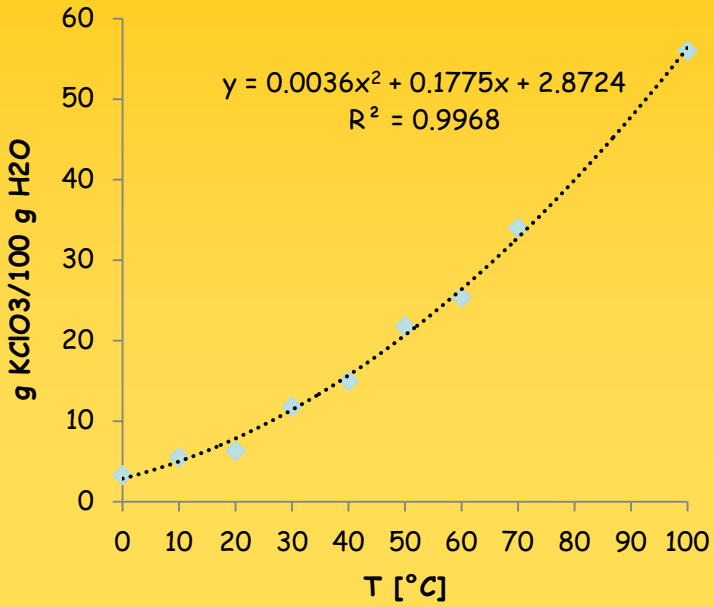
Základy zpracování geologických dat

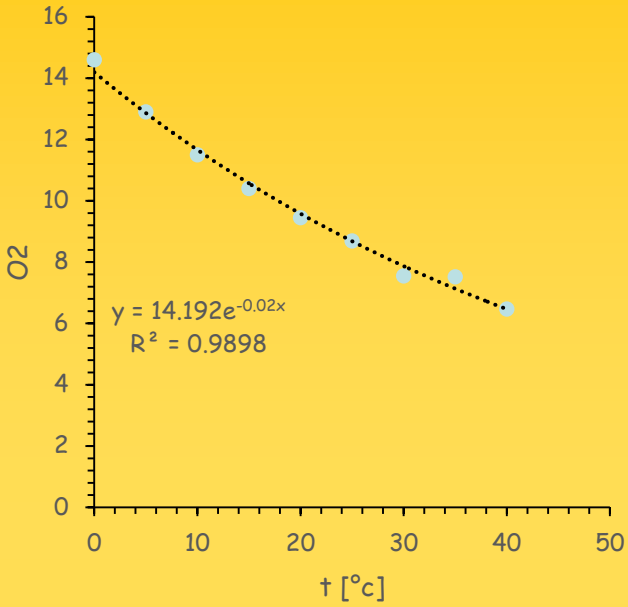
korelační analýza – nelineární závislost

testování statistických hypotéz

R. Čopjaková

Nelineární závislost





Korelační koeficient

Nepočítat Pearsonův korelační koeficient

Pro stanovení síly závislosti lze využít koeficient determinace v Excelu nebo spočítat Spearmanův koeficient pořadové korelace

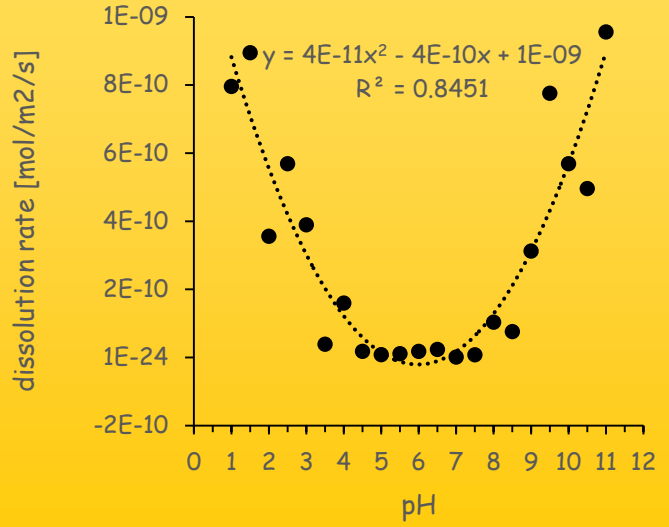
Spearmanův koeficient pořadové korelace

- Univerzální – nejen pro lineární závislost
- Chci-li spočítat hodnotu Spearmanova koeficintu, převedu naměřená data pro soubor X_i a Y_i na pořadové hodnoty X_{ip} a Y_{ip}.
- Spočtu rozdíly v pořadí jednotlivých párů $d_i = X_{ip} - Y_{ip}$, které použiji při výpočtu tohoto koeficientu

- Lze využít jen pro funkce monotónní (tedy fci rostoucí nebo klesající, nerostoucí nebo neklesají; nelze tedy použít např. pro kvadratickou fci)

  - Např. závislost rozpustnosti jílových minerálů na pH – není fce monotónní

  - r= 0,92 (stanovený z koeficientu determinace)
  - SR ~ 0 – SR nelze použít



Spearmanův koeficient pořadové korelace

Reálná naměřená data (soubor X a Y) s nelineární závislostí převedu na pořadové hodnoty a spočtu Spearmanův koeficient pořadové korelace

RANK
RANK.EQ RANK.AVG

Stanovení pořadí v Excelu
e RANK - starší verze MS Office
(vyžaduje úpravu stejných pořadí)
ce RANK.AVG – nové MS Office
(nevyžaduje žádnou úpravu)
ce RANK.EQ – nové MS Office; = RANK
(vyžaduje úpravu)

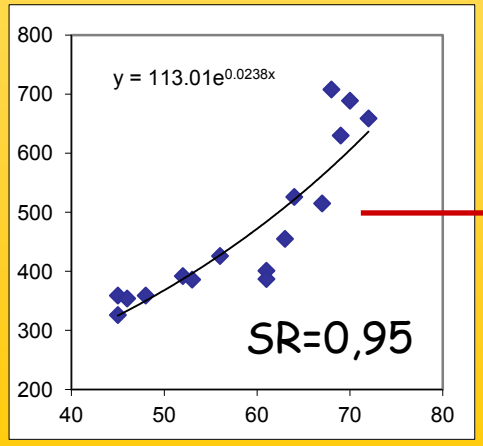
n	X	Y	rank X	rank Y	úprava X	úprava Y	di	di^2
1	45	350	1	3	1,5	3,5	-2	4
2	45	326	1	1	1,5	1	0,5	0,25
3	46	354	3	2	3	2	1	1
4	48	350	4	3	4	3,5	0,5	0,25
5	52	392	5	7	5	7	-2	4
6	53	386	6	5	6	5	1	1
7	56	426	7	9	7	9	-2	4
8	61	401	8	8	8,5	8	0,5	0,25
9	61	387	9	6	8,5	6	2,5	6,25
10	63	455	10	10	10	10	0	0
11	64	526	11	12	11	12	-1	1
12	67	515	12	11	12	11	1	1
13	69	630	14	13	14	13	1	1
14	72	659	16	14	16	14	2	4
15	70	689	15	15	15	15	0	0
16	68	703	13	16	13	16	-3	9
						suma		37

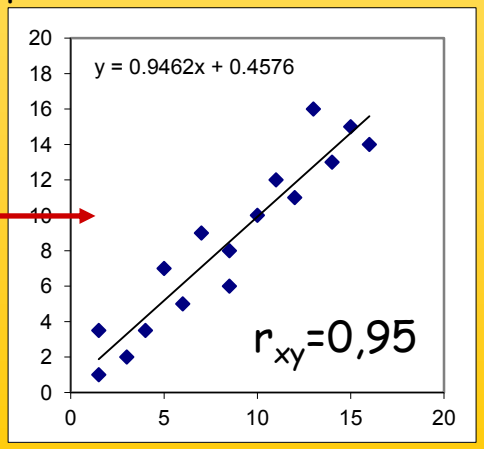
$$= 1 - 6*37/16(16^2-1) = 0,95$$

$$\frac{cov_{xy}}{s_x s_y} = \frac{20,05}{4,60*4,61} = 0,95$$

reálná data

pořadová data





Spočtu-li pearsonův koeficient korelace pro pořadové hodnoty (lineární závislost), bude velice blízký hodnotě Spearmanova koeficientu pořadové korelace pro naměřené hodnoty proměnné X a Y

Testování statistických hypotéz

Testování statistických hypotéz

- Existuje závislost mezi soubory dat? (např. vyšetřování substitucí v minerálech)

$r_{xy} = -0,6$

- Je některá hodnota souboru odlehlá? (Mám ji ze souboru vyřadit a nepracovat s ní při výpočtu dalších parametrů – střední hodnoty, Sx…?)

  4,0; 4,2; 4,4; 4,5; 4,5; 4,6; 4,7; 4,9; 5,1; 5,8 ?

- Chovají se naměřená data podle normálního rozdělení?

Testování statistických hypotéz

Při zpracování dat jsou časté úvahy typu:

- Liší se hodnoty naměřené na stejných přístrojích v různých laboratořích? (např. data z EMP v Brně a Barrandově)
- Liší se výsledky získané různými analytickými metodami (např. hodnoty naměřené přenosným terénním gama-spektrometrem a laboratorním gama-spektrometrem)
- Liší se hodnoty naměřené v různých časových intervalech (sezónní vlivy v hydrogeologii)
- Liší se hodnoty naměřené v různých místech (např. srovnání chemického složení – protolitu- ortorul sněžnických a gieraltovských orlicko-kladského krystalinika)
- Liší se hodnoty naměřené látky od deklarované hodnoty (např. prověřování standardů, či kontrola kvality analýz)

K řešení těchto problémů lze ve statistice využít metody testování statistických hypotéz, s jejichž pomocí lze hledat odpovědi na tyto otázky a činit závěry.

Testování statistických hypotéz

Základní pojmy

- **hypotéza H_0** – nulová (testovaná) hypotéza, kterou testujeme
- **hypotéza H_A** – alternativní hypotéza, kterou přijmeme, zamítneme-li hypotézu Ho
- α – **hladina významnosti** – volí se malá do 0,05; nejčastěji 0,05 - tedy 5-ti% (nebo 0,01 tedy 1%) pravděpodobnost chyby 1. druhu; vysoce významné výsledky testování pro $\alpha = 0{,}005$ a méně
- **kritická hodnota** pro test nulové hypotézy = **hodnota kvantilu hraniční pro oblast zamítání H_0 na zvolené hladině významnosti** α (kde α vyjadřuje pravděpodobnost, že náhodná veličina překročí tuto hodnotu).

Chyby při testování

Skutečnost	**Rozhodnutí statistického testu**	
	Zamítneme H_0	Nezamítneme H_0
H_0 je správná	**Chyba I. druhu**	Správné rozhodnutí
H_0 neplatí	Správné rozhodnutí	**Chyba II. druhu**

Chyba 1. druhu - α

- zamítneme-li platící hypotézu H_0, dopustíme se chyby I. druhu
- je spojena se zamítnutím nulové hypotézy, která ve skutečnosti platí; její pravděpodobnost se nazývá hladina významnosti α
- platí-li hypotéza alternativní H_A a testovanou hypotézu H_0 nezamítáme, dopouštíme se chyby II. Druhu

Chyba 2. druhu

- Značí se β
- je pravděpodobnost nesprávného přijetí nulové hypotézy
- 1- β se nazývá síla testu
- závisí na velikosti výběru (s větším souborem klesá)

Testování statistických hypotéz

Obecný postup testování

- zvolíme hladinu významnosti α
- formulujeme nulovou hypotézu H_0 a alternativní hypotézu H_A
- zvolíme vhodné testovací kritérium (test)
- vypočteme velikost test. kritéria T
- stanovíme kritickou hodnotu (hodnotu kvantilu hraniční pro oblast zamítání H_0) pro zvolenou hladinu významnosti - $k\alpha$
- porovnáme velikost testovacího kritéria s kritickou hodnotou

  obvykle:

  jestliže $T \leq k\alpha$, akceptujeme nulovou hypotézu H_0 na námi zvolené hladině významnosti

  jestliže $T > k\alpha$, zamítneme nulovou hypotézu a říkáme, že platí H_A

Oboustranný, jednostranný test

- oboustranná hypotéza (oboustranný test)

  $H_0: X_1 = X_0$

  $H_A: X_1 \neq X_0$

- jednostranná hypotéza (jednostranný test)

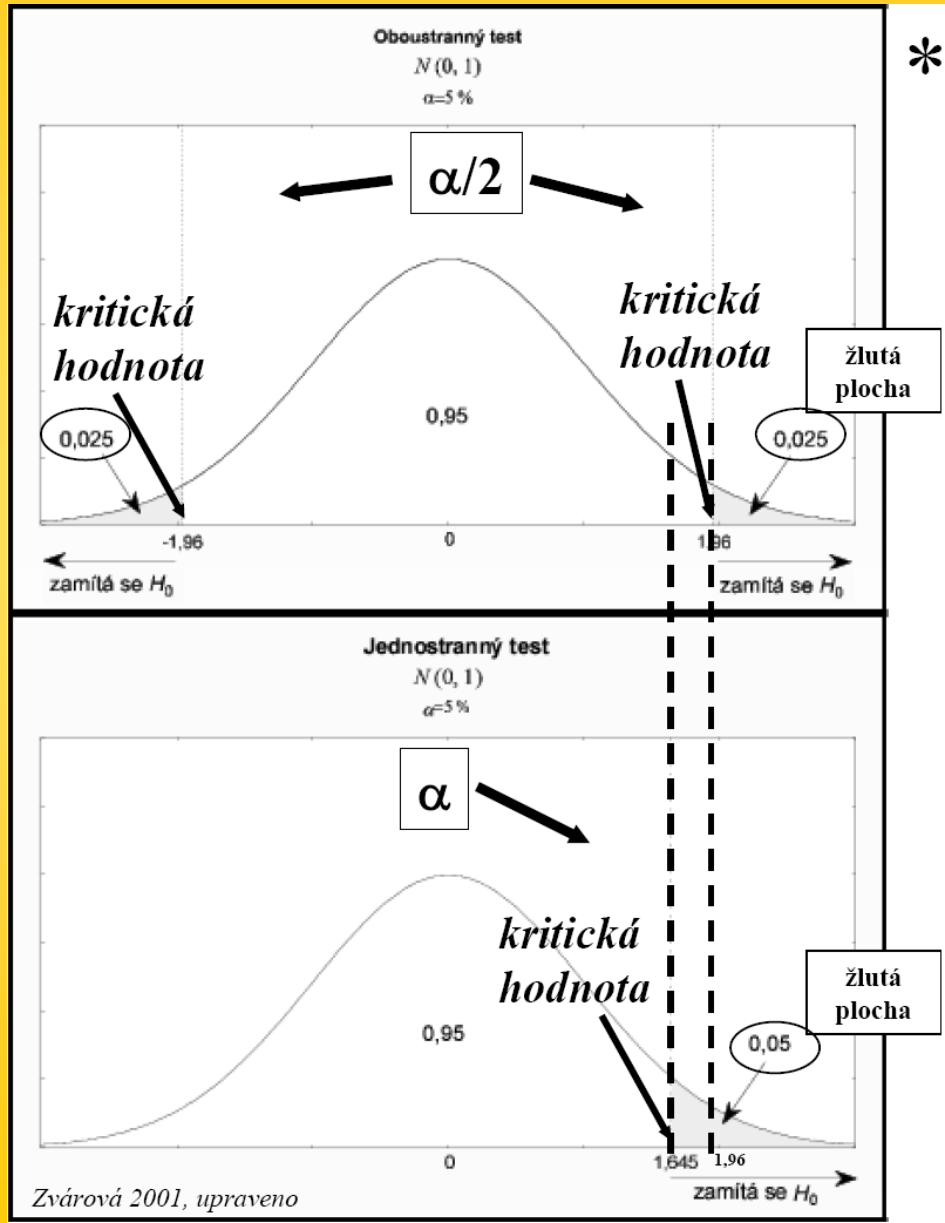
  $H_0: X_1 = X_0$

  $H_A: X_1 < X_0$, případně $X_1 > X_0$

Oboustranný, jednostranný test

V případě *oboustranného testu*:
musíme rozdělit danou hladinu významnosti α na dvě části reprezentující dva možné konce distribuce.
Značíme kα(2), např. t0,05(2)
Stanovíme tedy hodnotu kvantilu 0,975

V případě *jednostranného testu* *(pravostranný – Ha:$X_1 > X_0$)*
uvažujeme pouze jeden konec distribuce a danou hladinu významnosti proto nedělíme.
Značíme kα(1), např. t0,05(1)
Stanovíme tedy hodnotu kvantilu 0,95



Testování statistických hypotéz

Testy: **parametrické**
neparametrické

- **parametrický** test – pro soubory s normálním rozdělením nebo téměř normálním rozdělením pravděpodobností

  Známe-li rozdělení pravděpodobností základního souboru

- **neparametrický** test – i pro soubory a jiným než normálním rozložením pravděpodobností

  Neznáme-li rozdělení pravděpodobností základního souboru
  - širší použití než parametrické
  - řešení nezávisí na typu rozdělení základního souboru
  - lze použít i pro silně nenormální rozdělení, kdy parametrické testy předpokládající normální rozdělení selhávají

Test nezávislosti dat ~ síly korelačního koeficientu

- **Otázka – Existuje závislost mezi dvěma soubory data? Je spočtená hodnota korelačního koeficientu statisticky významná?**
  Když r_{xy} se blíží 1 či -1 pak jistě ano
  Ale co když r_{xy} je např. 0,5? – závislé na počtu měření

- ověření předpokladu o nulové hodnotě korelačního koeficientu (ověření nezávislosti dat)
  Ho: $r_{xy} = 0$

- Spočtení **testovacího kritéria**

- Stanovení kritické hodnoty pro zvolenou hladinu významnosti α a počet stupňů volnosti n-2; Tk(1-α/2; n-2) (**oboustranná varianta testu**)
  V excelu např. pro α =5% stanovím pomocí funkce
  T.INV (pro daný kvantil a hladinu významnosti; 1-α/2 = 0,975)
  T.INV.2T (pro danou hladinu významnosti a stupně volnosti; α = 0,05)
  TINV (starší verze MS Office; totéž jako T.INV.2T)

- Pokud t ≤Tk pak přijmeme Ho a tedy existenci závislosti mezi veličinami v souboru považujeme za neprokázanou.

Příklad

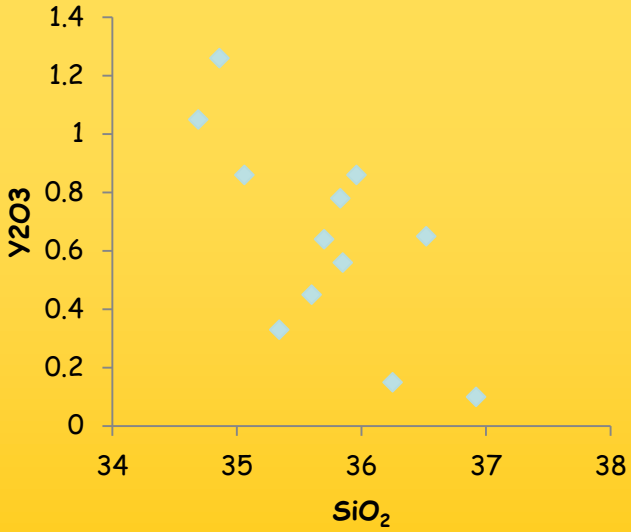
Otestujte, zda existuje statisticky významná závislost mezi obsahem Y_2O_3 a SiO_2 v granátu; r_{xy}=-0,70757
Pracujte při hladině významnosti 0,05; počet analýz je 12

SiO_2	Y_2O_3
36.52	0.65
35.96	0.86
35.6	0.45
35.83	0.78
36.25	0.15
36.92	0.1
35.85	0.56
35.7	0.64
34.69	1.05
35.06	0.86
35.34	0.33
34.86	1.26

- Nulová hypotéza Ho: $r_{xy} = 0$
- Spočtení testovacího kritéria

= -3,166

- Stanovení kritické hodnoty (z pravého konce distribuční fce)
  studentova rozděleníTk(1-$\alpha/2$; n-2)
  T.INV(0,975;10) = 2,228
  nebo T.INV.2T a TINV (0,05;10) = 2,228



- Velikost testovacího kritéria (beru jeho absolutní hodnotu) je větší než kritická hodnota
  3,166 > 2,228
- Ho zamítám; přijímám H_A – mezi soubory je statisticky významná závislost