Základy zpracování geologických dat

testování statistických hypotéz

neparametrické testy

R. Čopjaková

Neparametrické testy

- Používají se, i pokud nejsou předpoklady normality dat splněné
- Tedy i pro jiné typy rozdělení pravděpodobností než normální
- V případě souboru dat s normálním rozdělením jsou vhodnější, když v souboru jsou některé hodnoty výrazně odlehlé

Wilcoxonův test pro párované hodnoty

- Jedná se o neparametrickou obdobu t-testu na párové hodnoty
- Při ověřování, zda dva výběrové soubory lze považovat za výběry z jednoho základního souboru
- Pro dva závislé soubory dat, kde $n_1 = n_2$
- Jedná se o test pořadový - srovnává soubory dat pomocí seřazení souborů dat podle pořadí, nikoli přes průměry
- Lze využít i pro soubory dat s jiným než normálním rozdělením pravděpodobností, zatímco t-test na párové hodnoty lze využít pouze pro soubory dat s normálním rozdělením

Nulová hypotéza versus alternativní hypotéza

- Ho: Není systematická diference uvnitř párů
- (medián rozdílů M je nulový).
- H_A: Je systematická diference uvnitř párů (medián rozdílů M je různý od nuly).

Wilcoxonův test pro párované hodnoty

Výpočet testovacího kritéria

Počítá s rozdíly mezi naměřenými hodnotami, které tvoří pár. Testovací kritérium se počítá **ze součtu rozdílů v pořadí párovaných hodnot**.

Absolutní hodnoty rozdílů mezi páry se seřadí podle velikosti, jednotlivým členům se **přiřadí pořadové číslo** (páry s rozdílem 0 vyloučím před přidělením pořadí). Přidělená pořadová čísla se rozdělí do dvou sloupců, do sloupce + (u rozdílu kladného) a do sloupce – (u rozdílu záporného). Zvlášť se dělá **součet kladných rozdílů v pořadí párovaných hodnot a** zvlášť **záporných**. **Výsledkem jsou testovací statistiky T+ a T-**

Jako **testovací kritérium** se uvažuje **hodnota menšího** z těchto dvou součtů.

Stanovení kritické hodnoty

Kritické hodnoty získám z tabulek pro příslušnou hladinu významnosti α a n (**počet párů s nenulovým rozdílem**) – $T_k(\alpha;n)$

Rozhodnutí

Je-li menší z obou T+ a T- menší nebo rovno T_k, tedy min (T+,T-) ≤ T_k, pak zamítáme Ho

Existuje i obdoba tohoto testu pro jeden výběrový soubor a jeho srovnání s deklarovanou hodnotou

Wilcoxonův test pro párované hodnoty - příklad

Porovnejme koncentrace Th (ppm) v horninách stanovené dvěma různými analytickými metodami pomocí Wilcoxonova testu pro párované hodnoty na 5% α.

Ho: není rozdíl mezi obsahy Th stanového s použitím dvou různých přístrojů
H_A: je rozdíl mezi obsahy Th stanového s použitím dvou různých přístrojů

					RANK.AVG		
Th1	Th2	di	IdiI	vyloučení nul	pořadí IdiI	minus	plus
9.3	11.3	-2	2	2	13	13	0
11	11	0	0			0	0
10	10.2	-0.2	0.2	0.2	2	2	0
10.1	9.6	0.5	0.5	0.5	4.5	0	4.5
9.6	10.1	-0.5	0.5	0.5	4.5	4.5	0
7.8	9.7	-1.9	1.9	1.9	12	12	0
9.5	9.6	-0.1	0.1	0.1	1	1	0
7.9	8.8	-0.9	0.9	0.9	9	9	0
9.2	11.3	-2.1	2.1	2.1	14	14	0
9.7	10.5	-0.8	0.8	0.8	8	8	0
9.6	10.6	-1	1	1	10	10	0
8.7	10.3	-1.6	1.6	1.6	11	11	0
9.4	9.7	-0.3	0.3	0.3	3	3	0
11.1	10.5	0.6	0.6	0.6	6	0	6
9.3	10	-0.7	0.7	0.7	7	7	0
			test krit (menší z hodnot)			94.5	**10.5**
			krit hodnota (n=počet nenulových párů)				21

Pro nové MS Office použiji funkci RANK.AVG

Pro starší verze MS Office použiji funkci RANK – nutná úprava pořadí pro hodnoty se stejným pořadím

Testovací kritérium T = 10,5
Kritická hodnota Tk = 21
T 10,5 < Tk 21 Ho zamítám

obsahy Th stanového s použitím dvou různých přístrojů, se statisticky významně liší

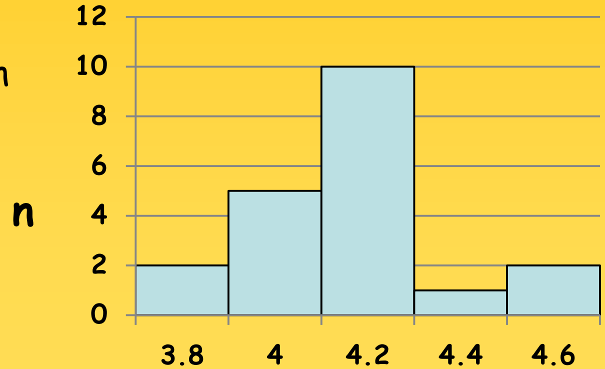
Kolmogorov-Smirnovův test shody - opakování

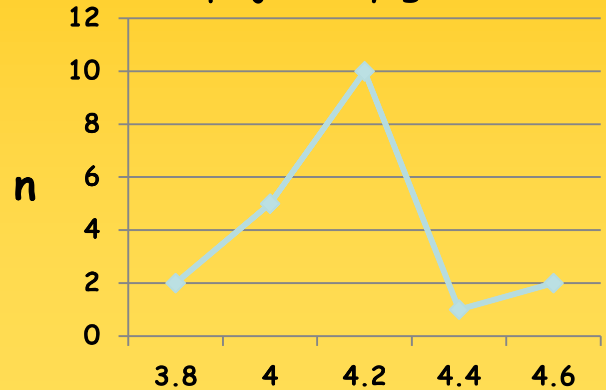
Vztah mezi histogramem a polygonem četností

sloupcový graf

spojnicový graf

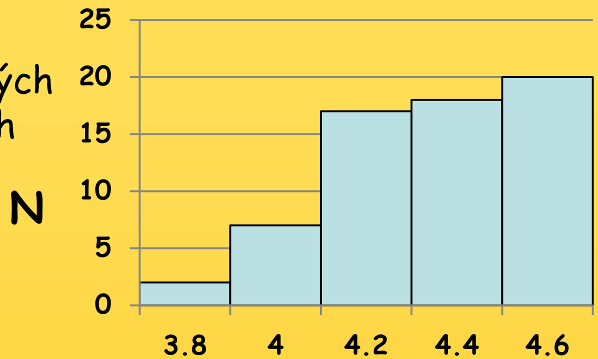
Histogram absolutních četností

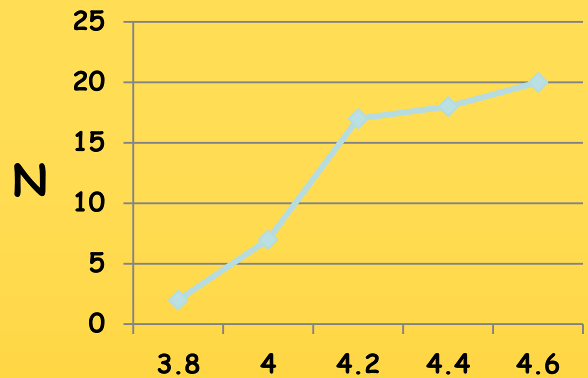




Polygon absolutních četností

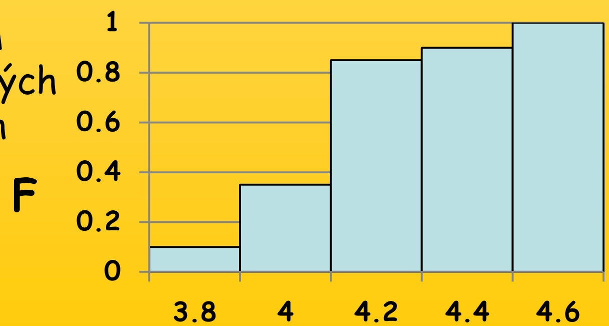
Histogram kumulovaných absolutních četností

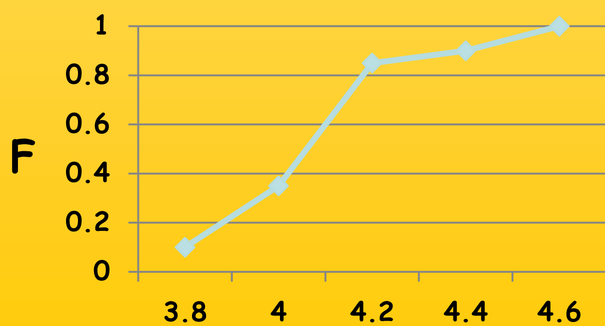




Polygon kumulovaných absolutních četností

Histogram kumulovaných relativních četností





Polygon kumulovaných relativních četností

Kolmogorov-Smirnovův test shody

nejprve varianta pro jeden výběr

Pro testování shody rozdělení pravděpodobnosti náhodného výběru s teoretickým, očekávaným rozdělením pravděpodobností.

Tedy ptáme-li se na otázku?

Má soubor dat normální rozdělení?
Má soubor dat logaritmicko-normální rozdělení?
Má soubor dat rovnoměrné rozdělení?

Obdobné využití jako chí-kvadrát test

Lze použít i v případech, kdy se nedoporučuje χ^2 test

Tedy i pokud více než 20% intervalových četností je menších než 5, či když přítomné intervaly s nulovou četností

Ho : pe1 = po1, . . . , pek = pok pro všechny intervaly
H_A : pej ≠ poj alespoň pro některý interval

Kolmogorov-Smirnovův test shody

nejprve varianta pro jeden výběr

Provedeme n nezávislých opakování pokusu. Výsledky rozdělíme do tříd a utvoříme kumulované četnosti nebo relativní kumulované četnosti pro experimentální data.

Namodelujeme kumulované četnosti nebo relativní kumulované četnosti pro teoretické rozdělení pravděpodobnosti, podle něhož očekáváme, že se naměřený soubor dat má chovat.

Najdeme maximální rozdíl mezi hodnotami kumulovaných četností $N_{1e}, \dots, N_{ke}$ (kde k představuje označení třídy) empirického souboru dat a očekávaného, teoretického rozdělení $N_{1o}, \dots, N_{ko}$. Nebo maximální rozdíl mezi hodnotami relativních kumulovaných četností $F_{1e}, \dots, F_{ke}$ empirického souboru dat a očekávaného, teoretického rozdělení $F_{1o}, \dots, F_{ko}$.

Tedy: $N_{1e} - N_{1o}$ nebo $F_{1e} - F_{1o}$

$N_{2e} - N_{2o}$ $F_{2e} - F_{2o}$

$\vdots$ $\vdots$

$N_{ke} - N_{ko}$ $F_{ke} - F_{ko}$

Testovací kritérium: $D_1 = \max|N_{ej} - N_{oj}|$, čili $D_1 = \max|F_{ej} - F_{oj}|$, kde

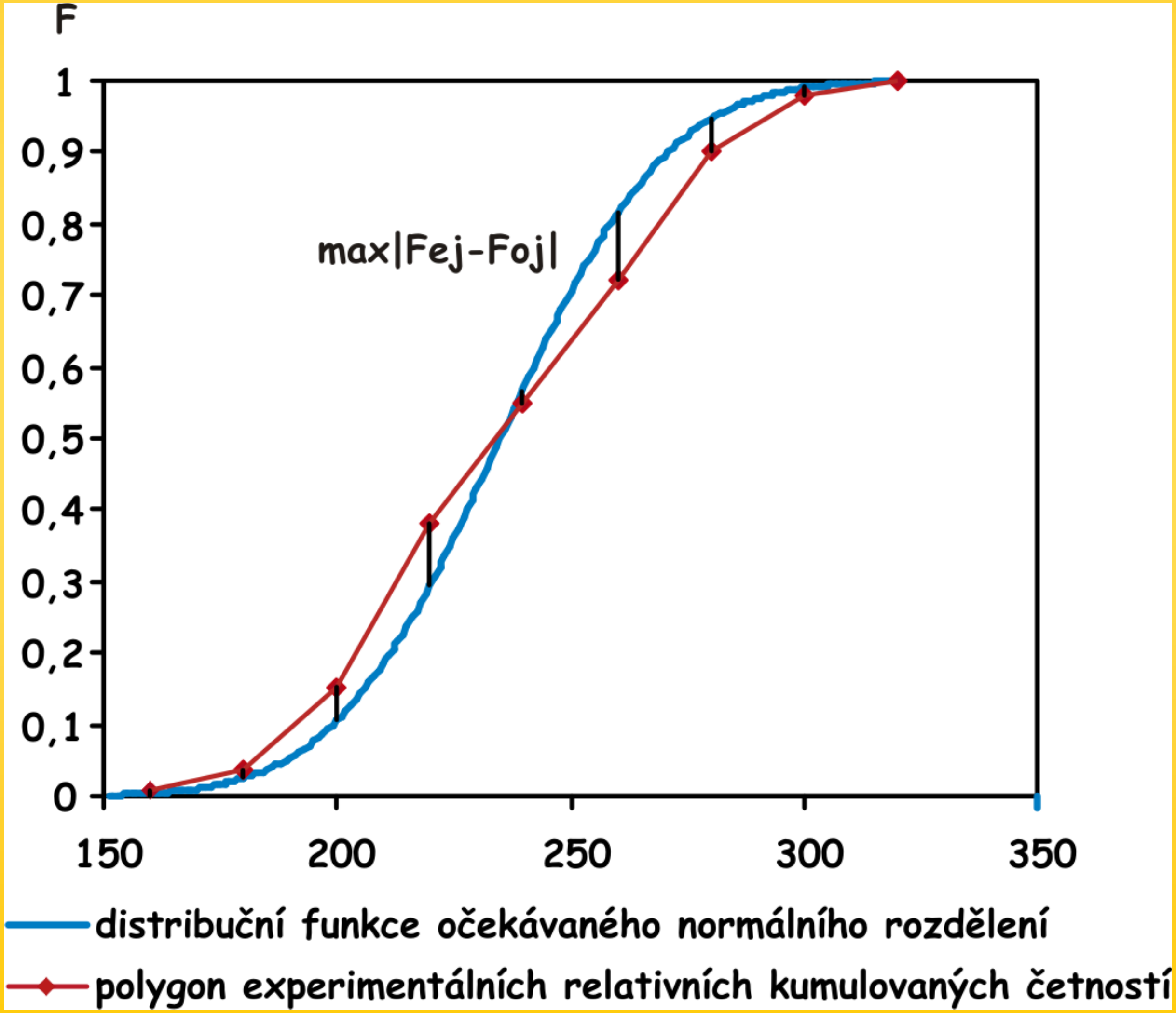
N_{ej} (F_{ej}) = experimentální kumulativní absolutní (relativní) četnost v j-tém řádku

N_{oj} (F_{oj}) = očekávaná kumulativní absolutní (relativní) četnost v j-tém řádku

Kolmogorov-Smirnovův test shody

nejprve varianta pro jeden výběr

grafické vyjádření stanovení testovacího kritéria



Kolmogorov-Smirnovův test shody

nejprve varianta pro jeden výběr

Stanovení kritické hodnoty D_k

pro n<=40 určíme ze statistických tabulek

pro n>40 dopočteme ze vztahu:

pro hladinu významnosti α 0,05 a 0,01

Pokud $D_1 \leq D_k$ pak přijmu Ho

Pokud $D_1 > D_k$ pak zamítnu Ho

Kolmogorov-Smirnovův test shody

tentokrát pro dva výběry

Užívá se pro hodnocení shody rozdělení četností dvou srovnávaných výběrů.

Lze použít i tam, **kde nelze použít dvouvýběrový t-test pro nepárová data** – není-li splněna podmínka normality obou výběrových souborů

Podmínky použití:

V případě malých výběrů (n_1 a $n_2 < 40$) je podmínkou jejich stejný rozsah $n_1 = n_2 < 40$

V případě velkých výběrů (n_1 a $n_2 > 40$) nemusí mít stejný rozsah

($n_1 \neq n_2 > 40$)

Ho: dva výběrové soubory mají shodné rozdělení četností

H_A: dva výběrové soubory nemají shodné rozdělení četností

Kolmogorov-Smirnovův test shody

tentokrát pro dva výběry

- výsledky obou souborů měření rozdělíme do intervalů – stanovení experimentálních četností v jednotlivých intervalech pro oba soubory dat (stejný počet intervalů a hranice u obou souborů dat)
- vypočítáme jednotlivé kumulované relativní četnosti pro oba soubory dat (nutnost pracovat s relativními četnostmi protože soubory dat nemusí mít stejný rozsah – počet měření)
- stanovíme absolutní hodnoty rozdílů kumulovaných relativních četností v každém intervalu
- Jako testovací kritérium vezmu maximální hodnotu těchto rozdílů

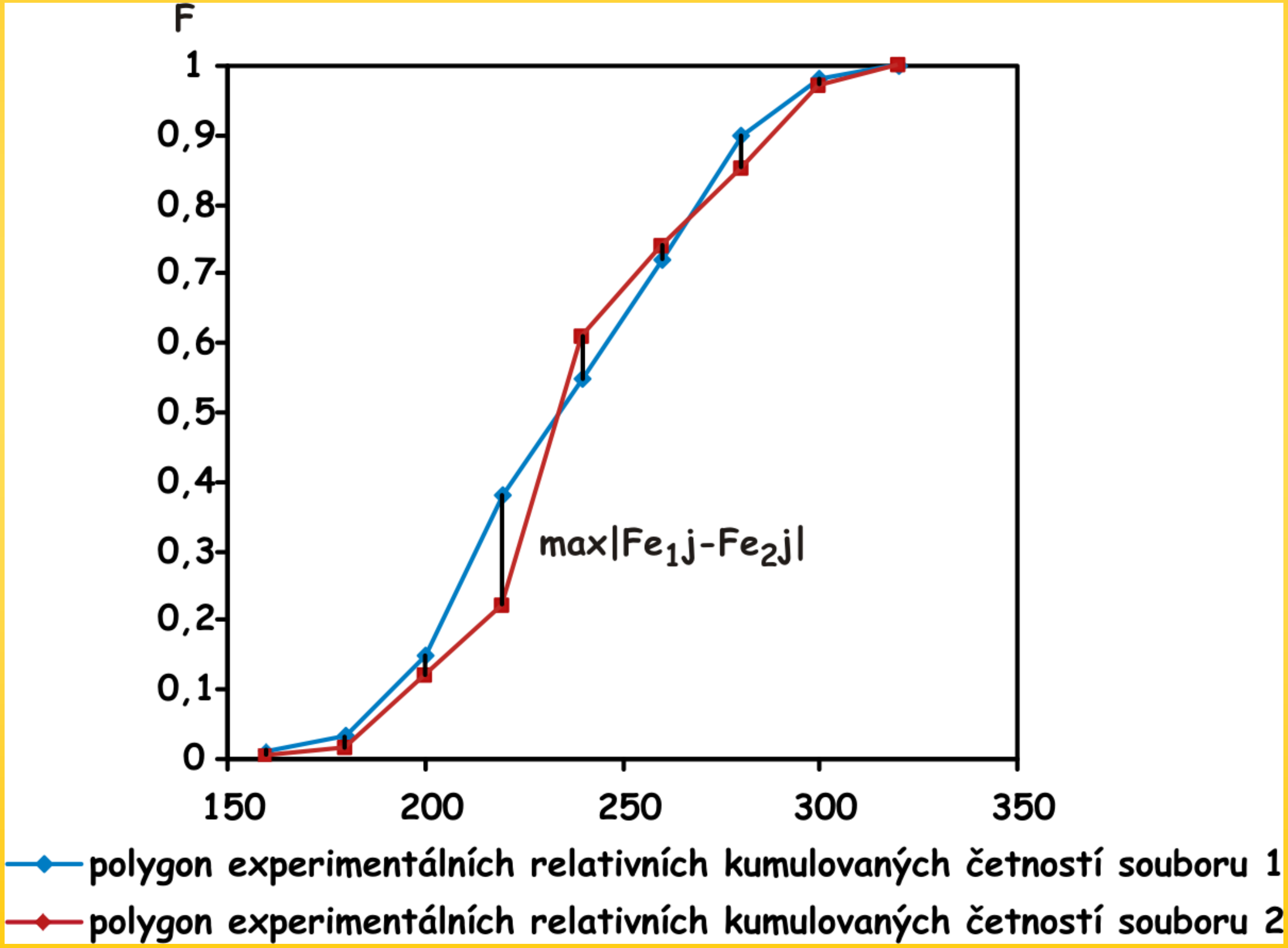
kde F_{1j}, F_{2j} jsou relativní kumulované četnosti souborů 1 a 2

- Kritická hodnota se dopočte podle vzorců (pro oboustrannou variantu testu a hladinu významnosti α 0,05 a 0,01):
- je-li $D_1 \leq D_k$ přijmeme H_0

Kolmogorov-Smirnovův test shody

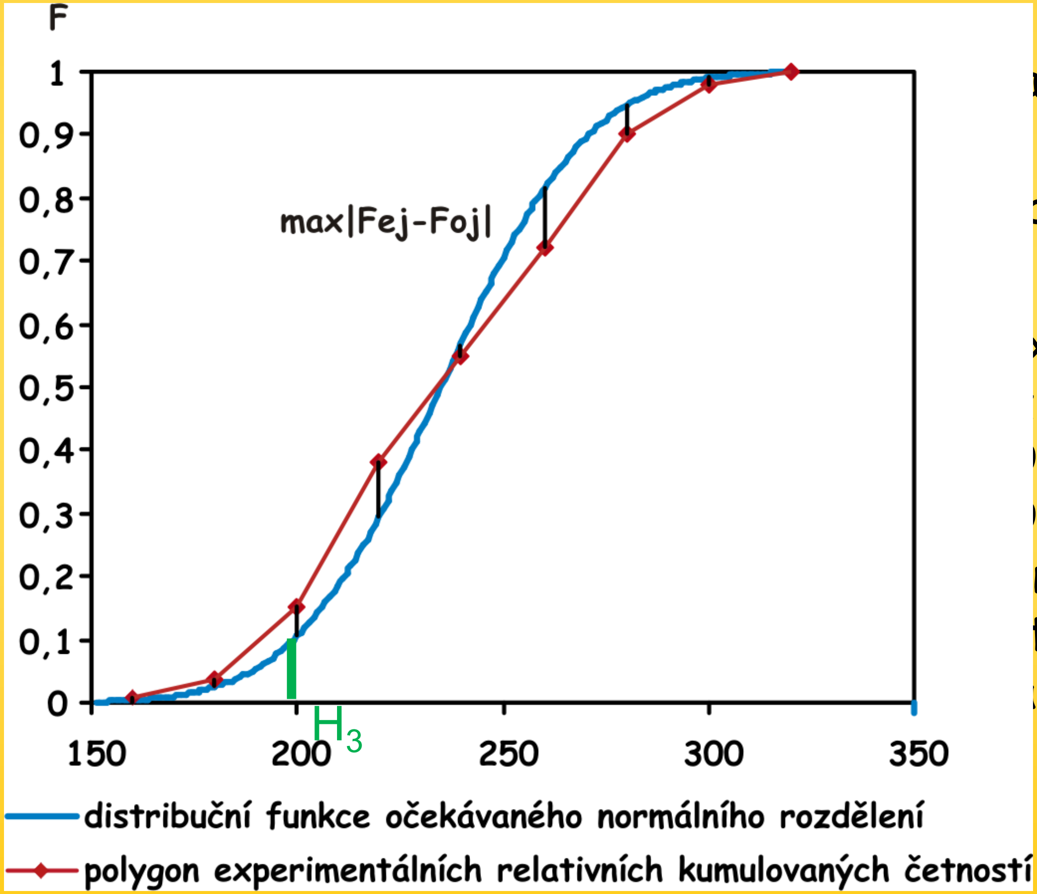
tentokrát pro dva výběry

grafické vyjádření stanovení testovacího kritéria



Princip konstrukce očekávaných četností pro normální rozdělení

při testování rozdělení pravděpodobností
Kolmogorov-Smirnovův test pro jeden výběr



anovení hodnot distribuční funkce
jednotlivé horní hranice;
př. v bodě H_3: $F(_{H3})$

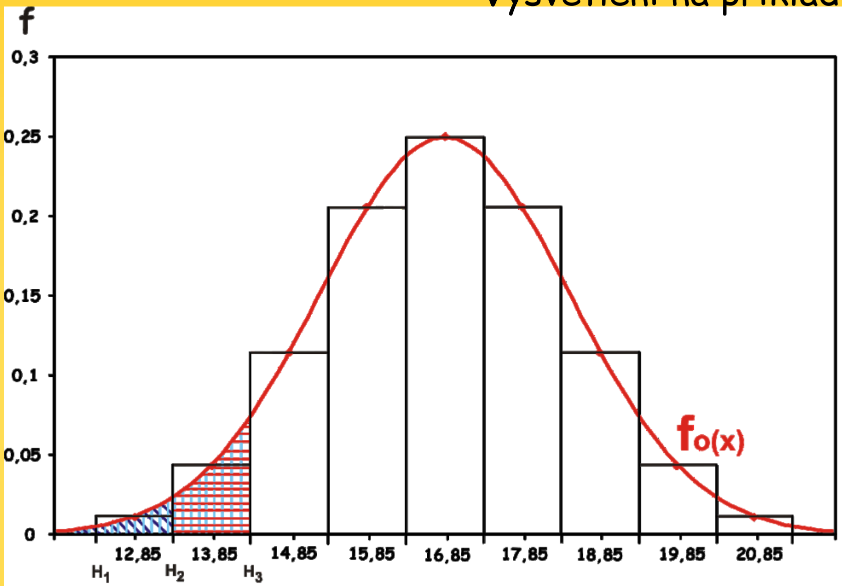
xcelu - např. pro normální
dělení - pomocí funkce
RM.DIST (nové MS Office) či
RMDIST (starší MS Office) –
novení hodnoty pravděpodobnosti
tribuční funkce pro všechny
nice intervalů

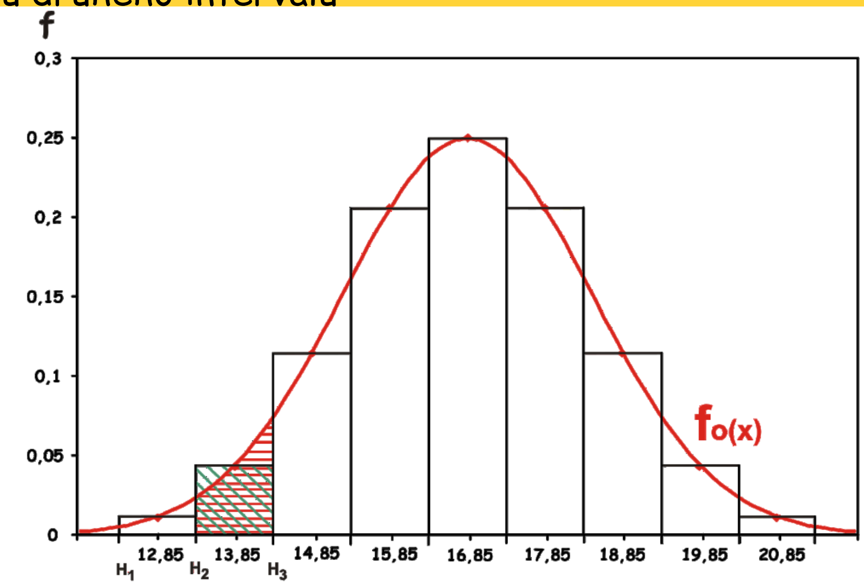
Princip konstrukce očekávaných četností pro normální rozdělení

při testování rozdělení pravděpodobností

Chí-kvadrát test

Vysvětlení na příkladu druhého intervalu





Hodnota distribuční funkce v bodě H2: $F_{(H_2)}$

Hodnota distribuční funkce v bodě H3: $F_{(H_3)}$

Pravděpodobnost hodnoty z 2. intervalu: $P_2 = F_{(H_3)} - F_{(H_2)}$

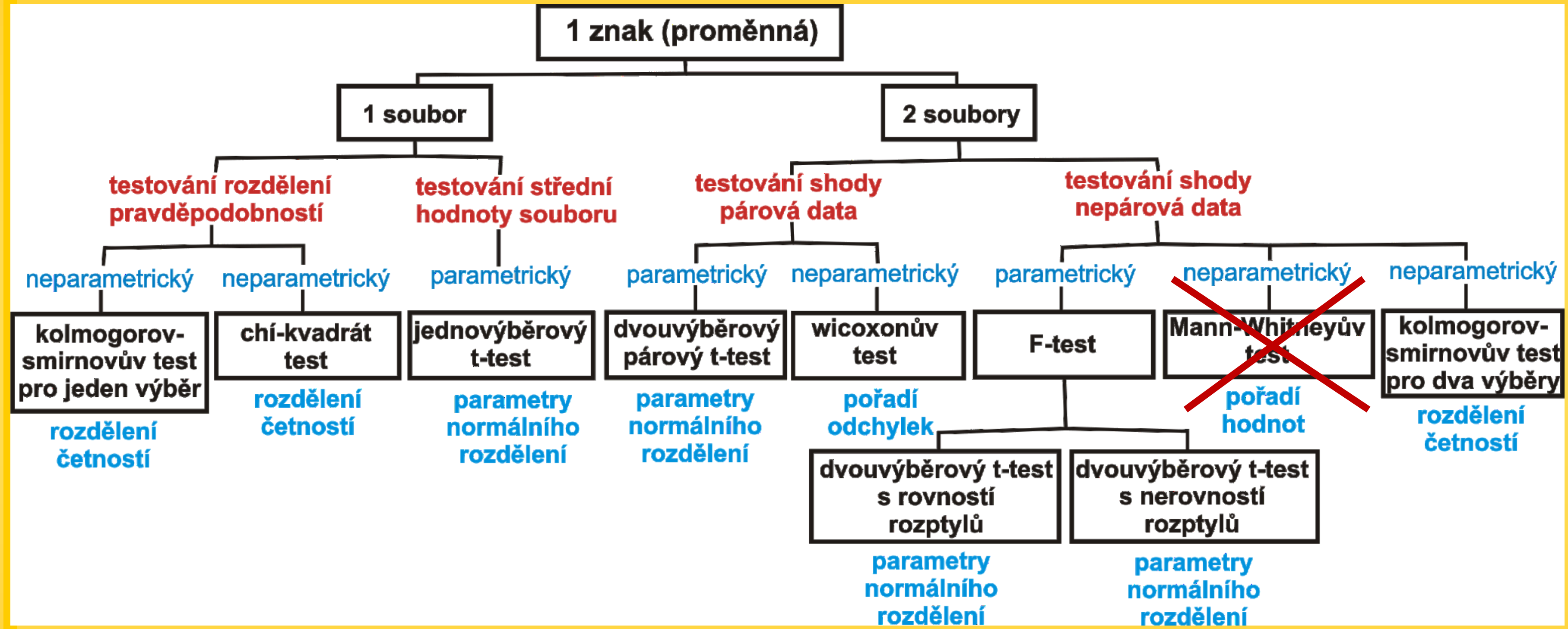
velikost červené plochy $\cong$ velikost zelené plochy

převedení očekávaných relativních četností na absolutní Pj*n pro všechny intervaly j = 1,2,....k (u chí-kvadrát testu dobré shody)

v excelu - např. pro normální rozdělení - funkce NORM.DIST (nové MS Office) či NORMDIST (starší MS Office) – stanovení hodnoty pravděpodobnosti distribuční funkce pro všechny hranice intervalů a odtud dopočtení pravděpodobnosti, že nastane hodnota z příslušného intervalu.

Úprava pravděpodobností pro všechny intervaly (j=1…k) na absolutní četnosti: $n_{oj}=p_jn$

Testování statistických hypotéz - shrnutí



Děkuji za pozornost