Pravděpodobnost a matematická statistika

Doc. RNDr. Gejza Dohnal, CSc.
dohnal@nipax.cz

Pravděpodobnost a matematická statistika 2010

1.týden (20.09.-24.09.) Data, typy dat, variabilita, frekvenční analýza (histogramy, četnosti absolutní, relativní, prosté, kumulativní), základní statistické charakteristiky (průměr, výběr.rozptyl, minimum, maximum, medián, kvartily, boxplot), sešikmenná rozdělení (vzájemná poloha mediánu a střední hodnoty), chvosty, kvantily

2. týden (27.09.-01.10.) Princip statistické indukce, výběr, vlastnosti výběru, experiment. Náhodná veličina, rozdělení pravděpodobnosti a jeho souvislost s histogramem. Pravděpodobnost, pravidla pro počítání s pravděpodobností, podmíněná pravděpodobnost, závislost náhodných veličin.

3.týden (04.10.-08.10.) Využití závislosti při stanovení pravděpodobnosti - věta o úplné pravděpodobnosti a Bayesova věta

4.týden (11.10.-15.10.) Rozdělení chyb měření - normální rozdělení a počítání s ním. Odhady parametrů normálního rozdělení. Intervaly spolehlivosti pro normální data. Jednovýběrové testy o střední hodnotě

5.týden (18.10.-24.10.) Výběrový poměr jako odhad pravděpodobnosti sledovaného jevu. Alternativní rozdělení, binomické rozdělení. Intervalový odhad výběrového poměru. Výběry s vracením a bez vracení (binomické a hypergeometrické rozdělení)

6.týden (25.10.-29.10.) odpadá

7.týden (01.11.-05.11.) Poruchy v čase (Poissonův proces). Poissonovo rozdělení, exponenciální rozdělení, jeho výhody a nevýhody, modelování doby do poruchy pomocí Weibullova rozdělení, lognormálního rozdělení, případně useknuté normální rozdělení.

8.týden (08.11.-12.11.) Testy dobré shody, Q-Q graf (pouze vysvětlení), testy normality. Některé neparametrické testy

9.týden (15.11.-19.11.) Dvě náhodné veličiny - srovnání dvou výběrů (dvouvýběrové testy)

10. týden (22.11.-26.11.) Dvě náhodné veličiny. Dvourozměrné četnosti jako odhad dvourozměrného rozdělení, frekvenční tabulka. Marginální rozdělení (vše pouze diskrétně s tabulkou)

11. týden (29.11.-03.12.) Závislost náhodných veličin, míry závislosti (kovariance, korelace), test významnosti korelačního koeficientu

12. týden (06.12.-10.12.) Regrese, lineární regresní model (přímková, kvadratická, polynomická regrese), analýza reziduí, pásy spolehlivosti

13. týden (13.12.-17.12.) Více výběrů, jednoduché třídění, ANOVA.

14. týden (20.12.-22.12.) Rezerva, opakování, testy normality (náhrada za 28.10.)

Jednoduché třídění, ANOVA

Celkem N pozorování náhodné veličiny X, rozdělených do k skupin:

$$X_{1,1}, \ldots X_{1,n1},$$
$$X_{2,1}, \ldots X_{2,n2},$$
$$\ldots\ldots$$
$$X_{k,1}, \ldots X_{k,nk},$$ podle nějakého hlediska (faktoru).

Nebo k skupin pozorování nezávislých náhodných veličin $X_1, \ldots X_k$:

$$X_{1,1}, \ldots X_{1,n1},$$
$$X_{2,1}, \ldots X_{2,n2},$$
$$\ldots\ldots$$
$$X_{k,1}, \ldots X_{k,nk}.$$ V obou případech je $N = n_1 + \ldots + n_k$.

Jednoduché třídění, ANOVA

Předpokládáme model: $X_{ij} = \mu + \mu_i + \varepsilon_{ij}$, kde

μ je společná střední hodnota bez ohledu na vliv faktoru (pro všechny náhodné veličiny)

μ_i je vliv i-tého faktoru (i-té náhodné veličiny)

ε_{ij} je "chyba" j-tého pozorování při i-tém faktoru (i-té náhodné veličiny)

H_0: hodnoty faktoru nemají vliv na veličinu X

H_A: hodnty veličiny X závisejí na hodnotách působícího faktoru

Testujeme hypotézu H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, proti alternativní hypotéze H_A: existují indexy i, j tak, že $\mu_i \neq \mu_j$.

Jednoduché třídění, ANOVA

Z napozorovaných dat můžeme odhadnout

společnou střední hodnotu: $$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j} = \bar{x}$$

střední hodnotu při i-tém faktoru: $$\hat{\mu} + \hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} = \bar{x}_i$$

chybu j-tého pozorování při i-tém faktoru (reziduum): $e_{i,j} = x_{i,j} - \bar{x}$

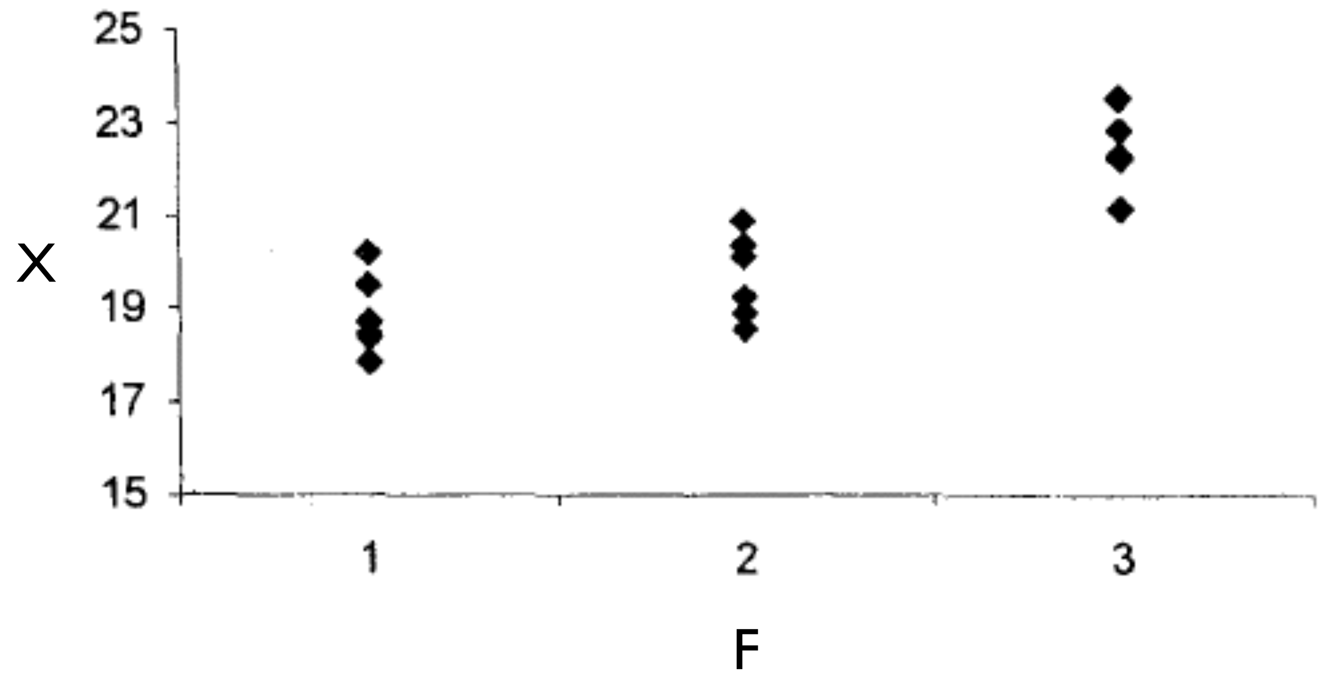
celkovou míru variability: $$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{i,j}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2$$

míru variability způsobenou variabilitou faktorů:

$$SS_F = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

míru zbytkové variability (uvnitř skupin): $SS_E = SS_T - SS_F$

Jednoduché třídění, ANOVA



Zdroj variability	Součet čtverců	Stupně volnosti	Průměrný čtverec	Statistika F	p hodnota
faktor	SS_F	k-1	$MS_F = \frac{SS_F}{k-1}$	$F = \frac{MS_F}{MS_E}$	
reziduální	SS_E	N-k	$MS_E = \frac{SS_E}{N-k}$		
celkový	SS_T	N-1			

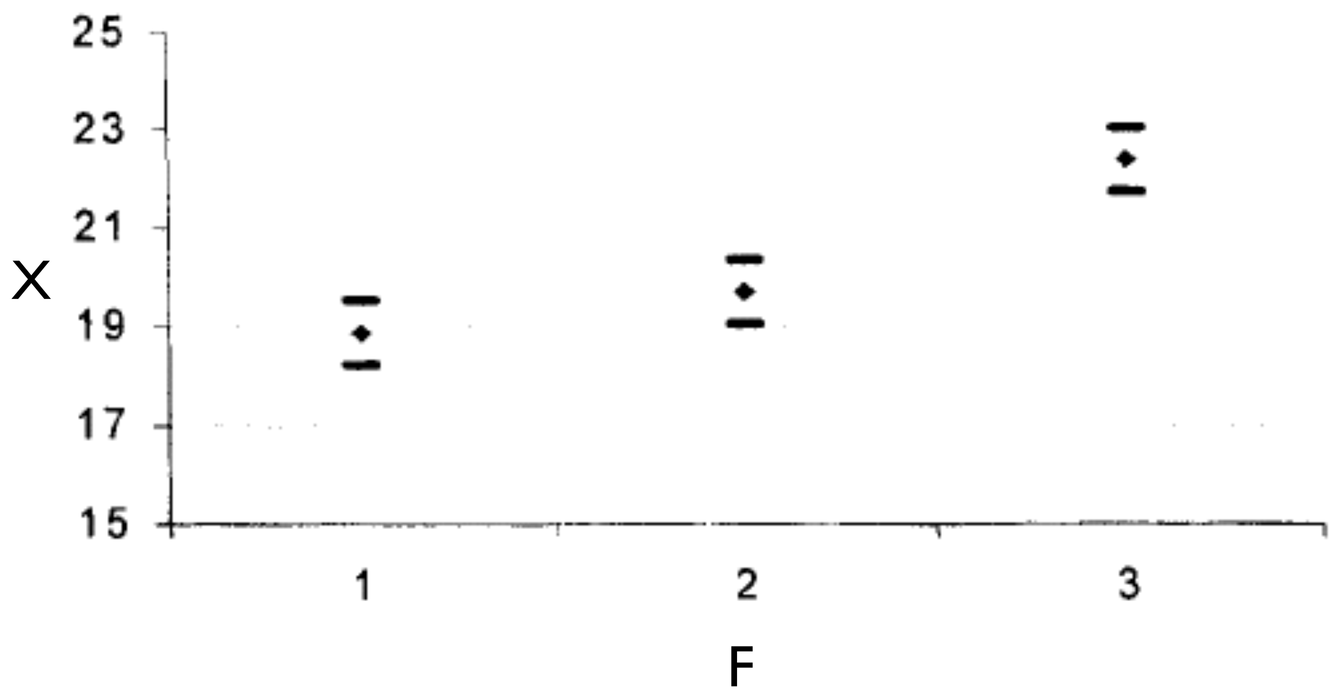
F má Fisherovo rozdělení prpsti

Jednoduché třídění, ANOVA

Bonferroniho metoda mnohonásobného srovnání úrovní faktorů:

Úroveň	Průměr	Dolní mez	Horní mez
A_1	18,867	18,217	19,516
A_2	19,700	19,051	20,349
A_3	22,383	21,734	23,033

$$\bar{x}_i \pm t_{1-\alpha/2}\sqrt{\frac{MS_E}{2k}}$$



Jednoduché třídění, ANOVA

Vliv katalyzátoru na výtěžek chemického procesu

Odezva: výtěžek procesu (množství vyráběné látky)

Faktor: druh katalyzátoru, 4 hodnoty

Počet replikací: 6

Počet měření: 4x6 = 24

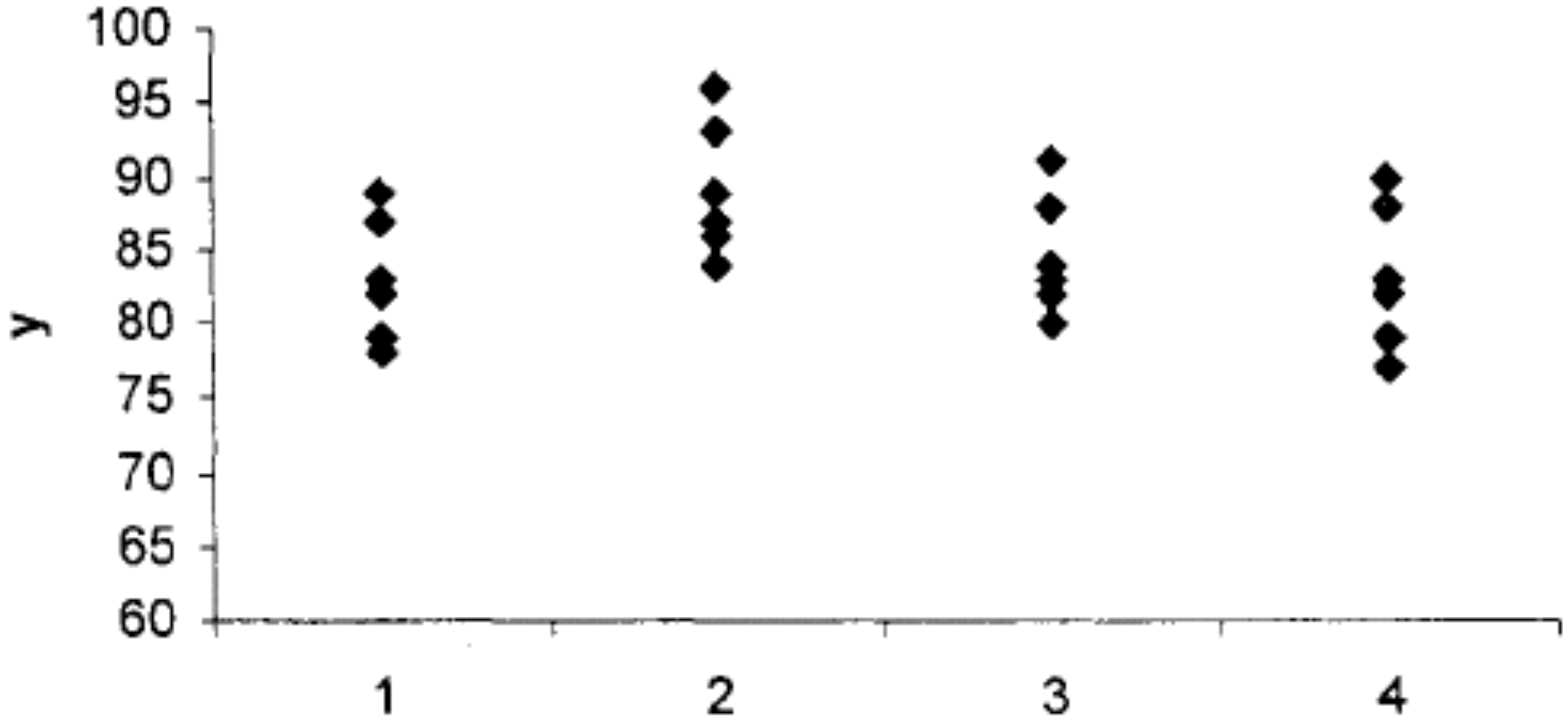
Vedlejší faktor: vliv várky vstupní suroviny

várka	1	2	3	4	5	6	průměr	rozptyl
A1	87	79	82	89	83	78	83	18,8
A2	93	84	89	96	86	87	89,2	20,57
A3	88	80	84	91	83	82	84,7	16,67
A4	88	77	83	90	82	79	83,2	25,37

Metoda vyhodnocení: ANOVA pro 2 faktory (bez opakování)

Jednoduché třídění, ANOVA

Vliv katalyzátoru na výtěžek chemického procesu



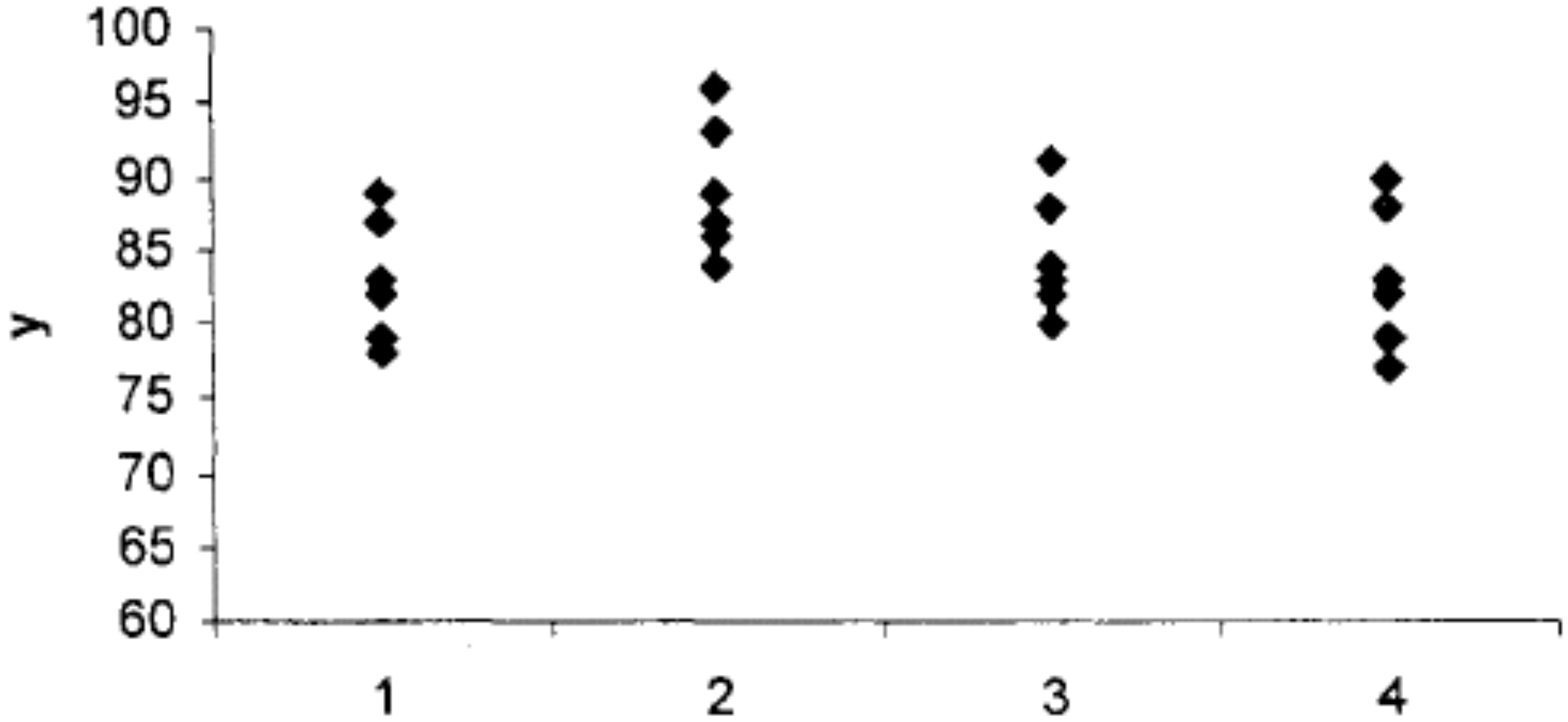
Postup výpočtu: Testujeme hypotézu H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

ANOVA

Zdroj variability	*SS*	*St.vol*	*MS*	*F*	*Hodnota P*	*F krit*
Faktor A	149	3	49,66667	49,66667	5,03E-08	3,287383
Bloky	392	5	78,4	78,4	3,28E-10	2,901295
Reziduální	15	15	1			
Celkový	556	23				

Jednoduché třídění, ANOVA

Vliv katalyzátoru na výtěžek chemického procesu



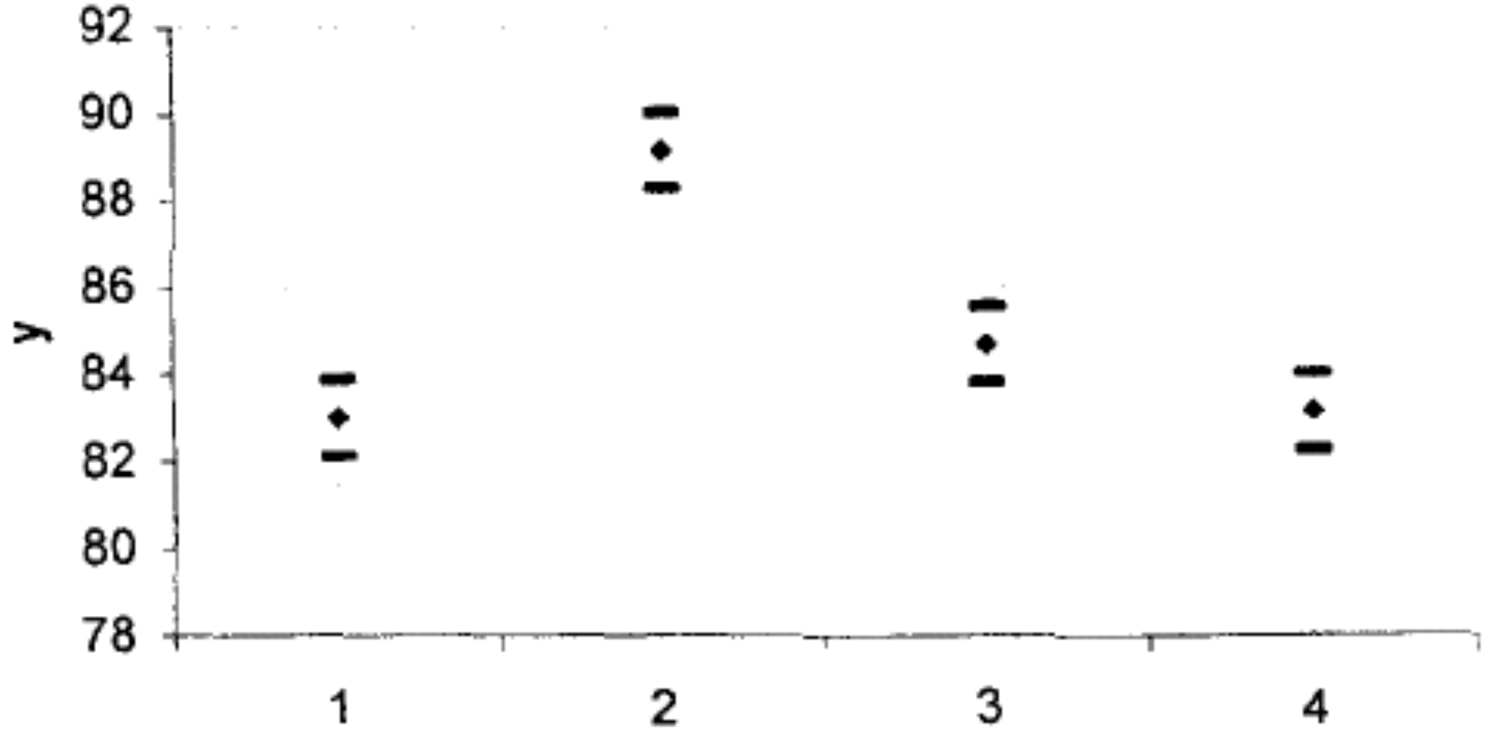
Postup výpočtu: Testujeme hypotézu H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Bonferroniho metoda mnohonásobného srovnání srovnání úrovní faktorů:

Úroveň	Průměr	Dolní mez	Horní mez
A_1	83,0	82,1	83,9
A_2	89,2	88,3	90,0
A_3	84,7	83,8	85,5
A_4	83,2	82,3	84,0

Jednoduché třídění, ANOVA

Vliv katalyzátoru na výtěžek chemického procesu



Postup výpočtu: Testujeme hypotézu H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Bonferroniho metoda mnohonásobného srovnání srovnání úrovní faktorů:

Úroveň	Průměr	Dolní mez	Horní mez
A_1	83,0	82,1	83,9
A_2	89,2	88,3	90,0
A_3	84,7	83,8	85,5
A_4	83,2	82,3	84,0