

## Cvičení 8: Parametrické úlohy o více nezávislých náhodných výběrech

**Úkol 1.:** V jisté továrně se měřil čas, který potřeboval každý ze tří dělníků k uskutečnění téhož pracovního úkonu. Čas v minutách:

1. dělník: 3,6 3,8 3,7 3,5
2. dělník: 4,3 3,9 4,2 3,9 4,4 4,7
3. dělník: 4,2 4,5 4,0 4,1 4,5 4,4.

Na hladině významnosti 0,05 testujte hypotézu, že výkony těchto tří dělníků jsou stejné. Zamítnete-li nulovou hypotézu, určete, výkony kterých dělníků se liší na dané hladině významnosti 0,05.

### Návod:

Úloha vede na analýzu rozptylu jednoduchého třídění. Postupujeme podle skript Základní statistické metody, odstavec 8.1.

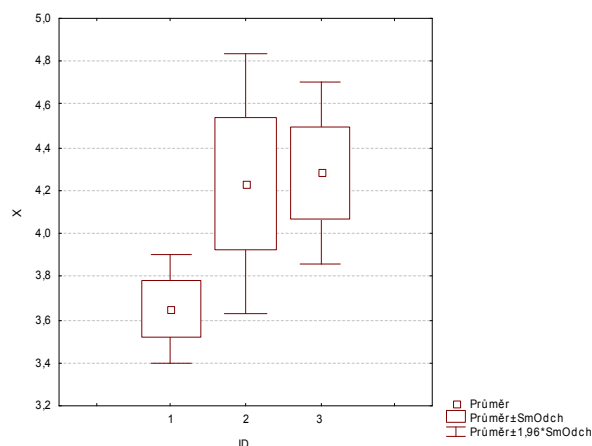
Načteme datový soubor `cas_delniku.sta`. Proměnná  $X$  obsahuje zjištěné časy, proměnná  $ID$  nabývá hodnoty 1 pro 1. dělníka, hodnoty 2 pro 2. dělníka a hodnoty 3 pro 3. dělníka.

Statistiky – Základní statistiky/tabulky – Rozklad & jednofakt. ANOVA – Proměnné - Závislé  $X$ , Grupovací  $ID$ , OK, Kódy pro grupovací proměnné – Vše, OK, Výpočet: Tabulka statistik (zobrazí se průměry, směrodatné odchylky a rozsahy všech tří výběrů).

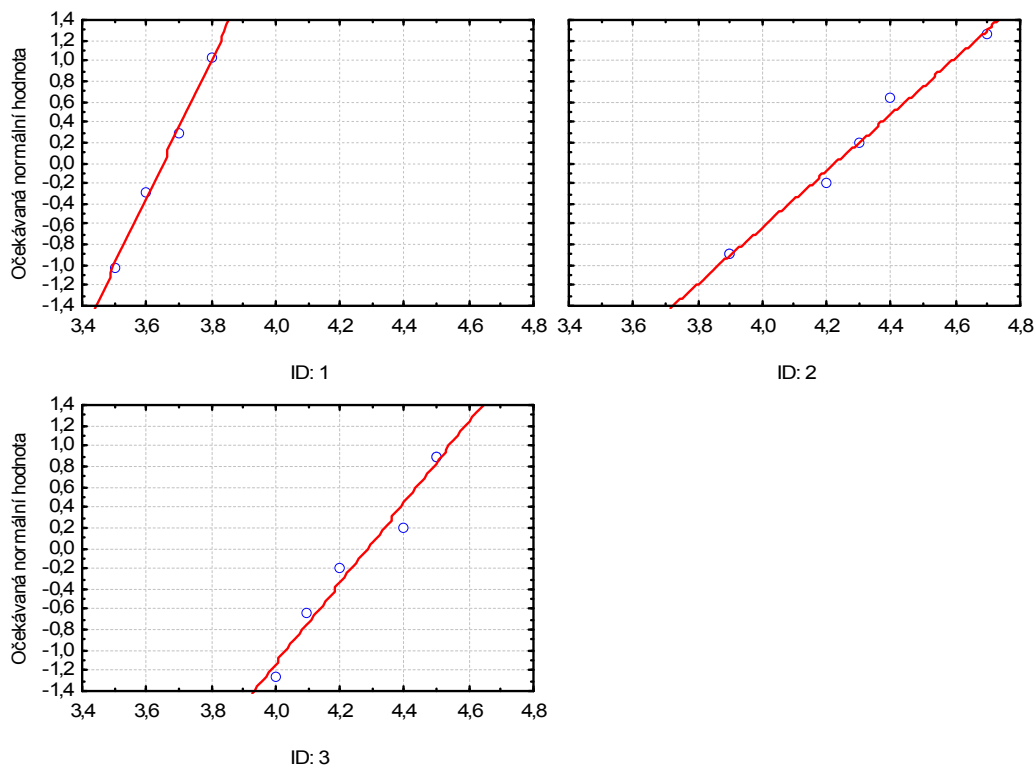
| ID       | X<br>průměr | X<br>N | X<br>Sm.odch. |
|----------|-------------|--------|---------------|
| 1        | 3,650000    | 4      | 0,129099      |
| 2        | 4,233333    | 6      | 0,307679      |
| 3        | 4,283333    | 6      | 0,213698      |
| Vš.skup. | 4,106250    | 16     | 0,353023      |

**Komentář:** Na uskutečnění daného pracovního úkonu potřebuje nejkratší čas 1. dělník. Podává také nejvyrovnanější výkony – směrodatná odchylka proměnné  $X$  je u něj nejmenší. Naopak nejpomalejší je 3. dělník.

Nyní vytvoříme krabicové diagramy: Návrat do Statistiky podle skupin – Kategoriz. krabicový graf (současné zobrazení krabicových diagramů pro všechny tři výběry )



Pomocí N-P plot orientačně posoudíme normalitu všech tří výběrů:  
 Návrat do Statistiky podle skupin – ANOVA & testy – Kategoriz. norm. pravd. grafy



**Komentář:** Ve všech třech případech se tečky jen málo odchyľují od přímky, lze soudit, že data pocházejí z normálního rozložení.

Provedení testu o shodě rozptylů:  
 Návrat do Statistiky podle skupin – Leveneovy testy

| Leveneův test homogenity rozptylů (cas_delniku.sta) |             |             |             |             |             |             |          |          |
|---|-------------|-------------|-------------|-------------|-------------|-------------|----------|----------|
| Označ. efekty jsou význ. na hlad. $p < ,05000$      |             |             |             |             |             |             |          |          |
| Proměnná  | SČ<br>efekt | SV<br>efekt | PČ<br>efekt | SČ<br>chyba | SV<br>chyba | PČ<br>chyba | F        | p        |
| X   | 0,042708    | 2           | 0,021354    | 0,183333    | 13          | 0,014103    | 1,514205 | 0,256356 |

**Komentář:** Testová statistika Levenova testu nabývá hodnoty 1,5142, stupně volnosti čitatele = 2, jmenovatele = 13, odpovídající p-hodnota = 0,256, tedy na hladině významnosti 0,05 se nezamítá hypotézu o shodě rozptylů.

Provedení testu o shodě středních hodnot:  
 Návrat do Statistiky podle skupin – Analýza rozptylu.

| Analýza rozptylu (cas_delniku.sta)             |             |             |             |             |             |             |          |          |
|--|-------------|-------------|-------------|-------------|-------------|-------------|----------|----------|
| Označ. efekty jsou význ. na hlad. $p < ,05000$ |             |             |             |             |             |             |          |          |
| Proměnná                                       | SČ<br>efekt | SV<br>efekt | PČ<br>efekt | SČ<br>chyba | SV<br>chyba | PČ<br>chyba | F        | p        |
| X  | 1,117708    | 2           | 0,558854    | 0,751667    | 13          | 0,057821    | 9,665327 | 0,002680 |

**Komentář:** Skupinový součet čtverců  $S_A = 1,1177$ , počet stupňů volnosti  $f_A = 2$ , reziduální součet čtverců  $S_E = 0,7517$ , počet stupňů volnosti  $f_E = 13$ , testová statistika  $F_A = \frac{S_A/f_A}{S_E/f_E}$  nabývá hodnoty 9,6653, počet stupňů volnosti čitatele = 2, jmenovatele = 13, odpovídající p-hodnota = 0,00268, tedy na hladině významnosti 0,05 se zamítá hypotéza o shodě středních hodnot .

Provedení metody mnohonásobného porovnávání (Scheffého test – viz skriptu Základní statistické metody, věta 8.2.2.1.):

Návrat do do Statistiky podle skupin – Post- hoc – Scheffého test.

|    |     | Scheffeho test; proměn.:X(cas_delniku.sta)<br>Označ. rozdíly jsou významné na hlad. p < ,05000 |                 |                 |
|----|-----|--|-----------------|-----------------|
| ID |     | {1}<br>M=3,6500  | {2}<br>M=4,2333 | {3}<br>M=4,2833 |
| 1  | {1} |  | 0,008391        | 0,004705        |
| 2  | {2} | 0,008391   |                 | 0,937504        |
| 3  | {3} | 0,004705   | 0,937504        |                 |

**Komentář:** Tabulka obsahuje p-hodnoty pro testování hypotéz o shodě středních hodnot všech dvojic výběrů. Výsledek Scheffého metody ukazuje, že na hladině významnosti 0,05 se liší výkony dělníků (1,2), (1,3) a neliší se (2,3).

**Úkol 2.:** V cestovní kanceláři zkoumali u 609 náhodně vybraných klientů, o jaké ubytování měli zájem (varianty apartmán, bungalov, hotel, stan) a zjišťovali též pohlaví klienta.

| Typ ubytování | apartmán | bungalov | hotel | stan |
|---------------|----------|----------|-------|------|
| Počet žen     | 12       | 27       | 208   | 33   |
| Počet mužů    | 100      | 68       | 36    | 152  |

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozdíly v typech ubytování mezi muži a ženami jsou způsobeny pouze náhodnými vlivy.

### Návod:

Postupujeme podle skriptu Základní statistické metody, Věta 8.5.1.1. Testujeme hypotézu  $H_0: \vartheta_1 = \dots = \vartheta_4$  proti alternativní hypotéze  $H_1$ : aspoň jedna dvojice parametrů je různá.

Načteme datový soubor klienti\_CK.sta. Proměnná POHLAVI obsahuje hodnotu 0 pro ženu, 1 pro muže. Proměnná TYP UBYTOVANI má hodnotu 1 pro apartmán, hodnotu 2 pro bungalov, hodnotu 3 pro hotel a hodnotu 4 pro stan.

Nejprve zjistíme podíly mužů v jednotlivých typech ubytování.

Statistiky – Základní statistiky/tabulky - Rozklad & jednofakt. ANOVA - OK - Proměnné - Závislé POHLAVI, Grupovací TYP UBYTOVANI, OK, Kódy pro grupovací proměnné – Vše, OK – Popisné statistiky - Výpočet: Tabulka statistik – ponecháme zaškrtnuto N - OK.

| typ ubytovani | pohlavi průměr | pohlavi N |
|---------------|----------------|-----------|
| apartmán      | 0,892857       | 112       |
| bungalov      | 0,602941       | 68        |
| hotel         | 0,147541       | 244       |
| stan          | 0,821622       | 185       |
| Vš.skup.      | 0,540230       | 609       |

**Komentář:** Vidíme, že z těch klientů, kteří se ubytovali v apartmánu, bylo 89,3% mužů, mezi obyvateli bungalovů bylo 60,3% mužů, z ubytovaných v hotelu bylo mužů pouze 14,7% a z těch, kteří bydleli pod stanem, bylo 82,1% mužů.

Ověříme splnění podmínek dobré aproximace:  $n_j m^* > 5$  pro všechna  $j = 1, \dots, r$ . Vážený průměr  $m^*$  se nachází v posledním řádku výstupní Rozkladové tabulky popisných statistik. Jeho hodnotu okopírujeme do políček pro průměry relativní četnosti ubytovaných v jednotlivých typech ubytování, poslední řádek odstraníme a k tabulce přidáme jednu novou proměnnou, do jejíhož Dlouhého jména napíšeme  $=v2*v3$ .

| typ ubytovani | pohlavi průměr | pohlavi N | NProm<br>$=v2*v3$ |
|---------------|----------------|-----------|-------------------|
| apartmán      | 0,540230       | 112       | 60,505747         |
| bungalov      | 0,540230       | 68        | 36,735632         |
| hotel         | 0,540230       | 244       | 131,816092        |
| stan          | 0,540230       | 185       | 99,942529         |

**Komentář:** Vidíme, že podmínky dobré aproximace jsou splněny.

Dále provedeme testování hypotézy o shodě parametrů čtyř alternativních rozložení. Statistiky – Základní statistiky/tabulky – Kontingenční tabulky – OK - Specif. tabulky – List 1 POHLAVI, List 2 TYP UBYTOVANI, OK– Možnosti - Statistiky dvourozm tabulek - zaškrtneme Pearson & M-L Chi –square – Detailní výsledky – Detailní 2-rozm. tabulky

| Statist.          | Chi-kvadr. | sv   | p        |
|-------------------|------------|------|----------|
| Pearsonův chí-kv. | 267,6070   | df=3 | p=0,0000 |
| M-V chí-kvadr.    | 294,9782   | df=3 | p=0,0000 |

**Komentář:** Testová statistika Q (viz skripta Základní statistické metody, vzorec 8.15.) se realizuje hodnotou 267,6070, počet stupňů volnosti je 3, odpovídající p-hodnota = 0,0000, tedy na asymptotické hladině významnosti 0,05 hypotézu  $H_0$  zamítáme. S rizikem omylu nejvýše 0,05 jsme tedy prokázali, že rozdíly v podílech klientů a klientek ubytovaných v různých typech ubytovacích zařízení nelze vysvětlit pouze náhodnými vlivy.

Nakonec provedeme metodu mnohonásobného porovnávání, abychom zjistili, které dvojice typů ubytování se liší na asymptotické hladině významnosti 0,05. Návrat do do Statistiky podle skupin – Post- hoc – Scheffěův test.

| typ ubytovani | {1}      | {2}      | {3}      | {4}      |
|---------------|----------|----------|----------|----------|
|               | M=,89286 | M=,60294 | M=,14754 | M=,82162 |
| apartmán {1}  |          | 0,000016 | 0,000000 | 0,471207 |
| bungalov {2}  | 0,000016 |          | 0,000000 | 0,000797 |
| hotel {3}     | 0,000000 | 0,000000 |          | 0,000000 |
| stan {4}      | 0,471207 | 0,000797 | 0,000000 |          |

**Komentář:** Tabulka obsahuje p-hodnoty pro testování hypotéz o shodě středních hodnot všech dvojic výběrů. Výsledek Scheffého metody ukazuje, že z hlediska podílu mužů se na hladině významnosti 0,05 neliší pouze ubytování v apartmánu a ve stanu.

### Příklady k samostatnému řešení

**Příklad 1.:** Studenti byli vyučováni předmětu za využití pěti pedagogických metod: tradiční způsob, programová výuka, audiotechnika, audiovizuální technika a vizuální technika. Z každé skupiny byl vybrán náhodný vzorek studentů a všichni byli podrobeni témuž písemnému testu. Výsledky testu:

| metoda        | počet bodů |      |      |      |      |                |
|---------------|------------|------|------|------|------|----------------|
| tradiční      | 76,2       | 48,3 | 85,1 | 63,7 | 91,6 | 87,2           |
| programová    | 85,2       | 74,3 | 76,5 | 80,3 | 67,4 | 67,9 72,1 60,4 |
| audio         | 67,3       | 60,1 | 55,4 | 72,3 | 40   |                |
| audiovizuální | 75,8       | 81,6 | 90,3 | 78   | 67,8 | 57,6           |
| vizuální      | 50,5       | 70,2 | 88,8 | 67,1 | 77,7 | 73,9           |

Na hladině významnosti 0,05 testujte hypotézu, že znalosti všech studentů jsou stejné a nezávisí na použité pedagogické metodě. V případě zamítnutí hypotézy zjistěte, které výběry se liší na hladině významnosti 0,05.

### Řešení:

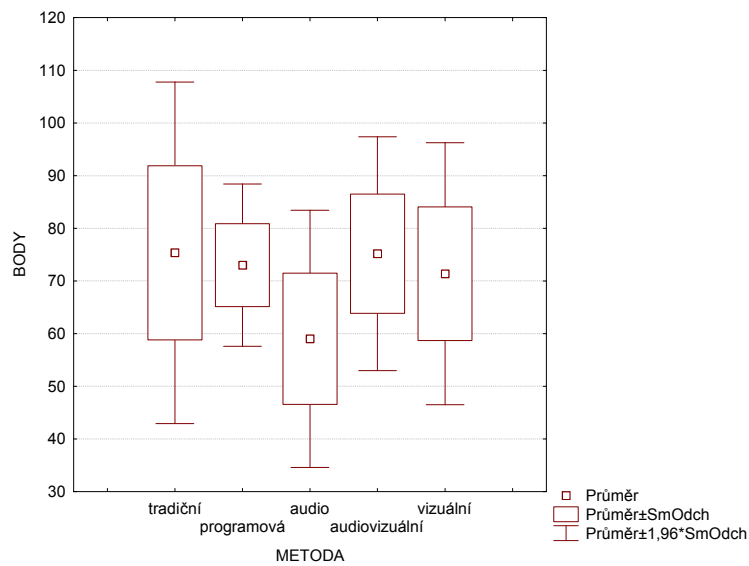
Načteme datový soubor pet\_metod.sta. Proměnná BODY obsahuje dosažené počty bodů a proměnná METODA označení příslušné pedagogické metody.

Nejprve vypočteme průměry, směrodatné odchylky a rozsahy všech tří výběrů:

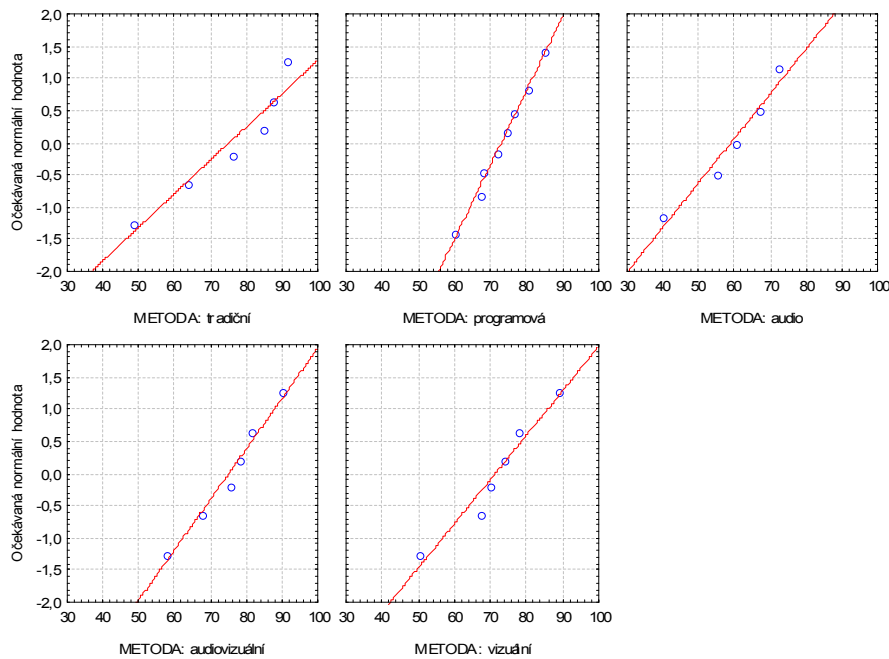
| Rozkladová tabulka popisných statistik (pet_metod.sta)<br>N=31 (V seznamu záv. prom. nejsou ChD) |                |           |                  |
|--|----------------|-----------|------------------|
| METODA   | BODY<br>průměr | 3ODY<br>N | BODY<br>Sm.odch. |
| tradiční   | 75,35000       | 6         | 16,53901         |
| programová   | 73,01250       | 8         | 7,86501          |
| audio  | 59,02000       | 5         | 12,45941         |
| audiovizuální  | 75,18333       | 6         | 11,32862         |
| vizuální   | 71,36667       | 6         | 12,69199         |
| Vš.skup.   | 71,30968       | 31        | 12,69534         |

**Komentář:** Nejlepších výsledků dosahují studenti vyučovaní tradiční metodou, podávají však nejméně vyrovnané výkony (počty bodů v této skupině mají největší směrodatnou odchylku). Naopak nejhoršího výsledku dosáhli studenti vyučovaní audio metodou. Nejvyrovnanější výkony pozorujeme u studentů vyučovaných programovou metodou.

Vytvoříme krabicové diagramy:



Pomocí N-P grafů vizuálně posoudíme normalitu všech pěti výběrů:



**Komentář:** Ze vzhledu N-P grafů je patrné, že předpoklad normality je ve všech pěti případech oprávněný.

Provedeme Levenův test (testování homogenity rozptylů všech pěti výběrů)

| Proměnná | Leveneův test homogenity rozptylů (pet_metod.sta) |          |          |          |          |          |          |          |
|----------|---|----------|----------|----------|----------|----------|----------|----------|
|          | SČ efekt  | SV efekt | PČ efekt | SČ chyba | SV chyba | PČ chyba | F        | p        |
| BODY     | 162,4883  | 4        | 40,62208 | 1289,544 | 26       | 49,59783 | 0,819029 | 0,524791 |

**Komentář:** Testová statistika F se realizuje hodnotou 0,819, počet stupňů volnosti čitatele = 4, jmenovatele = 26, odpovídající p-hodnota = 0,5248, na hladině významnosti 0,05 tedy nezamítáme hypotézu o shodě rozptylů.

Budeme testovat hypotézu o shodě středních hodnot všech pěti výběrů:

| Analýza rozptylu (pet_metod.sta)             |             |             |             |             |             |             |          |          |
|--|-------------|-------------|-------------|-------------|-------------|-------------|----------|----------|
| Označ. efekty jsou význ. na hlad. p < ,05000 |             |             |             |             |             |             |          |          |
| Proměnná                                     | SČ<br>efekt | SV<br>efekt | PČ<br>efekt | SČ<br>chyba | SV<br>chyba | PČ<br>chyba | F        | p        |
| BODY   | 966,3737    | 4           | 241,5934    | 3868,773    | 26          | 148,7990    | 1,623623 | 0,198252 |

**Komentář:** Testová statistika F se realizuje hodnotou 1,6236, počet stupňů volnosti čitatele = 4, jmenovatele = 26, odpovídající p-hodnota = 0,1983, na hladině významnosti 0,05 tedy nezamítáme hypotézu o shodě středních hodnot. Znamená to, že s rizikem omylu nejvýše 5% se neprokázal rozdíl v účinnosti jednotlivých pedagogických metod..

**Příklad 2.:** Pan Novák může cestovat z místa bydliště do místa pracoviště třemi různými způsoby: tramvají (způsob A), autobusem (způsob B) a metrem s následným přestupem na tramvaj (způsob C). Máme k dispozici jeho naměřené časy cestování do práce v době ranní špičky (včetně čekání na příslušný spoj) v minutách:

způsob A: 32, 39, 42, 37, 34, 38:

způsob B: 30, 34, 28, 26, 32,

způsob C: 40, 37, 31, 39, 38, 33, 34

Pro všechny tři způsoby dopravy vypočítejte průměrné časy cestování. Na hladině významnosti 0,05 testujte hypotézu, že doba cestování do práce nezávisí na způsobu dopravy. V případě zamítnutí nulové hypotézy zjistěte, které způsoby dopravy do práce se od sebe liší na hladině významnosti 0,05.

### Řešení:

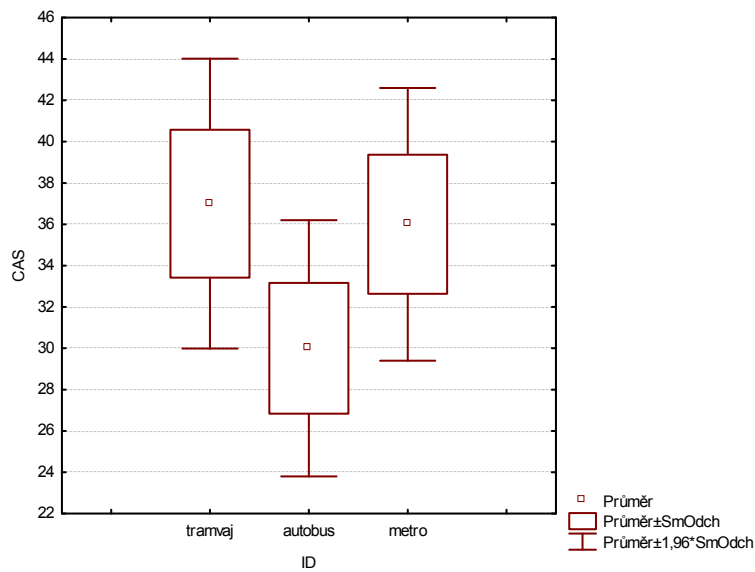
Načteme datový soubor doby\_cestovani.sta. Proměnná CAS obsahuje zjištěné doby cestování a proměnná ID označení příslušného způsobu dopravy.

Nejprve vypočteme průměry, směrodatné odchylky a rozsahy všech tří výběrů:

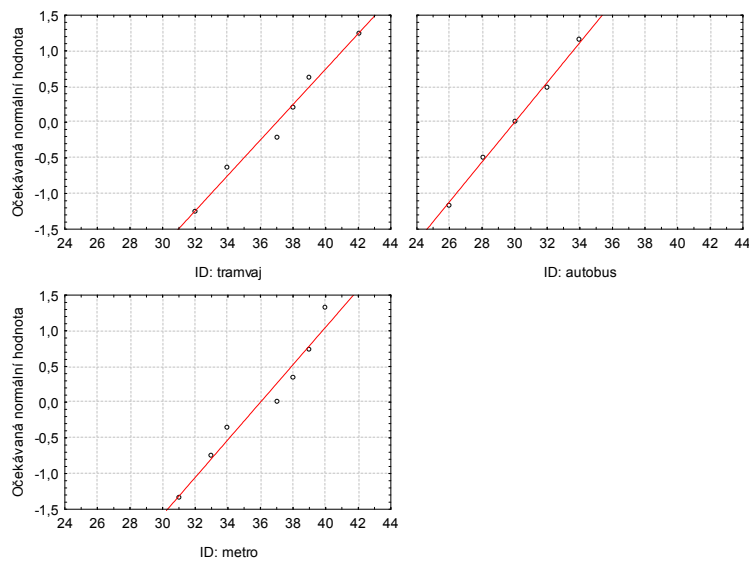
| Rozkladová tabulka popisných statistik (doby_cestovani.sta) |               |          |                 |
|---|---------------|----------|-----------------|
| N=18 (V seznamu záv. prom. nejsou ChD)                      |               |          |                 |
| ID  | CAS<br>průměr | CAS<br>N | CAS<br>Sm.odch. |
| tramvaj   | 37,00000      | 6        | 3,577709        |
| autobus   | 30,00000      | 5        | 3,162278        |
| metro   | 36,00000      | 7        | 3,366502        |
| Vš.skup.  | 34,66667      | 18       | 4,379095        |

**Komentář:** Nejkratší průměrnou dobu do zaměstnání pan Novák cestuje, když použije autobus, naopak nejdéle cestuje tramvají. Variabilita dob jednotlivých způsobů cestování je vcelku vyrovnaná.

Vytvoříme krabicové diagramy:



Pomocí N-P grafů vizuálně posoudíme normalitu všech tří výběrů:



**Komentář:** Ze vzhledu N-P grafů je patrné, že předpoklad normality je ve všech třech případech oprávněný.

Provedeme Levenův test (testování homogenity rozptylů všech tří výběrů)

| Leveneův test homogenity rozptylů (doby_cestovani.sta) |             |             |             |             |             |             |          |          |
|--|-------------|-------------|-------------|-------------|-------------|-------------|----------|----------|
| Označ. efekty jsou význ. na hlad. p < ,05000           |             |             |             |             |             |             |          |          |
| Proměnná   | SČ<br>efekt | SV<br>efekt | PČ<br>efekt | SČ<br>chyba | SV<br>chyba | PČ<br>chyba | F        | p        |
| CAS  | 0,609524    | 2           | 0,304762    | 43,39048    | 15          | 2,892698    | 0,105356 | 0,900665 |



**Komentář:** Testová statistika F se realizuje hodnotou 0,1054, počet stupňů volnosti čitatele = 2, jmenovatele = 15, odpovídající p-hodnota = 0,9007, na hladině významnosti 0,05 tedy nezamítáme hypotézu o shodě rozptylů.

Budeme testovat hypotézu o shodě středních hodnot všech tří výběrů:

| Analýza rozptylu (doby_cestovani.sta)        |             |             |             |             |             |             |          |          |
|--|-------------|-------------|-------------|-------------|-------------|-------------|----------|----------|
| Označ. efekty jsou význ. na hlad. p < ,05000 |             |             |             |             |             |             |          |          |
| Proměnná                                     | SČ<br>efekt | SV<br>efekt | PČ<br>efekt | SČ<br>chyba | SV<br>chyba | PČ<br>chyba | F        | p        |
| CAS  | 154,0000    | 2           | 77,00000    | 172,0000    | 15          | 11,46667    | 6,715116 | 0,008267 |

**Komentář:** Testová statistika F se realizuje hodnotou 6,7151, počet stupňů volnosti čitatele = 2, jmenovatele = 15, odpovídající p-hodnota = 0,0083, na hladině významnosti 0,05 tedy zamítáme hypotézu o shodě středních hodnot. Znamená to, že s rizikem omylu nejvýše 5% se prokázal rozdíl v dobách cestování pana Nováka do zaměstnání autobusem, tramvají a metrem.

Scheffého metodou mnohonásobného porovnávání zjistíme, které dvojice způsobů cestování do zaměstnání se liší na hladině významnosti 0,05:

| Scheffeho test; proměn.:CAS (doby_cestovani.sta) |                 |                 |                 |
|--|-----------------|-----------------|-----------------|
| Označ. rozdíly jsou významné na hlad. p < ,05000 |                 |                 |                 |
| ID   | {1}<br>M=37,000 | {2}<br>M=30,000 | {3}<br>M=36,000 |
| tramvaj {1}                                      |                 | 0,013410        | 0,869732        |
| autobus {2}                                      | 0,013410        |                 | 0,028046        |
| metro {3}  | 0,869732        | 0,028046        |                 |

**Komentář:** Z tabulky vyplývá, že s rizikem omylu nejvýše 5% se neliší pouze cestování tramvají a metrem.

**Příklad 3.:** U 856 žáků ZŠ bylo zjišťováno celkové IQ (proměnná IQ\_CELK). Na asymptotické hladině významnosti 0,05 testujte hypotézu, že pravděpodobnost výskytu dítěte s nadprůměrným IQ\_CELK (tj. nad 100 bodů) je stejná ve skupinách matek se základním, středoškolským a vysokoškolským vzděláním (proměnná VZDEL\_M).

### Řešení:

Máme tři nezávislé náhodné výběry, j-tý pochází z rozložení  $A(\vartheta_j)$ ,  $j = 1, 2, 3$ . Testujeme hypotézu  $H_0: \vartheta_1 = \vartheta_2 = \vartheta_3$ .

$$n_1 = 361, n_2 = 386, n_3 = 109, n = 856$$

$$m_1 = 111/361 = 30,75\%, m_2 = 227/386 = 58,81\%, m_3 = 85/109 = 77,98\%, m_* = (111+227+85)/856 = 423/856 = 49,42\%.$$

Podmínky dobré aproximace:

$$361 \cdot \frac{423}{856} = 178,39, 386 \cdot \frac{423}{856} = 190,75, 109 \cdot \frac{423}{856} = 53,86$$

Testová statistika

$$Q = \frac{1}{M_*(1-M_*)} \sum_{j=1}^r n_j M_j^2 - n \frac{M_*}{1-M_*} = 99,53$$

$$\text{Kritický obor: } W = \langle \chi^2_{0,95}(2), \infty \rangle = \langle 5,991, \infty \rangle.$$

Protože testové kritérium se realizuje v kritickém oboru,  $H_0$  zamítáme na asymptotické hladině významnosti 0,05.

Metoda mnohonásobného porovnávání prokázala, že na asymptotické hladině významnosti 0,05 se liší všechny tři skupiny.