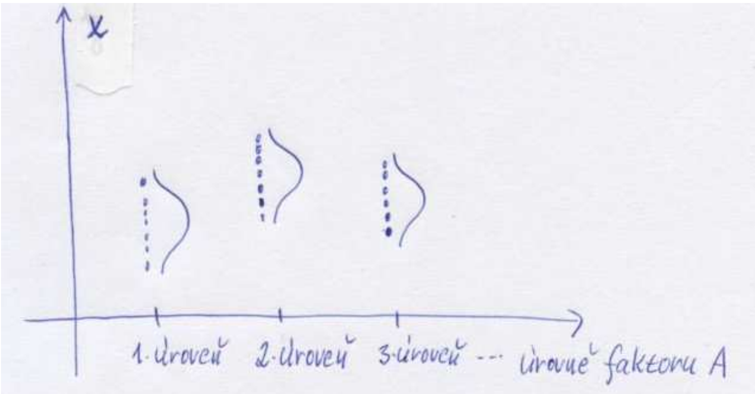
Analýza rozptylu jednoduchého třídění

Motivace: Zajímáme se o problém, zda lze určitým faktorem (tj. nominální náhodnou veličinou A) vysvětlit variabilitu pozorovaných hodnot náhodné veličiny X, která je intervalového či poměrového typu. Např. zkoumáme, zda metoda výuky určitého předmětu (faktor A) ovlivňuje počet bodů dosažených studenty v závěrečném testu (náhodná veličina X).
Předpokládáme, že faktor A má $r \geq 3$ úrovní a přitom i-té úrovni odpovídá n_i pozorování $X_{i1},\dots,X_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, i = 1, ..., r a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy $X_{ij} = \mu_i + \varepsilon_{ij}$, kde ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, i = 1, …, r, j = 1, …, n_i.
Výsledky lze zapsat do tabulky

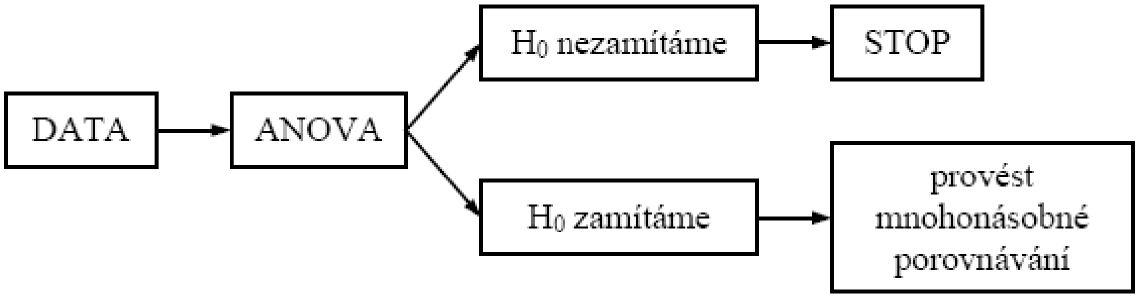
faktor A	výsledky
úroveň 1	$X_{11},\dots,X_{1n_1}$
úroveň 2	$X_{21},\dots,X_{2n_2}$
…	…
úroveň r	$X_{r1},\dots,X_{rn_r}$

Ilustrace:



Na hladině významnosti α testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné, tj. H_0: $\mu_1 = \dots = \mu_r$ proti alternativní hypotéze H_1, která tvrdí, že aspoň jedna dvojice středních hodnot se liší.
Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit $\binom{r}{2}$ dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test. Hypotézu o shodě všech středních hodnot bychom pak zamítli, pokud aspoň v jednom případě z $\binom{r}{2}$ porovnávání se prokáže odlišnost středních hodnot. Odtud je vidět, že k neoprávněnému zamítnutí nulové hypotézy (tj. k chybě 1. druhu) může dojít s pravděpodobností větší než α. Proto ve 30. le-

tech 20. století vytvořil R. A. Fisher metodu ANOVA (analýza rozptylu, v popsané situaci konkrétně analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.
Pokud na hladině významnosti α zamítneme nulovou hypotézu, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží metody mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.



Označení:

V analýze rozptylu jednoduchého třídění se používá tzv. tečková notace.

$n = \sum_{i=1}^{r} n_i$ … celkový rozsah všech r výběrů

$X_{i.} = \sum_{j=1}^{n_i} X_{ij}$ … součet hodnot v i-tém výběru

$M_{i.} = \frac{1}{n_i} X_{i.}$ … výběrový průměr v i-tém výběru

$X_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}$ … součet hodnot všech výběrů

$M_{..} = \frac{1}{n} X_{..}$ … celkový průměr všech r výběrů

Testování hypotézy o shodě středních hodnot

Náhodné veličiny X_{ij} se řídí modelem

$$M0: X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

pro $i = 1, \dots, r$, $j = 1, \dots, n_i$, přičemž
ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$,
μ je společná část střední hodnoty závisle proměnné veličiny,
α_i je efekt faktoru A na úrovni i.
Parametry μ, α_i neznáme.

Požadujeme, aby platila tzv. reparametrizační rovnice: $\sum_{i=1}^{r} n_i \alpha_i = 0$. (Pokud je třídění vyvážené, tj. pokud mají všechny výběry stejný rozsah: $n_1 = n_2 = \dots = n_r$, pak lze použít zjednodušenou podmínku $\sum_{i=1}^{r} \alpha_i = 0$.)

Zavedeme součty čtverců

$S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - M_{..})^2$ … celkový součet čtverců (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru),

počet stupňů volnosti $f_T = n - 1$,

$S_A = \sum_{i=1}^{r} n_i (M_{i.} - M_{..})^2$ … skupinový součet čtverců (charakterizuje variabilitu mezi jednotlivými náhodnými výběry),

počet stupňů volnosti $f_A = r - 1$.

Sčítanec $(M_{i.} - M_{..})$ představuje bodový odhad efektu α_i.

$S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - M_{i.})^2$ … reziduální součet čtverců (charakterizuje variabilitu uvnitř jednotlivých výběrů),

počet stupňů volnosti $f_E = n - r$.

Lze dokázat, že $S_T = S_A + S_E$.

(Důkaz je proveden např. ve skriptech Budíková, Mikoláš, Osecký: Popisná statistika v poznámce 5.20.)

Kdyby nezáleželo na faktoru A, platila by hypotéza $\alpha_1 = \ldots = \alpha_r = 0$ a dostali bychom model

$$M1: X_{ij} = \mu + \varepsilon_{ij}.$$

Během analýzy rozptylu tedy zkoumáme, zda výběrové průměry $M_1, \ldots, M_r$ se od sebe liší pouze v mezích náhodného kolísání kolem celkového průměru M nebo zda se projevuje vliv faktoru A.

Rozdíl mezi modely M0 a M1 ověřujeme pomocí testové statistiky

$F_A = \dfrac{S_A / f_A}{S_E / f_E}$, která se řídí rozložením F(r-1,n-r), je-li model M1 správný. Hypotézu o nevýznamnosti faktoru A tedy zamítneme na hladině významnosti α, když platí: $F_A \geq F_{1-\alpha}(r-1,n-r)$.

Výsledky výpočtů zapisujeme do tabulky analýzy rozptylu jednoduchého třídění.

Zdroj variability	součet čtverců	stupně volnosti	podíl	F_A
skupiny	S_A	$f_A = r - 1$	S_A/f_A	$\dfrac{S_A/f_A}{S_E/f_E}$
reziduální	S_E	$f_E = n - r$	S_E/f_E	-
celkový	S_T	$f_T = n - 1$	-	-

Sílu závislosti náhodné veličiny X na faktoru A můžeme měřit pomocí poměru determinace: $P^2 = \frac{S_A}{S_T}$. Nabývá hodnot z intervalu $\langle 0,1 \rangle$.

Testování hypotézy o shodě rozptylů

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných r výběrech.

a) Levenův test: Položme $Z_{ij} = |X_{ij} - M_{i.}|$. Označíme

$$M_{Zi} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij},$$

$$M_Z = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Z_{ij},$$

$$S_{ZE} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Z_{ij} - M_{Zi})^2,$$

$$S_{ZA} = \sum_{i=1}^{r} n_i (M_{Zi} - M_Z)^2$$

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_{ZA} = \frac{S_{ZA}/(r-1)}{S_{ZE}/(n-r)} \approx F(r-1, n-r).$$

Hypotézu o shodě rozptylů tedy zamítáme na asymptotické hladině významnosti α, když $F_{ZA} \geq F_{1-\alpha}(r-1, n-r)$.

(Levenův test je vlastně založen na analýze rozptylu absolutních hodnot centrovaných pozorování. Vzhledem k tomu, že náhodné veličiny $X_{ij} - M_{i.}$ nejsou stochasticky nezávislé a absolutní hodnoty těchto veličin nemají normální rozložení, je Levenův test pouze aproximativní.)

Modifikací Levenova testu je Brownův – Forsytheův test. Modifikace spočívá v tom, že místo výběrového průměru i-tého výběru se při výpočtu veličiny Z_{ij} používá medián i-tého výběru.

b) Bartlettův test: Platí-li hypotéza o shodě rozptylů a rozsahy všech výběrů jsou větší než 6, pak statistika

$$B = \frac{1}{C}\left[(n-r)\ln S_*^2 - \sum_{i=1}^{r}(n_i - 1)\ln S_i^2\right] \approx \chi^2(r-1), \text{ kde } C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{n_i - 1} - \frac{1}{n-r}\right)$$

a S_*^2 je vážený průměr výběrových rozptylů.

H_0 zamítáme na asymptotické hladině významnosti α, když $B \geq \chi^2_{1-\alpha}(r-1)$.

(Bartlettův test je poměrně slabý a je citlivý na porušení normality. Nedá se použít pro malé rozsahy výběrů.)

Post – hoc metody mnohonásobného porovnávání

Zamítneme-li na hladině významnosti α hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti α, tj. na hladině významnosti α testujeme H_0: $\mu_l = \mu_k$ proti H_1: $\mu_l \neq \mu_k$ pro všechna l, k = 1, .., r, l ≠ k.

a) Mají-li všechny výběry týž rozsah p (říkáme, že třídění je vyvážené), použijeme Tukeyovu metodu. Testová statistika má tvar $\dfrac{|M_{k.} - M_{l.}|}{\frac{S_*}{\sqrt{p}}}$. Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když $\dfrac{|M_{k.} - M_{l.}|}{\frac{S_*}{\sqrt{p}}} \geq q_{1-\alpha}(r, n-r)$, kde hodnoty $q_{1-\alpha}(r, n\text{-}r)$ jsou kvantily studentizovaného rozpětí a najdeme je ve statistických tabulkách. (Studentizované rozpětí je náhodná veličina $Q = \dfrac{X_{(n)} - X_{(1)}}{s}$.)

Existuje modifikace Tukeyovy metody pro nestejné rozsahy výběrů, nazývá se Tukeyova HSD metoda. V tomto případě má testová statistika tvar $\dfrac{|M_{k.} - M_{l.}|}{S_*\sqrt{\frac{1}{2}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}}$. Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když $\dfrac{|M_{k.} - M_{l.}|}{S_*\sqrt{\frac{1}{2}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}} \geq q_{1-\alpha}(r, n-r)$.

b) Nemají-li všechny výběry stejný rozsah, použijeme Schefférho metodu: rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když

$$|M_{k.} - M_{l.}| \geq S_*\sqrt{(r-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right)F_{1-\alpha}(r-1, n-r)}.$$

Výhodou Schefférho testu je, že k jeho provedení nepotřebujeme speciální statistické tabulky s hodnotami kvantilů studentizovaného rozpětí, ale stačí běžné statistické tabulky s kvantila Fisherova – Snedecorova rozložení.

V případě vyváženého třídění, kdy lze aplikovat Tukeyovu i Schefférho metodu, použijeme tu, která je citlivější. Tukeyova metoda tedy bude výhodnější, když $q_{1-\alpha}^2(r, n\text{-}r) < 2(r\text{-}1)F_{1-\alpha}(r\text{-}1, n\text{-}r)$.

Metody mnohonásobného porovnávání mají obecně menší sílu než ANOVA.

Může nastat situace, kdy při zamítnutí H_0 nenajdeme metodami mnohonásobného porovnávání významný rozdíl u žádné dvojice středních hodnot. K tomu dochází zvláště tehdy, když p-hodnota pro ANOVU je jen o málo nižší než zvolená hladina významnosti. Pak slabší test patřící do skupiny metod mnohonásobného porovnávání nemusí odhalit žádný rozdíl.

Plánované porovnávání - testování významnosti kontrastů

Plánované porovnávání je navrženo před prováděním ANOVY. Provádí se pomocí kontrastů, tj. pomocí lineárních kombinací středních hodnot. Kontrast $q = \sum_{i=1}^{r} c_i \mu_i$, kde $\sum_{i=1}^{r} c_i = 0$ a $\sum_{i=1}^{r} c_i^2 > 0$. Odhadem kontrastu je veličina $\hat{q} = \sum_{i=1}^{r} c_i M_{i.}$. Testování H_0: $q = 0$ proti H_1: $q \neq 0$ je založeno na statistice $F_q = \dfrac{\hat{q}^2}{S_*^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}}$, která se v případě platnosti H_0 řídí rozložením F(1, n-r). Nulovou hypotézu zamítáme na hladině významnosti α, když platí $F_q \geq F_{1-\alpha}(1,n-r)$.

Porovnávání s kontrolou

V tomto případě neporovnáváme jednotlivé skupiny mezi sebou, ale každou skupinu porovnáme s kontrolní skupinou. Na hladině významnosti α testujeme H_0: $\mu_i = \mu_{kontrola}$ proti H_1: $\mu_i \neq \mu_{kontrola}$. Provedeme tedy celkem r-1 porovnání. Používá se Dunettův test, jehož testové kritérium je $\dfrac{|M_{i.} - M_{kontrola}|}{S_*}$. Nulovou hypotézu zamítáme na hladině významnosti α, když $\dfrac{|M_{i.} - M_{kontrola}|}{S_*} \geq q_{1-\alpha}(r, n-r)$.

Příklad: U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky (v kg):

odrůda	hmotnost
A	0,9 0,8 0,6 0,9
B	1,3 1,0 1,3
C	1,3 1,5 1,6 1,1 1,5
D	1,1 1,2 1,0

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

Řešení:

Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné.
Vypočítáme výběrové průměry v jednotlivých výběrech:
$M_{1.} = 0{,}8$, $M_{2.} = 1{,}2$, $M_{3.} = 1{,}4$, $M_{4.} = 1{,}1$,
celkový průměr $M_{..} = 1{,}14$,
výběrové rozptyly:
$S_1^2 = 0{,}02$, $S_2^2 = 0{,}03$, $S_3^2 = 0{,}04$, $S_4^2 = 0{,}01$,
vážený průměr výběrových rozptylů:

$$S_*^2 = \frac{\sum_{i=1}^{r}(n_i - 1)S_i^2}{n-r} = \frac{3\cdot 0{,}02 + 2\cdot 0{,}03 + 4\cdot 0{,}04 + 2\cdot 0{,}01}{11} = \frac{3}{110} = 0{,}0\overline{27},$$

reziduální součet čtverců: $S_E = (n-r)S_*^2 = 11\cdot\frac{3}{110} = 0{,}3$,

skupinový součet čtverců:

$$S_A = \sum_{i=1}^{r} n_i (M_{i.} - M_{..})^2 =$$

$$= 4\cdot(0{,}8-1{,}14)^2 + 3\cdot(1.2-1{,}14)^2 + 5\cdot(1{,}4-1{,}14)^2 + 3\cdot(1{,}1-1{,}14)^2 = 0{,}816$$

celkový součet čtverců: $S_T = S_A + S_E = 0{,}816 + 0{,}3 = 1{,}116$,

testová statistika $F_A = \frac{S_A/f_A}{S_E/f_E} = \frac{0{,}816/3}{0{,}3/11} = 9{,}97$,

Kritický obor $W = \langle F_{0,95}(3,11), \infty) = \langle 3{,}59, \infty)$. Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,05.

Vypočteme poměr determinace: $P^2 = \frac{S_A}{S_T} = \frac{0{,}816}{1{,}116} = 0{,}7312$

Výsledky zapíšeme do tabulky ANOVA:

Zdroj variability	Součet čtverců	Stupně volnosti	podíl	F_A
skupiny	$S_A = 0{,}816$	3	$S_A/3 = 0{,}272$	$\frac{S_A/(r-1)}{S_E/(n-r)} =$ 9,97
reziduální	$S_E = 0{,}3$	11	$S_E/11 = 0{,}02727$	-
celkový	$S_T = 1{,}116$	14	-	-

Nyní pomocí Scheffého metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

Srovnávané odrůdy	Rozdíly $\lvert M_{k.} - M_{l.}\rvert$	Pravá strana vzorce
A, B	0,4	0,41
A, C	0,67	0,36
A, D	0,3	0,41

B, C	0,2	0,40
B, D	0,1	0,44
C, D	0,3	0,40

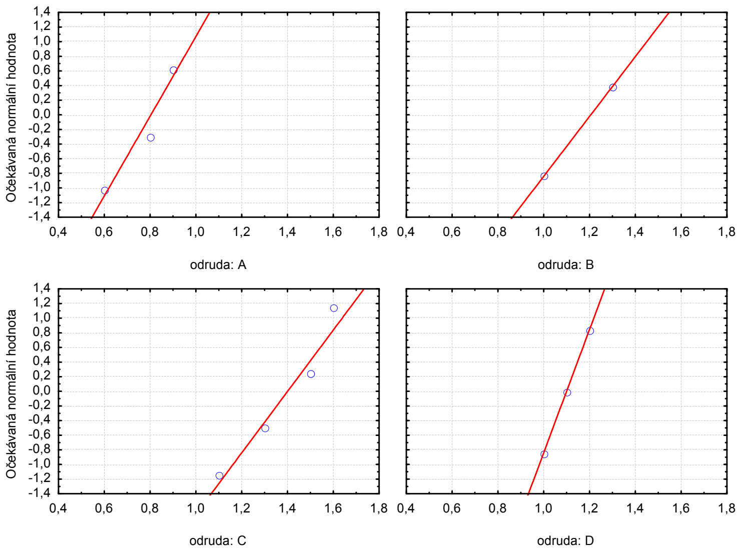
Na hladině významnosti 0,05 se liší odrůdy A a C.

Řešení pomocí systému STATISTICA

Otevřeme nový datový soubor o dvou proměnných X a odrůda a 15 případech. Do proměnné X zapíšeme zjištěné hmotnosti, do proměnné odrůda kódy pro dané odrůdy (1 pro A, 2 pro B, 3 pro C a 4 pro D).

	1 X	2 odruda
1	0,9	A
2	0,8	A
3	0,6	A
4	0,9	A
5	1,3	B
6	1	B
7	1,3	B
8	1,3	C
9	1,5	C
10	1,6	C
11	1,1	C
12	1,5	C
13	1,1	D
14	1,2	D
15	1	D

Ověříme normalitu daných čtyř náhodných výběrů pomocí N-P plotu:



Odchylky od normality jsou jen nepatrné.

Vypočteme výběrové průměry a výběrové rozptyly:

Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – X, Grupovací - odrůda – OK – Skupiny tabulek - zaškrtneme Rozptyly - Výpočet.

Rozkladová tabulka popisných statistik (priklad8301)
N=15 (V seznamu záv. prom. nejsou ChD)

odruda	X průměr	X N	X Sm.odch.	X Rozptyl
A	0,800000	4	0,141421	0,020000
B	1,200000	3	0,173205	0,030000
C	1,400000	5	0,200000	0,040000
D	1,100000	3	0,100000	0,010000
Vš.skup.	1,140000	15	0,282337	0,079714

Nyní ověříme předpoklad shody rozptylů.
Na záložce Skupiny tabulek zaškrtneme Levenův test – Výpočet.

Leveneův test homogenity rozpylů (priklad8301)
Označ. efekty jsou význ. na hlad. p < ,05000

Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	0,018667	3	0,006222	0,065333	11	0,005939	1,047619	0,410027

Vidíme, že p-hodnota Levenova testu je 0,41, tedy větší než hladina významnosti 0,05. Hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

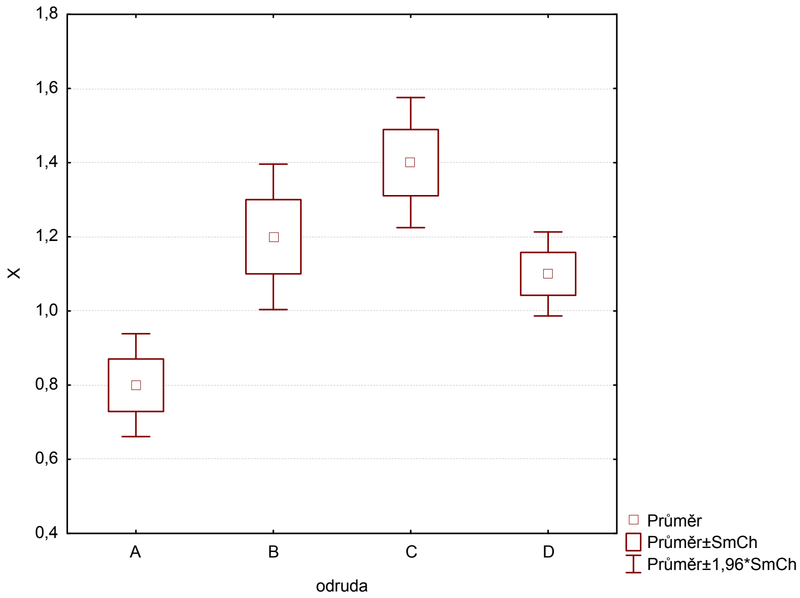
Přistoupíme k testu hypotézy o shodě středních hodnot.
Na záložce Skupiny tabulek zaškrtneme Analýza rozptylu – Výpočet.

Analýza rozptylu (priklad8301)
Označ. efekty jsou význ. na hlad. p < ,05000

Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	0,816000	3	0,272000	0,300000	11	0,027273	9,973333	0,001805

Jelikož p-hodnota = 0,001805 je menší než hladina významnosti 0,05, hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05.

Výpočet doplníme krabicovými diagramy:



Nyní aplikujeme Schefféhо metodu mnohonásobného porovnávání, abychom zjistili, které dvojice odrůd se liší na hladině významnosti 0,05. Na záložce Post – hoc zvolíme Schefféův test.

	Scheffeho test; proměn.:X (priklad8301) Označ. rozdíly jsou významné na hlad. p < ,05000			
odruda	{1} M=,80000	{2} M=1,2000	{3} M=1,4000	{4} M=1,1000
A {1}		0,059165	0,001950	0,190463
B {2}	0,059165		0,464537	0,905502
C {3}	0,001950	0,464537		0,163499
D {4}	0,190463	0,905502	0,163499	

Tabulka obsahuje p-hodnoty pro vzájemné porovnání středních hodnot hmotnosti všech čtyř odrůd. Vidíme, že na hladině významnosti 0,05 se liší odrůdy A, C.

Význam předpokladů v analýze rozptylu

a) Nezávislost jednotlivých náhodných výběrů – velmi důležitý předpoklad, musí být splněn, jinak dostaneme nesmyslné výsledky.
b) Normalita – ANOVA není příliš citlivá na porušení normality, zvlášť pokud mají všechny výběry rozsah nad 20 (důsledek centrální limitní věty). Při výraznějším porušení normality se doporučuje Kruskalův – Wallisův test.
c) Shoda rozptylů – mírné porušení nevadí, při větším se doporučuje Kruskalův – Wallisův test. Test shody rozptylů má smysl provádět až po ověření předpokladu normality.