ANOVA

Otázky a motivace

- **Situace: měříme hodnoty číselné proměnné u jednotek, které jsou klasifikovány do *K* nepřekrývajících se skupin.**

- **Liší se od sebe populační průměry (střední hodnoty) ve skupinách nebo jsou zjištěné rozdíly výběrových průměrů způsobené pouze náhodou?**
- **Jsou hodnoty analyzované proměnné ovlivněné sledovaným faktorem (skupinami)?**

- ***Úloha porovnání průměrů v několika skupinách je rozšířením situace dvouvýběrového t-testu pro více skupin***

Variabilita

Rozptyl proměnné *X*

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

- **Proč není proměnná *X* konstantní? Co způsobuje její variabilitu?**
- **Je možné vysvětlit tuto variabilitu vlivem zkoumaných skupin?**

Nulová a alternativní hypotéza

H_0: průměry (střední hodnoty) ve skupinách se rovnají, tj.

$$\mu_1 = \mu_2 = \ldots = \mu_K$$

H_A: průměry (střední hodnoty) nějakých dvou skupin se nerovnají, tj.

$$\exists\ i,j: \mu_i \neq \mu_j$$

Alternativní možnost formulace:

H_0: $\mu_i = \mu$ pro všechna *i* a nějakou konstantu μ (tj. existuje společná střední hodnota μ)

H_A: $\mu_i \neq \mu$ alespoň pro jednu skupinu *i* (společná hodnota μ neexistuje)

Rozklad variability

Celkový součet čtverců (Total Sum of Squares)

$$TSS = \sum(X_{ik} - \bar{X})^2$$

- součet čtverců rozdílů měření od společného průměru

Součet čtverců uvnitř skupin (Within Groups Sum of Squares)

$$WSS = \sum(X_{ik} - \bar{X}_k)^2$$

- součet čtverců rozdílů měření od průměrů svých skupin

Součet čtverců mezi skupinami (Between Groups Sum of Squares)

$$BSS = \sum n_k(\bar{X}_k - \bar{X})^2$$

- součet čtverců rozdílů všech měření zaměřených skupinovými průměry od společného průměru

TSS = WSS + BSS

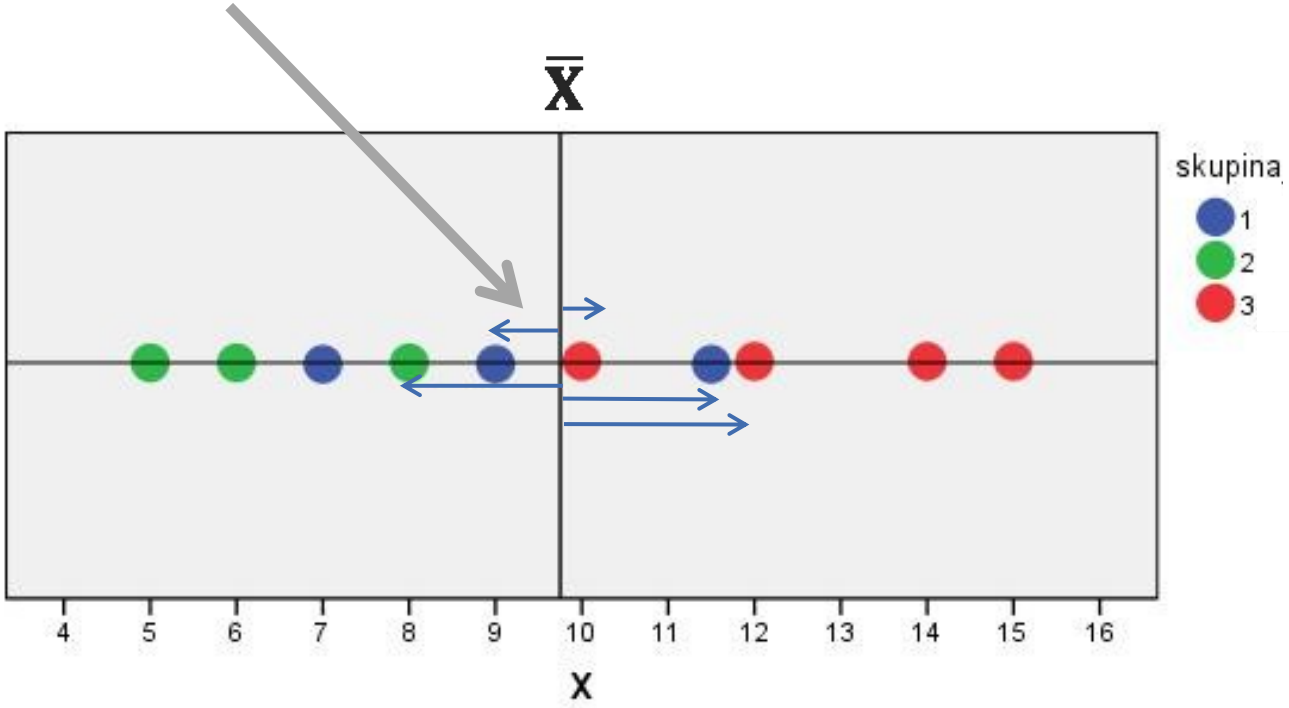
celková variabilita = variabilita uvnitř skupin + variabilita mezi skupinami

Grafické znázornění rozkladu (1)

Celkový součet čtverců (Total Sum of Squares)

$$TSS = \sum(X_{ik} - \bar{X})^2$$

součet čtverců rozdílů měření od společného průměru



Grafické znázornění rozkladu (2)

Součet čtverců mezi skupinami

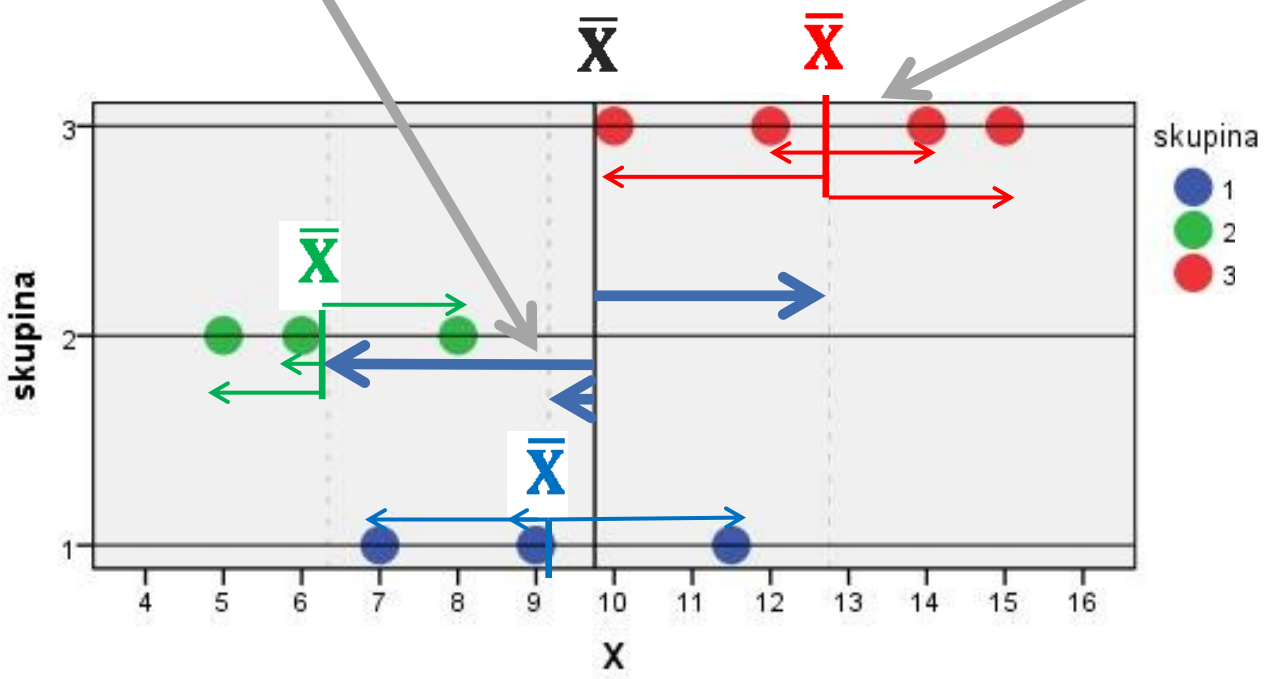
$$BSS = \sum n_k(\bar{X}_k - \bar{X})^2$$

součet čtverců rozdílů všech měření zaměřených skupinovými průměry od společného průměru

Součet čtverců uvnitř skupin

$$WSS = \sum(X_{ik} - \bar{X}_k)^2$$

součet čtverců rozdílů měření od průměrů svých skupin



Korelační poměr - motivace

- **Jak silná je vazba mezi číselnou proměnnou a nezávislou nominální proměnnou?**
- **Jak dobře vysvětluje rozdělení souboru do zkoumaných skupin variabilitu číselné proměnné?**
- **Který ze zkoumaných faktorů nejsilněji ovlivňuje analyzovanou proměnnou?**

Korelační poměr

$$\eta^2 = \frac{\textit{variabilita mezi skupinami}}{\textit{celková variabilita}}$$

$$\eta^2 = \frac{\sum n_k(\bar{X}_k - \bar{X})^2}{\sum(X_{ik} - \bar{X})^2} = \frac{BSS}{TSS} = 1 - \frac{WSS}{TSS} \quad \in \langle 0, 1 \rangle$$

- roven nule v případě, že jsou všechny průměry ve skupinách stejné (BSS=0)
- roven jedné v případě, že jsou všechna data uvnitř každé skupiny stejná (WSS = 0)
- stonásobek korelačního poměru se často vyjadřuje v procentech, charakterizuje **procento variability proměnné *X* vysvětlené faktorem**
- patří do skupiny měr, které mají obecnou vlastnost poměru vysvětlené variance zvoleným modelem, tzv. **koeficienty determinace**

Fisherův test

Fisherův F-test

- **testová statistika:**

$$F = \frac{\sum n_k(\bar{X}_k - \bar{X})^2/(k-1)}{\sum(X_{ik} - \bar{X}_k)^2/(n-k)}, \quad df = (k-1, n-k)$$

= průměrná čtvercová odchylka mezi skupinami / průměrná čtvercová odchylka uvnitř skupin

Pozn.: Fisherův F-test je zobecněním dvouvýběrového T-testu. V případě dvou skupin je testová statistika F druhou mocninou testové statistiky T.

Tabulka ANOVA

součty čtverců

stupně volnosti

průměrné čtvercové odchylky

testová statistika *F*

dosažená hladina významnosti

ANOVA

Průměr výdaje

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,102	3	,034	3,434	,022
Within Groups	,671	68	,010		
Total	,773	71			

celkem

uvnitř skupin

mezi skupinami

Předpoklady

- **pozorování jsou mezi sebou navzájem nezávislá**
- **skutečné hodnoty a chyby jsou navzájem nezávislé**
- **výběry pocházejí z normálního rozdělení**
- **výběry jsou navzájem nezávislé (mají prázdný průnik)**
- **ve skupinách jsou stejné rozptyly**

Pozn.: Simulační studie a zkušenosti z aplikací ukazují, že první předpoklad je kritický a jeho nedodržení silně ovlivňuje aplikabilitu. Předpoklady normálního rozdělení a shody rozptylů nemají na výsledky rozhodující vliv. Metoda je proti jejich použití značně robustní.

Ověřování předpokladů (1)

Předpoklad shody rozptylů ve skupinách

Leveneho test

H_0: rozptyly ve skupinách se rovnají, tj.

$$\sigma_1 = \sigma_2 = \ldots = \sigma_K$$

H_A: rozptyly některých dvou skupin se nerovnají, tj.

$$\exists\, i,j: \sigma_i \neq \sigma_j$$

Test of Homogeneity of Variances

Průměr výdaje

Levene Statistic	df1	df2	Sig.
1,249	3	68	,299

Pozn.: Statistika F je poměrně robustní vzhledem k odchylkám od tohoto předpokladu v případě, že velikost skupin je stejná nebo téměř stejná. Pokud se však velikosti skupin i rozptyly liší, je vhodnější užít robustnější testy.

Welschův a Brown-Forsythův test

Robustnější testy vzhledem k odchylkám od shody rozptylů

- Welschův test
- Brown-Forsythův test

Robust Tests of Equality of Means

Průměr vývodu

	Statistic[a]	df1	df2	Sig.
Welch	3,661	3	37,405	,021
Brown-Forsythe	3,434	3	64,093	,022

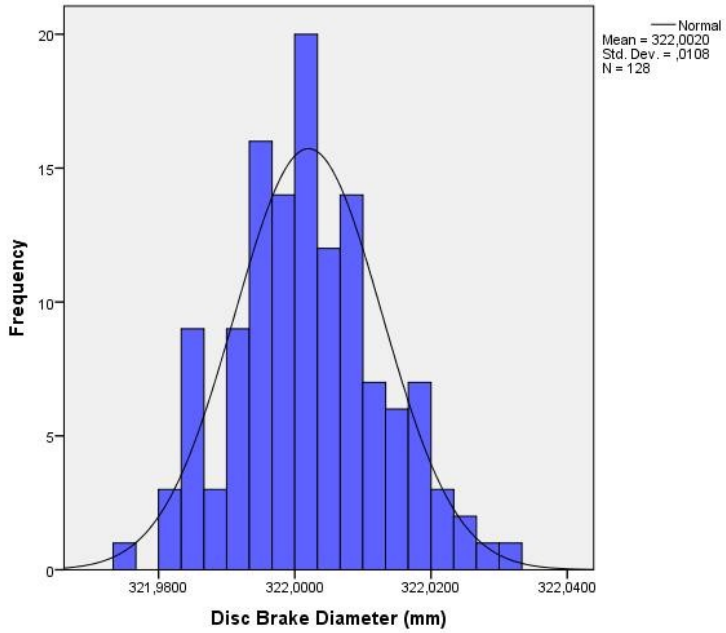
a. Asymptotically F distributed.

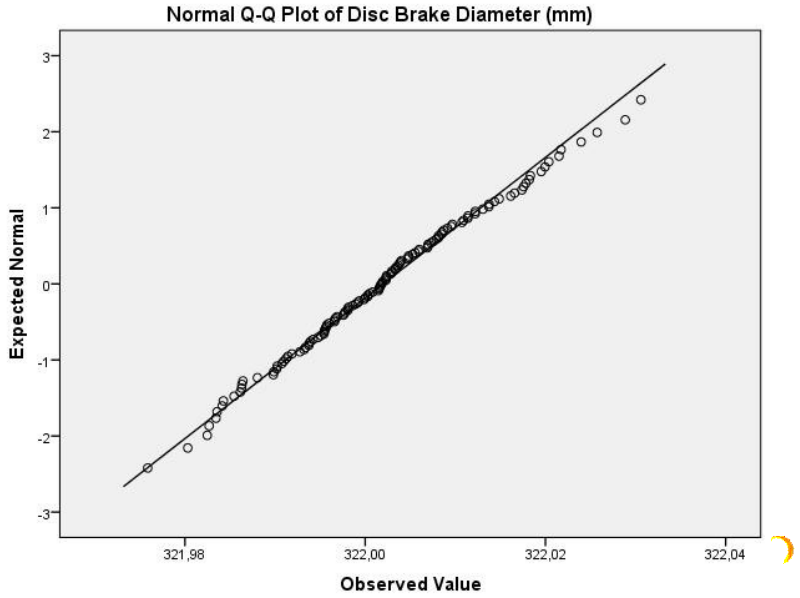
Ověřování předpokladů (2)

Předpoklad normálního rozložení

Ověřování:

a) grafické metody (histogram, boxplot, Q-Q Plot …)





Ověřování předpokladů (3)

b) testy:

- **Kolmogorov-Smirnov** – **test založený na porovnání distribučních funkcí: teoretické pro normální rozložení a kumulativní empirické distribuční funkce**
- **Shapiro-Wilk** – **test založený na porovnání kvantilových hodnot (pořádkových statistik) teoretické a uspořádané statistické řady**

H_0: proměnná má normální rozložení

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Obsluha	Statistic	df	Sig.	Statistic	df	Sig.
Průměr vývodu	AA	,202	18	,051	,935	18	,235
	BB	,183	18	,115	,906	18	,072
	CC	,190	18	,086	,929	18	,190
	DD	,129	18	,200*	,969	18	,780

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Kruskal-Wallisův test

Analyze – Nonparametric Tests – Legacy dialogs – N Independent Samples

- Dva a více nezávislých výběrů
- H_0: Všechny pozorované skupiny pochází ze stejného rozdělení
- H_A: Alespoň jedna z pozorovaných skupin pochází z jiného rozdělení
- Liší-li se rozdělení jen svými parametry považujeme je také za rozdílná, tj. platí H_A.
- testovaná rozdělení není třeba specifikovat
- neparametrická alternativa ANOVA