2

Metódy spracovania a hodnotenia hydrologických dát

Mgr. Monika Šulc, PhD.et PhD.

Hydrometria - *Prehľad základných pojmov*

Prietok (Q) – množstvo vody, ktoré preteká plochou prietočného profilu za jednotku času

Prietok je **najdôležitejšou** hydrologickou veličinou.

Vyjadruje sa v **$l.s^{-1}$** alebo (najčastejšie) v **$m^3.s^{-1}$.**

Prietok sa v krajine nemeria pravidelne – je spravidla odvodenou hodnotou.

Prietok sa odvodzuje z hodnoty **vodného stavu (H).**

Vodný stav je výška vodnej hladiny v mernom profile nad zvolenou úrovňou, t.j. **0 hodnotou.**

•Prehľad základných pojmov

Chronologická čiara vývoja financií študenta vždy ráno o 8.00 hod

podmieňujúcich prvkov za konkrétne časové obdobie.

•Metódy hodnotenia hydrologických prvkov

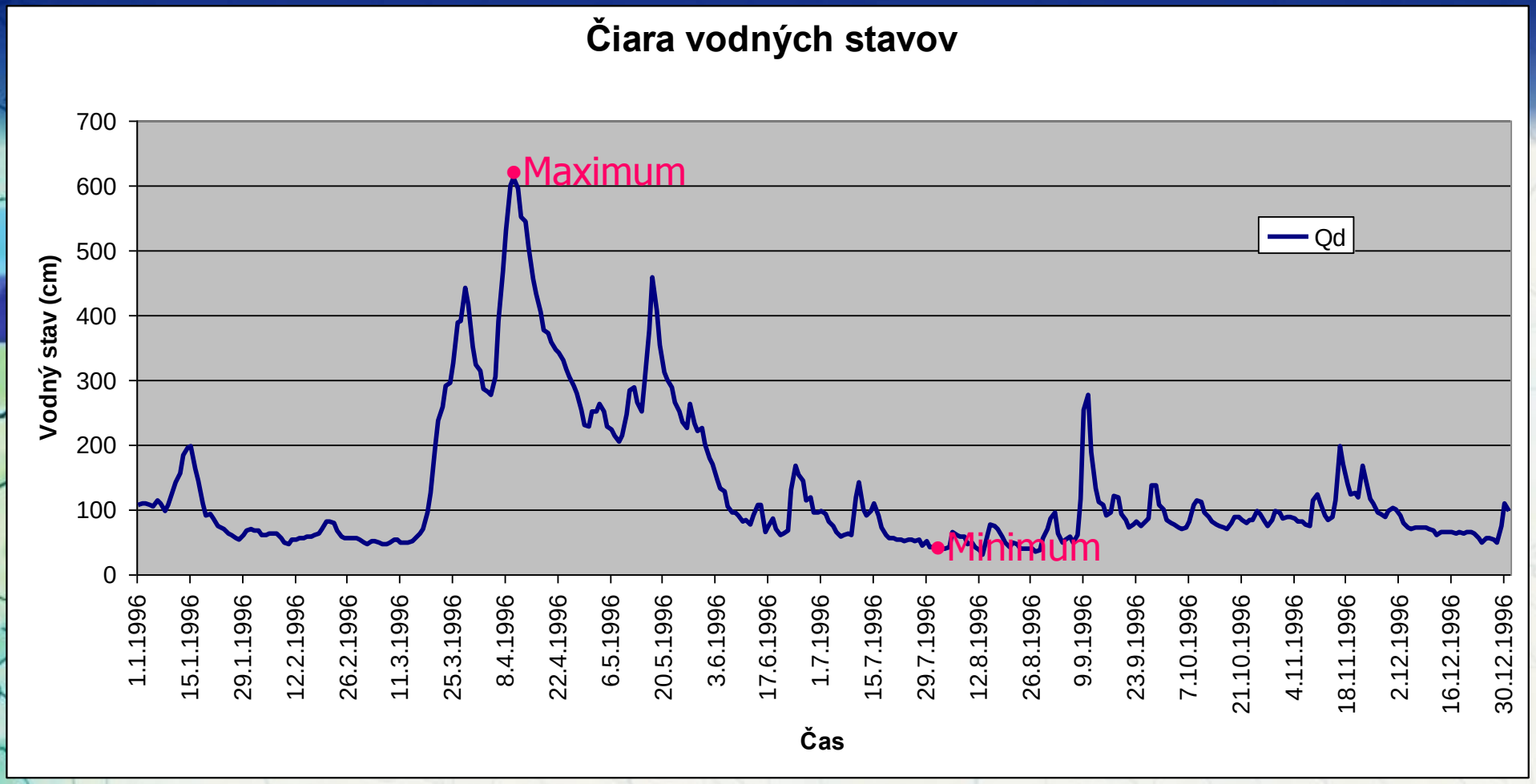
•Vodné stavy, prietoky

•Po nameraní hodnôt **vodných stavov** je základným spôsobom ich vyhodnotenia vykreslenie **chronologickej čiary vodných stavov (prietokov)** za určité zvolené obdobie.

•Z množiny **nameraných dát** sa týmto zoradením stáva **hydrologický rad.**

•Už z **chronologickej čiary** je možné zistiť **maximálny** a **minimálny** vodný stav za sledované obdobie. Rozdiel týchto hodnôt nazývame **variačné rozpätie – amplitúda.**

•Ďalšie hodnoty zaujímavé z hľadiska hodnotenia hydrologického radu získame využitím jednoduchých **metód matematickej štatistiky.**



- **Priemerné hodnoty** – mesačná, ročná, dlhodobá
- **Medián, modus ...**

- Pre riešenie **praktických** hydrologických úloh je potrebné hydrologické dáta **detailnejšie** spracovať.

- Najpoužívanejšou metódou hodnotenia hydrologických prvkov je **čiara prekročenia**, a to buď **empirická** alebo **teoretická**.

- **Empirická čiara** prekročenia je zostrojená z reálne nameraných údajov.

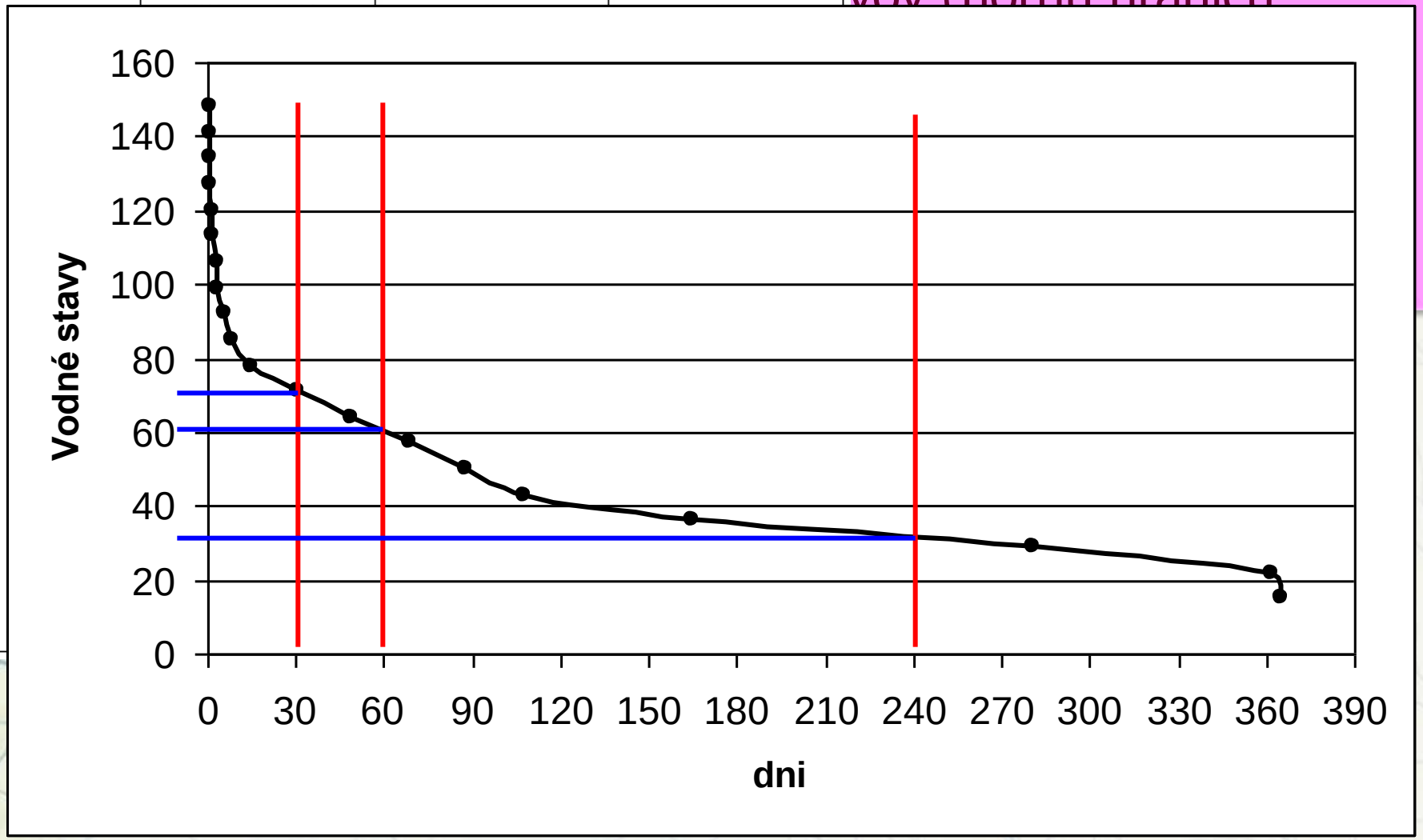
- **Teoretická čiara** (krivka – hydrologický model) **prekročenia** je zostrojená s využitím ďalších, najmä štatistických závislostí.

- Čiara prekročenia vyjadruje **dĺžku** (dobu), počas ktorej je **dosiahnutá a prekročená** hodnota určitého vodného stavu.

Tabuľka 10.1. Početnosť výskytu vodných stavov rieky Hron

Poradové číslo intervalu	Interval od	do	Počet vodných stavov	Kumulatívna početnosť
1	154	148	1	1
2	147	141	0	1
3	140	134	0	1
4	133	127	0	1

ho radu patrí do
latívnu početnosť.
ru prekročenia
hodnoty kumulatívnej
vov (hornú hranicu



•Hodnoty prvkov hydrologického radu sú náhodné hodnoty a platia pre ne zákony **počtu pravdepodobnosti**.

•Ak vychádzame z **empirickej pravdepodobnosti**, potom pravdepodobnosť výskytu zistíme pomerom:

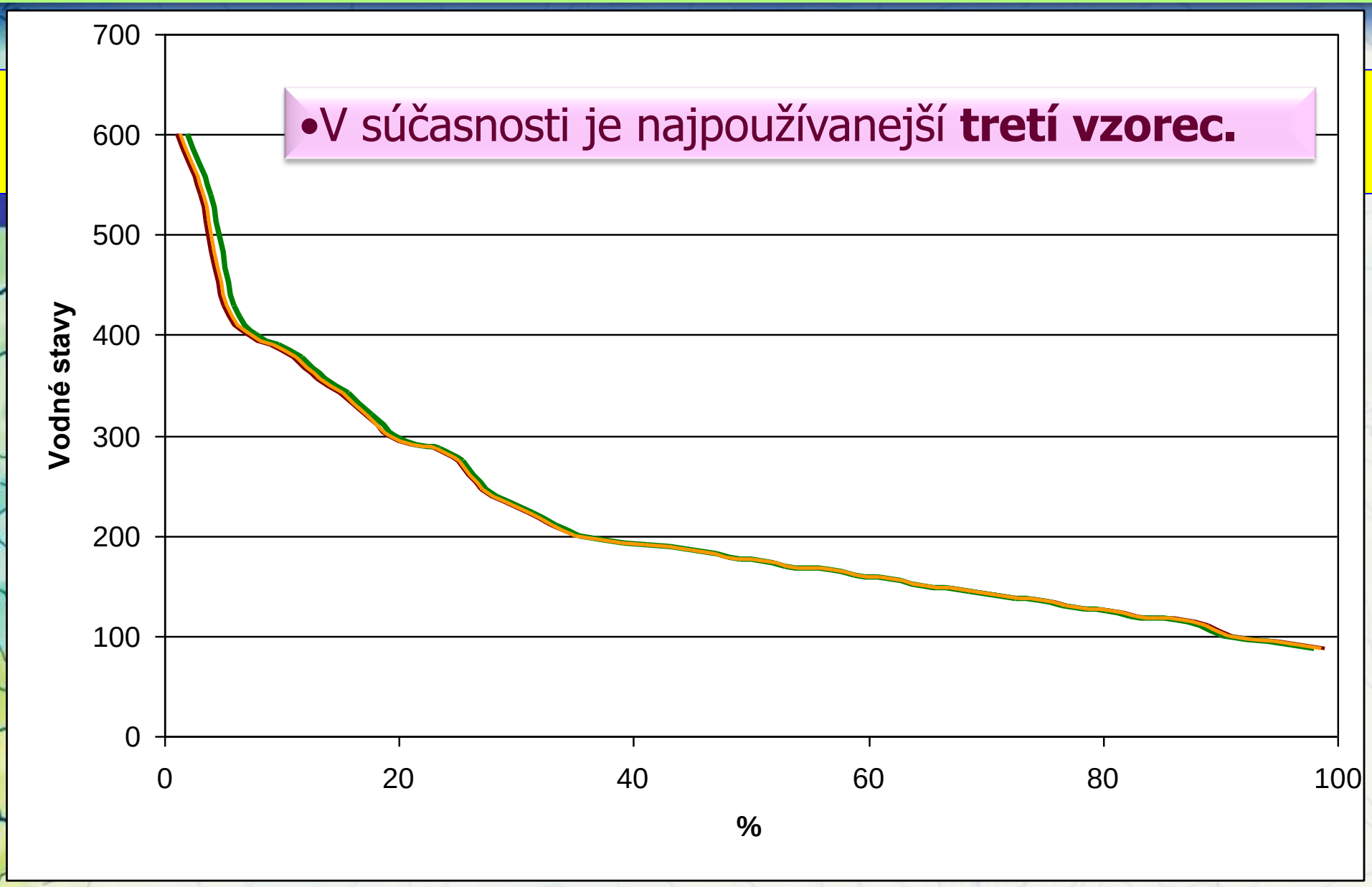
$$p = \frac{m}{n}$$

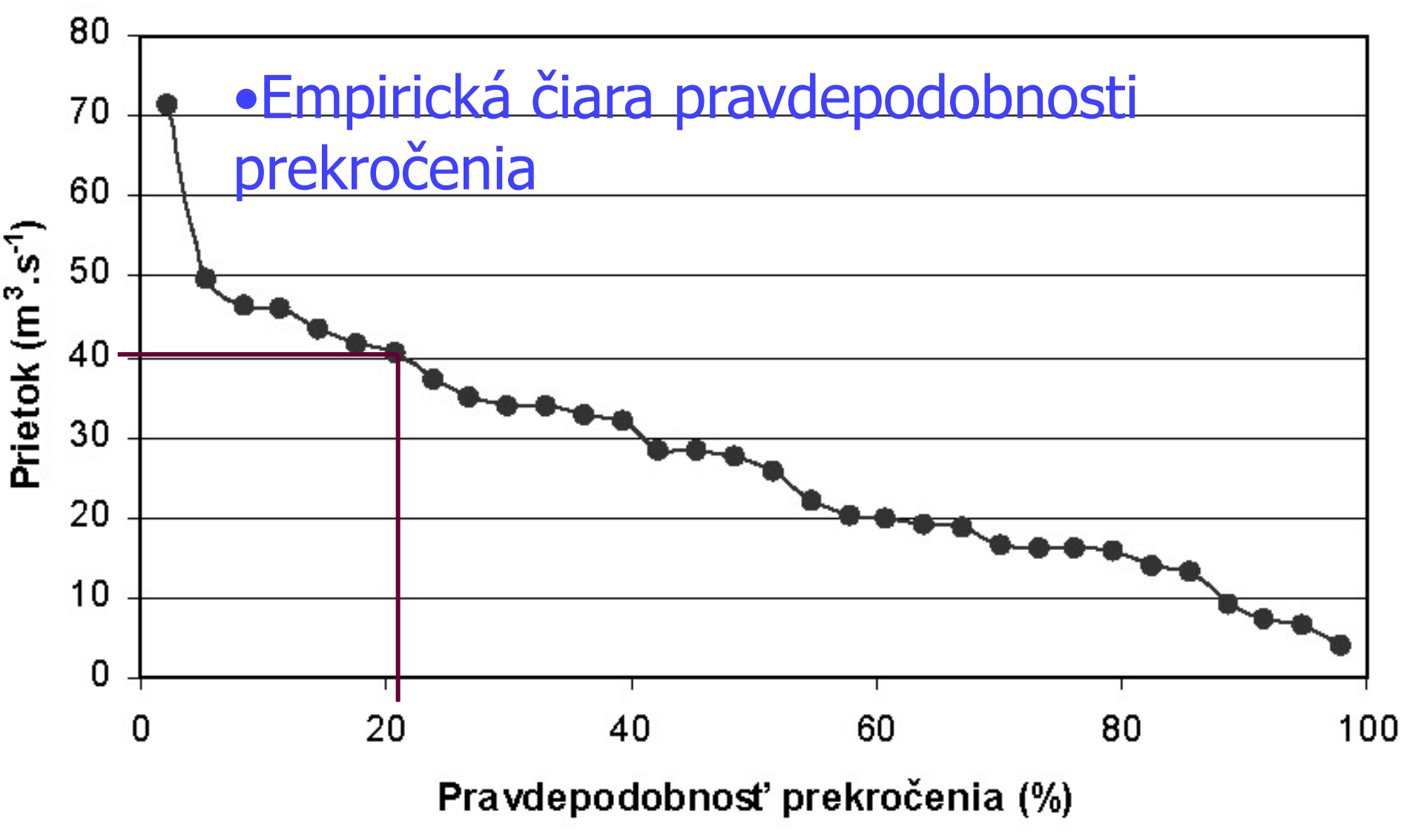
•kde **m** je počet kladných výsledkov javu pri **n**–násobnom opakovaní pokusu.

•Aplikácia tohto vzťahu v hydrológii vyzerá tak, že množinu prietokov zoradíme do **klesajúceho radu** a každej hodnote priradíme hodnotu pravdepodobnosti, pričom **m** je poradové číslo prvku v rade a **n** je celkový počet prvkov.

•Použitím vzorca by však **posledný, n-tý** prvok mal pravde-podobnosť výskytu **1, čiže 100%** a to je v hydrologickej praxi nereálne.

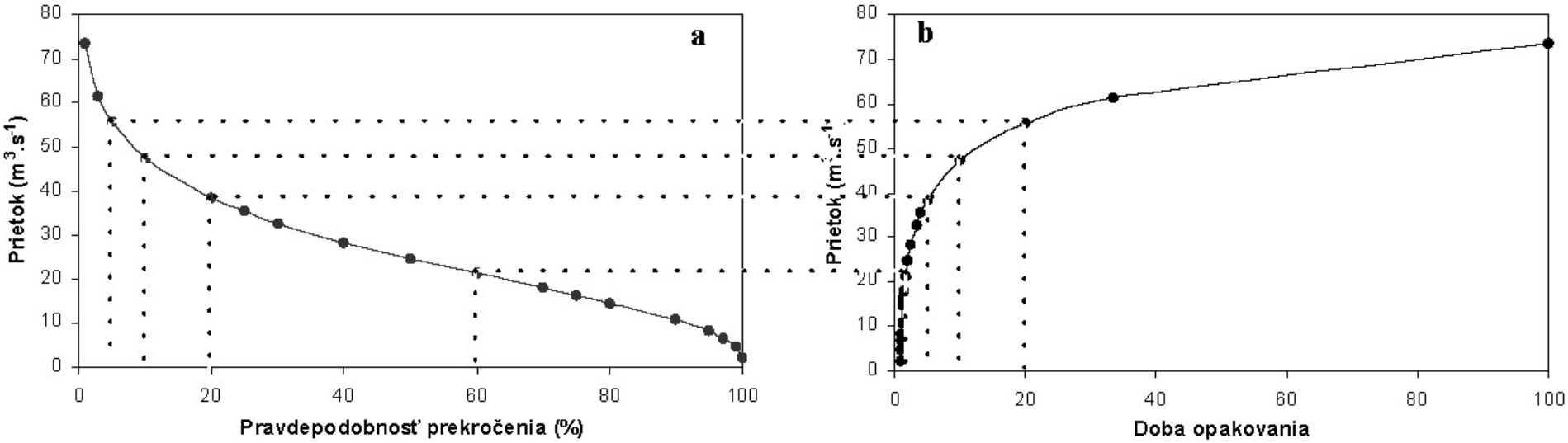
•Vzorec preto rôzni autori **upravili** nasledovne:




•Empirická čiara pravdepodobnosti prekročenia
Prietok ($m^3.s^{-1}$)
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100
Pravdepodobnosť prekročenia (%)

•Doteraz sme zaoberali **empirickým rozdelením** početnosti. Ale empíria má svoje **limity...**

•Od **empirického** rozdelenia musíme prejsť k **teoretickému** ...



$$P = \frac{1}{N}.100\%$$

alebo

$$N = \frac{1}{P}.100\%$$

•V hydrológii v praxi využívame **teoretické rozdelenie** početnosti, ktoré môžeme považovať za **matematický model** daného empirického rozdelenia, ktorý zostrojíme na základe parametrov získaných z empirického radu.

•Teoretické rozdelenie početnosti môže byť **symetrické** alebo **asymetrické**.

•Symetrické rozdelenie početnosti vyjadruje Gaussova – Laplaceova krivka **normálneho rozdelenia**. Keďže krivka je symetrická, **priemer, medián a modus** sú totožné.

•V hydrologickej praxi sa stretávame prevažne s **asymetrickým** rozdelením početnosti, keď sú hydrologické rady ohraničené konečnými **maximálnymi** a **minimálnymi** hodnotami.

•Z asymetrických kriviek rozdelenia početnosti je najznámejšia jedna z dvanástich kriviek štatistika Pearsona, známa ako **Pearsonova krivka III. typu.**

•Tvar a priebeh krivky sú určené **tromi parametrami**:

- **1.** aritmetickým priemerom radu
- **2.** koeficientom variácie **Cv**
- **3.** koeficientom asymetrie **Cs**

•Pri výpočte koeficientu variácie zavedieme do vzorca pre výpočet Cv hodnotu **ki,** ako:

$$k_i = \frac{x_i}{\overline{x}}$$

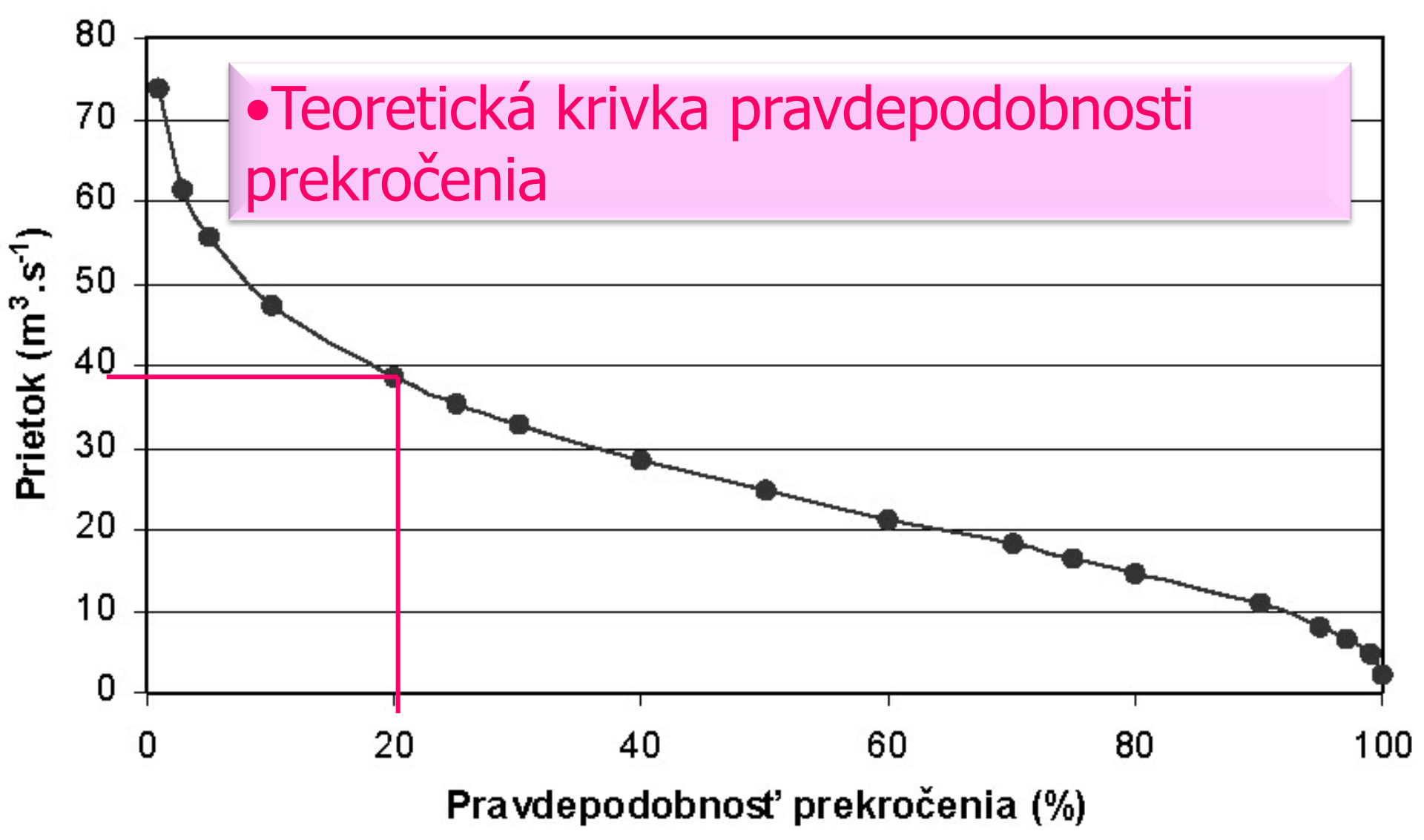
•kde $\mathbf{x_i}$ je aktuálna hodnota prietoku a **x**je priemerná hodnota celého radu

- Matematický aparát „modelu“ potom vyzerá nasledovne:

$$Cv = \sqrt{\frac{\sum (k_i - 1)^2}{n - 1}}$$

$$Cs = \frac{\sum (k_i - 1)^3}{(n - 1).Cv^3}$$

$$Qp = \bar{x}.(1 + \Phi_{s,p})$$


•Teoretická krivka pravdepodobnosti prekročenia
Prietok ($m^3.s^{-1}$)
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
Pravdepodobnosť prekročenia (%)

•Ďalšie **štatistické rozdelenia:**

- **Logaritmicko – normálne rozdelenie**, ktoré je vhodné pri súboroch s veľkou asymetriou (Cs > 3Cv). U nás je využívané najmä pri maximálnych prietokoch.
- **Goodrichovo exponenciálne rozdelenie** (na hodnotenie maximálnych prietokov) najmä pri väčšom počte ročných kulminácií.
- **Weibullovo rozdelenie** (na hodnotenie minimálnych prietokov).
- **Gumbelovo rozdelenie.**
- Na základe Pearsonovej krivky odvodili svoju krivku **Krickij** a **Menkeľ**, používa sa aj **logaritmický variant** Pearsonovej krivky.

•Jednotlivé parametre, potrebné na zostrojenie kriviek, môžeme získať viacerými spôsobmi. Najčastejšie sa používajú **metóda momentov** a **metóda kvantilov**.

•Pri metóde momentov sa parametre Qa, Cv a Cs vypočítajú zo známych vzťahov.

•Krivku prekročenia konštruujeme s využitím **Foster–Rybkinových** tabuliek.

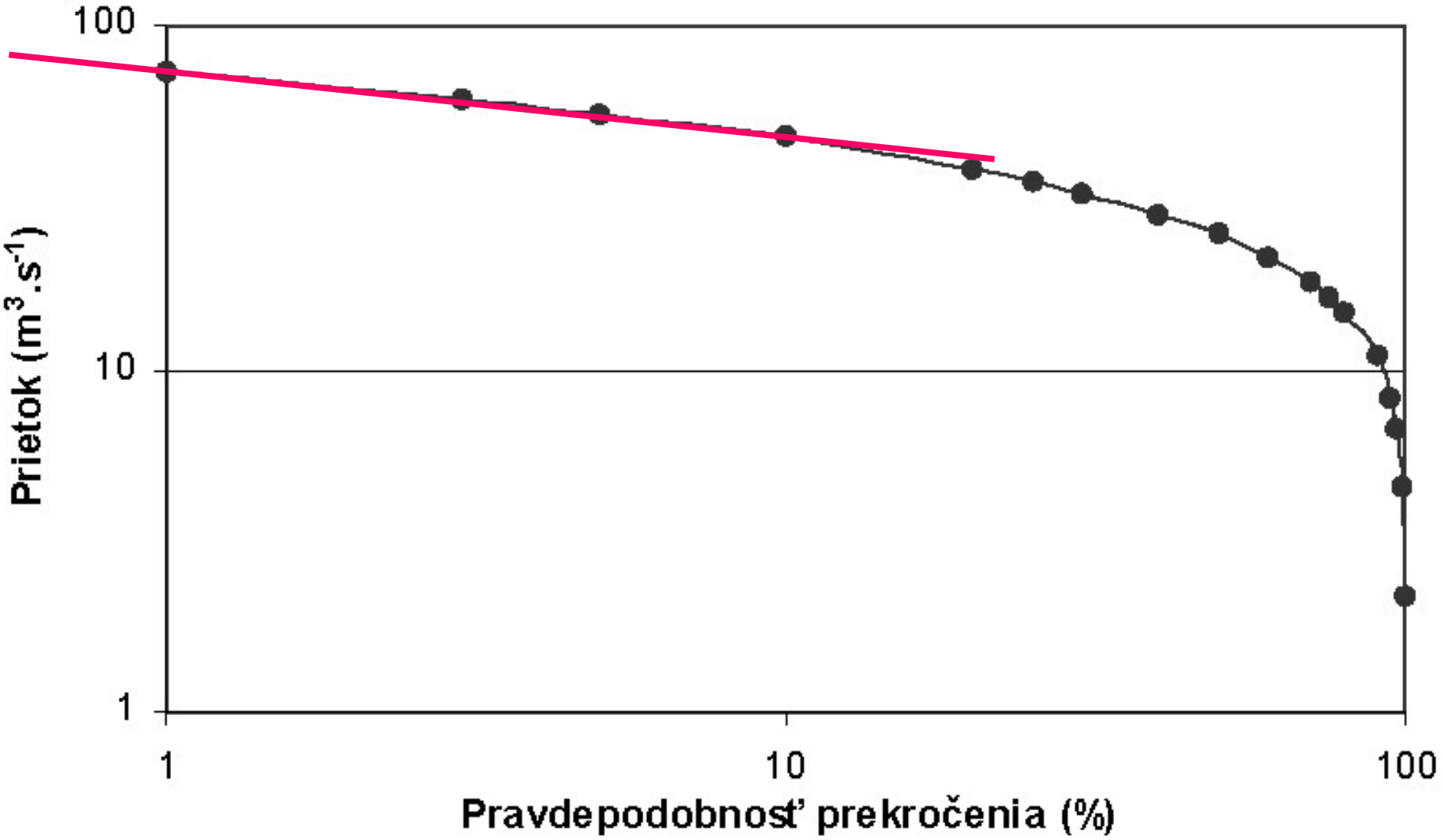
•Sú v nich uvedené odchýlky súradníc kriviek – od priemeru pre x = 1 a Cv = 1 pre rôzne hodnoty Cs.

Tabuľka 11.2 Odchýlky poradníc krivky prekročenia od priemeru pri Cv = 1 podľa S. Fostera a J. Rybkina

Cs	Pravdepodobnosť prekročenia p (v %)																				
	0,01	0,05	0,1	1	3	5	10	20	25	30	40	50	60	70	75	80	90	95	97	99	99,9
0,00	3,72	3,29	3,09	2,33	1,88	1,64	1,28	0,84	0,67	0,52	0,25	0,00	– 0,52	– 0,52	– 0,67	– 0,84	– 1,28	– 1,64	– 1,88	– 2,33	– 3,09
0,05	3,83	3,38	3,16	2,36	1,90	1,65	1,28	0,84	0,66	0,52	0,24	– 0,01	– 0,26	– 0,52	– 0,68	– 0,84	– 1,28	– 1,67	– 1,86	– 2,29	– 3,02
0,10	3,94	3,46	3,23	2,40	1,92	1,67	1,29	0,84	0,66	0,51	0,24	– 0,02	– 0,27	– 0,53	– 0,68	– 0,85	– 1,27	– 1,61	– 1,84	– 2,25	– 2,95
0,15	4,05	3,54	3,31	2,44	1,94	1,68	1,30	0,84	0,66	0,50	0,23	– 0,02	– 0,28	– 0,54	– 0,68	– 0,85	– 1,26	– 1,60	– 1,82	– 2,22	– 2,88
0,20	4,16	3,62	3,38	2,47	1,96	1,70	1,30	0,83	0,65	0,50	0,22	– 0,03	– 0,28	– 0,55	– 0,69	– 0,85	– 1,26	– 1,58	– 1,79	– 2,18	– 2,81
0,25	4,27	3,70	3,45	2,50	1,98	1,71	1,30	0,82	0,64	0,49	0,21	– 0,04	– 0,29	– 0,56	– 0,70	– 0,85	– 1,25	– 1,56	– 1,77	– 2,14	– 2,74
0,30	4,38	3,79	3,52	2,54	2,00	1,72	1,31	0,82	0,64	0,48	0,20	– 0,05	– 0,30	– 0,56	– 0,70	– 0,85	– 1,24	– 1,55	– 1,75	– 2,10	– 2,61
0,35	4,50	3,88	3,59	2,58	2,02	1,73	1,32	0,82	0,64	0,48	0,20	– 0,06	– 0,30	– 0,56	– 0,70	– 0,85	– 1,24	– 1,53	– 1,72	– 2,06	– 2,60
0,40	4,61	3,96	3,66	2,61	2,04	1,75	1,32	0,82	0,63	0,47	0,19	– 0,07	– 0,31	– 0,57	– 0,71	– 0,85	– 1,23	– 1,52	– 1,70	– 2,03	– 2,54
0,45	4,72	4,04	3,74	2,64	2,06	1,76	1,32	0,82	0,62	0,46	0,18	– 0,08	– 0,32	– 0,58	– 0,71	– 0,85	– 1,22	– 1,51	– 1,68	– 2,00	– 2,47
0,50	4,83	4,12	3,81	2,68	2,08	1,77	1,32	0,81	0,62	0,46	0,18	– 0,08	– 0,33	– 0,58	– 0,71	– 0,85	– 1,22	– 1,49	– 1,66	– 1,96	– 2,40
0,55	4,94	4,20	3,88	2,72	2,10	1,78	1,32	0,80	0,62	0,45	0,16	– 0,09	– 0,34	– 0,58	– 0,72	– 0,85	– 1,21	– 1,47	– 1,64	– 1,92	– 2,32
0,60	5,05	4,29	3,96	2,75	2,12	1,80	1,33	0,80	0,61	0,44	0,16	– 0,10	– 0,34	– 0,59	– 0,72	– 0,85	– 1,20	– 1,45	– 1,61	– 1,88	– 2,20
0,65	5,16	4,38	4,03	2,78	2,14	1,81	1,33	0,80	0,60	0,44	0,15	– 0,11	– 0,35	– 0,60	– 0,72	– 0,85	– 1,19	– 1,44	– 1,59	– 1,84	– 2,20
0,70	5,28	4,46	4,10	2,82	2,15	1,82	1,33	0,78	0,59	0,43	0,14	– 0,12	– 0,36	– 0,60	– 0,72	– 0,85	– 1,18	– 1,42	– 1,57	– 1,81	– 2,14
0,75	5,39	4,54	4,17	2,86	2,16	1,83	1,34	0,78	0,58	0,42	0,13	– 0,12	– 0,36	– 0,60	– 0,72	– 0,86	– 1,18	– 1,40	– 1,54	– 1,78	– 2,08
0,80	5,50	4,63	4,24	2,89	2,18	1,84	1,34	0,78	0,58	0,41	0,12	– 0,13	– 0,37	– 0,60	– 0,73	– 0,86	– 1,17	– 1,38	– 1,52	– 1,74	– 2,02
0,85	5,62	4,72	4,31	2,92	2,20	1,85	1,34	0,78	0,58	0,40	0,12	– 0,14	– 0,38	– 0,60	– 0,73	– 0,86	– 1,16	– 1,36	– 1,49	– 1,70	– 1,96
0,90	5,73	4,80	4,38	2,96	2,22	1,86	1,34	0,77	0,57	0,40	0,11	– 0,15	– 0,38	– 0,61	– 0,73	– 0,85	– 1,15	– 1,35	– 1,47	– 1,66	– 1,90
0,95	5,84	4,88	4,46	2,99	2,24	1,87	1,34	0,76	0,56	0,39	0,10	– 0,16	– 0,38	– 0,62	– 0,73	– 0,85	– 1,14	– 1,34	– 1,44	– 1,62	– 1,84
1,00	5,96	4,97	4,53	3,02	2,25	1,88	1,34	0,76	0,55	0,38	0,09	– 0,16	– 0,39	– 0,62	– 0,73	– 0,85	– 1,13	– 1,32	– 1,42	– 1,59	– 1,79
1,10	6,18	5,13	4,67	3,09	2,28	1,89	1,34	0,74	0,54	0,36	0,07	– 0,18	– 0,41	– 0,62	– 0,74	– 0,85	– 1,10	– 1,28	– 1,38	– 1,52	– 1,68
1,20	6,41	5,30	4,81	3,15	2,31	1,91	1,34	0,73	0,52	0,35	0,05	– 0,19	– 0,42	– 0,63	– 0,74	– 0,84	– 1,08	– 1,24	– 1,33	– 1,45	– 1,58
1,30	6,54	5,46	4,95	3,21	2,34	1,92	1,34	0,72	0,51	0,33	0,04	– 0,21	– 0,43	– 0,63	– 0,74	– 0,84	– 1,06	– 1,20	– 1,28	– 1,38	– 1,48
1,40	6,87	5,63	5,09	3,27	2,37	1,94	1,34	0,71	0,49	0,31	0,02	– 0,22	– 0,44	– 0,64	– 0,73	– 0,83	– 1,04	– 1,17	– 1,23	– 1,32	– 1,39
1,50	7,09	5,80	5,23	3,33	2,39	1,95	1,33	0,70	0,47	0,30	0,00	– 0,24	– 0,45	– 0,64	– 0,73	– 0,82	– 1,02	– 1,13	– 1,19	– 1,26	– 1,31
1,60	7,31	5,96	5,37	3,39	2,42	1,96	1,33	0,68	0,46	0,28	– 0,02	– 0,25	– 0,46	– 0,64	– 0,73	– 0,81	– 0,99	– 1,10	– 1,14	– 1,20	– 1,24
1,70	7,54	6,12	5,50	3,44	2,44	1,97	1,32	0,66	0,44	0,26	– 0,03	– 0,27	– 0,47	– 0,64	– 0,72	– 0,81	– 0,97	– 1,06	– 1,10	– 1,14	– 1,17
1,80	7,76	6,28	5,64	3,50	2,46	1,98	1,32	0,64	0,42	0,24	– 0,05	– 0,28	– 0,48	– 0,64	– 0,72	– 0,80	– 0,94	– 1,02	– 1,06	– 1,09	– 1,11
1,90	7,98	6,44	5,77	3,55	2,49	1,99	1,31	0,63	0,40	0,22	– 0,07	– 0,29	– 0,48	– 0,64	– 0,72	– 0,79	– 0,92	– 0,98	– 1,01	– 1,04	– 1,05
2,00	8,21	6,60	5,91	3,60	2,51	2,00	1,30	0,61	0,39	0,20	– 0,08	– 0,31	– 0,49	– 0,64	– 0,71	– 0,78	– 0,90	– 0,95	– 0,97	– 0,99	– 1,00
2,10			6,06	3,65	2,53	2,00	1,29	0,60	0,38	0,19	– 0,10	– 0,32	– 0,49	– 0,64	– 0,70	– 0,76	– 0,88	– 0,93	– 0,93	– 0,94	– 0,95
2,20			6,20	3,70	2,55	2,01	1,28	0,58	0,37	0,17	– 0,11	– 0,33	– 0,49	– 0,63	– 0,69	– 0,75	– 0,85	– 0,90	– 0,90	– 0,90	– 0,91
2,30			6,34	3,75	2,56	2,01	1,27	0,56	0,35	0,15	– 0,12	– 0,34	– 0,49	– 0,62	– 0,68	– 0,73	– 0,82	– 0,86	– 0,86	– 0,87	– 0,87
2,40			6,47	3,79	2,57	2,01	1,25	0,54	0,33	0,13	– 0,14	– 0,35	– 0,50	– 0,62	– 0,66	– 0,71	– 0,79	– 0,82	– 0,82	– 0,83	– 0,83
2,50			6,60	3,83	2,58	2,01	1,24	0,53	0,32	0,12	– 0,15	– 0,36	– 0,50	– 0,61	– 0,65	– 0,70	– 0,77	– 0,79	– 0,79	– 0,80	– 0,80
2,60			6,73	3,87	2,59	2,01	1,23	0,51	0,30	0,10	– 0,17	– 0,37	– 0,50	– 0,60	– 0,64	– 0,68	– 0,74	– 0,76	– 0,76	– 0,77	– 0,77
2,70			6,86	3,91	2,60	2,01	1,21	0,49	0,28	0,08	– 0,18	– 0,38	– 0,50	– 0,60	– 0,63	– 0,67	– 0,72	– 0,73	– 0,73	– 0,74	– 0,74
2,80			6,99	3,95	2,61	2,02	1,20	0,47	0,27	0,06	– 0,20	– 0,38	– 0,50	– 0,59	– 0,62	– 0,65	– 0,70	– 0,71	– 0,71	– 0,71	– 0,71
2,90			7,12	3,99	2,62	2,02	1,19	0,45	0,26	0,04	– 0,21	– 0,39	– 0,50	– 0,58	– 0,61	– 0,64	– 0,67	– 0,68	– 0,68	– 0,69	– 0,69
3,00			7,29	4,02	2,63	2,02	1,18	0,42	0,25	0,03	– 0,23	– 0,40	– 0,50	– 0,57	– 0,61	– 0,62	– 0,65	– 0,66	– 0,66	– 0,67	– 0,67


AXIS 211 2004-09-25 08:35:05

•Krivky prekročenia, či už empirické alebo teoretické, môžeme vykresliť buď v **normálnej, semilogaritmickej** alebo v **logaritmickej sieti** pravouhlých súradníc.



Dvojrozmerná induktívna štatistika

Jednoduchá lineárna regresia, Pearsonov korelačný koeficient

Jednoduchá lineárna regresia

- Párová regresná analýza skúma lineárnu závislosť medzi dvoma kvantitatívnymi premennými (napr. hmotnosťou a výškou človeka) a je špecifickým prípadom viacnásobnej regresie. Jednoduchá regresia odhaduje regresné koeficienty β_0 a β_1 v rovnici:

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- kde:
  y_i – hodnota závislej premennej Y (kritéria) v *i*-tom pozorovaní
  x_i – hodnota nezávislej premennej X (prediktora) v *i*-tom pozorovaní
  β_0 – regresná konštanta (priesečník regresnej priamky s osou x)
  β_1 – regresný koeficient (smernica regresnej priamky)
  ε_i – náhodná chyba *i*-teho pozorovania

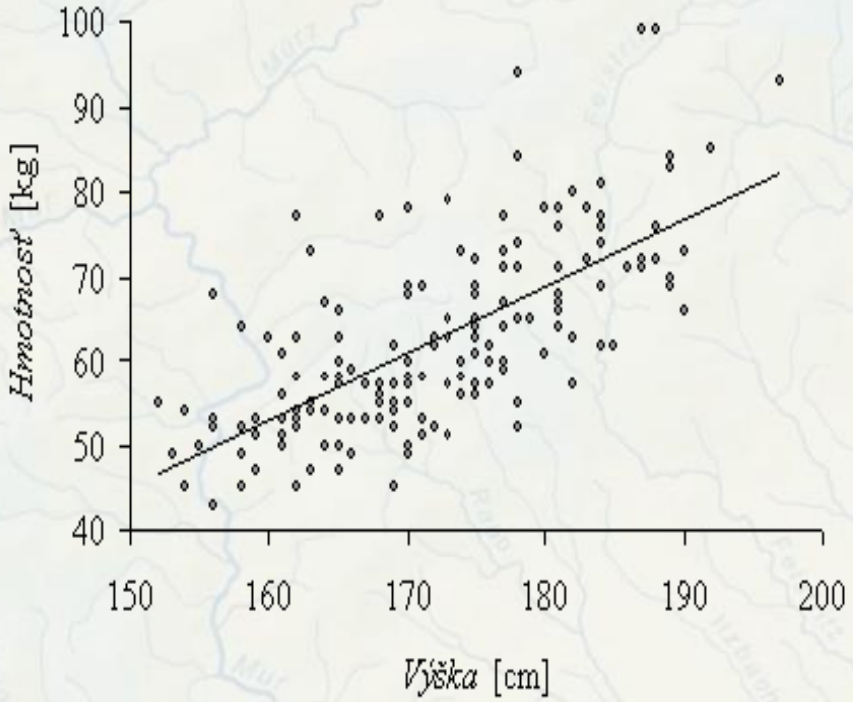
Jednoduchá lineárna regresia

- Regresný koeficient sa interpretuje v závislosti od typu výskumu. V prípade experimentu (v ktorom sa premennou X manipuluje), vyjadruje o koľko sa zvýši očakávaná hodnota premennej Y ak sa hodnota premennej X zvýši o 1 jednotku.
- V prípade pozorovacej štúdie sa koeficient interpretuje ako očakávaný rozdiel hodnôt premennej Y dvoch pozorovaní, ktorých hodnota premennej X sa líši o jednu jednotku.

- Za predpokladu, že údaje predstavujú náhodnú vzorku z populácie, sú vypočítané regresné koeficienty a korelačný koeficient najlepšími bodovými odhadmi neznámych parametrov. Okrem toho možno testovať hypotézy (nulová hypotéza, že koeficient sa rovná nule vyjadruje, že medzi premennými v základnom súbore neexistuje vzťah) a zostrojiť ich intervalové odhady.
- Testy hypotéz a intervalové odhady regresných koeficientov predpokladajú, že chyby ε_i sú vzájomne nezávislé (z čoho vyplýva, že aj y_i sú nezávislé), normálne rozdelené s priemerom 0 a rovnakým rozptylom pre všetky hodnoty X.

Jednoduchá lineárna regresia

- Na základe vzorky *n* pozorovaní premenných X a Y, metóda najmenších štvorcov odhadne neznáme parametre β_0 a β_1 tak, aby bol súčet druhých mocnín rezíduí minimálny. Rezíduum *ei* je rozdiel medzi skutočnou hodnotou závislej premennej y_i a hodnotou vypočítanou z regresnej funkcie dosadením hodnoty x_i:
- Rezíduum predstavuje vertikálnu vzdialenosť medzi bodom a regresnou priamkou:



Ciele regresnej analýzy môžu byť rôzne:

1. Nájdenie rovnice, ktorá opisuje vzťah medzi premennými
2. Odhad koeficientov - regresná analýza môže potvrdiť teóriu o vzťahu medzi premennými. Najčastejšie je záujem sústredený na znamienka a veľkosti koeficientov
3. Predikcia - Cieľom je predpovedať hodnoty závislej premennej

- **Korelačný koeficient** meria silu štatistickej závislosti medzi dvoma kvantitatívnymi premennými. Korelačná analýza na rozdiel od regresie nevyjadruje príčinno-následný vzťah Y=f(X).
- Premenná Y nezávisí na premennej X ale dve náhodné premenné X a Y sa spoločne menia.
- Regresná analýza predpokladá, že premenná Y je náhodná a premenná X fixná.
- Pod pojmom korelačný koeficient sa najčastejšie myslí **Pearsonov korelačný koeficient** (Pearson's product moment) z roku 1896, ktorý je mierou lineárnej závislosti dvoch premenných. Pearsonov korelačný koeficient ρ (ró) odhadnutý z náhodnej vzorky sa zapisuje *r* a vypočíta sa:

- Čitateľ sa nazýva **kovariancia** a vyjadruje ako sa súčasne menia hodnoty dvoch premenných. Kladná hodnota znamená, že sa menia spoločne jedným smerom, záporná hodnota znamená že sa menia opačným smerom a nula, že sa menia nezávisle.
- Vydelením kovariancie štandardnými odchýlkami sa vypočíta korelačný koeficient, ktorého hodnota sa nachádza v intervale od −1 do 1. Pearsonov korelačný koeficient sa rovná −1 v prípade, že všetky pozorovania ležia na klesajúcej priamke a 1 ak pozorovania ležia na stúpajúcej priamke.

- **Interpretácia korelačného koeficientu** závisí od kontextu.
- Hodnota 0,8 pri overení fyzikálneho zákona použitím presných meracích prístrojov je veľmi nízka, v sociálnych vedách je však veľmi vysoká.
- Cohen (1988) vytvoril jednoduchú pomôcku pre interpretáciu korelačných koeficientov v psychologickom výskume: Korelácia (v absolútnej hodnote) pod 0,1 je triviálna, 0,1–0,3 malá, 0,3–0,5 stredná a nad 0,5 veľká.

- Hodnota r 2 (R-squared) sa nazýva **koeficient determinácie** a vyjadruje podiel spoločnej variability medzi dvoma premennými. Test významnosti Pearsonovho korelačného koeficientu a intervalový odhad vyžadujú nezávislé pozorovania a bivariačné normálne rozdelenie.
- Pearsonov korelačný koeficient je silne ovplyvniteľný extrémnymi hodnotami (outliers) a to v oboch smeroch.
- Jediný extrémista vo veľkom súbore môže významne znížiť silnú závislosť, ale aj vyrobiť silnú závislosť tam, kde žiadna nie je.
- Touto citlivosťou na extrémne hodnoty netrpia poradové korelačné koeficienty.
- Dôležité závery sa nesmú robiť iba na základe hodnoty koeficientu. Vždy je nutné preskúmať X-Y graf.
- Z grafu možno zistiť aj nelineárny ale silný vzťah medzi premennými. V takom prípade treba vzťah linearizovať transformáciou premenných (napr. logaritmovaním Y), ktoré sa následne použijú na výpočet korelácie.

- Príklady:
  Existuje vzťah medzi množstvom konzumácie kávy (X) a krvným tlakom (Y)?
  Aká silná je závislosť medzi veľkosťou predaja výrobku (Y) a výdavkami na reklamu (X)?
  Aký nárast predaja možno očakávať, ak zvýšime výdavky na reklamu o 1 mil. Sk?

Testy štatistických hypotéz (testy štatistickej významnosti)

- Štatistická hypotéza je tvrdenie týkajúce sa základného súboru. V prípade parametrických testov je hypotéza tvrdenie o neznámej hodnote parametra základného súboru.
- Iba na základe výskumu celého základného súboru by bolo možné s úplnou istotou rozhodnúť o správnosti alebo nesprávnosti hypotézy. Takýto vyčerpávajúci výskum by však bol neekonomický, technicky neuskutočniteľný alebo neetický.
- Preto sa výskumu podrobuje iba časť základného súboru - výberový súbor (vzorka). Proces overovania správnosti alebo nesprávnosti hypotézy pomocou výsledkov získaných náhodným výberom sa nazýva testovanie štatistických hypotéz.
- Základným predpokladom štatistickej indukcie je náhodný výber.

Postup testovania hypotéz:

- 1) **Formulácia nulovej hypotézy (H_0)**
  Konečným cieľom väčšiny štatistických testov je zhodnotenie vzťahu medzi premennými. Nulová hypotéza potom vyjadruje nezávislosť premenných. Napríklad nulová hypotéza t-testu vyjadruje rovnosť priemerov dvoch základných súborov. Rozdiel zistený vo vzorke sa považuje za náhodný (je dôsledkom náhodného výberu).

- 2) **Formulácia alternatívnej hypotézy (H_a)**
  Väčšinou chceme dokázať pravdivosť alternatívnej hypotézy, ktorá najčastejšie vyjadruje štatistickú závislosť premenných. Pravdivosť alternatívnej hypotézy sa dokazuje vždy iba nepriamo a to tak, že ukážeme, že nulová hypotéza je nepravdepodobná a alternatívna (jediná zostávajúca) je teda pravdepodobná.

- 3) **Stanovenie hladiny významnosti (α)**
  Hladina významnosti je pravdepodobnosť chyby I. druhu, ktorú urobíme ak zamietneme nulovú hypotézu, ktorá v skutočnosti platí. Teda ak prídeme k záveru, že medzi premennými existuje vzťah, pričom medzi nimi vzťah nie je. Alfa sa tradične stanovuje na 5 % (= 0,05) alebo 1 %.

- 4) **Výpočet testovacej štatistiky a pravdepodobnosti**

  Zo vzorky sa vypočíta testovacia štatistika, ktorá má za predpokladu pravdivosti nulovej hypotézy príslušné rozdelenie pravdepodobnosti (Chi-kvadrát, t).

- *P*-hodnota (***P*-Value**, **Probability Level**) predstavuje pravdepodobnosť, že testovacia štatistika za predpokladu pravdivosti nulovej hypotézy dosiahne pri najmenšom tak extrémnu hodnotu ako je hodnota vypočítaná zo vzorky.

- *P*-hodnota je pravdepodobnosť, že vzťah zistený z našich údajov je iba dôsledkom nešťastnej vzorky a ak by sme vybrali ďalšiu náhodnú vzorku, nemuseli by sme nájsť nič.

- *P*-hodnota je najnižšia hodnota hladiny významnosti, ktorá vedie k zamietnutiu nulovej hypotézy.

- *P*-hodnota je odhadovaná pravdepodobnosť zamietnutia pravdivej nulovej hypotézy. Čím menšia je *P*, tým viac sme presvedčení, že nulová hypotéza nie je pravdivá a mala by byť zamietnutá.

5) Rozhodnutie

- Ak $P < \alpha$, nulová hypotéza sa voči príslušnej alternatívnej hypotéze zamietne. Znamená to, že rozdiel nameraný vo vzorke je príliš veľký na to aby bol iba náhodný. Medzi premennými teda existuje vzťah.

- Ak $P \geq \alpha$, nulovú hypotézu nemožno zamietnuť. Znamená to, že rozdiel nameraný vo vzorke môže byť iba náhodný. Často sa v takomto prípade nesprávne hovorí, že nulová hypotéza sa prijíma. Správny je výrok, že nemáme dostatočné dôkazy na to, aby sme nulovú hypotézu zamietli. Teda nemáme dostatok dôkazov na to, aby sme tvrdili, že medzi premennými existuje vzťah.

- V praxi sa veľmi často hladina významnosti nestanovuje vopred, teda *P*-hodnota sa interpretuje samostatne. Väčšina autorov uvádza $P < 0,05$ ako štatisticky významný a $P < 0,01$ ako štatisticky vysoko významný vzťah.

Situácie, ktoré môžu nastať pri testovaní hypotéz

	Rozhodnutie	
Skutočnosť	H_0 nezamietnutá	H_0 zamietnutá
H_0 pravdivá	Správne rozhodnutie ($p = 1-\alpha$)	Chyba I. druhu ($p = \alpha$)
H_0 nepravdivá	Chyba II. druhu ($p = \beta$)	Správne rozhodnutie ($p = 1-\beta$)

- **H0**: Nulová hypotéza

p: Pravdepodobnosť nastatia danej situácie

α: Significance level (hladina významnosti)

1−α: Confidence level (spoľahlivosť)

1−β: Power (sila testu)

Zhluková analýza

- **Zhluková analýza (Cluster analysis) sa zaoberá tým, ako by mali byť objekty (štatistické jednotky) zaradené do skupín tak, aby bola čo najväčšia podobnosť v rámci skupín a čo najväčšia rozdielnosť medzi skupinami.**
- Zhluková analýza sa používa napr. pri segmentácii trhu, pričom klasifikácia spotrebiteľov je založená na kombinácii viacerých premenných. Premennými, teda segmentačnými kritériami môžu byť: pohlavie, vek, vzdelanie, životný štýl, náboženstvo, skúsenosti s produktom, veľkosť spotreby, frekvencia spotreby a pod.

- Pri zohľadnení iba jednej premennej (1-D) je nájdenie zhlukov veľmi jednoduché: hodnoty premennej sa nanesú na číselnú os a zhluky sa identifikujú vizuálne (napr. podľa veku nájdeme v súbore 2 skupiny respondentov: jednu okolo 15 rokov a druhú okolo 40 rokov).
- Podobne použitím X-Y grafu možno jednoducho identifikovať zhluky pri zohľadnení 2 premenných (2-D). V priestore (3-D) sa pomocou interaktívneho X-Y-Z grafu tiež dajú nájsť zhluky vizuálne. Vizuálne identifikovať zhluky pri zohľadnení viac ako 3 premenných súčasne sa však už nedá. Práve vtedy sa používa zhluková analýza.

Zhluková analýza

- Zhluková analýza zahŕňa množstvo metód. Rozlišujú sa dve základné skupiny:
  1. Hierarchické zhlukovacie metódy
  2. Nehierarchické zhlukovacie metódy
- ***Hierarchické zhlukovacie metódy*** vychádzajú z jednotlivých objektov, ktoré reprezentujú zhluky. Ich spájaním sa v každom kroku počet zhlukov postupne zmenšuje až sa nakoniec všetky zhluky spoja do jedného celku.
- Hierarchické metódy vedú k hierarchickej (stromovej) štruktúre, ktorá sa graficky zobrazuje ako stromový diagram (dendrogram). Stromové zhlukovacie metódy začínajú výpočtom vzdialenosti medzi objektmi.
- **Euklidovská vzdialenosť** medzi objektmi *i* a *j* s *n* charakteristikami (premennými) sa vypočíta:

- Alternatívnu vzdialenosť prestavuje **vzdialenosť Manhattan (City-block)**:
- Euklidovská vzdialenosť vyjadruje vzdušnú vzdialenosť medzi dvoma objektmi a vzdialenosť Manhattan najkratšiu vzdialenosť, ktorú musí chodec prejsť aby sa v meste dostal z jedného miesta na druhé. Výhoda vzdialenosti Manhattan spočíva v znížení dopadu extrémnych prípadov (outliers) na výsledky.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

$$d_{ij} = \sum_{k=1}^{n} | x_{ik} - x_{jk} |$$

- Euklidovská vzdialenosť vyjadruje vzdušnú vzdialenosť medzi dvoma objektmi a vzdialenosť Manhattan najkratšiu vzdialenosť, ktorú musí chodec prejsť aby sa v meste dostal z jedného miesta na druhé. Výhoda vzdialenosti Manhattan spočíva v znížení dopadu extrémnych prípadov (outliers) na výsledky.
- Existujú ešte viaceré iné typy vzdialeností, ktoré sa používajú napr. pri kategorických premenných.

- Keď už máme vypočítané vzdialenosti medzi všetkými dvojicami objektov musíme určiť pravidlo podľa ktorého sa budú objekty spájať do zhlukov, teda ako sa bude určovať vzdialenosť medzi zhlukmi. Existujú viaceré **pravidlá spájania**:
- **Single linkage (Nearest Neighbour) – jednoduché spájanie (najbližší sused)**
  Vzdialenosť medzi dvoma zhlukmi je definovaná ako vzdialenosť dvoch najbližších členov.
- **Complete linkage (Furthest Neighbour) – kompletné spájanie (najvzdialenejší sused)**
  Vzdialenosť medzi dvoma zhlukmi je definovaná ako vzdialenosť dvoch najvzdialenejších členov.
- **Unweighted pair-group average (Group Average) – nevážený párový priemer (priemer skupín)**
  Vzdialenosť medzi zhlukmi je definovaná ako priemerná vzdialenosť medzi všetkými pármi, pričom 1.člen je z 1.zhluku a 2.člen z 2.zhluku.

- **Weighted pair-group average (Simple Average)** – **vážený párový priemer (jednoduchý priemer)**
  Podobná ako predošlá z tým rozdielom, že veľkosti zhlukov (počty objektov) sa berú ako váhy.
- **Unweighted pair-group centroid (Centroid)** – **nevážený centroid (centroid)**
  Vzdialenosť medzi dvoma zhlukmi je definovaná ako vzdialenosť centroidov týchto dvoch zhlukov. Centroid je vektor priemerov (každá súradnica je priemer príslušných súradníc objektov v zhluku).
- **Weighted pair-group centroid (Median)** – **vážený centroid (medián)**
  Podobná ako predošlá z tým rozdielom, že veľkosti zhlukov (počty objektov) sa berú ako váhy.
- **Wardova metóda**
  Táto metóda sa zreteľne odlišuje od všetkých ostatných pretože na určenie vzdialenosti medzi zhlukmi využíva prístup analýzy rozptylu. S touto metódou sa zhluky vytvárajú tak, aby sa vnútrozhlukový súčet štvorcov minimalizoval.

Nehierarchické zhlukovacie metódy

- nevytvárajú stromovú štruktúru. Najznámejšia nehierarchická zhlukovacia metóda je **metóda k-priemerov (k-means)**.
- Táto metóda sa vyznačuje tým, že vyprodukuje presne k-zhlukov tak, aby bol vnútroskupinový súčet štvorcov minimálny. Najvhodnejšia je na formovanie malého počtu zhlukov z veľkého počtu pozorovaní. Vyžaduje však intervalové premenné bez extrémnych hodnôt (outliers).
- Nominálne premenné sa dajú použiť ale môžu spôsobovať problémy. Užitočnou metódou je **neurčité zhlukovanie (Fuzzy clustering)**, ktoré na rozdiel od ostatných zhlukovacích metód, umožňuje čiastočné zaradenie objektu do viacerých zhlukov a to pomocou pravdepodobnosti.
- Cieľom je zabrániť skresleniu zhlukovania kvôli prítomnosti nezaraditeľných objektov. Takéto indivíduum sa nepriradí ku žiadnemu zhluku (od každého sa príliš odlišuje), ale priradia sa mu pravdepodobnosti s ktorými sa bude nachádzať v jednotlivých zhlukoch.
- Metóda sa často používa pri odhaľovaní podvodov v rôznych oblastiach. Napr. v bankovníctve sa bez vopred formulovanej definície podozrivej operácie z miliónov operácií klientov identifikuje pár desiatok takých, ktoré sa od zvyšných (zoskupených do niekoľkých zhlukov) pri použití viacerých premenných (napr. obrat, typ operácie, konštantný symbol, čas od zadania po jej splatnosť atď.) výrazne odlišujú.

Jednofaktorová analýza rozptylu

- Jednosmerná analýza rozptylu (One-Way ANOVA) je najjednoduchšou formou ANOVA (ANalysis Of VAriance). Jednosmerná (jednoduchá) ANOVA skúma vzťah medzi intervalovou a nominálnou premennou (faktorom), napr. množstvom cholesterolu v krvi a typom diéty alebo predajnosťou výrobku a druhom obalu.

- Cieľom analýzy rozptylu je odhaliť, či vo vzorke zistené rozdiely priemerov jednotlivých skupín (podľa úrovne faktora) sú štatisticky významné (medzi premennými je vzťah) alebo môžu byť iba náhodné (medzi premennými nie je vzťah).

- Overuje sa to tak, že sa celková variabilita (suma štvorcov odchýlok hodnôt premennej od jej priemeru) rozdelí na vnútroskupinovú (náhodná chyba) a medziskupinovú (daná rozdielom priemerov skupín). F-štatistika sa vypočíta ako pomer medziskupinovej a vnútroskupinovej variability a použije sa na testovanie nulovej štatistickej hypotézy o rovnosti priemerov.

- Ak je *P*-hodnota nižšia ako zvolená hladina významnosti (tradične 5 % = 0,05), nulová hypotéza sa zamietne. Znamená to, že rozdiel medzi aspoň jednou dvojicou priemerov vypočítaných zo vzorky je príliš veľký na to, aby mohol byť iba dôsledkom náhodného výberu, je teda štatisticky významný – medzi premennými je vzťah.

- Ak je *P*-hodnota rovná alebo vyššia ako zvolená hladina významnosti, nulovú hypotézu nemožno zamietnuť. Znamená to, že rozdiel medzi každou dvojicou priemerov vypočítaných zo vzorky môže byť iba dôsledkom náhodného výberu, nie je teda štatisticky významný – medzi premennými nie je vzťah.

- Jednofaktorová analýza rozptylu predstavuje rozšírenie dvojvzorkového t-testu. V prípade porovnávania dvoch skupín (faktor má iba dve úrovne) ANOVA vedie k rovnakému výsledku ako t-test.

- ANOVA sa typicky zameriava na testovanie významnosti nie sily asociácie. Pritom v prípade veľkých vzoriek sa môže stať, že priemery skupín sa významne líšia, ale tieto rozdiely sú malé.

- Preto pri použití ANOVA treba uvádzať aj silu asociácie pre významné efekty. Veľkosť efektu, ktorý má faktor na závislú premennú sa meria pomocou Eta2 a Omega2.

- **Eta2 ($\eta 2$)** sa vypočíta ako podiel medziskupinovej a celkovej sumy štvorcov a vyjadruje podiel celkovej variability, ktorá sa prisudzuje faktoru. Eta2 je obdobou koeficientu determinácie R2, ktorý sa používa v regresnej analýze. Nevýhodou Eta2 je skreslenosť odhadu efektu v populácii - efekt systematicky nadhodnocuje.

- **Omega2 ($\omega 2$)** predstavuje alternatívu mieru veľkosti efektu k Eta2 s rovnakou interpretáciou poskytujúcou neskreslené odhady efektu. Eta2 predstavuje stupeň asociácie medzi efektom a závislou premennou vo vzorke a Omega2 odhad stupňa asociácie v základnom súbore.

Sila testu a veľkosti vzoriek

- Sila testu (pravdepodobnosť zachytenia existujúceho významného rozdielu) závisí od:

1. Variability
2. Veľkosti vzorky
3. Pravdepodobnosti chyby I. druhu (α)
4. Veľkosti efektu

- Čím je vyššia variabilita hodnôt premennej, tým je nižšia sila testu. Zvýšenie zvyšných troch faktorov zvyšuje silu testu. Veľkosť efektu je veľkosť rozdielu parametrov (napr. rozdiel priemerov pri t-teste), ktorý možno zachytiť napr. experimentom. Na zachytenie veľkého efektu stačí menšia vzorka ako na zachytenie malého efektu. Voliť sa má vždy taká veľkosť efektu, ktorá je pre daný výskum užitočná.
- Vzťahmi medzi variabilitou, veľkosťou vzorky, alfou, veľkosťou efektu a silou testu sa zaoberá **analýza sily (Power Analysis)**. Analýzou sily sa treba zaoberať už pri plánovaní výskumu. Nedostatočná veľkosť vzorky, môže spôsobiť nezachytenie relevantného efektu. Príliš veľká vzorka stojí zbytočne veľa času a peňazí s minimálnym úžitkom. Analýza sily je značne komplikovaná, preto sa veľmi často vôbec nerobí. Na analýzu sily je nutné použiť kvalitný software.

Nevýhody testovania hypotéz

- Výsledok testovania hypotéz je rozhodnutie o tom, či zamietnuť alebo nezamietnuť nulovú hypotézu. Veľmi často je takýto výsledok nepostačujúci - napr. v prípade testovania efektívnosti novej liečby. Výskumník sa zaujíma o silu efektu nie o to, či sa efekt rovná presne 0. Porovnávanie *P*-hodnôt (aj v rámci jednej štúdie) bez doplňujúcich informácií a následným vyvodením záverov nemusí byť správne. Ak napríklad vo viacfaktorovej ANOVA faktor A má *P*=0,0001 a faktor B *P*=0,049, nemôžeme jednoducho povedať, že faktor A má silnejší efekt ako faktor B. Samotná *P*-hodnota 0,001 môže v skutočnosti znamenať 3 situácie:

- 1) triviálny (z praktického hľadiska nevýznamný) efekt v základnom súbore zistený z veľkej vzorky

- 2) silný efekt v základnom súbore zistený zo stredne veľkej vzorky

- 3) obrovský efekt v základnom súbore zistený z malej vzorky.

Výhody intervalových odhadov

- Intervalové odhady odpovedajú na otázku v akých hraniciach možno očakávať skutočný efekt v základnom súbore. Poskytujú teda viac informácií ako testy hypotéz. V prípade, že chceme zistiť, či je liek proti vysokému krvnému tlaku účinný, môžeme použiť párový t-test.
- Vzorke pacientov zmeriame tlak pred a po podávaní lieku. Ak sa priemerný rozdiel tlakov významne odlišuje od 0, potom má liek účinok. Silu účinku však možno určiť len intervalovým odhadom priemerného rozdielu. Intervalový odhad nám s danou spoľahlivosťou (pravdepodobnosťou v %) povie, aký pokles tlaku môžeme očakávať v základnom súbore tvorenom pacientmi s vysokým krvným tlakom. Veľkou výhodou intervalových odhadov je ich vypovedacia schopnosť. Z intervalu, ktorý je príliš široký (vykazuje veľkú chybu) jasne vidno, že veľkosť vzorky je nedostatočná. Naopak, z intervalu ktorý je úzky, pričom vyjadruje triviálny efekt vidno, že štatistická významnosť je dosiahnutá veľkou vzorkou (teda príliš veľkou silou testu).
- Jediný intervalový odhad poskytuje dostatok informácií na priame uskutočnenie teoreticky nekonečného množstva testov hypotéz. Ak je 95%-ný interval spoľahlivosti rozdielu dvoch priemerov od 10 do 15, znamená to zamietnutie nulovej hypotézy (na 5%-nej hladine významnosti), že rozdiel priemerov dvoch základných súborov sa rovná 0 (pretože 0 sa nachádza mimo intervalu od 10 do 15.) Pre ten istý interval však s 95%-nou spoľahlivosťou nemožno zamietnuť hypotézu, že rozdiel priemerov základných súborov sa rovná 12 (lebo 12 patrí do intervalu od 10 do 15).

Dvojrozmerná induktívna štatistika - poradové premenné

- **Neparametrické korelačné koeficienty**
- Korelačný koeficient meria silu štatistickej závislosti medzi dvoma číselnými premennými. Hodnoty všetkých korelačných koeficientov sa nachádzajú v intervale od −1 do 1. Hodnoty blízko 0 znamenajú žiadny vzťah a absolútne hodnoty blízko 1 silný vzťah. Kladné hodnoty znamenajú, že premenné majú tendenciu meniť sa rovnakým smerom, záporné hodnoty rôznym smerom.

- **Interpretácia** korelačného koeficientu závisí od kontextu. Napr. hodnota 0,8 pri overení fyzikálneho zákona použitím presných meracích prístrojov je veľmi nízka, v sociálnych vedách je však veľmi vysoká. Cohen (1988) vytvoril jednoduchú pomôcku pre interpretáciu korelačných koeficientov v psychologickom výskume: Korelácia pod 0,1 je triviálna, 0,1–0,3 malá, 0,3–0,5 stredná a nad 0,5 veľká. Počítať a interpretovať korelačné koeficienty treba vždy až po prezretí X-Y grafu.

Dvojrozmerná induktívna štatistika - poradové premenné

- **Kendallov poradový korelačný koeficient** (1948) meria silu závislosti medzi dvoma poradovými premennými a poskytuje neparametrický test nezávislosti (test významnosti koeficientu).

- **Kendalovo tau** vyjadruje rozdiel medzi pravdepodobnosťou, že hodnoty dvoch premenných sú v rovnakom poradí oproti pravdepodobnosti, že hodnoty nie sú v rovnakom poradí. V prípade väčšieho výskytu nerozhodných párov sa použije **tau-b**. V prípade absencie nerozhodných párov sa tau-b rovné tau. Ak jedna premenná nadobúda odlišný počet unikátnych hodnôt ako druhá (kontingenčná tabuľka *m*×*n*), treba uprednostniť **tau-c**, ktoré sa tiež nazýva **Stuartovo tau-c** alebo **Kendall-Stuartovo tau-c**:

$$\tau_c = \frac{n_c - n_d}{n^2 (k-1)/2k}$$

- **Kendallovo tau** vyjadruje rozdiel medzi pravdepodobnosťou, že hodnoty dvoch premenných sú v rovnakom poradí oproti pravdepodobnosti, že hodnoty nie sú v rovnakom poradí. V prípade väčšieho výskytu nerozhodných párov sa použije **tau-b**. V prípade absencie nerozhodných párov sa tau-b rovné tau. Ak jedna premenná nadobúda odlišný počet unikátnych hodnôt ako druhá (kontingenčná tabuľka *m*×*n*), treba uprednostniť **tau-c**, ktoré sa tiež nazýva **Stuartovo tau-c** alebo **Kendall-Stuartovo tau-c**:

$$\tau_c = \frac{n_c - n_d}{n^2 (k-1)/2k}$$

Analýza hlavných komponentov

- Analýza hlavných komponentov **(Principal Components Analysis - vytvorená v roku 1901 Pearsonom)** je analytický nástroj, ktorý sa zvyčajne používa na redukciu rozmernosti (počtu premenných) veľkého počtu vzájomne súvisiacich premenných na hlavné komponenty, pri čo najmenšej strate informácií (variability).

- PCA vypočíta súbor vzájomne nezávislých premenných (hlavných komponentov), ktoré sú lineárnou kombináciou (váženým priemerom) originálnych premenných.

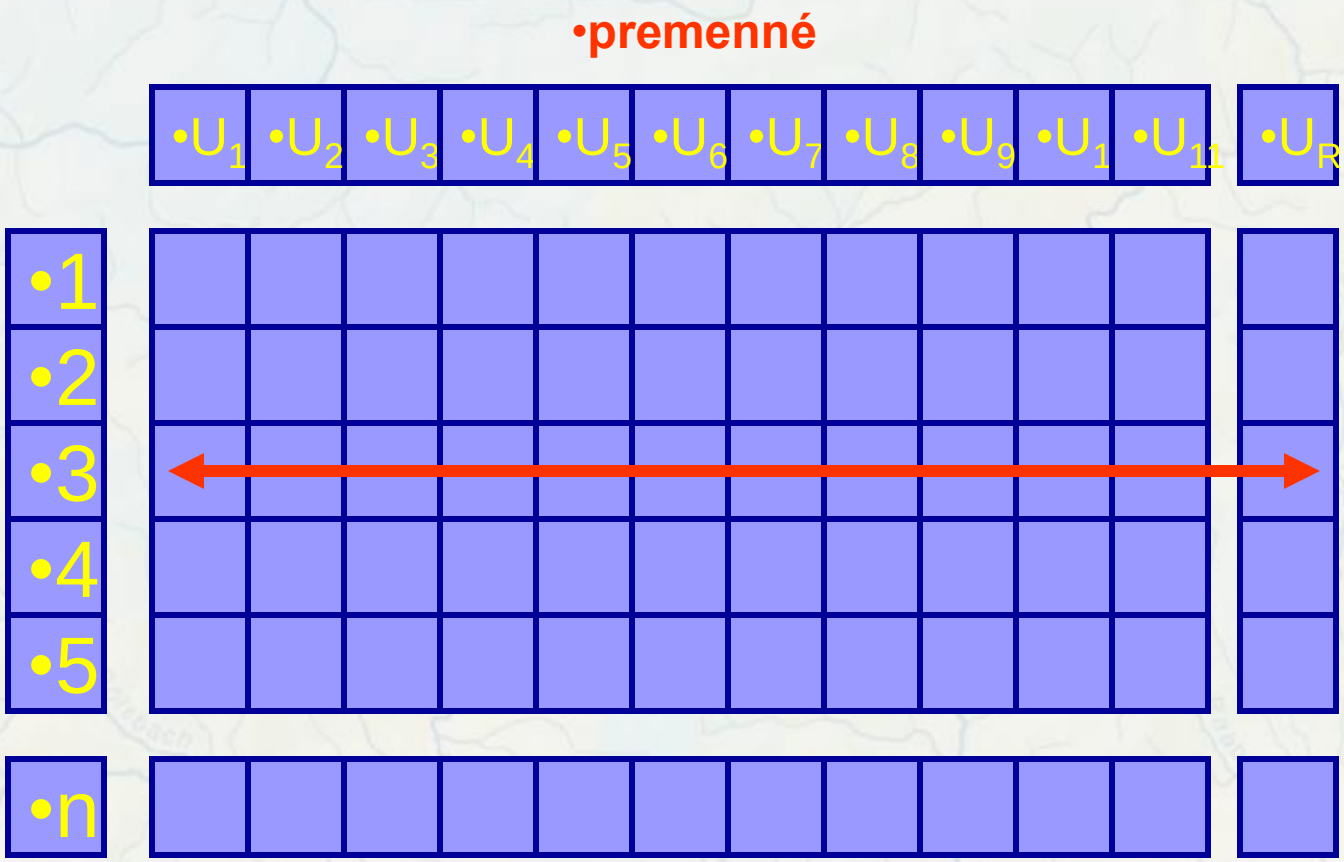
- Prvý hlavný komponent vysvetľuje najväčšiu časť variability premenných, druhý komponent vysvetľuje druhú najväčšiu časť variability, atď. až kým je vysvetlená všetka variabilita. Komponenty sú vzájomne nezávislé a niekoľko z nich často vysvetľuje okolo 80 % variability. Tieto sa potom skúmajú, graficky znázornia, prípadne použijú ako vstupy do lineárnej regresie, diskriminačnej analýzy alebo zhlukovej analýzy. PCA na rozdiel od príbuznej faktorovej analýzy (FA) prinesie vždy rovnaké výsledky.

Analýza hlavných komponentov

- FA aj PCA sa snažia zredukovať rozmernosť skupiny údajov. Hlavný rozdiel medzi FA a PCA je ten, že PCA vysvetľuje všetku variabilitu medzi originálnymi premennými (vyjadrenú v korelačnej matici) a FA iba variabilitu, ktorú majú premenné spoločnú.
- Cieľom PCA (Rao 1964) je odvodenie malého množstva lineárnych kombinácií (hlavných komponentov) z množiny premenných pri zachovaní čo najviac informácií obsiahnutých v pôvodných premenných. Cieľom FA (Mulaik 1972) je vysvetliť korelácie alebo kovariancie medzi premennými pomocou malého množstva nepozorovateľných, latentných premenných. Latentné premenné nemožno všeobecne vypočítať ako lineárnu kombináciu originálnych premenných. FA predpokladá lineárne vzťahy medzi premennými nebyť nekorelovanej náhodnej chyby (unikátnej variability) v každej premennej, pričom lineárne vzťahy aj množstvo unikátnej variability možno odhadnúť.

Faktorová analýza
(FA)

Viacrozmerné metódy



- **Metódy analýzy skrytých vzťahov**

Viacrozmerné metódy

- Metódy analýzy skrytých vzťahov
  - premenné nemožno logicky rozdeliť do dvoch skupín na závislé a nezávislé
  - cieľom je pochopiť alebo identifikovať prečo a ako sú premenné navzájom korelované t.j. ako sa navzájom ovplyvňujú
  - ak sú premenné navzájom prepojené – korelované, možno rovnaký objem informácií vystihnúť menším počtom premenných – zníženie dimenzie

Viacrozmerné metódy

Počet premenných	Typ údajov	
	Kvantitatívne	Kvalitatívne
Dve	Jednoduchá korelácia	Analýza dvojrozmerných kontingenčných tabuliek
		Loglineárne modely
Viac ako dve	Analýza hlavných komponentov	Analýza viacrozmerných kontingenčných tabuliek
		Loglineárne modely
	Faktorová analýza	Korešpondenčná analýza

- **Metódy analýzy skrytých vzťahov**

Faktorová analýza

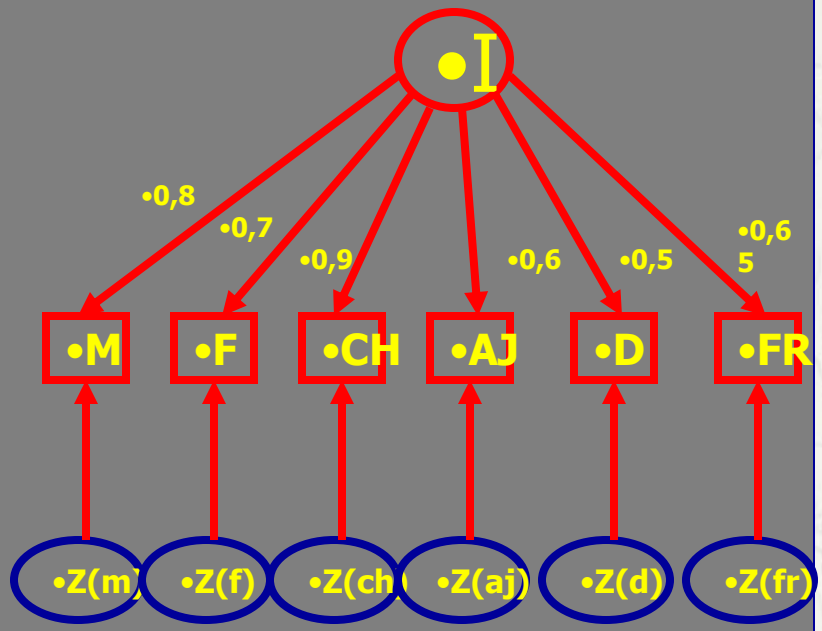
- Charakteristika
  - predmetom analýzy je skupina kvantitatívnych premenných
  - merateľné veličiny môžeme vyjadriť ako lineárne funkcie menšieho počtu skrytých – spoločných faktorov a jedného špecifického faktora

Faktorová analýza

- Charakteristika
  - k dispozícii máme výsledky testov študentov z rôznych predmetov
    - matematika (M)
    - fyzika (F)
    - chémia (CH)
    - anglický jazyk (AJ)
    - dejepis (D)
    - francúzština (FR)
  - môžeme predpokladať, že výsledky testu sú funkciou:
    - všeobecnej inteligencie študenta (I)
    - jeho záujmu o daný predmet (Z)

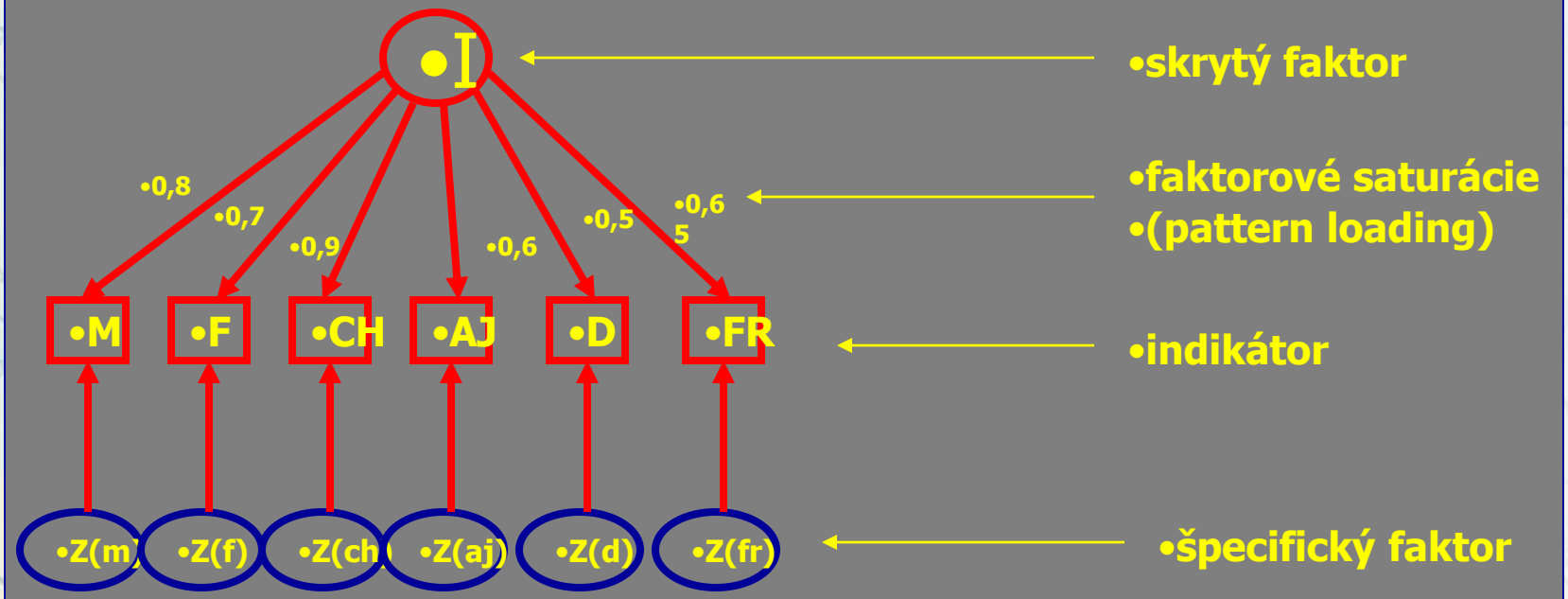
Faktorová analýza

- Charakteristika
  - na základe uvedených predpokladov platí napr.:
    - M = 0,8 I + Z(m)
    - F = 0,7 I + Z(f)
    - CH = 0,9 I + Z(ch)
    - AJ = 0,6 I + Z(aj)
    - D = 0,5 I + Z(d)
    - FR = 0,65 I + Z(fr)



Faktorová analýza

- Charakteristika



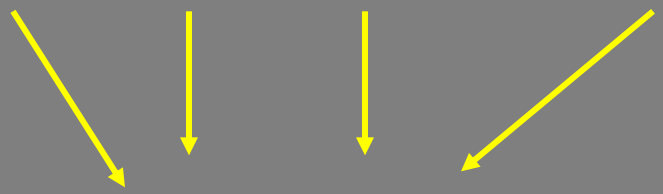
Faktorová analýza

- Princípy
  - indikátory sú navzájom korelované, pretože zdieľajú minimálne jeden spoločný znak
    - ktorý je zodpovedný za koreláciu medzi indikátormi
    - nemôže byť priamo zmeraný
    - pôsobí minimálne na dva indikátory súčasne
      - sa nazýva spoločný alebo skrytý faktor
  - variabilita indikátorov nevysvetlená skrytým faktorom je spôsobená špecifickými vplyvmi
      - tzv. špecifickými faktormi resp. náhodnou chybou

Faktorová analýza

- Princípy
  - každý indikátor možno vyjadriť ako

- $X_1 = a_{11} f_1 + a_{12} f_2 + a_{13} f_3 + \ldots + a_{1q} f_q + e_1$
- $X_2 = a_{21} f_1 + a_{22} f_2 + a_{23} f_3 + \ldots + a_{2q} f_q + e_2$
- $X_3 = a_{31} f_1 + a_{32} f_2 + a_{33} f_3 + \ldots + a_{3q} f_q + e_3$
- ….
- $X_k = a_{k1} f_1 + a_{k2} f_2 + a_{k3} f_3 + \ldots + a_{kq} f_q + e_k$



- saturácia, váha

Faktorová analýza

- Princípy
  - cieľom je teda odhadnúť model, ktorý je podobný všeobecnému lineárnemu modelu
    - avšak pri lineárnom modely poznáme X aj Y, čo nám umožňuje nájsť jedinečné riešenie pre β a E
      - **$X = \alpha F + E$**
      - **$Y = \beta X + E$**
  - pri FA máme len X, z ktorých vychádzame pri hľadaní riešenia pre F, α a E
    - pre FA tak možno určiť nekonečné množstvo riešení
    - každé z nájdených riešení bude odhadovať údaje rovnako kvalitne

Faktorová analýza

- Princípy
  - odhad vychádza z rozkladu variability
    - celkovú variabilitu každého indikátora možno rozložiť na dve zložky

$$D(X_j) = s_j^2 = (a_{j1}^2 + a_{j2}^2 + \ldots + a_{jq}^2) + u_j^2$$

$$D(X_j) = s_j^2 = h_j^2 + u_j^2$$

$$D(X_j) = s_j^2 = \text{komunalita} + \text{unicita}$$

- komunalita – časť rozptylu indikátora, ktorú je možné vysvetliť pôsobením skrytých faktorov
- unicita – časť rozptylu indikátora, ktorú možno vysvetliť len pôsobením špecifických faktorov alebo náhody

Faktorová analýza

- Princípy
  - ak poznáme odhady rozptylov, môžeme odhadnúť saturácie
  - východiskom je korelačná matica indikátorov
    - **$R = R_h + E$**
    - R_h – redukovaná korelačná matica
      - diagonála obsahuje odhady komunalít
      - mimo diagonály sú koeficienty korelácie
    - E – reziduálna korelačná matica
      - na diagonále sú rozptyly špecifických faktorov

Faktorová analýza

- Princípy
  - predpoklady
    - R je korelačná matica indikátorov s viacerými štatisticky významnými koeficientmi korelácie
    - spoločné faktory sú navzájom nekorelované
    - špecifické faktory sú navzájom nekorelované
    - spoločné a špecifické faktory sú navzájom nekorelované

Faktorová analýza

- Postup
  - inicializačný odhad komunalít
  - extrakcia spoločných faktorov
  - určenie počtu spoločných faktorov
  - rotácia faktorov
  - odhad faktorových saturácií, komunalít, unicít
  - interpretácia spoločných faktorov
  - odhad faktorových skóre

Faktorová analýza

- Postup
  - inicializačný odhad komunalít
    - najvyšší korelačný koeficient danej premennej s ostatnými premennými
    - štvorec viacnásobného koeficienta determinácie
    - priemerný korelačný koeficient
    - najvyššia korelácia – pomer štvorca j-teho stĺpcového súčtu k celkovej sume štvorcov všetkých koeficientov
    - iteratívny odhad faktorov

Faktorová analýza

- Postup
  - extrakcia spoločných faktorov
    - metóda HK (principal components factoring)
      - inicializačné komunality = 1
      - korelačná matica s komunalitami je vstupom pre klasickú PCA
    - metóda hlavných osí (principal axis factoring)
      - iteratívny odhad inicializačných komunalít
      - PCA, kým zmena komunality nie je menšia ako stanovené kritérium
    - metóda maximálnej vierohodnosti
    - image factor analysis
    - alpha factor analysis

Faktorová analýza

- Postup
  - určenie počtu spoločných faktorov
    - analýza scree grafu – podiel komunality
    - vlastné číslo > 1
    - Bartletov test:
      - H_o : posledných k-q faktorov nie je štat. významných
      - H_1 : neplatí H_o

Faktorová analýza

- Postup
  - rotácia faktorov
    - cieľom je získať lepšie interpretovateľný odhad faktorov
    - typy
      - ortogonálna (nekorelované)
        - » **VARIMAX**
        - » **EQUAMAX**
        - » **QUARTIMAX**
        - » **PARSIMAX**
      - šikmá (korelované)
        - » **PROCRUSTES**
        - » **PROMAX**

Faktorová analýza

- Postup
  - odhad faktorových saturácií, komunalít, unicít
  - interpretácia spoločných faktorov
    - vychádza vo všeobecnosti z dvoch matíc
      - matica faktorových saturácií (factor pattern matrix)
        - » koeficienty pre výpočet indikátorov zo skrytých faktorov
      - matica faktorovej štruktúry (factor structure matrix)
        - » koeficienty korelácie medzi faktormi a indikátormi
  - pre ortogonálne rotácie sú obe matice zhodné tzv. factor loading matica

- **Doštudovať:**

1. Ďalšie hydrologické veličiny
2. Teplotný a ľadový režim tokov
3. Klasifikácia tokov podľa režimu odtoku (Ľvovič, Pardé) a Dub feat. Zaťko, Šimo

- **Skriptá, s. 98 – 109.**