CENTRUM BIOSTATISTIKY A ANALÝZ

Lékařská a Přírodovědecká fakulta, Masarykova univerzita

Biostatistika

– základní kurz –

http://www.cba.muni.cz ➲ http://www.cba.muni.cz/vyuka

BioStatistika na Přf a LF MU

Centrum biostatistiky a analýz

- Pracoviště specializované na analýzu biologických dat
- Výuka analýzy biologických dat – řada kurzů

Kontakt:

Ladislav Dušek – dusek@cba.muni.cz

Jiří Jarkovský – jarkovsky@cba.muni.cz

Kurz biostatistiky

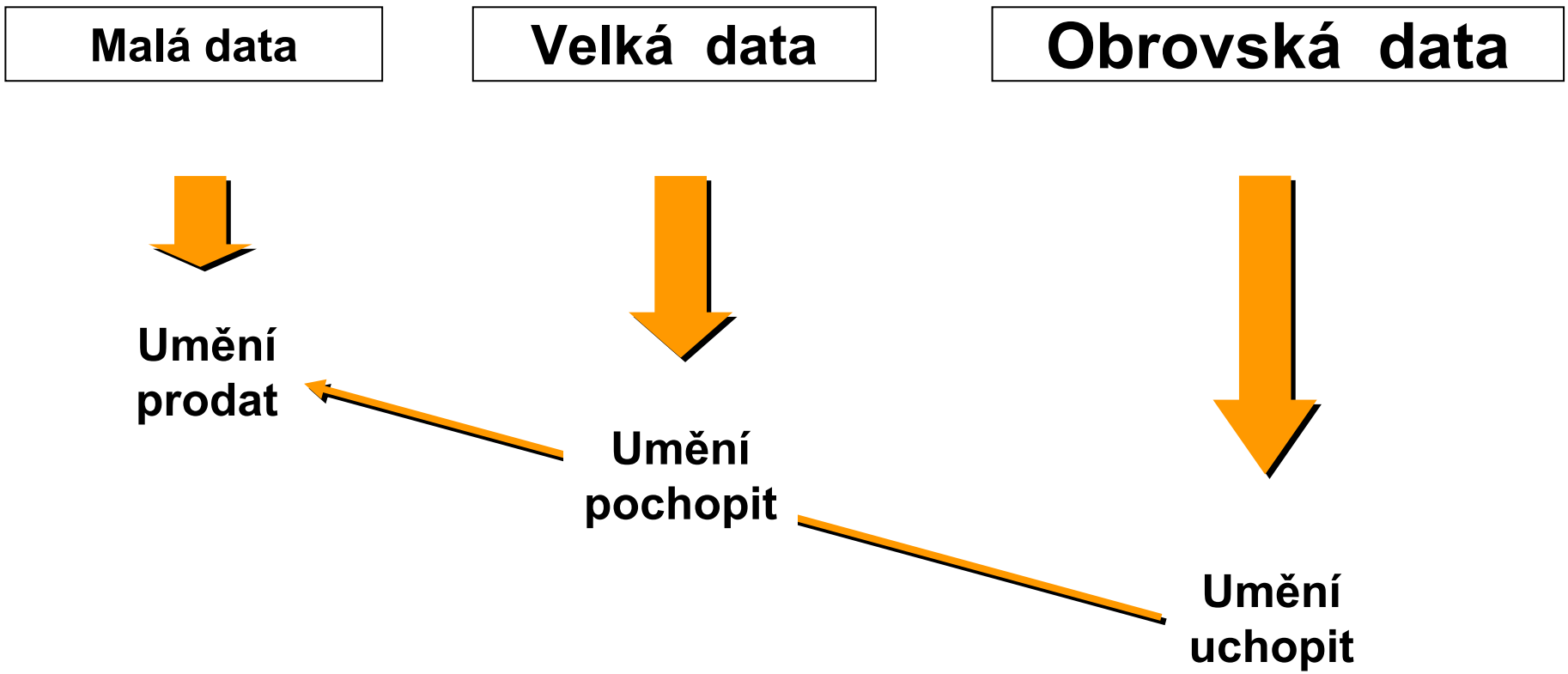
- Sada přednášek pokrývajících základní oblasti analýzy biologických dat, zejména z praktického hlediska:
  - Způsoby ukládání dat, typy dat a jejich statistický popis
  - Hypotézy o datech a jejich testování
  - Vztahy proměnných a jejich statistické hodnocení
  - Predikce a příčinné vztahy proměnných
  - Grafické zobrazení dat a výsledků analýz
  - Příklady aplikace na reálných datech
  - Přehled základních statistických SW
  - Složitější metody statistické analýzy – přehled metod

- Cyklus přednášek ukončen zkouškou
  - Písemná zkouška – příklady
  - Zaměřeno na postup řešení, číselný výsledek méně významný

1. Data a informace

BIOSTATISTIKA - BIOMETRIKA

Věda zabývající se hodnocením **biologických dat** = záznamů o biologických systémech a jejich chování



DATA – ukázka uspořádání datového souboru

Parametry (znaky)

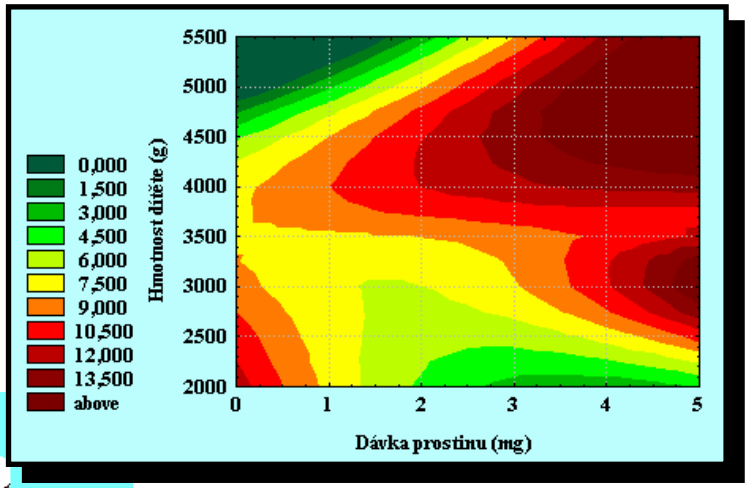
Opakování

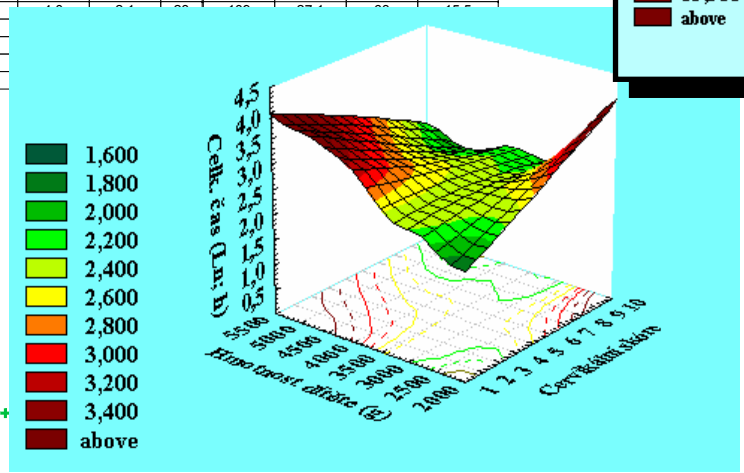
Pacient	Clovek	aLeu	aTy%	aSe%	aNeu%	aLy%	aTy	aSe	aNeu	aLy	aHtc	aCLsk	aCLNeus	aCLOZ	aCLNeuO
		cell.10^6/	%	%	%	%	cell.10^6/	cell.10^6/	cell.10^6/	cell.10^6/	%	mV.s.10^3	mV.s.10^3	mV.s.10^3	mV.s.10^3
3	1	4									33	72		32	
4	2	7,6	8	58	66	24	0,6	4,4	5,0	1,8	33	95	19	48	10
8	3	4	3	52	55	40	0,1	2,1	2,2	1,6	22	77	35	33	15
11	4	6,1	5	59	64	35	0,3	3,6	3,9	2,1	33	103	26	49	13
12	5	6,9	3	85	88	9	0,2	5,9	6,1	0,6	37	81	13	45	7
14	6	5,9	15	55	70	19	0,9	3,3	4,1	1,1	32	137	33	61	15
16	7	8	18	75	93	7	1,4	6,0	7,4	0,6	34	151	20	59	8
20	8	9,6	3	72	75	23	0,3	6,9	7,2	2,2	40	77	11	38	5
21	9	6	10	67	77	19	0,6	4,0	4,6	1,1	32	120	26	52	11
22	10	3,3	4	55	59	39	0,1	1,8	2,0	1,3	28	81	42	24	12
37	11	3,8	10	60	70	30	0,4	2,3	2,7	1,1	32	111	42	29	11
38	12	6,4	2	76	78	17	0,1	4,9	5,0	1,1	25	366	73	115	23
39	13	6,8	1	57	58	39	0,1	3,9	3,9	2,7	20	234	59	71	18
49	14	8,5	7	67	74	26	0,6	5,7	6,3	2,2	30	156	25	108	17
51	15	9,3	7	57	64	35	0,7	5,3	6,0	3,3	35	129	21	23	4
52	16	2,2	10	56	66	34	0,2	1,2	1,5	0,7	33	46	30	12	8
55	17	9,9	3	78	81	10	0,3	7,7	8,0	0,1	30	189	24	140	18
56	18	5	2	80	82	13	0,1	4,0	4,1	0,7	26	101	25	54	13
6	1	8,8	11	72	83	12	1,0	6,3	7,3	1,1	44	268	36,6	145	19,9
9	2	9,2	2	66	68	28	0,2	6,1	6,3	2,6	42	168	26,9	76	12,2
13	3	10,0	7	83	90	8	0,7	8,3	9,0	0,8	54	181	20,1	81	9
15	4	9,6	1	75	76	23	0,1	7,2	7,3	2,2	45	343	47	124	16,9
17	5	6,0									45	40		21	
19	6	7,2	2	78	80	18	0,1	5,6	5,8	1,3	44	103	17,8	63	10,9
24	7	8,2	1	72	73	25	0,1	5,9	6,0	2,1	41	209	34,9	57	9,6
26	8	10,3	1	85	86	3	0,1	8,8	8,9	0,3	41	364	41,1	112	12,6
29	9	5,0	1	74	75	21	0,1	3,7	3,8	1,1	39	83	22,1	32	8,5
30	10	11,9	1	51	52	47	0,1	6,1	6,2	5,6	33	83	13,4	52	8,4
31	11	7,2	3	53	56	29	0,2	3,8	4,0	2,1	28	109	27,1	63	15,5
32	12	10,8	36	50	76	8	3,9	5,4	9,3	0,9	27	146	15,7	106	11,4
33	13	11,8	22	54	76	16	2,6	6,4	9,0	1,9	45	246	27,4	63	7
34	14	17,0	1	82	83	16	0,2	13,9	14,1	2,7	34	440	31,2	119	8,4
40	15	10,0	8	72	80	4	0,8	7,2	8,0	0,4	37	176	22,0	52	6,5

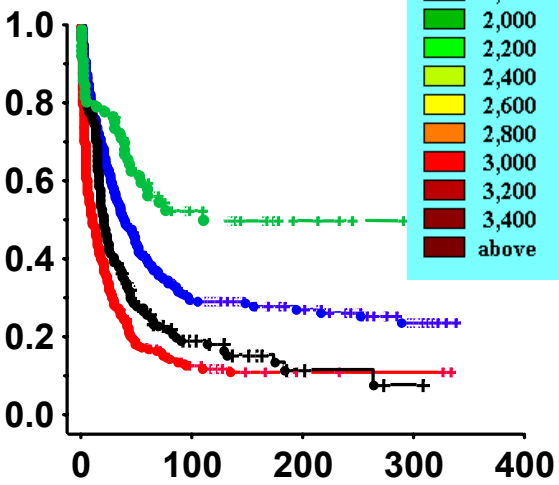
BIOSTATISTIKA - BIOMETRIKA

Pacient	Clovek	aLeu	aTy%	aSe%	aNeu%	aLy%	aTy	aSe	aNeu	aLy	aHtc	aCLsk	aCLNeus	aCLOZ	aCLNeuO
		cell.10^6/	%	%	%	%	cell.10^6/	cell.10^6/	cell.10^6/	cell.10^6/	%	mV.s.10^3	mV.s.10^3	mV.s.10^3	mV.s.10^3
3	1	4									33	72		32	
4	2	7,6	8	58	66	24	0,6	4,4	5,0	1,8	33	95	19	48	10
8	3	4	3	52	55	40	0,1	2,1	2,2	1,6	22	77	35	33	15
11	4	6,1	5	59	64	35	0,3	3,6	3,9	2,1	33	103	26	49	13
12	5	6,9	3	85	88	9	0,2	5,9	6,1	0,6	37	81	13	45	7
14	6	5,9	15	55	70	19	0,9	3,3	4,1	1,1	32	137	33	61	15
16	7	8	18	75	93	7	1,4	6,0	7,4	0,6	34	151	20	59	8
20	8	9,6	3	72	75	23	0,3	6,9	7,2	2,2	40	77	11	38	5
21	9	6	10	67	77	19	0,6	4,0	4,6	1,1	32	120	26	52	11
22	10	3,3	4	55	59	39	0,1	1,8	2,0	1,3	28	81	42	24	12
37	11	3,8	10	60	70	30	0,4	2,3	2,7	1,1	32	111	42	29	11
38	12	6,4	2	76	78	17	0,1	4,9	5,0	1,1	25	366	73	115	23
39	13	6,8									20	234	59	71	18
49	14	8,5									30	156	25	108	17
51	15	9,3									35	129	21	23	4
52	16	2,2									33	46	30	12	8
55	17	9,9									30	189	24	140	18
56	18	5	2	80	82	13	0,1	4,0	4,1	0,7	26	101	25	54	13
6	1	8,8	11	72	83	12	1,0	6,3	7,3	1,1	44	268	36,6	145	19,9
9	2	9,2	2	66	68	28	0,2	6,1	6,3	2,6	42	168	26,9	76	12,2
13	3	10,0	7	83	90	8	0,7	8,3	9,0	0,8	54	181	20,1	81	9
15	4	9,6	1	75	76	23	0,1	7,2	7,3	2,2	45	343	47	124	16,9
17	5	6,0									45	40		21	
19	6	7,2	2	78	80	18	0,1	5,6	5,8	1,3	44	103	17,8	63	10,9
24	7	8,2	1	72	73	25	0,1	5,9	6,0	2,1	41	209	34,9	57	9,6
26	8	10,3	1	85	86	3	0,1	8,8	8,9	0,3	41	364	41,1	112	12,6
29	9	5,0	1	74	75	21	0,1	3,7	3,8	1,1	39	83	22,1	32	8,5
30	10	11,9	1	51	52	47	0,1	6,1	6,2	5,6	33	83	13,4	52	8,4
31	11	7,2	3	53	56	29	0,2	3,8							
32	12	10,8	36	50	76	8	3,9	5,4							
33	13	11,8	22	54	76	16	2,6	6,4							
34	14	17,0	1	82	83	16	0,2	13,9							
40	15	10,0	8	72	80	4	0,8	7,2							

Data





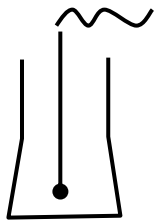


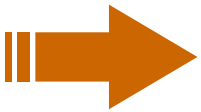
Schopnost: vidět data – komunikovat – interpretovat - prodávat

BIOSTATISTIKA - BIOMETRIKA

Věda zabývající se variabilitou

Variabilita opakovaných měření



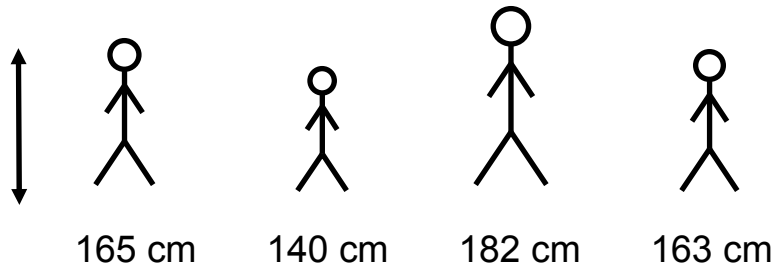


Data

2,1
2,8
3,2
1,2
5,2
2,9

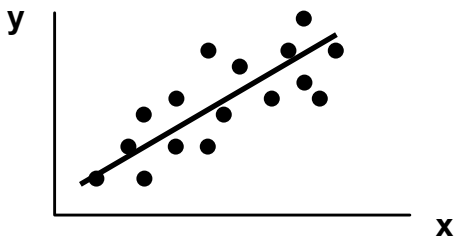
chyba

Variabilita znaku v populaci



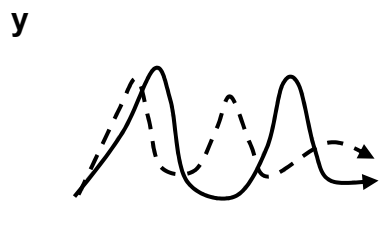
rozptyl znaku, přirozená variabilita

Variabilita modelovaných dat



chyba = nepřesnost modelu

Variabilita časových řad



fluktuace, časová proměnlivost

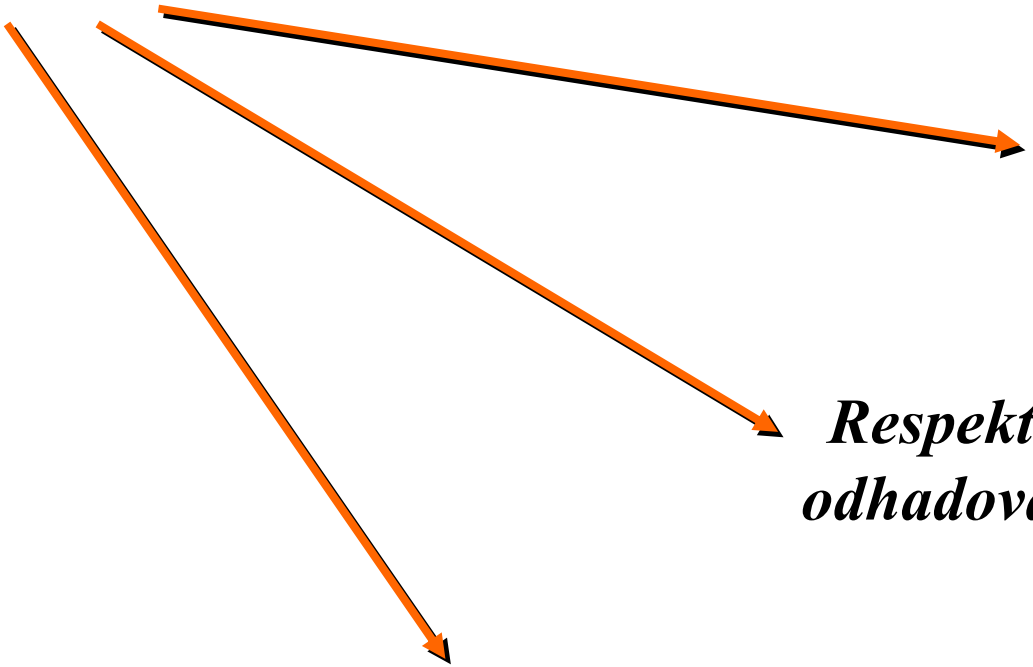
Variabilita ve skladbě biologických společenstev

DRUH 1	**15**
DRUH 2	**30**
DRUH 3	**40**
DRUH 4	**14**

biodiverzita

Pojem VARIABILITA má mnoho významů ……

…. a ty určují přístup k jejímu hodnocení



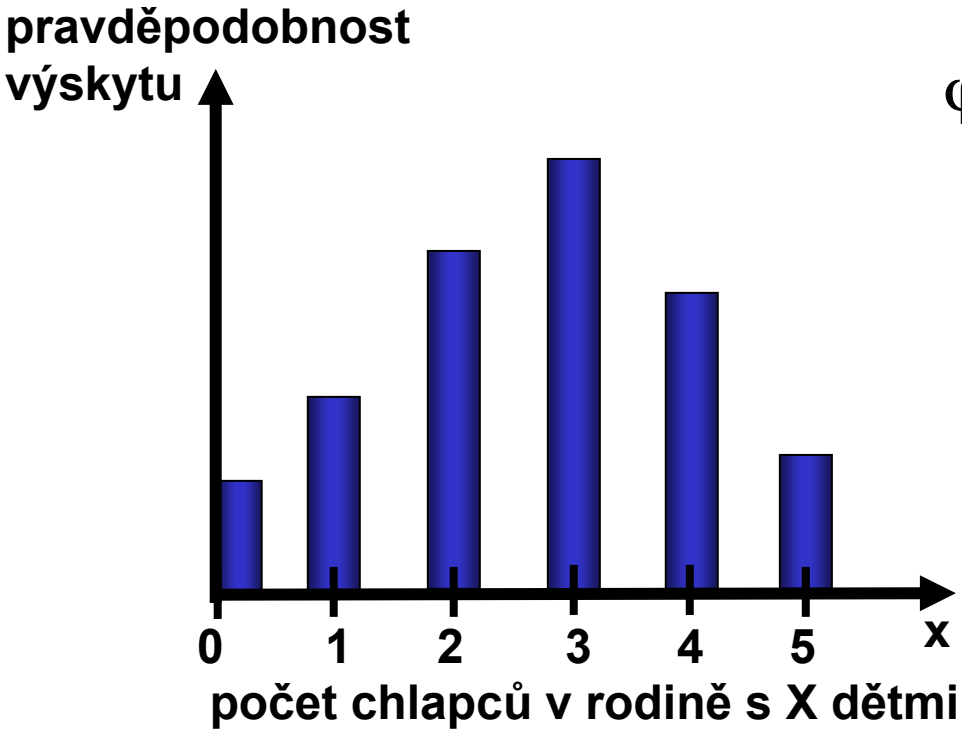
Maskování a minimalizace vlivu

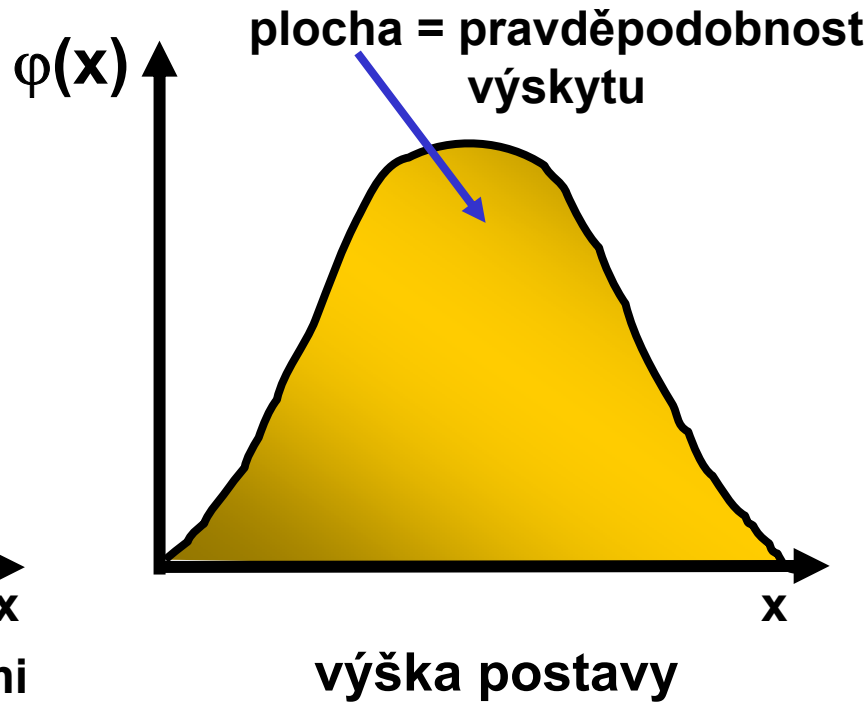
Respektování a odhadování vlivu

Přímé využití k predikcím chování systému

Variabilita = základ „biologického principu neurčitosti“

- existuje pravděpodobnost výskytu jevů (nedeterministické závěry)
- „vše je možné“: pouze jev s pravděpodobností 0 nikdy nenastane
- pravděpodobnost lze zkoumat retrospektivně i prospektivně





BIOSTATISTIKA - BIOMETRIKA

Věda přinášející novou kvalitu

- Popisná analýza dat („exploratorní“ analýzy)
- Data mining („investigativní“ analýzy)

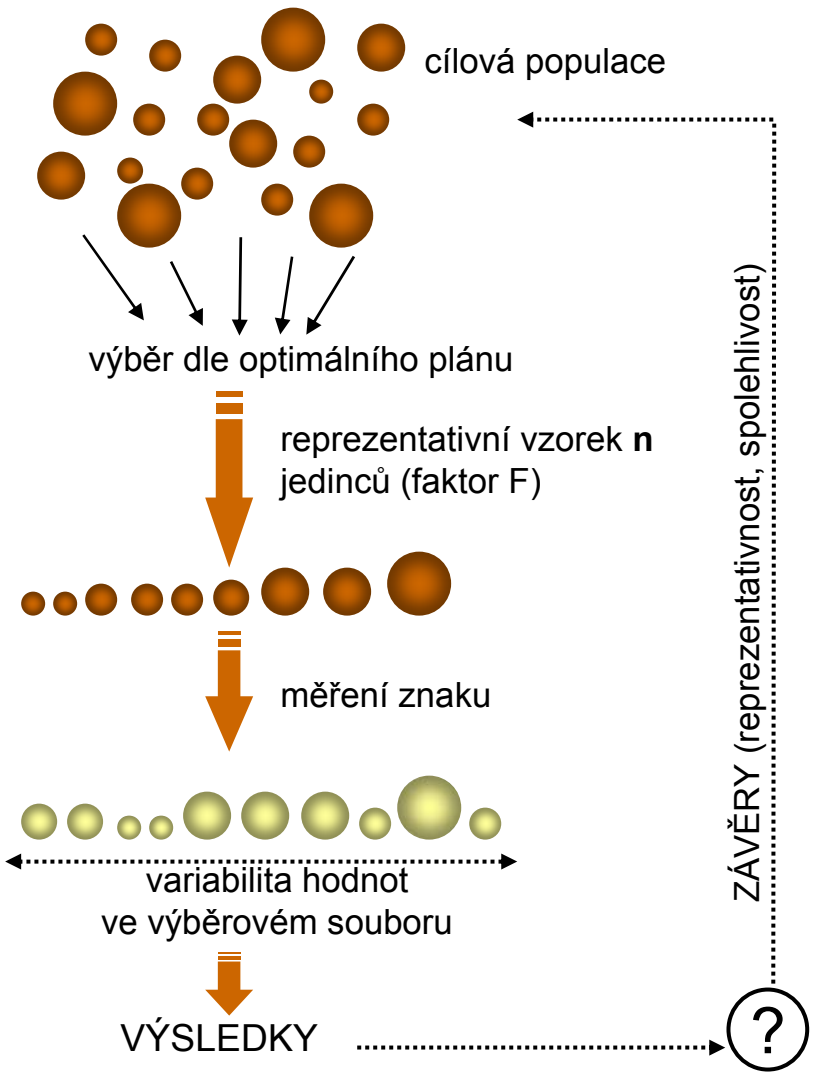
- Srovnávací analýzy, testy hypotéz
- Experimentální plány („experimental design“)
- QA/QC

- Stochastické modelování, hodnocení prognóz
- Vícerozměrné analýzy, „pattern recognition“

- Analýza biodiverzity (species community associations, ….)
- Analýza časových řad, analýzy trendů
- Analýza biomedicínských dat

Experimentální design: nezbytná výbava biologa

Účel analýzy:
Popisný

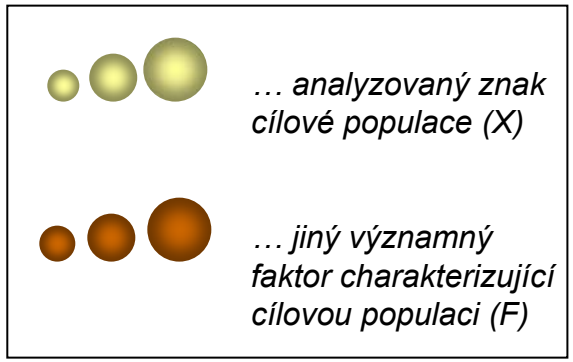


?

Reprezentativnost

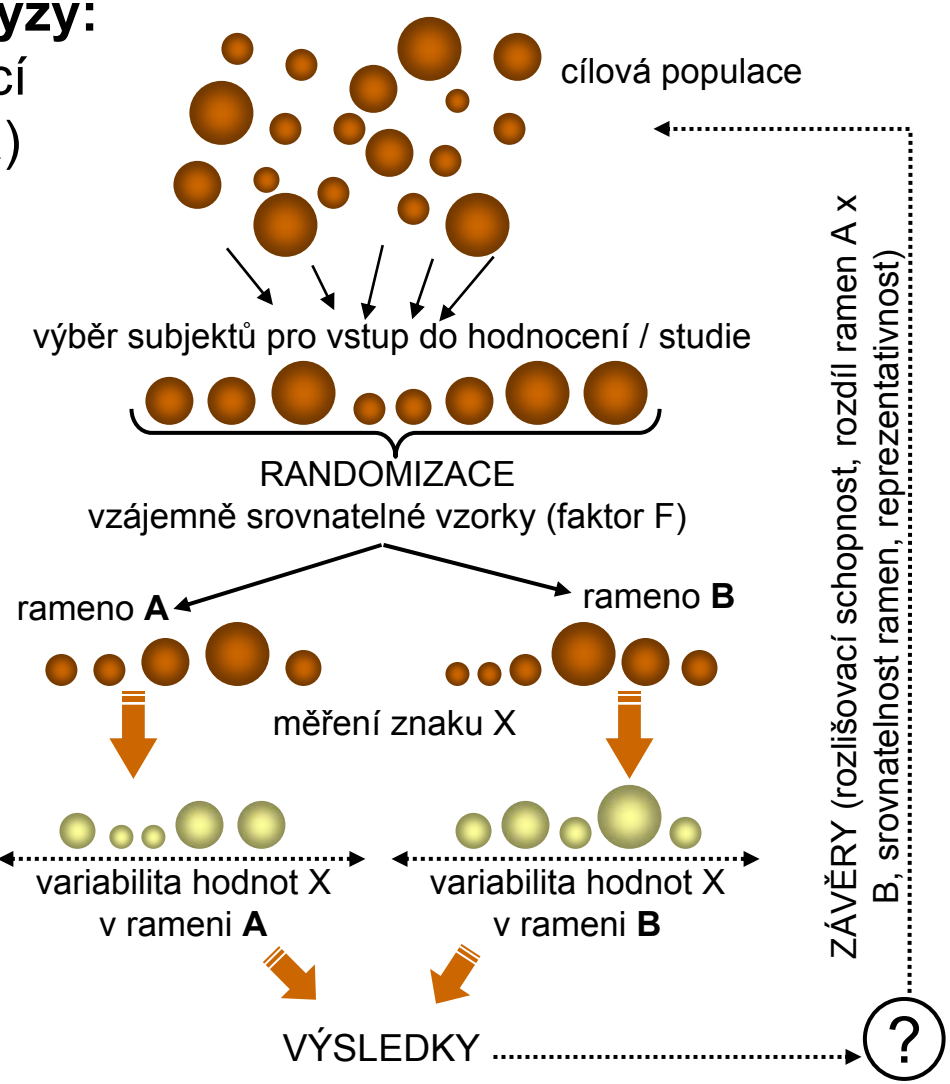
Spolehlivost

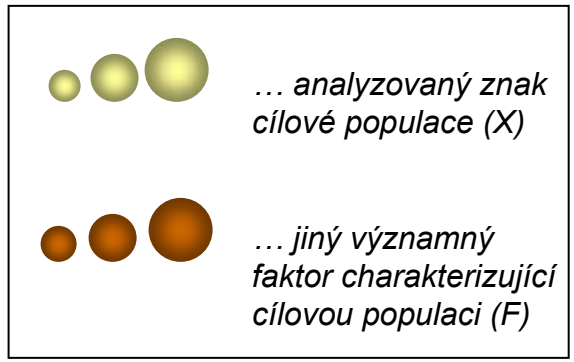
Přesnost



Experimentální design: nezbytná výbava biologa

Účel analýzy:
Srovnávací
(2 ramena)





?

Srovnatelnost

Spolehlivost

Přesnost

Stochastické modelování: predikce neurčitých jevů

- Prospektivně – modelově - postihuje chování jevů při respektování variability

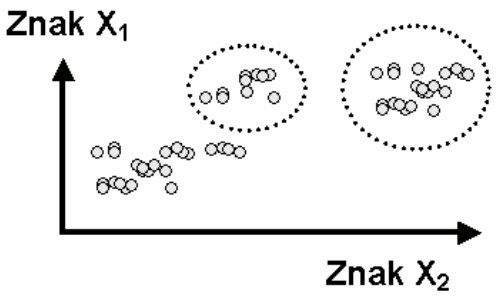
Pravděpodobnostní vztahy

Anamnéza x Výsledek vyšetření pacienta

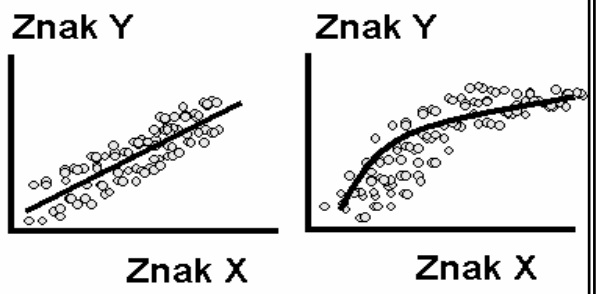
	Karcinom	Benigní léze	Benigní riziková	Zdravá	
Pozitivní anamnéza	2,22	34,44	0,00	63,33	100%
Negativní anamnéza	1,06	28,23	0,96	69,75	100%

$p < 0.05$

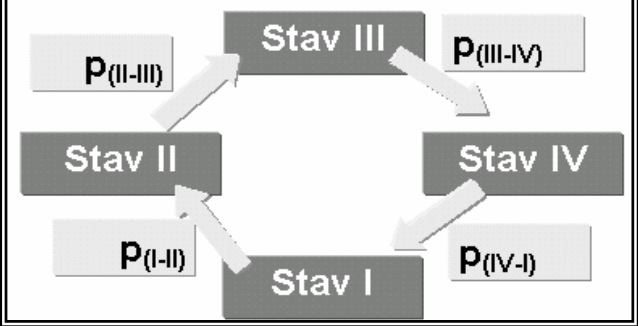
Vícerozměrná diskriminace



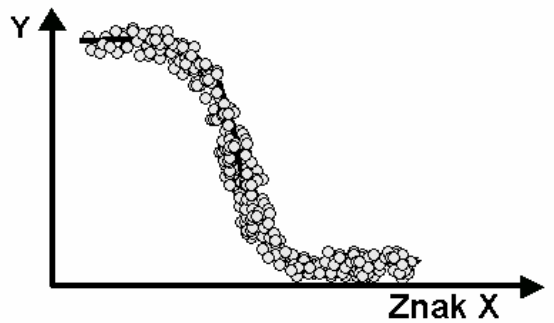
Funkční vztahy znaků



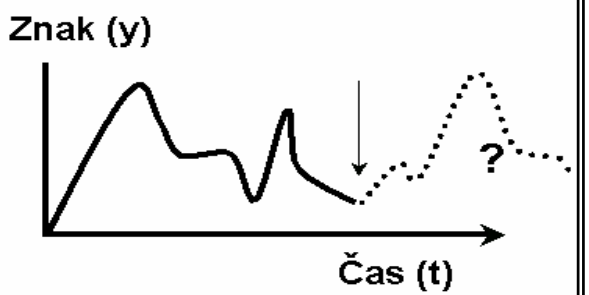
Markovovy řetězce



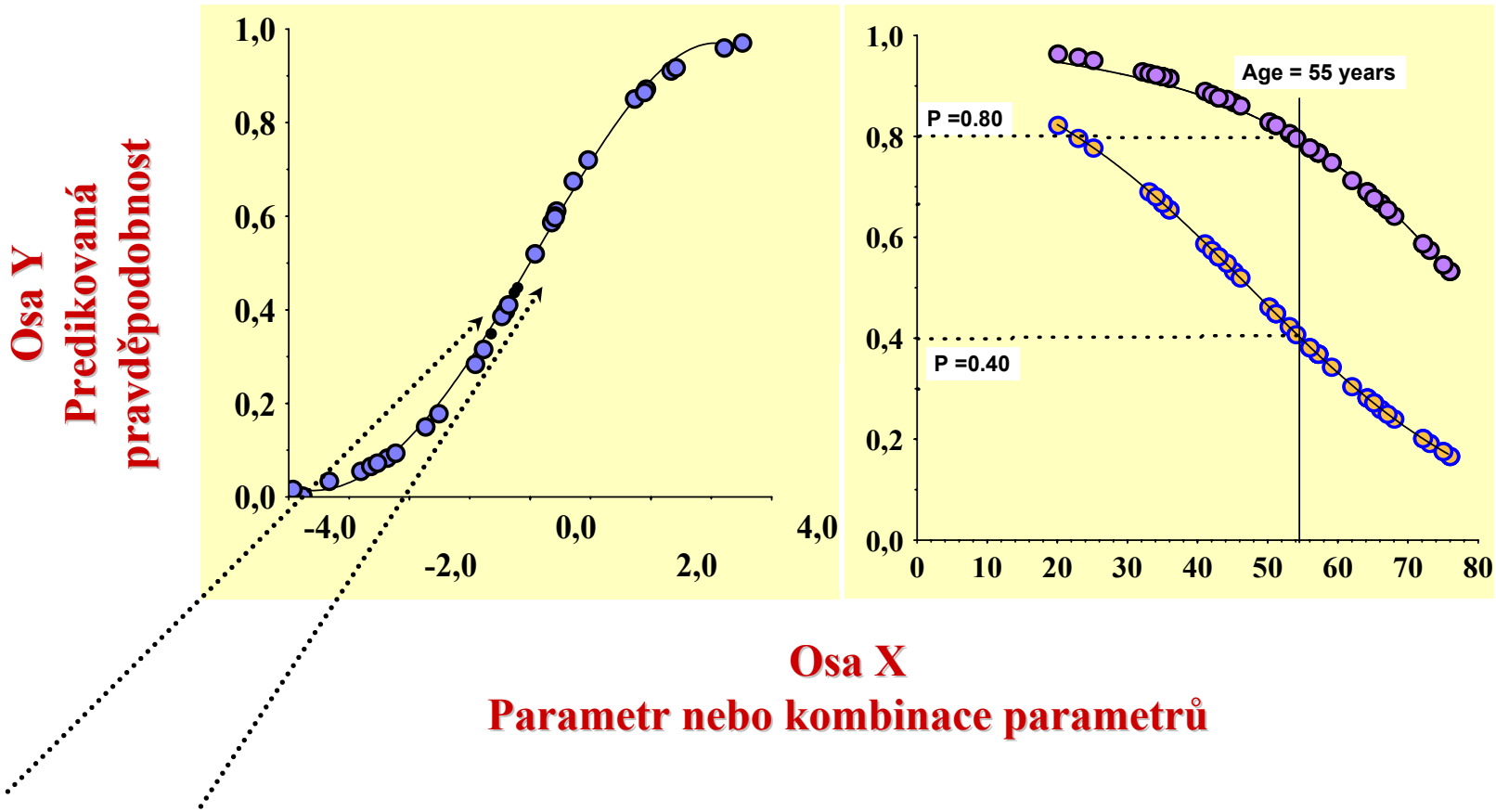
Logistické modely



Chování systému v čase

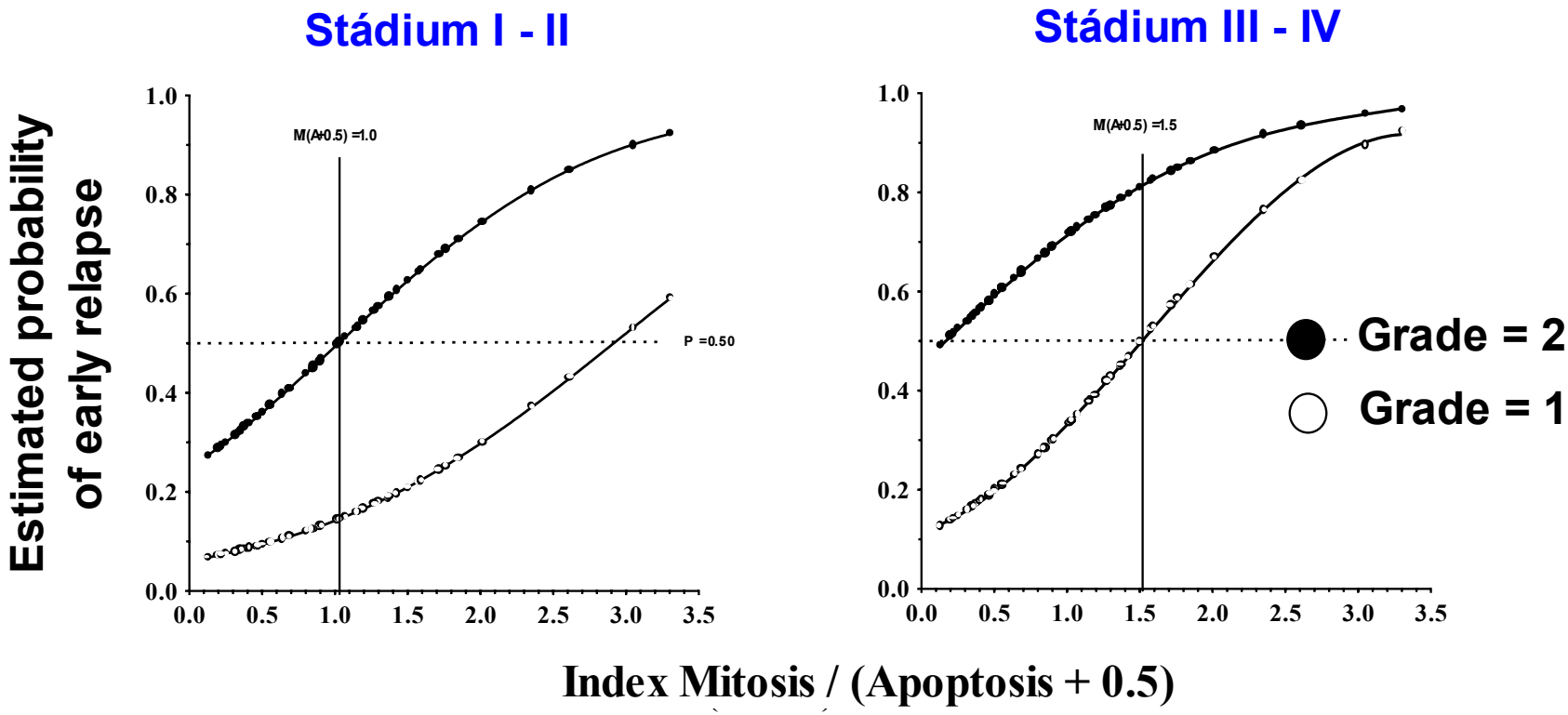


Stochastické modelování: predikce neurčitých jevů



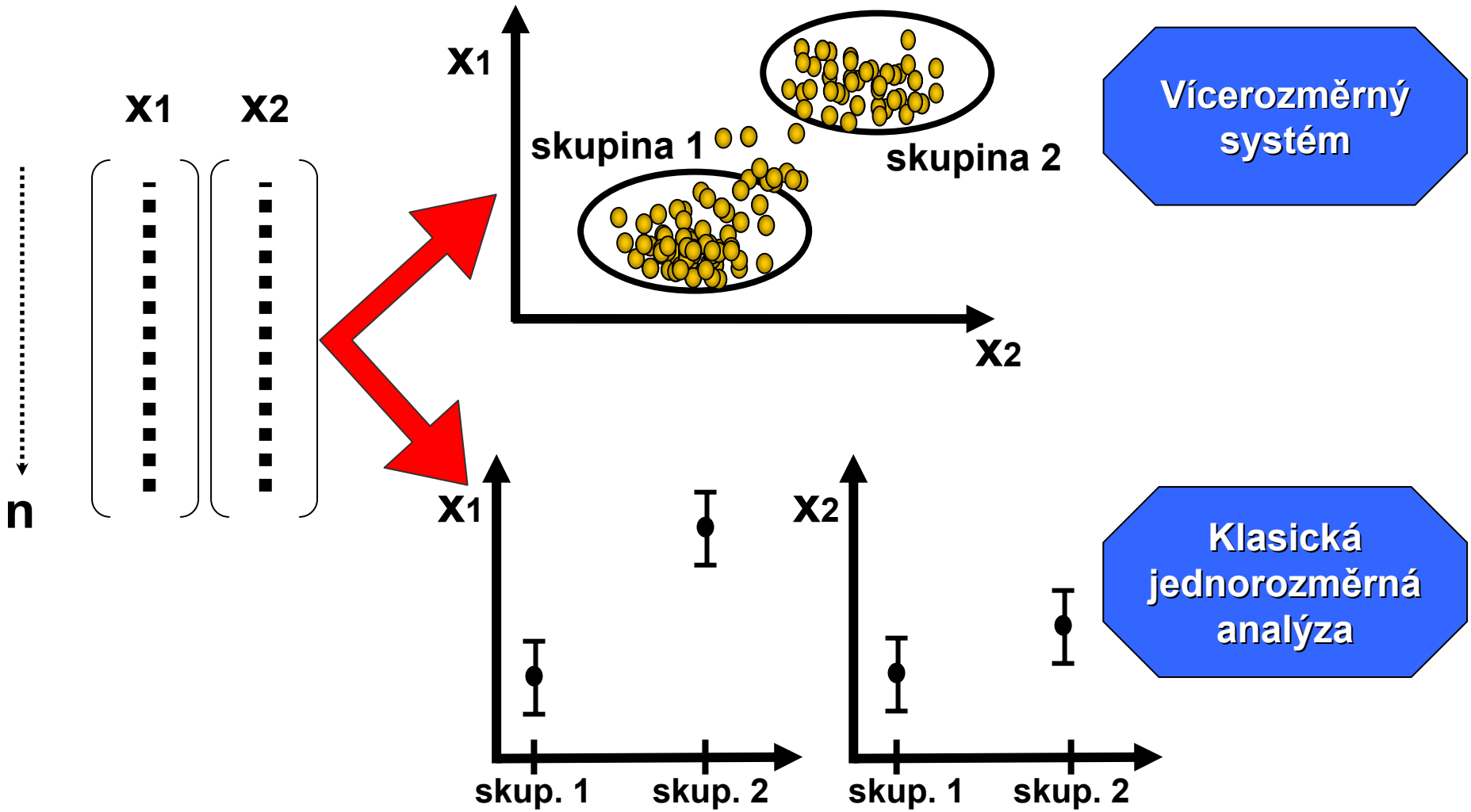
Data konkrétních pacientů (subjektů)
k přímému hodnocení

Maligní lymfomy: Pravděpodobnost časného relapsu

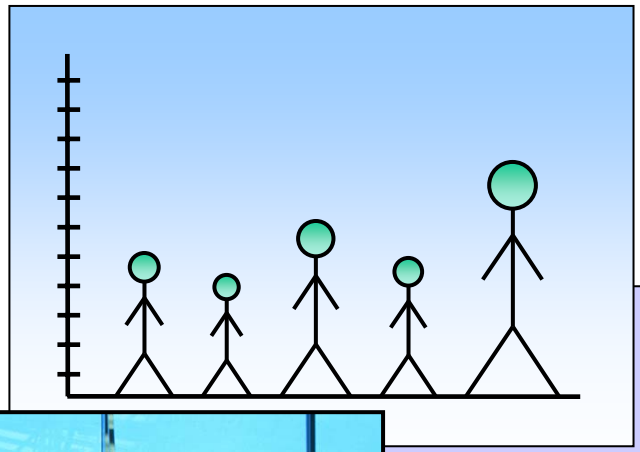


Schopnost: vytvářet prakticky využitelné nástroje

Vícerozměrné vnímání skutečnosti – nová kvalita analýzy dat



Biologové analýzou dat proti variabilitě nebojují !



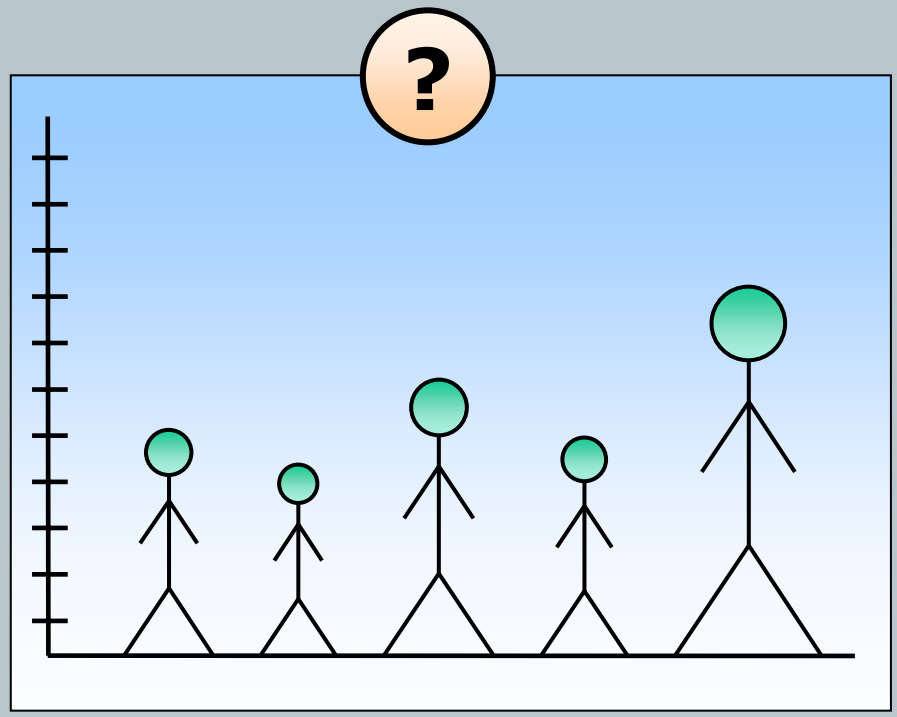
VARIABILITA

?

CHYBA

INFORMACE

Běžná sumarizace dat „likviduje“ individualitu jedince



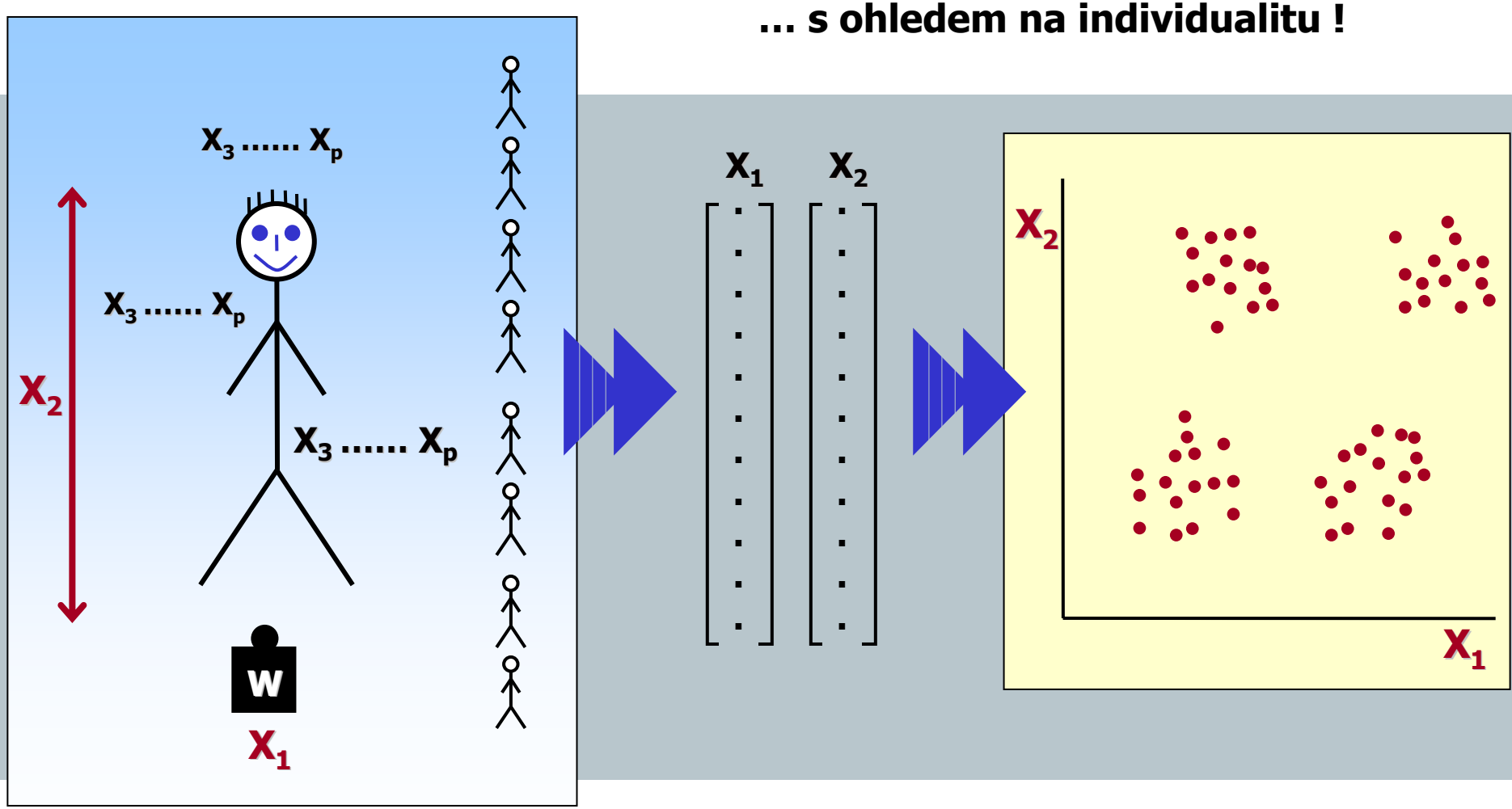
Průměr ± SE

BĚŽNÁ STATISTICKÁ SUMARIZACE

- ✓ *Zpřehlednění dat*
- ✓ *Neodliší původní měření*

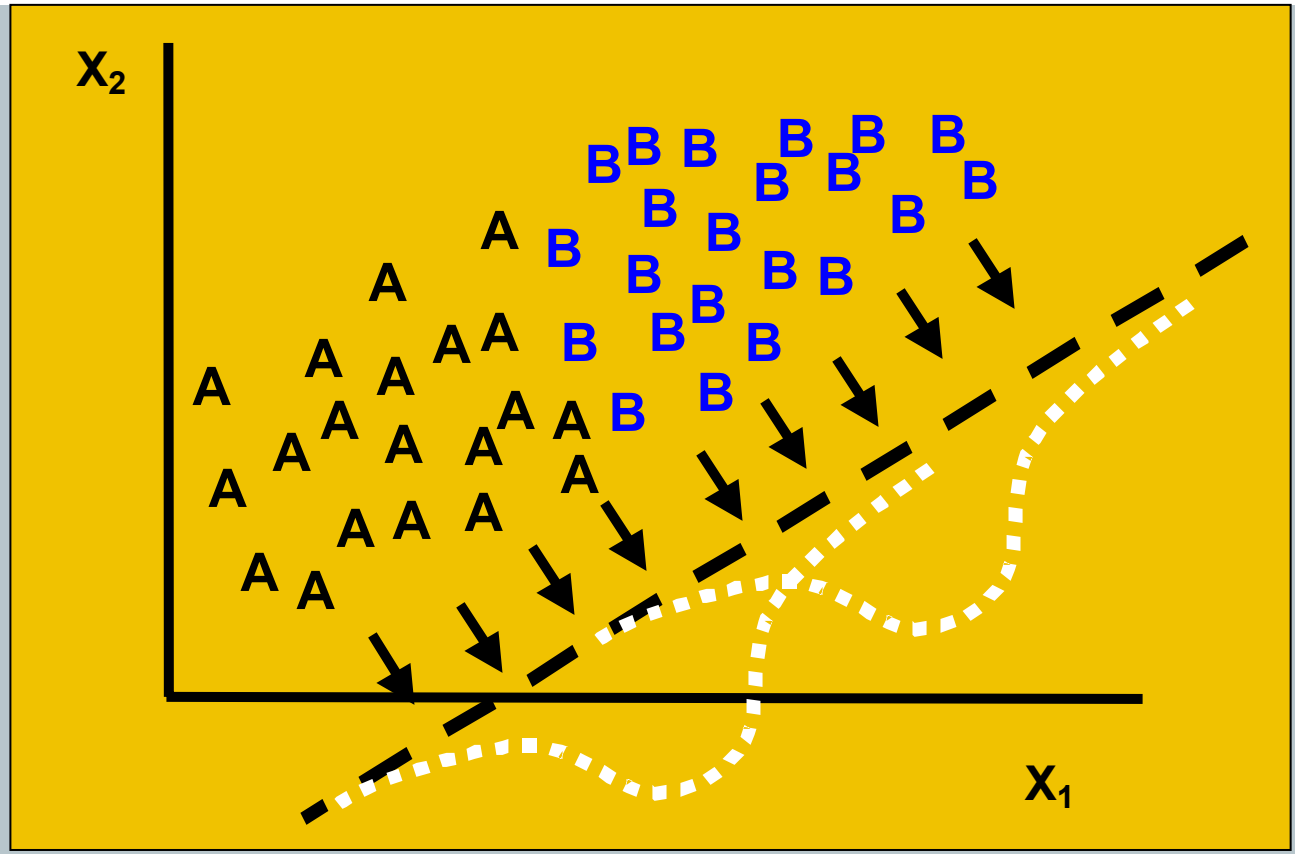
Vícerozměrné hodnocení

… s ohledem na individualitu !



Vícerozměrné hodnocení – nová kvalita

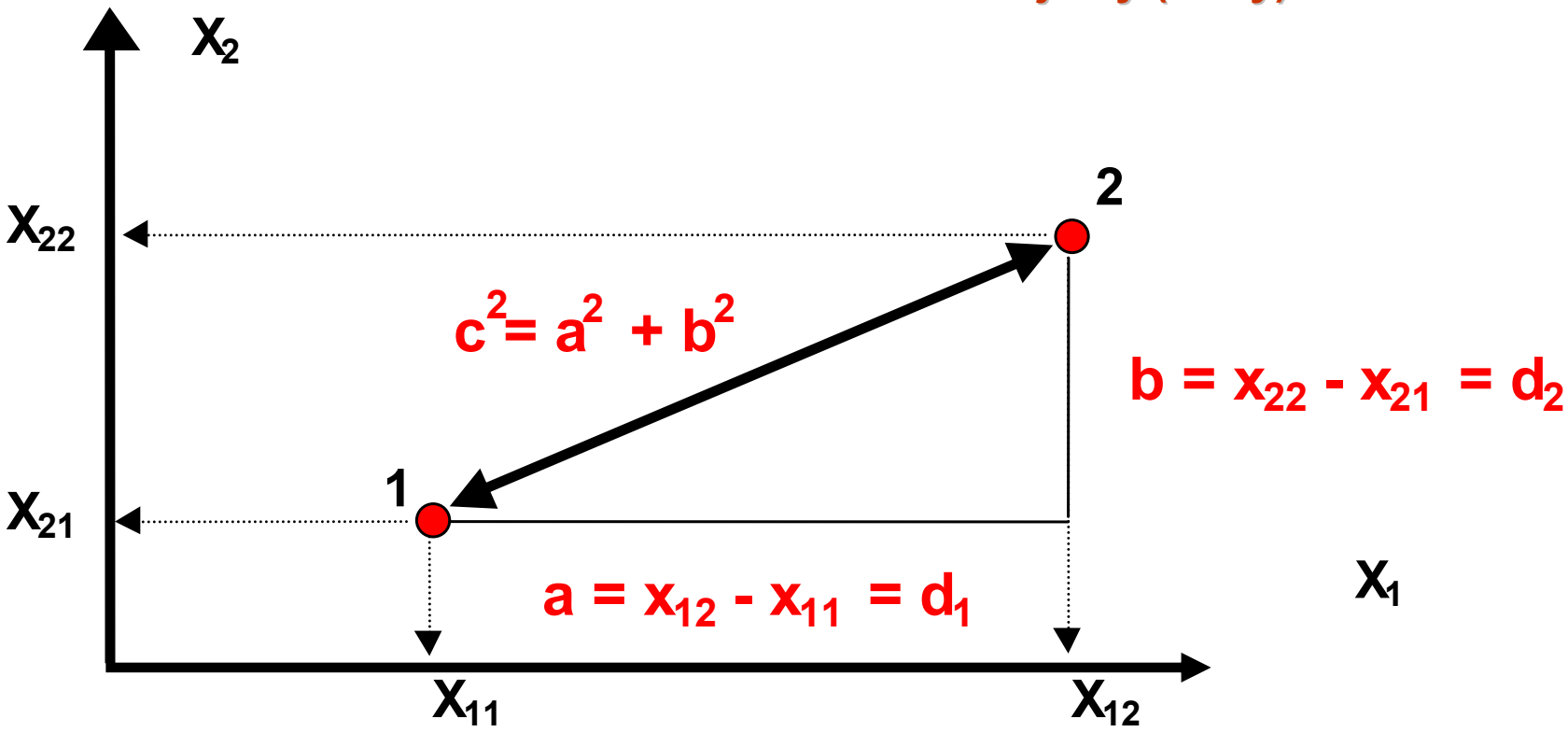
Pouze kombinované parametry mají odpovídající informační sílu



příklad: X1 =

Vícerozměrné hodnocení vychází z jednoduchých principů

příklad: vícerozměrná vzdálenost měření mezi dvěma objekty (body)



Vícerozměrné modelování je strategickou disciplínou

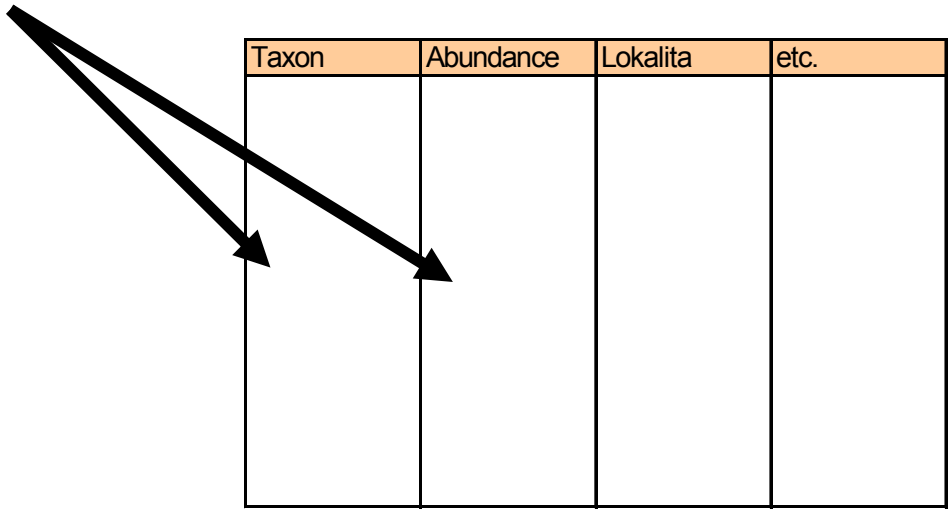
$X_1 \ldots\ldots X_n$	$X_{n+1} \ldots\ldots X_p$	$X_{p+1} \ldots\ldots X_2$
technické parametry automobilu	řidičovy schopnosti a jeho stav	rychlost, povrch, situace

X_1 X_2 X_3 X_4 X_5 ..... X_p

2. Data a jejich prezentace – základ statistické analýzy

Zásady pro ukládání dat

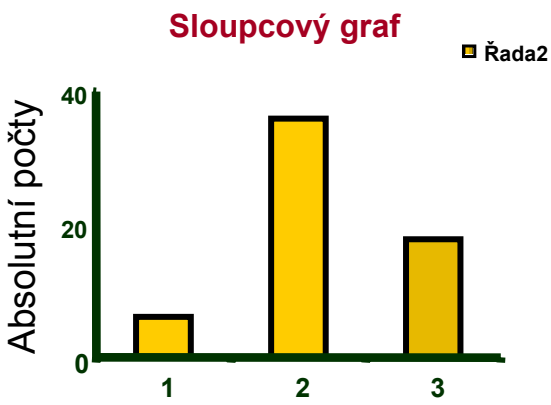
- Správné a přehledné uložení dat je základem jejich pozdější analýzy
- Je vhodné rozmyslet si předem jak budou data ukládána
- Pro počítačové zpracování dat je nezbytné ukládat data v tabulární formě
- Nejvhodnějším způsobem je uložení dat ve formě databázové tabulky
  - Každý sloupec obsahuje pouze jediný typ dat, identifikovaný hlavičkou sloupce (např. rozepsané taxonomické zařazení, abundance, místo a vlastnosti odběru atd.)

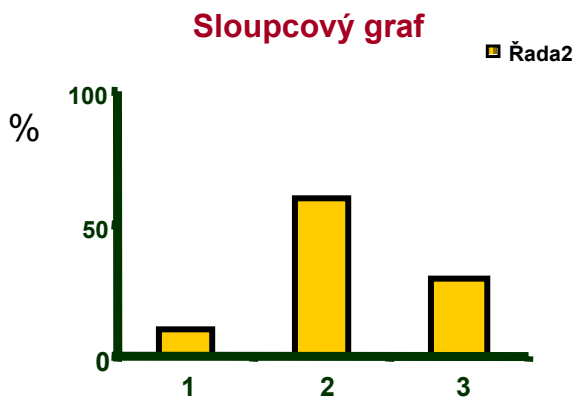


- Takto uspořádaná data je v tabulkových nebo databázových programech možné převést na libovolnou výstupní tabulku

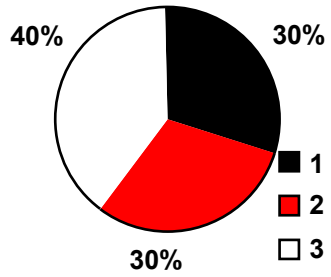
Grafická prezentace dat - umění komunikace

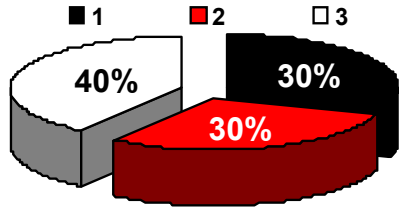
1. Výskyt kategorií (1, 2, 3,)



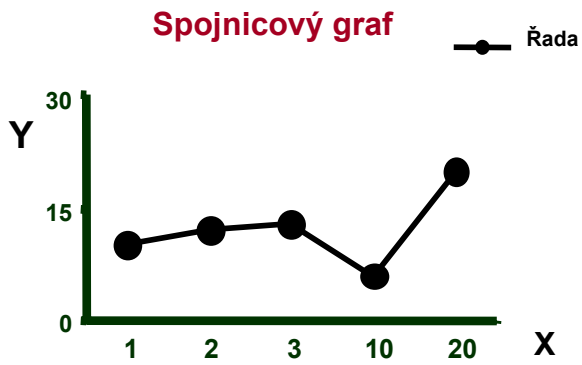


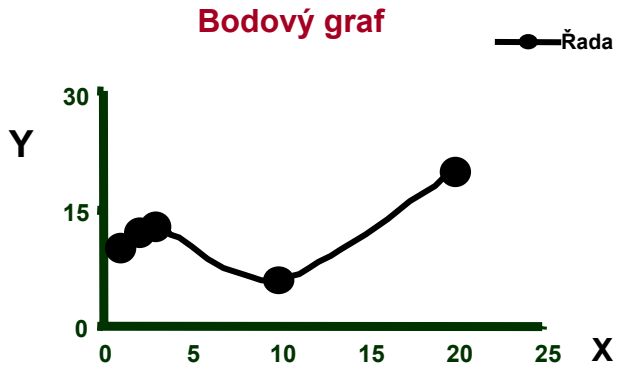
Koláčový (výsečový) graf

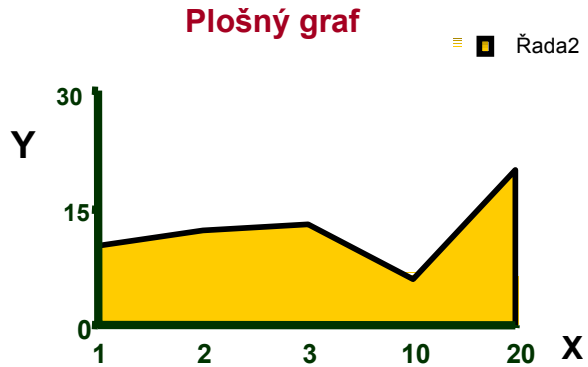




2. Vývoj hodnot (v čase) Y vs. X (t)

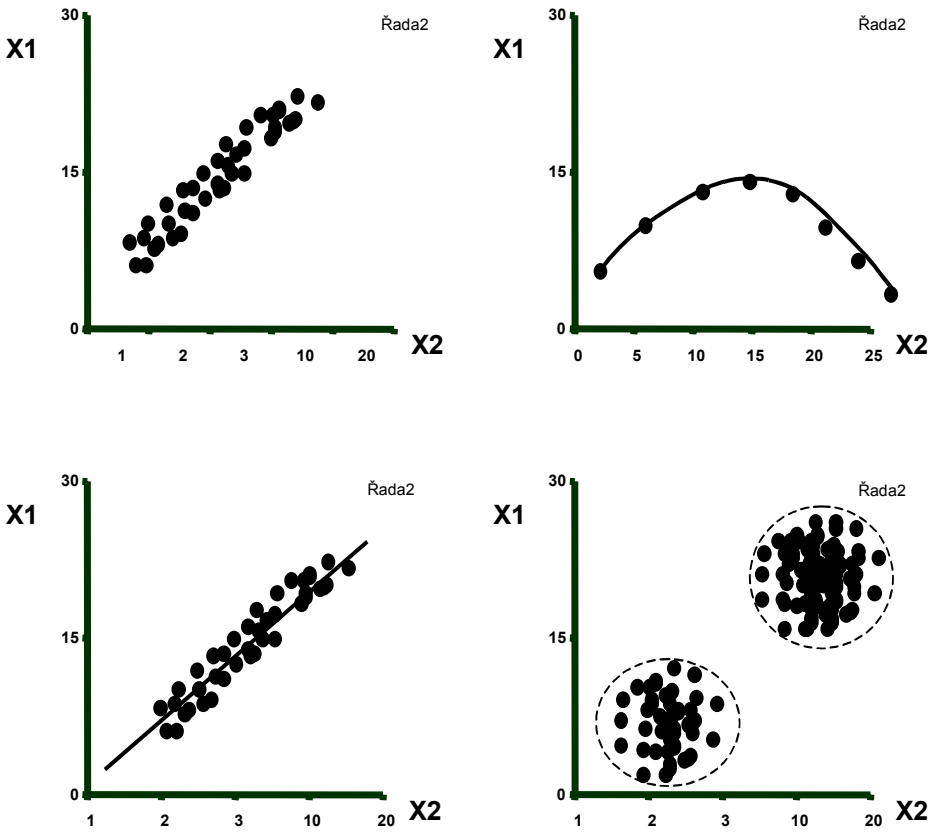




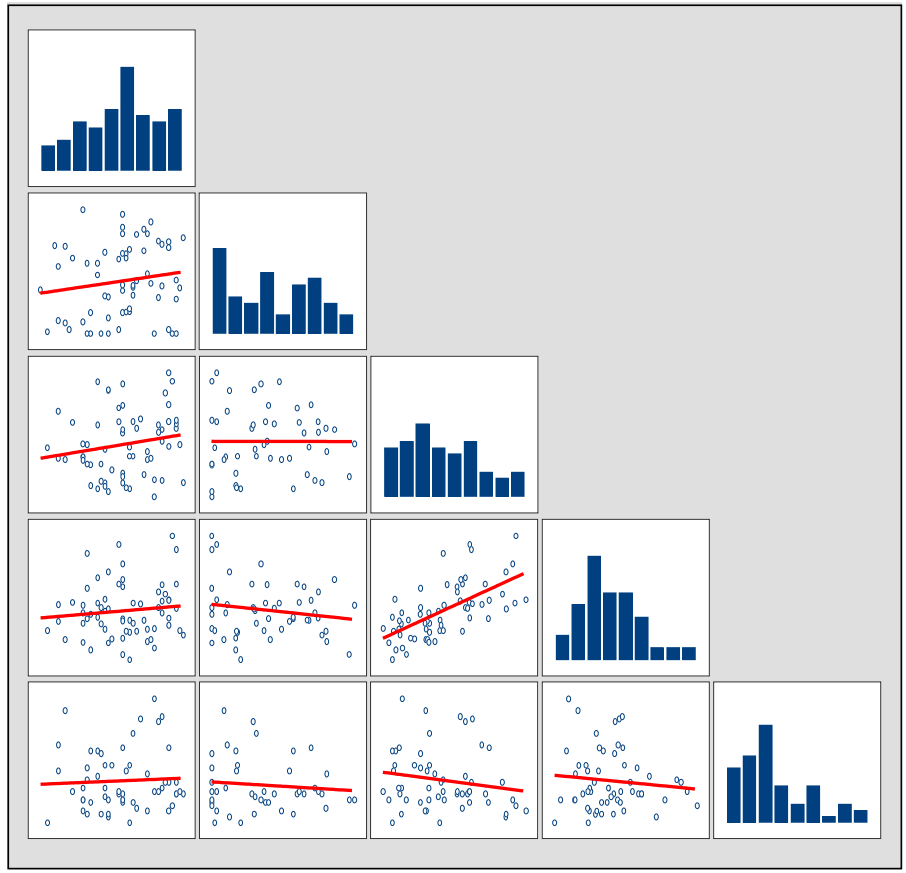


3. Vztahy mezi proměnnými - korelace

Bodový - korelační diagram



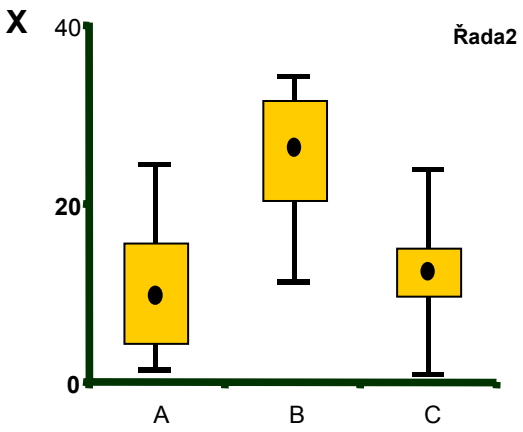
Bodový - korelační diagram

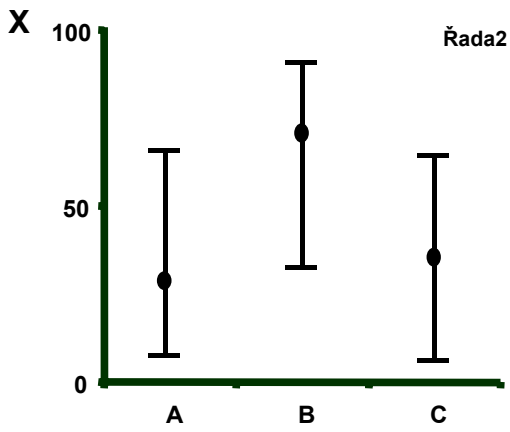


Grafická prezentace dat – umění komunikace

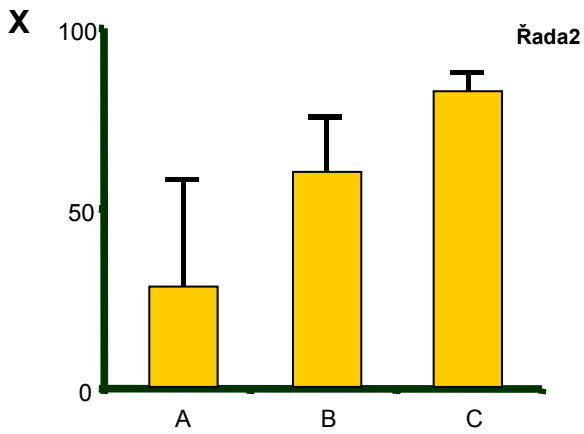
4. Kvantitativní hodnoty parametru(ů) - X - v rámci kategorií A, B, C

Krabicový graf

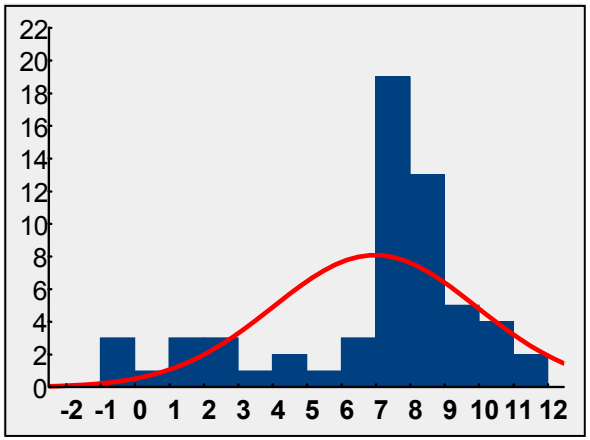


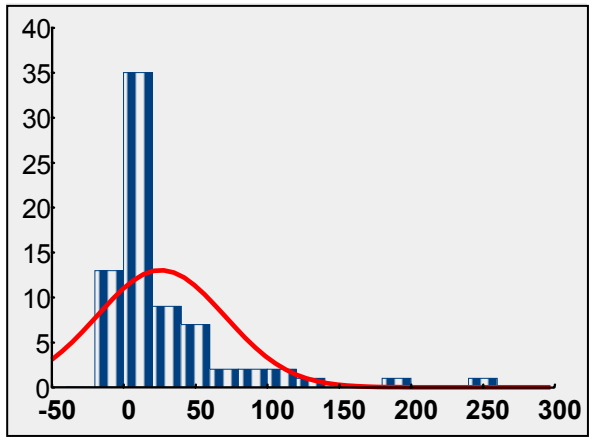


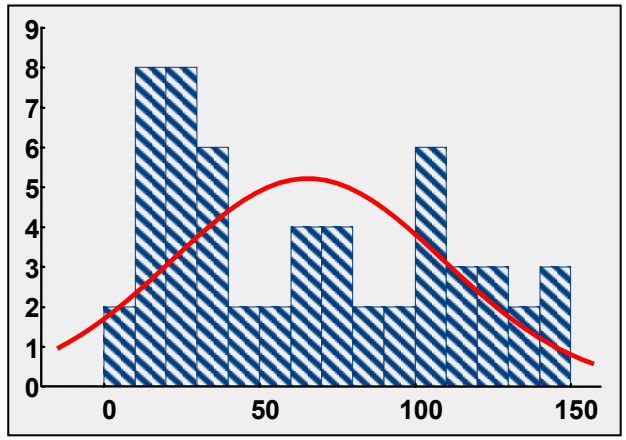
Sloupcový graf



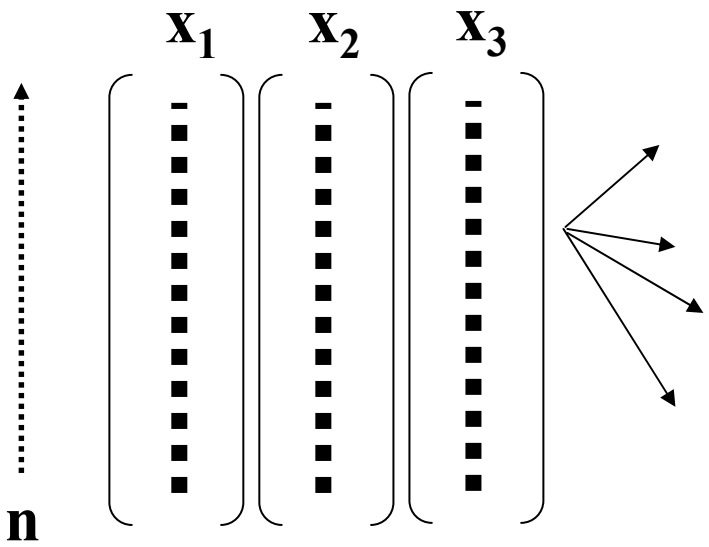
5. Histogram

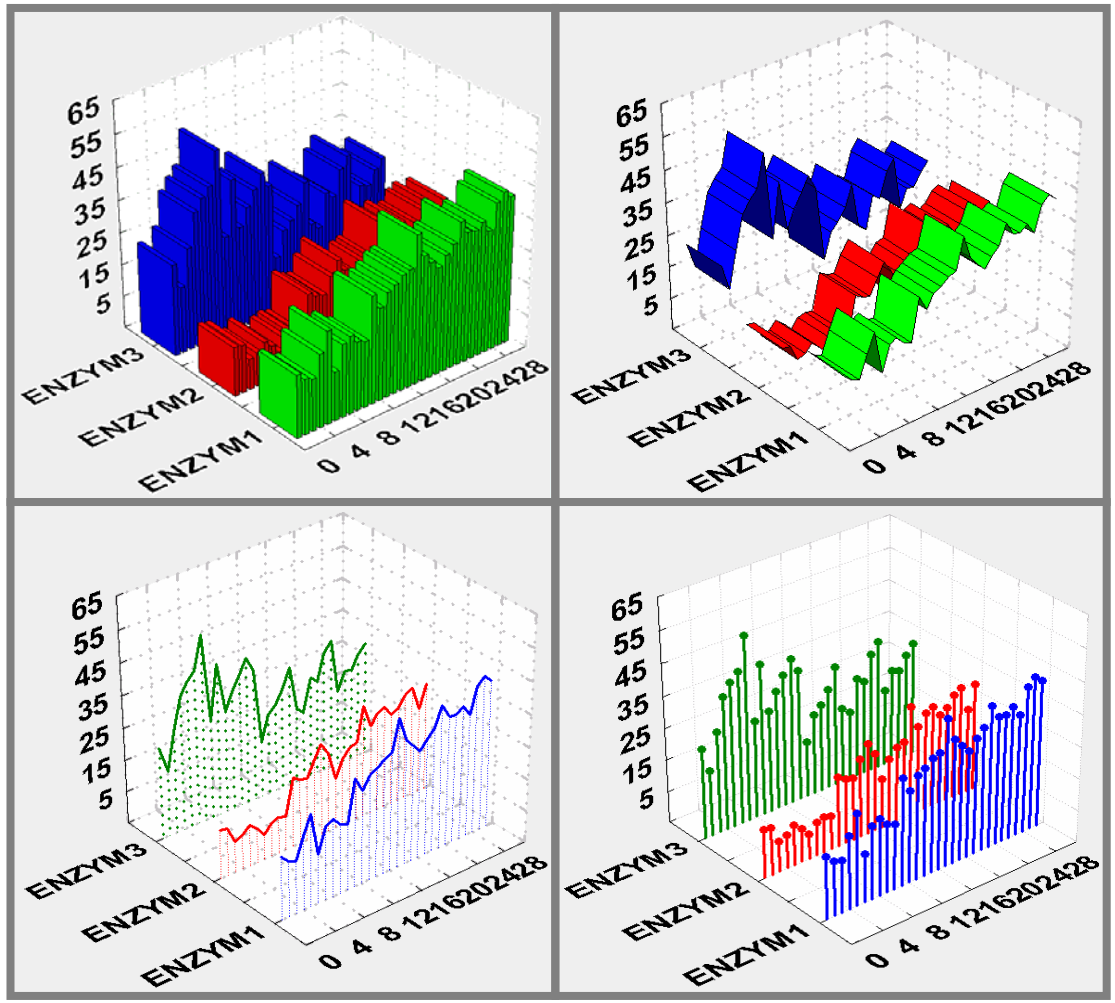






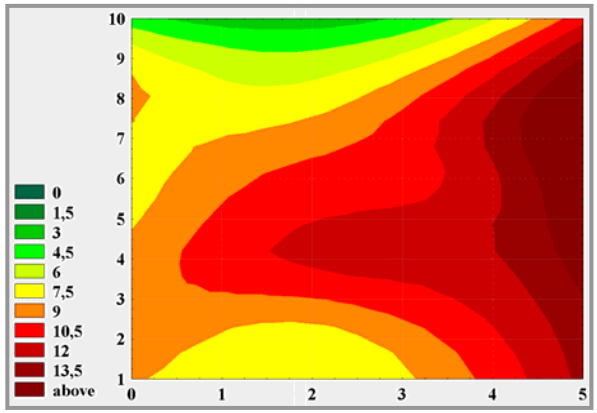
6. Zviditelnění primárních dat

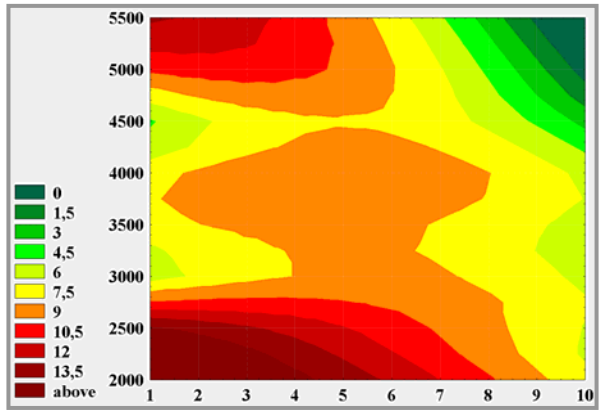


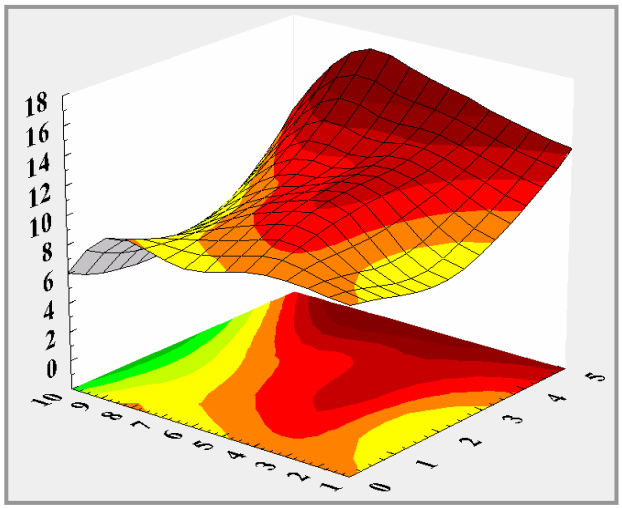


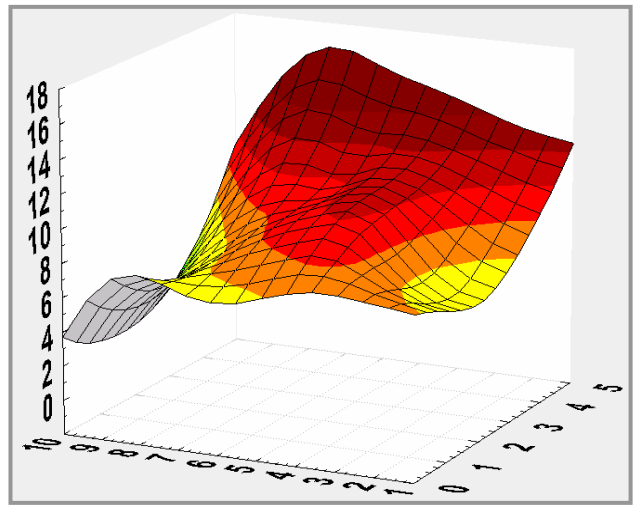
Grafická prezentace dat – umění komunikace

7. Vztahy mezi proměnnými - interakce dvou parametrů, reakční plochy



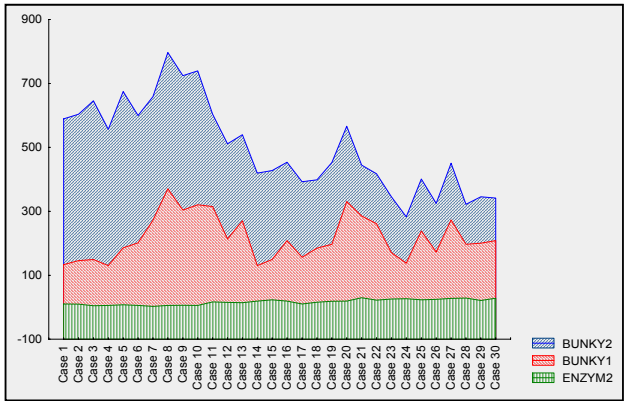


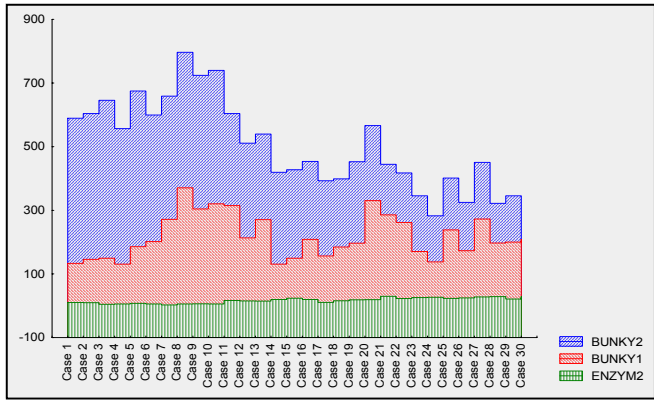


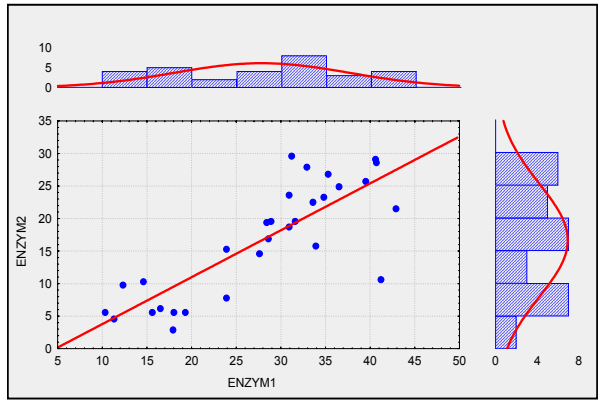


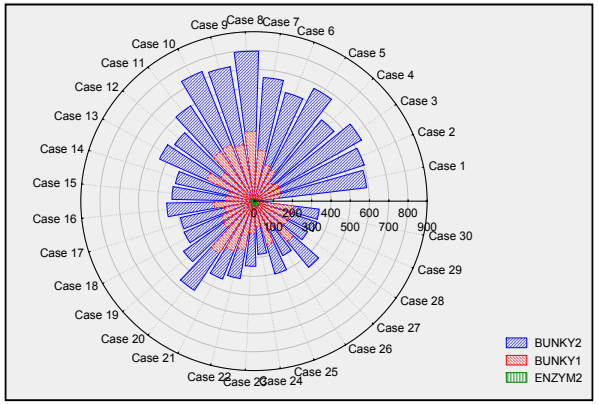
Grafická prezentace dat – umění komunikace

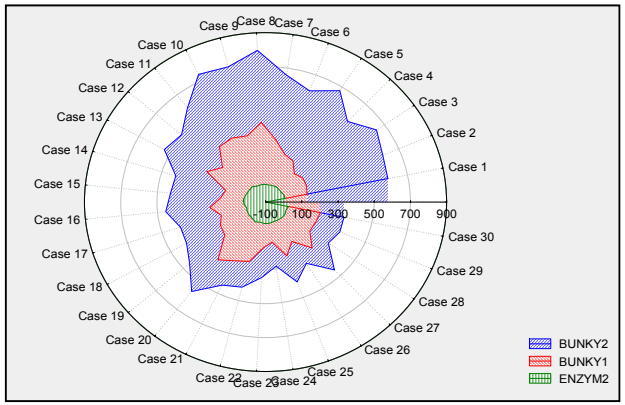
8. Grafické zviditelnění má nekonečně mnoho možností

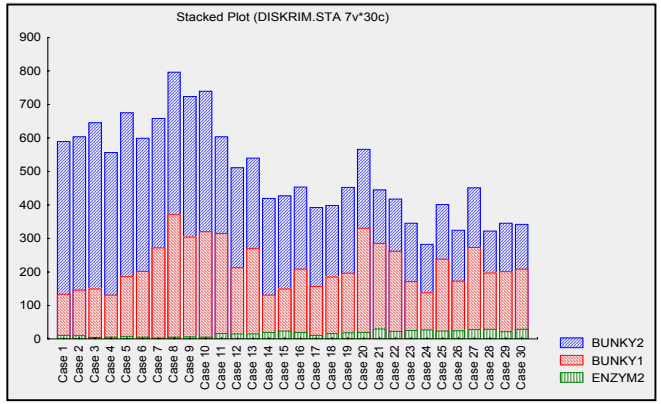




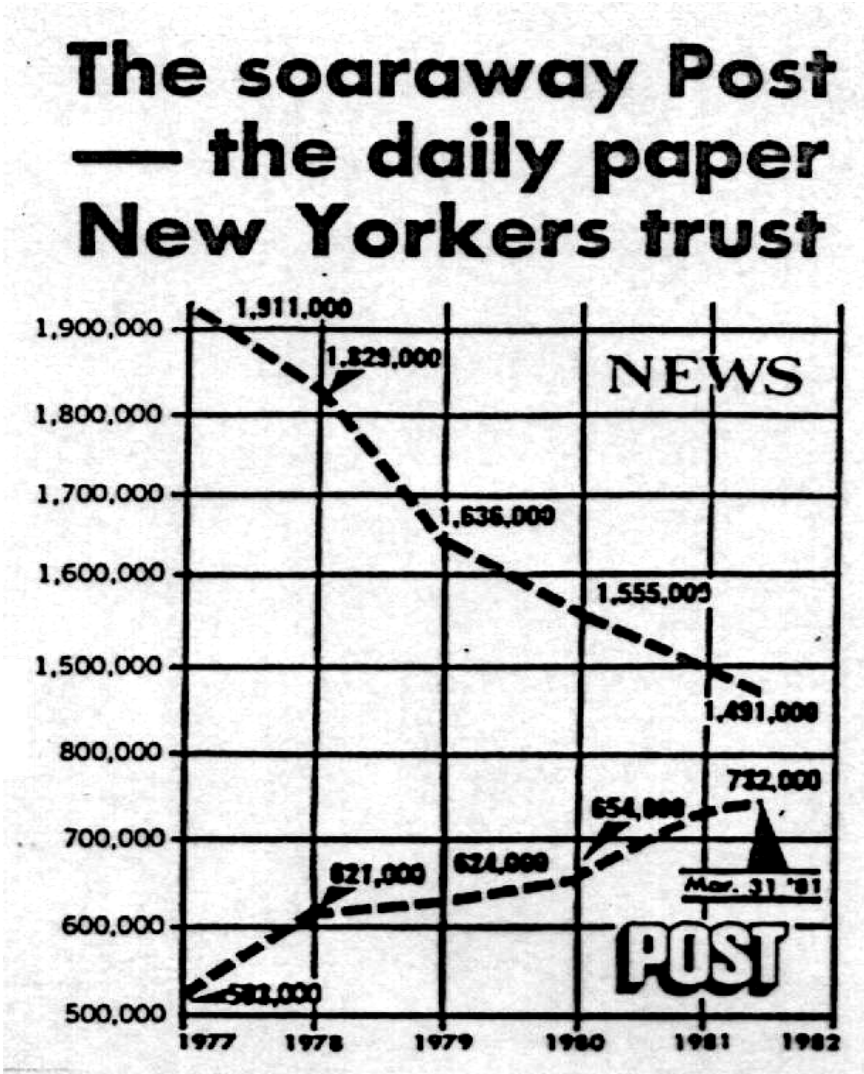


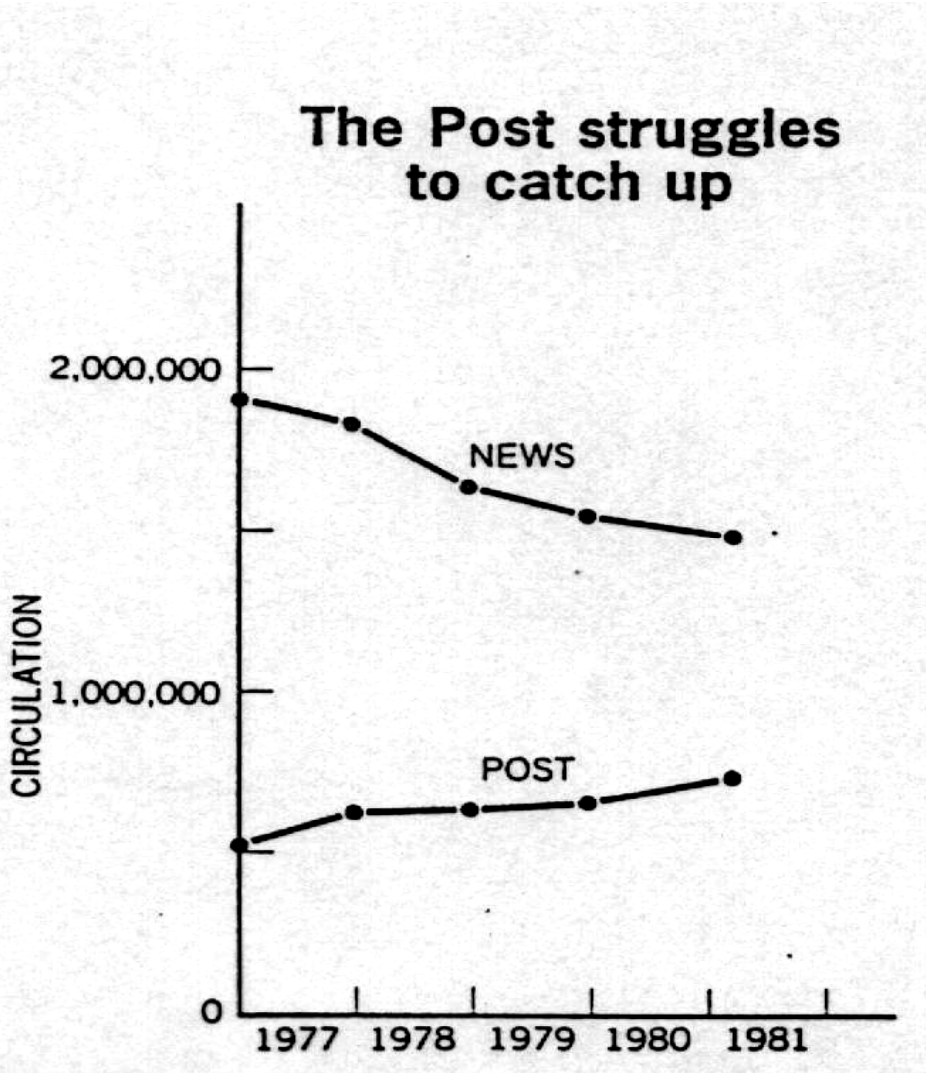




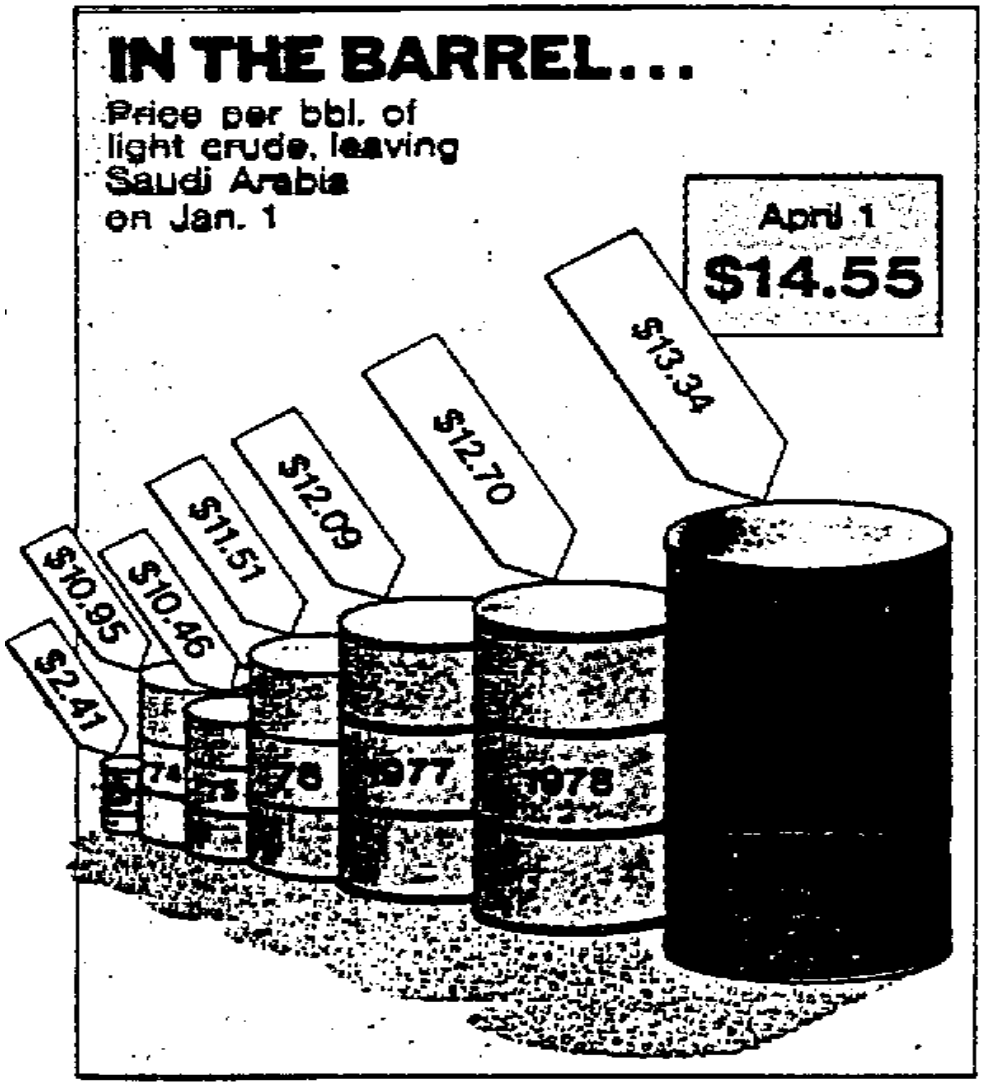


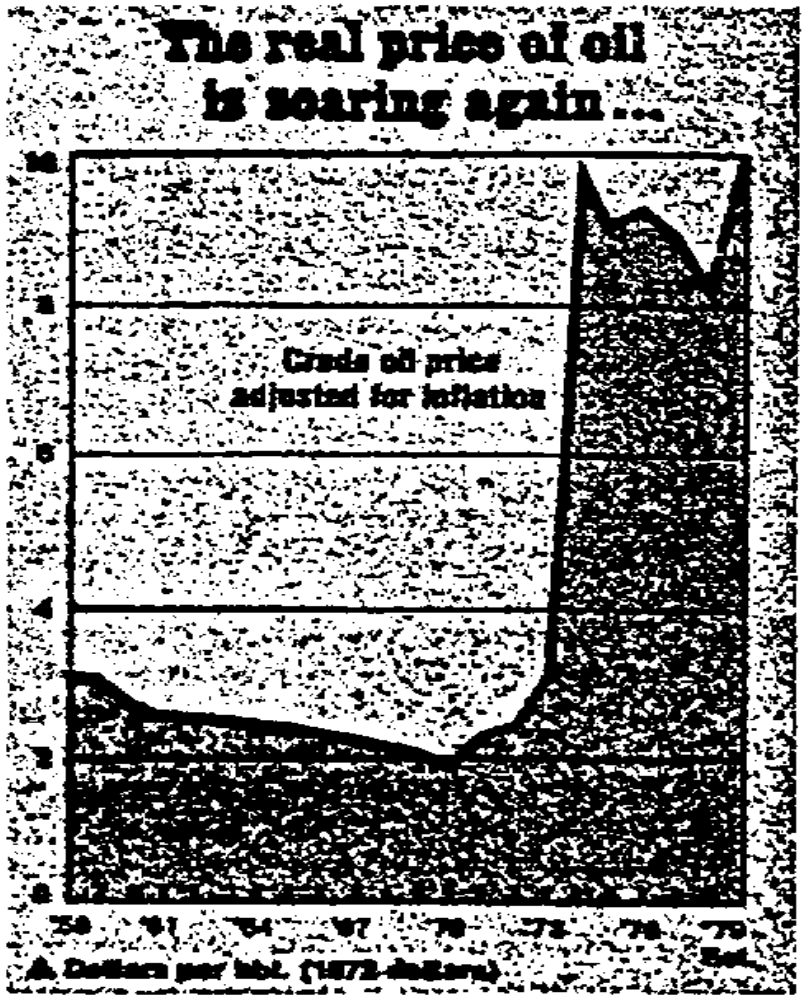
Nesprávné užití grafů - problém rozsahu číselné osy



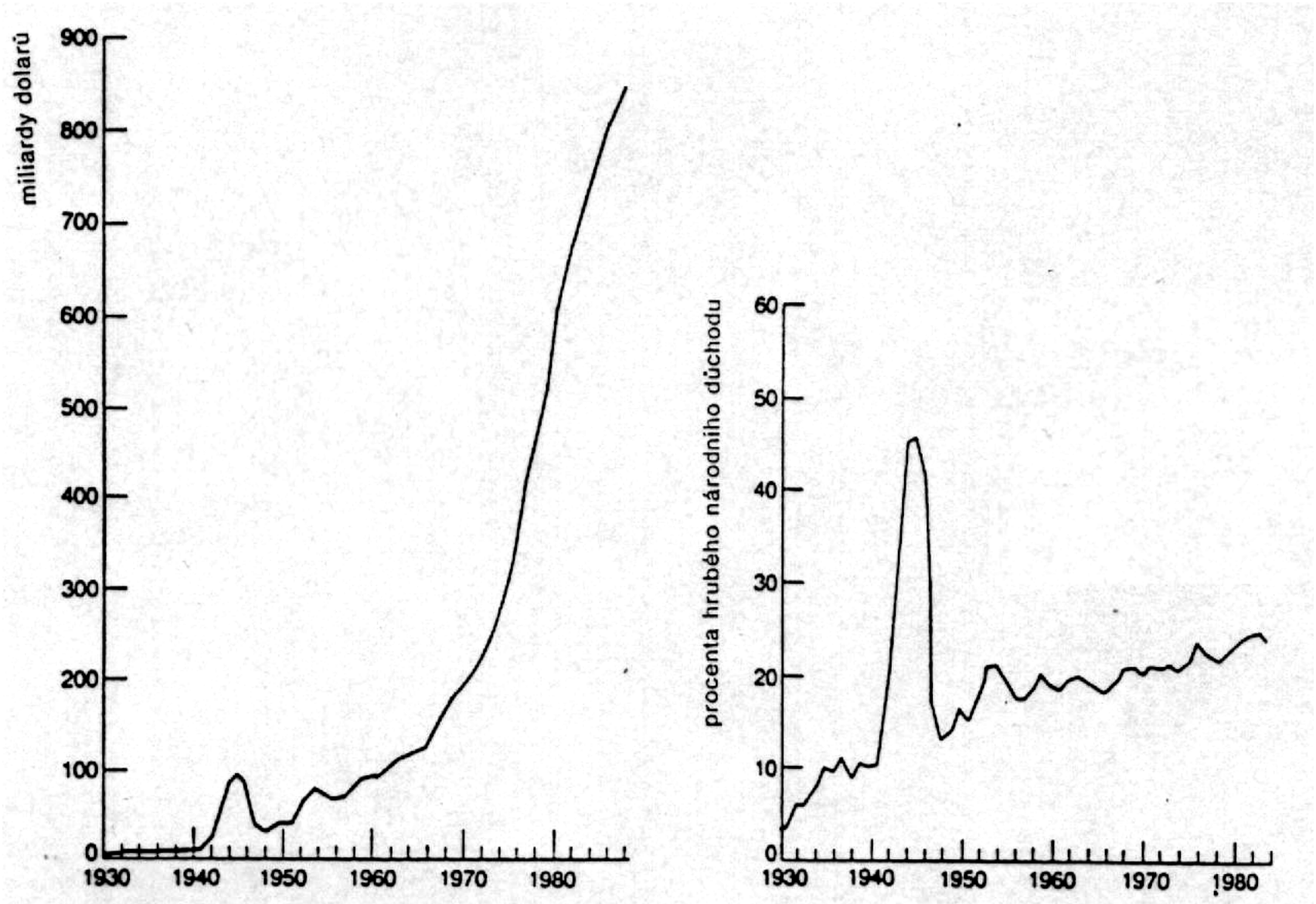


Nesprávné užití grafů - grafické zastírání trendu

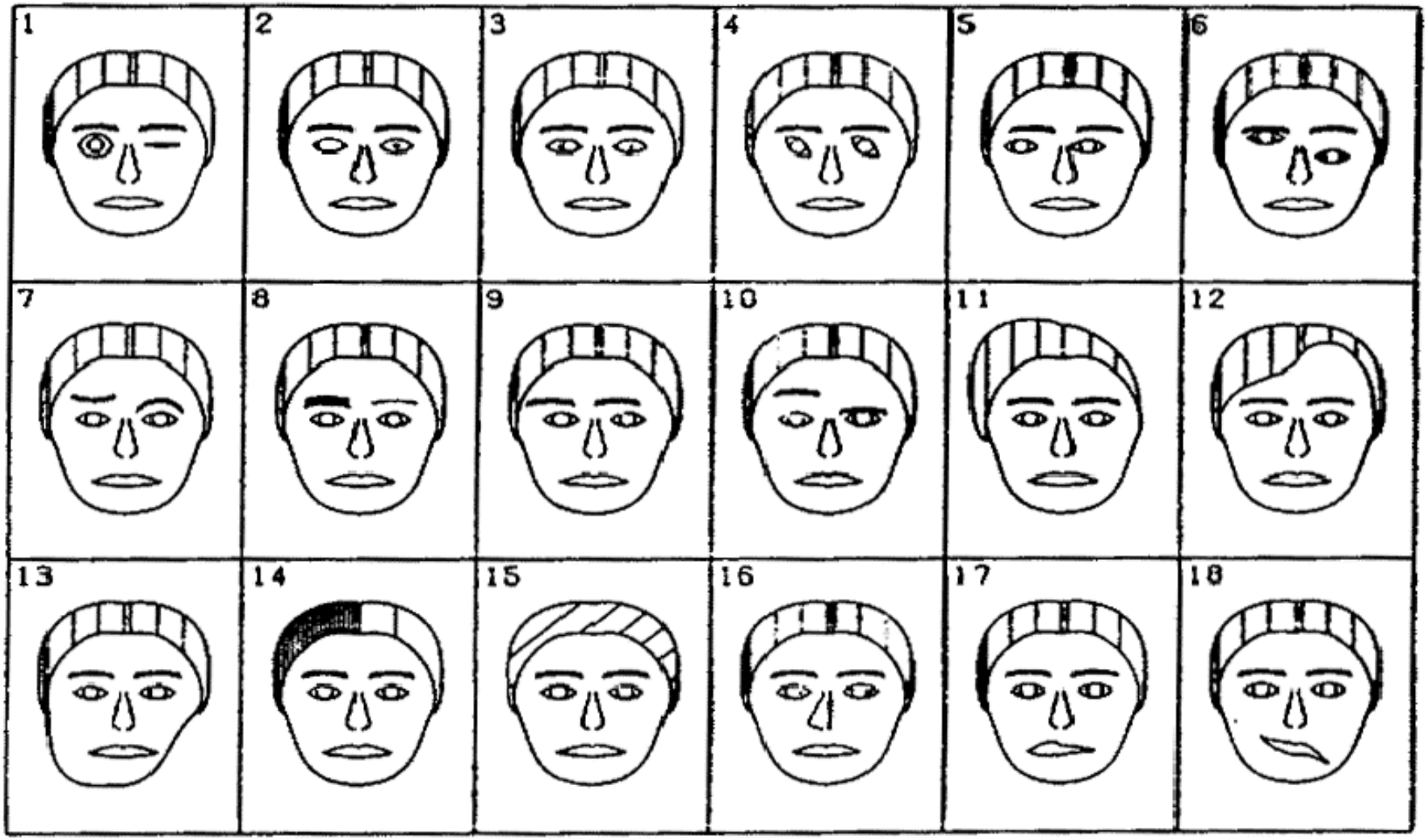




Nesprávné užití grafů - problém standardizace hodnot



Grafy zaměřené na vícerozměrné soubory dokáží zviditelnit i veliké soubory dat



3. JAK vznikají informace

Primárním důvodem analýzy dat je získání nezkreslené a přehledné **INFORMACE**

Ukázka uspořádaného datového souboru

cislo	stadium	vek	tran1_3	tran1_4	tran1_5	tran1_6	alb_pbsct	ldh_vstup	sternum	typ_myel
1	3	33	104.36	23.24	104.3	57.77	33	6.02	0.4	IgG
2	3	33	184.88	7.84	105.5	13.82	26	4.01	30	IgG
3	1	34	123.41	9.8	73.3	13.05	32	3.73	45.2	IgG
4	2	43	52.17	6.66	18.03	17.19	42	4.67	40.8	IgG
5	1	45	8.22	2.2	8.22		32	8.25	2	B-J
6	3	46	403.08		115.31		29.7	7.17	38.8	IgA
7	2	49		4.5		12.25	34	4.99	6.4	IgG
8	2	50	33.13	9.64	33.13		35	3.99	14	IgG
9	3	52	257.08	12.0					2	
10	2	53	78.33	11.3					6	
11	3	53	61.43	4.67					2	B-J
12	3	53	135.8	6.7	135.8	59.3	38		26	IgG
13	3	54	129.16	13.33	92.6	38.24	32	4.18	20	IgG
14	3	54	66.89	6.74	33.58	17.3	38	8.44	7.2	B-J
15	3	54	82.86	4.32	18.9	16.4	37	3.6	50	B-J
16	3	55	71.37	6.34	23.91	5.34	43	8.75	27	B-J
17	3	60	14.6	0.9	14.6	11.88	44	5.35	7.5	IgG
18	3	61	94.07	5.62	94.07	1.51	33	4.29	6.4	B-J
19	3	62	86.84	7.53	32.13	2.61	29	4.55	34	IgA

Primární data → Sumarizace

- v jedné skupině („one-sample“)
- ve dvou skupinách („two-sample“)
- ve více skupinách („multiple sample“)

JAK vznikají informace ? – základní pojmy

Skutečnost

Náhoda
(vybere jednu z možností pokusu)

Jev

podmnožina všech možných výsledků pokusu/děje, o které lze říct, zda nastala nebo ne

Pozorovatel

Rozliší, co nastalo
a) ***podle možností***
b) ***podle toho, jak potřebuje***

Jevové pole

třída všech jevů, které jsme se rozhodli nebo jsme schopni sledovat

Skutečnost + Jevové pole = Měřitelný prostor

Experimentální jednotka - *objekt, na kterém se provádí šetření*

Populace - *soubor experimentálních jednotek* **Znak -** *vlastnost sledovaná na objektu*

Sledovaná veličina - *číselná hodnota vyjadřující výsledek náhodného experimentu*

Znak se stává náhodnou veličinou, pokud se jeho hodnota zjišťuje vylosováním objektu ze základního souboru

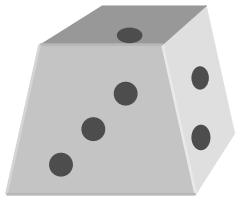
Výběr - výběrová populace - cílová populace

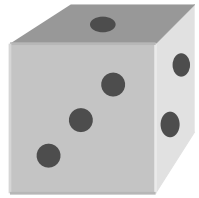
Náhodný výběr **Reprezentativnost**

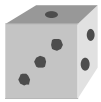
JAK vznikají informace ?

„Empirical approach“

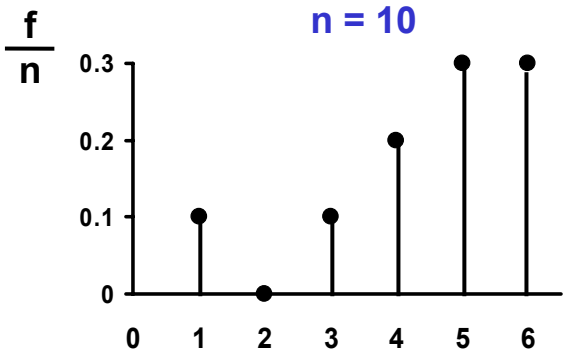
„Classical approach“

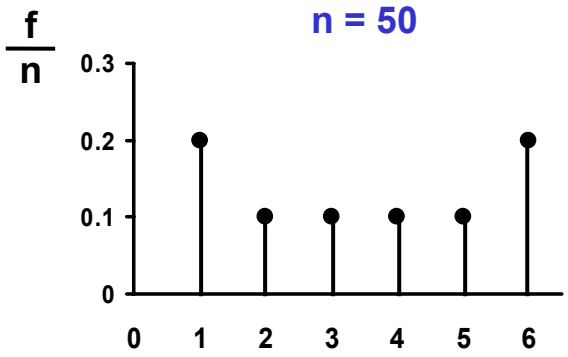


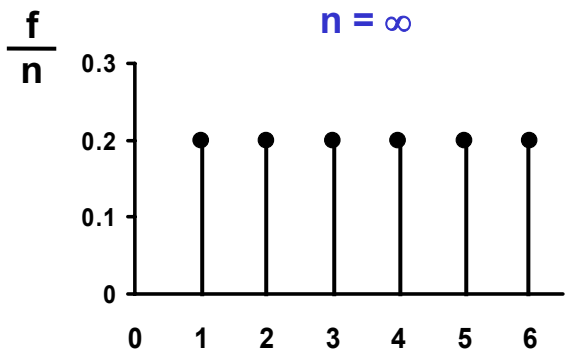




Empirický postup







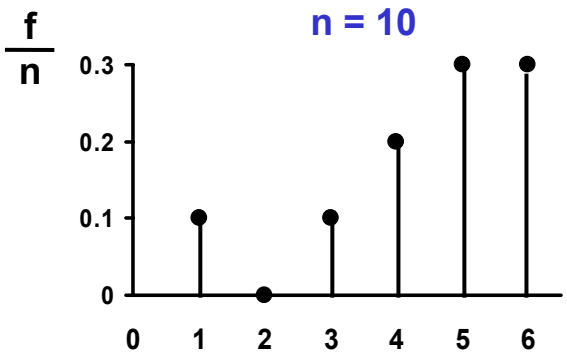
možné jevy: čísla 1 – 6

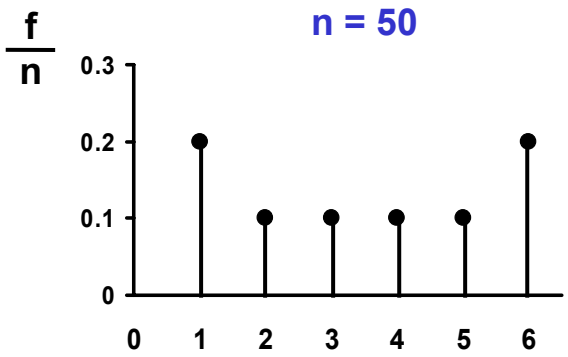
n – počet hodů (opakování)

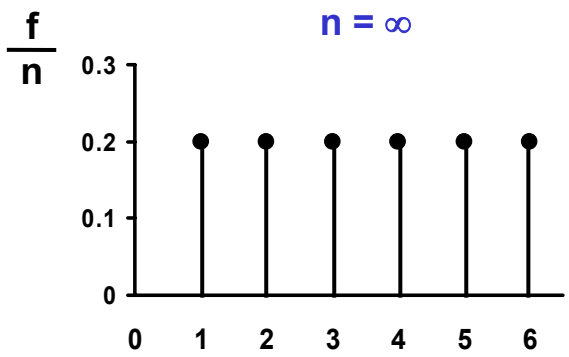
U složitých stochastických systémů se pravda získá až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit

JAK vznikají informace ?

Empirický postup







možné jevy: čísla 1 – 6

n – počet hodů (opakování)

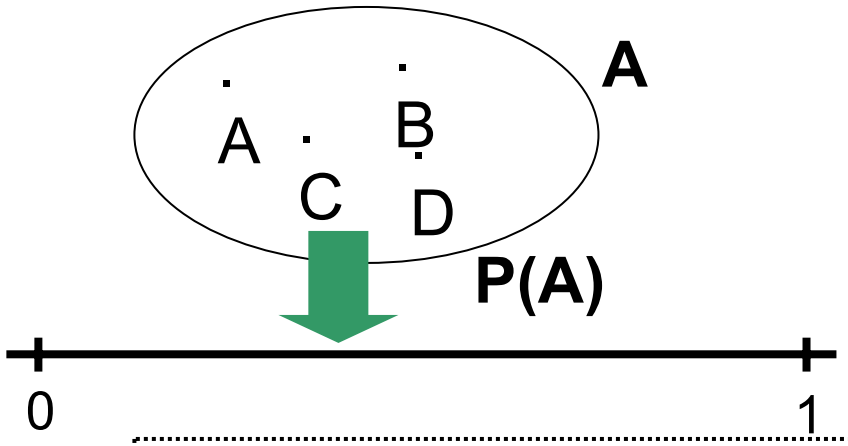
Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější) …. diskutabilní je ale ovšem míra zobecnění konkrétního experimentu

Empirický zákon velkých čísel

Při opětovné nezávislé realizaci téhož náhodného experimentu se podíl výskytů sledovaného jevu mezi všemi dosud provedenými realizacemi zpravidla ustaluje kolem konstanty.

Pravděpodobnost

je libovolná reálná funkce definovaná na jevovém poli A, která každému jevu A přiřadí nezáporné reálné číslo P(A) z intervalu 0 - 1.



Z praktického hlediska je pravděpodobnost **idealizovaná relativní četnost**

P (A) = 1 ..... jev jistý

P (A) = 0 ..... jev nemožný

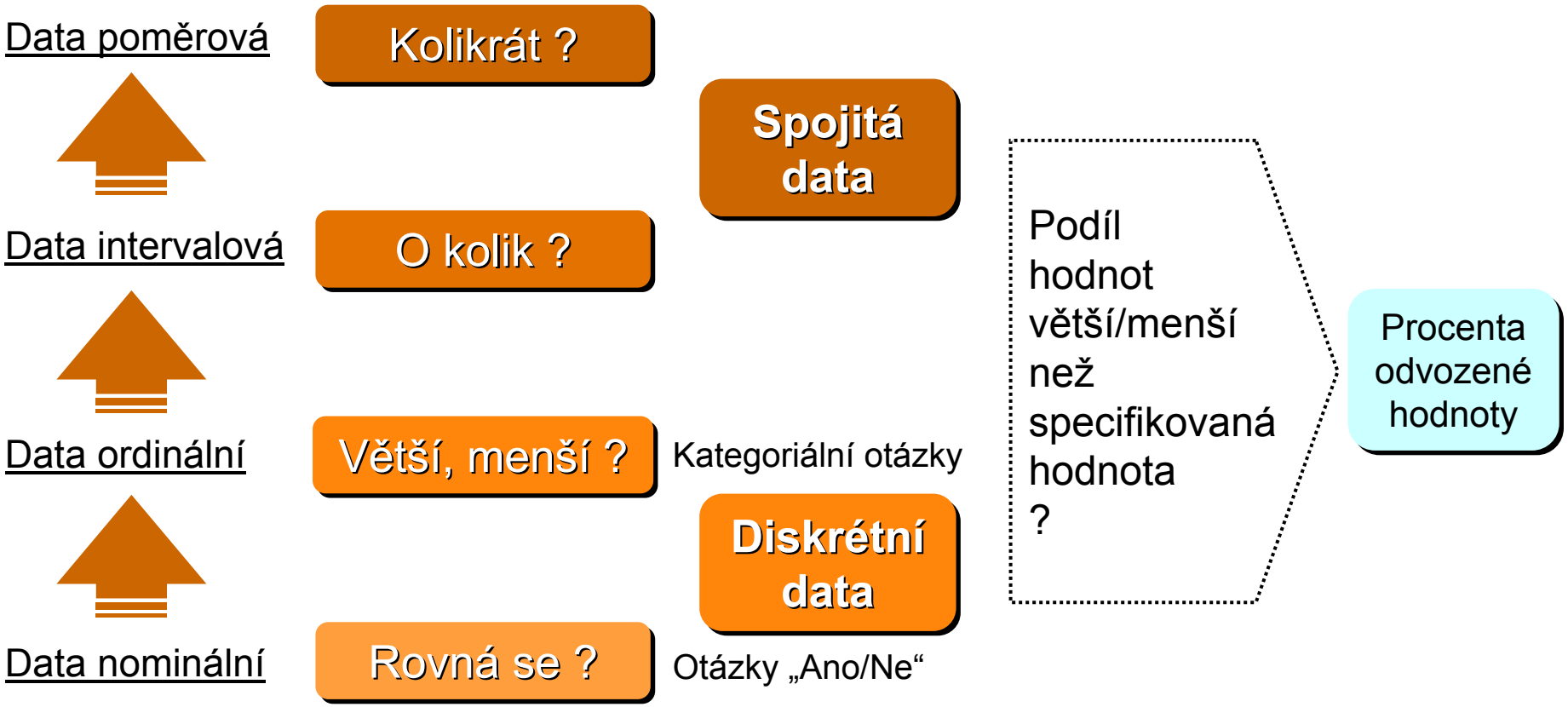
P (A ∩ B) = P (A) . P (B)..... nezávislé jevy

P (A ∩ B) = P (A) . P (B/A) .....závislé jevy

P (A / B) = P (A ∩ B) / P (B) .....podmíněná pravděpodobnost

4. Základní typy dat

Jak vznikají informace ?
– různé typy dat znamenají různou informaci



Samotná znalost typu dat ale na dosažení informace nestačí .....

Jak vznikají informace ?
– různé typy dat znamenají různou informaci

Statistika středu

Data poměrová

Data intervalová

Data ordinální

Data nominální

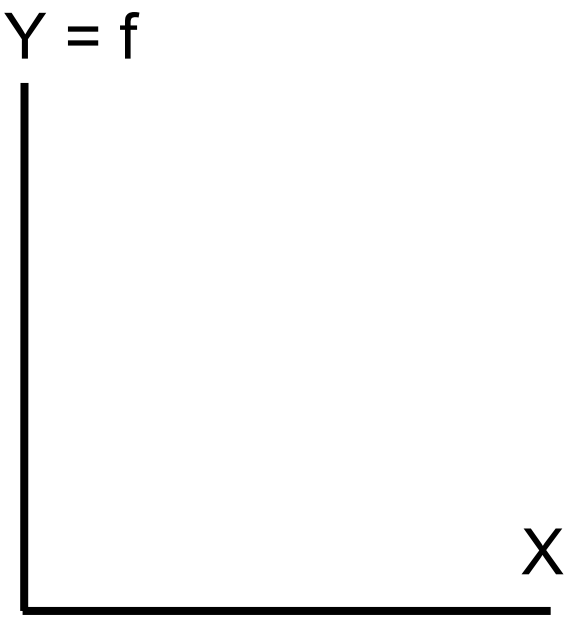
PRŮMĚR

MEDIÁN

MODUS

Spojitá data

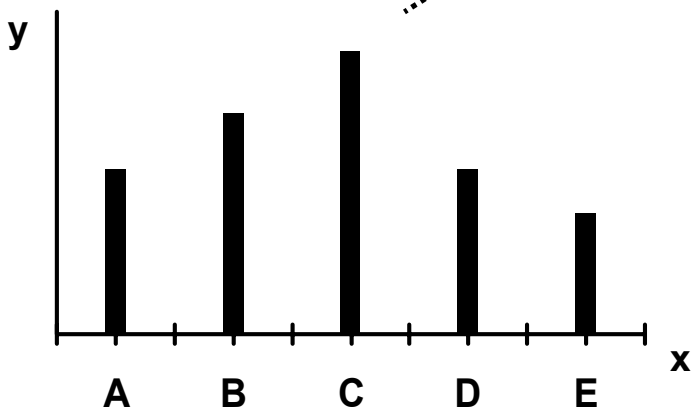
Diskrétní data

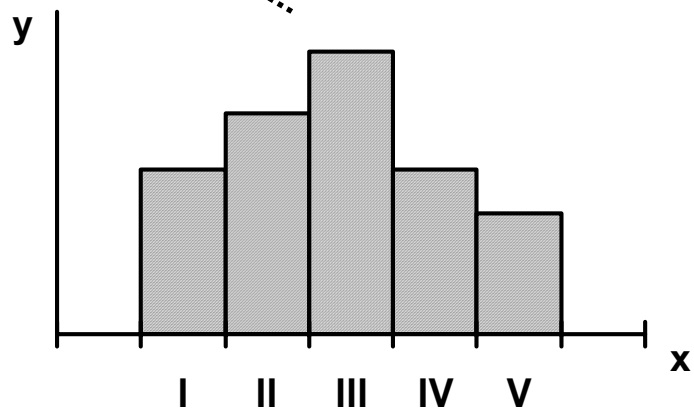


JAK vznikají informace ?
- opakovaná měření informují rozložením hodnot

Y: frekvence
- absolutní / relativní

KOLIK se naměřilo





CO se naměřilo

X: měřený znak

Diskrétní data ← → **Spojitá data**

Odvozená data

Pozor na odvozené indexy

Příklad I:

Znak X: Hmotnost
Znak Y: Plocha

Příklad II:

X: Průměrný počet výrobků v prodejně
Y: Odhad prostoru průměrně nabízeného k vystavení výrobku

	průměr :	(min - max)	
X:	1,2 :	(1,15 - 1,24)	→ **+ / - 3,8 %**
Y:	1,8 :	(1,75 - 1,84)	→ **+ / - 2,5 %**

$$X/Y = 0{,}667 : \left(\frac{1{,}15}{1{,}84} - \frac{1{,}24}{1{,}75}\right)$$ → **+ / - 6,2 %**

Nová veličina má jinou šířku rozpětí než ty, ze kterých je odvozená

Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

DISKRÉTNÍ DATA

Primární data

Počty epizod pro n = 100 hemofiliků

0
0
1
2
1
1
3
1
1
2
.
.
.
.
.
.
n = 100

Frekvenční sumarizace

N: 100 dětí (hemofiliků)
x: znak: počet krvácivých epizod za měsíc

x	n(x)	p(x)	N(x)	F(x)
0	20	0,2	20	0,2
1	10	0,1	30	0,3
2	30	0,3	60	0,6
3	40	0,4	100	1,0

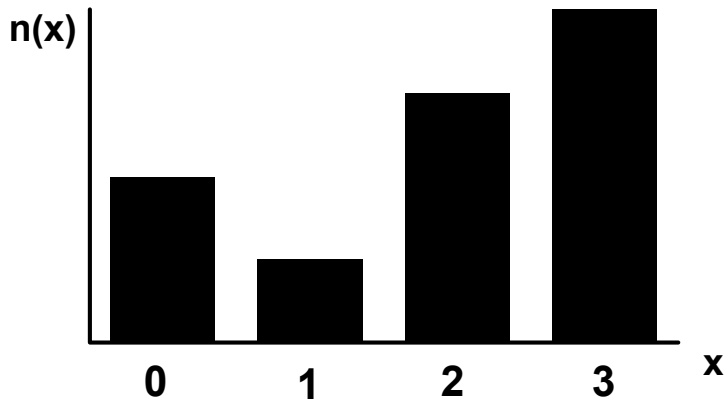
n(x) – absolutní četnost x
p(x) – relativní četnost; p(x) = n(x) / n
N(x) – kumulativní četnost hodnot nepřevyšujících x;

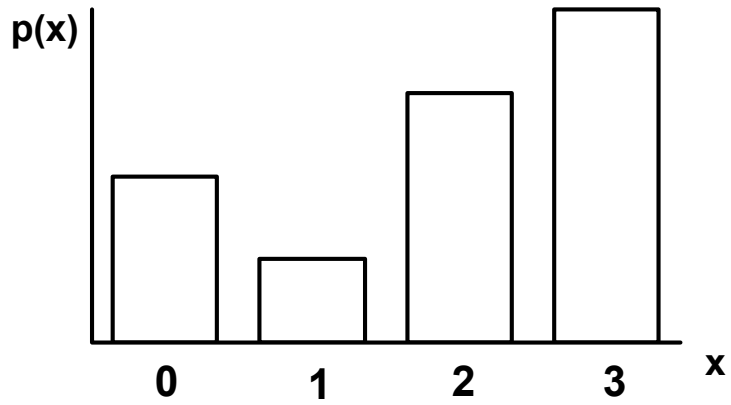
$$N(x) = \sum_{t \le x} n(t)$$

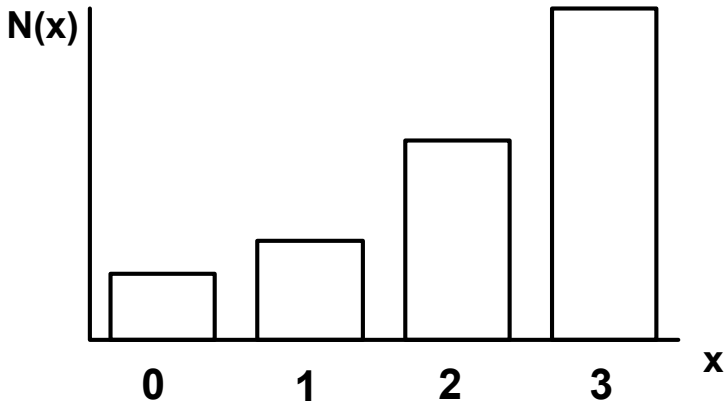
F(x) – kumulativní relativní četnost hodnot nepřevyšujících x; F(x) = N(x) / n

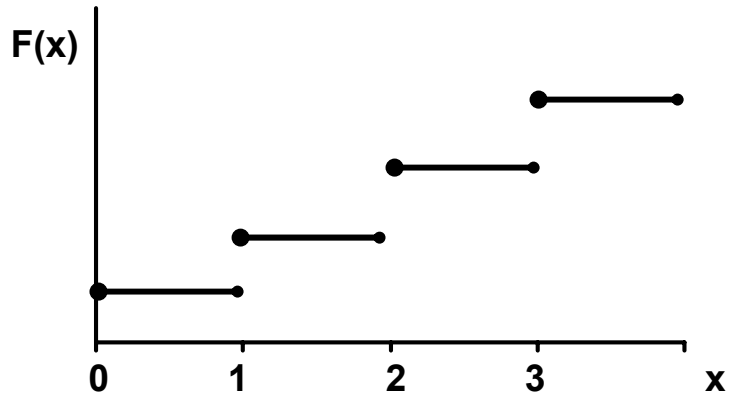
Jak vznikají informace ?
- frekvenční sumarizace diskrétních dat

Grafické výstupy z frekvenční tabulky









Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

SPOJITÁ DATA

Příklad: **x: koncentrace látky v krvi n = 100 pacientů**

Primární data

Hodnoty pro n = 100 osob

1,21
1,48
1,56
0,31
1,21
1,33
0,33
0,21
1,32
1,11
.
.
.
.
.
.
n = 100

Frekvenční sumarizace

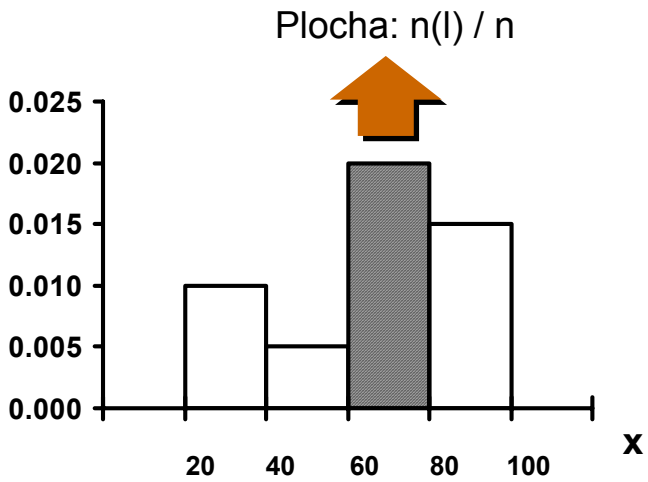
n = 100 opakovaných měření (100 pacientů)
x: koncentrace sledované látky v krvi (20 – 100 jednotek)

interv	d(l)	n(l)	n(l)/n	N(x'')	F(x'')
<20, 40)	20	20	0,2	20	0,2
<40, 60)	20	10	0,1	30	0,3
<60, 80)	20	40	0,4	70	0,7
<80, 100)	20	30	0,3	100	1,0

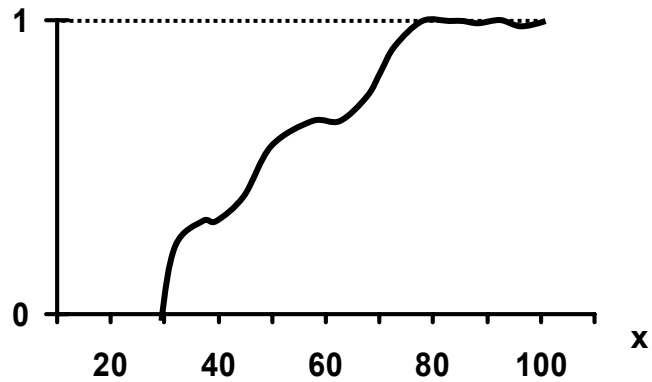
d(l) – šířka intervalu
n(l) – absolutní četnost
n(l) / n – intervalová relativní četnost
N(x'') – intervalová kumulativní četnost do horní hranice X''
F(x'') – intervalová relativní kumulativní četnost do horní hranice X''

Jak vznikají informace ?
- frekvenční sumarizace spojitých dat

Histogram



Výběrová distribuční funkce



$$f(x) = \frac{n(l) / n}{d(l)}$$ ⇒ **Intervalová hustota četnosti**

F(x) ⇒ **Intervalová relativní kumulativní četnost**

Histogram = standardní nástroj zviditelnění spojitých dat

1. Data X: 14,1; 8,4; 12,1; 18,2; 20,4; … … … … n
2. Setřídění dat podle velikosti
3. Kategorizace hodnot X - vytvoření intervalů
4. Frekvenční tabulka
5. Histogram

"Absolute frequency histogram"

$$f(x) = \frac{n(I)}{d(I)}$$

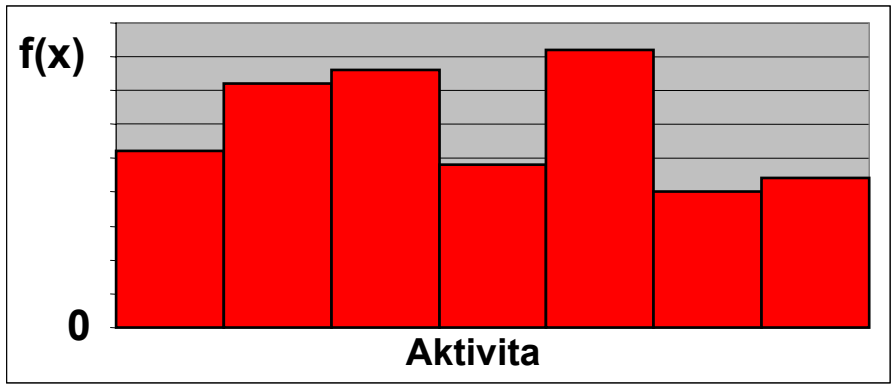
"Relative frequency histogram"

$$f(x) = \frac{n(I) / n}{d(I)}$$

Spojitá data – postup frekvenčních analýz

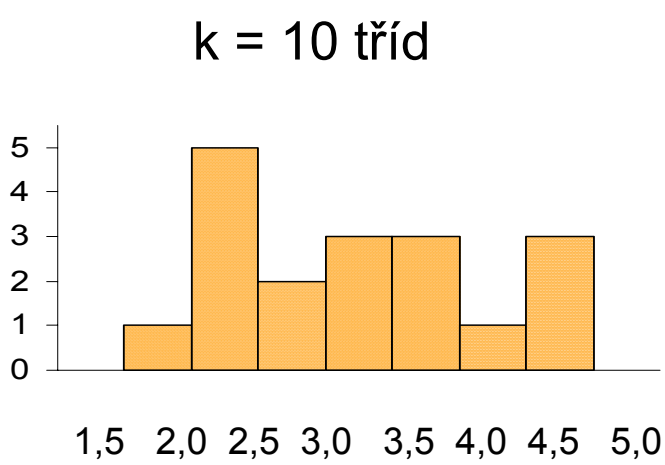
Aktivita enzymu (X)

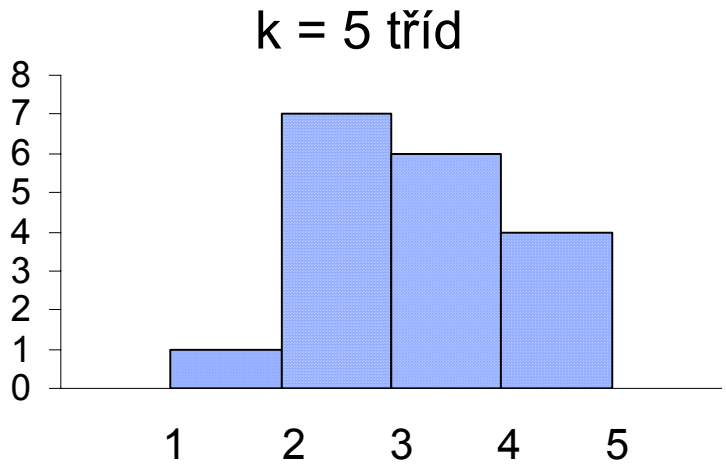
- I. Utřídit podle velikosti
- II. Rozdělit do intervalů o rozumné šířce
- III. Vyhodnotit počty hodnot v jednotlivých intervalech - absolutní četnosti
- IV. Vyhodnotit podíly (relativní četnosti) hodnot v jednotlivých intervalech
- V. Grafické znázornění - histogram

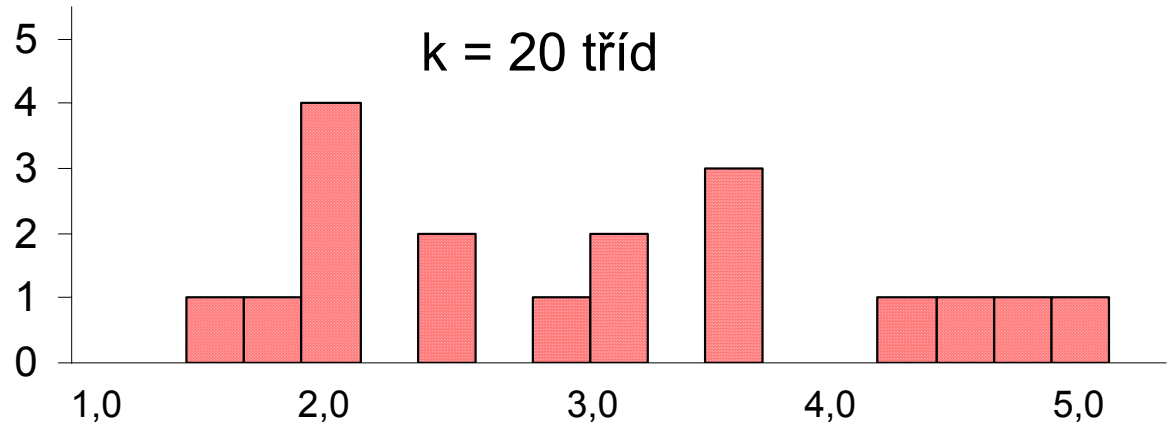


Počet intervalů X: dán daty a hodnotitelem
Šířka intervalů: pokud možno stejná

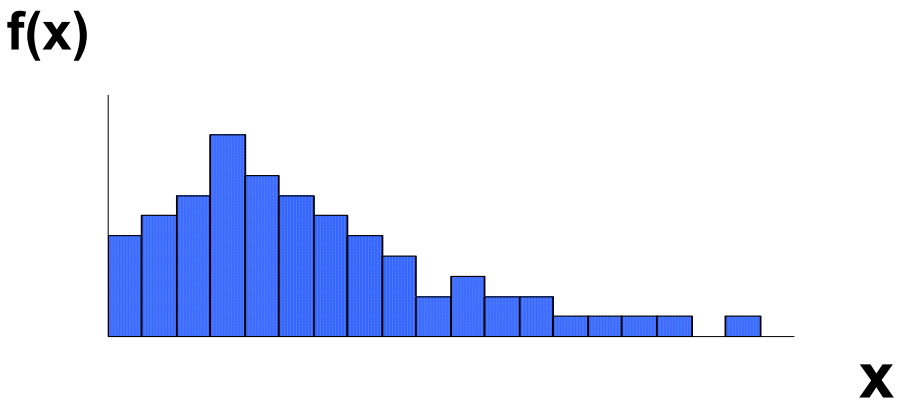
Počet zvolených tříd a velikost souboru určují kvalitu výstupu

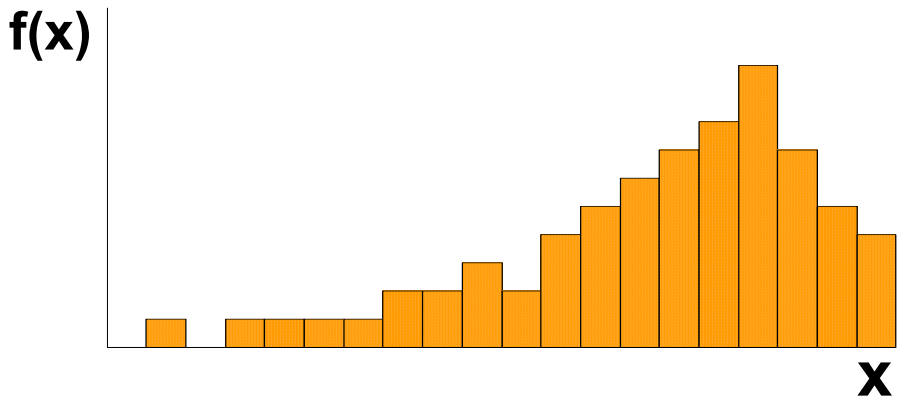


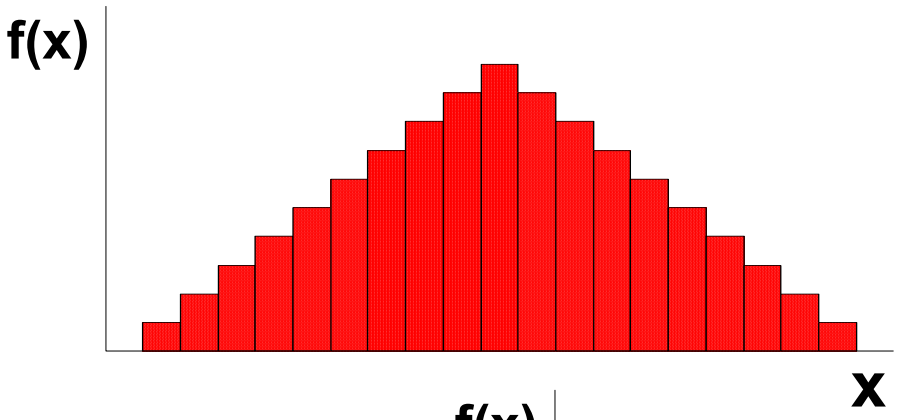


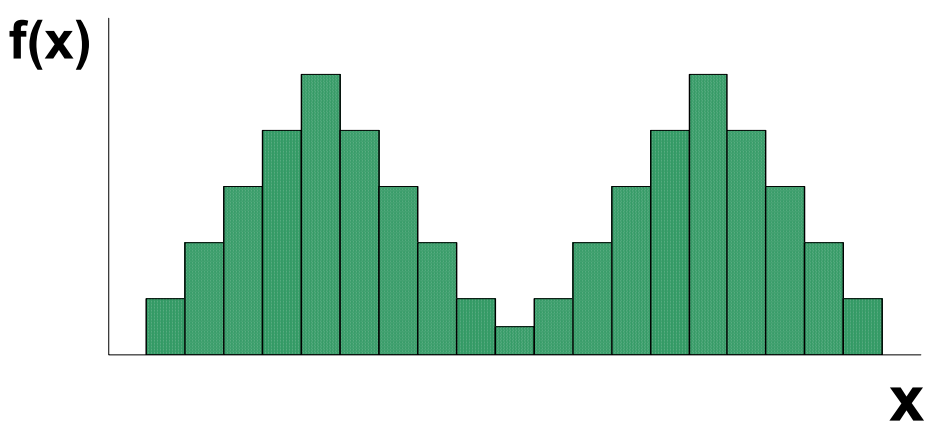


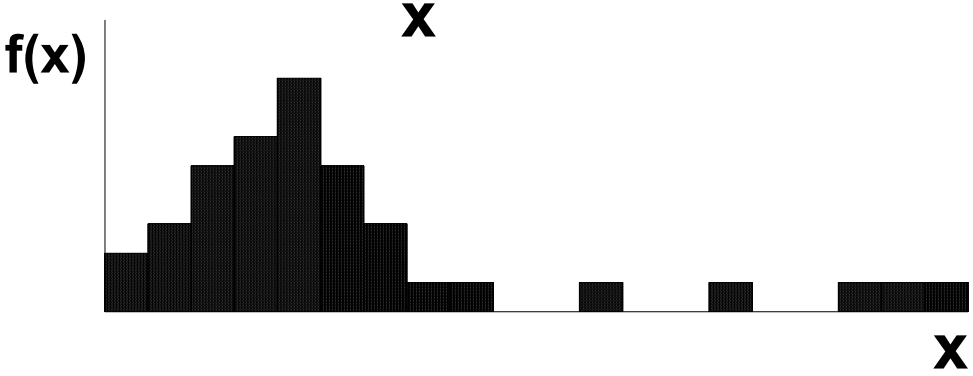
Histogram vyjadřuje tvar výběrového rozložení



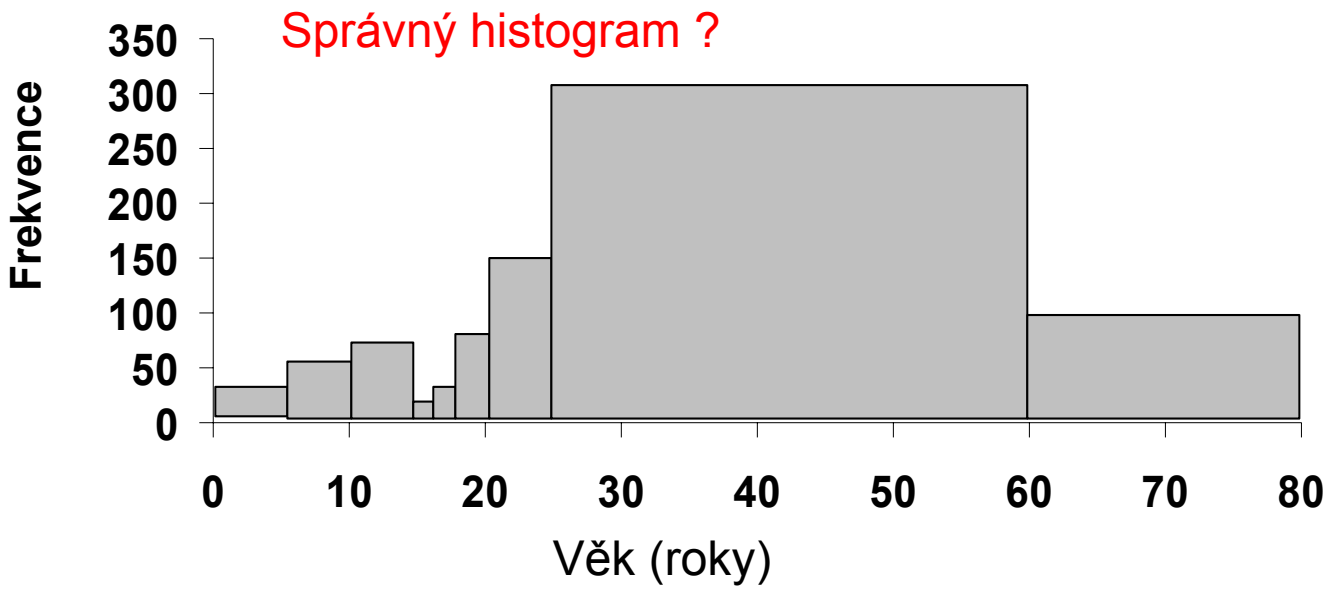




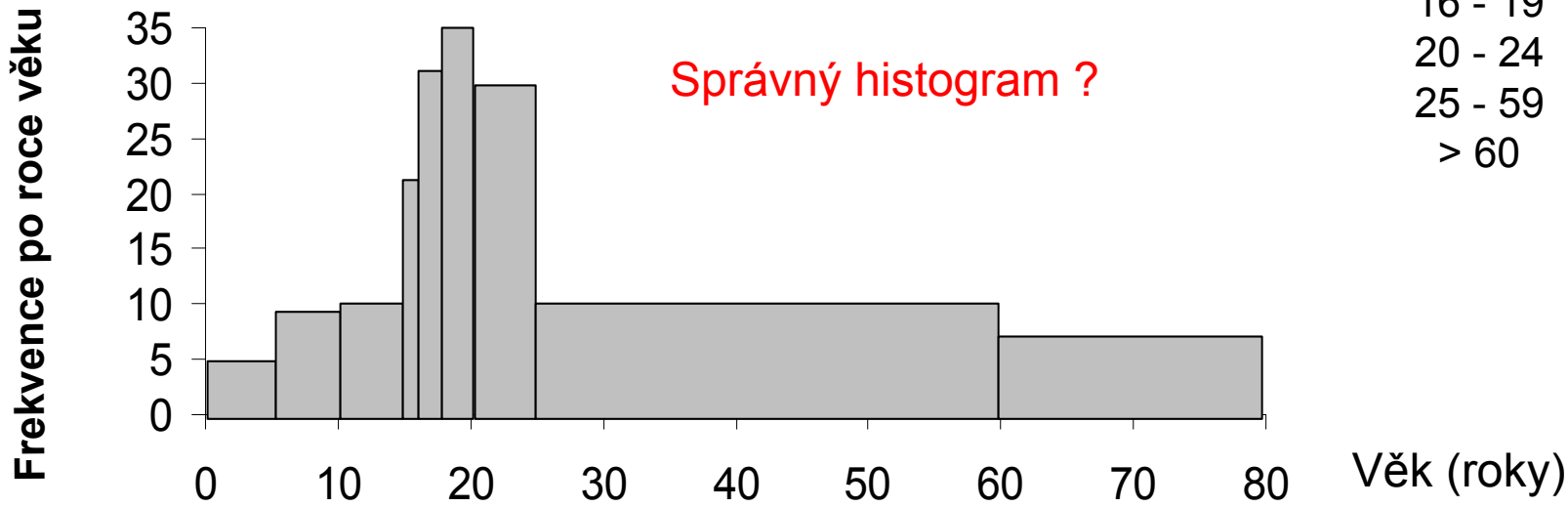




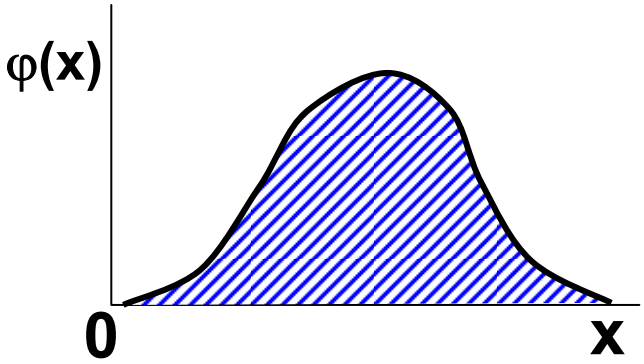
Příklad: věk účastníků vážných dopravních nehod



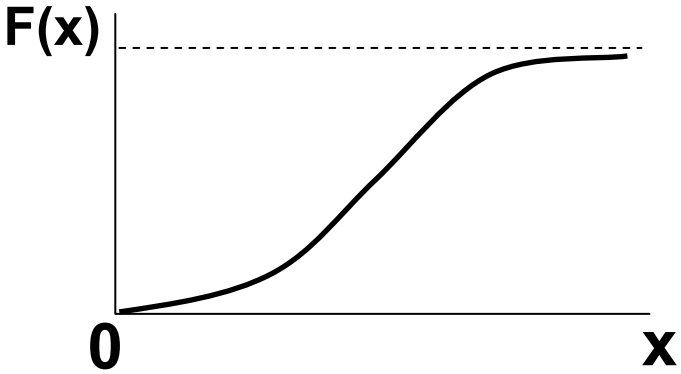
Věk	f
0 - 4	28
5 - 9	46
10 - 15	58
16 - 19	20
20 - 24	114
25 - 59	316
> 60	103



Pojem ROZLOŽENÍ - příklad spojitých dat



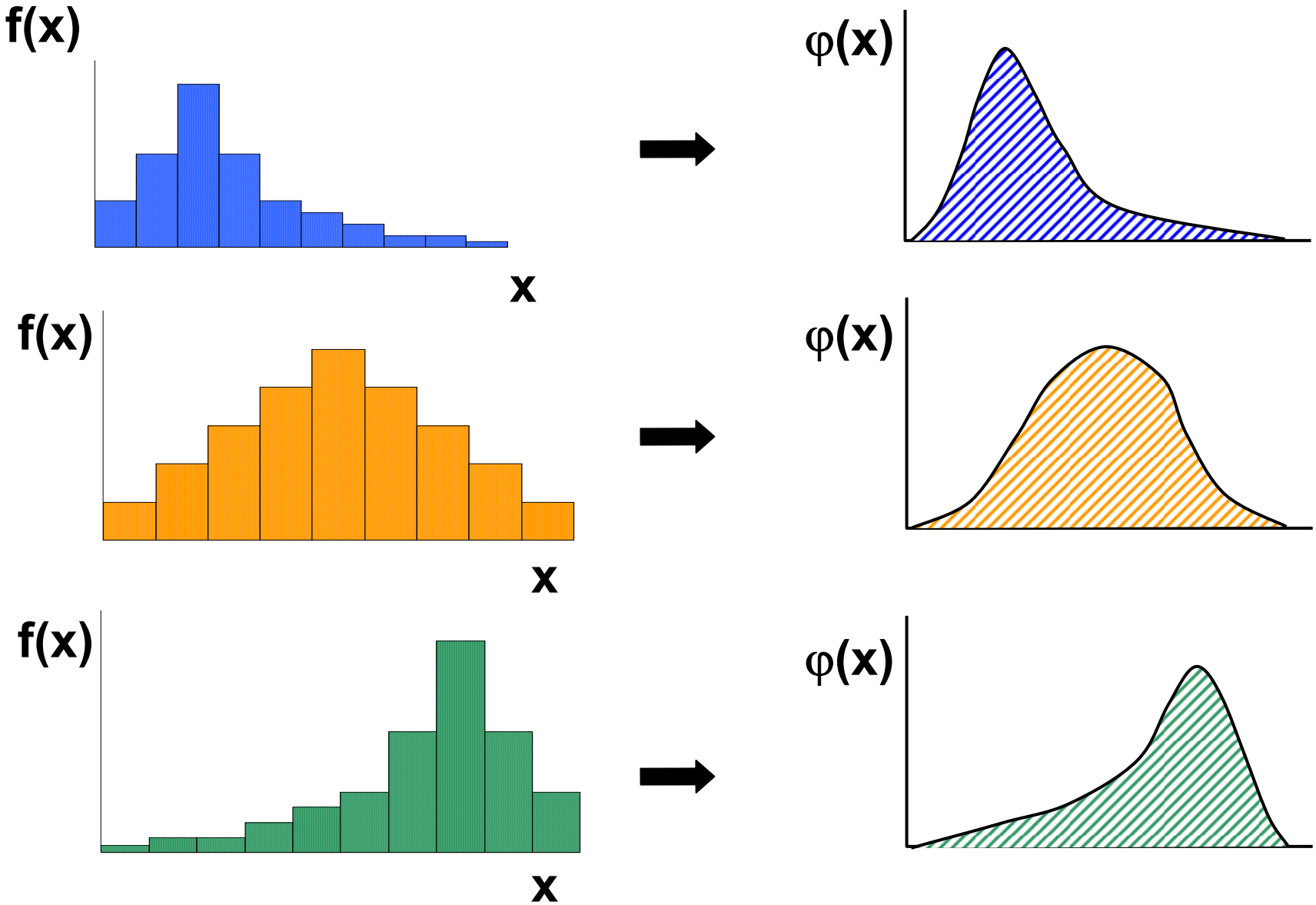
Rozložení



Distribuční funkce

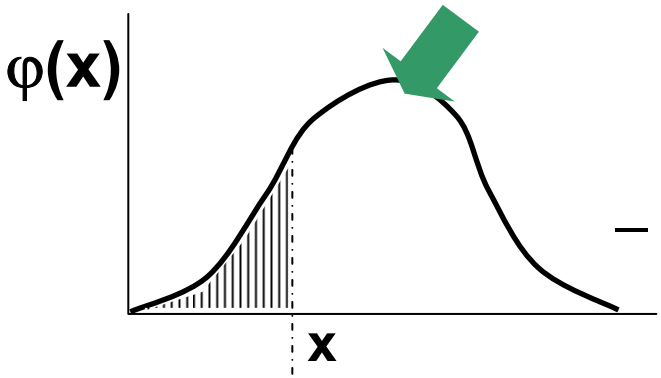
Je - li dána distribuční funkce, je dáno rozložení

Výběrové rozložení hodnot lze modelově popsat a definovat tak pravděpodobnost výskytu X



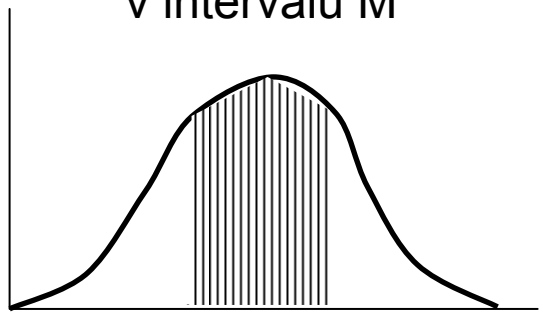
Distribuční funkce jako užitečný nástroj pro práci s rozložením

Plocha = relativní četnost



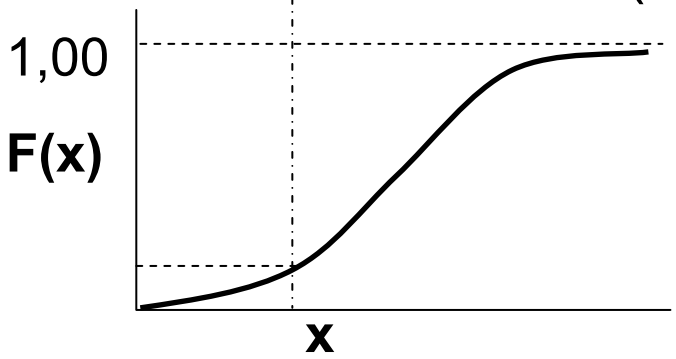
$$\int_{-\infty}^{\infty} \varphi(x)\ d(x) = 1$$

F(x):
Pravděpodobnost, že se X vyskytuje v intervalu M



M

$$P(X \leq x) = \Phi(x) = F(x'')$$



Φ(x) … distribuční funkce

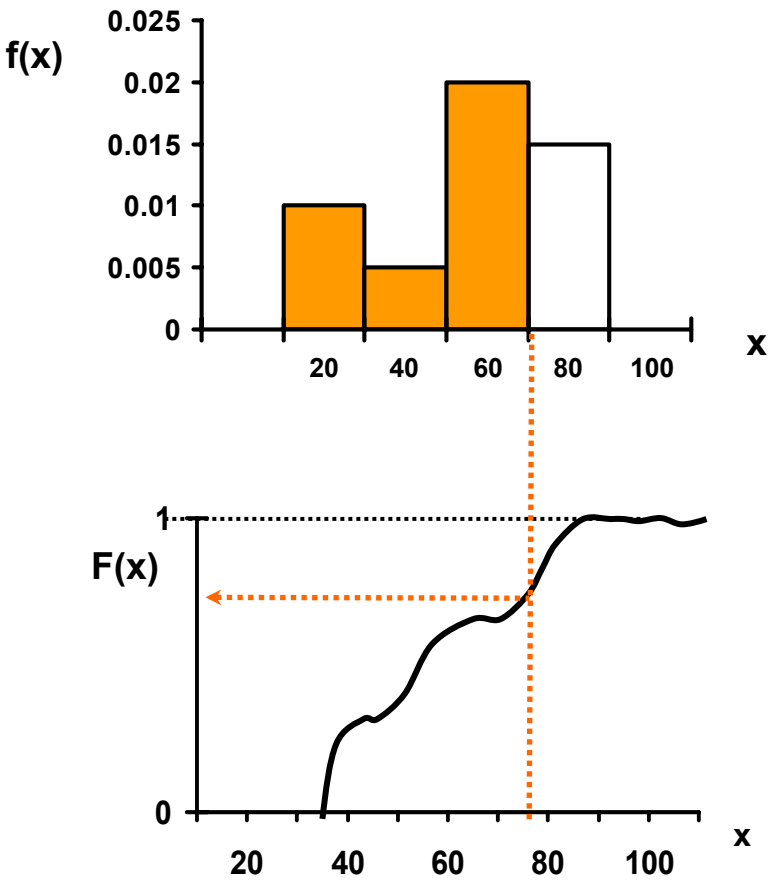
$$P(X \leq x) = \int_{M} \varphi(x)\ d(x)$$

Známe-li distribuční funkci, pak známe rozložení sledované veličiny.

Pro jakoukoli množinu hodnot (**M**) lze určit P, že X do této množiny patří.

Jak vznikají informace ?
- frekvenční sumarizace spojitých dat

Grafické výstupy z frekvenční tabulky – spojitá data



Uspořádání čísel podle velikosti a konstrukce rozložení umožňuje pravděpodobnostní zařazení každé jednotlivé hodnoty

KVANTIL

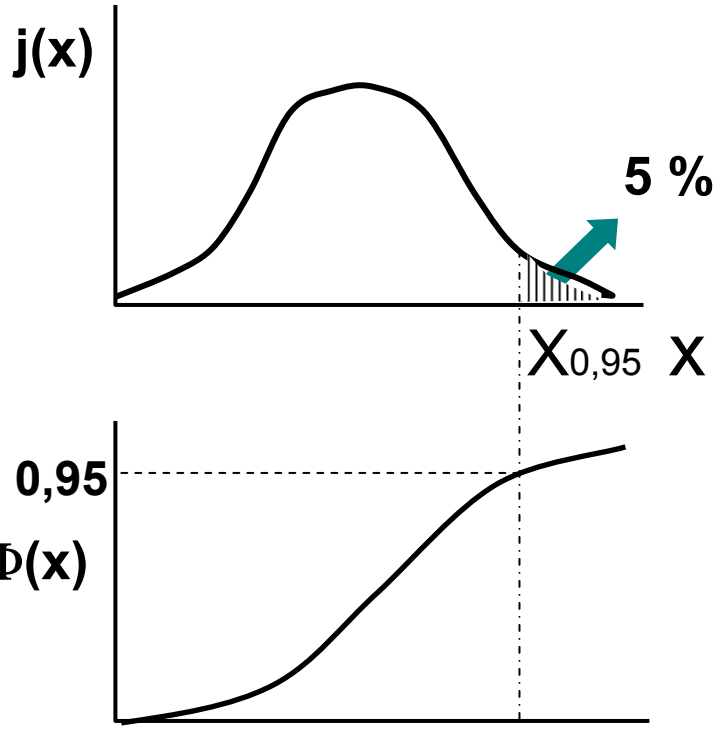
$X_{0.1}$; $X_{0.9}$; $X_{0.5}$; X_{θ}

Otázka: Jak velké musí být X, aby 5 % všech hodnot bylo nad ním?

θ = 0,95 … Pravděpodobnost

Hledáme: $P(X \leq x_\theta) = 0{,}95 = \theta$

$x_\theta = (X_{0,95}) = ?$



$F(x\theta) = \theta$

Kvantil je číslo, jehož hodnota distribuční funkce je rovna P, pro kterou je kvantil definován

Jakékoliv číslo na ose x je kvantilem