__Analýza variance (ANOVA) - jednocestná; faktor s pevným efektem; mnohonásobná srovnání__

1. Analýzu variance (ANOVu) používáme při studiu problémů, kdy máme závislou proměnou spojitého typu a nezávislé proměnné jsou kategoriální (faktory). Faktory jsou dvojího typu, pevné (fixní) a náhodné, a závisí na nich to, jak se ANOVA vypočítá (kromě jednocestné ANOVy, kdy je výpočet shodný jak pro model s pevnými tak i pro model s náhodnými faktory).

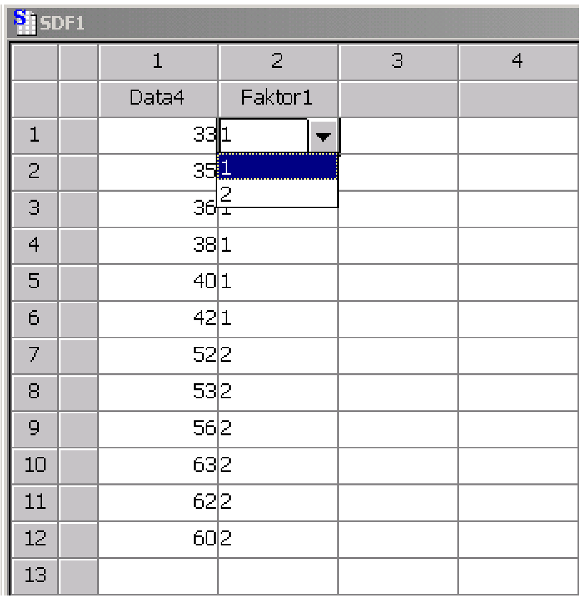
V prvním příkladu studujeme, zda výška rostliny závisí na množství zálivky. Při pokusu zaléváme jednu skupinu rostlin standardním množstvím vody a druhou skupinu dvojnásobným množstvím.

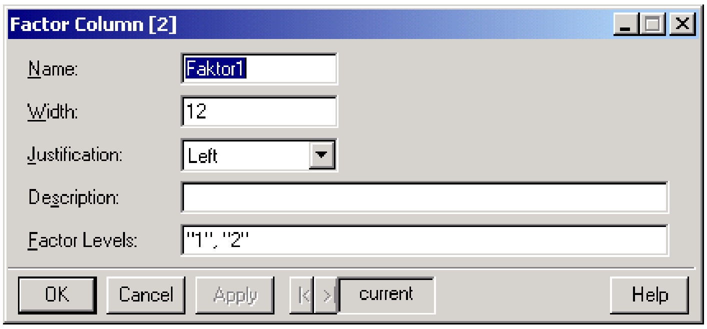
zálivka	1	1	1	1	1	1	2	2	2	2	2	2
výška rostlin [cm]	33	35	36	38	40	42	52	53	56	63	62	60

Zálivka: 1 - kontrola, 2 - dvojnásobná zálivka
Pokud data neimportujeme či nezadáváme přímo do tabulky datové sestavy, můžeme použít příkazového řádku. Použijeme sestavu SDF1, kdy proměnná Data4 obsahuje výšky rostlin a Faktor1 jsou úrovně zálivky.

```
SDF1$Data4<-c(33, 35, 36, 38, 40, 42, 52, 53, 56, 63, 62, 60)
SDF1$Faktor1<-factor(rep(c(1,2), c(6, 6)))
```

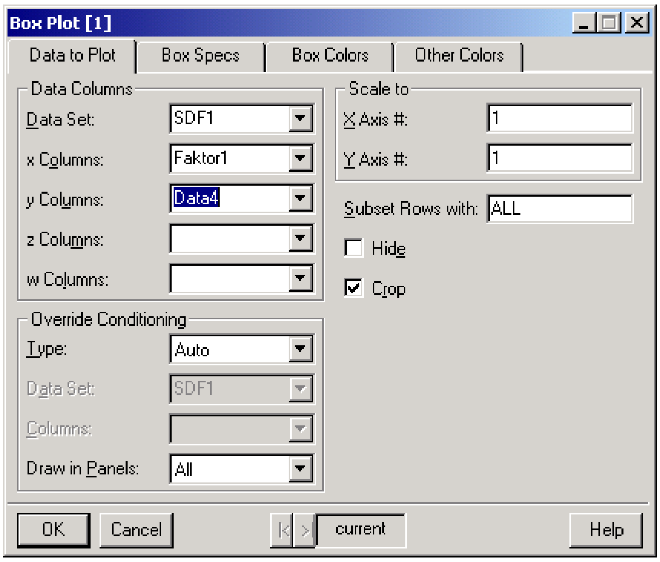
funkce `rep` má dva parametry: opakované číslo (či sekvence čísel) a počet opakování

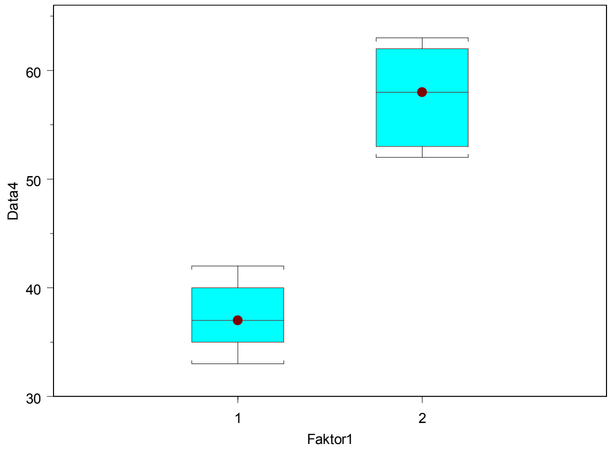




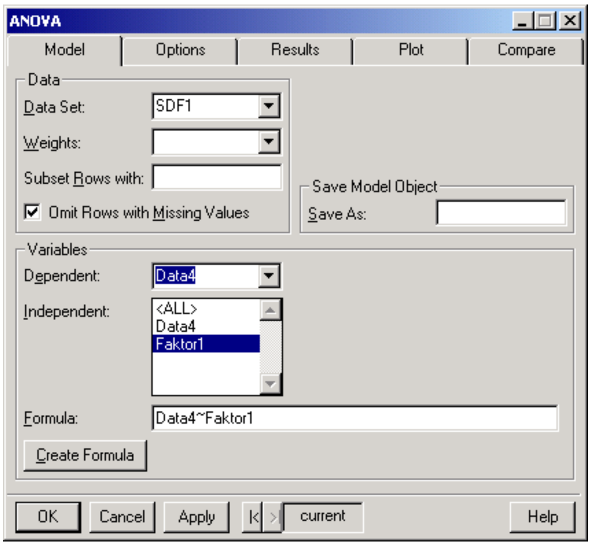
Výsledná datová sestava je na obrázku. Vzhledem k tomu, že jsme proměnnou "Faktor1" definovali jako typ "factor", po kliknutí do pole jsou automaticky k dispozici použité hladiny faktoru. Další hladiny faktoru můžeme doplnit do pole "Factor Levels", toto okno je k dispozici po kliknutí pravým tlačítkem do sloupce a zvolení položky Properties... . Při importu dat je důležité převést proměnné s faktorem na datový typ "faktor".

Základem analýzy variance je stanovit, na jaké hladině pravděpodobnosti chyby I. řádu můžeme zamítnout, že míra zálivky nemá vliv na růst rostlin (nulová hypotéza).
Před samotnou analýzou je vhodné provést alespoň grafickou kontrolu zadaných dat. Krabicový diagram pro hladiny faktoru nalezneme v menu Graph>2D Plot...>Box Plot (x, grouping-optional) a poté do pole "x Column" jako shlukující proměnnou Faktor1.



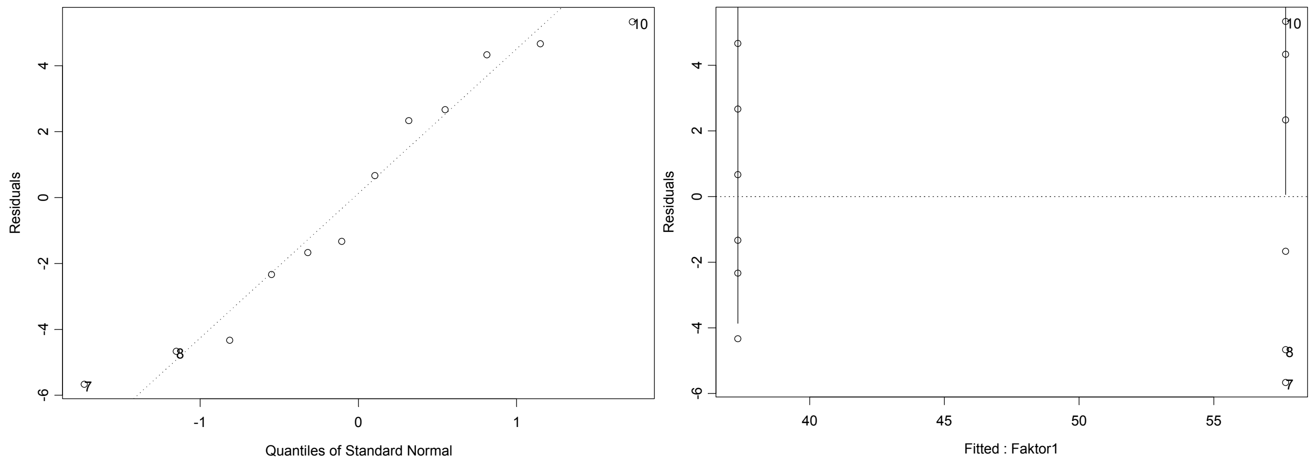


Náš příklad se zalévanými rostlinami je model ANOVy s pevnými efekty. V S+ je tato analýza pod menu Statistics>ANOVA>Fixed Effects... Závislá proměnná je Data4, nezávislá Faktor1. Po výběru proměnných se automaticky vytvoří požadovaný vzorec, v našem případě Data4~Faktor1. Pokud potřebujeme data transformovat, použijeme široké možnosti tvorby vzorců pod nabídkou "Create Formula".



Na záložce "Results" zaškrtneme "Short Output" , "Type I Sum of Squares" a "Means". Dále bychom měli zkontrolovat, jak vypadají reziduály z analýzy. Pokud s nimi budeme dále pracovat, hodnoty reziduálů uložíme zaškrtnutím položky "Saved Results>Residuals" a vybráním Datové sestavy kam budou uloženy. Na záložce "Plot" zaškrtneme "Residuals vs. Fit" a "Residuals Normal QQ". K záložce "Compare" se vrátíme později, až budeme řešit složitější design než v případě jednocestné ANOVy s dvěma hladinami faktoru.

Výsledek analýzy:



Grafické posouzení reziduálů ukazuje, že není narušen předpoklad normality rozložení reziduálů a také variance reziduálů pro každý faktor nejsou výrazně odlišné. Smyslem ANOVy je zjistit, zda je variance mezi skupinami (mezi hladinami faktoru) signifikantně vyšší než variance uvnitř skupin (v rámci jednotlivých hladin faktoru). Testujeme tedy podíl variance mezi skupinami a variance uvnitř skupin (F hodnota).

```
*** Analysis of Variance Model ***

Short Output:
Call:
   aov(formula = Data4 ~ Faktor1, data = SDF1, qr = T, na.action
        = na.exclude)

Terms:
                 Faktor1 Residuals
 Sum of Squares 1240.333   164.667
Deg. of Freedom        1        10

Residual standard error: 4.057914
Estimated effects are balanced

          Df Sum of Sq  Mean Sq  F Value         Pr(F)
  Faktor1  1  1240.333 1240.333 75.32389 5.729512e-006
Residuals 10   164.667   16.467

Tables of means
Grand mean

 47.5

 Faktor1
```

```
      1      2
 37.333 57.667
```

Oddíl "Call" shrnuje, co bylo počítáno. Výsledky analýzy jsou v oddílu "Terms". Stupně volnosti jsou pro faktor se dvěma hladinami rovny 1 (počet hladin-1), stupně volnosti pro reziduální variabilitu odpovídají počtu měření (12) sníženému o počet hladin faktoru (2), tedy 10. "Mean Square" je "Sum of Squares" dělená odpovídajícím počtem stupňů volnosti. Testovací kritérium, F hodnota, je průměrný středního čtverce faktoru a průměrného čtverce reziduálů. "Pr (F)" udává hladinu pravděpodobnosti odpovídající získané hodnotě F statistiky. Můžeme tedy zamítnout nulovou hypotézu a na dosažené hladině pravděpodobnosti přijmout alternativní hypotézu, že vyšší míra zálivky znamená průkazně vyšší růst rostlin.

Poslední částí výsledků je tabulka průměrů pro jednotlivé hladiny faktoru.

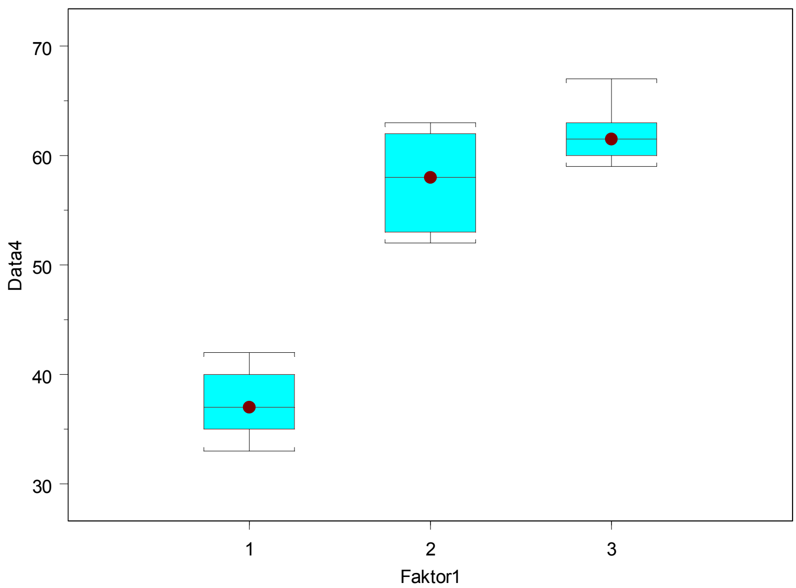
V případě, že naše data nesplňují předpoklady pro použití parametrické ANOVy, je k dispozici neparametrická ANOVA (Kruskal-Wallisova). Ta je založena na pořadí hodnot a testování polohy mediánů. V S+ je dostupná v menu Statistics>Compare Samples>k Samples>Kruskal-Wallis Rank Test.

2. Mnohonásobná porovnání, jednocestná analýza variance s faktorem s pevným efektem

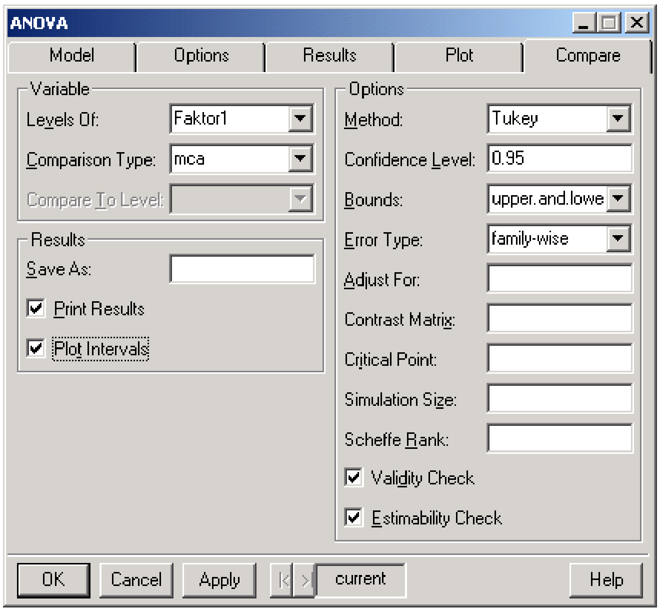
Do předchozího datového souboru přidáme další hladinu faktoru zalévání (trojnásobná zálivka).

zálivka	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3
výška rostlin [cm]	33	35	36	38	40	42	52	53	56	63	62	60	59	61	62	67	60	63

Po zobrazení krabicového diagramu pro hladiny faktoru můžeme předpokládat, že rozdíly budou mezi první hladinou faktoru a zbývajícími dvěma, avšak mezi druhou a třetí již rozdíl nejspíš nebude.



Analýzu zadáme jako v předchozím příkladě. Pro zodpovězení otázky, které hladiny faktoru se mezi sebou liší, musíme spočítat mnohonásobná srovnání. Pokud mnohonásobná srovnání počítáme souběžně s analýzou variance, nalezneme je na záložce "Compare". Pokud je počítáme samostatně, jsou v menu Statistics>ANOVA>Multiple Comparisons... (v tomto případě je však potřeba mít analýzu předem uloženou jako "Model Object"). V tomto okně je několik možností jaké testy provést. Volba testů je různá při párových srovnání před vlastní ANOVou (a priori) a jiné jsou pro párové testy až na základě průkazného výsledku ANOVy (a posteriori). Před použitím těchto testů je vhodné zkontrolovat v nápovědě jak je dané srovnání počítáno. My v tomto příkladě použijeme *Tukey* test. V podokně "Variable" vybereme který faktor studujeme a v podokně "Results" zaškrtneme grafické zobrazení konfidenčních intervalů (Plot Intervals).



Grafy reziduí opět nevykazují nějaké narušení předpokladů ANOVy a výsledek analýzy je vysoce průkazný.

```
          Df  Sum of Sq  Mean Sq  F Value         Pr(F)
  Faktor1  2   2081.333 1040.667 76.27036 1.379691e-008
Residuals 15    204.667   13.644

Faktor1
     1      2      3
 37.333 57.667 62.000
```

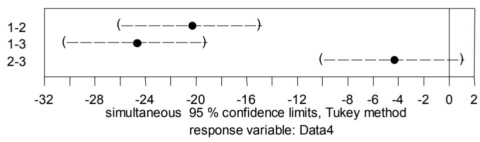
Následná tabulka mnohonásobného srovnání potvrzuje náš předpoklad. Na naší stanovené hladině významnosti 0.05 jsou rozdíly mezi normálním a zvýšenými zalévacími režimy, ale již není rozdíl mezi dvoj a troj násobnou dávkou zálivky.

```
95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.5979
response variable: Data4
```

```
intervals excluding 0 are flagged by '****'

    Estimate Std.Error Lower Bound Upper Bound
1-2   -20.30      2.13      -25.90      -14.80 ****
1-3   -24.70      2.13      -30.20      -19.10 ****
2-3    -4.33      2.13       -9.87        1.21
```



To samé zobrazuje i graf konfidenčních intervalů.

V případě, že máme nás nezajímají mnohonásobná srovnání mezi všemi proměnnými navzájem a jde nám např. o porovnání efektu oproti kontrole, můžeme z nabídky mnohonásobných srovnání vybrat metodu mcc, kdy vybereme co je kontrola.

Pokud chceme spočítat průměry (ale můžeme i jiné charakteristiky) pro určité skupiny použijeme v S+ funkci `tapply` (proměnná, definice skupin, požadovaná charakteristika). V případě předchozího příkladu a spočtení průměrů píšeme: `tapply (Data4, Faktor1, mean)`. Další použití funce `tapply` si ukážeme v příkladu se dvěma faktory.

Výpočty stupňů volnosti v tabulce ANOVY

	df
ošetření	k-1 (k – počet hladin faktoru ošetření)
error	k(n-1) (n – počet opakování v rámci jedné hladiny faktoru)
total	kn-1

Tento materiál byl vytvořen na základě příkladů a textů P. Sklenáře pro kurz biostatistiky v NCSS.