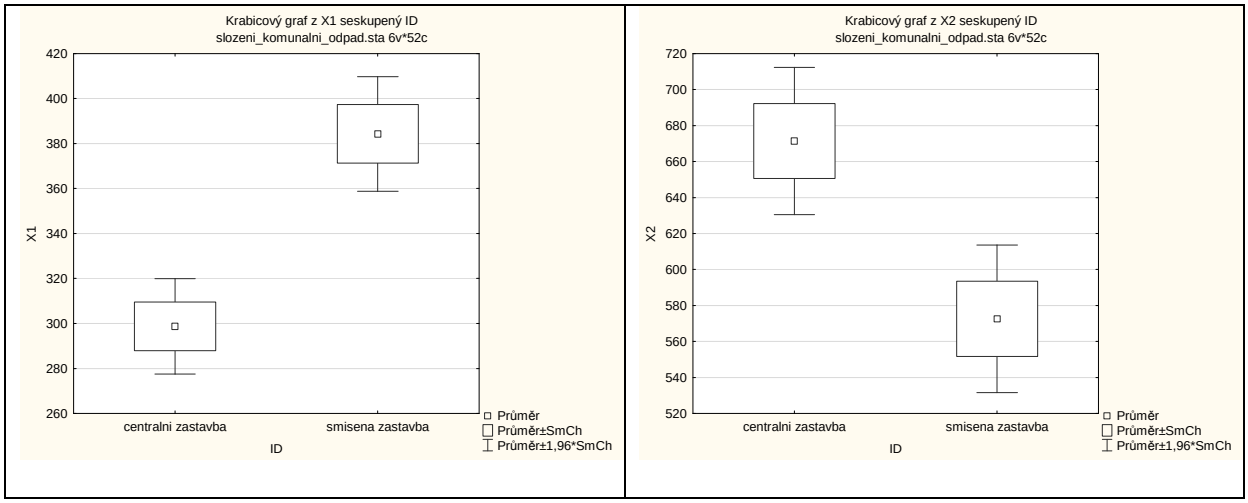
V rámci předběžných úvah o způsobu zpracování tuhého komunálního odpadu byl analyzován obsah 24 náhodně vybraných kontejnerů umístěných v centrální zástavbě, která je vytápěna převážně dálkovým topením a obsah 28 náhodně vybraných kontejnerů ve smíšené zástavbě, kde se vedle dálkového topení hojně vyskytují i lokální topeniště. Byly zjišťovány hodnoty pěti proměnných:
X1 … měrná hmotnost
X2 … podíl hrubé frakce (zůstává v sítu s oky 40 mm)
X3 … podíl jemné frakce (propadá sítem s oky 8 mm)
X4 … vlhkost (v promile)
X5 … výhřevnost (v kJ/kg)
Výsledky analýz jsou uloženy v datovém souboru slozeni_komunalni_odpad.sta.

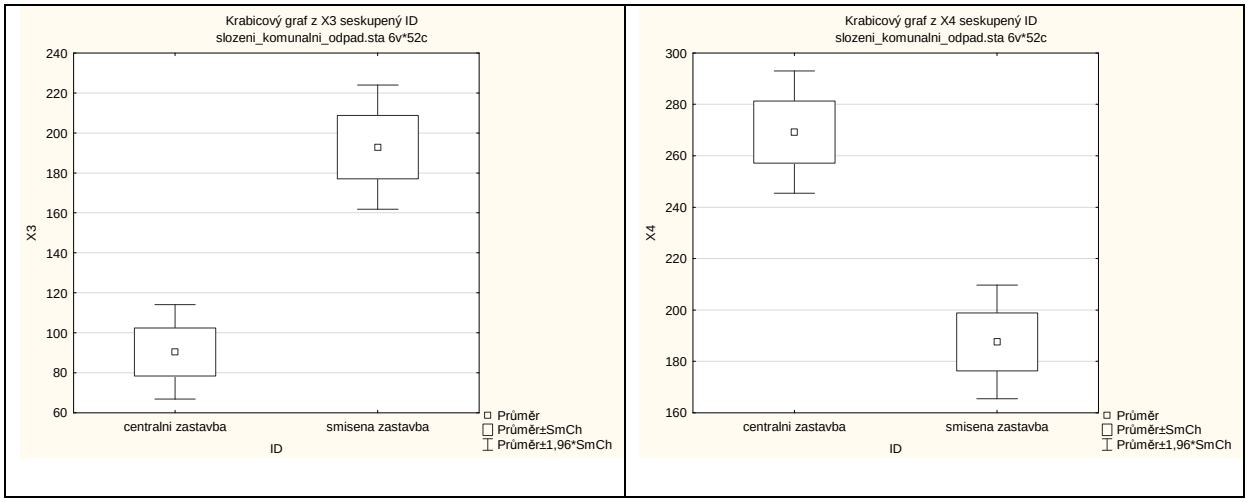
Úkol 1.: Ve obou skupinách vypočtěte průměry a směrodatné odchylky proměnných X1, X2, X3, X4, X5. Vytvořte krabicové grafy proměnné Xi obou skupinách, i = 1, 2, 3, 4, 5.
Řešení: Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X1, X2, X3, X4, X5 – OK – Anal. skupin – zaškrtneme Zapnuto a Sloučit tabulkové výsledky v jedné tabulce a zrušíme Výsledky za všech. skupiny – zadáme Skupin. proměnná ID – OK – Detailní výsledky – zrušíme Minimum a maximum – Výpočet

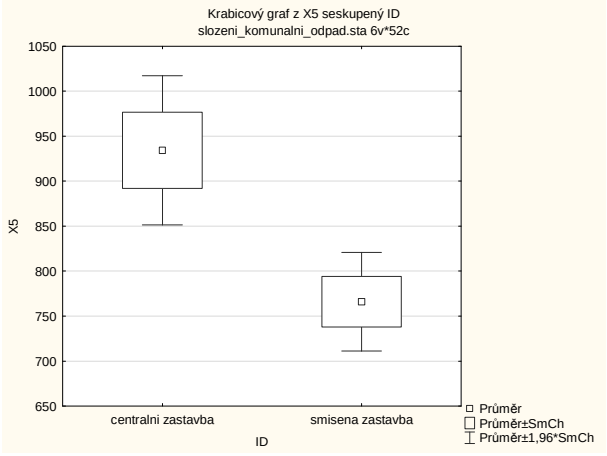
	Souhrnné výsledky Popisné statistiky (slozeni_komunalni_odpad.sta)			
Proměnná	ID	N platných	Průměr	Sm.odch.
X1	centralni zastavba	24	298,7500	53,0228
X2	centralni zastavba	24	671,3750	102,0672
X3	centralni zastavba	24	90,5000	58,8565
X4	centralni zastavba	24	269,2083	59,3816
X5	centralni zastavba	24	934,3333	206,9185
X1	smisena zastavba	28	384,2857	68,8223
X2	smisena zastavba	28	572,5714	110,6250
X3	smisena zastavba	28	192,8571	83,7747
X4	smisena zastavba	28	187,6429	59,6733
X5	smisena zastavba	28	766,0357	148,0639

Komentář: Ve smíšené zástavbě je v průměru vyšší měrná hmotnost odpadu a vyšší podíl jemné frakce, u podílu hrubé frakce, výhřevnosti a vlhkosti je tomu naopak.

Grafy – 2D grafy – Krabicové grafy – Typ grafu: Vícenásobný – Proměnné – Závisle proměnné X1 – Grupovací proměnná ID – Detaily – Střední bod – Průměr – v části Krabicový zvolíme Hodn.: SmCh, v části Svorka zvolíme Hodn.: SmCh, koeficient 1,96 - Odlehlé hodnoty – Vypnuto – OK
Tentýž postup zopakujeme pro proměnné X2, X3, X4, X5.







Úkol 2.: Na hladině významnosti 0,05 testujte hypotézu, že proměnné X1, X2, X3, X4, X5 se v obou skupinách řídí normálním rozložením.
Řešení: Statistiky – Základní statistiky a tabulky – Tabulky četností – OK - X1, X2, X3, X4 – OK - Anal. skupin – zaškrtneme Zapnuto a Sloučit tabulkové výsledky v jedné tabulce a zrušíme Výsledky za všech. skupiny – zadáme Skupin. proměnná ID – OK – OK – záložka Normalita – zaškrtneme S-W test a zrušíme K-S test – Testy normality

Proměnná	Souhrnné výsledky Testy normality (slozeni_komunalni_odpad.sta) ID	N	max D	Lilliefors p	W	p
X1: merna hmotnost	centralni zastavba	24	0,096522	p > .20	0,974771	0,783707
X2: podil hrube frakce	centralni zastavba	24	0,174723	p < ,10	0,924437	0,073260
X3: podil jemne frakce	centralni zastavba	24	0,136194	p > .20	0,927926	0,087624
X4: vlhkost (v promile)	centralni zastavba	24	0,149222	p < ,20	0,945015	0,210778
X5: vyhrevnost (v kJ/kg)	centralni zastavba	24	0,134568	p > .20	0,957839	0,396561
X1: merna hmotnost	smisena zastavba	28	0,140229	p < ,15	0,945321	0,150917
X2: podil hrube frakce	smisena zastavba	28	0,130313	p > .20	0,954143	0,251352
X3: podil jemne frakce	smisena zastavba	28	0,149243	p < ,10	0,954354	0,254385
X4: vlhkost (v promile)	smisena zastavba	28	0,151968	p < ,10	0,937937	0,097953
X5: vyhrevnost (v kJ/kg)	smisena zastavba	28	0,160347	p < ,10	0,930071	0,061903

Komentář: Ani v jednom případě nebyla hypotéza o normalitě zamítnuta na hladině významnosti 0,05.

Úkol 3.: Na hladině významnosti 0,05 testujte hypotézu, že varianční matice proměnných X1, X2, X3, X4, X5 jsou v obou skupinách shodné.
Řešení: Statistiky – ANOVA – Jednofaktorová ANOVA – OK – Proměnné – Seznam závislých proměnných X1, X2, X3, X4, X5 - Kategor. nezávislá proměnná (faktor) ID – OK – OK – Více výsledků – záložka Předpoklady – Boxův M test

	Boxův M test (slozeni_komunalni_odpad.sta) Efekt: "ID" (Vypočteno pro všechny proměnné)			
	Boxovo M	Chí-kv.	SV	p
Boxovo M	19,96967	17,82128	15	0,272178

Komentář: p-hodnota je 0,2722, což je větší než 0,05, tedy dále budeme varianční matice pro centrální zástavbu a pro smíšenou zástavbu považovat za shodné.

Lze konstatovat, že důležité předpoklady vícerozměrného dvouvýběrového t-testu jsou splněny.

Úkol 4.: Na hladině významnosti 0,05 testujte hypotézu, že vektory středních hodnoty proměnných X1, X2, X3, X4, X5 jsou v obou skupinách shodné.
Řešení: Statistiky – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK – Proměnné – Závisle proměnné X1, X2, X3, X4, X5, Grupovací proměnná ID – OK – na záložce Možnosti zaškrtneme Vícerozměrný test (Hotellingovo T^2) – Výpočet

Proměnná	t-testy; grupováno:ID (slozeni_komunalni_odpad.sta) Skup. 1: centralni zastavba; Skup. 2: smisena zastavba Hotellingovo 79,1167 F(5,46)=14,557 p<,00000										
	Průměr centralni zastavba	Průměr smisena zastavba	t	sv	p	Poč.plat centralni zastavba	Poč.plat. smisena zastavba	Sm.odch. centralni zastavba	Sm.odch. smisena zastavba	F-poměr Rozptyly	p Rozptyly
X1	298,7500	384,2857	-4,95502	50	0,000009	24	28	53,0228	68,8223	1,684743	0,207528
X2	671,3750	572,5714	3,32653	50	0,001654	24	28	102,0672	110,6250	1,174718	0,699898
X3	90,5000	192,8571	-5,01506	50	0,000007	24	28	58,8565	83,7747	2,025989	0,089454
X4	269,2083	187,6429	4,92477	50	0,000010	24	28	59,3816	59,6733	1,009850	0,989295
X5	934,3333	766,0357	3,40703	50	0,001304	24	28	206,9185	148,0639	1,952991	0,096578

Komentář: Testová statistika vícerozměrného dvouvýběrového t-testu nabývá hodnoty 14,557, odpovídající p-hodnota je velmi blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu, že vektory středních hodnot proměnných X1, X2, X3, X4, X5 jsou v obou skupinách shodné. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že mezi centrální zástavbou a smíšenou zástavbou existuje rozdíl z hlediska složení komunálního odpadu.

Úkol 5.: Pomocí simultánních testů zjistěte, které složky vektorů středních hodnot proměnných X1, X2, X3, X4, X5 v centrální a smíšené zástavbě se liší na hladině významnosti 0,05.

Řešení: Simultánní testy založené na statistice $T_{0j} = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} \cdot \frac{(M_{1j} - M_{2j})^2}{S_{*j}^{\ 2}}$

STATISTICA neposkytuje. (V našem případě n = 52, p = 5, n_1 = 24, n_2 = 28, tedy $\frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} = \frac{30912}{1300}$.) S pomocí STATISTIKY však můžeme vypočítat vektory výběrových průměrů a směrodatných odchylek – viz tabulku v úkolu 4. V této tabulce ponecháme pouze proměnné obsahující průměry a směrodatné odchylky. Dále za poslední proměnnou vložíme dvě nové proměnné T_{0j} a kvantil. Do Dlouhého jména proměnné T_{0j} napíšeme:

=(30912/13000)*(v1-v2)^2/((23*v3^2+27*v4^2)/50)

Do Dlouhého jména proměnné kvantil napíšeme:

=VF(0,95;5;46)

Proměnná	t-testy; grupováno:ID (slozeni_komunalni_odpad.sta) Skup. 1: centralni zastavba; Skup. 2: smisena zastavba Hotellingovo 79,1167 F(5,46)=14,557 p<,00000					
	Průměr centralni zastavba	Průměr smisena zastavba	Sm.odch. centralni zastavba	Sm.odch. smisena zastavba	T0j =(30912/13	kvantil =VF(0,95;5
X1	298,7500	384,2857	53,0228	68,8223	4,51761518	2,41735604
X2	671,3750	572,5714	102,0672	110,6250	2,03610872	2,41735604
X3	90,5000	192,8571	58,8565	83,7747	4,62775957	2,41735604
X4	269,2083	187,6429	59,3816	59,6733	4,46261189	2,41735604
X5	934,3333	766,0357	206,9185	148,0639	2,13584049	2,41735604

Komentář: Vidíme, že statistiky T01, T03 a T04 se realizují v kritickém oboru $W = \langle 2{,}4174; \infty)$. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že centrální a smíšená zástavba se liší v měrné hmotnosti, podílu jemné frakce a vlhkosti komunálního odpadu.