

Prezentace pro 9. přednášku

6BZST1

Základy statistiky

doc. RNDr. Lenka Komárková, Ph.D.



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

VÝSTUPY Z UČENÍ

Student bude schopen:

- chápat princip metody ANOVA a rozlišovat meziskupinovou a vnitroskupinovou variabilitou
- vyhodnotit závislost kvantitativního znaku na kvalitativním
- využít Bonferroniho korekci v rámci mnohonásobného porovnávání

ANALÝZA ROZPTYLU (ANOVA)

ANALÝZA ROZPTYLU JEDNODUCHÉHO TŘÍDĚNÍ

- porovnání populačních průměrů u více než dvou nezávislých výběrů
 - zobecnění dvouvýběrového t-testu
- analýza závislosti kvantitativního znaku na kvalitativním
 - kvalitativní znak se v kontextu analýzy rozptylu nazývá **faktor**
- **příklady:**
 - *Závisí výše platu na dosaženém vzdělání (ZŠ, SŠ, VŠ)?*
 - *Závisí oktanové číslo benzínu na dodavateli (A, B, C, D)?*
- zkratka pro analýzu rozptylu: **ANOVA**
 - **AN**ALYSIS **O**F **VA**RIANCE
- **ANOVA jednoduchého třídění** \equiv **jednofaktorová ANOVA**
 - one-way ANOVA

ANOVA

VYUŽITÍ

- **V zemědělství**
 - Výnos brambor vs. odrůda
 - Výnos obilí vs. typ hnojení (při kontrole typu pole)
- **V průmyslu**
 - Koksovací čas vs. šířka pece (10 , 20, 30 cm)
 - Pevnost cementu vs. způsob míchání
 - Životnost výrobku vs. metoda výroby
- **V ekonomii**
 - Cena výrobku vs. dodavatel
 - Výše platu vs. vzdělání (popř. i pohlaví)
 - Rodinné výdaje vs. počet členů domácnosti
 - Výkon dělníků vs. směna (ranní, odpolední, noční)
- **V medicíně**
 - Doba léčby vs. způsob léčby
 - Srážlivost (koagulace) krve vs. dieta

ANOVA

TESTOVÝ PROBLÉM

- **Nulová hypotéza H_0 :**
 - populační průměry **jsou shodné**
 - střední hodnoty **se rovnají**
 - kvantitativní znak **nezávisí** na kvalitativním znaku

- **Alternativní hypotéza H_1 :**
 - populační průměry **nejsou shodné**
 - **alespoň jednou** se střední hodnoty **liší**
 - kvantitativní znak **závisí** na kvalitativním znaku

MODEL ANOVA

$$\text{Model: } Y_i = \mu_i + \varepsilon$$

- Y_i kvantitativní znak pro i-tý výběr (skupinu)
- μ_i populační průměr (střední hodnota) pro i-tý výběr
- ε náhodná chyba (nevysvětlitelná složka)

Testový problém:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots$$

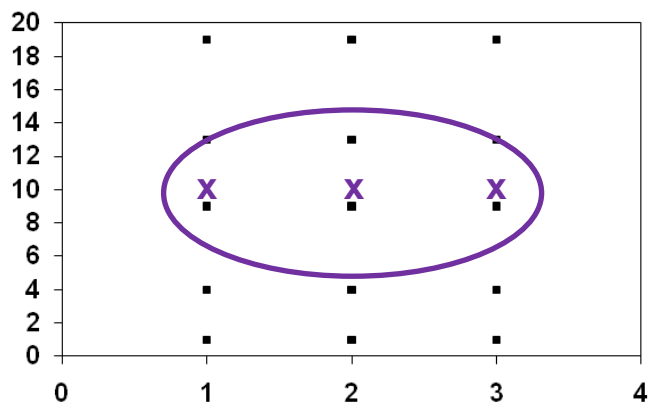
$$H_1: \mu_1 \neq \mu_2 \text{ nebo } \mu_1 \neq \mu_3 \text{ nebo } \mu_2 \neq \mu_3 \dots$$

ANOVA

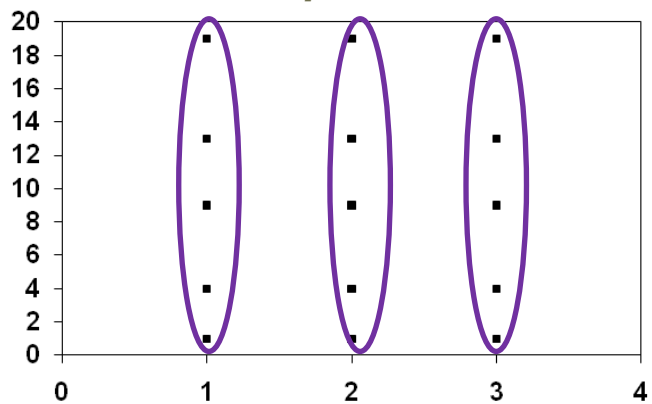
ZÁKLADNÍ PRINCIP

Porovnávání meziskupinové a vnitroskupinové variability

SHODA MEZI SKUPINOVÝMI PRŮMĚRY

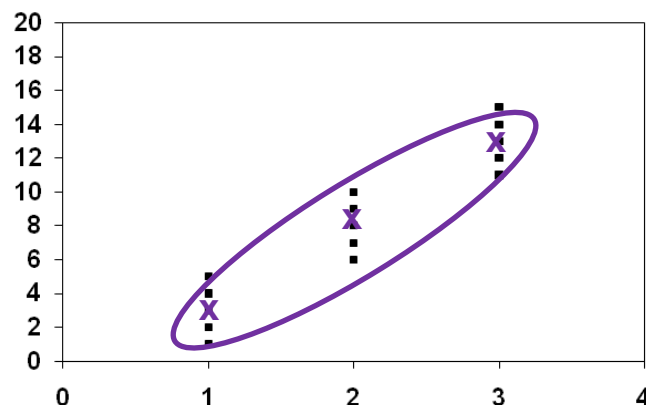


Malá meziskupinová variabilita

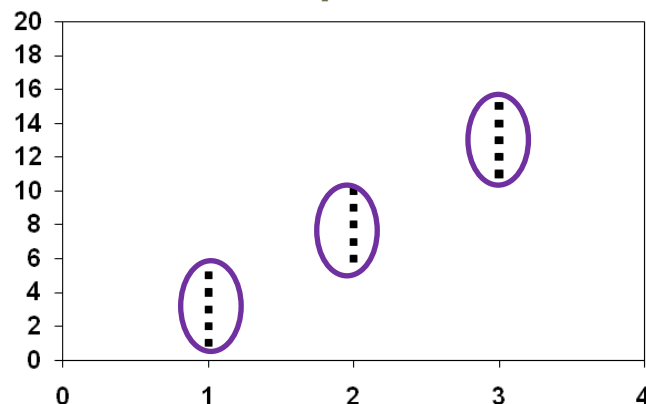


Velká vnitroskupinová variabilita

ROZDÍL MEZI SKUPINOVÝMI PRŮMĚRY



Velká meziskupinová variabilita



Malá vnitroskupinová variabilita

ANOVA

ZÁKLADNÍ PRINCIP

- **Meziskupinová (*between*) variabilita**
 - Variabilita průměrů mezi jednotlivými skupinami
- **Vnitroskupinová (*within*) variabilita**
- **≡ reziduální (*residual*) variabilita**
 - Variabilita pozorování uvnitř každé skupiny
- **Porovnání těchto 2 zdrojů variability**
 - Test je založen na podílu meziskupinové a vnitroskupinové variability

ANOVA

PŘÍKLAD

- Závisí čas nutný ke koksování na šířce pece, v níž se koksování provádí?
- Kvantitativní znak (odezva) – koksovací čas
- Kvalitativní znak (faktor) – šířka pece (*malá, střední, velká*)

Koksovací čas pro pece různé šířky

Šířka pece (cm)	Koksovací časy (min)					
10	3,5	3,0	2,7	2,2	2,3	2,4
20	7,1	6,9	7,5	5,2	4,6	6,8
30	10,8	10,6	11,0	7,6	7,1	7,3

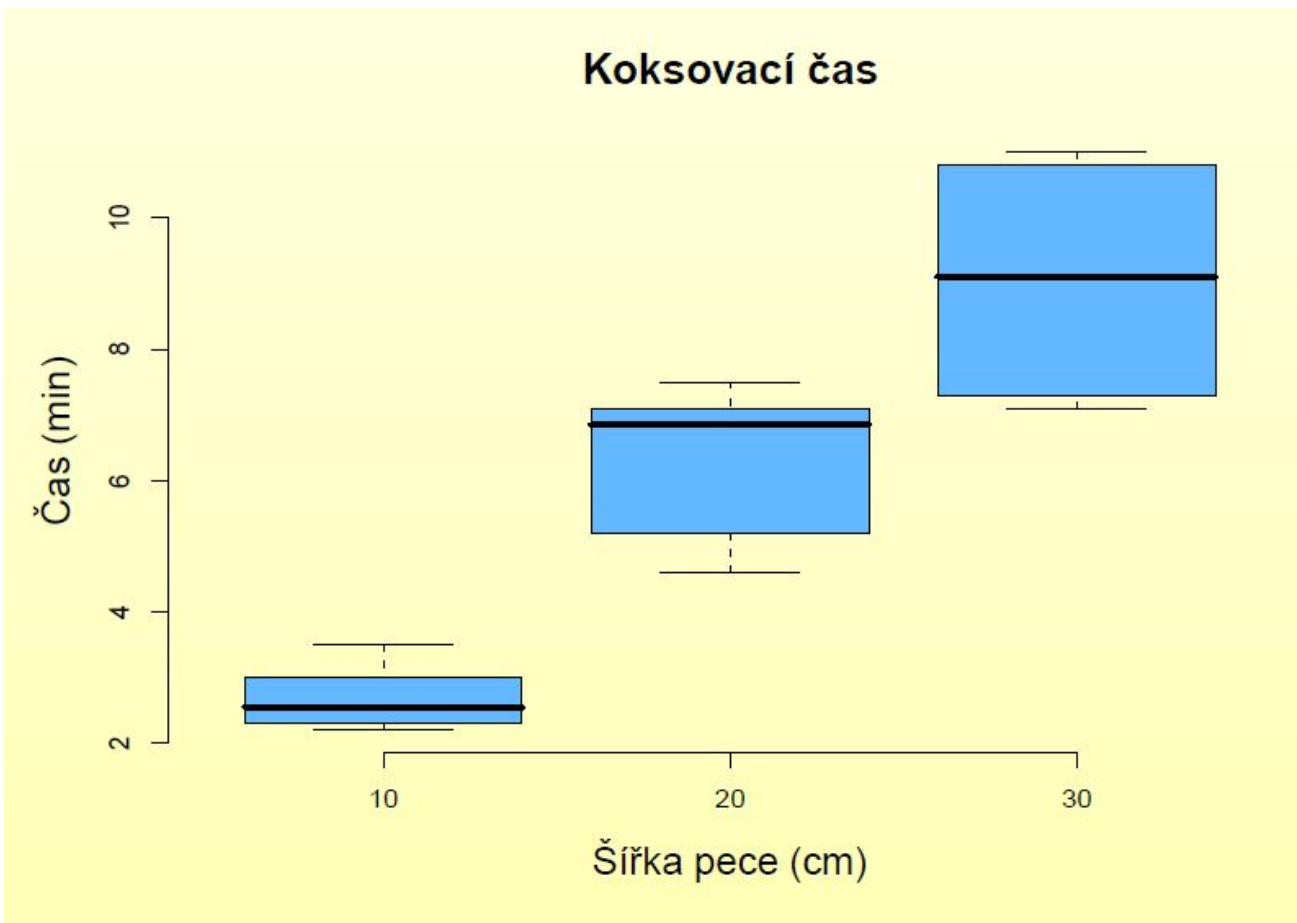
Zdroj: Johnson, R. A. (1994). *Miller and Freund's Probability and Statistics for Engineers, Fifth edition*. Prentice-Hall.

ANOVA

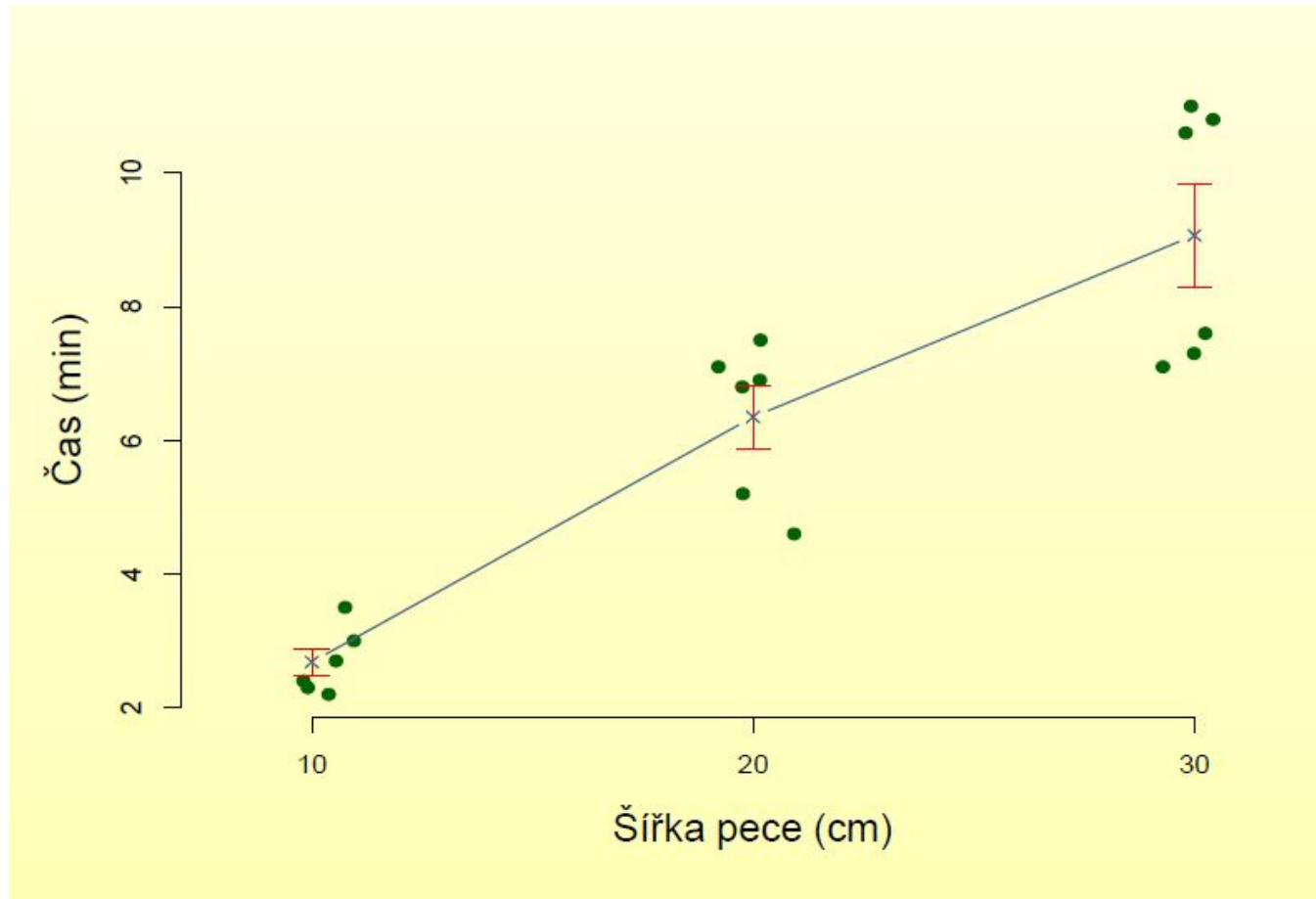
PŘÍKLAD

- $H_0: \mu_1 = \mu_2 = \mu_3$
 - Průměrný koksovací čas je pro všechny tři šířky pece stejný.
 - Průměrný koksovací čas nezávisí na šířce pece.
- $H_1: \mu_1 \neq \mu_2$ nebo $\mu_1 \neq \mu_3$ nebo $\mu_2 \neq \mu_3$
 - Průměrný koksovací čas je pro alespoň jednu šířku pece jiný.
 - Průměrný koksovací čas závisí na šířce pece.

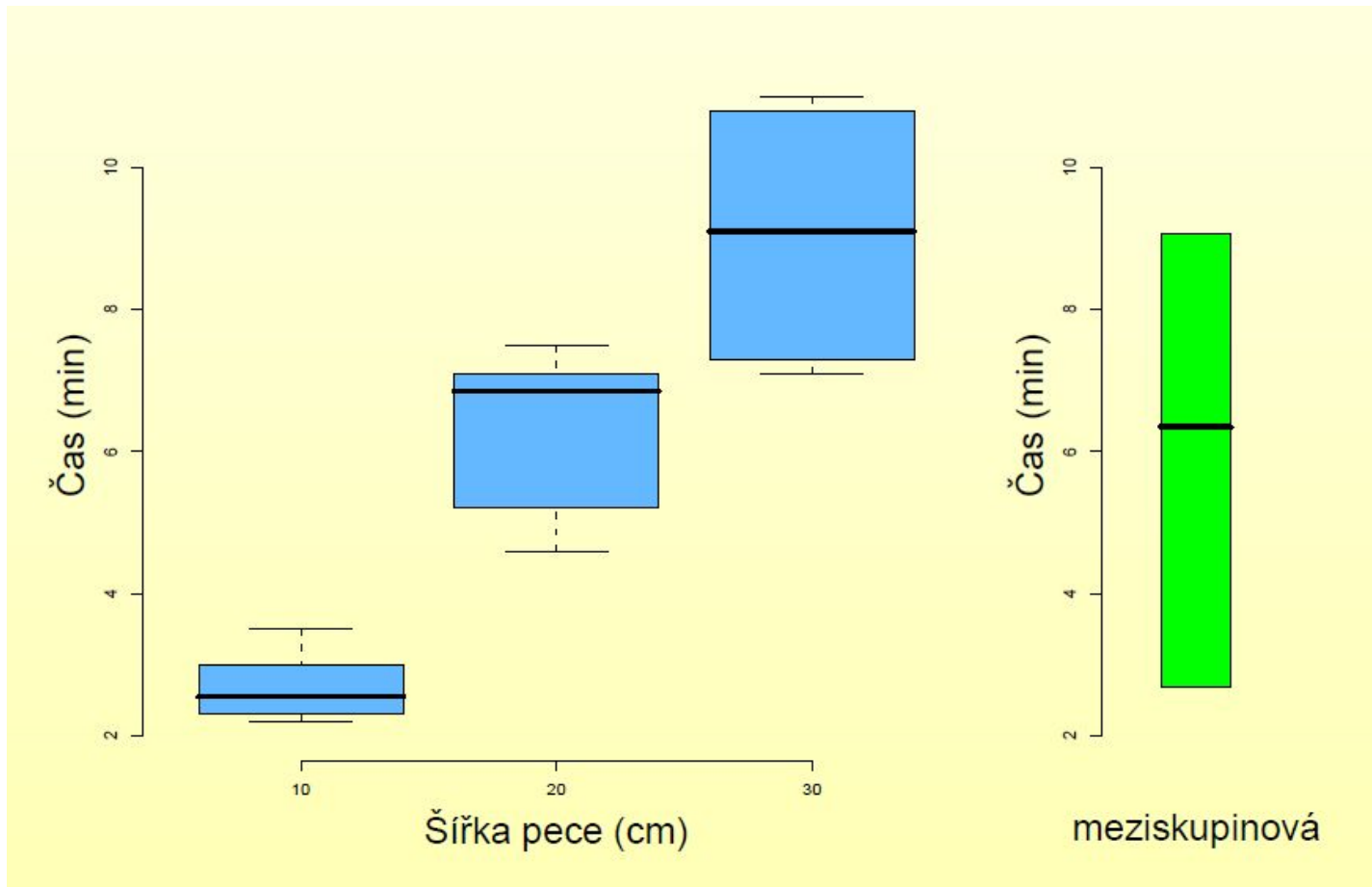
KRABICOVÉ DIAGRAMY



BODOVÉ GRAFY



MEZI- VS. VNITROSKUPINOVÁ VARIABILITA



ANOVA

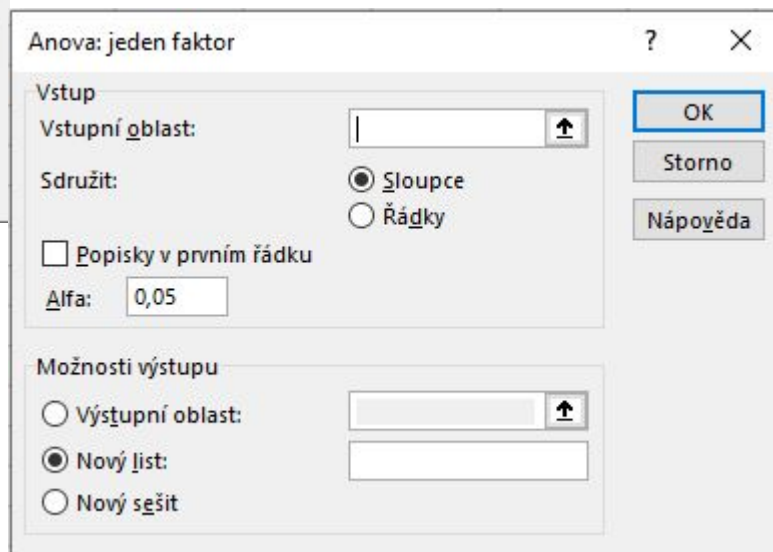
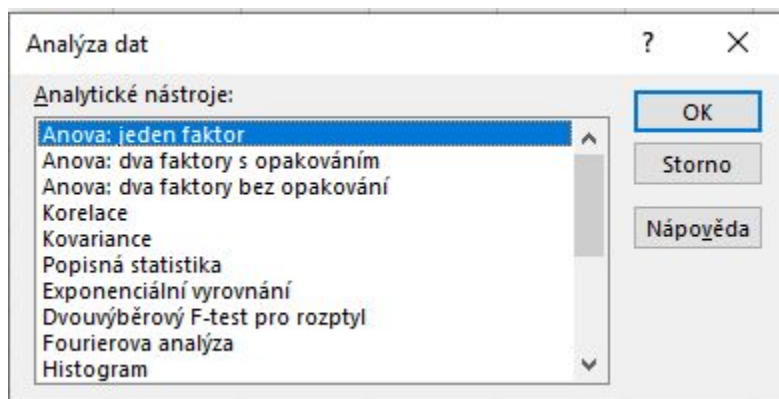
VÝSLEDEK TESTU

Šířka pece (cm)	Koksovací časy (min)						Průměr
10	3,5	3,0	2,7	2,2	2,3	2,4	2,683
20	7,1	6,9	7,5	5,2	4,6	6,8	6,350
30	10,8	10,6	11,0	7,6	7,1	7,3	9,067

- Testová statistika **F = 35,202**
- Dosažená hladina testu **p < 0,001**
- Na 5% hladině významnosti jsme prokázali, že čas nutný ke koksování závisí na šířce pece.

ANOVA V MS EXCEL

Doplněk Analýza dat



MNOHONÁSOBNÉ POROVNÁVÁNÍ

- provádí se v případě **zamítnutí nulové hypotézy** o shodě populačních průměrů
- odpovídá na otázku, které skupiny se mezi sebou liší
- v případě použití dvouvýběrových testů **kumulace chyby I. druhu**
 - **Bonferroniho princip**: hladinu významnosti je nutné stanovit jako podíl chyby I. druhu a počtu vytvořených párů
 - **Scheffého, Tukeyova metoda** – méně konzervativní přístupy

BONFERRONIHO KOREKCE

- **Konzervativní** korekce dvouvýběrových testů vzhledem k **mnohonásobnému testování**
- Každý jednotlivý dvouvýběrový test provádíme na hladině **α/M**
 - M počet provedených dvouvýběrových testů
- Dvojici prohlásíme za významně odlišnou, pokud p-hodnota příslušného dvouvýběrového testu je **$\leq \alpha/M$**
 - Je-li počet M vyšší, prohlásíme za signifikantní pouze „**velké**“ rozdíly

SCHEFFÉHO METODA

TUKEYOVA METODA

- **Alternativa** k paralelnímu použití Bonferroniho korekci dvouvýběrových testů
- **Mnohem méně konzervativní**
 - *Podarí se nám prohlásit za signifikantní i „menší“ rozdíly*
- Rozdíl mezi dvěma populačními skupinovými průměry prohlásíme za *signifikantní*, pokud $p \leq \alpha$
- Celková pravděpodobnost chyby I. druhu zůstává **pod požadovanou hladinou α**

ANOVA

PŘEDPOKLADY

- Všechna měření musí být vzájemně **nezávislá**
 - uvnitř skupin
 - mezi skupinami
- Měření v každé skupině jsou **normálně rozdělená** s populačním průměrem μ_i
- Ve všech skupinách mají měření **stejný rozptyl (homoskedasticita)** kolem populačního průměru

ANOVA

PŘEDPOKLADY

Předpoklady o náhodných chybách:

- Chyby jsou nezávislé a náhodně fluktuují kolem 0
- Chyby jsou stejně rozdělené
 - mají normální rozdělení
 - *Ize obejít přes velký počet dat v jednotlivých skupinách*
 - mají stejný rozptyl, tzv. **homoskedasticita**
 - *Ize obejít přes Welchovu modifikaci ANOVA jednoduchého třídění*

Praktické ověřování:

- dělá se většinou přes tzv. rezidua (odchylky hodnot od skupinového průměru)
- zamyšlením se nad mechanismem, který byl použit ke sběru dat nebo k provedení experimentu
 - *Graficky*
 - *Testem*
- **Grafické metody jsou mnohdy cennější než formální test!**

OVĚŘOVÁNÍ PŘEDPOKLADŮ

Normalita

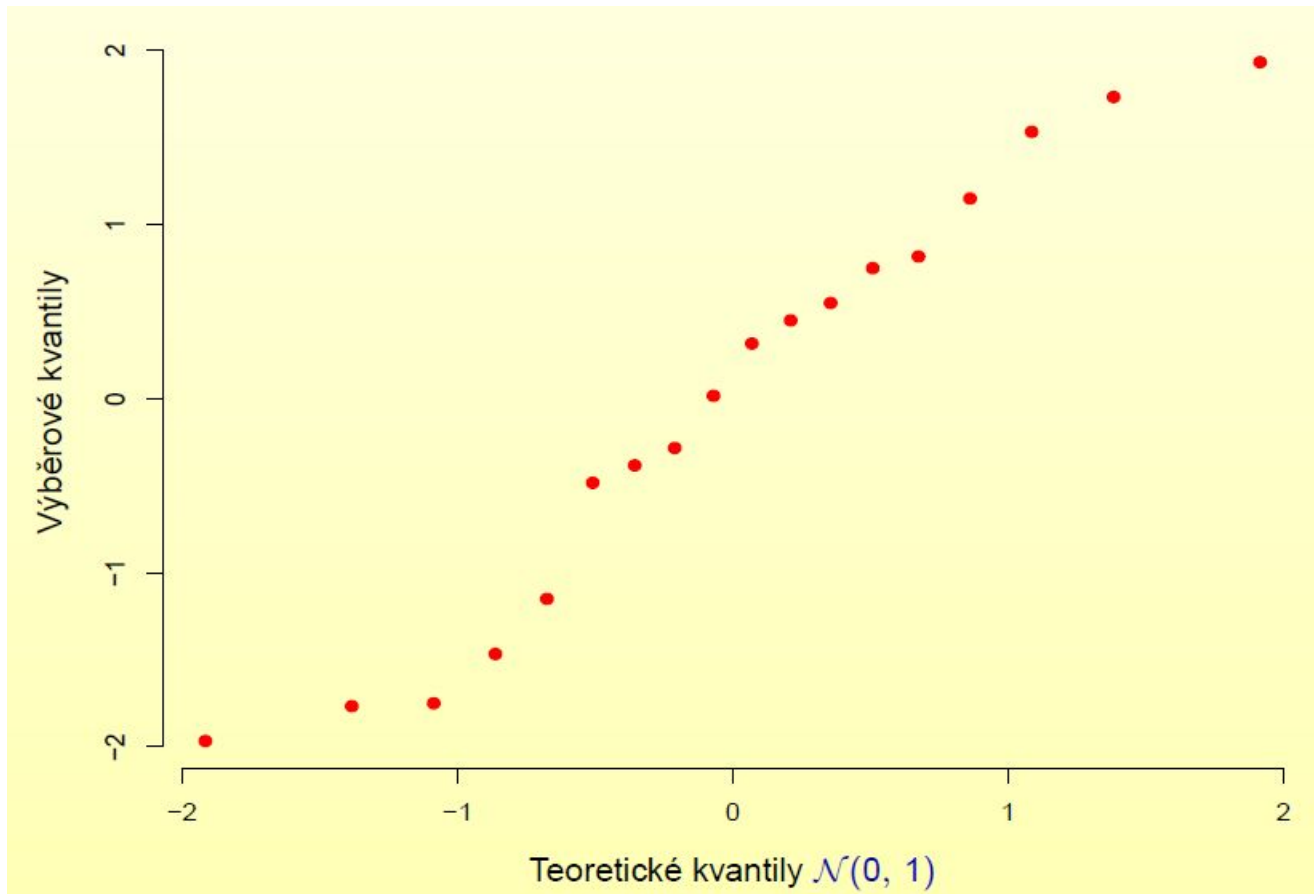
- **Normální QQ graf**
 - *bud'* pozorování pro jednotlivé skupiny
 - *nebo* reziduí
- **Shapiro-Wilkův test**
 - *bud'* pro pozorování zvlášt' pro každou skupinu
 - *nebo* pro rezidua

Homoskedasticita

- **Reziduální graf**
 - rezidua vs. faktor
 - rezidua vs. skupinové průměry
- **Leveneův test**
 - *bud'* pro pozorování
 - *nebo* pro rezidua

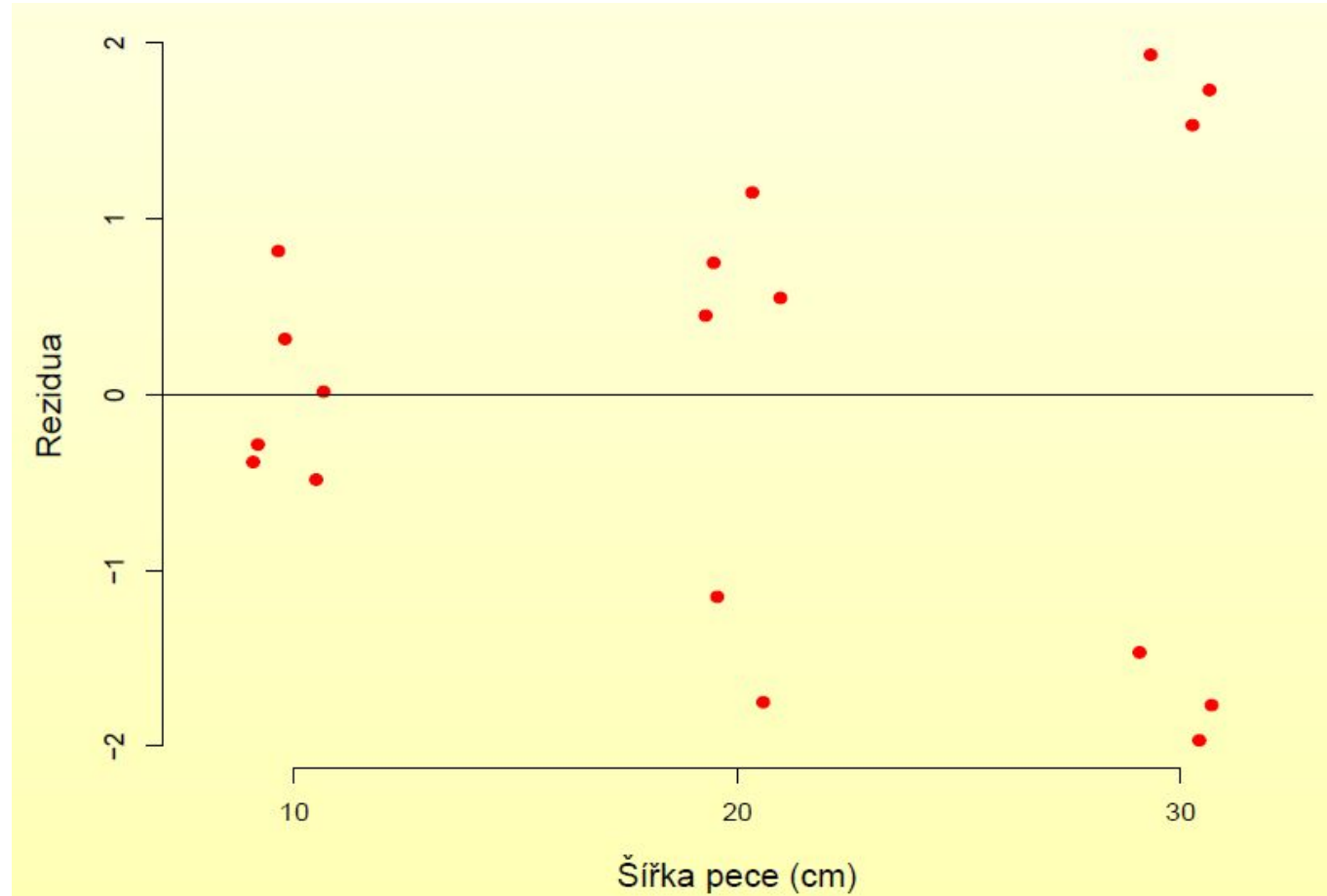
OVĚŘOVÁNÍ NORMALITY

QQ-GRAF



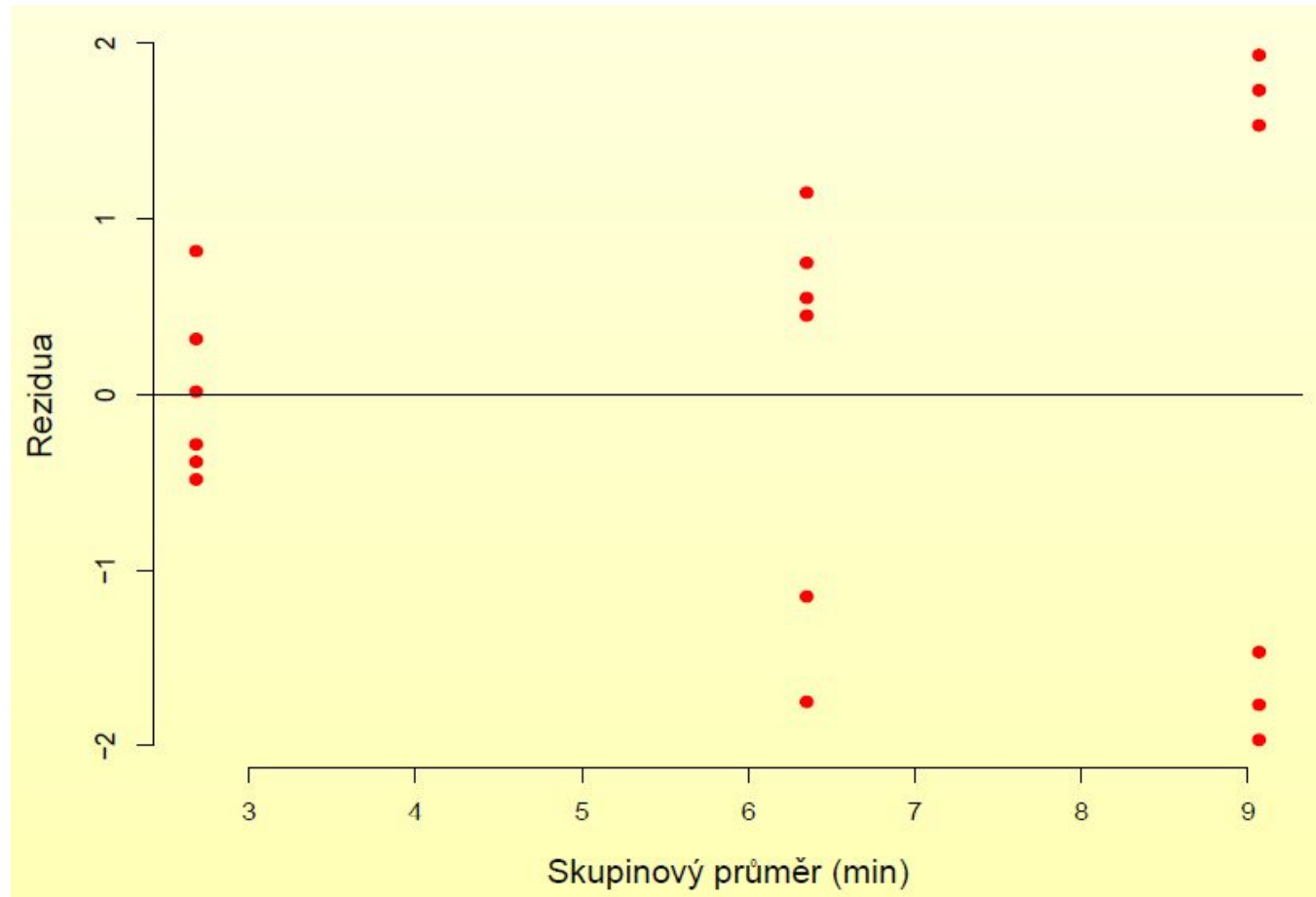
OVĚŘOVÁNÍ HOMOSKEDASTICITY

REZIDUA VS. FAKTOR



OVĚŘOVÁNÍ HOMOSKEDASTICITY

REZIDUA VS. SKUPINOVÉ PRŮMĚRY



OVĚŘOVÁNÍ HOMOSKEDASTICITY

LEVENĚŮV TEST (INFORMATIVNĚ)

- $H_0: \sigma_1 = \sigma_2 = \sigma_3$
 - Homoskedasticita
 - Směrodatná odchylka koksovacích časů je pro všechny tři šířky pece stejná.
- $H_1: \sigma_1 \neq \sigma_2$ nebo $\sigma_1 \neq \sigma_3$ nebo $\sigma_2 \neq \sigma_3$
 - Heteroskedasticita
 - Směrodatná odchylka koksovacího času se pro alespoň jednu šířku pece liší.
- Testová statistika $F = 22,37$
- Dosažená hladina testu $p < 0,001$

PORUŠENÍ PŘEDPOKLADŮ

Porušení normality

- drobné odchylky od normality nevadí
- nevadí v případě velkého počtu dat ve skupinách
- **řešení:**
 - transformace dat
 - Kruskalův-Wallisův test

Porušení homoskedasticity

- drobné porušení homoskedasticity nevadí
- závažný je rozptyl rostoucí s velikostí skupinového průměru
- **řešení:**
 - transformace dat
 - Welchova ANOVA

VÝBĚR ANOVA TESTU

KOKSOVACÍ ČAS VS. ŠÍŘKA PECE

- Neporušena normalita
- Porušena homoskedasticita



- logaritmická transformace nebo
- Welchova modifikace ANOVA testu:
 - Testová statistika $F = 47,602$
 - Dosažená hladina testu $p < 0,001$
- Na 5% hladině významnosti jsme prokázali, že čas nutný ke koksování **závisí** na šířce pece.

KRUSKALŮV-WALLISŮV TEST (INFORMATIVNĚ)

- porovnání populačních **mediánů** u více než dvou nezávislých výběrů
- **neparametrická verze ANOVA** jednoduchého třídění
 - místo pozorování se používá jeho pořadí
- zobecněná verze dvouvýběrového **Wilcoxonova** testu

- $H_0: med_1 = med_2 = med_3 = \dots$
- $H_1: med_1 \neq med_2$ nebo $med_1 \neq med_3$ nebo $med_2 \neq med_3 \dots$

NEZÁVISLÉ VS. ZÁVISLÉ VÝBĚRY

Zobecnění srovnávání dvou výběrů

Normální výběry	Počet výběrů	
	právě dva	alespoň dva
nezávislé	Dvouvýběrový t-test	ANOVA jednoduchého třídění
závislé	Párový t-test	ANOVA s bloky

- porovnání populačních charakteristik u více než dvou závislých výběrů
- zobecnění párových testů (studie typu před-během-po)
 - místo párování máme tzv. **blokový přístup**

NEBLOKOVÝ PŘÍSTUP

- Problém: Nadnárodní firma, která má dvacet poboček po celém světě, potřebuje rozhodnout, který ze tří výrobních postupů je optimální pro danou výrobu
- Neblokový přístup:
 - 20 poboček je rozděleno náhodně do 3 skupin
 - Pobočky v rámci jedné skupiny vyzkouší všechny jeden výrobní postup, např:
 - 7 poboček testuje postup I
 - 7 poboček testuje postup II
 - 6 poboček testuje postup III
- Analýza: ANOVA jednoduchého třídění
 - $(n = 20; n_1 = 7, n_2 = 7, n_3 = 6)$

BLOKOVÝ PŘÍSTUP

- Problém: Nadnárodní firma, která má dvacet poboček po celém světě, potřebuje rozhodnout, který ze tří výrobních postupů je optimální pro danou výrobu
- Bloková studie (paralelní přístup):
 - Pracovníci jedné pobočky jsou rozděleni do tří skupin
 - Každá skupina testuje jeden výrobní postup
 - V rámci experimentu jsou v každé pobočce testovány všechny tři výrobní postupy najednou (paralelně)
 - Bloková studie s $B = 20$ bloky (pobočkami) zkoumající vliv faktoru (výrobní postup) o $I = 3$ úrovních na odezvu (produktivita výroby apod.)

ZNÁHODNĚNÁ BLOKOVÁ STUDIE

- Blokovou studii lze vylepšit znáhodněním
- Příklad: sekvenční přístup a znáhodnění
 - Pořadí, v jakém jsou jednotlivé výrobní postupy v rámci jedné pobočky testovány, je zvoleno náhodně
 - i. Pobočka A: I → III → II
 - ii. Pobočka B: II → I → III
 - iii. atd.
 - Eliminuje, kromě jiného, vliv případného „učícího“ efektu na závěry statistické analýzy
- Příklad: paralelní přístup a znáhodnění
 - Přiřazení výrobních postupů jednotlivým pracovním skupinám se provede *náhodně* a ne na základě rozhodnutí experimentátora

VÝZNAM BLOKOVÁNÍ A ZNÁHODNĚNÍ

- **Uspořádání do bloků**
 - Zlepšuje statistické úsudky eliminací **známých zdrojů variability**
- **Znáhodnění**
 - Zlepšuje statistické úsudky eliminací **neznámých zdrojů variability**

TESTOVÁ OTÁZKA 1

Nulová hypotéza testu *ANOVA jednoduchého třídění* zní:

Vyberte libovolný počet možných odpovědí.

(Správná nemusí být žádná, ale také mohou být správné všechny.)

- a)** numerická proměnná nezávisí na kategoriální proměnné;
- b)** numerická proměnná závisí na kategoriální proměnné;
- c)** ošetření (faktor) nemá vliv na kvantitativní odezvu;
- d)** skupiny se neliší ve svých populačních průměrech.

TESTOVÁ OTÁZKA 2

K zamítnutí nulové hypotézy u ANOVA testu vede relativně velký podíl meziskupinové variability vůči vnitroskupinové variabilitě. Je to pravda?

Vyberte **ano/ne** na zadanou otázku.

KLÍČOVÁ SLOVA

Analýza rozptylu

ANOVA

jednoduchého třídění

Vícevýběrový test

Meziskupinová
variabilita

Vnitroskupinová
variabilita

Reziduální graf

Mnohonásobné
porovnávání

Bonferroniho
princip

Welchova ANOVA

Kruskalův-Wallisův
test

ANOVA s bloky

LITERATURA

- MAREŠ, Petr, RABUŠIC, Ladislav a SOUKUP, Petr. Analýza sociálněvědních dat (nejen) v SPSS. 1. vydání. Brno: Masarykova univerzita, 2015.
Kapitola 7.4.
- HENDL, Jan. Přehled statistických metod: analýza a metaanalýza dat. Páté, rozšířené vydání. Praha: Portál, 2015.
Kapitola 9.1 – 9.2.
- MAREK, Luboš a kol. Statistika v příkladech. Druhé vydání. Praha: Professional Publishing, 2015.
Kapitola 5.