

OBSAH

Charakteristika a význam Studentovho t-Testu

Vzorové údaje 1.2

Praktické použitie programu SAS (SAS Enterprise Guide)

Príklad 1.3 (SAS) - jednovýberový t-Test

Príklad 1.4 (SAS) - dvojjvýberový t-Test

Príklad 1.5 (SAS) - párový t-Test

Praktické použitie programu EXCEL (Microsoft 365)

Príklad 1.4 (Excel) - dvojjvýberový t-Test

Príklad 1.5 (Excel) - párový t-Test

Praktické použitie programu R (R Studio)

Príklad 1.6 (Program R) - jednovýberový t-Test

Príklad 1.7 (Program R) - dvojjvýberový t-Test

Príklad 1.8 (Program R) - párový t-Test

Ďalšie príklady použitia t-Testu

Zdroje a zoznam použitej literatúry

Charakteristika a význam Studentovho t-Testu

Štatistické hypotézy sú tvrdenia o jednotlivých parametroch a vlastnostiach základných súborov, ktoré môžeme overujeme štatistickými metódami na základe údajov z výberových súborov. Najčastejšie sa stretávame s hypotézami o parametroch, o pravdepodobnostnom rozdelení základného súboru a o ďalších vlastnostiach základných súborov. Hypotézy sa môžu týkať aj porovnávania parametrov, pravdepodobnostných rozdelení a vlastností dvoch, alebo viacerých základných súborov.

Dôležitým prvkom testovania hypotéz je overovanie ich platnosti. Testovaním potvrdzujeme, alebo zamietame stanovenú nulovú hypotézu H_0 oproti alternatívnej hypotézy H_1 . Každá odborná a vedecká činnosť by mala mať v rámci stanoveného cieľa jasne formulované základné hypotézy. K overeniu platnosti štatistických hypotéz najčastejšie používame štatistické testy.

Štatistické testy všeobecne rozdeľujeme na testy parametrické a testy neparametrické (existujú ale aj iné triedenia štatistických testov). Aplikácia každého parametrického štatistického testu predpokladá určité pravdepodobné rozdelenie základného súboru. Aplikácia neparametrického testu nepredpokladá pravdepodobné rozdelenie základného súboru.

Parametrické testy môžeme rozdeliť na jednovýberové, dvojvýberové a párové testy.

Jednovýberové parametrické testy sú:

- jednovýberové parametrické testy zhody pre strednú hodnotu normálneho rozdelenia so známym rozptylom,
- jednovýberové parametrické testy zhody pre strednú hodnotu normálneho rozdelenia s neznámym rozptylom,
- Jednovýberové parametrické testy zhody pre rozptyl normálneho rozdelenia.

Dvojvýberové parametrické testy sú:

- dvojvýberové parametrické testy zhody stredných hodnôt dvoch základných súborov s normálnym rozdelením so známym rozptylom,
- dvojvýberové parametrické testy zhody stredných hodnôt dvoch základných súborov s normálnym rozdelením s neznámym rozptylom (neznáme rozptyly sa môžu štatisticky preukazne líšiť a nelíšiť),
- dvojvýberové parametrické testy zhody rozptylov dvoch základných súborov s normálnym rozdelením (zhodu rozptylov dvoch základných súborov s normálnym rozdelením testujeme Fisherovým F- testom).

Párové testy

Párové testy predstavujú špeciálne testy pri ktorých testujeme (overujeme) štatistické hypotézy o dvoch navzájom možných súvisiacich znakov v rámci jedného základného súboru údajov (príkladom by mohla byť hmotnosť človeka pred a po absolvovaní diéty).

Medzi najviac používané parametrické testy patrí Studentov t-Test. Studentovo t- rozdelenie je symetrické okolo osi y ak dosahuje iba jeden vrchol. Studentovo t-rozdelenie sa dá aproximovať normálnym rozdelením ak počet stupňov voľnosti je väčší ako 30. Ak má náhodná premenná X normované normálne rozdelenie a náhodná premenná Y má Chí-kvadrát rozdelenie s n stupňami voľnosti a súčasne platí, že náhodné premenné X, Y sú nezávislé, potom premenná Z má Studentovo t- rozdelenie s n stupňami voľnosti.

Pre náhodnú premennú Z platí nasledovný vzťah:

$$Z = \frac{X}{\sqrt{Y/n}}$$

Studentov t-Test je štatistickou metódou, ktorá umožňuje overiť najčastejšie nasledujúce hypotézy:

1. hypotézu či normálne rozdelenie, z ktorého pochádza určitý náhodný výber má určitú konkrétnu hodnotu, pričom rozptyl premennej je neznámy,
2. hypotézu či dve normálne rozdelenia majúce rovnaký rozptyl, z ktorých sú dva nezávislé náhodné výbery, majú rovnaké stredné hodnoty, alebo rozdiel týchto stredných hodnôt je rovnaký presne stanovenej číselnej hodnote,
3. hypotézu či normálne rozdelenie, z ktorého pochádza určitý náhodný výber má nezávislé stredné hodnoty dvoch konkrétnych (párových) znakov.

V prvom prípade môže byť náhodný výber tvorený jednotlivými konkrétnymi hodnotami znaku, pre ktorý je vypočítaná stredná hodnota. Vtedy sa jedná o **jednovýberový t-Test**,

Pri ďalšej hypotéze sa jedná o **dvojvýberový t-Test**. Tento test je asi najznámejší a najviac používaný keď chceme zistiť či merania v jednej skupine (ich stredná hodnota) sú štatisticky významne odlišné od meraní (strednej hodnoty) v druhej skupine.

Náhodný výber v treťom prípade môže byť ale tvorený aj párovými dvojicami hodnôt, pri ktorých zisťujeme rozdiel ich stredných hodnôt. Takýto test sa nazýva **párový t-Test**.

Vzorové údaje 1.2 (zdroj: Databáza údajov projektu KEGA)

V rámci výučby predmetu Biometrika je vytváraný výberový súbor niektorých telesných mier študentov s ich základnými triediacimi znakmi, ktorý je aktualizovaný každý rok (dodržaný princíp anonymity, dobrovoľnosti s potvrdeným súhlasom jednotlivých študentov). Cieľom tejto úlohy je overiť 3 hypotézy: prvú hypotézu, skupina študentiek má hmotnosť presne 62 kilogramov a skupina študentov hmotnosť presne 80 kilogramov, druhú hypotézu, či existuje rozdiel v priemernej hodnote veku študentov podľa pohlavia a overiť tretiu hypotézu, či existuje štatisticky preukazný rozdiel 10 centimetrov medzi obvodom hrudníka a obvodom pása v triedení podľa pohlavia.

Súbor údajov **TM.xlsx** obsahuje nasledovné premenné:

PORADIE	poradové číslo záznamu
POHL	muž - M, žena - Z
VEK	vek v rokoch
ROK	rok merania
HMOT	hmotnosť (kg)
VYSKA	telesná výška (cm)
OBVODH	obvod hrudníka (cm)
OBVODP	obvod pása (cm)
OBVODZ	obvod zápästia (cm)
KRUZOK	študijný krúžok
BMI	index telesnej hmotnosti (Body Mass Index)
INDEXHP	relatívny index (pomer obvodu hrudníka a obvodu pása, %)

Vzor súboru údajov (10 záznamov):

PORADIE	POHL	VEK	ROK	HMOT	VYSKA	OBVODH	OBVODP	OBVODZ	KRUZOK	BMI	INDEXHP
1	Z	21	2008	52	160	90	70	15.0	1	20.3	128.6
2	Z	20	2008	54	174	86	68	16.0	1	17.8	126.5
3	Z	22	2008	57	174	83	63	16.0	1	18.8	131.7
4	Z	21	2008	62	168	100	76	15.0	1	22.0	131.6
5	M	21	2008	67	169	94	86	18.0	1	23.5	109.3
6	M	21	2008	72	192	94	89	17.0	1	19.5	105.6
7	M	20	2008	75	176	94	85	19.0	1	24.2	110.6
8	M	21	2008	81	187	100	87	21.0	1	23.2	114.9
9	M	22	2008	89	180	105	104	19.0	1	27.5	101.0
10	M	20	2008	65	186	97	78	18.0	3	18.8	124.4

Praktické použitie programu SAS (SAS Enterprise Guide)

Úlohy (Tasks): Analyze - ANOVA - t-Test

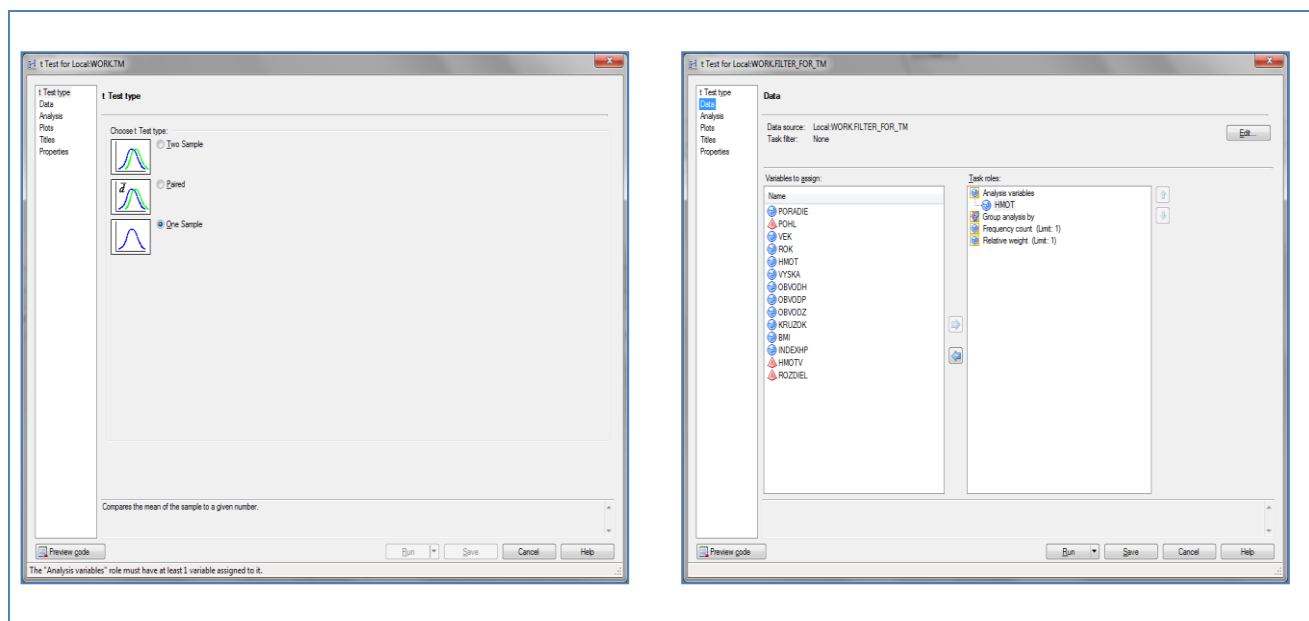
Príklad 1.3 (SAS) - jednovýberový t-Test

Pre overenie prvej hypotézy, či skupina študentov má hmotnosť presne 80 kilogramov a skupina študentiek hmotnosť presne 62 kilogramov, je užitočné vytvoriť si dva samostatné súbory podľa pohlavia študentov (využitie možnosti filtrovania a triedenia údajov v programe). Overenie hypotézy urobíme pre každé pohlavie samostatne. Úlohu je možné riešiť aj za celý hodnotený súbor spolu, ale v tom prípade je potrebné nastaviť skupinovú analýzu podľa pohlavia (analýzu treba uskutočniť ale 2x, spolu so správnym nastavením nulovej hypotézy v rámci pohlavia a z výsledkov si potom vyberieme iba príslušné pohlavie).

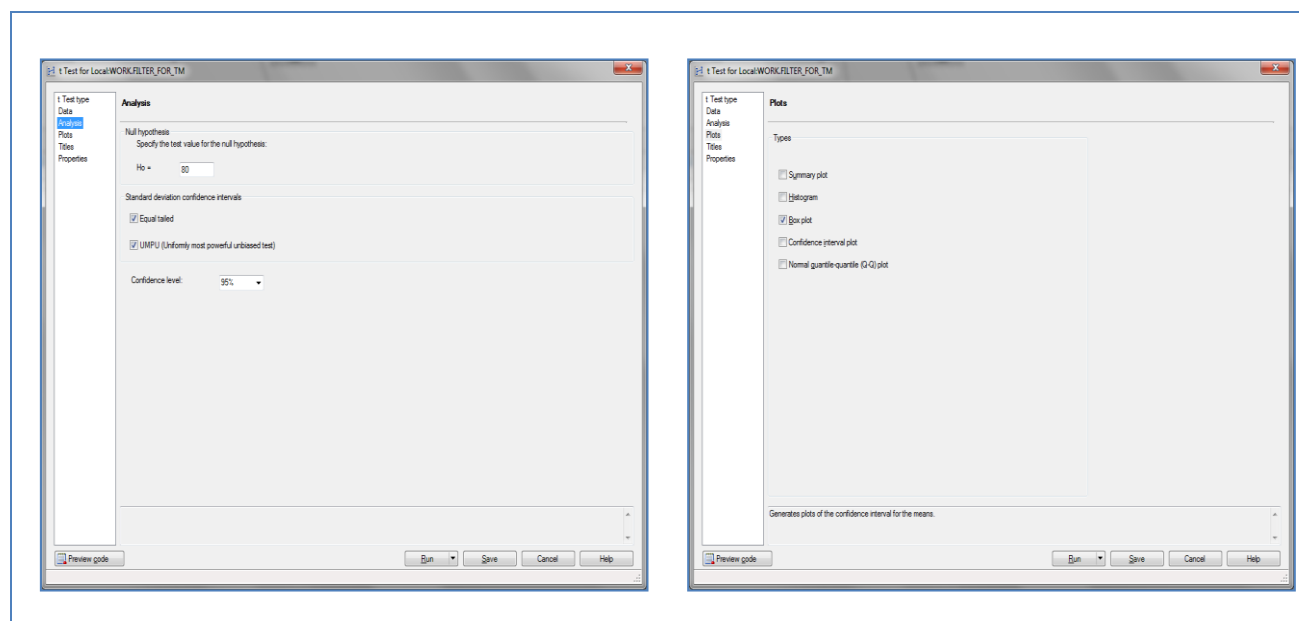
Použijeme **jednovýberový typ t-Testu** (One Sample). V základnom nastavení testu musí byť samozrejme, správne nastavená hodnota nulovej hypotézy 80 kilogramov v prípade študentov, 62 kilogramov v prípade študentiek a pre lepšie znázornenie výsledkov potvrdené nastavenie grafických možnosti výstupov samotného testu.

Zadaná úloha má nasledovné nastavenia a parametre:

Tab 2.1a Nastavenie analýzy - príklad 1.3 (zdroj: projekt KEGA, Candrák, 2021)



Tab 2.1b Nastavenie analýzy - príklad 1.3 (zdroj: projekt KEGA, Candrák, 2021)

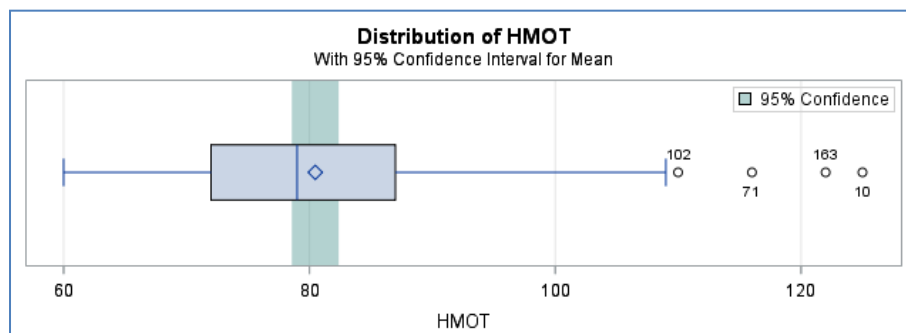


Výsledky jednovýberového t-Testu a ich interpretácia (pohlavie muži)

Tab 2.2 Výsledky t-Testu – muži (zdroj: projekt KEGA, Candrák, 2021)

t Test							
The TTEST Procedure							
Variable: HMOT							
N	Mean	Std Dev	Std Err	Minimum	Maximum		
164	80.4756	12.3266	0.9625	60.0000	125.0		
Mean	95% CL Mean	Std Dev	95% CL Std Dev	95% UMPU CL	Std Dev		
80.4756	78.5749 82.3763	12.3266	11.1213 13.8271	11.0994	13.7977		
DF	t Value	Pr > t					
163	0.49	0.6219					

Tab 2.3 Grafické znázornenie výsledkov – muži (zdroj: projekt KEGA, Candrák, 2021)



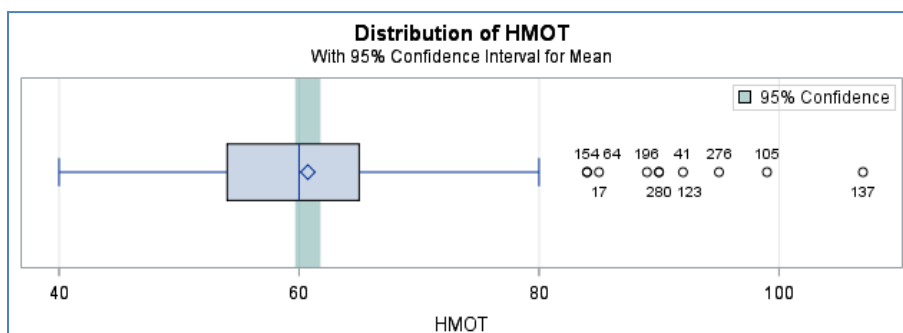
V prípade študentov, ktorí dosiahli priemernú hmotnosť 80,47 kilogramov **bola potvrdená nulová hypotéza**, že výberový priemer sledovaného znaku sa štatisticky nelíši od hodnoty 80 kilogramov ($p=0,6219$). Hodnota 80 kilogramov sa nachádza v rozpätí konfidenčných intervalov pri stanovenej spoľahlivosti 95 %.

Výsledky jednovýberového t-Testu a ich interpretácia (pohlavie ženy)

Tab 2.4 Výsledky t-Testu – ženy (zdroj: projekt KEGA, Candrák, 2021)

t Test						
The TTEST Procedure						
Variable: HMOT						
N	Mean	Std Dev	Std Err	Minimum	Maximum	
325	60.7292	9.6846	0.5372	40.0000	107.0	
Mean	95% CL Mean	Std Dev	95% CL Std Dev	95% UMPU CL	Std Dev	
60.7292	59.6724 61.7861	9.6846	8.9929 10.4924	8.9838	10.4813	
DF	t Value	Pr > t				
324	-2.37	0.0186				

Tab 2.5 Grafické znázornenie výsledkov – ženy (zdroj: projekt KEGA, Candrák, 2021)



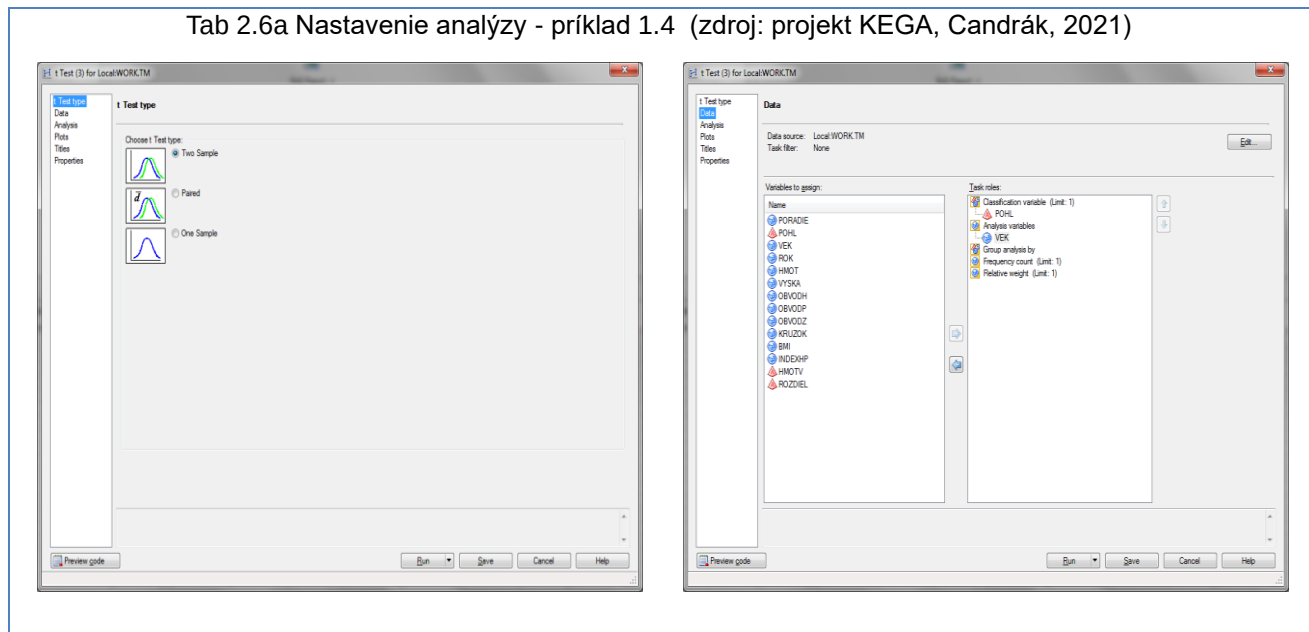
V prípade študentiek, ktoré dosiahli priemernú hmotnosť 60,72 kilogramov **nebola potvrdená nulová hypotéza**, že výberový priemer sledovaného znaku sa štatisticky nelíši od hodnoty 62 kilogramov. **Platí preto jednotková hypotéza, že výberový priemer sa štatisticky líši od hodnoty 62 kilogramov** ($p=0,0186$). Hodnota 62 kilogramov, nastavená v nulovej hypotéze, sa nenachádza v rozpätí konfidenčných intervalov pri stanovenej spoľahlivosti 95 %. Konfidenčný interval dosiahol pri študentkách hodnoty od 59,6724 do 61,7861 kilogramov.

Príklad 1.4 (SAS) - dvojvýberový t-Test

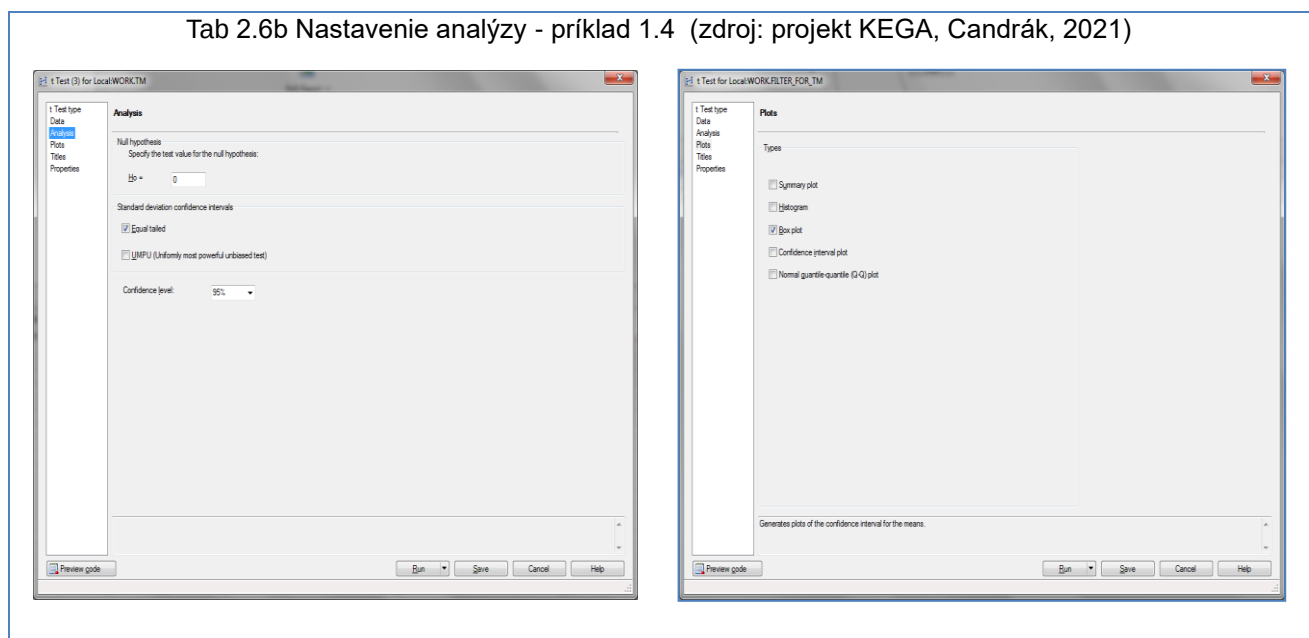
Overte hypotézu druhú hypotézu, či existuje rozdiel v priemernej hodnote veku študentov podľa pohlavia (**dvojvýberový t-Test, Two sample**). Analyzovaná premenná je vek a triediace kritérium je pohlavie študenta.

Hodnota nulovej hypotézy (rozdiel v priemernej hodnote veku je 0), ak predpokladáme, že neexistuje rozdiel medzi strednými hodnotami, alebo môžeme ju postaviť aj na nejakej predpokladanej konkrétnej hodnote rozdielu medzi dvoma skupinami študentov. Nastavenie niektorých parametrov a grafických výstupov môže byť identické ako to bolo v prípade jednovýberového T-testu.

Tab 2.6a Nastavenie analýzy - príklad 1.4 (zdroj: projekt KEGA, Candrák, 2021)



Tab 2.6b Nastavenie analýzy - príklad 1.4 (zdroj: projekt KEGA, Candrák, 2021)



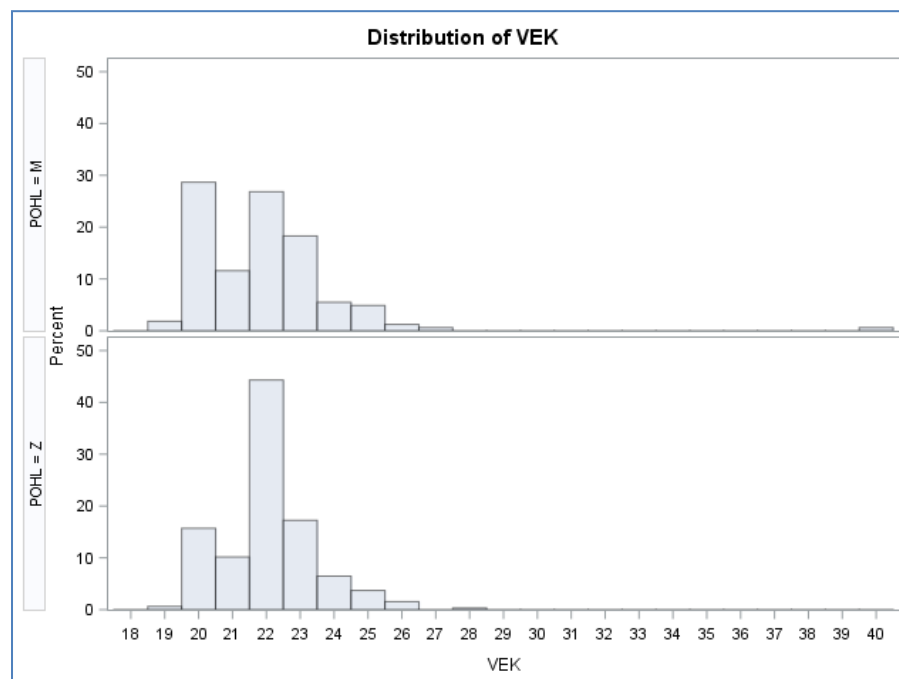
Výsledky dvojvýberového t-Testu a ich interpretácia

Základom správneho porovnania dvoch hodnotených skupín na základe dvojvýberového t-Testu je splnenie nasledovných predpokladov: nezávislosť pozorovaní, predpoklad normality údajov v rámci jednotlivých skupín (často stačí približná normalita) a predpoklad rovnakého rozptylu pre každú skupinu (ak nie je dodržaný tak môžeme uplatniť iba aproximatívny - približný t-Test).

Podmienka nezávislosti pozorovaní bola splnená a dodržiavaná už pri samotnom zhromažďovaní údajov o telesných mierach študentov. Problém by mohol nastať iba v prípade, ak by medzi študentami a študentkami bolo príliš veľa súrodeneckých vzťahov vo forme dvojčiat, čo by mohlo narušiť podmienku nezávislosti údajov. Nami zhromažďované údaje ale takýto problém nemajú a tak ich môžeme v rámci skupín aj medzi skupinami považovať za nezávislé.

Približne splnenú podmienku normality údajov dvoch skupín študentov ukazuje grafické znázornenie distribúcie veku študentov podľa pohlavia, ktoré môžeme získať pri výpočte základných štatistických parametrov v rámci opisnej (sumárnej) štatistiky.

Tab 2.7 Grafické znázornenie výsledkov – dvojvýberový test (zdroj: projekt KEGA, Candrák, 2021)



Dodržanie predpokladu približne rovnakej hodnoty rozptylu analyzovaného znaku, ako tretej podmienky, nachádzame vo výsledkoch vo forme F-testu. Preukaznosť tohto testu dáva možnosť

správneho výberu ponúkaných dvoch typov a hodnôt t-Testu (test pri rovnakom rozptyle znaku a test pri nerovnakom rozptyle znaku).

Spravidla ak je nulová hypotéza o rovnosti rozptylov pravdivá hodnota F-testu je veľmi blízko hodnote 1. Ak zamietnete nulovú hypotézu o rovnosti rozptylov, použijeme na testovanie rovnosti skupinových priemerov T-test pre nerovnaký rozptyl.

Samozrejme, ale stále platí, že tento test je platný iba pre nezávislé vzorky z normálneho rozdelenia. Dodržanie približnej normality sa vyžaduje aj pri analýzach veľkých vzoriek údajov.

V riešenej úlohe sme získali nasledovné detailné výsledky:

Tab 2.8 Detailné výsledky – dvojitý test (zdroj: projekt KEGA, Candrák, 2021)

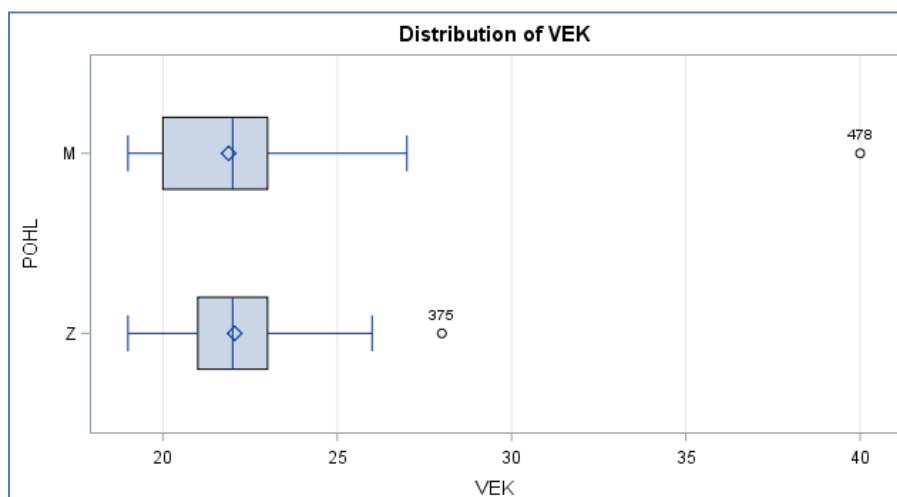
t Test						
The TTEST Procedure						
Variable: VEK						
POHL	N	Mean	Std Dev	Std Err	Minimum	Maximum
M	164	21.8841	2.1462	0.1676	19.0000	40.0000
Z	325	22.0585	1.3810	0.0766	19.0000	28.0000
Diff (1-2)		-0.1743	1.6764	0.1606		

POHL	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev
M		21.8841	21.5532 22.2151	2.1462	1.9363 2.4074
Z		22.0585	21.9078 22.2092	1.3810	1.2823 1.4962
Diff (1-2)	Pooled	-0.1743	-0.4898 0.1412	1.6764	1.5774 1.7888
Diff (1-2)	Satterthwaite	-0.1743	-0.5374 0.1887		

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	487	-1.09	0.2782
Satterthwaite	Unequal	233.11	-0.95	0.3451

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	163	324	2.42	<.0001

Tab 2.9 Grafické výsledky – dvojitý test (zdroj: projekt KEGA, Candrák, 2021)



Výsledky obsahujú základné štatistické charakteristiky veku študentov podľa pohlavia a základné štatistické charakteristiky rozdielu veku medzi študentmi a študentkami. Priemerný rozdiel

vo veku bol zistený -0,1743 roka, čo predstavuje, že študenti sú o túto hodnotu v priemere mladší ako študentky.

Konfidenčný interval pri stanovenej hranici spoľahlivosti 95 % sa pohybuje v rozpätí od -0,5374 do +0,1887 čo znamená, že priemerný rozdiel sa nachádza v tomto intervale.

Hodnota F-testu 2.42 ($p < 0,0001$) potvrdzuje, že hypotéza o rovnosti rozptylov je neplatí a hodnoty rozptylov znaku v dvoch skupinách študentov sú štatisticky vysoko rozdielne.

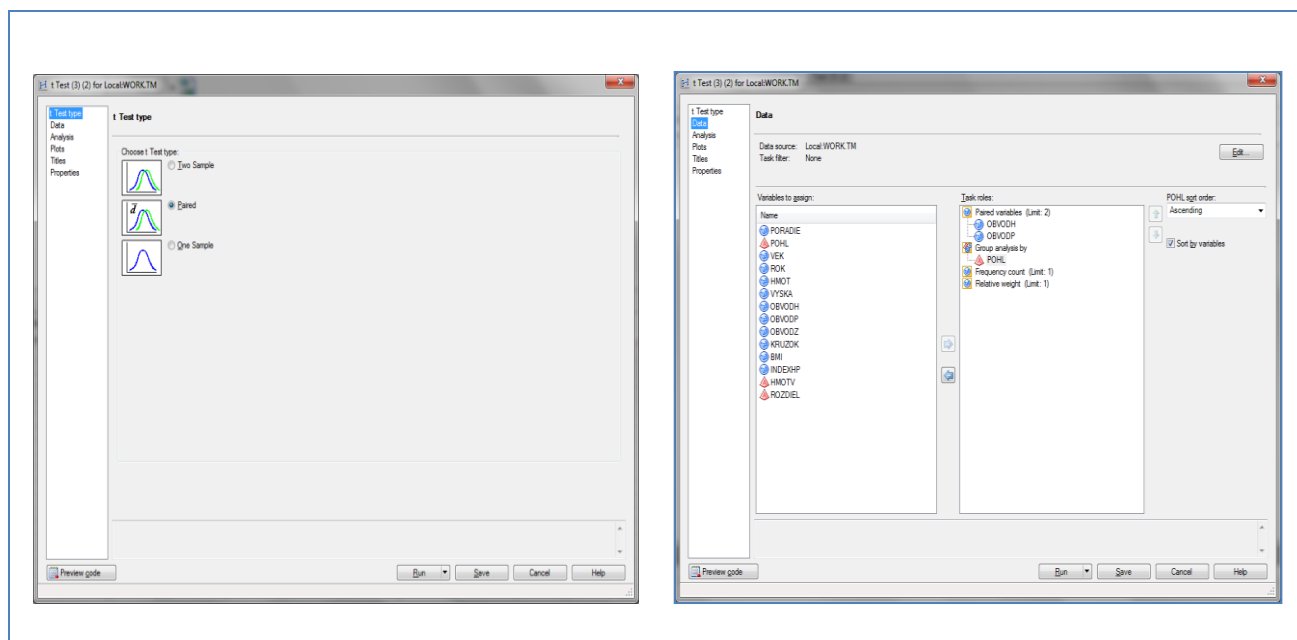
Z uvedeného vyplýva, že pre hlavnú interpretáciu výsledkov dvojvýberového t-Testu použijeme hodnotu t-Testu a jeho preukaznosti pre nerovnaké rozptyly (Satterthwaite, Unequal). Hodnota tohto t-Testu dosiahla hodnotu -0,95, pravdepodobnosť $p=0,3451$, čo **potvrdzuje nulovú hypotézu o rovnosti priemerných hodnôt veku študentov a študentiek**. Rozdiel priemerných hodnôt preto pokladáme prakticky za nulový, čiže neexistuje medzi nimi štatistická preukaznosť.

Príklad 1.5 (SAS) - párový t-Test

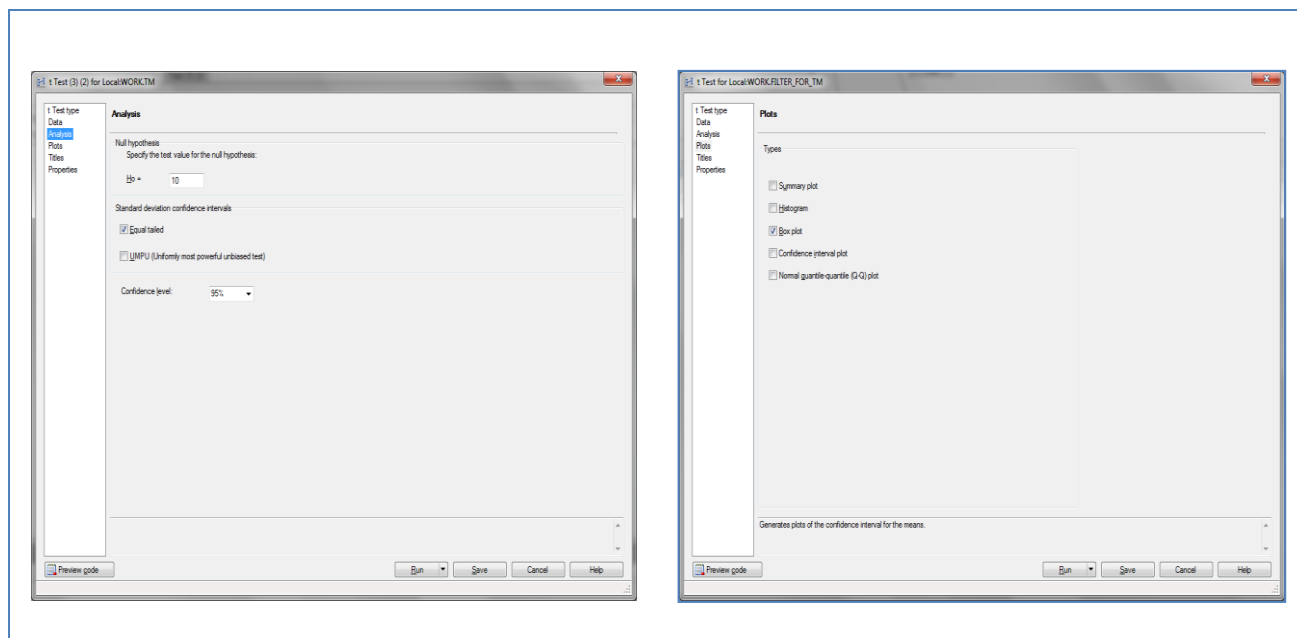
Overte tretiu hypotézu, či existuje štatisticky preukazný rozdiel 10 centimetrov medzi obvodom hrudníka a obvodom pása v triedení podľa pohlavia.

Nastavenie úlohy pre overenie tretej hypotézy či existuje štatisticky preukazný rozdiel 10 centimetrov medzi obvodom hrudníka a obvodom pása v triedení podľa pohlavia (**párový t-Test, Paired**):

Tab 2.10a Nastavenie analýzy - príklad 1.5 (zdroj: projekt KEGA, Candrák, 2021)



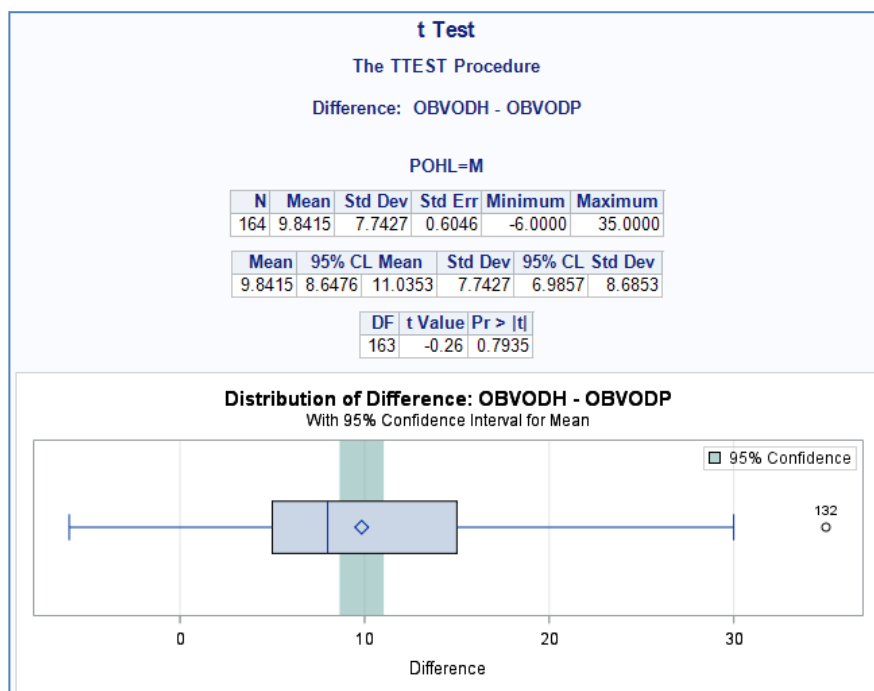
Tab 2.10b Nastavenie analýzy - príklad 1.5 (zdroj: projekt KEGA, Candrák, 2021)



Nastavenie grafických výstupov môže byť znovu identické ako to bolo v prípade predchádzajúcich t-Testov.

Výsledky párového t-Testu a ich interpretácia (pohlavie muži)

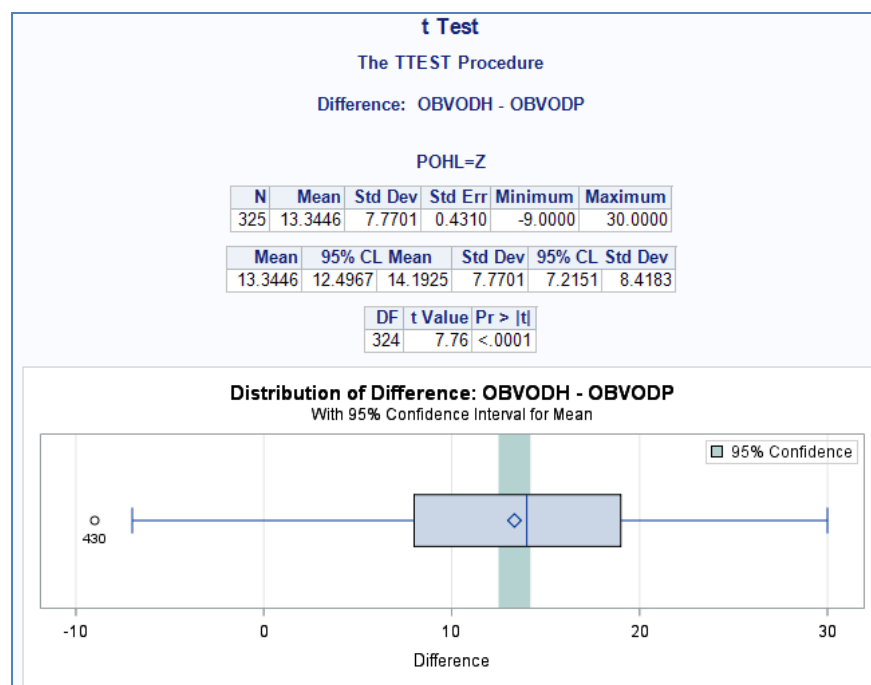
Tab 2.11 Detailné výsledky – párový test muží (zdroj: projekt KEGA, Candrák, 2021)



Na základe výsledkov párového t-Testu rozdielu priemerných hodnôt obvodu hrudníka a obvodu pása študentov (+9,8415 cm) sa **pri stanovenej nulovej hypotéze (+10 cm) táto hypotéza potvrdila** (hypotézu nezamietame). Hodnota t-Testu dosiahla -0,26 ($p=0,7935$), čo potvrdzuje, že skutočný rozdiel priemerných hodnôt obvodu hrudníka a obvodu pása sa **štatisticky vôbec nelíši** od hodnoty +10 centimetrov. Hodnota 10 centimetrov sa nachádza v rozpätí skutočne zistených konfidenčných intervalov rozdielu hodnôt (číselné aj grafické vyjadrenie).

Výsledky párového t-Testu a ich interpretácia (pohlavie žena)

Tab 2.12 Detailné výsledky – párový test ženy (zdroj: projekt KEGA, Candrák, 2021)



Na základe výsledkov párového t-Testu rozdielu priemerných hodnôt obvodu hrudníka a obvodu pása študentiek (+13,3446 cm) sa **pri stanovenej nulovej hypotéze (+10 cm) táto hypotéza nepotvrdila** (hypotézu zamietame). Hodnota t-Testu dosiahla 7,76 ($p<0,0001$), čo potvrdzuje, že skutočný rozdiel priemerných hodnôt obvodu hrudníka a obvodu pása sa **vysoko štatisticky líši** od hodnoty +10 centimetrov. Hodnota 10 centimetrov sa nenachádza v rozpätí skutočne zistených konfidenčných intervalov rozdielu hodnôt. Konfidenčné intervaly dosiahli hodnoty od 12,4967 do 14,1925), čo názorne potvrdzuje aj grafické vyjadrenie výsledkov analýzy.

Použitie t-Testu má ale isté obmedzenia pokiaľ chcem navzájom porovnávať priemerné hodnoty vo viac ako dvoch skupinách nie je možné dvojvýberový t-Test použiť. V takomto prípade je najvhodnejším štatistickým testom použitie F-testu v rámci jednofaktorovej analýzy rozptylu (jednofaktorová ANOVA, One-Way-Anova). Tento druh F-testu je v skutočnosti zovšeobecnením t-Testu pre porovnanie stredných hodnôt dvoch a viacerých skupín. Analýza rozptylu (ANOVA) je všeobecne štatistická metóda používaná na porovnanie priemerov a zistenie štatistických preukazností vplyvu dvoch alebo viacerých skupín výberových pozorovaní.

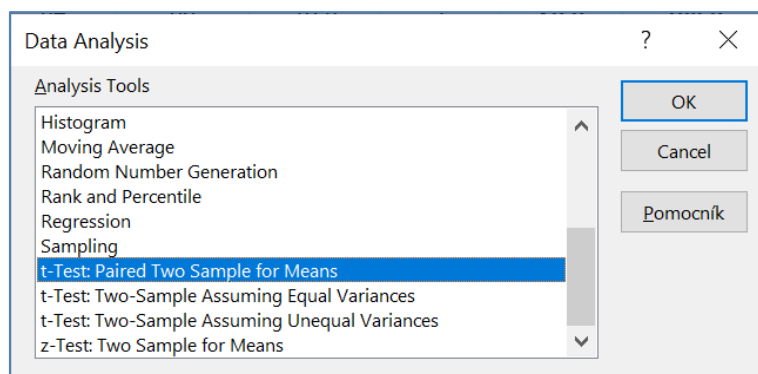
Použitie jednofaktorovej analýzy rozptylu predpokladá splnenie veľmi podobných predpokladov ako to je pri dvojvýberovom t-Teste.

Analýza rozptylu (ANOVA) analyzuje rozptyly údajov, tak aby sme mohli zistiť, či existuje rozdiel medzi priemerami jednotlivých skupín. Analýza pracuje s rozptylom analyzovaného znaku v rámci skupín, s rozptylom medzi skupinami a s celkovým rozptylom znaku. Podrobnejšie vysvetlenie tejto analýzy, konkrétne príklady riešenia a interpretáciu praktického uplatnenia najdeme v samostatnej časti Modulu 1.

Praktické použitie programu EXCEL (Microsoft 365)

V rámci doplnkov programu Microsoft Excel je možné využiť doplnok analytických nástrojov **Data Analysis**, ktorý umožňuje použiť párový t-Test a dvojvýberový t-Test s predpokladom rovnakého rozptylu resp. nerovnakého rozptylu medzi dvoma hodnotenými skupinami.

Tab 2.13 Použitie analýzy údajov v programe Excel



Použitie t-Testu je ale možné aj s pomocou priameho vkladania príslušných štatistických funkcií v rámci pracovných údajových hárkov.

Funkcia t-Test vráti pravdepodobnosť spojenú s t-Testom. V skutočnosti určí, či dve vzorky pravdepodobne pochádzajú z rovnakých dvoch súborov, ktoré majú rovnakú strednú hodnotu.

T.TEST(rozsah1; rozsah2; strany; typ) zdroj: Excel pre Microsoft 365 (Podpora pre Office)

rozsah1 – prvá vzorka údajov, alebo skupina buniek pre T-test

rozsah2 – druhá vzorka údajov, alebo skupina buniek pre T-test

strany – počet strán rozdelenia, 1 jednostranné rozdelenie, 2 obojstranné rozdelenie

typ – typ t-testu, 1 párový test, 2 dvojvýberový test s rovnosťou rozptylov, 3 dvojvýberový test s nerovnosťou rozptylov

Príklad 1.4 (Excel) - dvojvýberový t-Test

(Excel - Data Analysis - t-Test: Two-Samples Assuming Unequal Variances)

Použite **dvojvýberového t-Testu** pri overení hypotézy či existuje štatisticky preukazný rozdiel v priemernej hodnote obvodu zápästia podľa pohlavia študentov. Overujeme nulovú hypotézu, že priemerný rozdiel v obvodoch zápästia podľa pohlavia je 2 cm.

V rámci tejto úlohy máme dva výberové súbory: súbor študentov a súbor študentiek. Z praktického dôvodu je preto potrebné najskôr usporiadať (triediť) celý východiskový súbor podľa pohlavia študentov.

Výsledky (Excel)

Pred samotným použitím dvojvýberového t-Testu musíme ale najskôr pomocou F-testu zistiť aký rozptyl majú dva výberové súbory. Analýza údajov v programe Excel to umožňuje pomerne jednoducho:

Tab 2.16 Nastavenie F-testu v programe Excel

Tab 2.17 Výsledky F-testu testu v programe Excel

F-Test Two-Sample for Variances		
	OBVODZ_M	OBVODZ_Z
Mean	17.7439024	15.348
Variance	1.77143498	1.1712074
Observations	164	325
df	163	324
F	1.51248615	
P(F<=f) one-tail	0.00090764	
F Critical one-tail	1.24496982	

Výsledky F-testu dokazujú, že dva výberové súbory nemajú rovnaký rozptyl obvodu zápästia študentov ($F\text{-test} = 1,5124$ a hodnota $p = 0,0009$) a preto pri dvojvýberovom t-Teste musíme použiť test s nerovnakým rozptylom.

Tab 2.18 Výsledky párového testu v programe Excel

t-Test: Two-Sample Assuming Unequal Variances		
	OBVODZ_M	OBVODZ_Z
Mean	17.7439024	15.348
Variance	1.77143498	1.1712074
Observations	164	325
Hypothesized Mean Difference	2	
df	275	
t Stat	3.29859743	
P(T<=t) one-tail	0.00054985	
t Critical one-tail	1.65041343	
P(T<=t) two-tail	0.00109971	
t Critical two-tail	1.96862787	

Na základe získaných výsledkov dvojvýberového t-Testu sa nulová hypotéza nepotvrdila. Hodnota t-Testu je 3,2985 a hodnota $p = 0,00109$. Priemerný rozdiel obvodu zápästia je preto preukazne rozdielny od nášeho predpokladu (hodnoty 2 cm).

Príklad 1.5 (Excel) - párový t-Test

(Excel - Data Analysis - t-Test: Paired Two Sample for Means)

Použite **párového t-Testu** pri overení hypotézy či existuje štatisticky preukazný rozdiel medzi obvodom hrudníka a obvodom pásu v rámci celého hodnoteného súboru telesných mier študentov. Nulová hypotéza je, že rozdiel medzi obvodom hrudníka a obvodom pásu je 0.

Tab 2.14 Detailné nastavenie analýzy údajov v programe Excel

t-Test: Paired Two Sample for Means

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

☒ Labels

Alpha:

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

OK Cancel Pomocník

Výsledky (Excel)

Výsledky testu sa nám po správnom zadaní rozsahu analyzovaných znakov, stanovenej nulovej hypotéze a štatistickej preukaznosti 95 % (Alpha=0,05) následne môžu uložiť do nového hárku, alebo do presne určeného miesta aktívneho hárku. V rámci príkladu sme zvolili uloženie výsledkov do samostatného hárku.

Tab 2.15 Výsledky párového testu v programe Excel

t-Test: Paired Two Sample for Means		
	OBVODH	OBVODP
Mean	93.09816	80.92843
Variance	87.33871	144.8781
Observations	489	489
Pearson Correlation	0.752823	
Hypothesized Mean Difference	0	
df	488	
t Stat	33.94549	
P(T<=t) one-tail	7.3E-131	
t Critical one-tail	1.647982	
P(T<=t) two-tail	1.5E-130	
t Critical two-tail	1.964837	

Na základe výsledkov existuje preukazný rozdiel medzi priemernými hodnotami obvodu hrudníka a obvodu pásu v rámci celého hodnoteného súboru telesných mier študentov. Nulová hypotéza, že rozdiel medzi obvodom hrudníka a obvodom pásu je 0 sa nepotvrdila a preto musíme prijať alternatívnu hypotézu, že rozdiel je preukazne odlišný od hodnoty 0.

Praktické použitie programu R (R Studio)

Príklad importu súboru TM.xlsx a zobrazenie údajov v rámci prostredia programu R Studio:

Import Excel Data

File/URL:
C:/Kega 2019-2021/Modul1/TM.xlsx Browse...

Data Preview:

PORADIE (double)	POHL (character)	VEK (double)	ROK (double)	HMOT (double)	VYSKA (double)	OBVODH (double)	OBVODP (double)	OBVODZ (double)	KRUZOK (double)	BMI (double)
1	Z	21	2008	52	160	90	70	15	1	20.1
2	Z	20	2008	54	174	86	68	16	1	17.4
3	Z	22	2008	57	174	83	63	16	1	18.1
4	Z	21	2008	62	168	100	76	15	1	21.1
5	M	21	2008	67	169	94	86	18	1	23.1
6	M	21	2008	72	192	94	89	17	1	19.1

Previewing first 50 entries.

Import Options:

Name: TM Max Rows: ☒ First Row as Names
 Sheet: Data_TM Skip: 0 ☒ Open Data Viewer
 Range: A1:D10 NA:

Code Preview:

```
library(readxl)
TM <- read_excel("C:/Kega 2019-2021/Modul1/TM.xlsx",
  sheet = "Data_TM")
View(TM)
```

Reading Excel files using readxl Import Cancel

Zdrojový kód potrebný na import a zobrazenie excelovského súboru "TM.xlsx", ktorý je možné použiť priamo v programe R, alebo R Studio ako súbor s názvom **TM**:

```
# Import a zobrazenie údajov (databáza projektu KEGA)
library(readxl)
TM <- read_excel("C:/Kega 2019-2021/Modul1/TM.xlsx", sheet = "Data_TM")
View(TM)
```

Zdrojový kód potrebný na import a zobrazenie textového súboru "SMen.txt", ktorý je možné použiť priamo v programe R ako súbor s názvom **SMen**:

```
# Import a zobrazenie údajov (databáza projektu KEGA)
library(readxl)
SMen <- read_delim("C:/Pkhof/10-2021/SMen.txt",
  ";", escape_double = FALSE, locale = locale(encoding = "WINDOWS-1252"),
  trim_ws = TRUE)
View(SMen)
```

Zdrojový kód potrebný na import a zobrazenie súboru vo formáte programu SAS "ukazovatel.sas7bda"t, ktorý je možné použiť priamo v programe R ako súbor s názvom **ukazovatel**:

```
# Import a zobrazenie údajov (databáza projektu KEGA)
library(haven)
ukazovatel <- read_sas("C:/Biometrika/ZS-2021-2022/Data/ukazovatel.sas7bdat", NULL)
View(ukazovatel)
```

Program R umožňuje importovať aj mnoho ďalších typov súborov, prakticky skoro, so všetkých najčastejšie používaných programov, aplikácií, databázových a webovských systémov.

Príklad 1.6 (Program R) - jednovýberový t-Test

Overte dve nulové hypotézy, či priemerná hmotnosť študentov má konkrétnu hodnotu 68 kilogramov resp. 69 kilogramov (95 % štatistická pravdepodobnosť).

Použitie jednovýberového t-Testu

súbor: TM, ukazovatel HMOT, nulová hypotéza HMOT = 68, resp. HMOT =69

```
t.test(TM$HMOT, mu=68)
t.test(TM$HMOT, mu=69)
```

Výsledok jednovýberového t-Testu

súbor: TM, ukazovatel HMOT, nulová hypotéza HMOT = 68

One Sample t-test

data: TM\$HMOT

t = -1.0134, df = 488, **p-value = 0.3114**

alternative hypothesis: true mean is not equal to 68

95 percent confidence interval:

66.09481 68.60867

sample estimates:

mean of x 67.35174

V tomto prípade bola potvrdená nulová hypotéza. Priemerná hmotnosť všetkých študentov 67.35174 kilogramov, pri stanovenej hladine spoľahlivosti 95 %, nie je rozdielna hodnotu 68 kilogramom.

Výsledok jednovýberového t-Testu# súbor: TM, ukazovateľ HMOT, nulová hypotéza $HMOT = 69$

One Sample t-test

data: TM\$HMOT

 $t = -2.5766$, $df = 488$, **p-value = 0.01027**

alternative hypothesis: true mean is not equal to 69

95 percent confidence interval:

66.09481 68.60867

sample estimates:

mean of x 67.35174

V tomto prípade bola nulová hypotéza zamietnutá a preto platí alternatívna hypotéza. Priemerná hmotnosť všetkých študentov (67.35174 kilogramov) je pri stanovenej hladine spoľahlivosti 95 % štatisticky rozdielna hodnote 69 kilogramov.

Príklad 1.7 (Program R) - dvojitýberový t-Test

Overte nulovú hypotézu, či rozdiel v priemernej hmotnosti študentov a študentiek má hodnotu 0 (priemerné hodnoty sú približne rovnaké) .

Overte nulovú hypotézu, či rozdiel v priemernej hmotnosti študentov a študentiek má hodnotu 20 (priemerné hodnoty sú výrazne rozdielne).

Použitie dvojitýberového t-Testu

súbor: TM, ukazovateľ HMOT, dve skupiny podľa pohlavia,

nulové hypotézy rozdielu priemerných hodnôt = 0, resp. 20 kilogramov.

with (TM, t.test (HMOT [POHL == "M"], HMOT [POHL == "Z"]))

with (TM, t.test (HMOT [POHL == "M"], HMOT [POHL == "Z"], mu = 20))

Výsledok dvojitýberového t-Testu

súbor: TM, ukazovateľ HMOT, dve skupiny podľa pohlavia,

nulová hypotéza rozdielu priemerných hodnôt = 0.

Welch Two Sample t-test

data: HMOT[POHL == "M"] and HMOT[POHL == "Z"]

 $t = 17.914$, $df = 267.31$, **p-value < 2.2e-16**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: 17.57608 21.91668

sample estimates:

mean of x mean of y 80.47561 60.72923

Nulová hypotéza bola zamietnutá a preto platí alternatívna hypotéza. Priemerná hmotnosť študentov a študentiek nie je rovnaká (jej rozdiel sa nerovná hodnote 0, na stanovenej hladine štatistickej spoľahlivosti 95 %).

Výsledok dvojvýberového t-Testu

súbor: TM, ukazovateľ HMOT, dve skupiny podľa pohlavia,
nulová hypotéza rozdielu priemerných hodnôt = 20.

Welch Two Sample t-test

data: HMOT[POHL == "M"] and HMOT[POHL == "Z"]

t = -0.23008, df = 267.31, **p-value = 0.8182**

alternative hypothesis: true difference in means is not equal to 20

95 percent confidence interval: 17.57608 21.91668

sample estimates:

mean of x mean of y 80.47561 60.72923

Potvrdila sa nulová hypotéza, že rozdiel priemernej hmotnosti študentov a študentiek má hodnotu približne 20 (jej rozdiel nie je štatisticky preukazný oproti hodnote 20, štatistická spoľahlivosť 95 %).

Príklad 1.8 (Program R) - párový t-Test

Overte hypotézy, že rozdiel priemerných hodnôt obvodu hrudníka a obvodu pásu študentov sa rovná hodnote 0, resp. hodnote 12 centimetrov.

Použitie párového t-Testu

súbor: TM, ukazovatele OBVODH, OBVODP, nulové hypotézy rozdielu priemerných hodnôt = 0, resp. 12

t.test(Pair(OBVODH, OBVODP) ~ 1, data = TM)

t.test(Pair(OBVODH, OBVODP) ~ 1, data = TM, mu = 12)

Výsledok párového t-Testu

súbor: TM, ukazovatele OBVODH, OBVODP, nulová hypotéza rozdielu priemerných hodnôt = 0

Paired t-test

data: Pair(OBVODH, OBVODP)

t = 33.945, df = 488, **p-value < 2.2e-16**

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: 11.46532 12.87414

sample estimates:

mean of the differences 12.16973

Nulová hypotéza bola zamietnutá. Platí alternatívna hypotéza. Rozdiel priemerných hodnôt dvoch analyzovaných telesných mier sa nerovná 0. Rozdiel 12,16973 cm je vysoko štatisticky preukazný (štatistická spoľahlivosť 95 %).

Výsledok párového t-Testu

súbor: TM, ukazovatele OBVODH, OBVODP, nulová hypotéza rozdielu priemerných hodnôt = 12

Paired t-test

data: Pair(OBVODH, OBVODP)

t = 0.47345, df = 488, **p-value = 0.6361**

alternative hypothesis: true difference in means is not equal to 12

95 percent confidence interval: 11.46532 12.87414

sample estimates:

mean of the differences 12.16973

Nulová hypotéza bola potvrdená. Zistený rozdiel priemerných hodnôt dvoch analyzovaných telesných nie je významne rozdielny od hodnoty 12 (skutočný rozdiel 12,16973 cm nie je štatisticky preukazný oproti predpokladanej hodnote 12 centimetrov, štatistická spoľahlivosť 95 %).

Ďalšie príklady použitia t-Testu v programe R

(zdroj: R Documentation, Documentation for package 'stats' version 4.2.0).

```
require(graphics)
```

```
t.test(1:10, y = c(7:20))    # P = .00001855
```

```
t.test(1:10, y = c(7:20, 200)) # P = .1245    -- NOT significant anymore
```

```
## Classical example: Student's sleep data
```

```
plot(extra ~ group, data = sleep)
```

```
## Traditional interface
```

```
with(sleep, t.test(extra[group == 1], extra[group == 2]))
```

```
## Formula interface
```

```
t.test(extra ~ group, data = sleep)
```

```
## Formula interface to one-sample test
```

```
t.test(extra ~ 1, data = sleep)
```

```
## Formula interface to paired test
```

```
## The sleep data are actually paired, so could have been in wide format:
```

```
sleep2 <- reshape(sleep, direction = "wide",
```

```
                  idvar = "ID", timevar = "group")
```

```
t.test(Pair(extra.1, extra.2) ~ 1, data = sleep2)
```

Tab. 2.16 Kritické hodnoty $t(k)$ α Studentovho t – rozdelenia

α k	0,10	0,05	0,02	0,01	α k	0,10	0,05	0,02	0,01
1	6,314	12,706	31,821	63,657	38	1,686	2,024	2,429	2,712
2	2,920	4,303	6,965	9,925	39	1,685	2,023	2,426	2,708
3	2,353	3,182	4,541	5,841	40	1,684	2,021	2,423	2,704
4	2,132	2,776	3,747	4,604	41	1,683	2,020	2,421	2,701
5	2,015	2,571	3,365	4,032	42	1,682	2,018	2,418	2,698
6	1,943	2,447	3,143	3,707	43	1,681	2,017	2,416	2,695
7	1,895	2,365	2,998	3,499	44	1,680	2,015	2,414	2,692
8	1,860	2,306	2,896	3,355	45	1,679	2,014	2,412	2,690
9	1,833	2,262	2,821	3,250	46	1,679	2,013	2,410	2,687
10	1,812	2,228	2,764	3,169	47	1,678	2,012	2,408	2,685
11	1,796	2,201	2,718	3,106	48	1,677	2,011	2,407	2,682
12	1,782	2,179	2,681	3,055	49	1,677	2,010	2,405	2,680
13	1,771	2,160	2,650	3,012	50	1,676	2,009	2,403	2,678
14	1,761	2,145	2,624	2,977	51	1,675	2,008	2,402	2,676
15	1,753	2,131	2,602	2,947	52	1,675	2,007	2,400	2,674
16	1,746	2,120	2,583	2,921	53	1,674	2,006	2,399	2,672
17	1,740	2,110	2,567	2,898	54	1,674	2,005	2,397	2,670
18	1,734	2,101	2,552	2,878	55	1,673	2,004	2,396	2,668
19	1,729	2,093	2,539	2,861	56	1,673	2,003	2,395	2,667
20	1,725	2,086	2,528	2,845	57	1,672	2,002	2,394	2,665
21	1,721	2,080	2,518	2,831	58	1,672	2,002	2,392	2,663
22	1,717	2,074	2,508	2,819	59	1,671	2,001	2,391	2,662
23	1,714	2,069	2,500	2,807	60	1,671	2,000	2,390	2,660
24	1,711	2,064	2,492	2,797	61	1,670	2,000	2,389	2,659
25	1,708	2,060	2,485	2,787	62	1,670	1,999	2,388	2,657
26	1,706	2,056	2,479	2,779	63	1,669	1,998	2,387	2,656
27	1,703	2,052	2,473	2,771	64	1,669	1,998	2,386	2,655
28	1,701	2,048	2,467	2,763	65	1,669	1,997	2,385	2,654
29	1,699	2,045	2,462	2,756	66	1,668	1,997	2,384	2,652
30	1,697	2,042	2,457	2,750	67	1,668	1,996	2,383	2,651
31	1,696	2,040	2,453	2,744	68	1,668	1,995	2,382	2,650
32	1,694	2,037	2,449	2,738	69	1,667	1,995	2,382	2,649
33	1,692	2,035	2,445	2,733	70	1,667	1,994	2,381	2,648
34	1,691	2,032	2,441	2,728	71	1,667	1,994	2,380	2,647
35	1,690	2,030	2,438	2,724	72	1,666	1,993	2,379	2,646
36	1,688	2,028	2,434	2,719	73	1,666	1,993	2,379	2,645
37	1,687	2,026	2,431	2,715	74	1,666	1,993	2,378	2,644

zdroj: Markechová, D., Tirpáková, A., & Stehlíková, B. (2011)

Tab 2.16 Kritické hodnoty $t(k)$ α Studentovho t – rozdelenia (pokračovanie)

α k	0,10	0,05	0,02	0,01	α k	0,10	0,05	0,02	0,01
75	1,665	1,992	2,377	2,643	112	1,659	1,981	2,360	2,620
76	1,665	1,992	2,376	2,642	113	1,658	1,981	2,360	2,620
77	1,665	1,991	2,376	2,641	114	1,658	1,981	2,360	2,620
78	1,665	1,991	2,375	2,640	115	1,658	1,981	2,359	2,619
79	1,664	1,990	2,374	2,640	116	1,658	1,981	2,359	2,619
80	1,664	1,990	2,374	2,639	117	1,658	1,980	2,359	2,619
81	1,664	1,990	2,373	2,638	118	1,658	1,980	2,358	2,618
82	1,664	1,989	2,373	2,637	119	1,658	1,980	2,358	2,618
83	1,663	1,989	2,372	2,636	120	1,658	1,980	2,358	2,617
84	1,663	1,989	2,372	2,636	121	1,658	1,980	2,358	2,617
85	1,663	1,988	2,371	2,635	122	1,657	1,980	2,357	2,617
86	1,663	1,988	2,370	2,634	123	1,657	1,979	2,357	2,616
87	1,663	1,988	2,370	2,634	124	1,657	1,979	2,357	2,616
88	1,662	1,987	2,369	2,633	125	1,657	1,979	2,357	2,616
89	1,662	1,987	2,369	2,632	126	1,657	1,979	2,356	2,615
90	1,662	1,987	2,368	2,632	127	1,657	1,979	2,356	2,615
91	1,662	1,986	2,368	2,631	128	1,657	1,979	2,356	2,615
92	1,662	1,986	2,368	2,630	129	1,657	1,979	2,356	2,614
93	1,661	1,986	2,367	2,630	130	1,657	1,978	2,355	2,614
94	1,661	1,986	2,367	2,629	131	1,657	1,978	2,355	2,614
95	1,661	1,985	2,366	2,629	132	1,656	1,978	2,355	2,614
96	1,661	1,985	2,366	2,628	133	1,656	1,978	2,355	2,613
97	1,661	1,985	2,365	2,627	134	1,656	1,978	2,354	2,613
98	1,661	1,984	2,365	2,627	135	1,656	1,978	2,354	2,613
99	1,660	1,984	2,365	2,626	136	1,656	1,978	2,354	2,612
100	1,660	1,984	2,364	2,626	137	1,656	1,977	2,354	2,612
101	1,660	1,984	2,364	2,625	138	1,656	1,977	2,354	2,612
102	1,660	1,983	2,363	2,625	139	1,656	1,977	2,353	2,612
103	1,660	1,983	2,363	2,624	140	1,656	1,977	2,353	2,611
104	1,660	1,983	2,363	2,624	141	1,656	1,977	2,353	2,611
105	1,659	1,983	2,362	2,623	142	1,656	1,977	2,353	2,611
106	1,659	1,983	2,362	2,623	143	1,656	1,977	2,353	2,611
107	1,659	1,982	2,362	2,623	144	1,656	1,977	2,353	2,610
108	1,659	1,982	2,361	2,622	145	1,655	1,976	2,352	2,610
109	1,659	1,982	2,361	2,622	146	1,655	1,976	2,352	2,610
110	1,659	1,982	2,361	2,621	147	1,655	1,976	2,352	2,610
111	1,659	1,982	2,360	2,621	∞	1,645	1,960	2,326	2,576

zdroj: Markechová, D., Tirpáková, A., & Stehlíková, B. (2011)

Zdroje a použitá literatúra

Markechová, D., Tirpáková, A., & Stehlíková, B. (2011). *Štatistické metódy a ich aplikácie* (1. vyd.). Nitra: Fakulta prírodných vied UKF v Nitre.