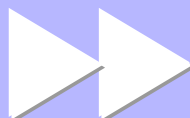




## Biostatistika



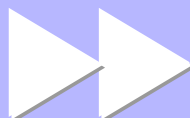
# Základní biostatistika

---

- Pokrytí základních oblastí analýzy biologických dat:
  - 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



## 1. Data a informace



# BIOSTATISTIKA - BIOMETRIKA

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

Malá data

Velká data

Obrovská data



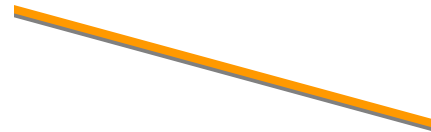
Umění  
prodat



Umění  
pochopit



Umění  
uchopit



# DATA – ukázka uspořádání datového souboru

## Parametry (znaky)

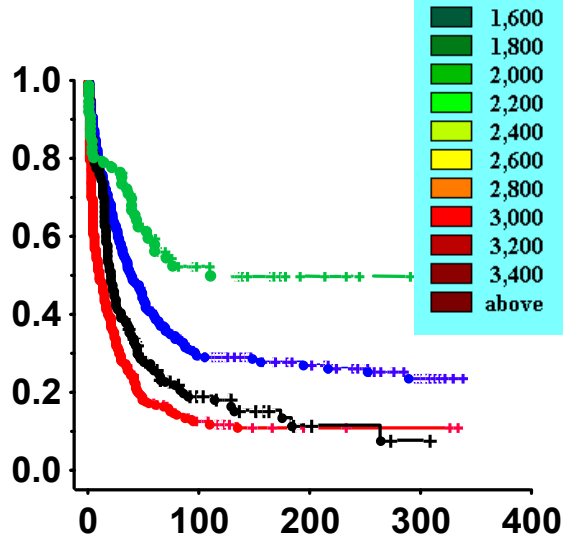
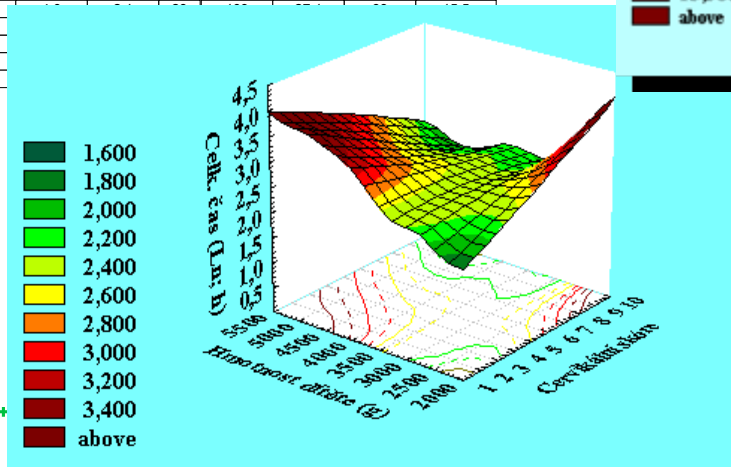
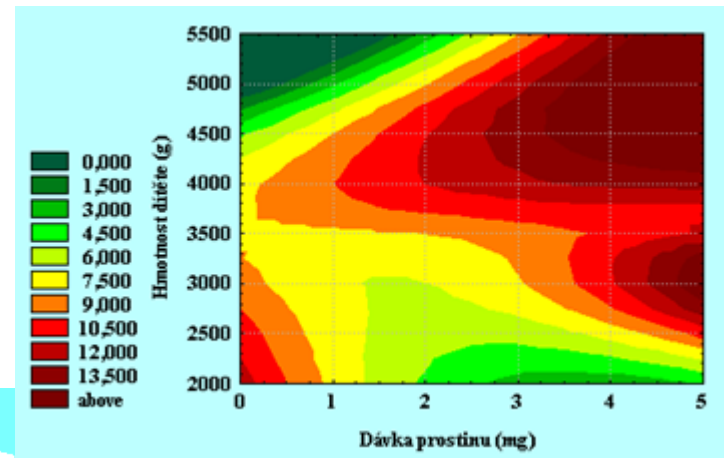
Opakování

Pacient	Clovek	aLeu	aTy%	aSe%	aNeu%	aLy%	aTy	aSe	aNeu	aLy	aHtc	aCLsk	aCLNeus	aCLOZ	aCLNeuO
		cell.10 <sup>6</sup> /	%	%	%	%	cell.10 <sup>6</sup> /	cell.10 <sup>6</sup> /	cell.10 <sup>6</sup> /	cell.10 <sup>6</sup> /	%	mV.s.10 <sup>3</sup>	mV.s.10 <sup>3</sup>	mV.s.10 <sup>3</sup>	mV.s.10 <sup>3</sup>
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 cell.10 <sup>6</sup> /	aTy% %	aSe% %	aNeu% %	aLy% %	aTy cell.10 <sup>6</sup> /	aSe cell.10 <sup>6</sup> /	aNeu cell.10 <sup>6</sup> /	aLy cell.10 <sup>6</sup> /	aHtc %	aCLsk mV.s.10 <sup>3</sup>	aCLNeus mV.s.10 <sup>3</sup>	aCLOZ mV.s.10 <sup>3</sup>	aCLNeuO mV.s.10 <sup>3</sup>
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

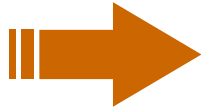
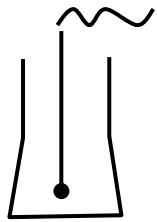


Schopnost: vidět data – komunikovat  
– interpretovat - prodávat

# 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



165 cm



140 cm



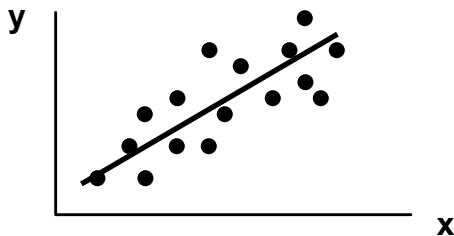
182 cm



163 cm

rozptyl znaku, přirozená variabilita

### Variabilita modelovaných dat



chyba = nepřesnost modelu

### Variabilita časových řad



čas

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*

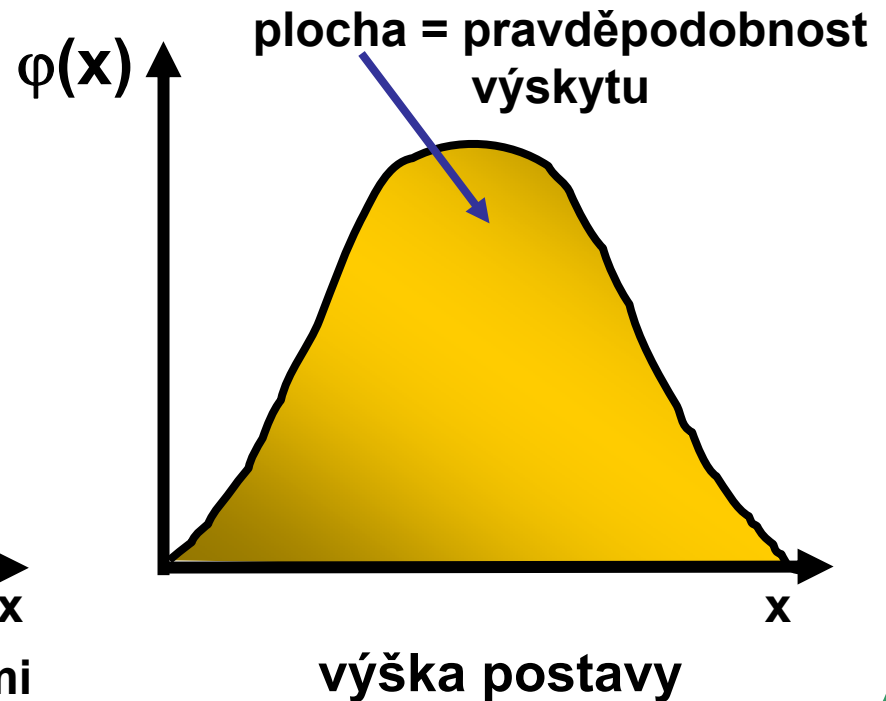
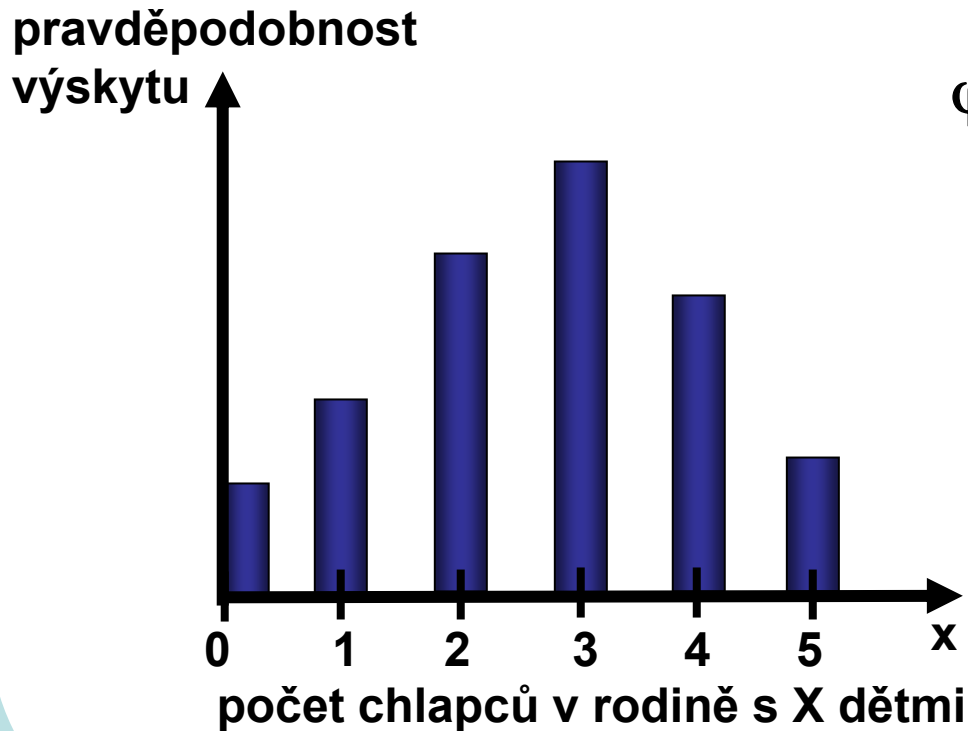
*Přímé využití k predikcím  
chování systému*



# 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, ....)**



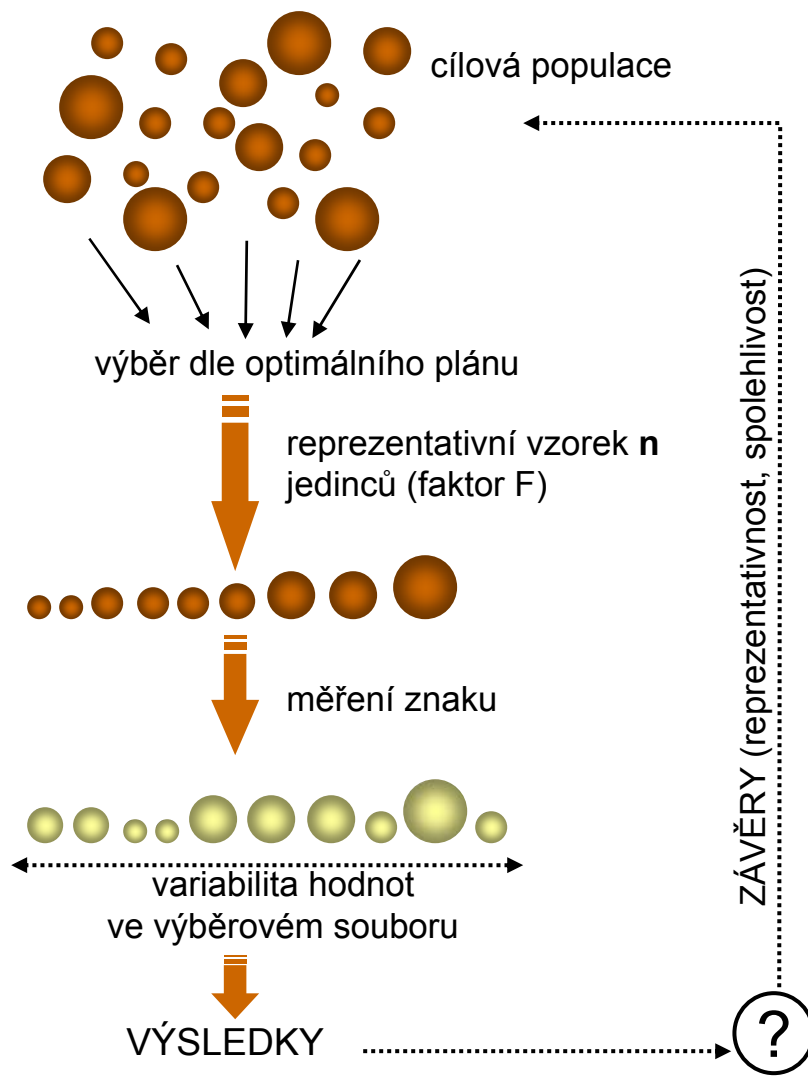
**Analýza časových řad, analýzy trendů**



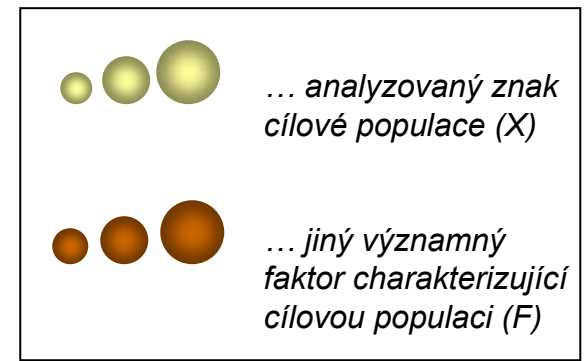
**Analýza biomedicínských dat**

# Experimentální design: nezbytná výbava

Účel analýzy:  
Popisný

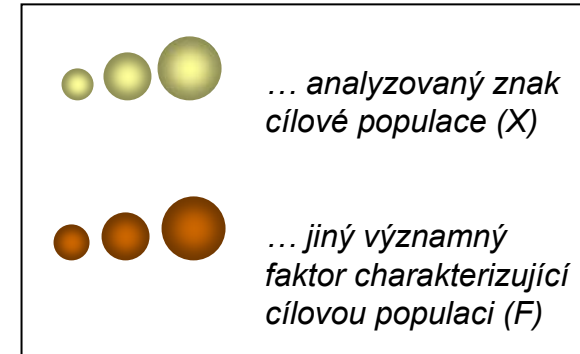
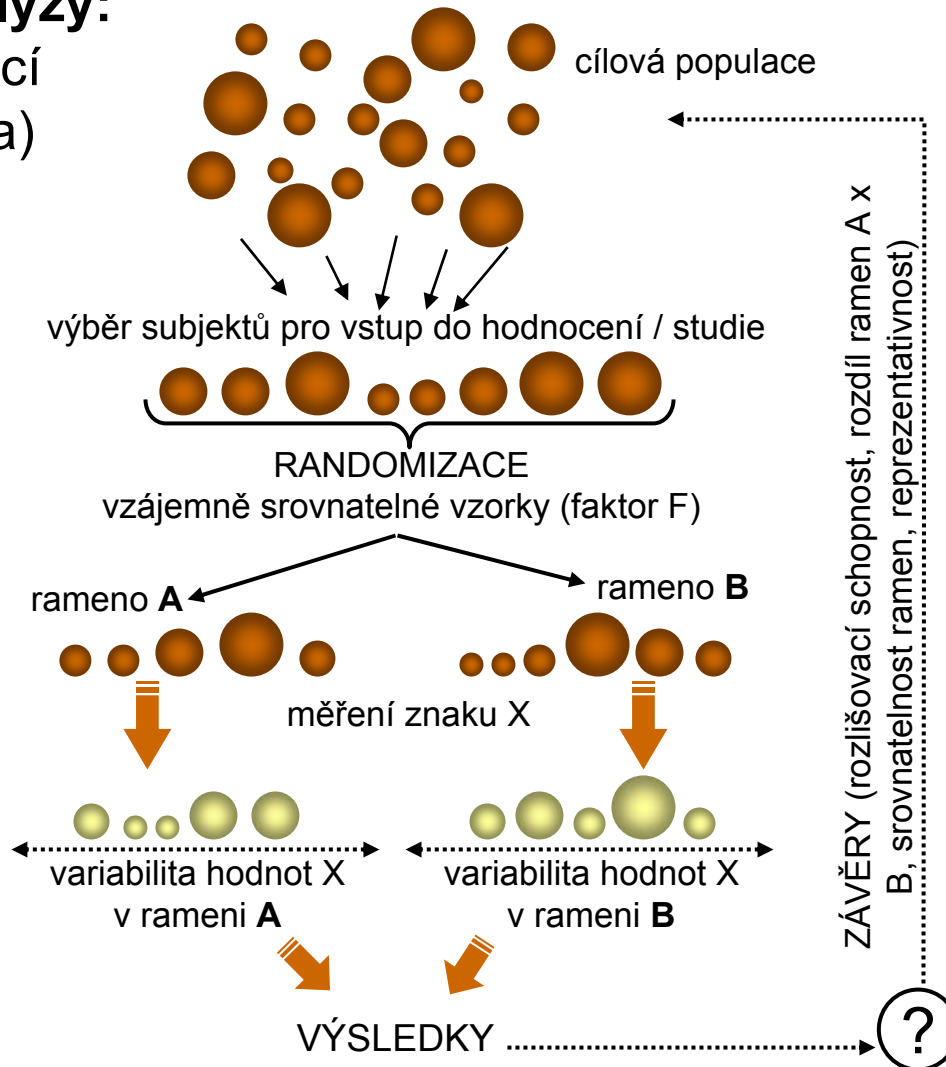


**?**  
**Reprezentativnost**  
**Spolehlivost**  
**Přesnost**



# Experimentální design: nezbytná výbava

Účel analýzy:  
Srovnávací  
(2 ramena)



**?**  
**Srovnatelnost**  
**Spolehlivost**  
**Přesnost**

# Stochastické modelování: predikce neurčitých jevů

✦ Prospektivně – modelově - postihuje chování jevů při respektování variability

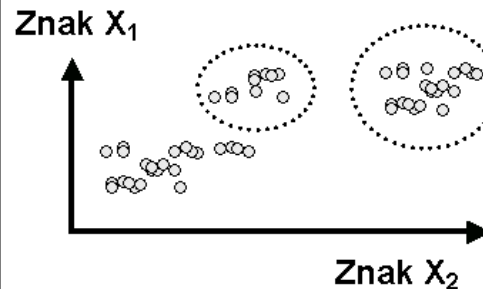
## Pravděpodobnostní vztahy

Anamnéza x Výsledek vyšetření pacienta

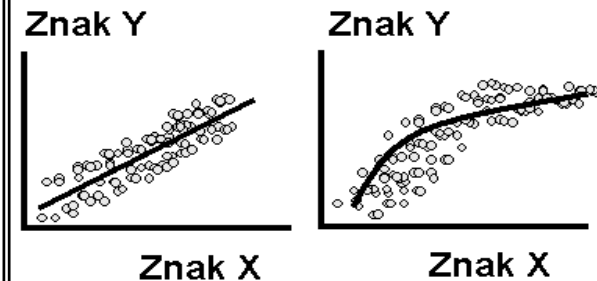
	Karcinom	Benigní léze	Benigní riziková	Zdravá	
Pozitivní anamnéza	2,22	34,44	0,00	63,33	100%
Negativní anamnéza	1,06	28,23	0,96	69,75	100%

$p < 0.05$

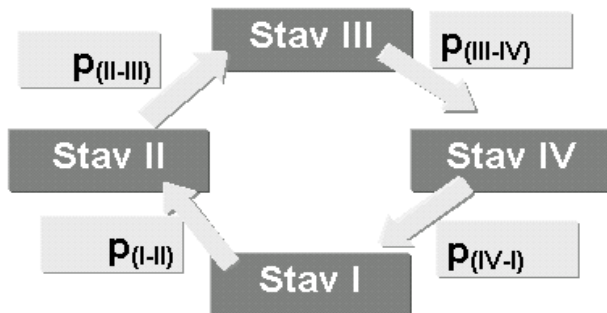
## Vícerozměrná diskriminace



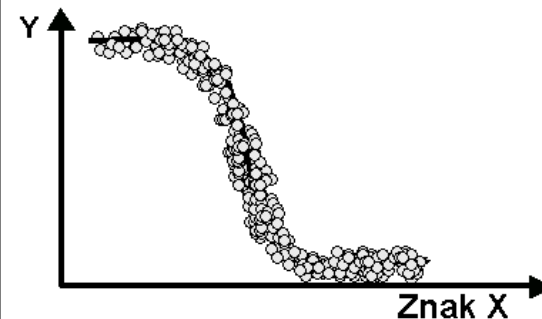
## Funkční vztahy znaků



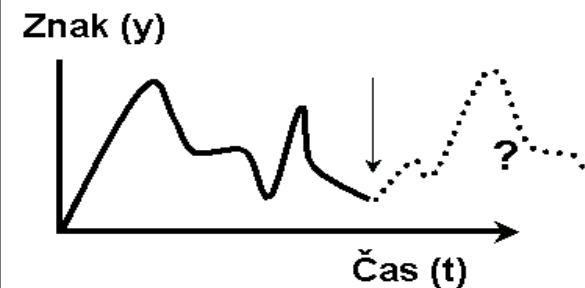
## Markovovy řetězce



## Logistické modely

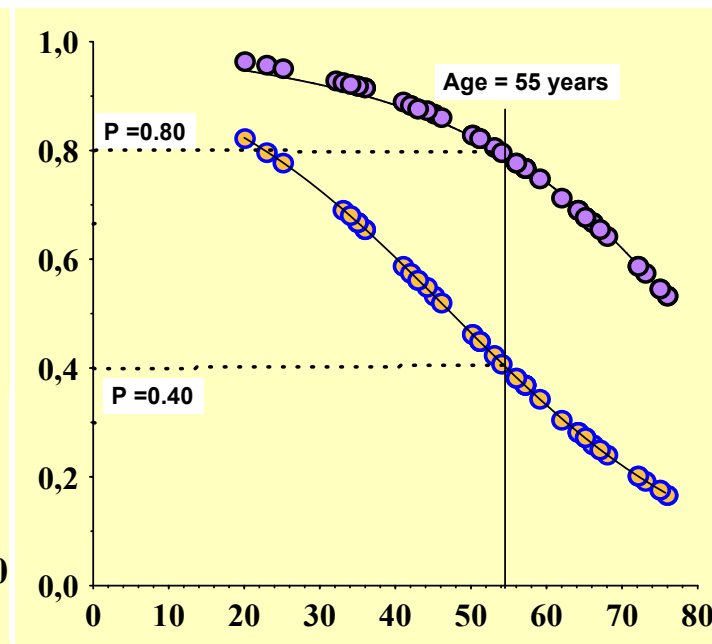
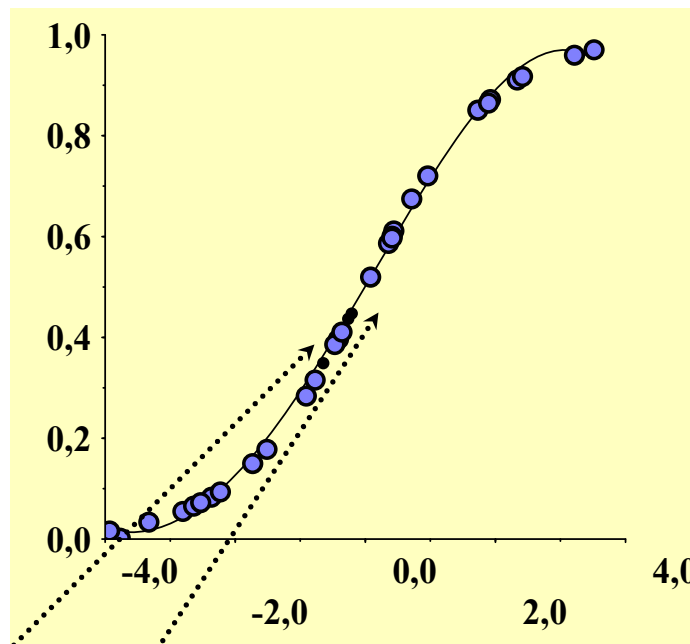


## Chování systému v čase



# Stochastické modelování: predikce neurčitých jevů

**Osa Y**  
**Predikovaná**  
**pravděpodobnost**



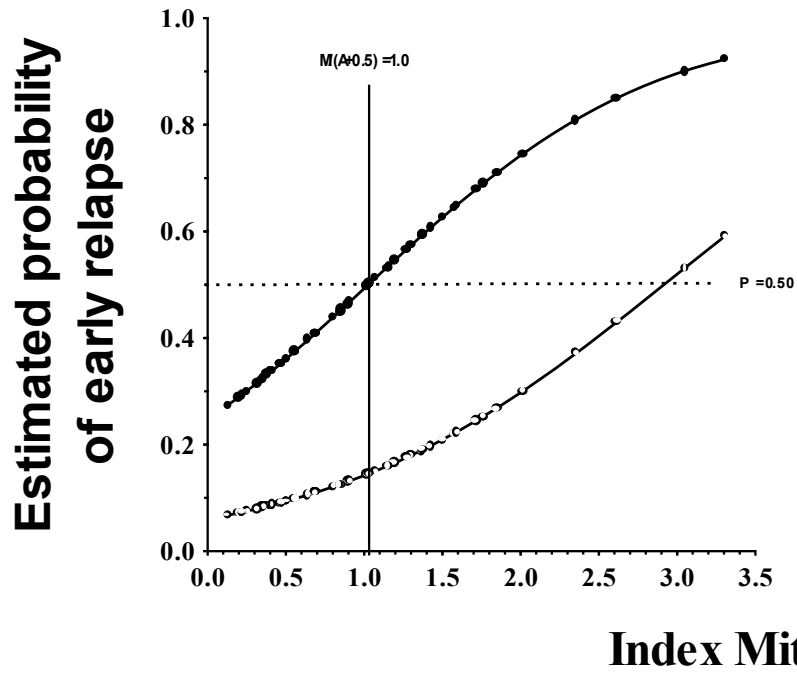
**Osa X**  
**Parametr nebo kombinace parametrů**

Data konkrétních pacientů (subjektů)  
k přímému hodnocení

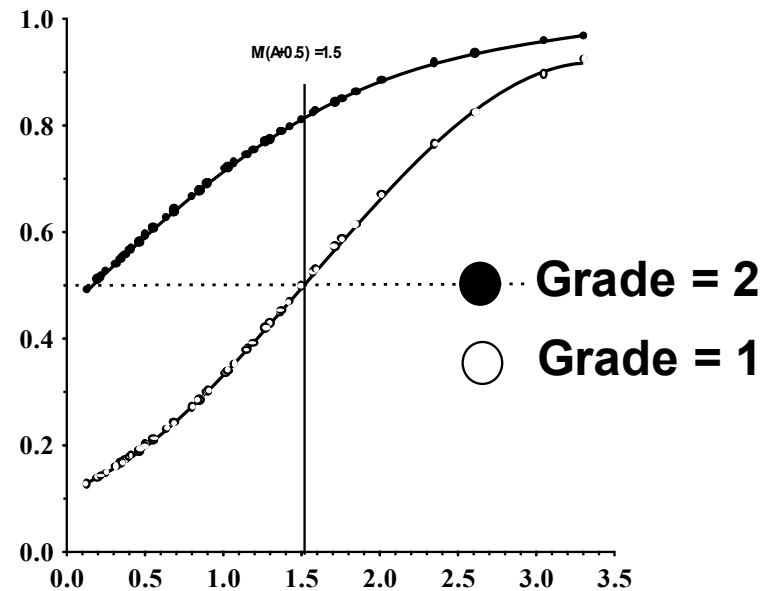
# Pravděpodobnostní prediktivní modely

## Maligní lymfomy: Pravděpodobnost časného relapsu

### Stádium I - II

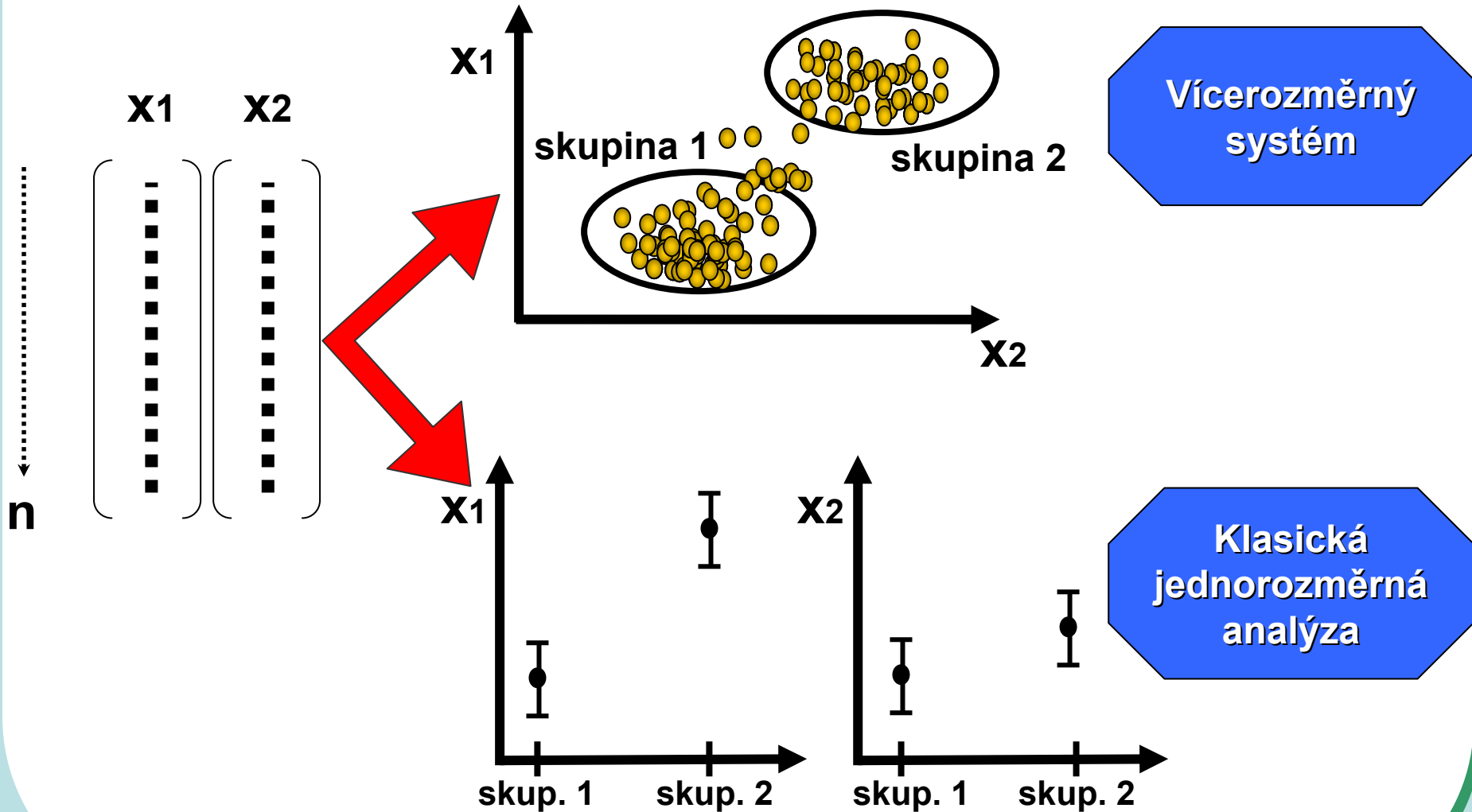


### Stádium III - IV



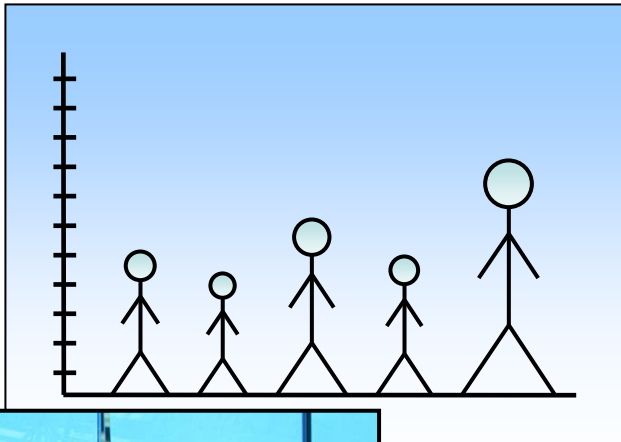
Schopnost: vytvářet prakticky využitelné nástroje

# Vícerozměrné vnímání skutečnosti – nová kvalita analýzy dat





# Biologové analýzou dat proti variabilitě nebojují !



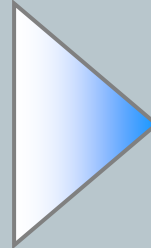
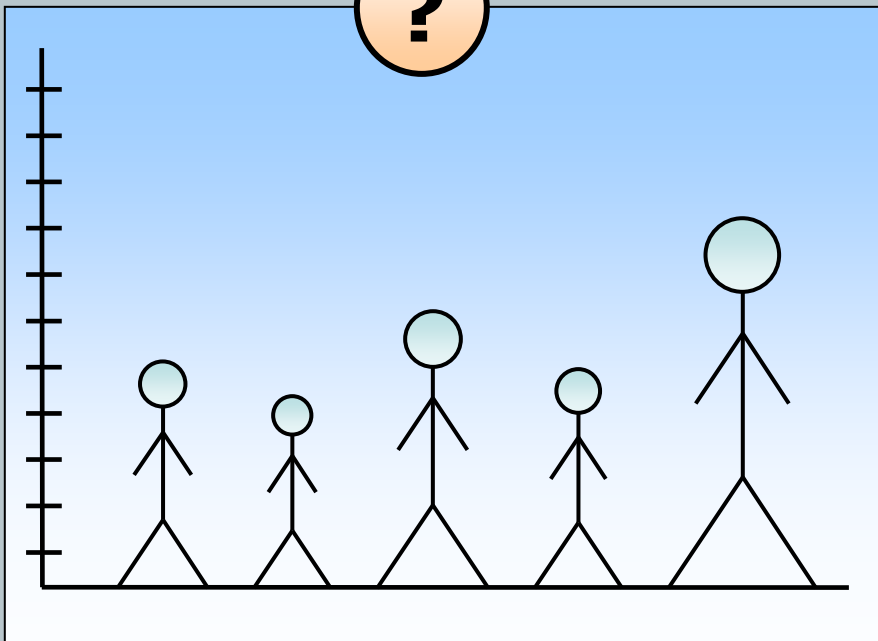
**VARIABILITA**



**CHYBA**

**INFORMACE**

# Běžná sumarizace dat „likviduje“ individualitu jedince



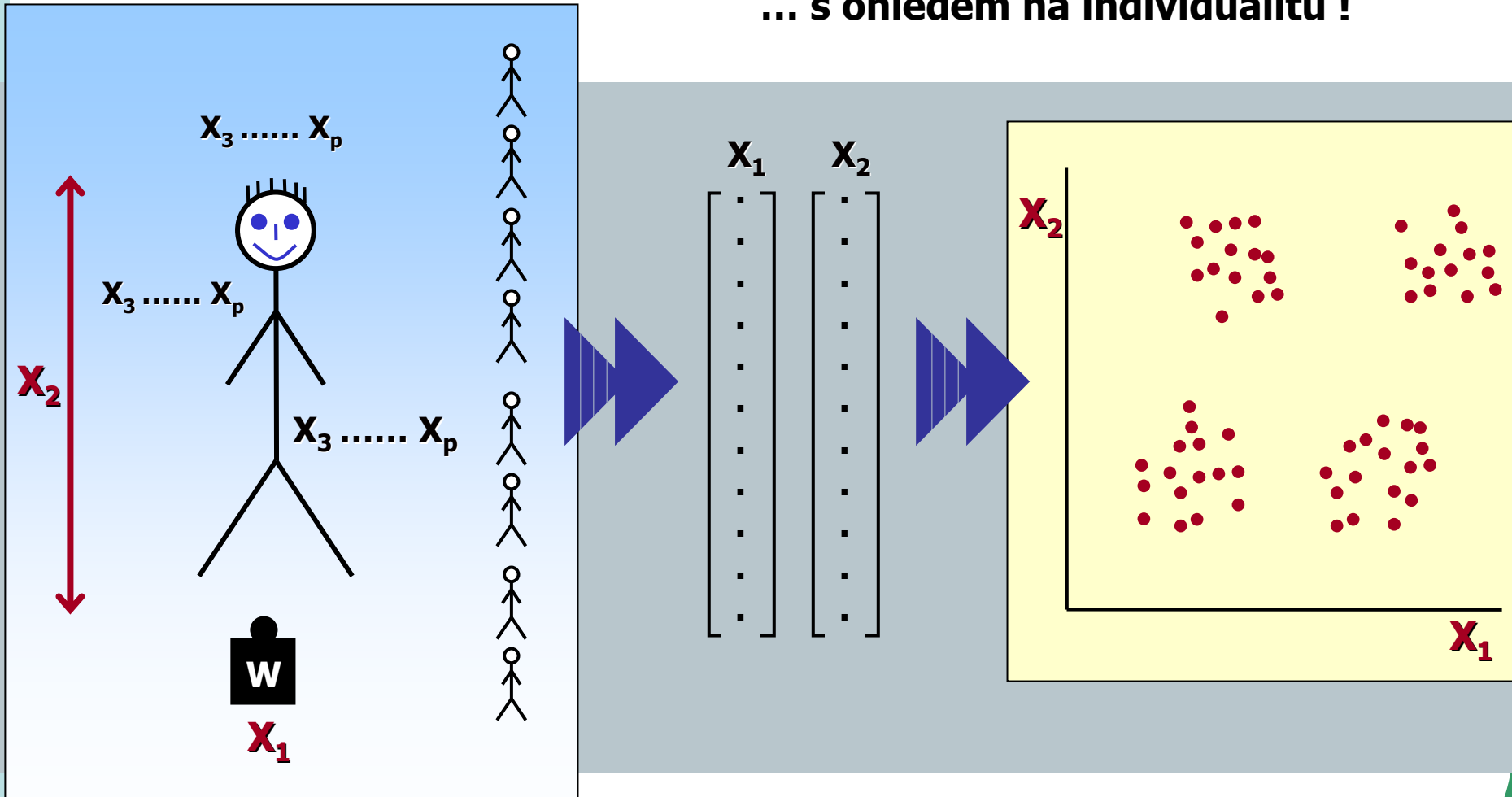
## Průměr $\pm$ SE

BĚŽNÁ STATISTICKÁ  
SUMARIZACE

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

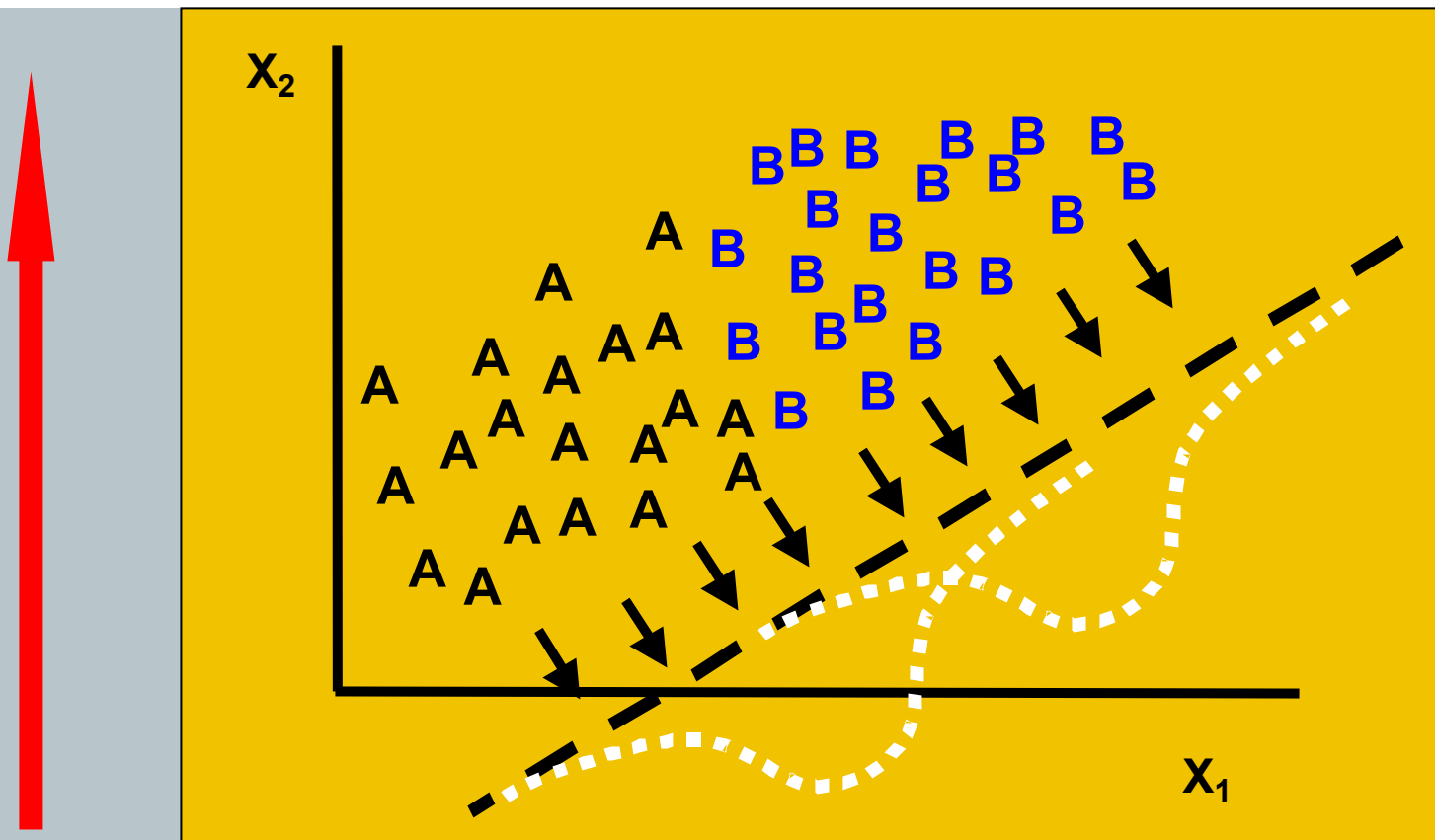
# Vícerozměrné hodnocení

... s ohledem na individualitu !



# Vícerozměrné hodnocení – nová kvalita

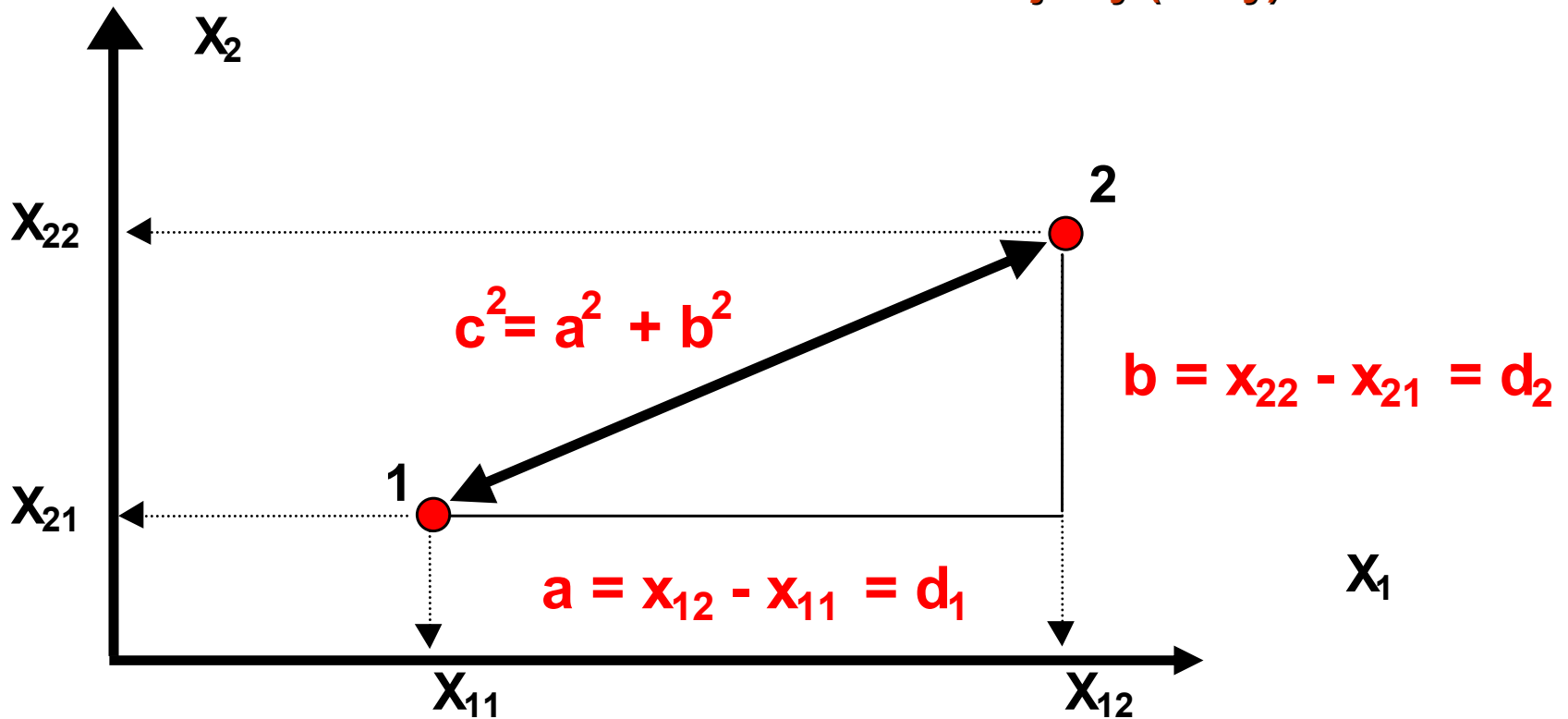
Pouze kombinované parametry mají  
odpovídající informační sílu



příklad:  $X_1 =$

# Vícerozměrné hodnocení vychází z jednoduchých principů

příklad: vícerozměrná vzdálenost měření mezi dvěma objekty (body)



# Vícerozměrné modelování je strategickou disciplínou

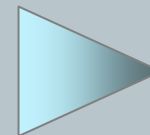


$X_1 \dots X_n$

technické parametry  
automobilu

$X_{n+1} \dots X_p$

řidičovy schopnosti  
a jeho stav



$X_{p+1} \dots X_2$

rychlost, povrch,  
situace

$X_1$

⋮

$X_2$

⋮

$X_3$

⋮

$X_4$

⋮

$X_5$

⋮

⋮

$X_p$

⋮

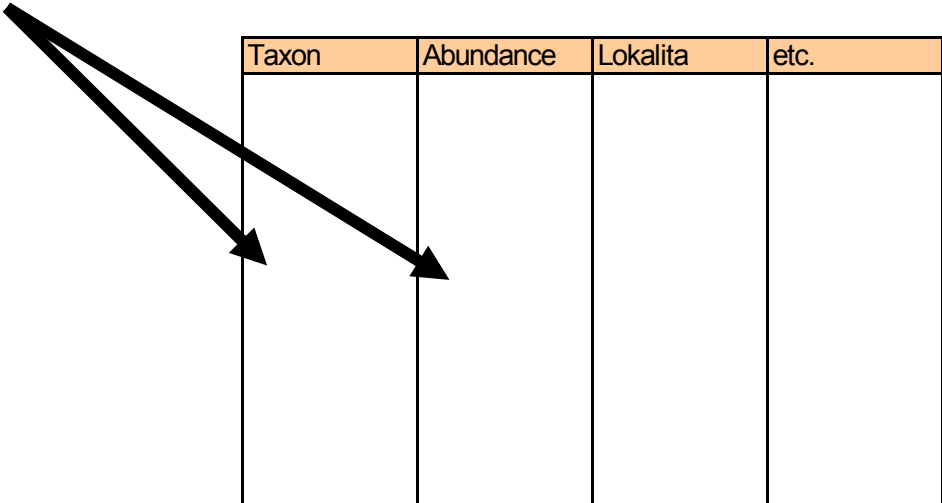


## 2. Data a jejich prezentace



# 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.)



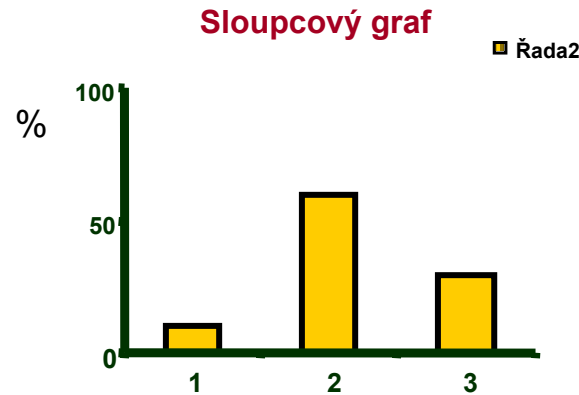
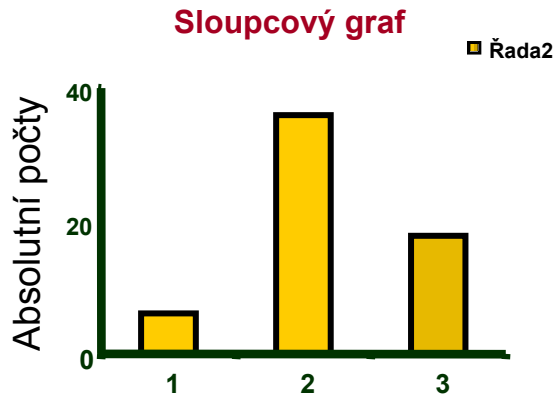
Taxon	Abundance	Lokalita	etc.

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

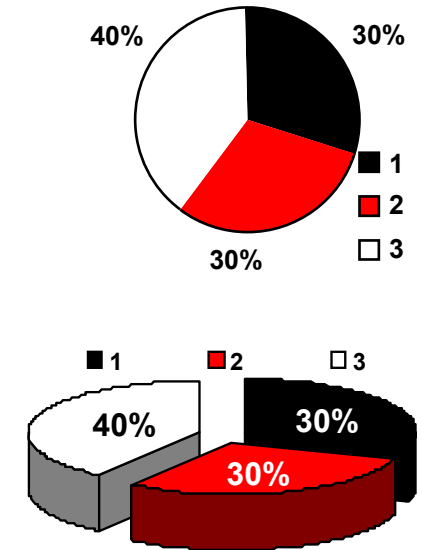


# Grafická prezentace dat - umění komunikace

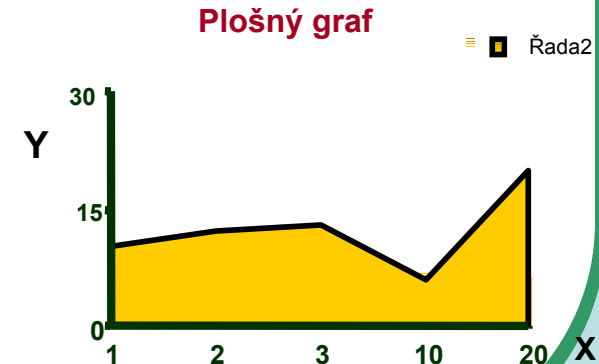
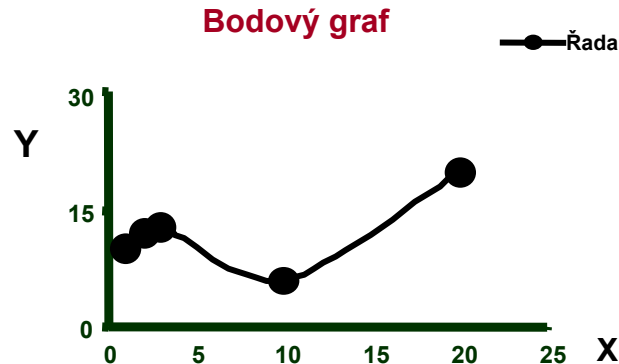
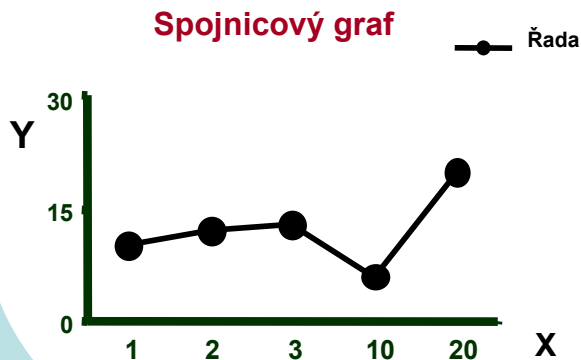
## 1. Výskyt kategorií (1, 2, 3,)



Koláčový (výsečový) graf



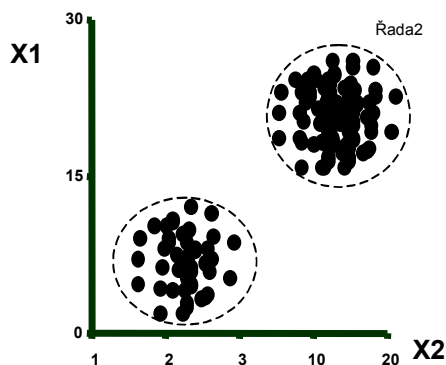
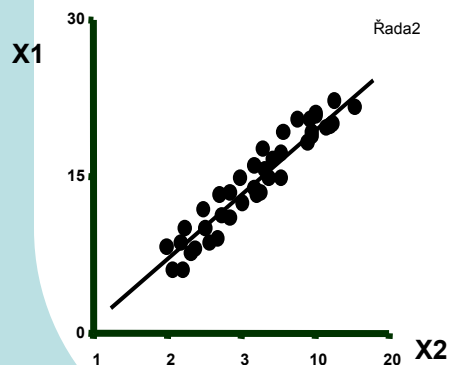
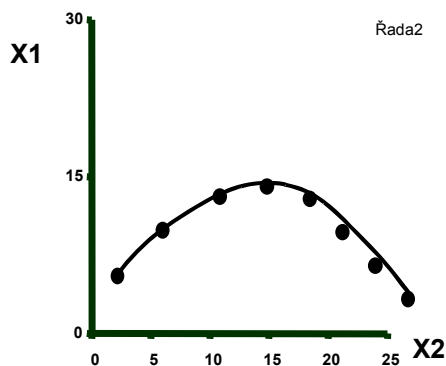
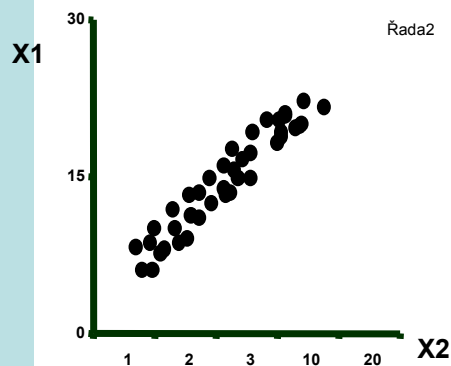
## 2. Vývoj hodnot (v čase) $Y$ vs. $X$ (t)



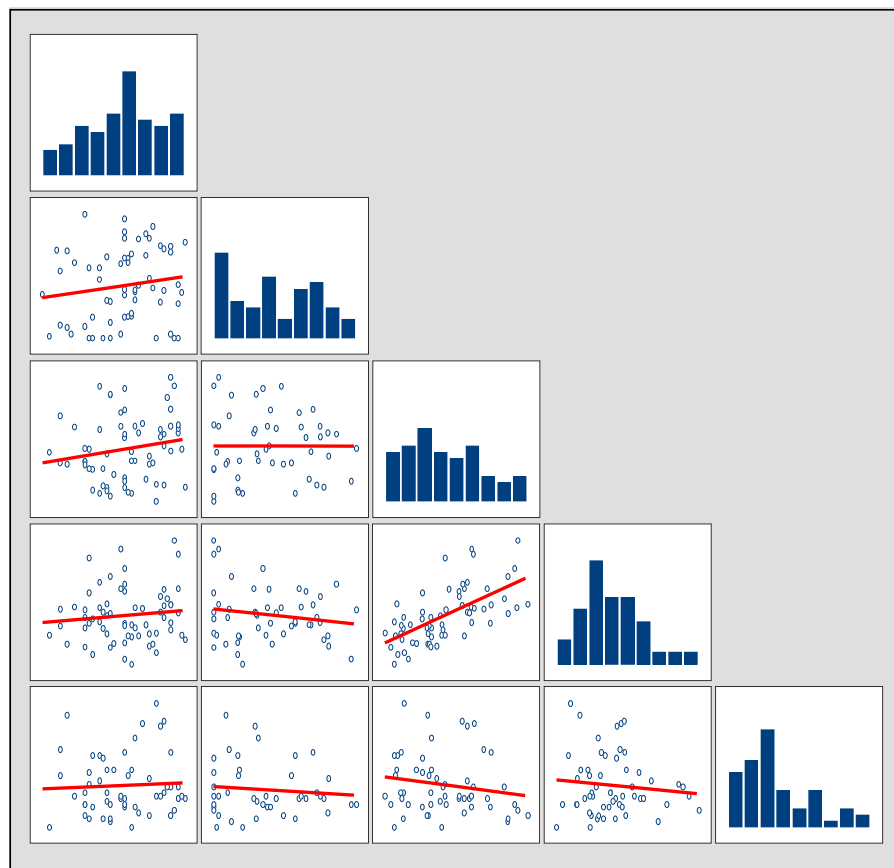
# Grafická prezentace dat – umění komunikace

## 3. Vztahy mezi proměnnými - korelace

Bodový - korelační diagram



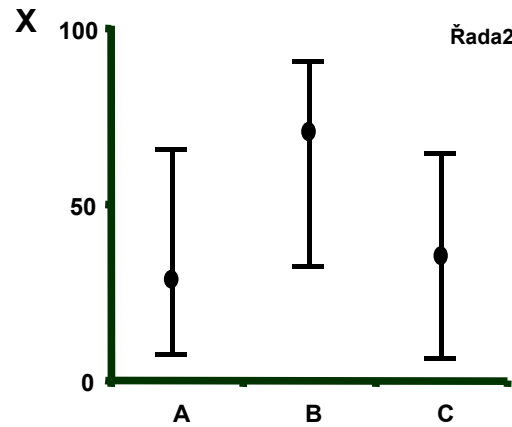
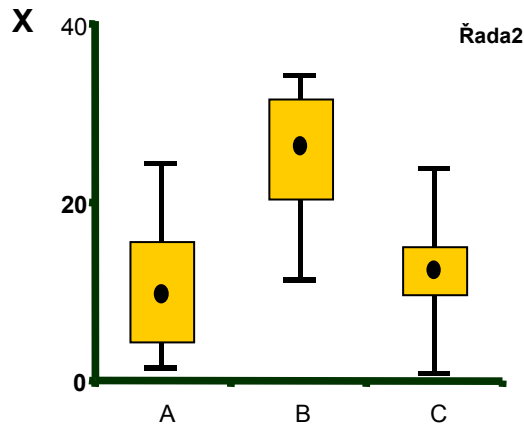
Bodový - korelační diagram



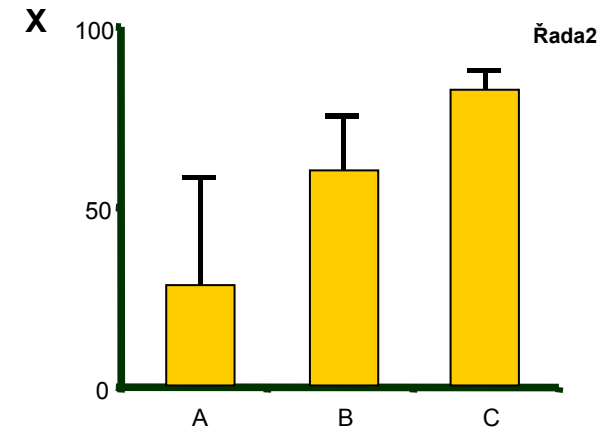
# Grafická prezentace dat – umění komunikace

## 4. Kvantitativní hodnoty parametru( $\hat{\mu}$ ) - $X$ - v rámci kategorií $A, B, C$

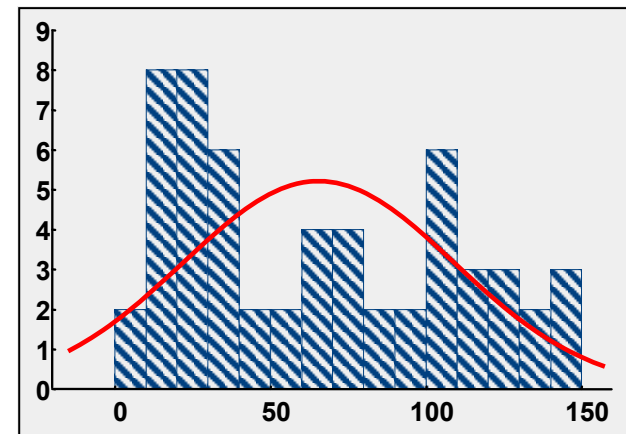
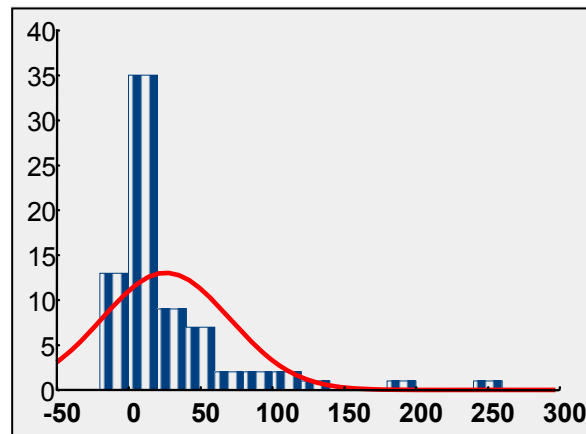
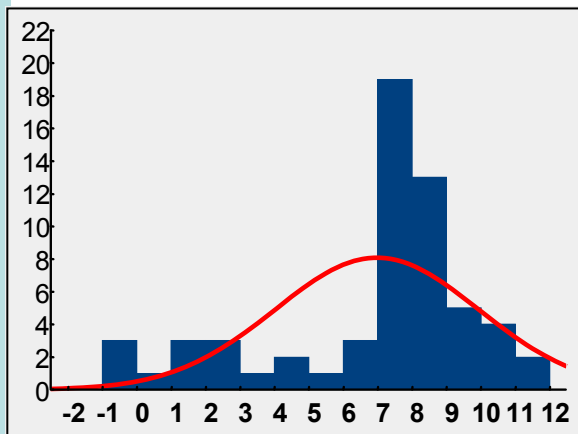
Krabicový graf



Sloupcový graf

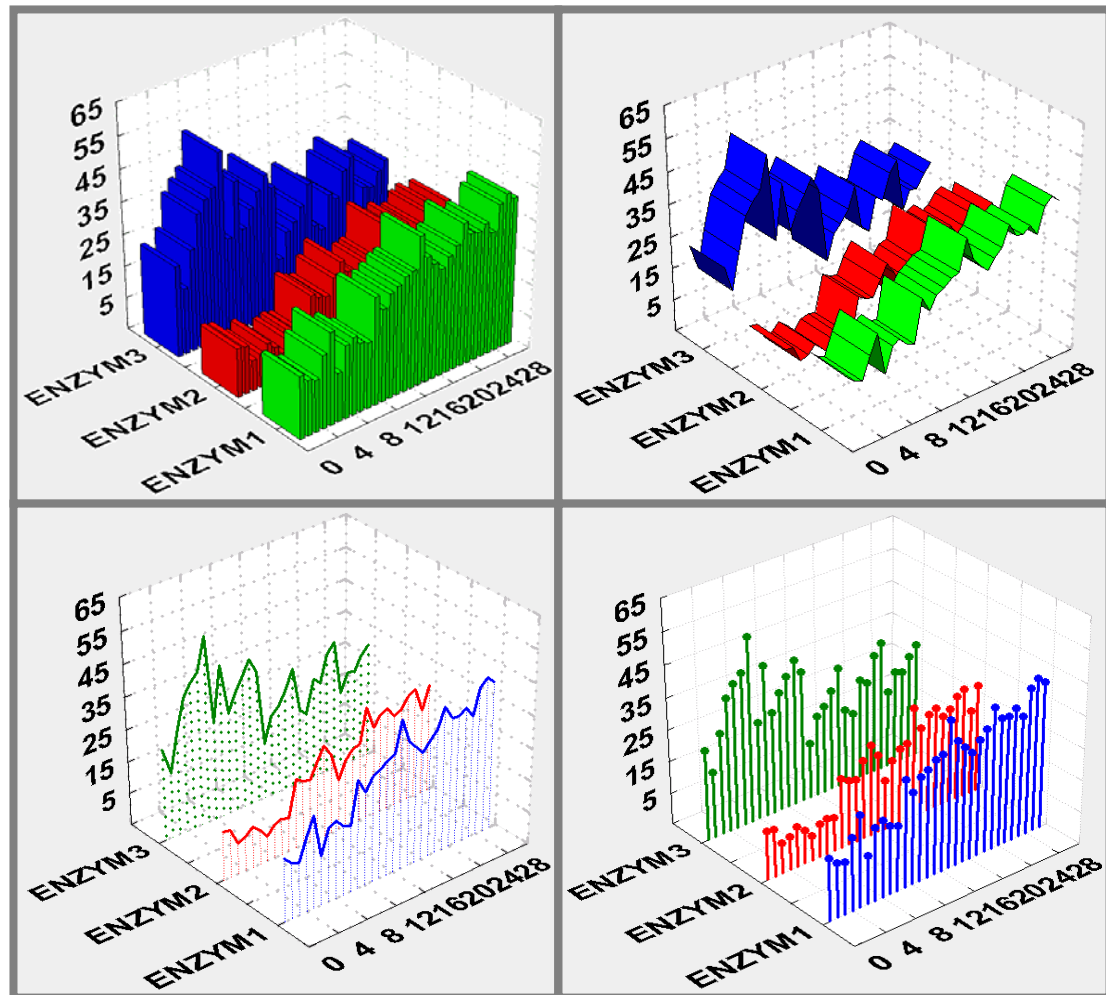
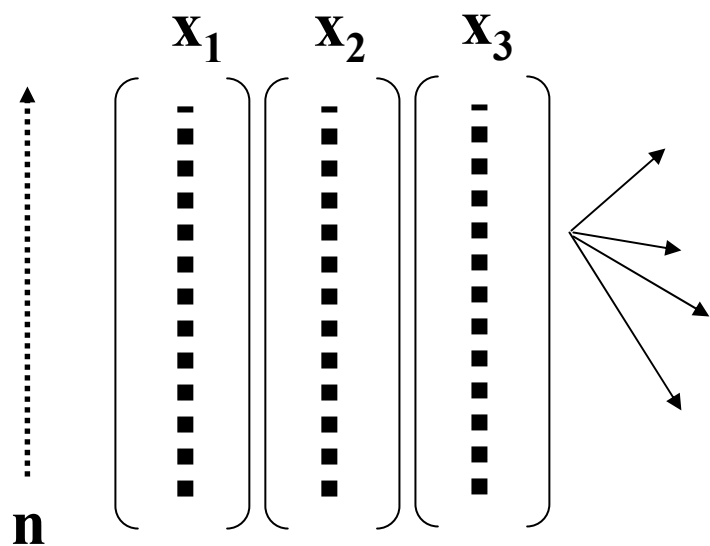


## 5. Histogram



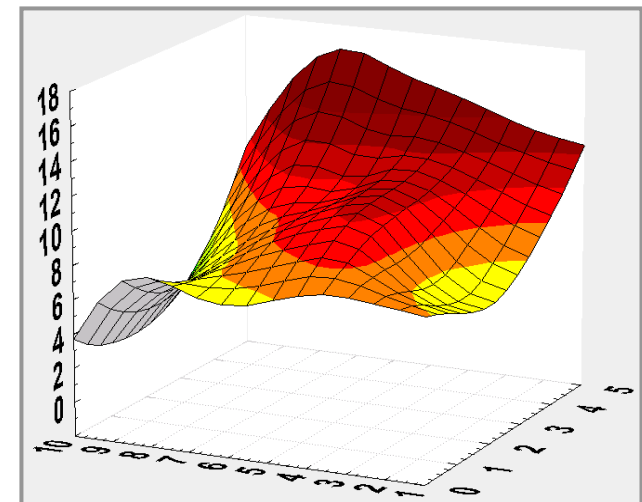
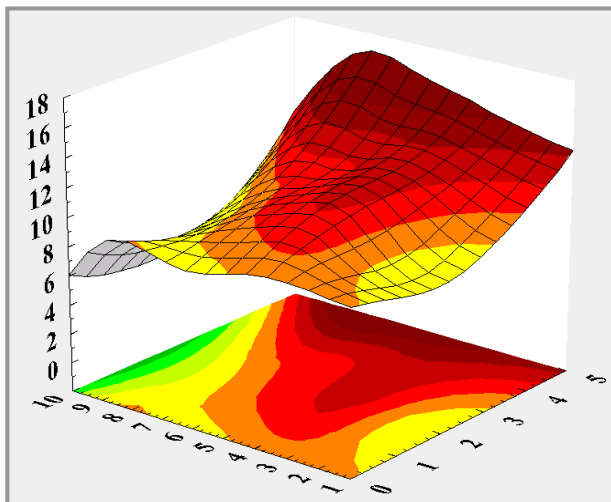
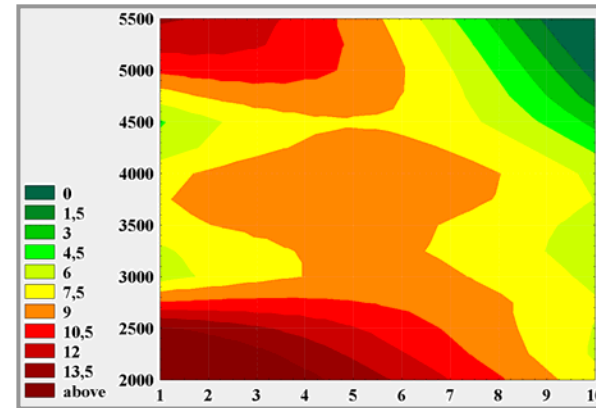
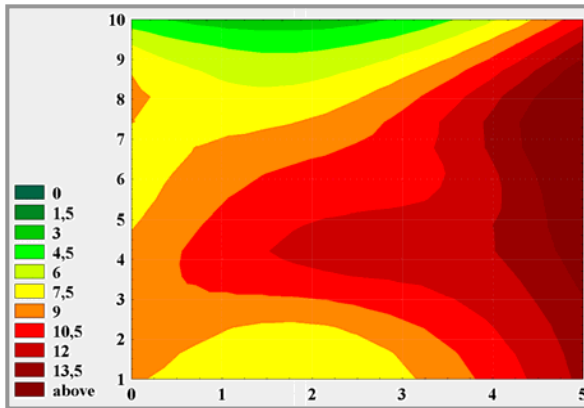
# Grafická prezentace dat - umění komunikace

## 6. Zviditelnění primárních dat



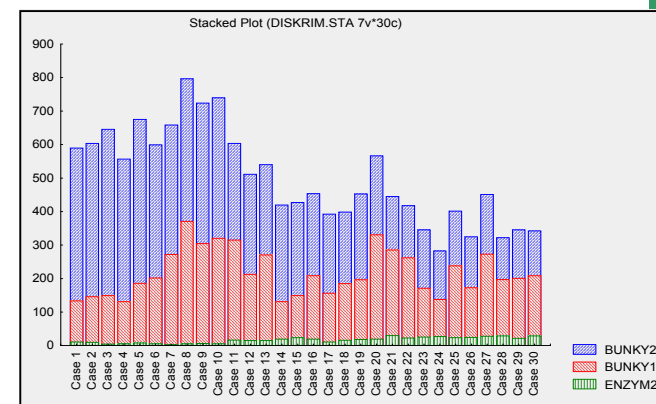
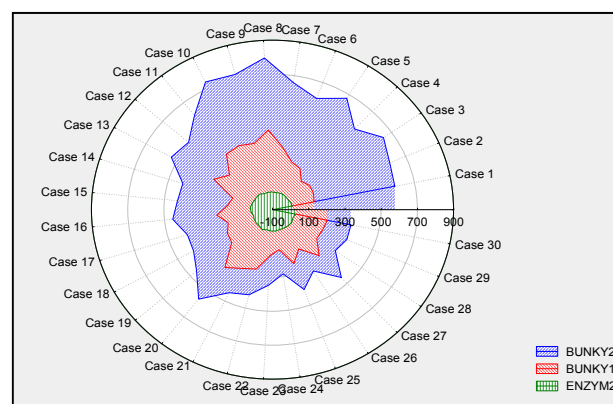
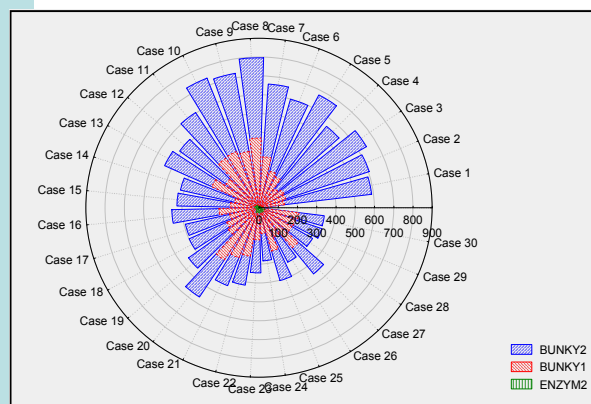
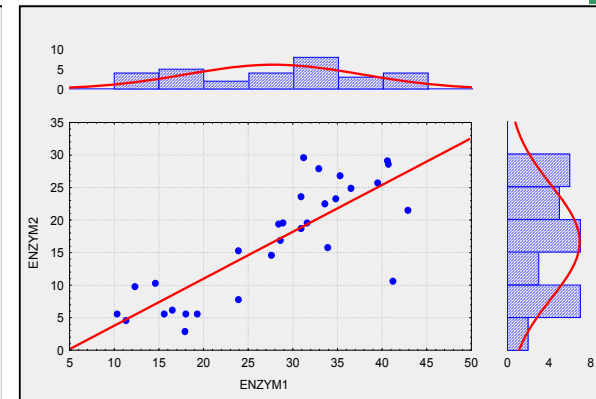
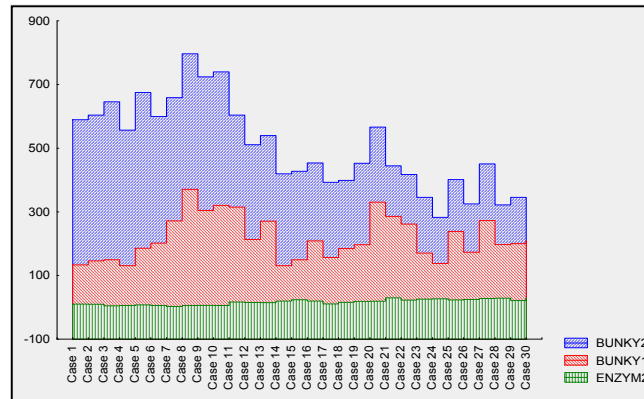
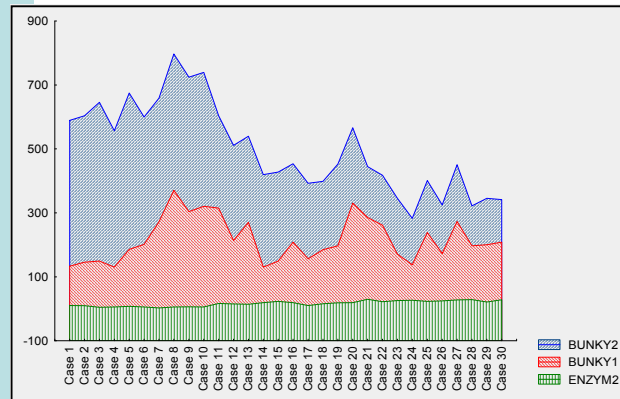
# Grafická prezentace dat – umění komunikace

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



# Grafická prezentace dat – umění komunikace

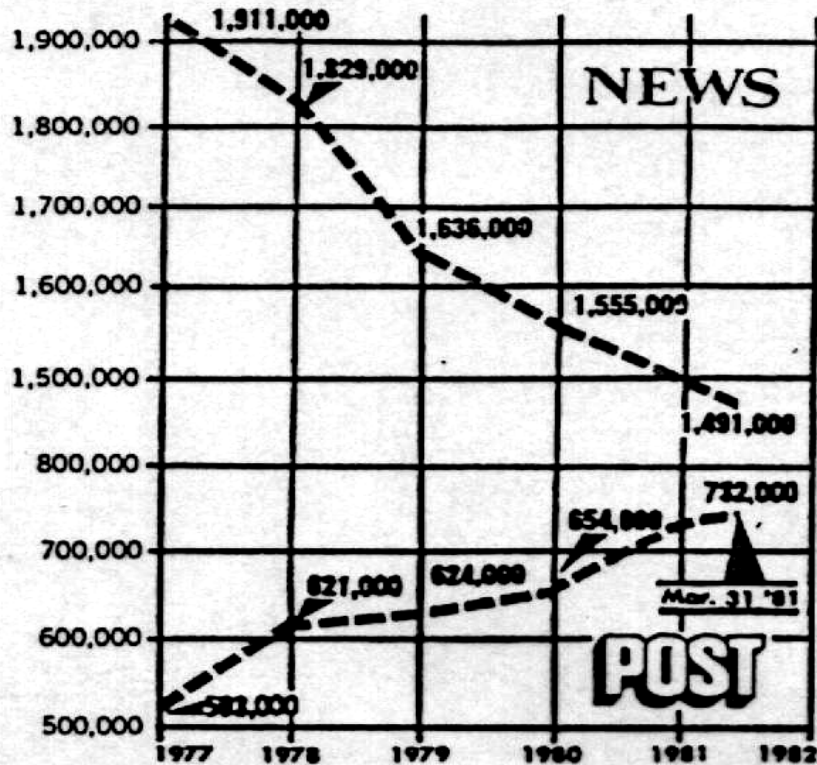
## 8. Grafické zviditelnění má nekonečně mnoho možností



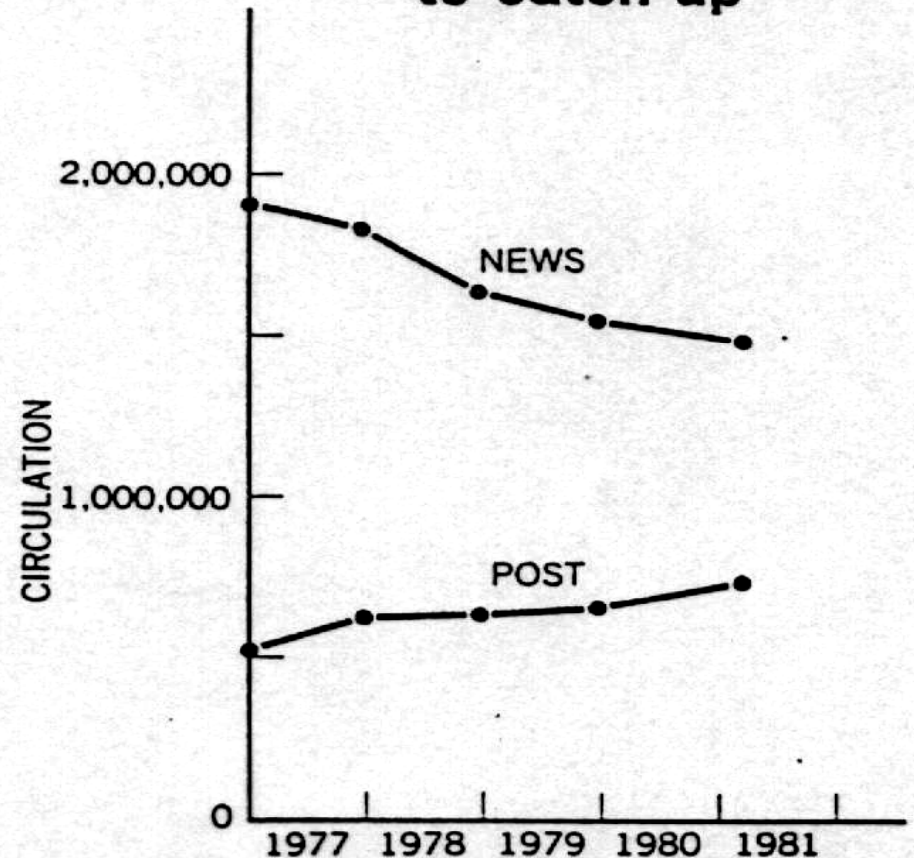
# Nesprávné užití grafů

## - problém rozsahu číselné osy

### The soaraway Post — the daily paper New Yorkers trust

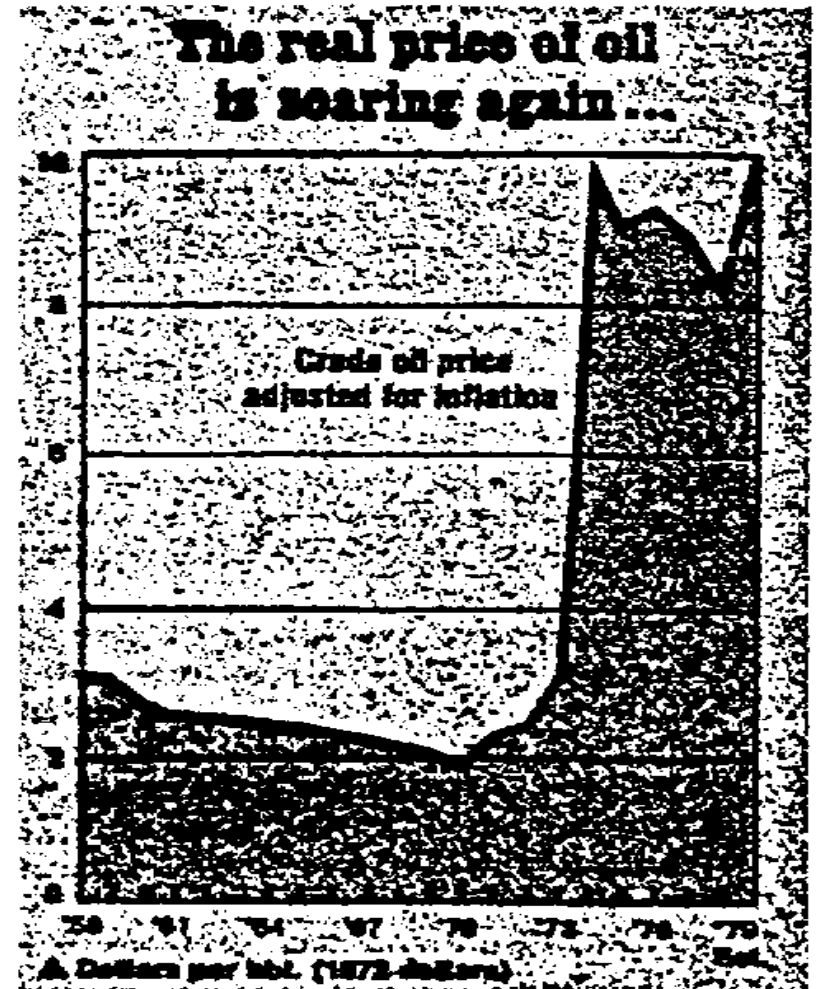
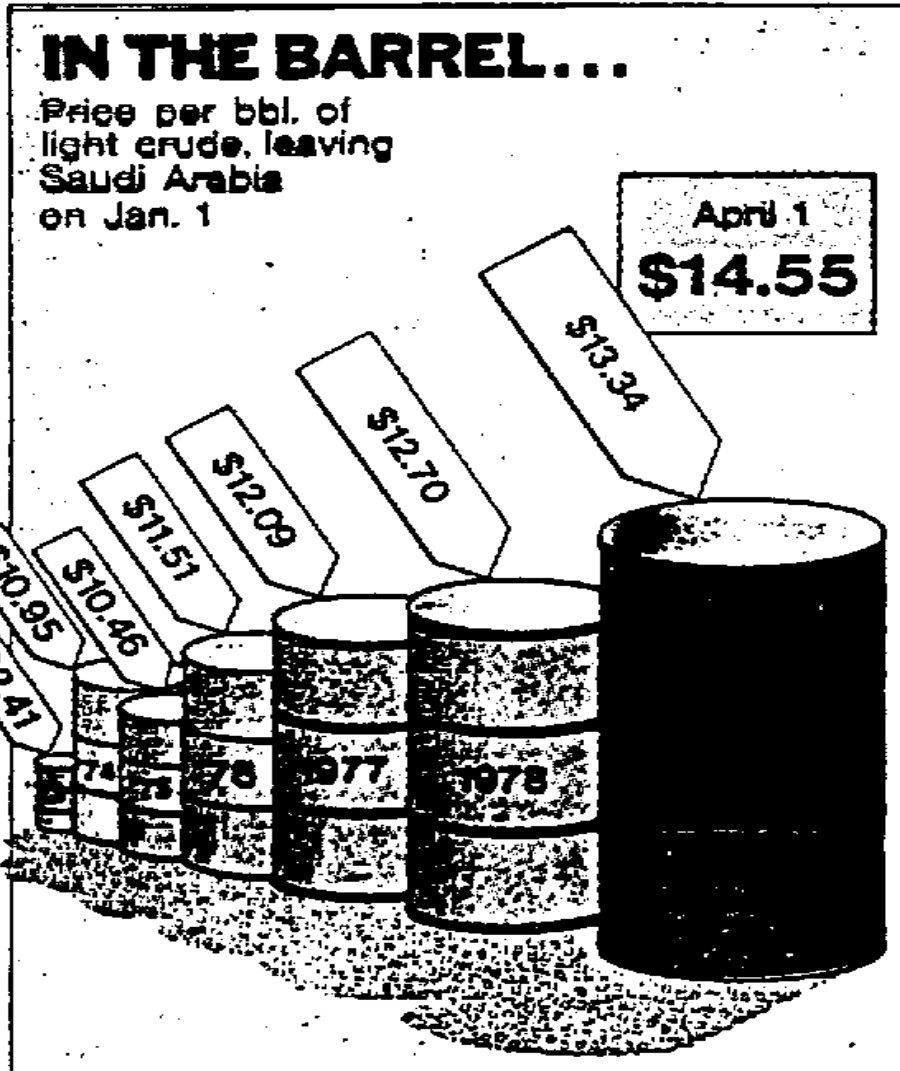


### The Post struggles to catch up



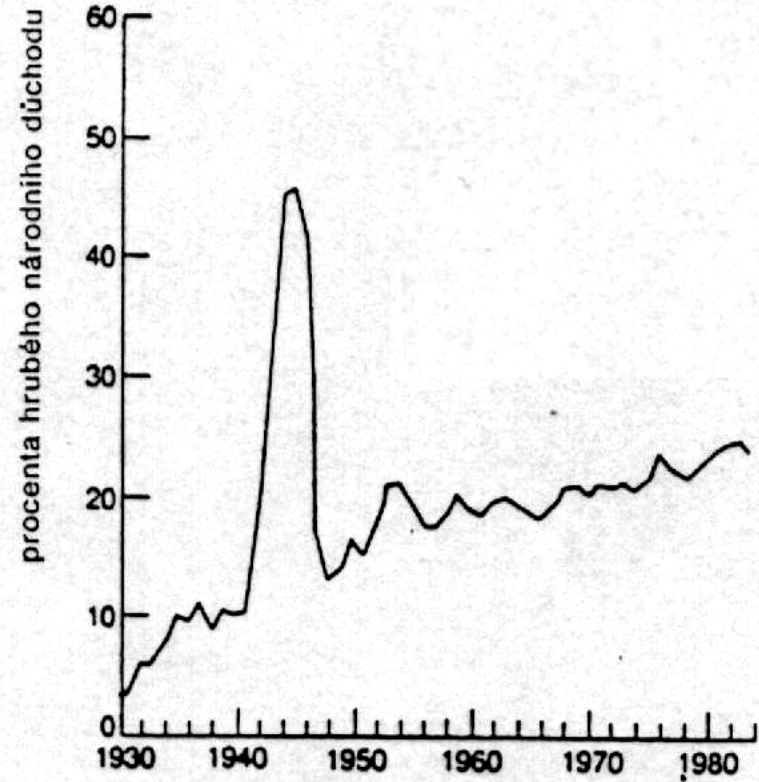
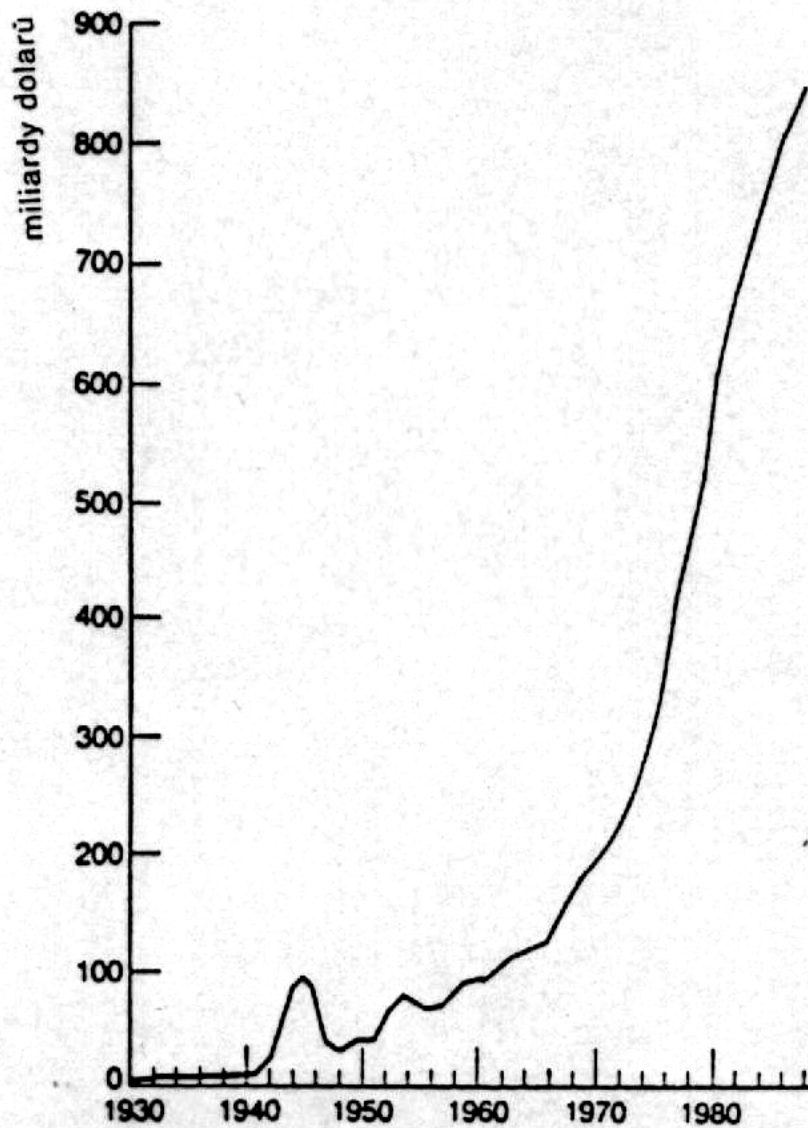


# Nesprávné užití grafů - grafické zastírání trendu

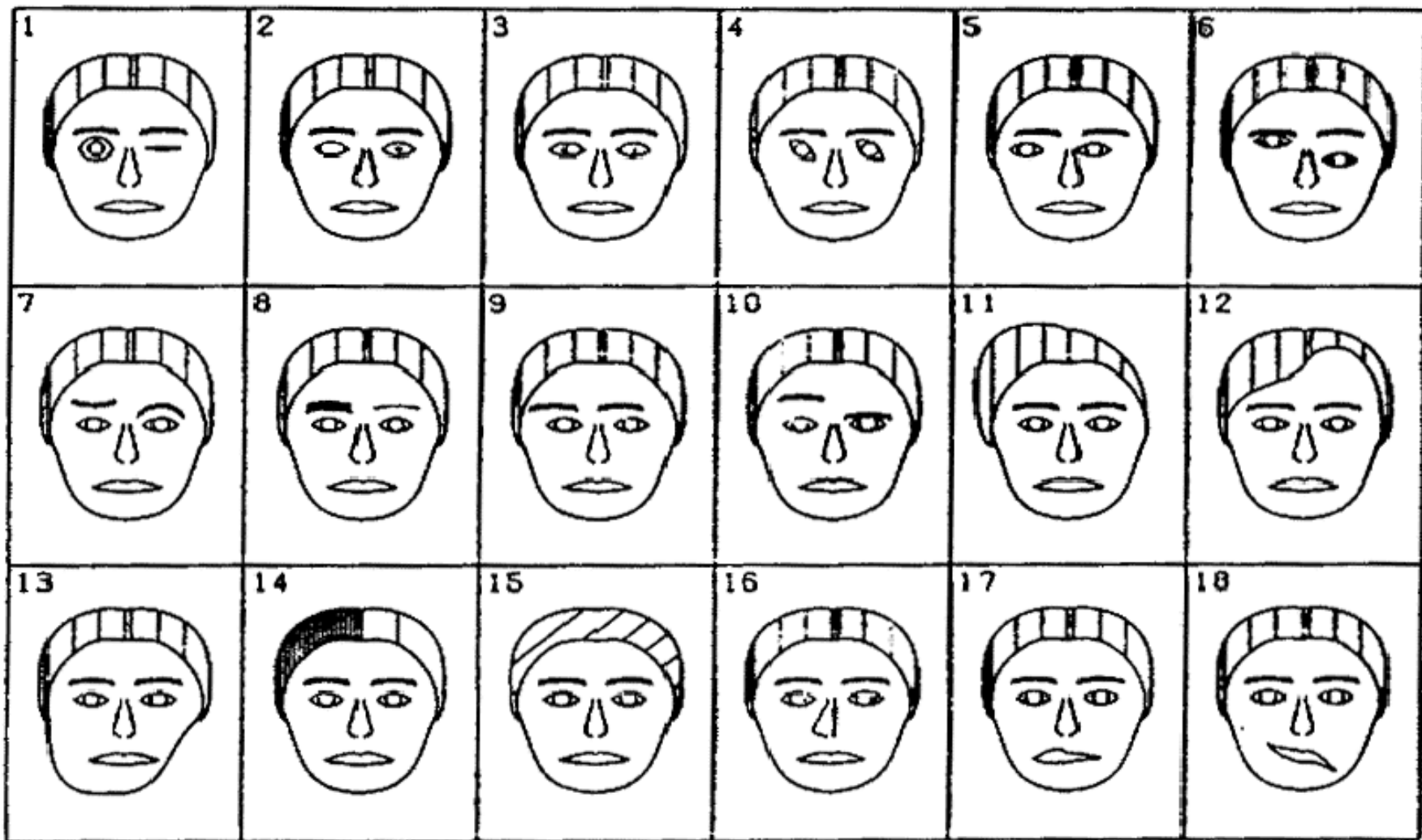




# Nesprávné užití grafů - problém standardizace hodnot

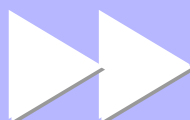


# Grafy zaměřené na vícerozměrné soubory dokáží zviditelnit i veliké soubory dat





## 3. Jak vznikají informace



# Primárním důvodem analýzy dat je získání nezkrácené a přehledné INFORMACE

Ukázka uspořádaného datového souboru

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

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

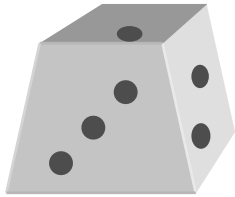
Výběr - výběrová populace - cílová populace

Náhodný výběr

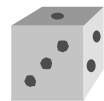
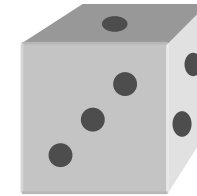
Reprezentativnost

# JAK vznikají informace ?

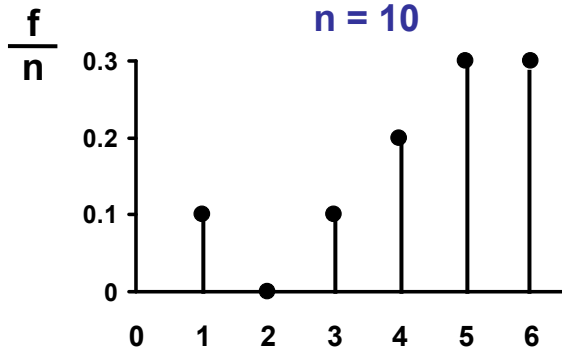
„Empirical approach“



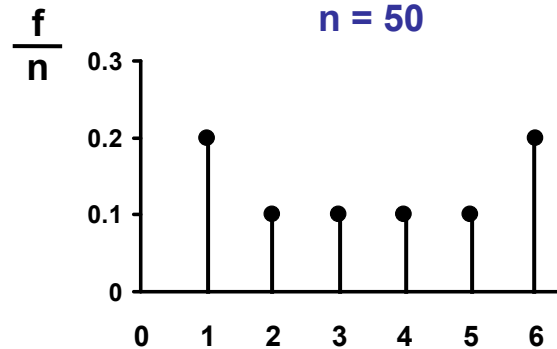
„Classical approach“



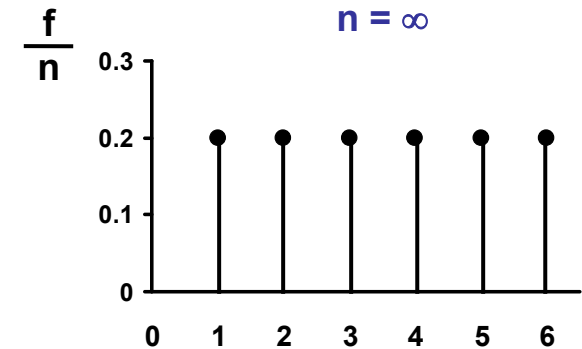
Empirický postup



možné jevy: čísla 1 – 6

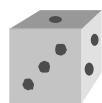


$n$  – počet hodů (opakování)

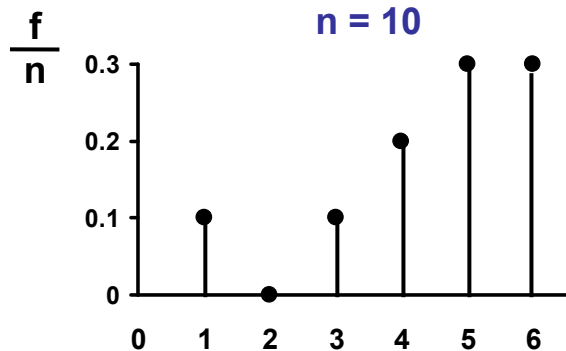


**U složitých stochastických systémů se pravda získá až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit**

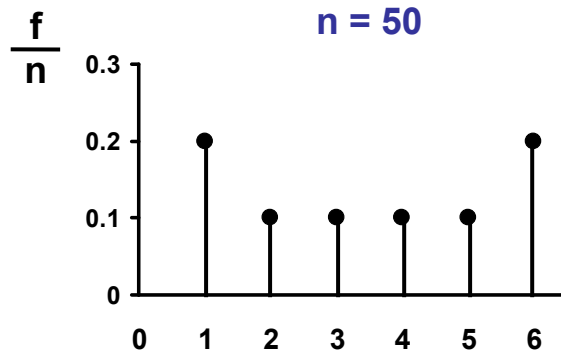
# JAK vznikají informace ?



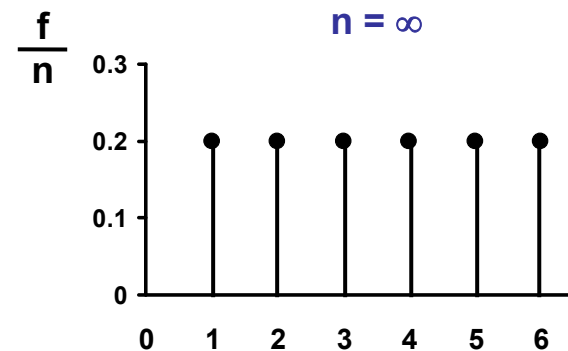
Empirický postup



možné jevy: čísla 1 – 6



$n$  – počet hodů (opakování)



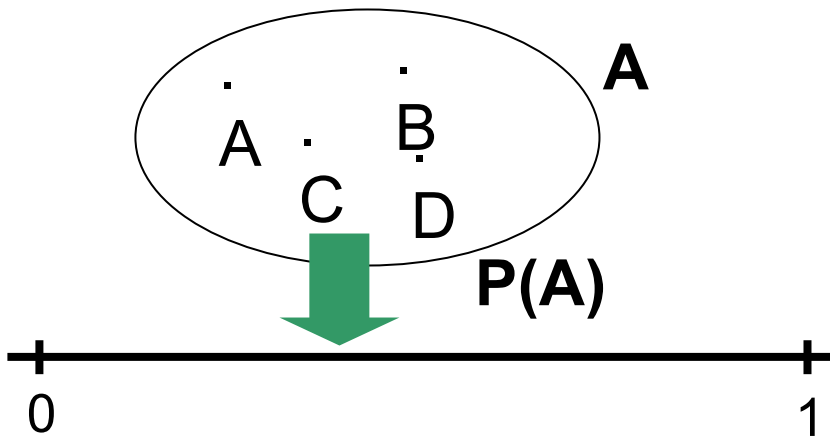
Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější) .... diskutabilní je ale ovšem míra zobecnění konkrétního experimentu

# 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
- $P(A \cap B) = P(A) \cdot P(B)$  ..... nezávislé jevy
- $P(A \cap B) = P(A) \cdot P(B/A)$  ..... závislé jevy
- $P(A / B) = P(A \cap B) / P(B)$  ..... podmíněná pravděpodobnost





## 4. Základní typy dat



# Jak vznikají informace ?

## – různé typy dat znamenají různou informaci

Data poměrová



Kolikrát ?

Data intervalová



O kolik ?

Data ordinální



Větší, menší ?

Data nominální

Rovná se ?

**Spojité data**

Kategoriální otázky

**Diskrétní data**

Otázky „Ano/Ne“

Podíl hodnot větší/menší než specifikovaná hodnota ?

Procenta odvozené hodnoty

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

# Jak vznikají informace ?

## – různé typy dat znamenají různou informaci

### Statistika středu

Data poměrová



**PRŮMĚR**

**Spojité data**

Data intervalová



**MEDIÁN**

**Diskrétní data**

Data ordinální



**MODUS**

Data nominální

$Y = f$

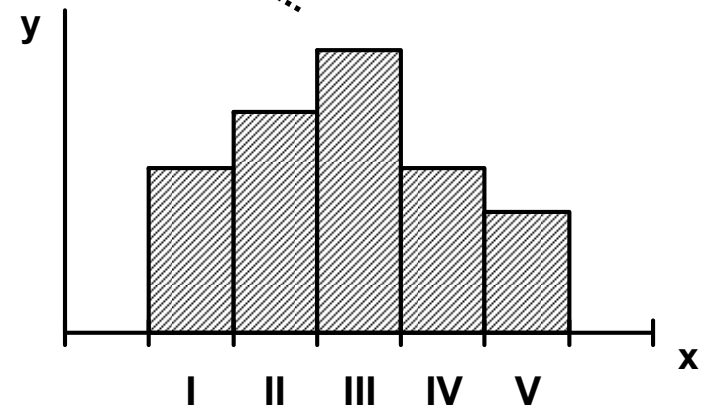
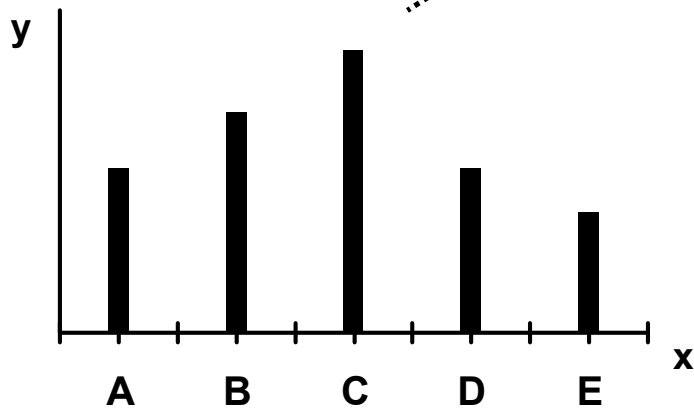
X

# JAK vznikají informace ?

- opakovaná měření informují rozložením hodnot

Y: frekvence  
- absolutní / relativní

**KOLIK** se  
naměřilo



**CO** se  
naměřilo

X: měřený znak

Diskrétní data

Spojité 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  
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;

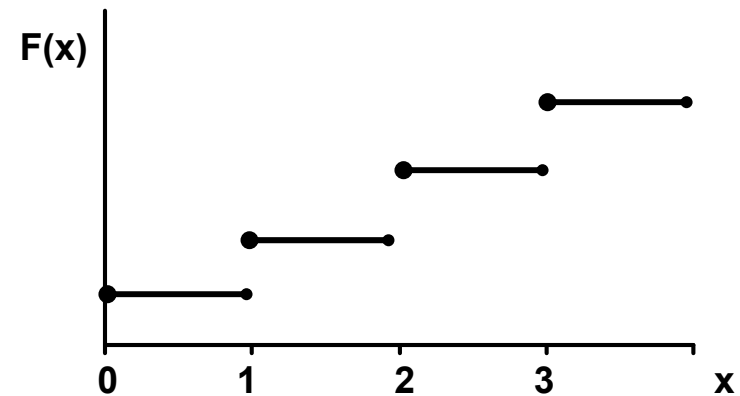
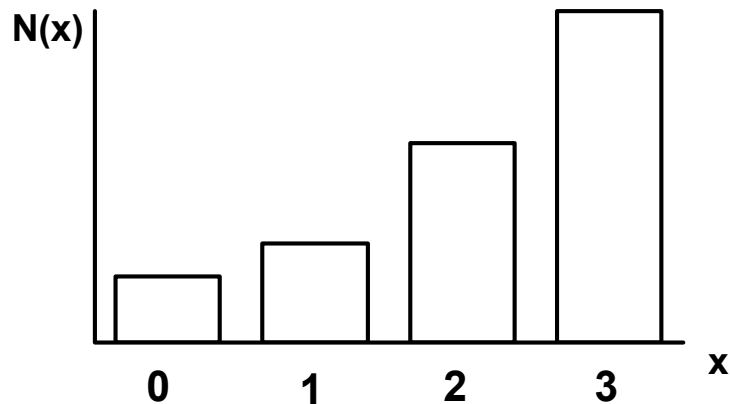
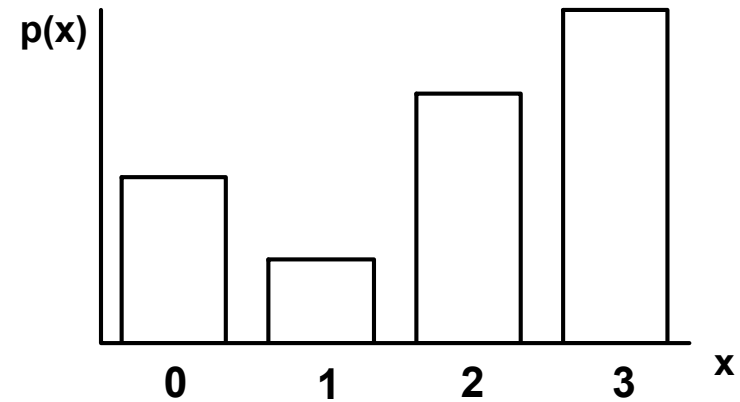
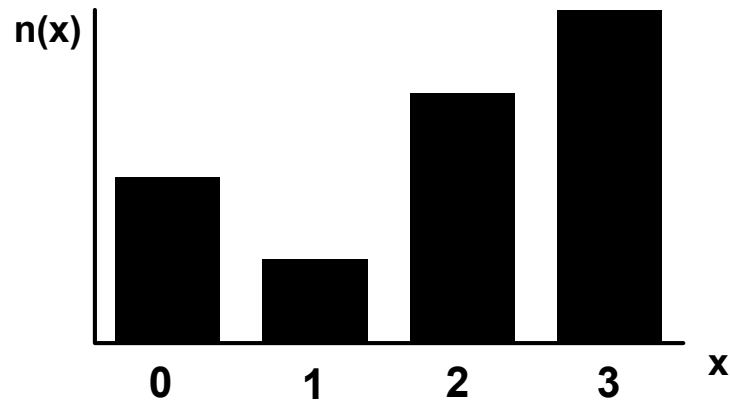
$$N(x) = \sum_{t \leq x} n(t)$$

**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

Frekvenční sumarizace

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



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''

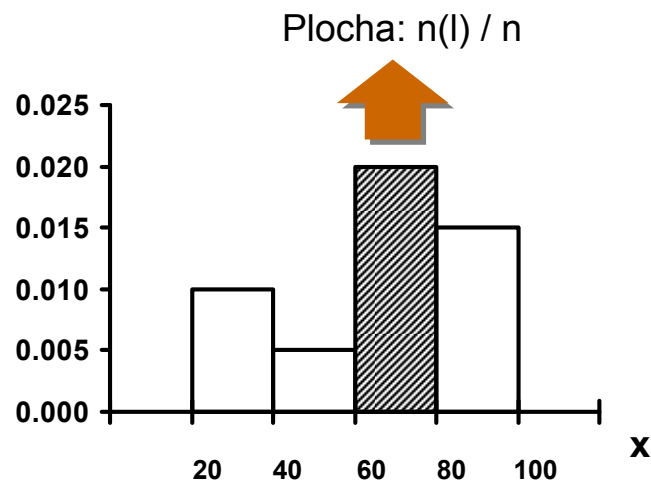
F(x'') – intervalová relativní kumulativní četnost do horní hranice X''



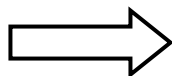
# Jak vznikají informace ?

## - frekvenční sumarizace spojitých dat

### Histogram

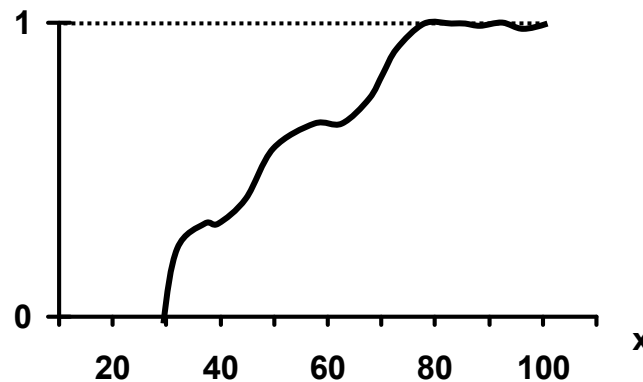


$$f(x) = \frac{n(l) / n}{d(l)}$$

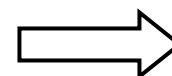


Intervalová  
hustota  
četnosti

### Výběrová distribuční funkce



$F(x)$



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"

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

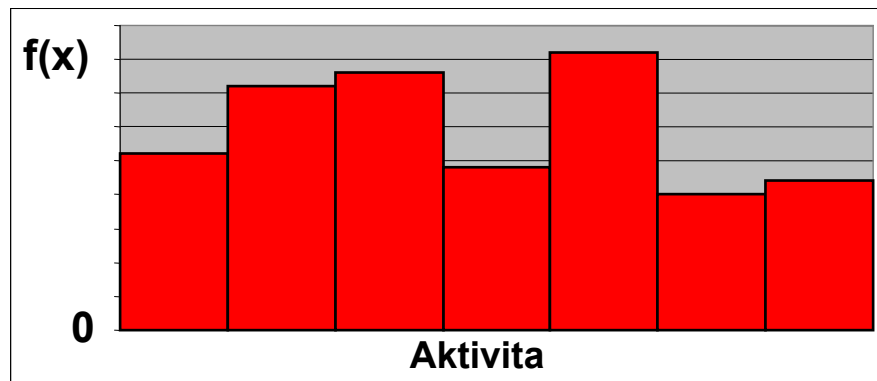
"Relative frequency histogram"

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

# Spojité data – postup frekvenčních analýz

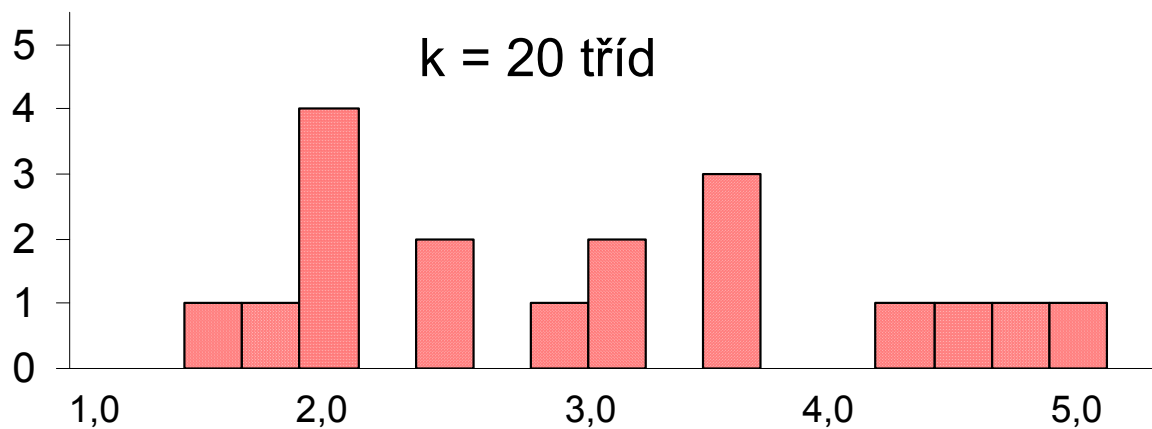
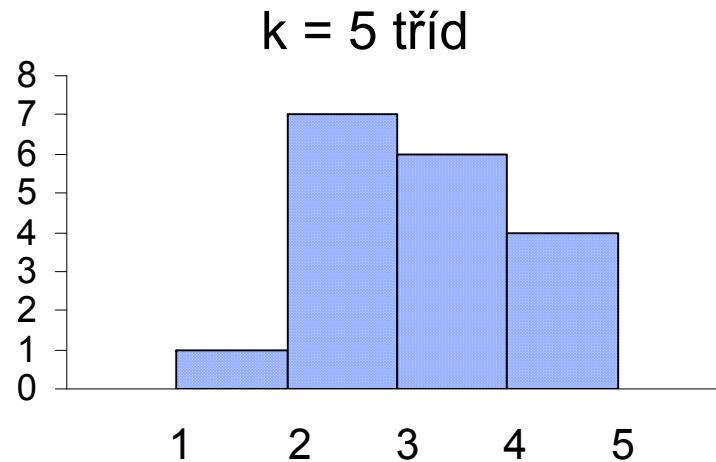
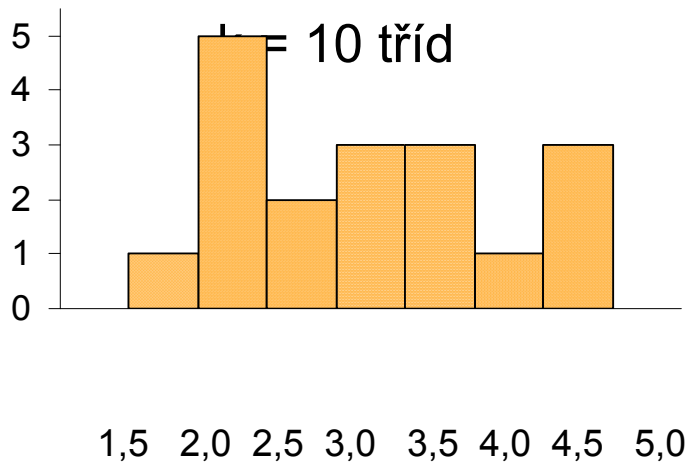
## Aktivita enzymu (X)

- I. Utřídít 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

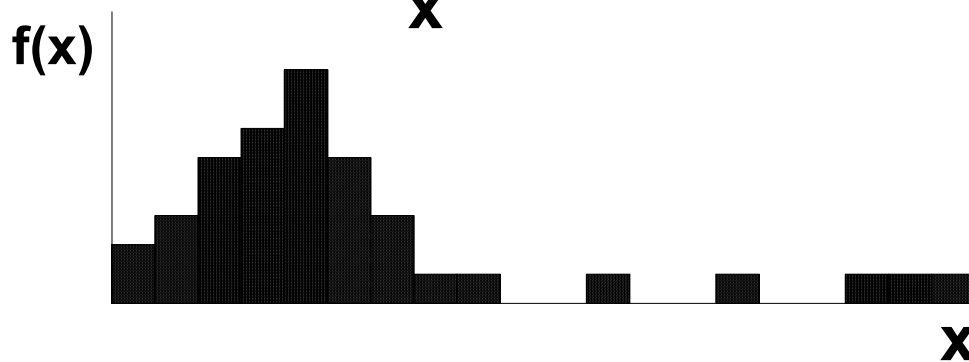
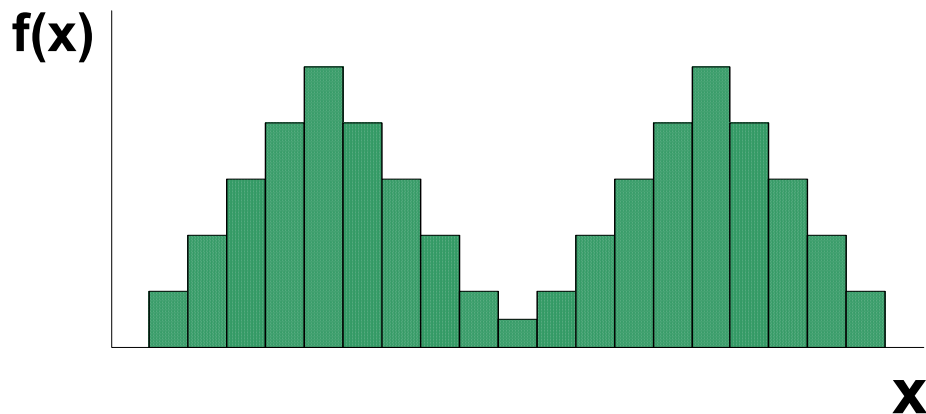
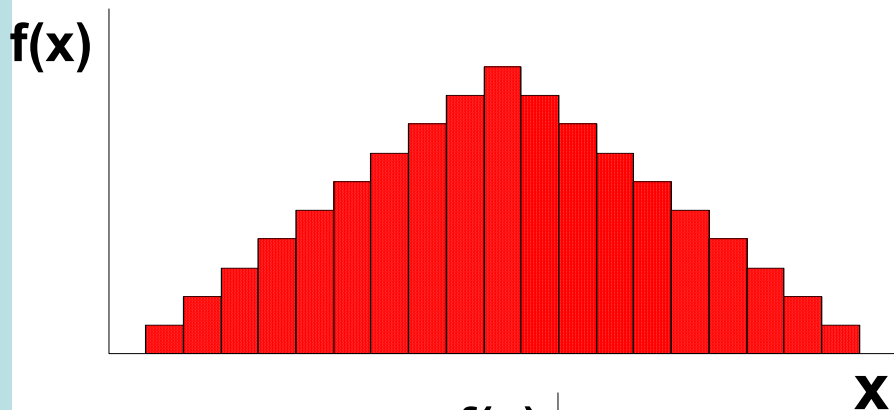
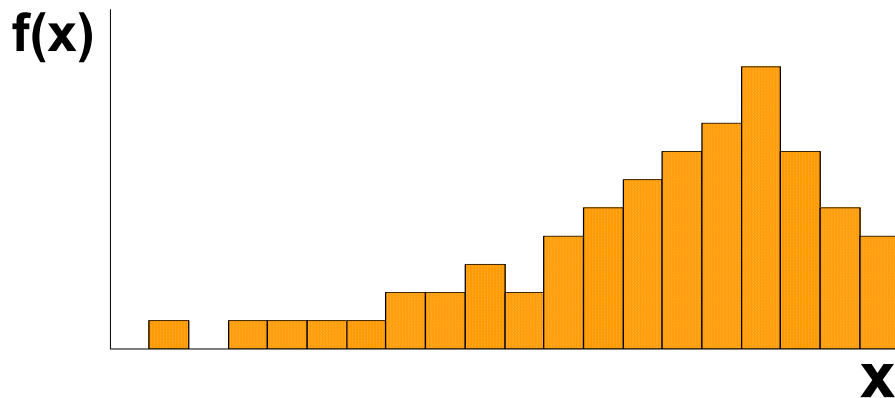
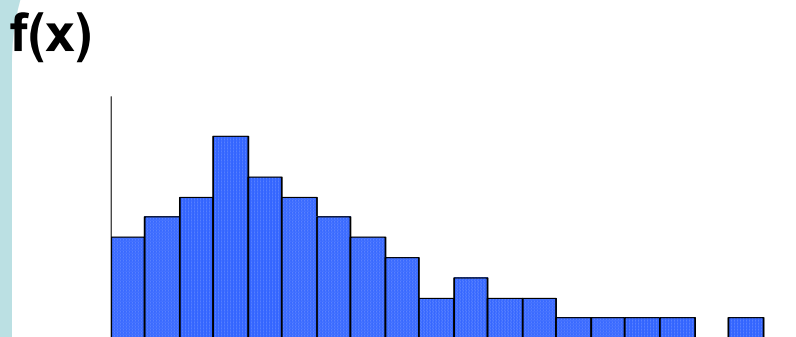


**Počet intervalů X: dán daty a hodnotitelem**  
**Šířka intervalů: pokud možno stejná**

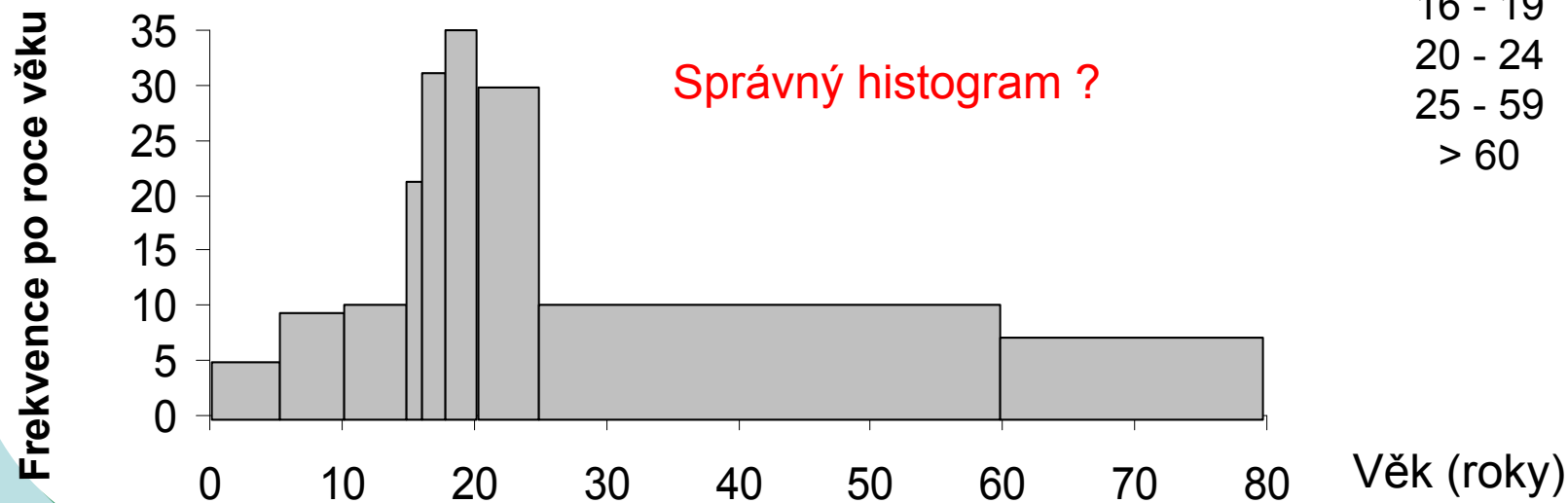
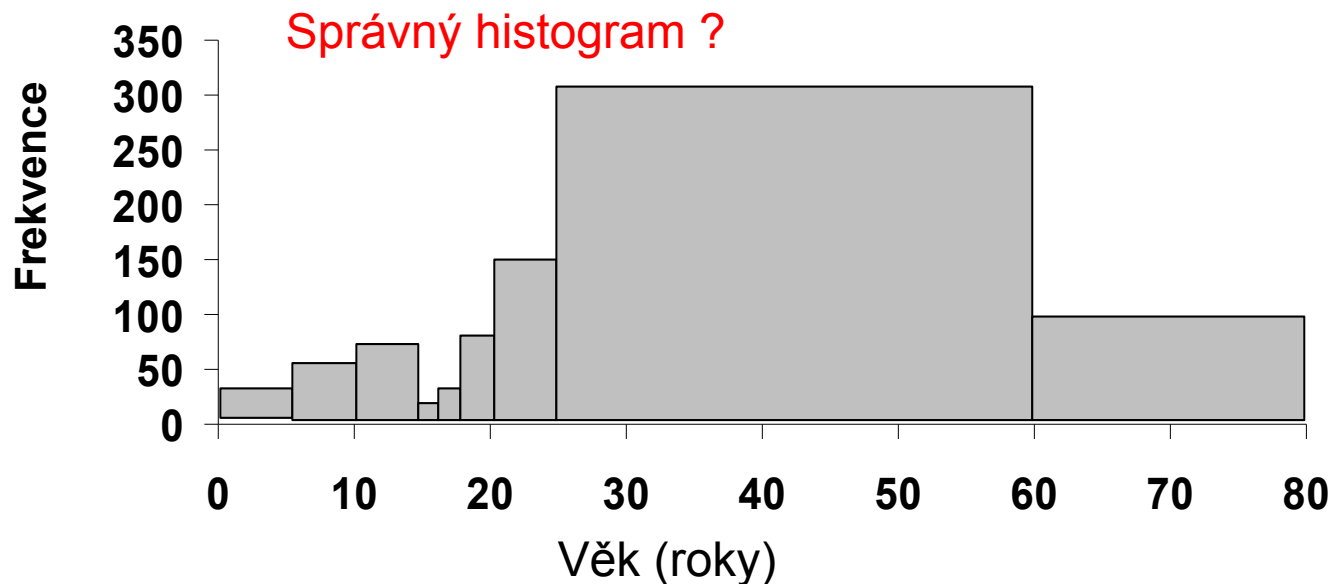
# Počet zvolených tříd a velikost souboru určují kvalitu výstupu



# Histogram vyjadřuje tvar výběrového rozložení

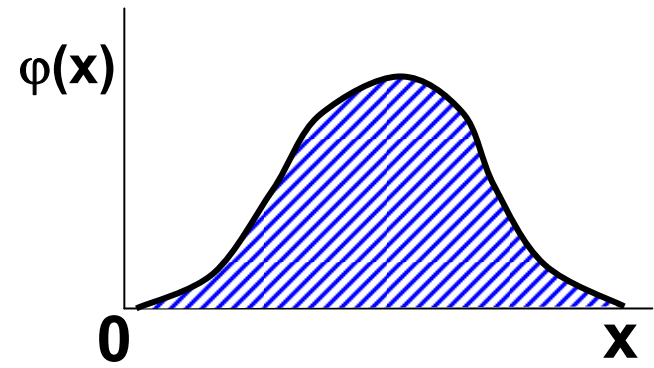


# Příklad: věk účastníků vážných dopravních nehod

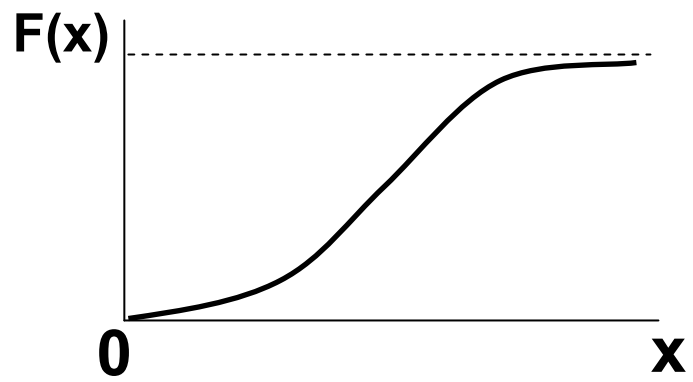


<u>Věk</u>	<u>f</u>
0 - 4	28
5 - 9	46
10 - 15	58
16 - 19	20
20 - 24	114
25 - 59	316
> 60	103

# Pojem ROZLOŽENÍ - příklad spojitých dat



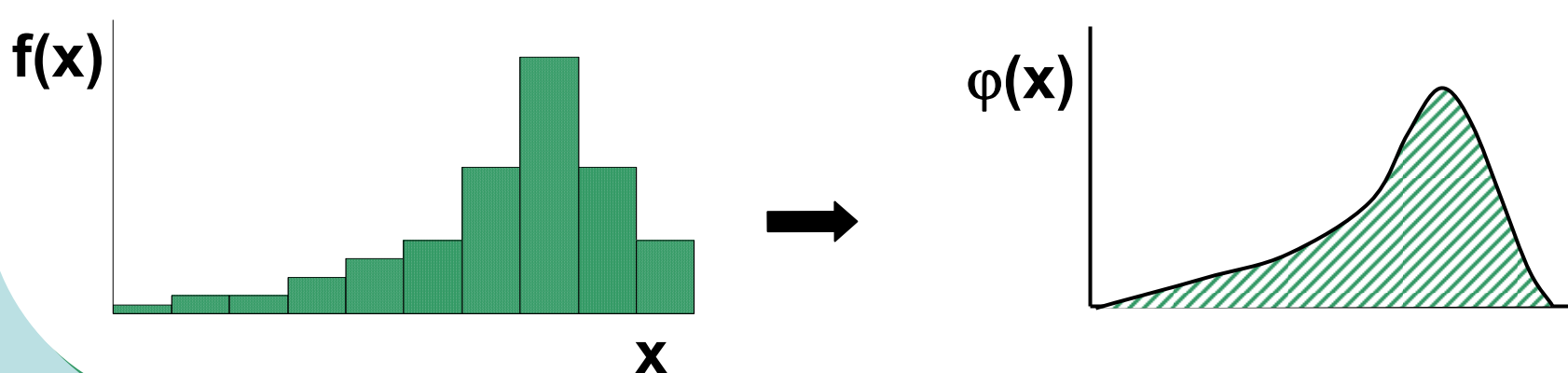
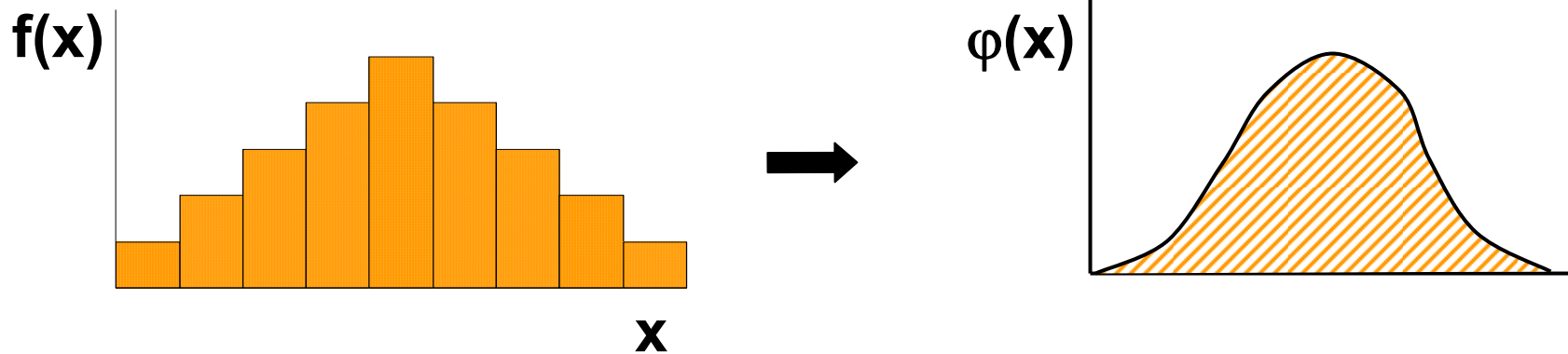
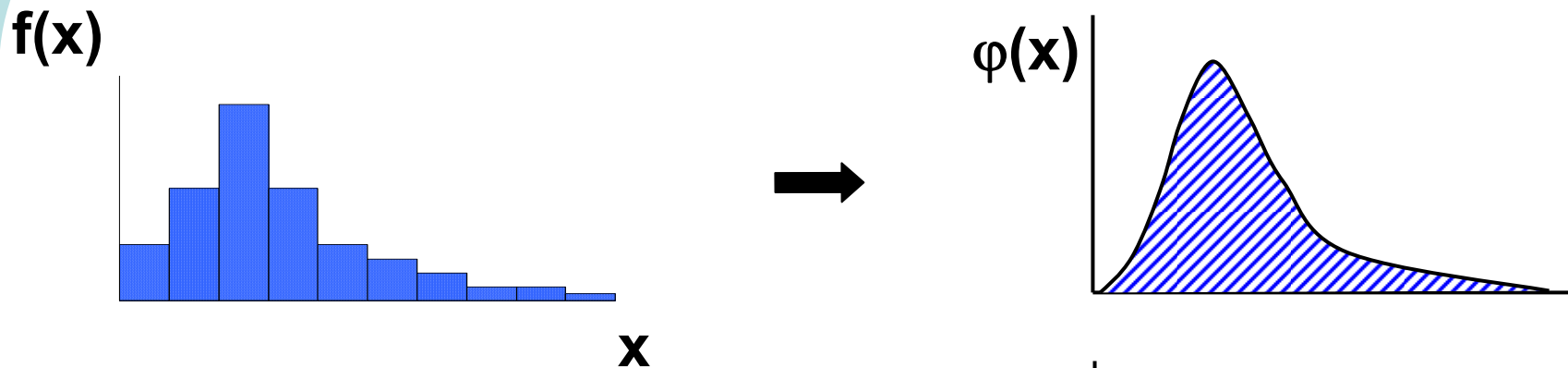
Rozložení



Distribuční funkce

**Je - li dána  
distribuční  
funkce,  
je dáno  
rozložení**

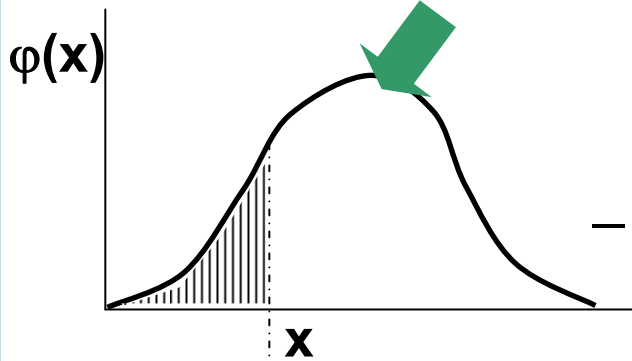
# Výběrové rozložení hodnot lze modelově popsat a definovat tak pravděpodobnost výskytu $X$





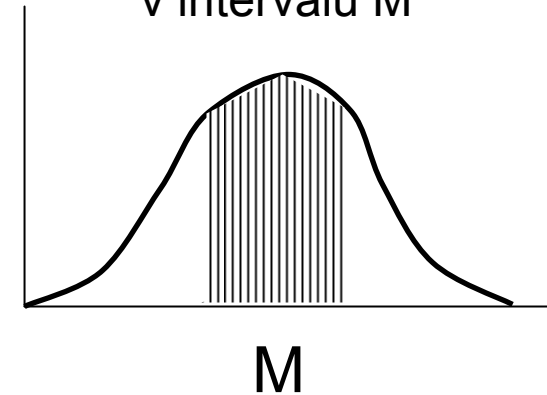
# Distribuční funkce jako užitečný nástroj pro práci s rozložením

Plocha = relativní četnost

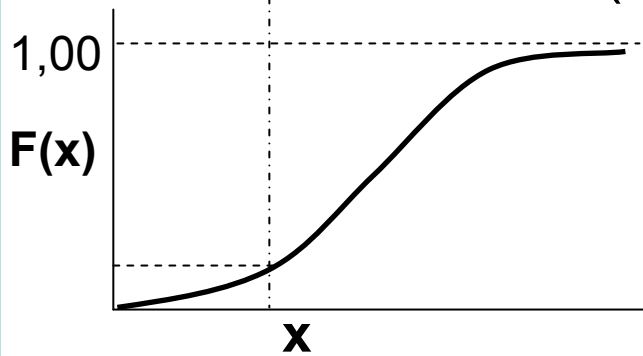


$$\int_{-\infty}^{\infty} \varphi(x) d(x) = 1$$

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



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



$\Phi(x)$  ... distribuční funkce

$$P(X \leq x) = \int_M \varphi(x) d(x)$$

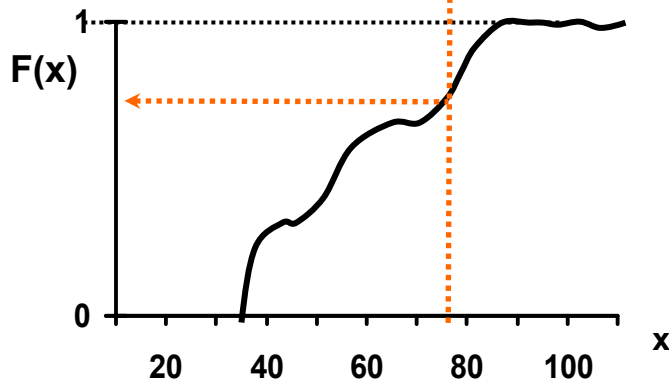
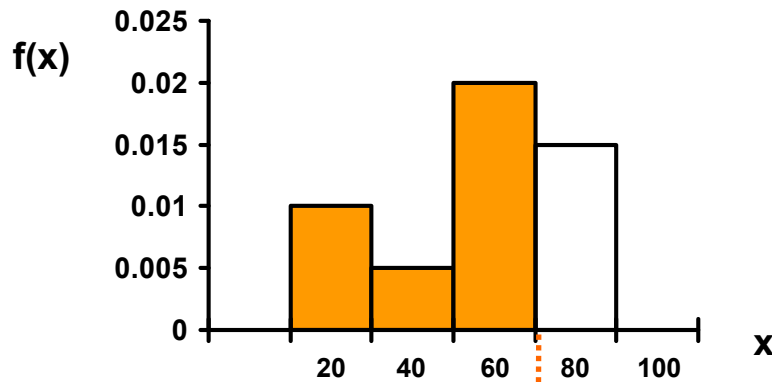
Známe-li distribuční funkci, pak známe rozložení sledované veličiny.

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

# Jak vznikají informace ?

## - frekvenční sumarizace spojitéch dat

### Grafické výstupy z frekvenční tabulky – spojitá data



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

**KVANTIL**

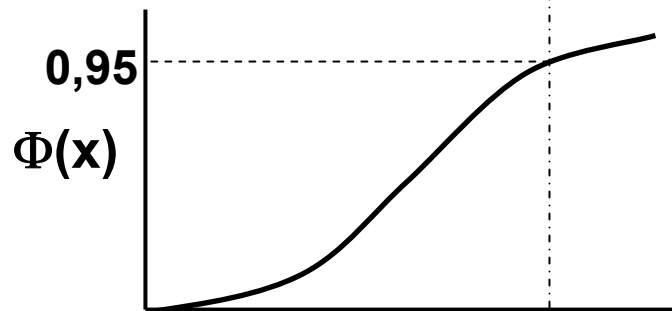
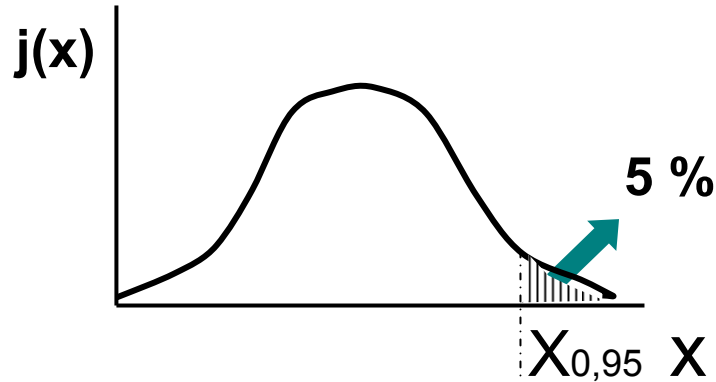
$X_{0.1}$ ;  $X_{0.9}$ ;  $X_{0.5}$ ;  $X_{\theta}$

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

$\theta = 0,95$  ... 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**

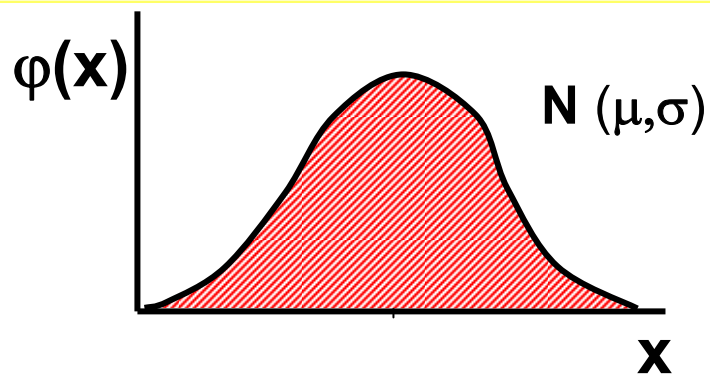


## 5. Modelová rozložení



# Rozložení hodnot jako model

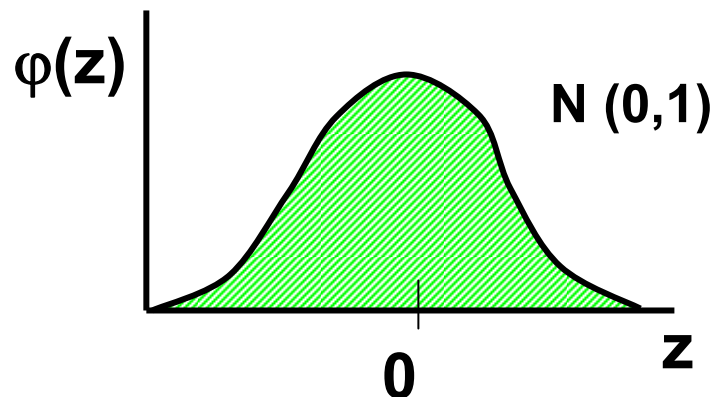
## Příklad - Normální rozložení



$$\varphi(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$z = \frac{x - \mu}{\sigma}$$

Standardizovaná forma

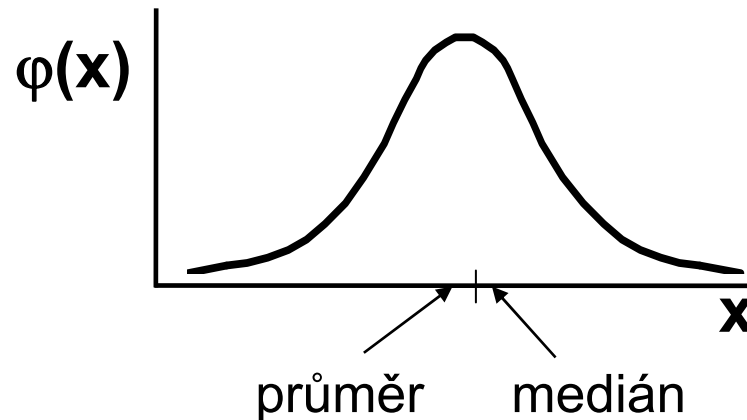


$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

# Parametry charakterizující normální rozložení a jejich význam

$$E(x) \sim \bar{x} \sim \mu$$
$$D(x) \sim s^2 \sim \sigma^2$$



a)

$$\mu \sim \bar{x}$$

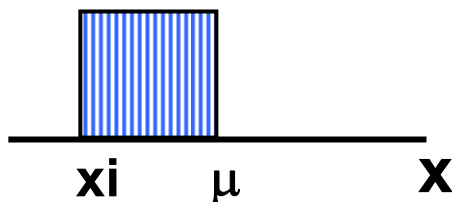
**průměr - ukazatel středu**

b)

$$\sigma^2 \sim s^2$$

rozptyl

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



c)

$$\sigma \sim s$$

**směrodatná odchylka**

$$s = \sqrt{s^2}$$

**Pravidlo  $\pm 3s$**

d)

**koeficient variance**

$$c = s / \bar{x}$$

# Interpretace parametrů normálního rozložení

Parametr středu

Parametr šířky

$$E(x) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

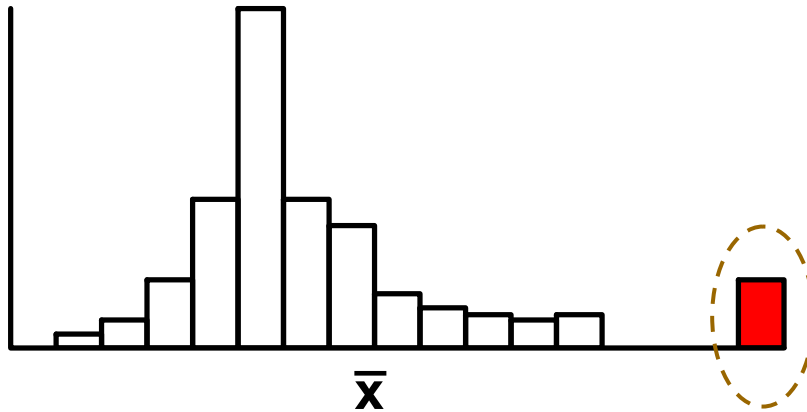
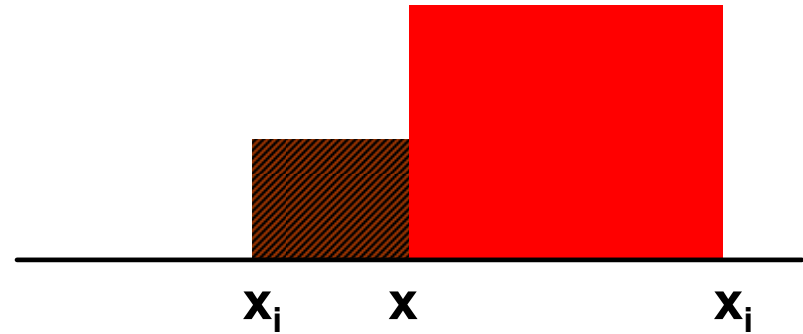
$$D(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \frac{\left[ \sum_{i=1}^n x_i \right]^2}{n}}{n-1} = s^2$$

Směrodatná odch. (S.D.)

$$\sqrt{s^2} = s$$

# Rozptyl není univerzálním ukazatelem variability

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



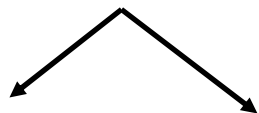
⇒ neúměrně zvýší  $s^2$



# Interpretace parametrů normálního rozložení

## Variační koeficient $c$ (koeficient variance)

Př.: 2 soubory dat - koncentrace Zn v rostlinné tkáni



$$\bar{x}_1 = 100$$
$$s_1 = 10$$

$$\bar{x}_2 = 10$$
$$s_2 = 2,6$$

$$c_1 = \frac{s_1}{\bar{x}_1} = 0,10$$

$$c_2 = \frac{s_2}{\bar{x}_2} = 0,26$$

# Normální rozložení jako model

## I. Použitelnost modelu

### A) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,4; 3,8

n = 7 opakování

medián = 1,8

$$\text{průměr} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{7} \sum_{i=1}^7 x_i = \frac{1}{7} (1,2 + 1,4 + 1,6 + 1,8 + 2,0 + 2,4 + 3,8) = \frac{1}{7} 14,2 = 2,03$$

$$\text{rozptyl (s}^2\text{)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^7 (x_i - 2,03)^2}{6} = 0,766$$

$$\text{sm. odchylka (s)} = \sqrt{s^2} = \sqrt{0,766} = 0,875$$



Je předpoklad normálního rozložení oprávněný ?  
Jaký předpokládáte možný rozsah hodnot tohoto znaku ?



# Normální rozložení jako model

## I. Použitelnost modelu

### B) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,2; 2,4; 3,8; 8,9

n = 9 opakování

medián = 2




$$\text{průměr} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{9} \sum_{i=1}^9 x_i = \frac{1}{9} (1,2 + 1,4 + 1,6 + 1,8 + 2,0 + 2,2 + 2,4 + 3,8 + 8,9) = \frac{1}{9} 25,3 = 2,81$$

$$\text{rozptyl (s}^2\text{)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^9 (x_i - 2,81)^2}{8} = 5,79$$

$$\text{sm. odchylka (s)} = \sqrt{s^2} = \sqrt{5,79} = 2,269$$

Jak hodnotíte model u těchto dat ?

# Stochastické rozložení jako model

- 1 Předpoklad: Znak  $x$  je rozložen podle daného modelu ✓
- 2 Znak  $x$  je naměřen o  $n$  hodnotách s modelovými parametry:  $\bar{x}$  a  $s$   Platnost modelu ? 
- 3 Znak  $x$  je převeden na formu odpovídající tabulkovému standardu:  
$$Z_i = \frac{x - \mu}{\sigma}$$
- 4 Využije se tabelované (modelové) distribuční funkce pro testy o rozložení hodnot  $x$

# Normální rozložení jako model - příklad

## Tabulky distribuční funkce

- Data z průzkumu jsou publikována jako:

Kosti prehistorického zvířete:

$n = 2000$

**průměrná délka = 60 cm**

**sm. odchylka ( $s$ ) = 10 cm**

✓ **Předpokládáme, že je oprávněný model normálního rozložení**


? Jaká je pravděpodobnost, že by velikost dané kosti překročila velikost 66 cm:  $P(x > 66)$  ?  $Z = \frac{x - \mu}{\sigma}$

$P(x > 66) = 1 - P(x \leq 66)$  a platí, že  $P(X \leq x) = F(X)$

tedy  $P(x > 66) = 1 - P(x \leq 66) = 1 - P\left(\frac{x - m}{s} \leq \frac{66 - 60}{10}\right) = 1 - F(0,6) = 0,27425$

? Kolik kostí mělo zřejmě délku větší než 66 cm ?  $P(x > 66) * n = 0,27425 * 2000 = 548$

? Jaký podíl kostí ležel svou délkou v rozsahu  $x$  od 60 cm do 66 cm ?

$P(60 < x < 66) = P\left(\frac{60 - 60}{10} < Z < \frac{66 - 60}{10}\right) = F(0,6) - F(0) = 0,22575$   22,6% kostí leží v rozsahu 60-66cm

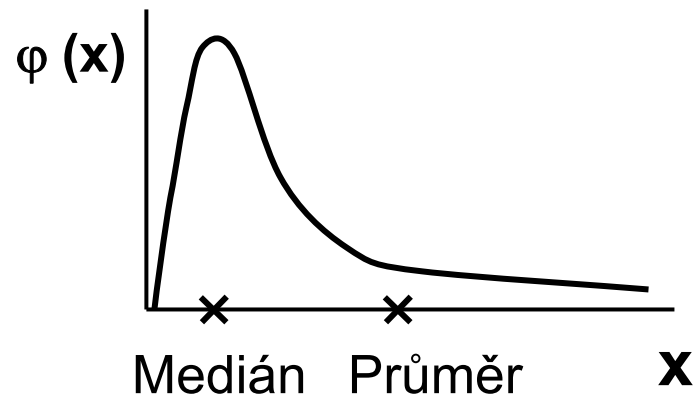
# Stručný přehled modelových rozložení I.

Rozložení	Parametry	Stručný popis
<b>Normální</b>	Průměr ( $\mu$ ) Rozptyl ( $\sigma^2$ )	Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci.
<b>Log-normální</b>	Medián Geometrický průměr Rozptyl ( $\sigma^2$ )	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
<b>Weibullovo</b>	$\alpha$ - parametr tvaru $\beta$ - parametr rozsahu hodnot	Změnou parametru $\alpha$ lze modelovat distribuci doby přežití, např. stresovaného organismu. Rozložení využívané i jako model k odhadu $LC_{50}$ nebo $EC_{50}$ u testů toxicity.
<b>Rovnoměrné</b>	Medián Geometrický průměr Rozptyl ( $\sigma^2$ )	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
<b>Triangulární</b>	$f(x) = [b - ABS(x - a)] / b^2$ $a - b < x < a + b$	Pravděpodobnostní funkce pro typ rozložení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové.
<b>Gamma</b>	Parametry distribuční funkce: $\alpha$ - parametr tvaru $\beta$ - parametr rozsahu hodnot	Umožňuje flexibilně modelování distribučních funkcí nejrůznějších tvarů. Např. $\chi^2$ rozložení je rozložení typu Gamma. Gamma rozložení s $a = 1$ je známo jako exponenciální rozložení.

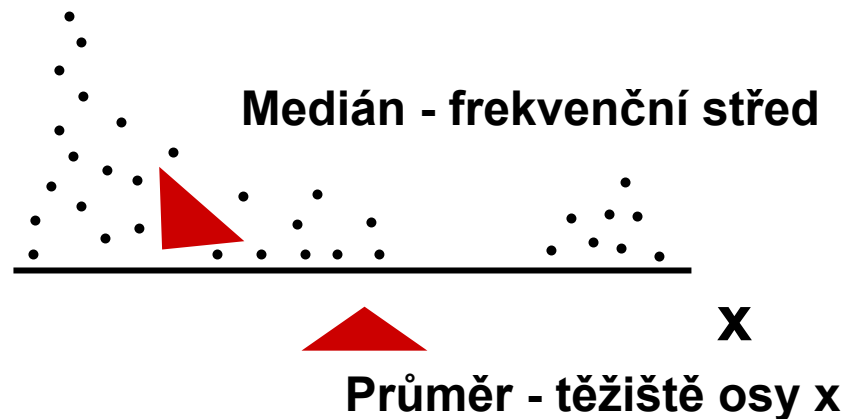
# Stručný přehled modelových rozložení II.

Rozložení	Parametry	Stručný popis
<b>Beta</b>	<p>Parametry distribuční funkce:</p> <p><math>\alpha</math> - parametr tvaru</p> <p><math>\beta</math> - parametr rozsahu hodnot</p>	<p>Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní při popisu změn hodnot proměnné v ohraničeném intervalu.</p>
<b>Studentovo</b>	<p>Stupně volnosti - uvažuje velikost vzorku</p> <p>Průměr</p> <p>Rozptyl</p>	<p>Simuluje normální rozložení pro menší vzorky čísel. Pro větší soubory (<math>n &gt; 100</math>) se limitně blíží k normálnímu rozložení.</p>
<b>Pearsonovo</b>	<p>Stupně volnosti - uvažuje velikost vzorku</p>	<p>Slouží především k porovnání četností jevů ve dvou a více kategoriích. Používá se k modelování rozložení odhadu rozptylu normálně rozložených dat.</p>
<b>Fisher-Snedecorovo</b>	<p>Dvojí stupně volnosti - uvažuje velikost dvou vzorků</p>	<p>Používá se k testování hodnot průměrů - F test pro porovnání dvou výběrových rozptylů; F test, ANOVA atd.</p>

# Log-normální rozložení jako častý model reálných znaků

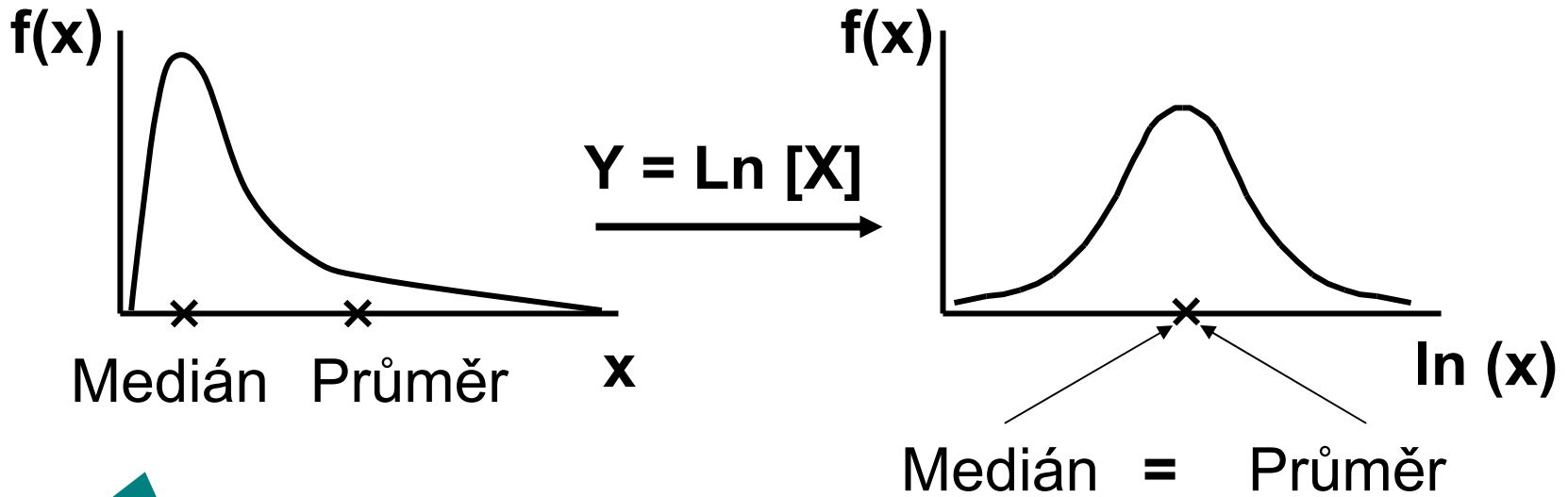


**U asymetrických rozložení je medián velmi vhodným alternativním ukazatelem středu**





# Log-normální rozložení lze jednoduše transformovat



**EXP (Y) = Geometrický průměr X**

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}$$

**$\bar{Y} \pm$  Standardní chyba**

# Transformace dat - legitimní úprava rozložení



Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu

## Logaritmická transformace

Logaritmická transformace je velmi vhodná pro data s odlehlými hodnotami na horní hranici rozsahu. Při porovnání průměrů u více souborů dat je pro tuto transformaci indikující situace, kdy se s rostoucím průměrem mění proporcionálně i směrodatná odchylka, a tedy jednotlivé proměnné mají stejný koeficient variance, ačkoli mají různý průměr.

Za takovéto situace přináší logaritmická transformace nejen zeslabení asymetrie původního rozložení, ale také vyšší homogenitu rozptylu proměnných. Pro transformaci se nejčastěji používá přirozený logaritmus a pokud jsou v původním souboru dat nulové hodnoty, je vhodné použít operaci  $Y = \ln(X+1)$ .

Je-li průměr logaritmovaných dat (tedy průměrný logaritmus) zpětně transformován do původních hodnot, výsledkem není aritmetický, ale geometrický průměr původních dat.

# Transformace dat - legitimní úprava rozložení



Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu

## Odmocninová transformace

Transformace je vhodná pro proměnné mající Poissonovo rozložení, tedy proměnné vyjadřující celkový počet nastání určitého jevu (spíše vzácného) v  $n$  nezávisle opakovaných pokusech. Obecněji lze tento typ transformace doporučit v případě normalizace dat typu počtu jedinců (buněk, apod.). Jde o transformaci:

$$Y = \sqrt{x} \quad \text{nebo} \quad Y = \sqrt{x+1} \quad \text{nebo} \quad Y = \sqrt{x} + \sqrt{x+1}$$

Transformace s přičtenou hodnotou 1 jsou efektivní, pokud  $X$  nabývá velmi malých nebo nulových hodnot. Situace indikující vhodnost odmocninové transformace je také proporcionalita výběrového rozptylu a průměru, tedy obecně jestliže  $s_x^2 = k$  (výběrový průměr).

# Transformace dat - legitimní úprava rozložení

## Arcsin transformace

Tzv. **úhlová transformace** - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi  $n$  hodnocenými jedinci - tedy pro data mající binomické rozložení. Pokud se určitý znak vyskytuje  $r$ -krát mezi  $n$  možnostmi (jedinci, opakováními), pak lze vyjádřit relativní četnost jeho výskytu jako  $p = r/n$  s variabilitou  $p \cdot (1-p)/n$ . Arcsin transformace odstraní ze souborů dat podíly blízké 0 nebo 1, a tak efektivně sníží variabilitu odhadů středu. Transformace však není schopná odstranit variabilitu vyvolanou rozdílným počtem opakování v jednotlivých variantách - v takovém případě lze doporučit provedení vážených transformací dat. Velmi častou formou této transformace je:

$$Y = \arcsin \sqrt{p}$$

- tedy transformace podílů do hodnot, jejichž sinus je roven druhé odmocnině původních hodnot. Pokud celkový počet jedinců (opakování), mezi kterými je výskyt znaku monitorován, je  $n < 50$ , pak lze doporučit velmi efektivní empirická opatření pro transformaci podílů blízkých 0 nebo 1. Pro tento případ lze nahrazovat nulové podíly hodnotou  $1/4n$  a 100 % podíly hodnotou  $(n-1/4)/n$ . Pokud se mezi hodnotami vyskytuje větší množství krajních hodnot (menší než 0,2 a větší než 0,8), lze doporučit transformaci:

$$Y = \frac{1}{2} \left[ \arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}} \right]$$

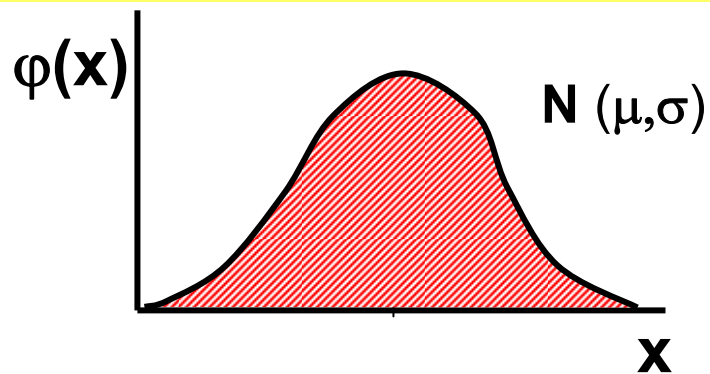


## 5. Modelová rozložení



# Rozložení hodnot jako model

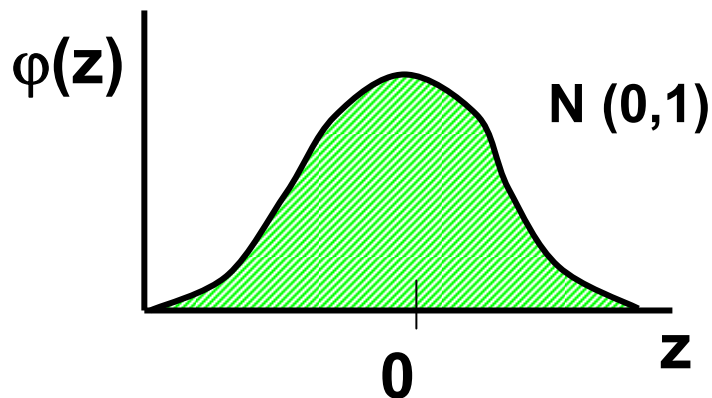
## Příklad - Normální rozložení



$$\varphi(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$z = \frac{x - \mu}{\sigma}$$

Standardizovaná forma

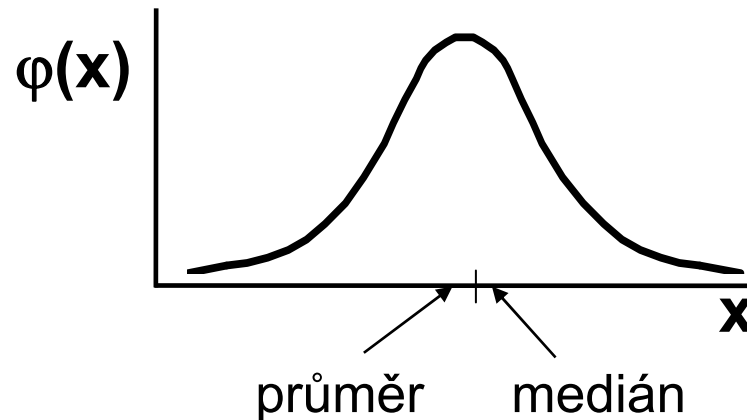


$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

# Parametry charakterizující normální rozložení a jejich význam

$$E(x) \sim \bar{x} \sim \mu$$
$$D(x) \sim s^2 \sim \sigma^2$$



a)

$$\mu \sim \bar{x}$$

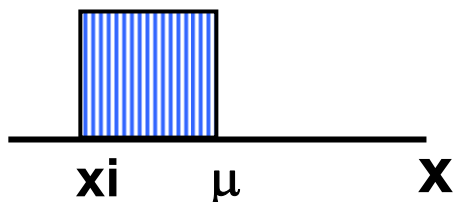
**průměr - ukazatel středu**

b)

$$\sigma^2 \sim s^2$$

rozptyl

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



c)

$$\sigma \sim s$$

**směrodatná odchylka**

$$s = \sqrt{s^2}$$

**Pravidlo  $\pm 3s$**

d)

**koeficient variance**

$$c = s / \bar{x}$$

# Interpretace parametrů normálního rozložení

Parametr středu

Parametr šířky

$$E(x) = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$D(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \frac{\left[ \sum_{i=1}^n x_i \right]^2}{n}}{n-1} = s^2$$

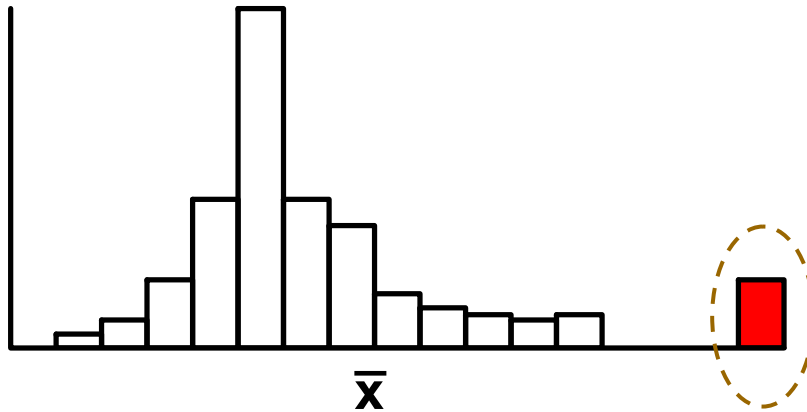
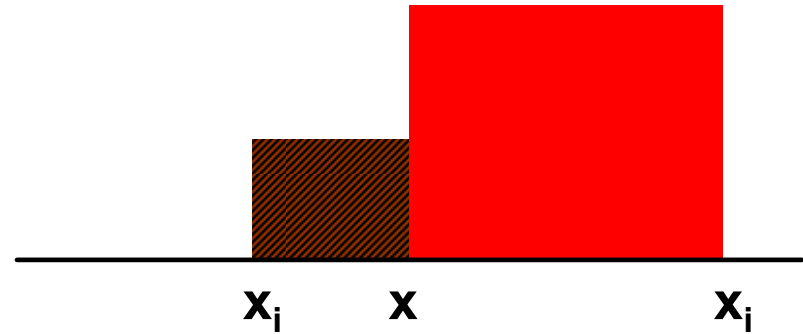
Směrodatná odch. (S.D.)

$$\sqrt{s^2} = s$$



# Rozptyl není univerzálním ukazatelem variability

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

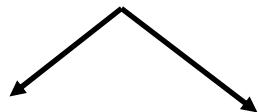


⇒ neúměrně zvýší  $s^2$

# Interpretace parametrů normálního rozložení

## Variační koeficient $c$ (koeficient variance)

Př.: 2 soubory dat - koncentrace Zn v rostlinné tkáni



$$\bar{x}_1 = 100$$
$$s_1 = 10$$

$$\bar{x}_2 = 10$$
$$s_2 = 2,6$$

$$c_1 = \frac{s_1}{\bar{x}_1} = 0,10$$

$$c_2 = \frac{s_2}{\bar{x}_2} = 0,26$$

# Normální rozložení jako model

## I. Použitelnost modelu

### A) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,4; 3,8

n = 7 opakování

medián = 1,8

$$\text{průměr} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{7} \sum_{i=1}^7 x_i = \frac{1}{7} (1,2 + 1,4 + 1,6 + 1,8 + 2,0 + 2,4 + 3,8) = \frac{1}{7} 14,2 = 2,03$$

$$\text{rozptyl (s}^2\text{)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^7 (x_i - 2,03)^2}{6} = 0,766$$

$$\text{sm. odchylka (s)} = \sqrt{s^2} = \sqrt{0,766} = 0,875$$



Je předpoklad normálního rozložení oprávněný ?  
Jaký předpokládáte možný rozsah hodnot tohoto znaku ?



# Normální rozložení jako model

## I. Použitelnost modelu

### B) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,2; 2,4; 3,8; 8,9

n = 9 opakování

medián = 2




$$\text{průměr} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{9} \sum_{i=1}^9 x_i = \frac{1}{9} (1,2 + 1,4 + 1,6 + 1,8 + 2,0 + 2,2 + 2,4 + 3,8 + 8,9) = \frac{1}{9} 25,3 = 2,81$$

$$\text{rozptyl (s}^2\text{)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^9 (x_i - 2,81)^2}{8} = 5,79$$

$$\text{sm. odchylka (s)} = \sqrt{s^2} = \sqrt{5,79} = 2,269$$

Jak hodnotíte model u těchto dat ?

# Stochastické rozložení jako model

- 1 Předpoklad: Znak  $x$  je rozložen podle daného modelu ✓
- 2 Znak  $x$  je naměřen o  $n$  hodnotách s modelovými parametry:  $\bar{x}$  a  $s$   Platnost modelu ? 
- 3 Znak  $x$  je převeden na formu odpovídající tabulkovému standardu:  
$$Z_i = \frac{x - \mu}{\sigma}$$
- 4 Využije se tabelované (modelové) distribuční funkce pro testy o rozložení hodnot  $x$

# Normální rozložení jako model - příklad

## Tabulky distribuční funkce

- Data z průzkumu jsou publikována jako:

Kosti prehistorického zvířete:

$n = 2000$

průměrná délka = 60 cm

sm. odchylka ( $s$ ) = 10 cm

✓ Předpokládáme, že je oprávněný model normálního rozložení


? Jaká je pravděpodobnost, že by velikost dané kosti překročila velikost 66 cm:  $P(x > 66)$  ?  $Z = \frac{x - \mu}{\sigma}$

$P(x > 66) = 1 - P(x \leq 66)$  a platí, že  $P(X \leq x) = F(X)$

tedy  $P(x > 66) = 1 - P(x \leq 66) = 1 - P\left(\frac{x - m}{s} \leq \frac{66 - 60}{10}\right) = 1 - F(0,6) = 0,27425$

? Kolik kostí mělo zřejmě délku větší než 66 cm ?  $P(x > 66) * n = 0,27425 * 2000 = 548$

? Jaký podíl kostí ležel svou délkou v rozsahu  $x$  od 60 cm do 66 cm ?

$P(60 < x < 66) = P\left(\frac{60 - 60}{10} < Z < \frac{66 - 60}{10}\right) = F(0,6) - F(0) = 0,22575$   22,6% kostí leží v rozsahu 60-66cm

# Stručný přehled modelových rozložení I.

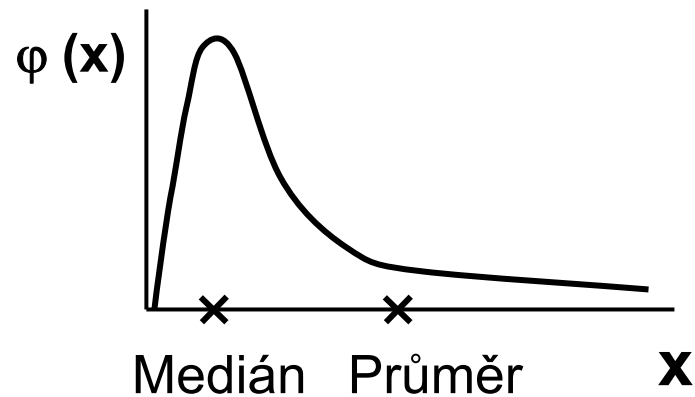
Rozložení	Parametry	Stručný popis
<b>Normální</b>	Průměr ( $\mu$ ) Rozptyl ( $\sigma^2$ )	Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci.
<b>Log-normální</b>	Medián Geometrický průměr Rozptyl ( $\sigma^2$ )	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
<b>Weibullovo</b>	$\alpha$ - parametr tvaru $\beta$ - parametr rozsahu hodnot	Změnou parametru $\alpha$ lze modelovat distribuci doby přežití, např. stresovaného organismu. Rozložení využívané i jako model k odhadu $LC_{50}$ nebo $EC_{50}$ u testů toxicity.
<b>Rovnoměrné</b>	Medián Geometrický průměr Rozptyl ( $\sigma^2$ )	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
<b>Triangulární</b>	$f(x) = [b - ABS(x - a)] / b^2$ $a - b < x < a + b$	Pravděpodobnostní funkce pro typ rozložení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové.
<b>Gamma</b>	Parametry distribuční funkce: $\alpha$ - parametr tvaru $\beta$ - parametr rozsahu hodnot	Umožňuje flexibilně modelování distribučních funkcí nejrůznějších tvarů. Např. $\chi^2$ rozložení je rozložení typu Gamma. Gamma rozložení s $\alpha = 1$ je známo jako exponenciální rozložení.

# Stručný přehled modelových rozložení II.

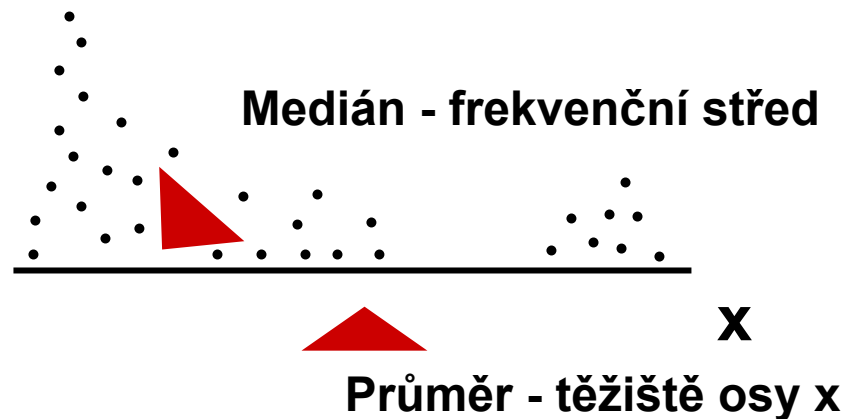
Rozložení	Parametry	Stručný popis
<b>Beta</b>	<p>Parametry distribuční funkce:</p> <p><math>\alpha</math> - parametr tvaru</p> <p><math>\beta</math> - parametr rozsahu hodnot</p>	<p>Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní při popisu změn hodnot proměnné v ohraničeném intervalu.</p>
<b>Studentovo</b>	<p>Stupně volnosti - uvažuje velikost vzorku</p> <p>Průměr</p> <p>Rozptyl</p>	<p>Simuluje normální rozložení pro menší vzorky čísel. Pro větší soubory (<math>n &gt; 100</math>) se limitně blíží k normálnímu rozložení.</p>
<b>Pearsonovo</b>	<p>Stupně volnosti - uvažuje velikost vzorku</p>	<p>Slouží především k porovnání četností jevů ve dvou a více kategoriích. Používá se k modelování rozložení odhadu rozptylu normálně rozložených dat.</p>
<b>Fisher-Snedecorovo</b>	<p>Dvojí stupně volnosti - uvažuje velikost dvou vzorků</p>	<p>Používá se k testování hodnot průměrů - F test pro porovnání dvou výběrových rozptylů; F test, ANOVA atd.</p>



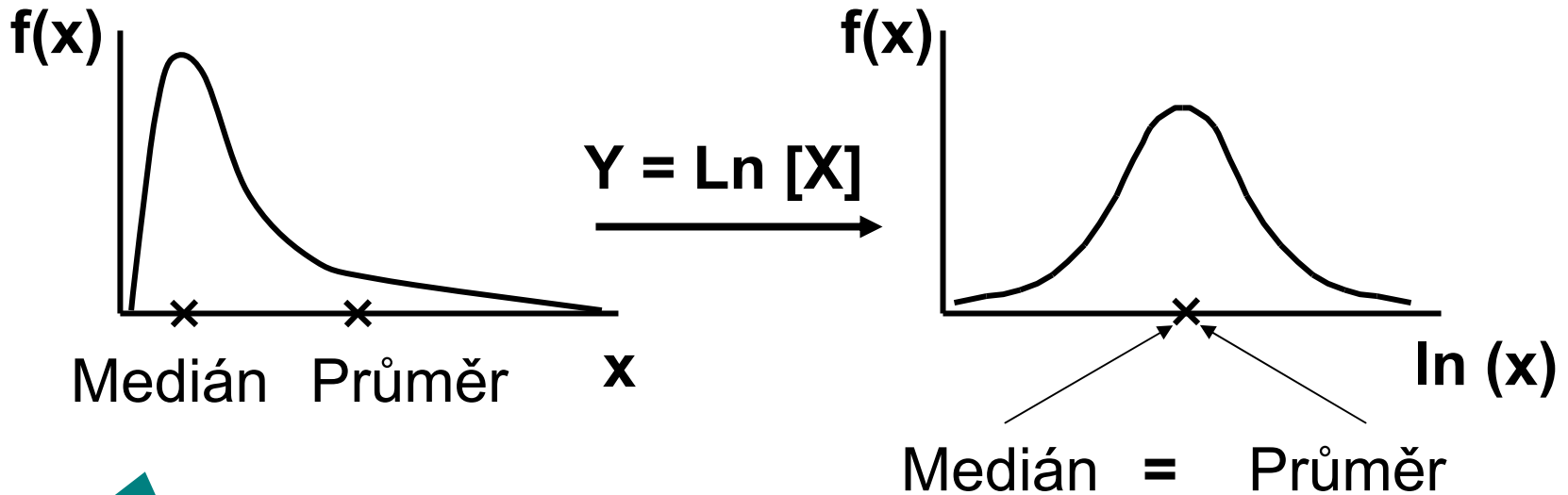
# Log-normální rozložení jako častý model reálných znaků



**U asymetrických rozložení je medián velmi vhodným alternativním ukazatelem středu**



# Log-normální rozložení lze jednoduše transformovat



$\text{EXP}(Y) = \text{Geometrický průměr } X$

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}$$

$\bar{Y} \pm \text{Standardní chyba}$

# Transformace dat - legitimní úprava rozložení



Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu

## Logaritmická transformace

Logaritmická transformace je velmi vhodná pro data s odlehlými hodnotami na horní hranici rozsahu. Při porovnání průměrů u více souborů dat je pro tuto transformaci indikující situace, kdy se s rostoucím průměrem mění proporcionálně i směrodatná odchylka, a tedy jednotlivé proměnné mají stejný koeficient variance, ačkoli mají různý průměr.

Za takovéto situace přináší logaritmická transformace nejen zeslabení asymetrie původního rozložení, ale také vyšší homogenitu rozptylu proměnných. Pro transformaci se nejčastěji používá přirozený logaritmus a pokud jsou v původním souboru dat nulové hodnoty, je vhodné použít operaci  $Y = \ln(X+1)$ .

Je-li průměr logaritmovaných dat (tedy průměrný logaritmus) zpětně transformován do původních hodnot, výsledkem není aritmetický, ale geometrický průměr původních dat.

# Transformace dat - legitimní úprava rozložení



Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu

## Odmocninová transformace

Transformace je vhodná pro proměnné mající Poissonovo rozložení, tedy proměnné vyjadřující celkový počet nastání určitého jevu (spíše vzácného) v  $n$  nezávisle opakovaných pokusech. Obecněji lze tento typ transformace doporučit v případě normalizace dat typu počtu jedinců (buněk, apod.). Jde o transformaci:

$$Y = \sqrt{x} \quad \text{nebo} \quad Y = \sqrt{x+1} \quad \text{nebo} \quad Y = \sqrt{x} + \sqrt{x+1}$$

Transformace s přičtenou hodnotou 1 jsou efektivní, pokud  $X$  nabývá velmi malých nebo nulových hodnot. Situace indikující vhodnost odmocninové transformace je také proporcionalita výběrového rozptylu a průměru, tedy obecně jestliže  $s_x^2 = k$  (výběrový průměr).

# Transformace dat - legitimní úprava rozložení

## Arcsin transformace

Tzv. **úhlová transformace** - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi  $n$  hodnocenými jedinci - tedy pro data mající binomické rozložení. Pokud se určitý znak vyskytuje  $r$ -krát mezi  $n$  možnostmi (jedinci, opakováními), pak lze vyjádřit relativní četnost jeho výskytu jako  $p = r/n$  s variabilitou  $p \cdot (1-p)/n$ . Arcsin transformace odstraní ze souborů dat podíly blízké 0 nebo 1, a tak efektivně sníží variabilitu odhadů středu. Transformace však není schopná odstranit variabilitu vyvolanou rozdílným počtem opakování v jednotlivých variantách - v takovém případě lze doporučit provedení vážených transformací dat. Velmi častou formou této transformace je:

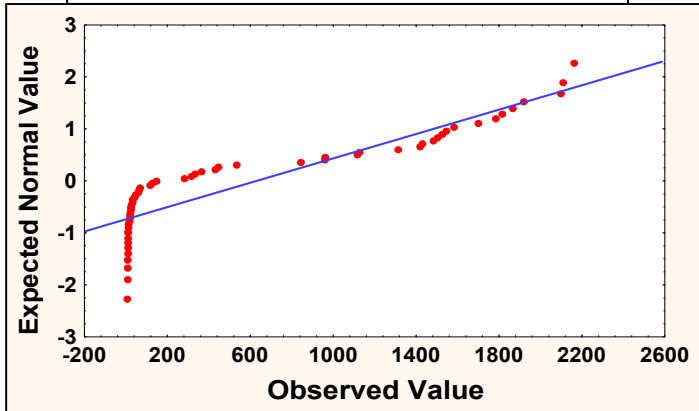
$$Y = \arcsin \sqrt{p}$$

- tedy transformace podílů do hodnot, jejichž sinus je roven druhé odmocnině původních hodnot. Pokud celkový počet jedinců (opakování), mezi kterými je výskyt znaku monitorován, je  $n < 50$ , pak lze doporučit velmi efektivní empirická opatření pro transformaci podílů blízkých 0 nebo 1. Pro tento případ lze nahrazovat nulové podíly hodnotou  $1/4n$  a 100 % podíly hodnotou  $(n-1/4)/n$ . Pokud se mezi hodnotami vyskytuje větší množství krajních hodnot (menší než 0,2 a větší než 0,8), lze doporučit transformaci:

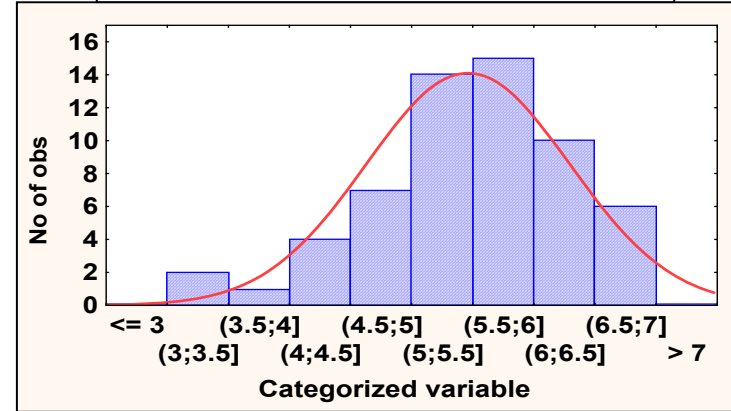
$$Y = \frac{1}{2} \left[ \arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}} \right]$$

# Testy o rozložení, grafický průzkum rozložení

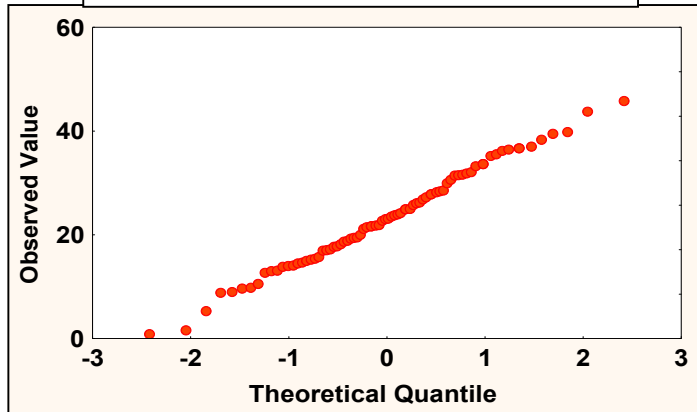
Normal probability plot



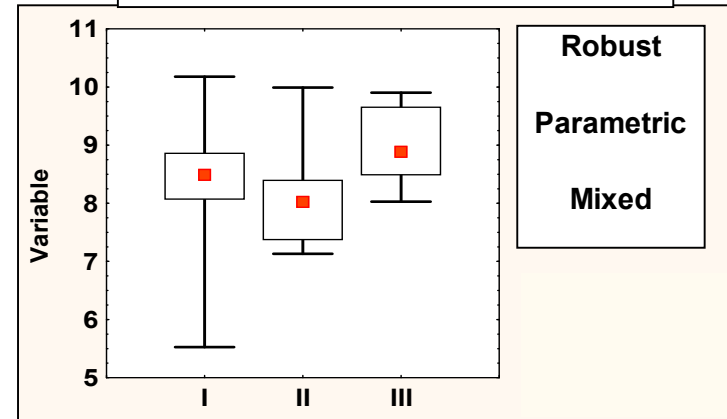
Histogram



Quantile - Quantile plot



Multiple BW plots



Testy o rozložení: Kolmogorov-Smirnov test, Shapiro-Wilks test,  $\chi^2$  test



## 6. Sumární statistika



# Sumární statistika I

## Znak X

- Medián

- Min Max

- kvantily(percentily)

- horní kvartil

- dolní kvartil

- Rozsah

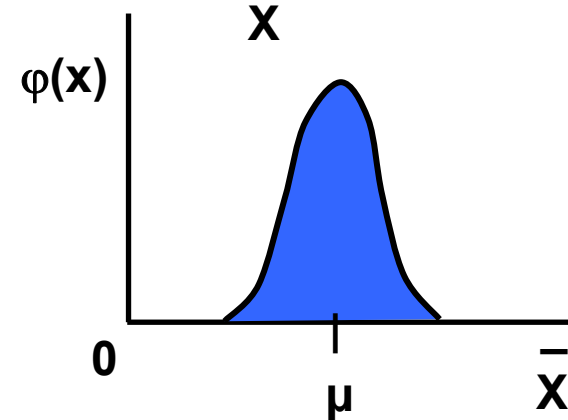
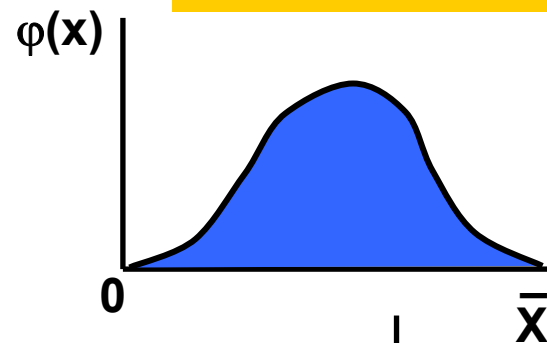
- mezikvartilová odchylka

## Střed znaku X

- průměr

- SD, SE

- interval spolehlivosti

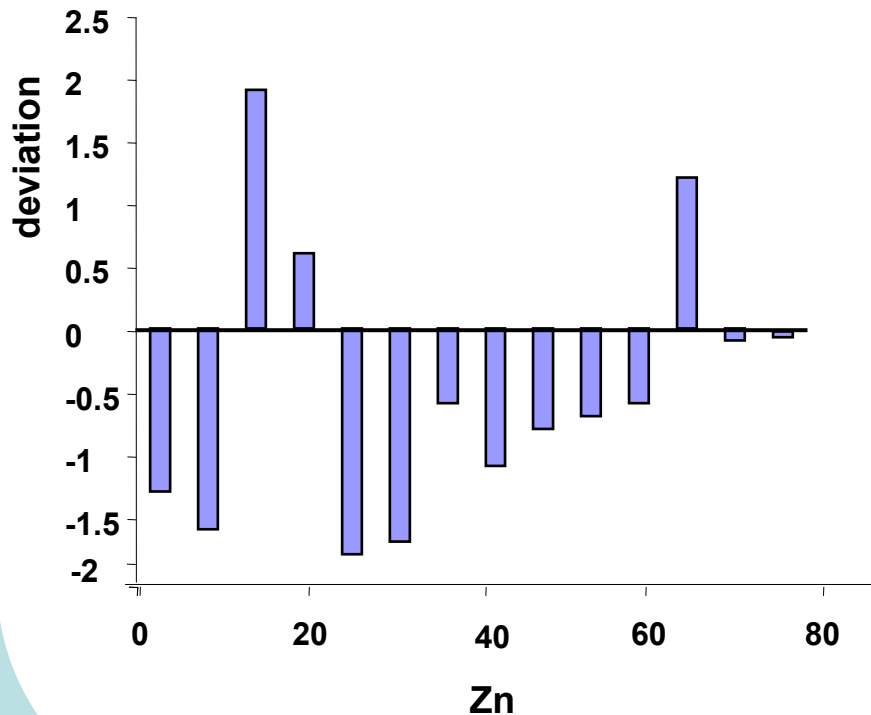




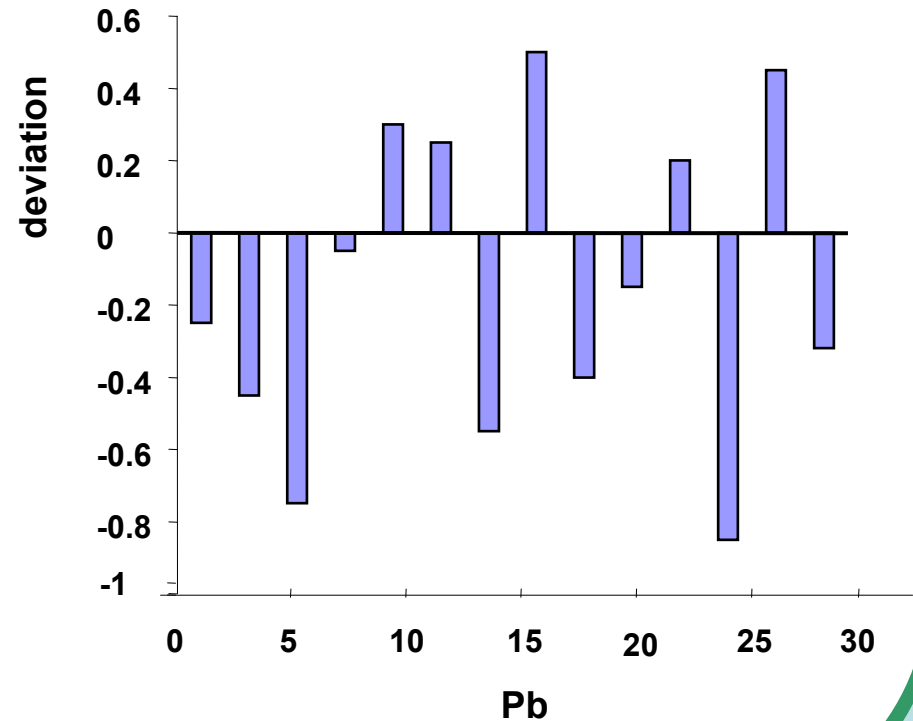
# Sumární statistika II

Následuje přehled jednoduchých grafů, které umožňují posouzení normality proměnných. Porovnejte jejich vypovídací schopnost (opět pro proměnné Zn a Pb).

## Rootgram

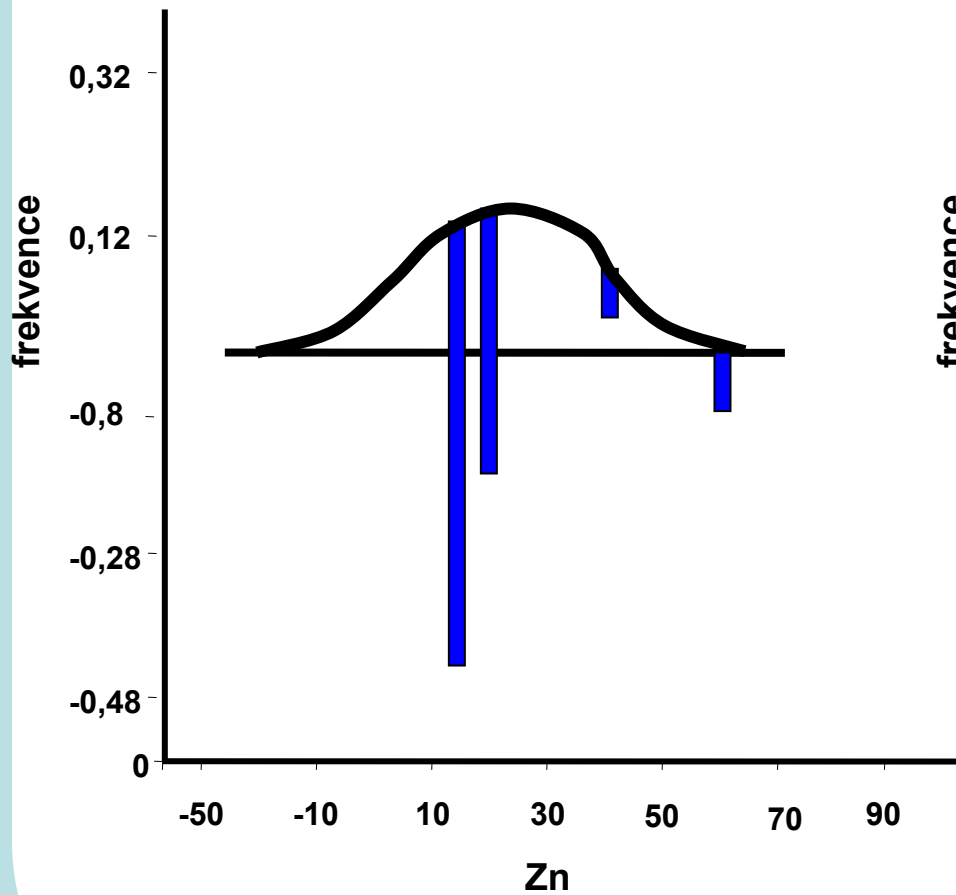


## Rootgram

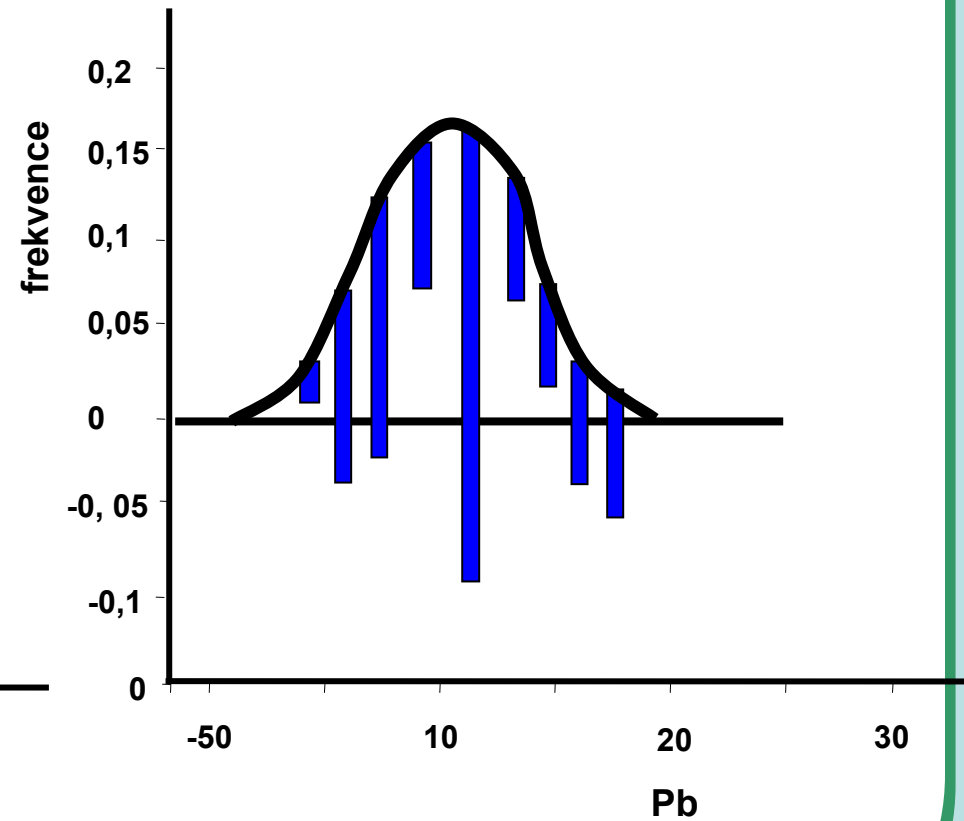


# Sumární statistika III

Hanging Histobars.

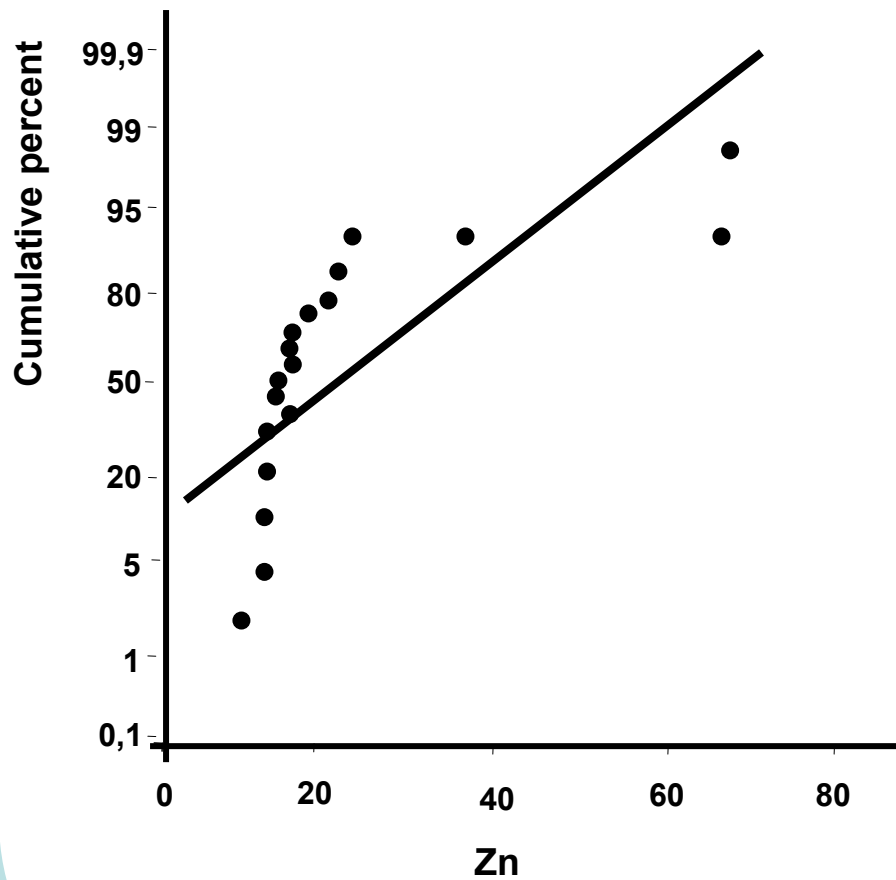


Hanging Histobars.

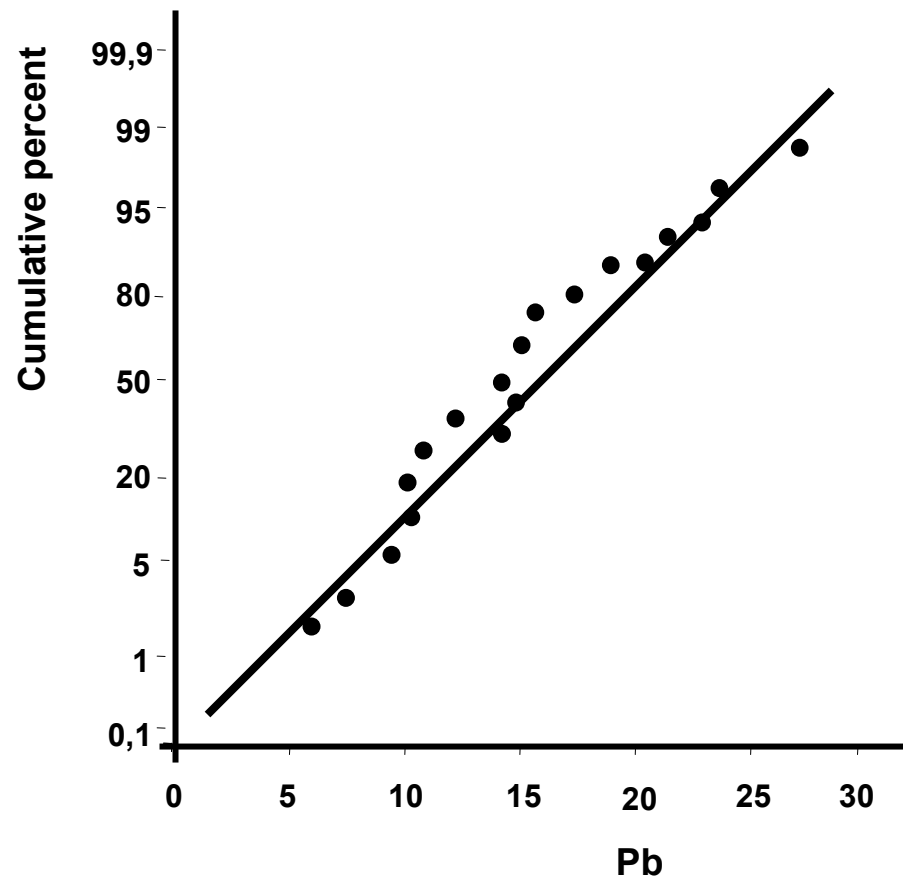


# Sumární statistika IV

## Normal Probability Plot

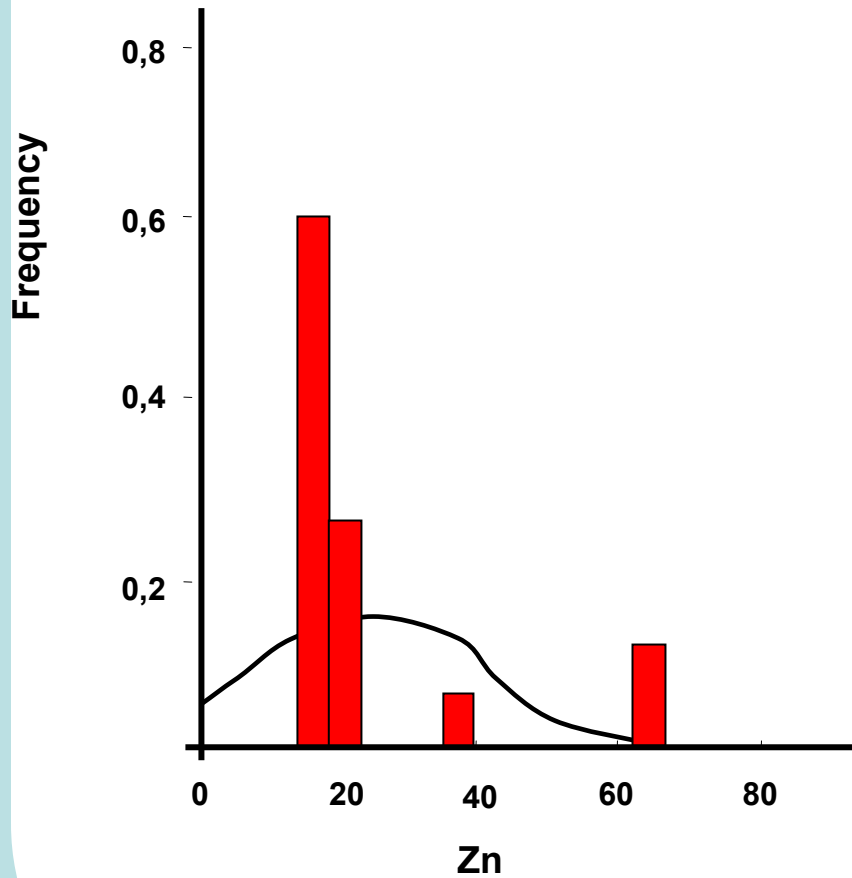


## Normal Probability Plot

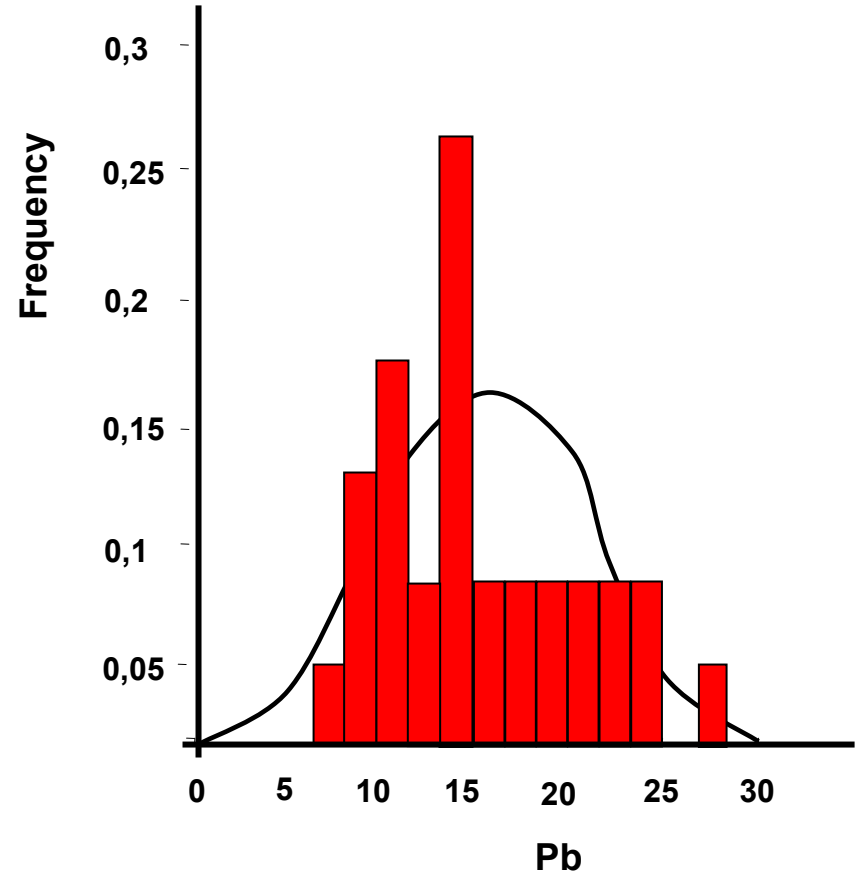


# Sumární statistika V

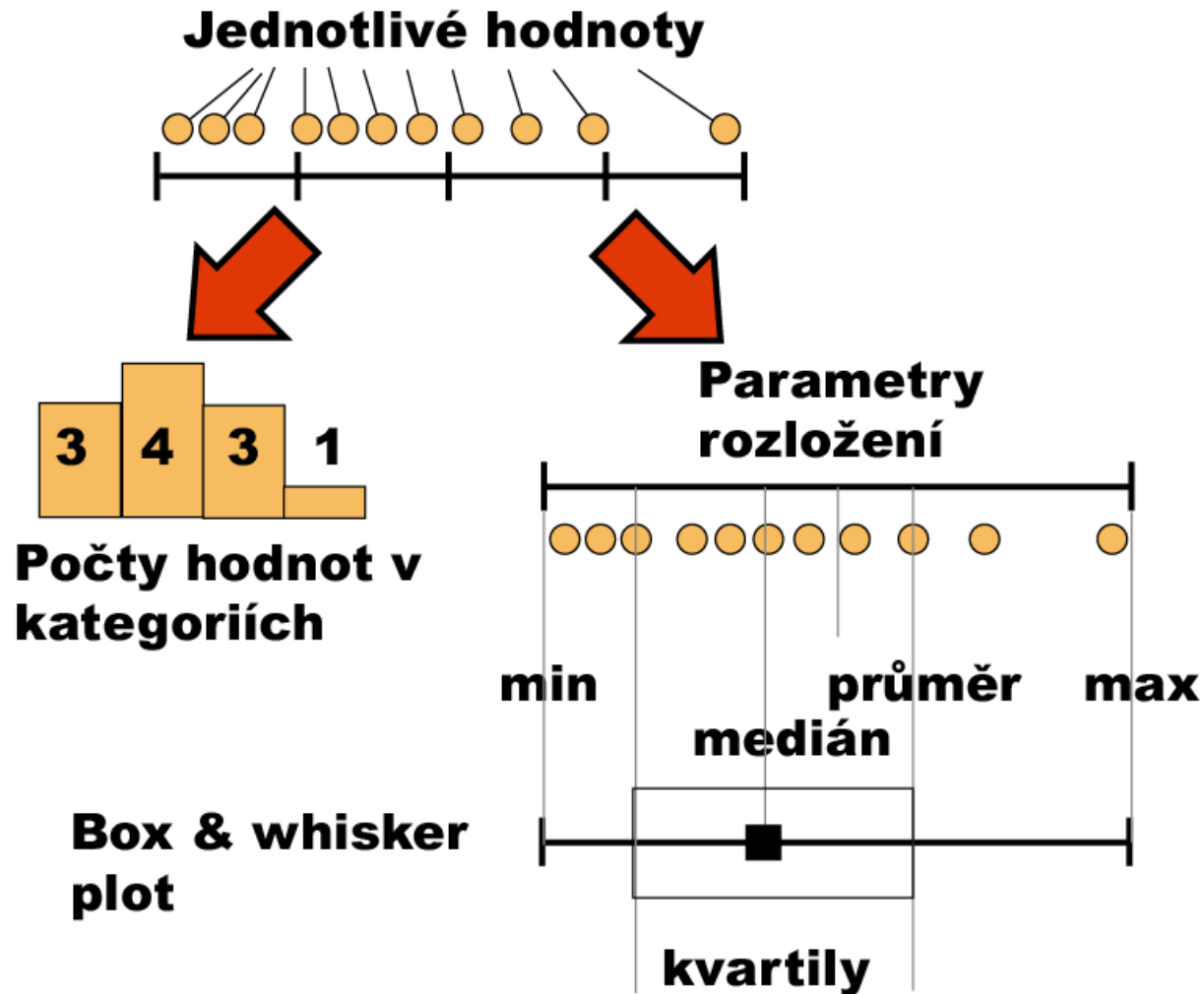
## Frequency Histogram



## Frequency Histogram



# Řada dat a její vlastnosti



# Frekvenční rozložení

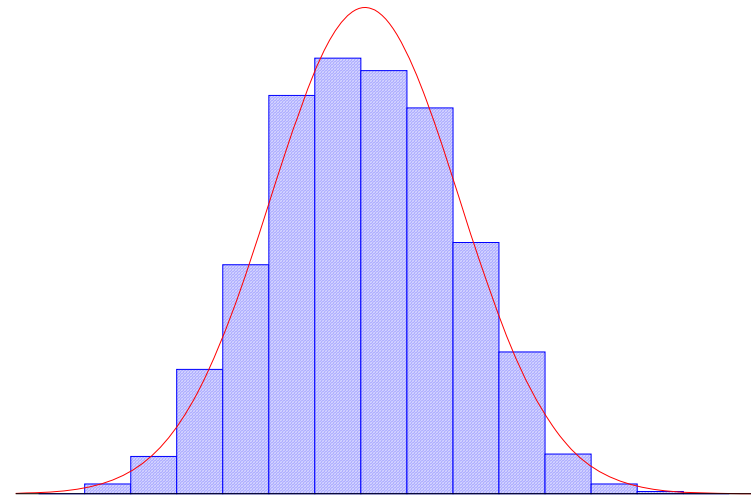
Kategorie	Četnost
B	5
C	8
D	1

## Kvalitativní data

Tabulka s četností jednotlivých kategorií.

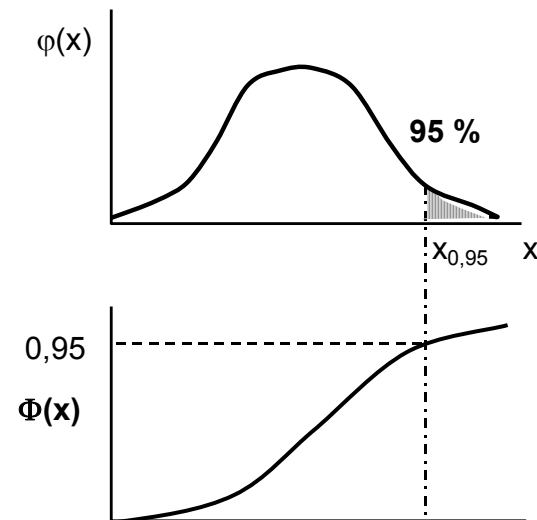
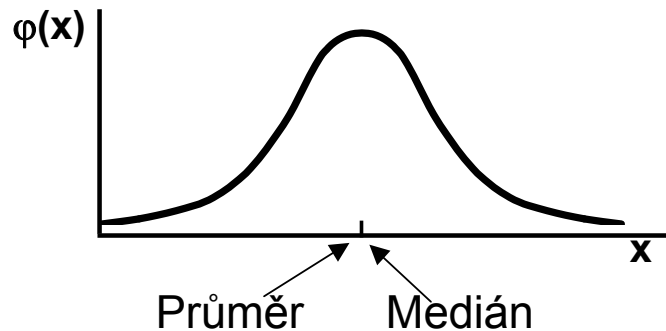
## Kvantitativní data

Četnost hodnot rozložení v jednotlivých intervalech.



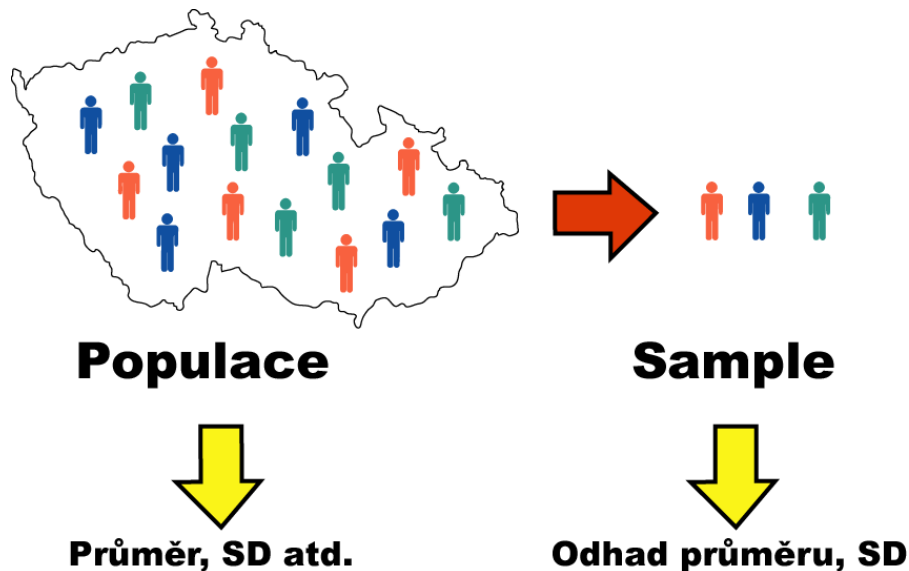
# Parametry rozložení

- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - Středu (medián, průměr, geometrický průměr)
  - Šířky rozložení (rozsah hodnot, rozptyl, směrodatná odchylka)
  - Tvaru rozložení (skewness, kurtosis)
  - Kvantily rozložení – kolik % řady dat leží nad a pod kvantilem



# Populace a vzorek

- Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozložení
- Z populace je prováděno vzorkování za účelem získání reprezentativního vzorku (**sample**) populace, toto vzorkování by mělo být náhodné, důležitá je také velikost vzorku, ze vzorku získáme odhady parametrů rozložení



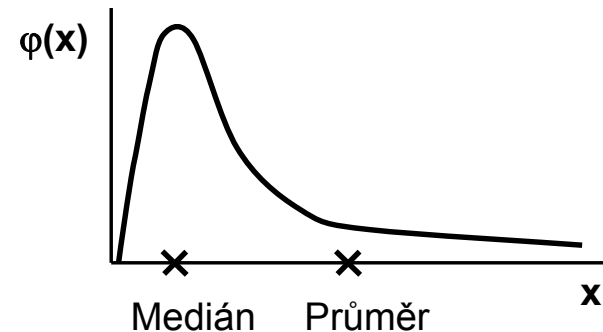
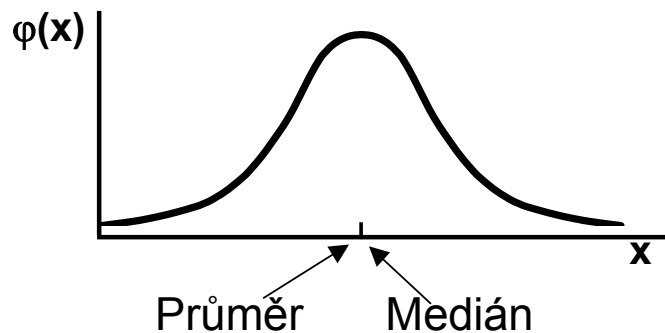


# Ukazatele středu rozložení I

- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde  $x_i$  jsou jednotlivé hodnoty a  $n$  jejich počet

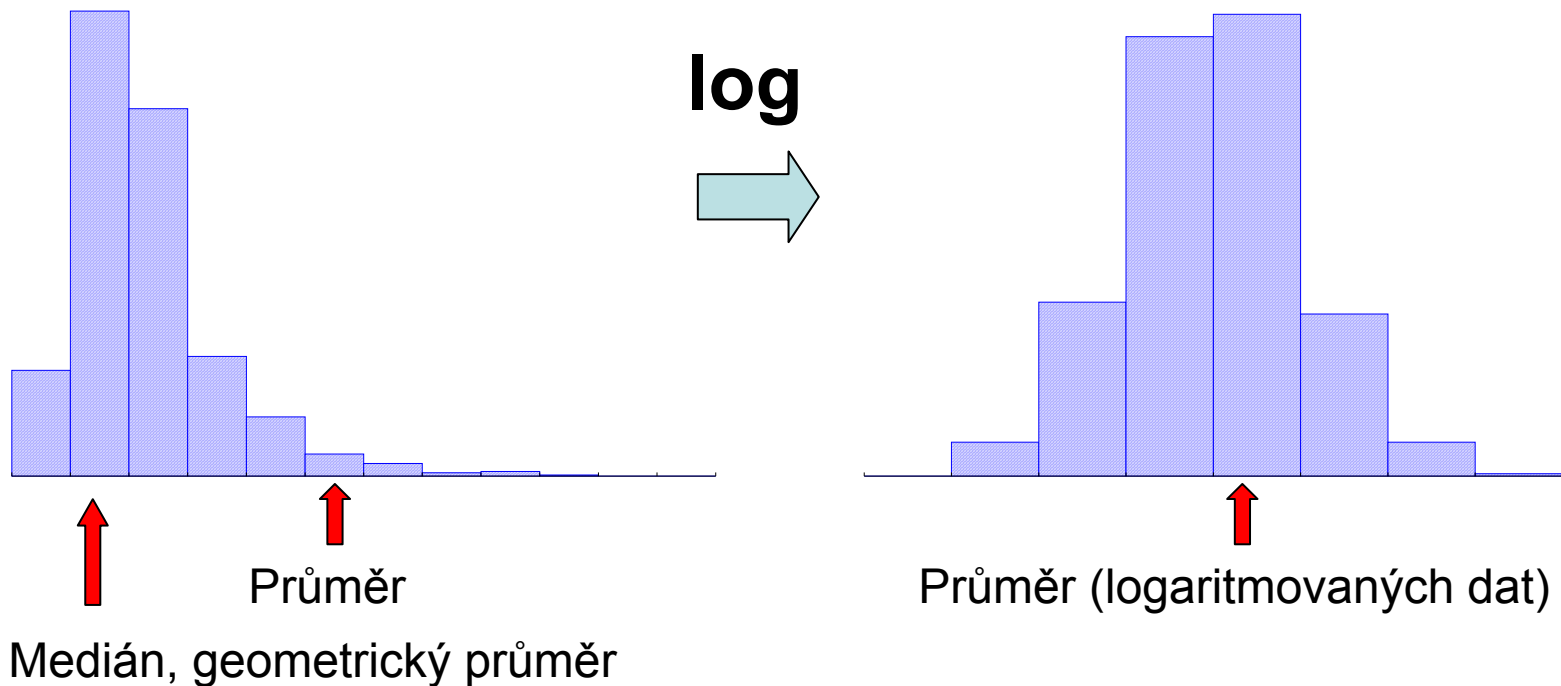
$$E(x) = \bar{x} = \sum_{i=1}^n \frac{x_i}{n}$$

- **Medián** – jde vlastně o 50% kvantil, tj. polovina hodnot leží nad a polovina pod mediánem
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné



# Ukazatele středu rozložení II.

- Geometrický průměr – antilogaritmus průměru logaritmovaných dat, je vhodný pro doleva asymetrická data (lognormální rozložení), která jsou v biologii velmi častá, jeho hodnota v podstatě odpovídá mediánu
- Takto asymetrická data je možné převést logaritmickou transformací na normální rozložení

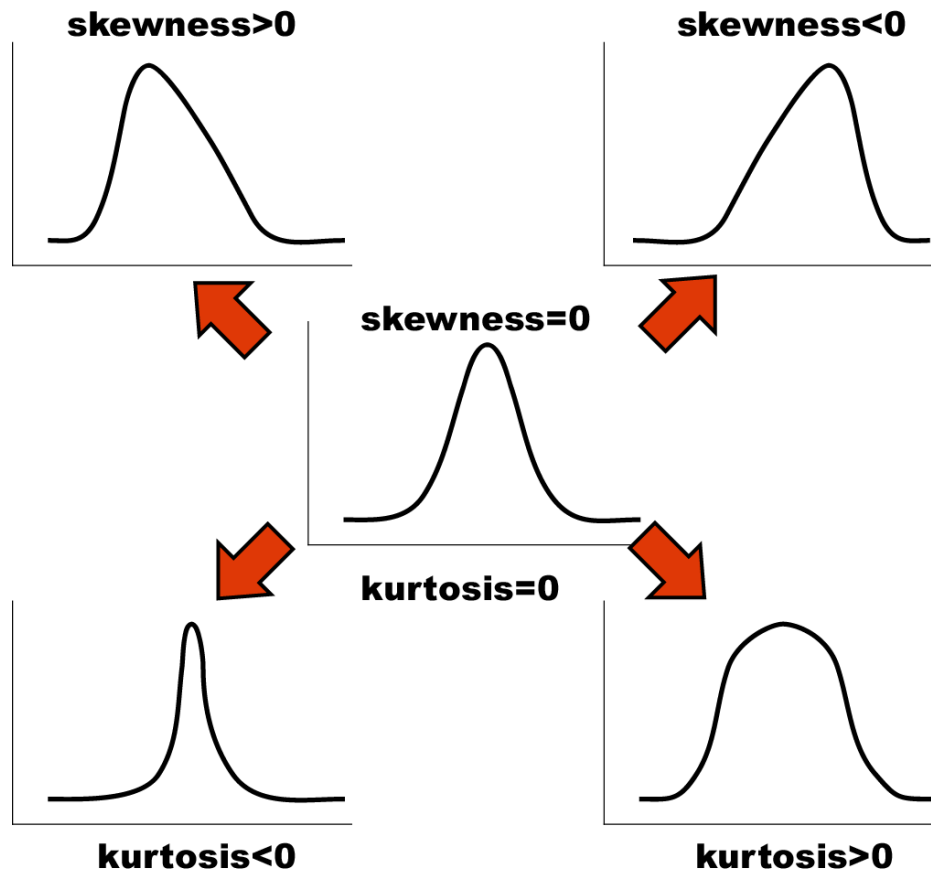


# Ukazatele šířky rozložení

- **Rozptyl** je ukazatelem šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru. 
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$
- Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení
- **Směrodatná odchylka** je druhá odmocnina z rozptylu
- **Koeficient variance** - podíl SD ku průměru (u normálního rozložení by se 95% hodnot mělo vejít do průměr  $\pm 3$  SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozložení – ukazatel problémů s normalitou dat

# Ukazatele tvaru rozložení

- **Skewness** – ukazatel „šikmosti“ rozložení, asymetrie rozložení
- **Kurtosis** – ukazatel „špičatosti/plochosti“ rozložení



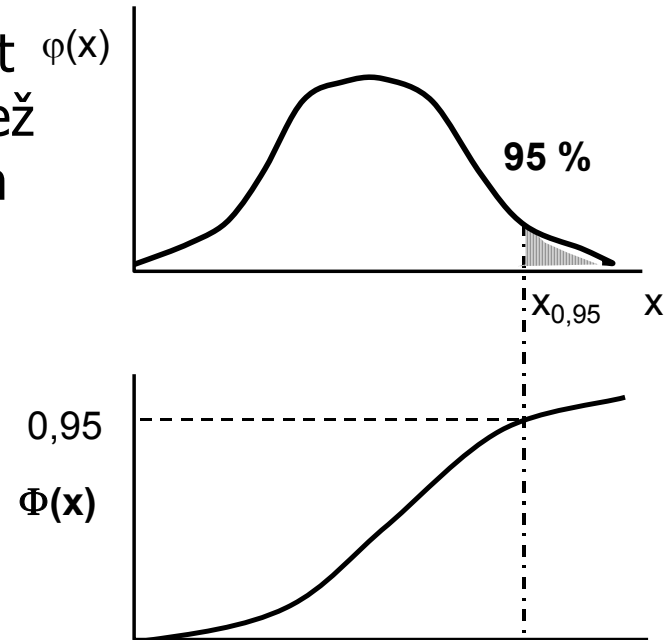
# Další parametry rozložení

---

- **Počet hodnot** – důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Střední chyba odhadu průměru** - je založena na směrodatné odchylce rozložení a **počtu hodnot**, vlastně jde o směrodatnou odchylku rozložení průměru. Říká jak přesný je náš výpočet průměru. Čím větší počet hodnot rozložení, tím je náš odhad skutečného průměru přesnější.
- **Suma hodnot**
- **Modus** – nejčastější hodnota, vhodný např. při kategoriálních datech
- **Minimum, maximum**
- **Rozsah hodnot**
- **Harmonický průměr** - převrácená hodnota průměru převrácených hodnot (vždy platí harmonický průměr < geometrický průměr < aritmetický průměr)

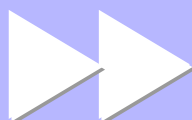
# Distribuční funkce

- Definice kvantilu dle distribuční funkce - Kvantil rozložení ( $X_{0,95}$ ) je číslo, jehož hodnota distribuční funkce je rovna pravděpodobnosti, pro kterou je kvantil definován ( $\Phi(x)$  ... distribuční funkce), tj. pokud vezmeme nějaký bod rozložení a porovnáme jej s tímto bodem (kvantilem), máme 95% pravděpodobnost, že bude menší než hodnota kvantilu ( $X_{0,95}$ ).
- Pomocí distribuční funkce můžeme určit jaký podíl hodnot rozložení je menší než daná hodnota – využití při statistických testech



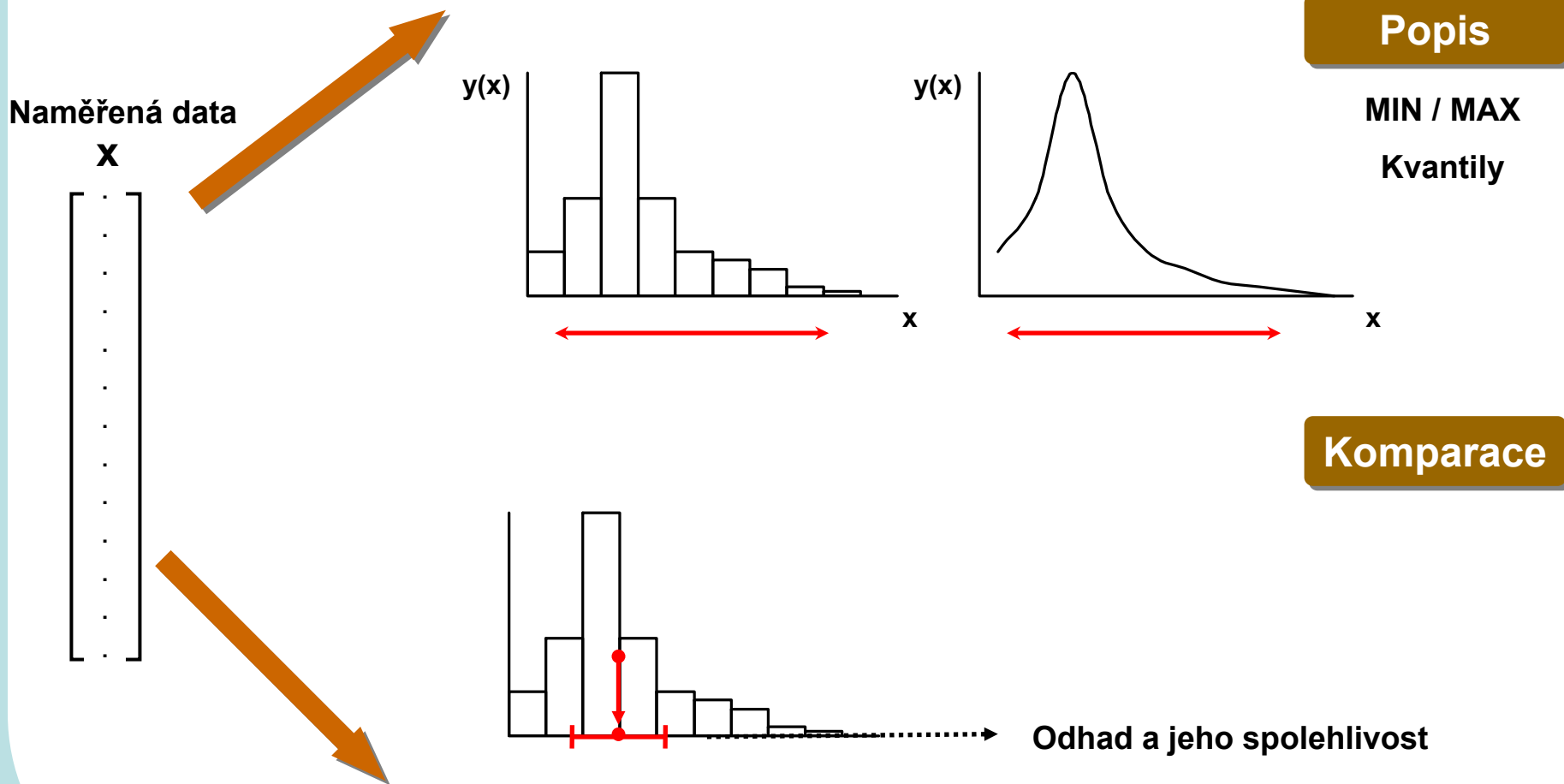


## 7. Strategie sumarizace a zviditelnění dat



# Zviditelnění dat a jeho zásadní strategie

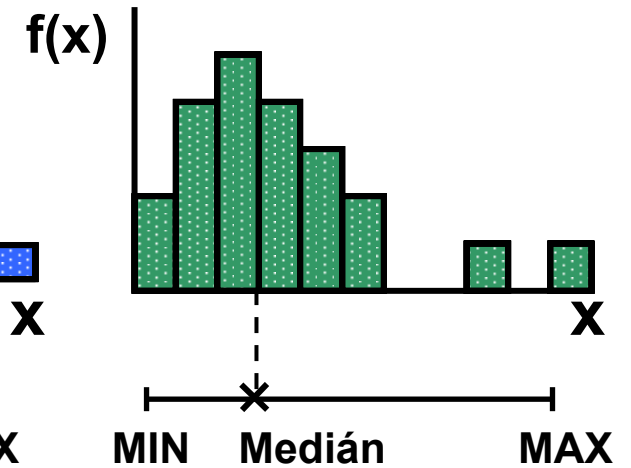
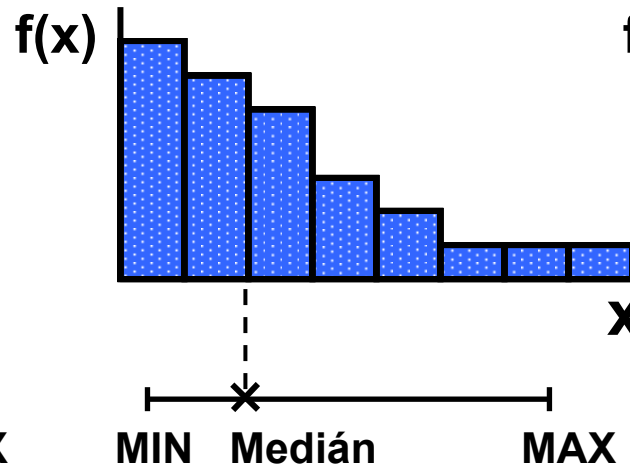
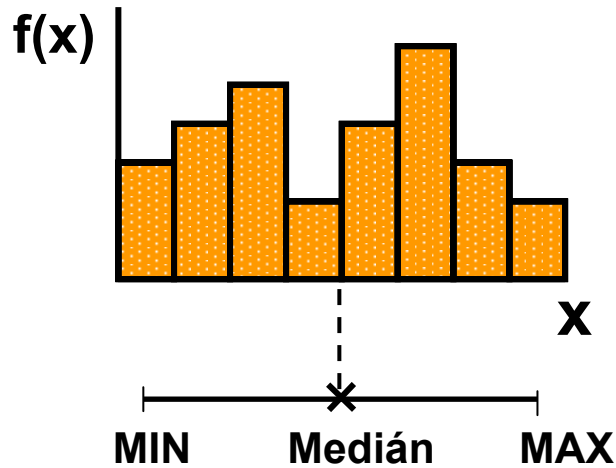
## A. Zviditelnění reálných dat – výběrové rozložení



## B. Sumarizace odhadem „zástupců“ primárních dat



# Formální popis tvaru rozložení



Z% Medián Y%  
kvantil kvantil

Z% Medián Y%  
kvantil kvantil

Z% Medián Y%  
kvantil kvantil

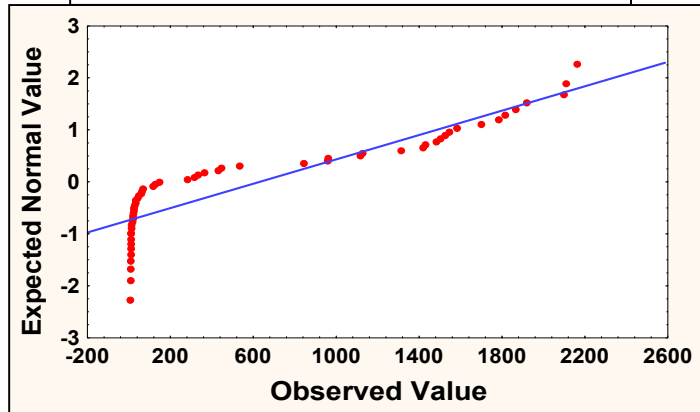
**Medián** = 50 % kvantil = frekvenční střed

**MAX - MIN** = rozsah (range)

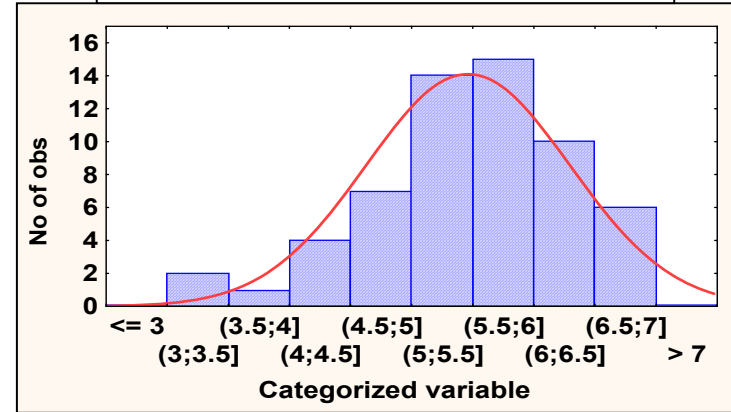
**Modus** = nejčastější hodnota

# Testy o rozložení, grafický průzkum rozložení

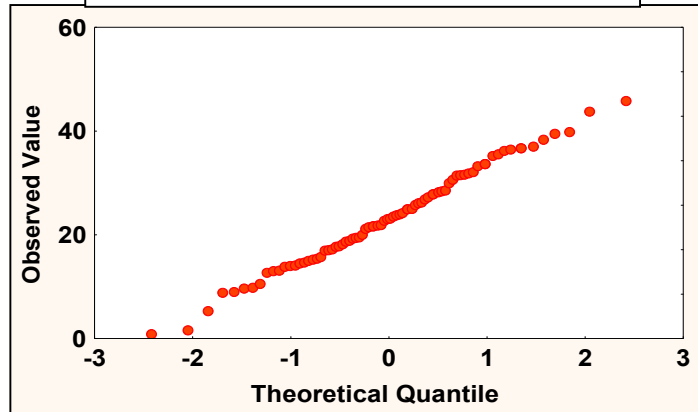
### Normal probability plot



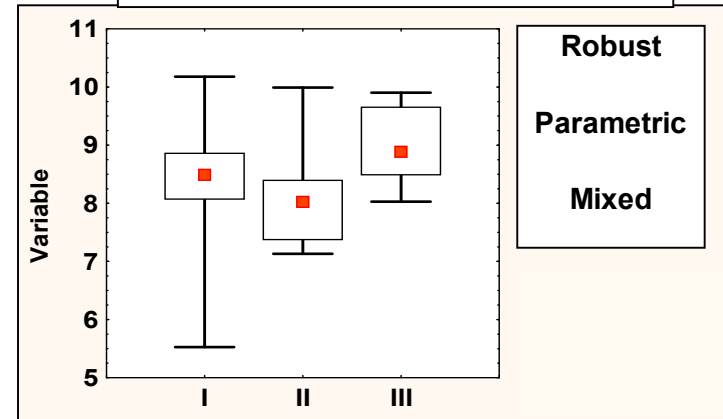
### Histogram



### Quantile - Quantile plot



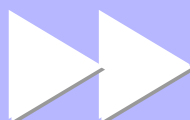
### Multiple BW plots



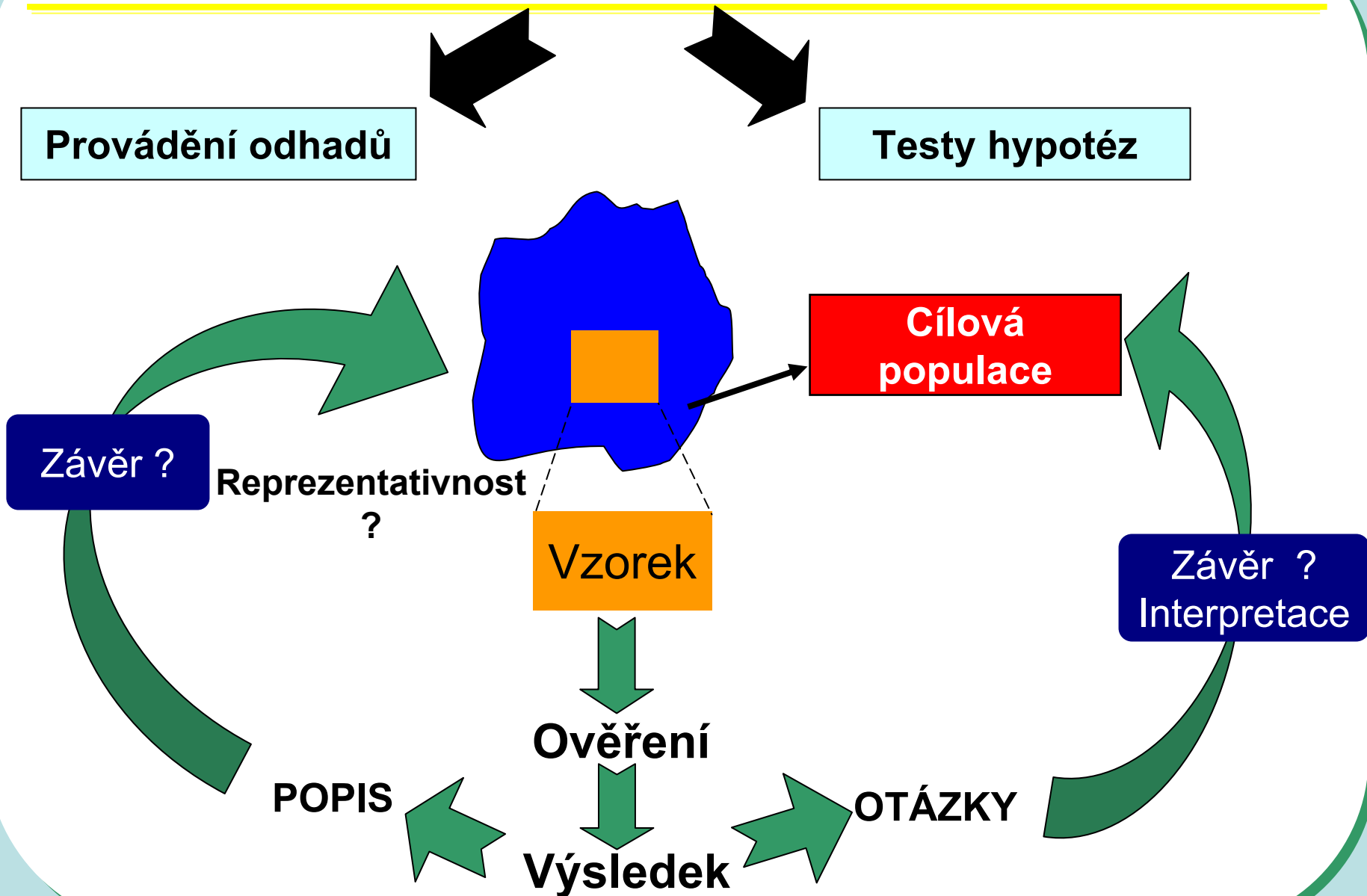
Testy o rozložení: Kolmogorov-Smirnov test, Shapiro-Wilks test, c2 test



## 8. Provádění odhadů

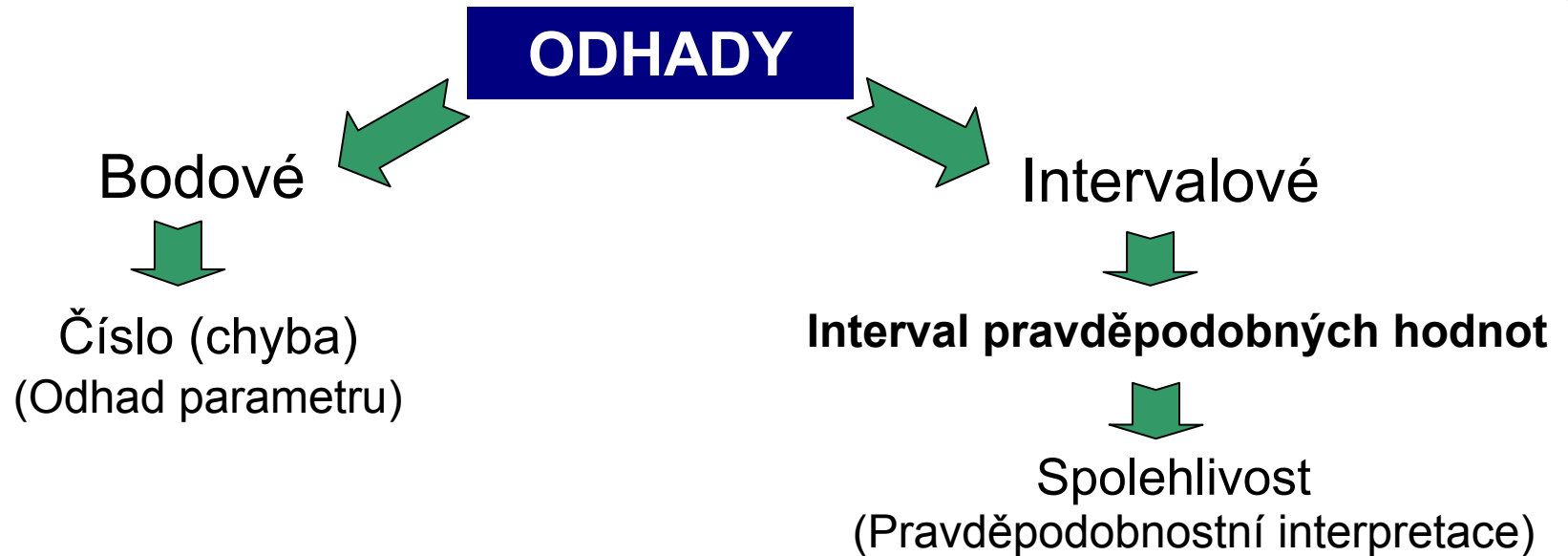


# Statistika v průzkumném studiu



# INTERVAL SPOLEHLIVOSTI

- velmi užitečná míra věrohodnosti odhadů -



➔ **Obecný tvar:**

$$P(L_1 < \text{Odhad} < L_2) \geq 1 - \alpha/2$$

**Odhadovaný  
parametr**

±

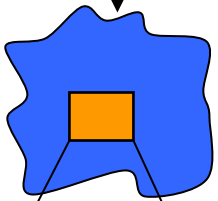
Kvantil  
modelového . SE (odhadu)  
rozložení

$K_V$  pro  $(1 - \alpha/2)$

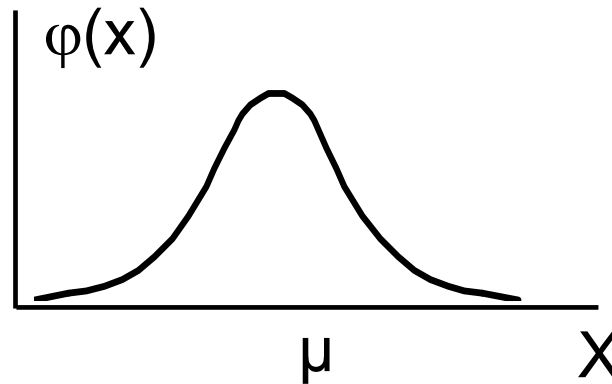
# NORMÁLNÍ ROZLOŽENÍ

## - model pro odhad průměru -

Cílová populace



Vzorek:  $n$



Prezentace

$n$ ;  $\bar{x}$ ;  $s$

$n$ ;  $\bar{x}$ ;  $\frac{s}{\sqrt{n}}$

$n$ ;  $\bar{x}$ ;  $c$

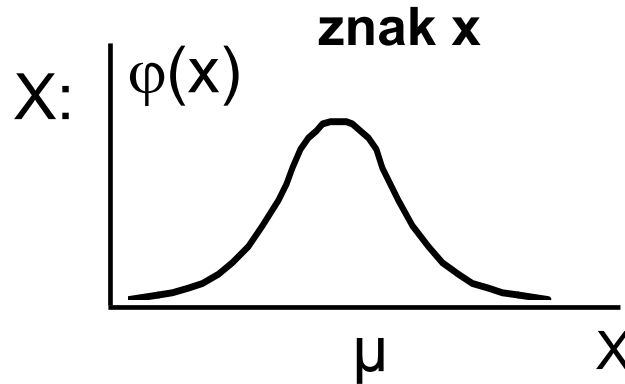
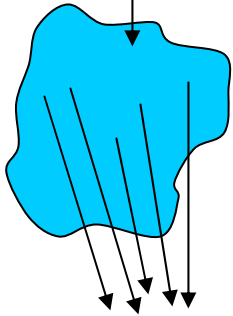
$n$ ;  $\bar{x}$ ; Interval  
spolehlivost  
 $i$  pro odhad  
průměru

$\bar{X}$  ..... odhad průměru

# NORMÁLNÍ ROZLOŽENÍ

- odhad průměru je rovněž normálně rozložen -

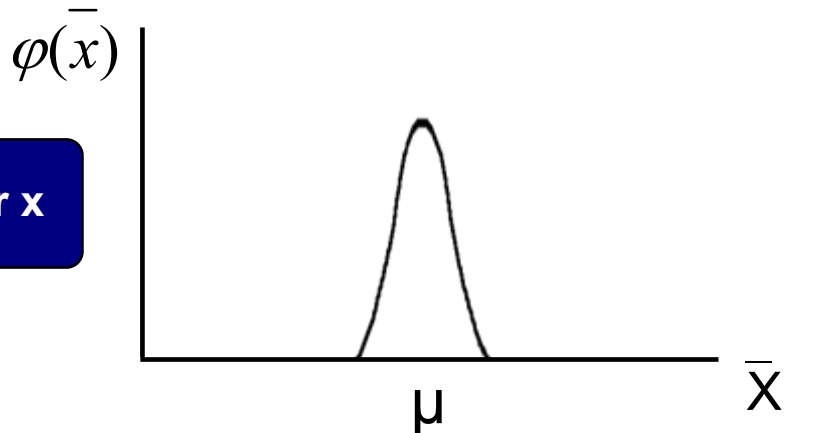
Cílová populace



$$x: \mu \pm 3s$$

Náhodné výběry o  $n = 100$

$\bar{x}_1$   $\bar{x}_2$   $\bar{x}_3$   $\bar{x}_4$  ....  $\bar{x}_i$



průměr  $x$

$$\mu \pm 3 \cdot \frac{s}{\sqrt{n}}$$

$$\frac{s}{\sqrt{n}}$$

~ Standardní chyba odhadu průměru

# ODHAD PRŮMĚRU

## I. Vztahy



### Bodový

$$\bar{x}; \left( \frac{s}{\sqrt{n}} \right)$$



### Intervalový

$$\bar{x} - t_{1-\alpha/2}^{(v=n-1)} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}^{(v=n-1)} \cdot \frac{s}{\sqrt{n}}$$

$$\mu : \bar{x} \pm t_{1-\alpha/2}^{(v=n-1)} \cdot \frac{s}{\sqrt{n}}$$

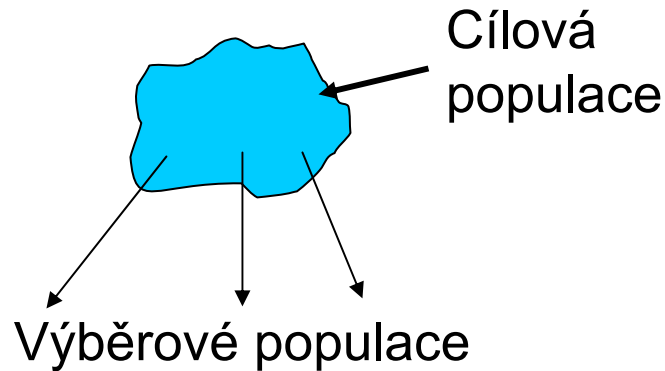
$$\mu : \bar{x} \pm t_{1-\alpha/2}^{(v=n-1)} \cdot s_{\bar{x}}$$

**t** ... příslušný kvantil Studentova rozložení  
**1 - α** ... spolehlivost hodnoceného intervalu



# Interval spolehlivosti odhadu průměru je pouze informací o přesnosti tohoto odhadu

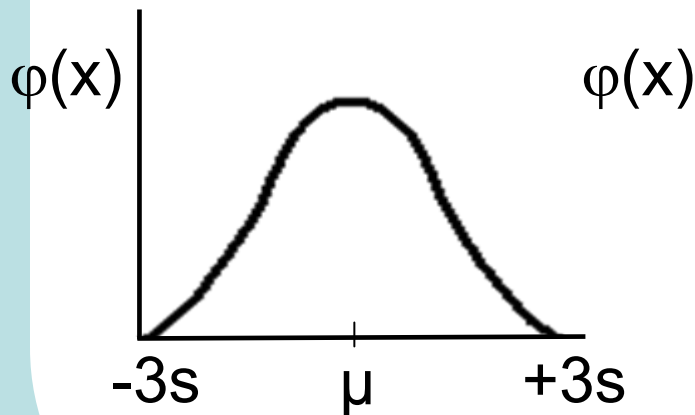
*Interval spolehlivosti je hodnocen pro  $(1 - \alpha)$  procentní spolehlivost*



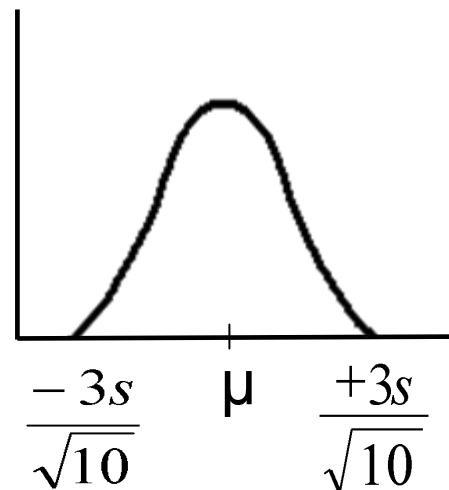
Šířku intervalu určuje:

- a) velikost vzorku
- b) rozptyl (variabilita) vzorku
- c) požadovaná spolehlivost

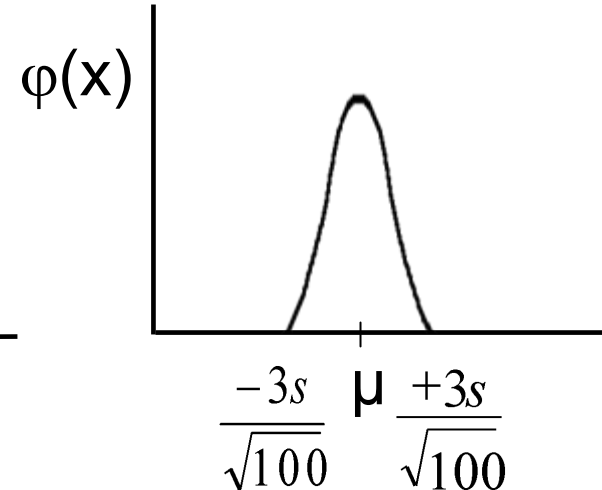
**Původní proměnná  $x$**



**Výběr  $n=10$  pro odhad průměru**



**Výběr  $n=100$  pro odhad průměru**



# ODHAD PRŮMĚRU

## II. Příklad

**X:** Cena výrobku v  $n = 21$  obchodech

**Data:**

$$n = 21; \bar{x} = 3,58; s^2 = 0,12$$

$$s_{\bar{x}} = \sqrt{0,12/21} = 0,075$$

95% Interval spolehlivosti:

$$t_{1-\alpha/2}^{(u = n-1)} = t_{0,975}^{(20)} = 2,086$$

$$\mu : \bar{x} \pm 2,086 \cdot s_{\bar{x}}$$

$$3,58 - 2,086 \cdot 0,075 \leq \mu \leq 3,58 + 2,086 \cdot 0,075$$

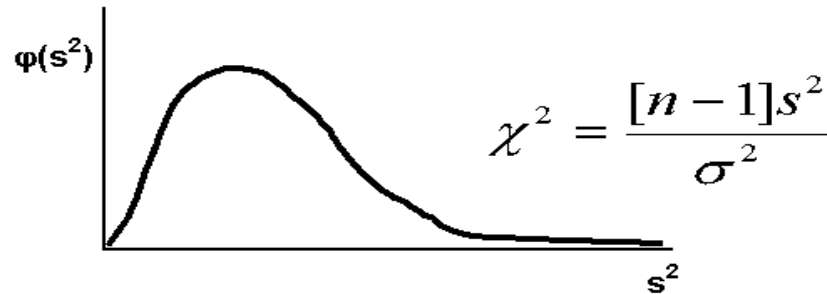
$$3,423 \leq \mu \leq 3,737$$



$$P(3,423 \leq \mu \leq 3,737) \geq 0,95$$

# Interval spolehlivosti pro odhad rozptylu

$$s^2 \sim \sigma^2 \text{ pro velká } n$$



## Interval spolehlivosti

$$\text{a) pro } \sigma^2 : \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}(n-1)}$$

$$\text{b) pro } \sigma : \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}(n-1)}}$$

$$\text{c) pro } \sigma/\sqrt{n} : \sqrt{\frac{(n-1)s^2}{n\chi^2_{\alpha/2}(n-1)}} \leq \frac{\sigma}{\sqrt{n}} \leq \sqrt{\frac{(n-1)s^2}{n\chi^2_{(1-\alpha/2)}(n-1)}}$$

$\sigma/\sqrt{n}$   
-směrodatná odchylka  
odhadu průměru (S.E.)

# Interval spolehlivosti pro odhad rozptylu - příklad

*Příklad: měření produkce metabolitu (x) u buněk dvou nádorových linií*

Linie 1

$n = 50$

$s^2(x) = 10 \text{ (mg/ml)}^2$

$s(x) = 3,16 \text{ mg/ml}$

$\bar{x} = 2 \text{ mg/ml}$

$\bar{s}_x = 0,447 \text{ mg/ml}$

**95% IS**

$$\frac{49 * 10}{77,22} \leq \sigma^2 \leq \frac{49 * 10}{31,56}$$

$$6,98 \leq \sigma^2 \leq 15,53$$

**c = 1,58**

Linie 1

$n = 100$

$s^2(x) = 16 \text{ (mg/ml)}^2$

$s(x) = 4 \text{ mg/ml}$

$\bar{x} = 2,8 \text{ mg/ml}$

$\bar{s}_x = 0,4 \text{ mg/ml}$

**95% IS**

$$\frac{99 * 16}{128,42} \leq \sigma^2 \leq \frac{99 * 16}{73,36}$$

$$12,33 \leq \sigma^2 \leq 13,49$$

**c = 1,43**

# Výpočet mediánu z frekvenčních dat a jeho odhady

- a) Určete medián tohoto souboru dat: 1,3,4,5,7,8 [4,5]
- b) Určete medián tohoto souboru dat: 5,1,8,3,4 [4]
- c) Tento příklad je ukázkou výpočtu mediánu u velkého souboru dat. V následující tabulce je uveden rozbor rozložení souboru dat od 179 krav, kde sledovanou veličinou byl počet dní od narození telete do znovuobnovení menstruačního cyklu. Uvedená data jsou velmi zjednodušena a jsou zde uvedena pouze pro ilustraci:

<b>Class limits</b> (days)	<b>0,5- 20,5</b>	<b>20,5- 40,5</b>	<b>40,5- 60,5</b>	<b>60,5- 80,5</b>	<b>80,5- 100,5</b>	<b>100,5- 120,5</b>	<b>120,5- 140,5</b>	<b>140,5- 160,5</b>	<b>160,5- 180,5</b>	<b>180,5- 200,5</b>	<b>200,5- 220,5</b>
<b>Frequency</b>	8	33	50	32	15	20	11	6	2	1	1
<b>Cumulative frequency</b>	8	41	91	123	138	158	169	175	177	178	179

**Frekvence zastoupení dosahuje nejvyšší hodnoty u třídy od 40,5 – 60,5 dnů. Druhý (menší) frekvenční pík lze pozorovat u intervalu od 100,5 do 120,5 dní. Existence dvou maxim (bimodální data) je důkazem nenormality tohoto konkrétního souboru.**

# Výpočet mediánu z frekvenčních dat a jeho odhady

Jelikož  $n = 179$ , pak je medián devadesátá hodnota od počátku souboru, a dále je zřejmé, že bude velmi blízko horní hranici třídy 40,5 – 60,5 dní. Za předpokladu, že 50 hodnot této třídy je v ní rovnoměrně rozmístěno lze použít následující vzorec:

$$M = X_L + \frac{gl}{f}, \text{ kde}$$

$X_L$  = hodnota  $X$  (sledované veličiny) na spodní hranici třídy obsahující medián: zde 40,5 dní

$g$  = pořadová hodnota mediánu minus kumulativní frekvence do horní hranice předchozí třídy, tj.  $90 - 41 = 49$

$l$  = třídní interval: 20 dní

$f$  = frekvence ve třídě obsahující medián

- Dosadíme-li do uvedeného vzorce, získáme odhad mediánu jako 60 dní. Průměr tohoto datového souboru je 69,9, což je významně odlišná hodnota, a potvrzuje znovu nenormální charakter dat.
- U velkých vzorků z normálních populací je výběrový odhad mediánu normálně rozložen kolem populační hodnoty se směrodatnou odchylkou  $1,253 \sigma / \sqrt{n}$ . U normálního rozložení, kde medián i průměr představují odhad stejné hodnoty, je medián méně přesný než průměr. Proto hlavní význam mediánu spočívá u nesymetrických distribucí.
- Existuje velmi jednoduchá metoda pro výpočet intervalu spolehlivosti pro odhad mediánu a jako horní a spodní hranice slouží pořadová čísla vypočítaná podle následujícího vztahu:

$$\frac{(n + 1)}{2} \pm \frac{z \sqrt{n}}{2}, \text{ kde}$$

$n$  představuje velikost datového souboru,  $z$  je kvantil standardizovaného normálního rozložení pro příslušnou pravděpodobnost. U našeho příkladu je  $n = 179$  a pro 95% interval spolehlivosti je  $z$  přibližně rovno 2. Horní a spodní limit pro odhad mediánu tedy je  $90 \pm \sqrt{179} = 77$  a 103. 95% interval spolehlivosti je tedy tvořen počty dní, které mají pořadí 77 a 103:

**77: Počet dní =  $40,5 + (36)(20)/50 = 55$  dní**

**103: Počet dní =  $60,5 + (12)(20)/32 = 68$  dní**

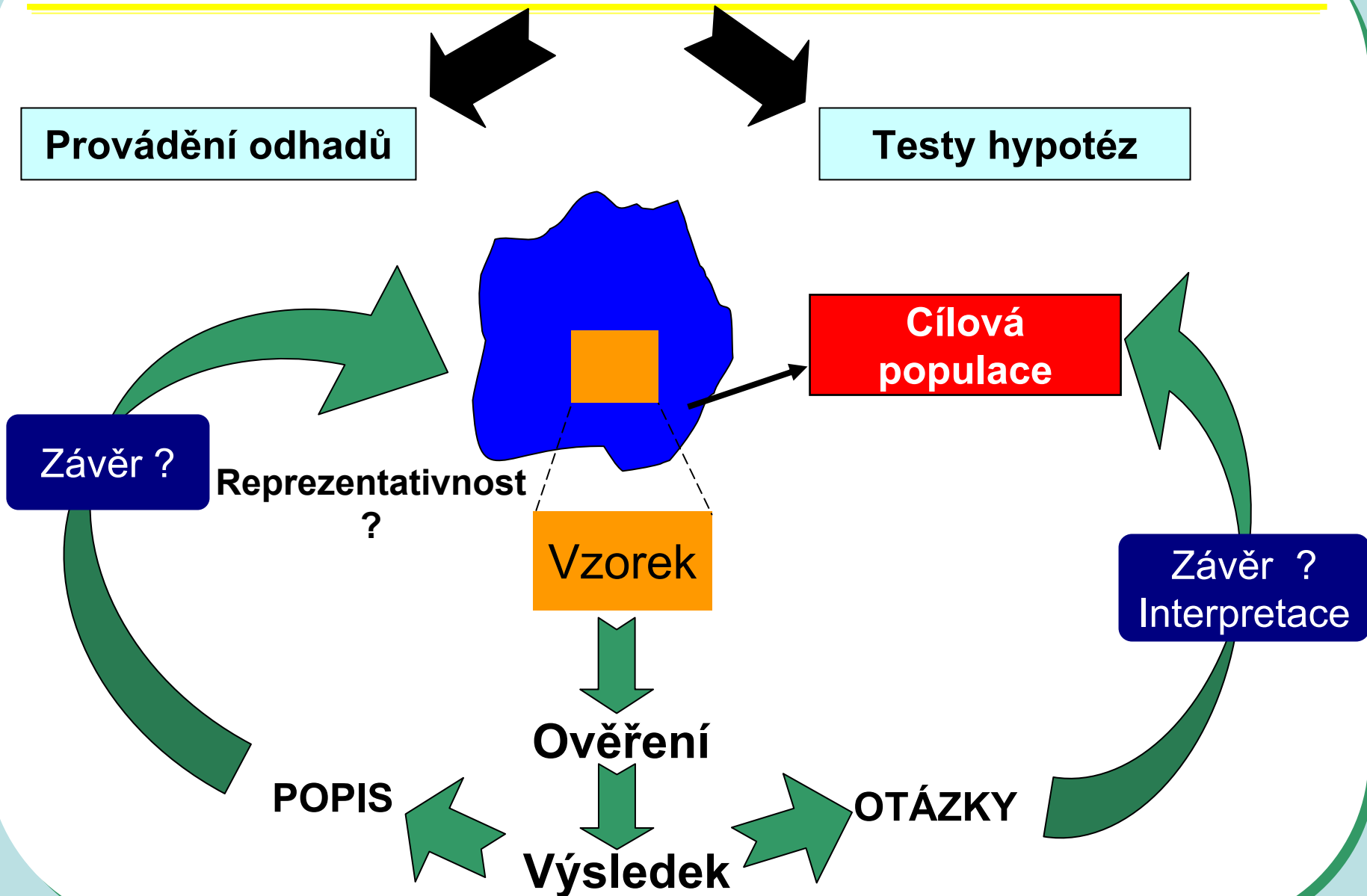
**Medián cílové populace byl tedy odhadnut 95% intervalem spolehlivosti jako hodnota ležící mezi 55 a 68 dny. Interpretujte tento výsledek.**



## 9. Testy hypotéz: úvod



# Statistika v průzkumném studiu





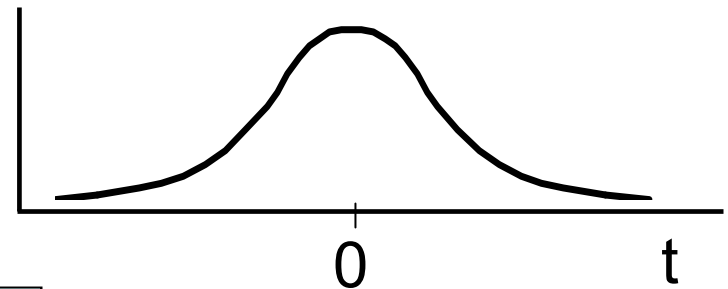
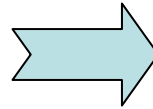
# Elementární prvky statistických testů

➤ Nulová hypotéza  $H_0$

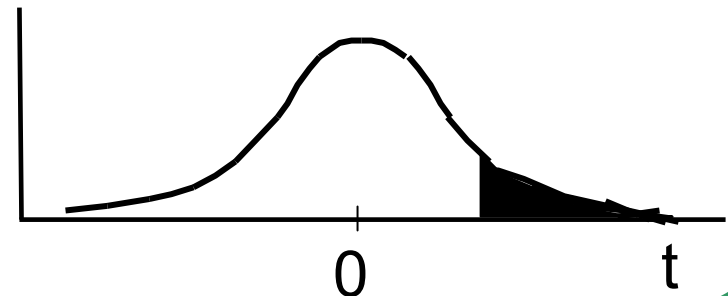
➤ Alternativní hypotéza  $H_A$

➤ Testová statistika

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$



➤ Kritický obor testované statistiky



# Experimentální design

Optimální velikost  
vzorku nebo  
počet opakování

Efektivní  
uspořádání  
experimentů

*Účelná minimalizace chyb*

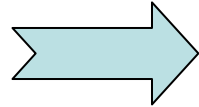
		Závěr testu	
		Platí	Neplatí
Skutečnost	Platí	$1-\alpha$	$\alpha$
	Neplatí	$\beta$	$1-\beta$

# Experimentální design



## Pravděpodobnost chyby 1. druhu

$\alpha$



P nesprávného zamítnutí nulové hypotézy



## Pravděpodobnost chyby 2. druhu

$\beta$



P nerozpoznání neplatné nulové hypotézy



## Síla testu

$1-\beta$



Pravděpodobnostně vyjádřená schopnost rozpoznat neplatnost hypotézy



## 10. Předpoklady a pojmy statistických testů



# Parametrické vs. neparametrické testy

## Parametrické testy

- Mají předpoklady o rozložení vstupujících dat (např. normální rozložení)
- Při stejném N a dodržení předpokladů mají vyšší sílu testu než testy neparametrické
- Pokud nejsou dodrženy předpoklady parametrických testů, potom jejich síla testu prudce klesá a výsledek testu může být zcela chybný a nesmyslný

## Neparametrické testy

- Nemají předpoklady o rozložení vstupujících dat, lze je tedy použít i při asymetrickém rozložení, odlehlých hodnotách, či nedetekovatelném rozložení
- Snížená síla těchto testů je způsobena redukcí informační hodnoty původních dat, kdy neparametrické testy nevyužívají původní hodnoty, ale nejčastěji pouze jejich pořadí

# One-sample vs. two sample testy

## One – sample testy

- Srovnávají jeden vzorek (one sample, jednovýběrové testy) s referenční hodnotou (popřípadě se statistickým parametrem cílové populace)
- V testu je tedy srovnáváno rozložení hodnot (vzorek) s jediným číslem (referenční hodnota, hodnota cílové populace)
- Otázka položená v testu může být vztažena k průměru, rozptylu, podílu hodnot i dalším statistickým parametrům popisujícím vzorek

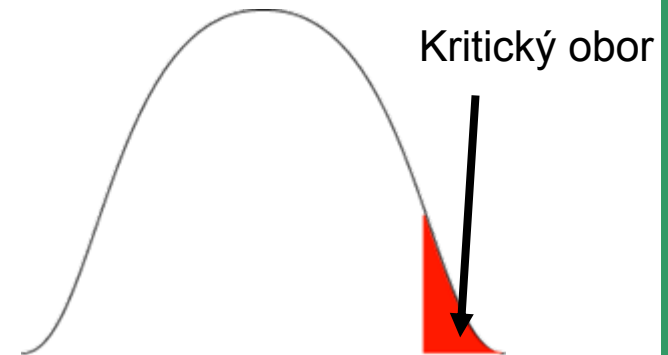
## Two – sample testy

- Srovnávají navzájem dva vzorky (two sample, dvouvýběrové vzorky)
- V testu jsou srovnávány dvě rozložení hodnot
- Otázka položená v testu může být opět vztažena k průměru, rozptylu, podílu hodnot i dalším statistickým parametrům popisujícím vzorek
- Kromě testů pro dvě skupiny hodnot existují samozřejmě i testy pro více skupin dat

# One-tailed vs. Two-tailed testy

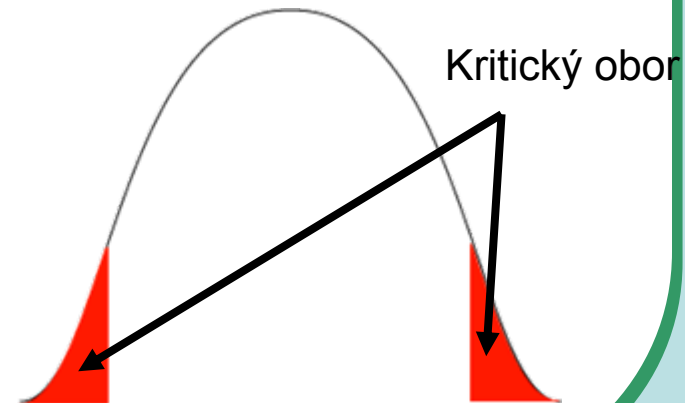
## One – tailed testy

- Hypotéza testu je postavena asymetricky, tedy ptáme se na **větší než/ menší než**
- Test může mít pouze dvojí výstup – jedna z hodnot je větší (menší) než druhá a všechny ostatní případy



## Two – tailed testy

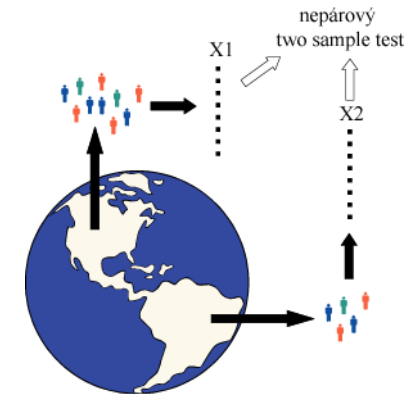
- Hypotéza testu se ptá na otázku **rovná se/nerovná se**
- Test může mít trojí výstup – **menší - rovná se – větší než**
- Situace **nerovná se** je tedy souhrnem dvou možných výstupů testu (**menší+větší**)



# Nepárový vs. Párový design

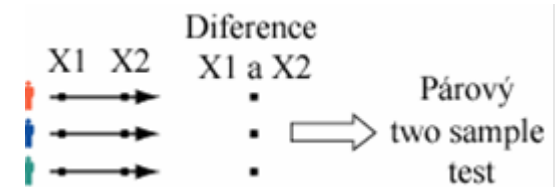
## Nepárový design

- Skupiny srovnávaných dat jsou na sobě zcela nezávislé (též nezávislý, independent design), např. lidé z různých zemí, nezávislé skupiny pacientů s odlišnou léčbou atd.
- Při výpočtu je nezbytné brát v úvahu charakteristiky obou skupin dat



## Párový design

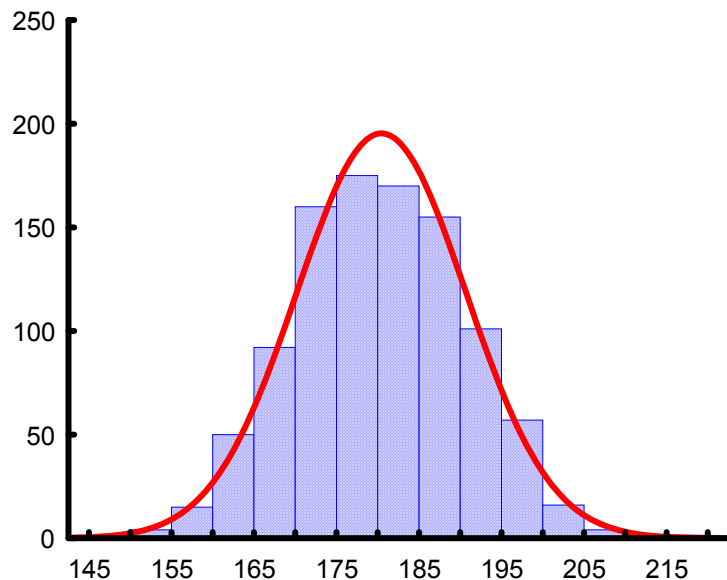
- Mezi objekty v srovnávaných skupinách existuje vazba, daná např. člověkem před a po operaci, reakce stejného kmene krys atd.
- Vazba může být buď přímo dána nebo pouze předpokládána (v tom případě je nutné ji ověřit)
- Test je v podstatě prováděn na diferencích skupin, nikoliv na jejich původních datech





# Testy normality

- Testy normality pracují s nulovou hypotézou, že není rozdíl mezi zpracovávaným rozložením a normálním rozložením. Vždy je ovšem dobré prohlédnout si i histogram, protože některé odchylky od normality, např. bimodalitu některé testy neodhalí.



## • Test dobré shody

V testu dobré shody jsou data rozdělena do kategorií (obdobně jako při tvorbě histogramu), tyto intervaly jsou normalizovány (převedeny na normální rozložení) a podle obecných vzorců normálního rozložení jsou k nim dopočítány očekávané hodnoty v intervalech, pokud by rozložení bylo normální. Pozorované normalizované četnosti jsou poté srovnány s očekávanými četnostmi pomocí  $\chi^2$  testu dobré shody. Test dává dobré výsledky, ale je náročný na  $n$ , tedy množství dat, aby bylo možné vytvořit dostatečný počet tříd hodnot.

## • Kolmogorov Smirnov test

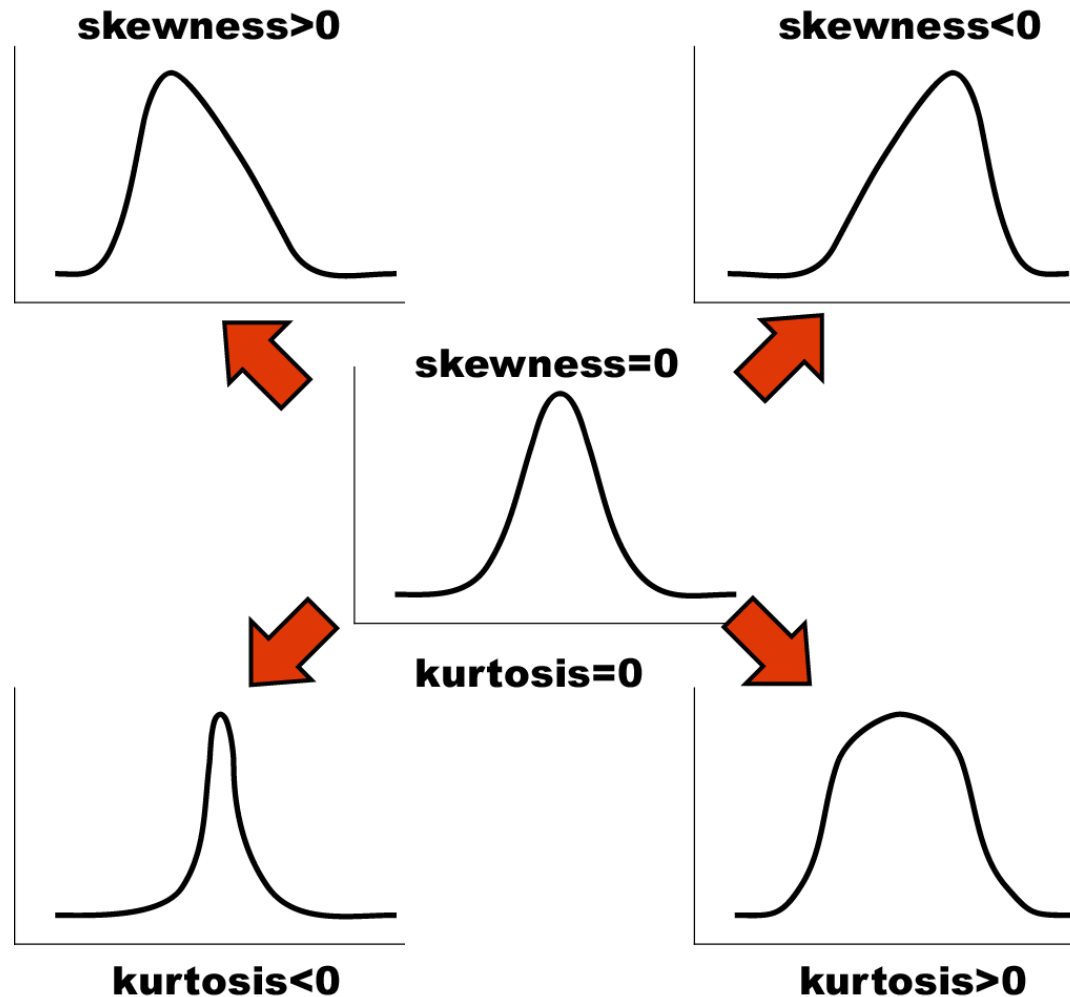
Tento test je často používán, dokáže dobře najít odlehlé hodnoty, ale počítá spíše se symetrií hodnot než přímo s normalitou. Jde o neparametrický test pro srovnání rozdílu dvou rozložení. Je založen na zjištění rozdílu mezi reálným kumulativním rozložením (vzorek) a teoretickým kumulativním rozložením. Měl by být počítán pouze v případě, že známe průměr a směrodatnou odchylku hypotetického rozložení, pokud tyto hodnoty neznáme, měla by být použita jeho modifikace – Lilieforsův test.

## • Shapiro-Wilk`'s test

Jde o neparametrický test použitelný i při velmi malých  $n$  (10) s dobrou silou testu, zvláště ve srovnání s alternativními typy testů, je zaměřen na testování symetrie.

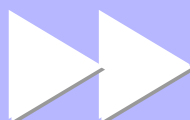
# Šikmost a špičatost jako testy normality

- Parametry normálního rozložení, skewness a kurtosis mohou být využity pro testování normality, ale pouze pro velké vzorky (šikmost – 100, špičatost – 500).



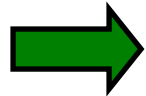


# 11. Statistické testy o parametrech jednoho výběru



# "One sample" testy

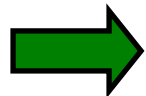
V případě one sample testů jde o srovnání výběru dat (tedy one sample) s cílovou populací. Pro parametrické testy musí mít datový soubor normální rozložení.



Průměr – cílová vs. výběrová populace

$H_0$	$H_A$	Testová statistika	Interval spolehlivosti
$\bar{x} \leq \mu$	$\bar{x} > \mu$	$t$	$t > t_{1-\alpha}^{(n-1)}$
$\bar{x} \geq \mu$	$\bar{x} < \mu$	$t$	$t < t_{\alpha}^{(n-1)}$
$\bar{x} = \mu$	$\bar{x} \neq \mu$	$t$	$ t  > t_{1-\alpha/2}^{(n-1)}$

$$t = \frac{\bar{X} - \mu}{s} \sqrt{n}$$

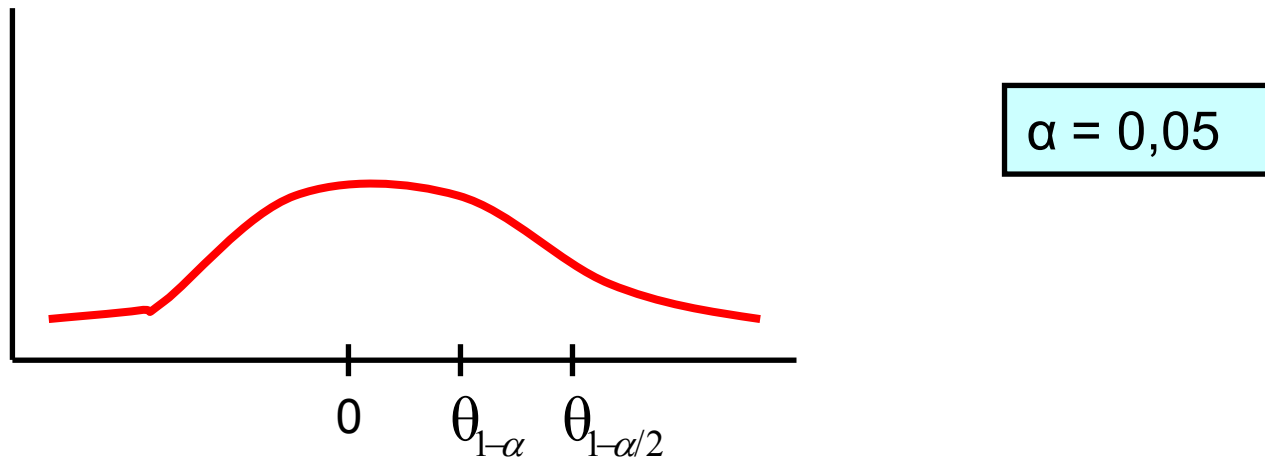


Rozptyl – cílová vs. výběrová populace

$H_0$	$H_A$	Testová statistika	Interval spolehlivosti
$s^2 \leq \sigma^2$	$s^2 > \sigma^2$	$\chi^2$	$\chi^2 > \chi_{1-\alpha}^2 (n-1)$
$s^2 \geq \sigma^2$	$s^2 < \sigma^2$	$\chi^2$	$\chi^2 < \chi_{\alpha}^2 (n-1)$
$s^2 = \sigma^2$	$s^2 \neq \sigma^2$	$\chi^2$	$\chi^2 > \chi_{1-\alpha/2}^2$ nebo $\chi^2 < \chi_{\alpha/2}^2$

$$\chi^2 = \frac{(n-1) \cdot s^2}{\sigma^2}$$

# "One sample" testing vs one/two tailed



➔ Pokud "two – tailed" test vyjde významný tak, že  $P < 0,05$ , pak dobře zvolený "one – tailed" test je významný při  $P < 0,025$ .

➔ ... tzn. že testová charakteristika  $> \theta_{1-\alpha/2}$  a "one – tailed" testy na hladině jsou v podstatě zbytečné.

➔ Pokud je pro "two – tailed" test  $P = 0,1$ , pak lze na hladině  $\alpha = 0,05$  prokázat nerovnost srovnávaných parametrů vhodně voleným "one – tailed" testem.

➔ ... tzn. že testová charakteristika  $\in \langle \theta_{1-\alpha}; \theta_{1-\alpha/2} \rangle$

# Srovnání odhadu průměru s předpokládanou hodnotou I

## Koncentrace antibiotika v cílovém orgánu

Při 1000 měřeních antibiotika byla zjištěna v cílovém orgánu průměrná koncentrace 202,5 jednotek a směrodatná odchylka 44 jednotek.

Požadovaná koncentrace antibiotika je 200 jednotek.

- 1) Je daný rozdíl 2,5 významný vzhledem k variabilitě znaku na hladině významnosti 5%?
- 2) Jaká je skutečná hladina významnosti?

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{2,5}{44} \sqrt{1000} = 1,797$$

# Srovnání odhadu průměru s předpokládanou hodnotou II

## Aktivita enzymu v buňkách

Při zjišťování aktivity enzymu v buňkách na vzorku 25 měření byl zjištěn průměr 3,5 jednotek a směrodatná odchylka 1.

1. otázka zní, zda se naměřené hodnoty našeho vzorku liší od výsledků dřívější rozsáhlé studie zaměřené na celou cílovou populaci, kde byla zjištěna průměrná aktivita 2,5 jednotky?

H0:  $x = \mu$  tedy two tailed test

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{3,5 - 2,5}{1} \sqrt{25} = 5$$

$t_{0,975}^{24} = 2,064 \Rightarrow t > t_{1-\alpha/2}^{24} \Rightarrow H_0$  zamítnuta při  $\alpha \leq 0,05$   
od jiné hodnoty bychom zachytili při daných hodnotách?

2. otázka – jakou minimální odchylku X od jiné hodnoty bychom zachytili při daných hodnotách?

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{d}{s} \sqrt{n} \Rightarrow d = \frac{t_{1-\alpha/2}^v}{\sqrt{n}} s \Rightarrow d = \frac{2,064}{5} s$$

3. za předpokladu, že z praktického hlediska je významná odchylka již 0,2 jednotky, jaký minimální počet měření musíme provést, abychom ji byli schopni prokázat ?

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{d}{s} \sqrt{n} \Rightarrow n = \left( \frac{t_{1-\alpha/2}^v}{d} s \right)^2$$

# Srovnání odhadu průměru s předpokládanou hodnotou III

**x**: Aktivita enzymu v buňkách

$n = 25$ ;  $\bar{x} = 3,5$ ;  $s = 1$

$\mu$ : Hodnota zjištěná při předcházejícím, dlouhodobém průzkumu

$$H_0 : \bar{x} = \mu$$

$$t = \frac{3,5 - 2,5}{1} \sqrt{25} = 5$$

$$\text{Kvantil } t_{0,975}^{(24)} = 2,064$$

$$t > t_{1-\alpha/2}^{(24)}$$

**$H_0$  zamítnuta při  $\alpha \leq 0,05$**



# Srovnání odhadu průměru s předpokládanou hodnotou IV

**Situace:** Odhad průměrné hodnoty znaku X

**?** Jakou minimální odchylku X od nějaké jiné hodnoty zachytíme jako významnou při daném  $n$ ,  $\alpha$ ,  $\beta$  ?

$$d = \sqrt{\frac{s^2}{n}} (t_{1-\alpha/2}^v + t_{1-\beta}^v)$$

**Necht'  $\alpha = 0,05$ ;  $\beta = 0,10$ ;  $n = 25$  ;  $s^2 = 1,5682$**

$$t_{1-\alpha/2}^{(24)} = 2,064$$

$$t_{1-\beta}^{(24)} = 1,318$$

$$d = \sqrt{\frac{1,5682}{25}} (2,064 + 1,318) = \underline{\underline{0,85}}$$

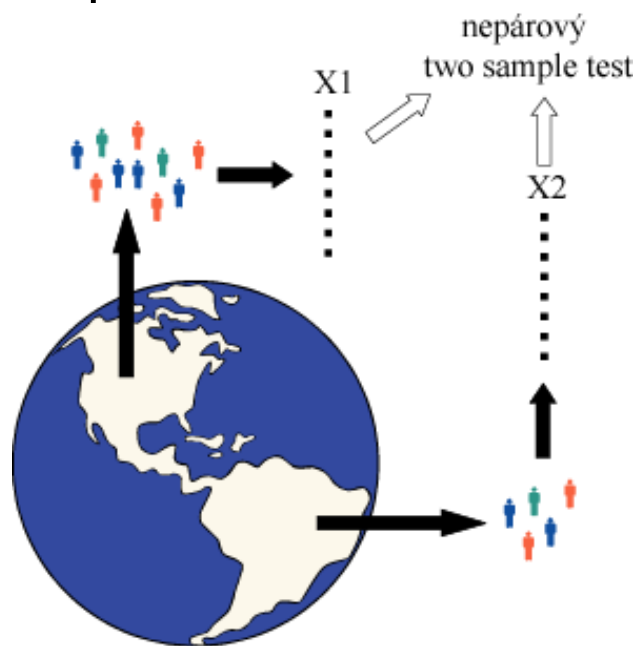


## 12. Statistické testy o parametrech dvou výběrů

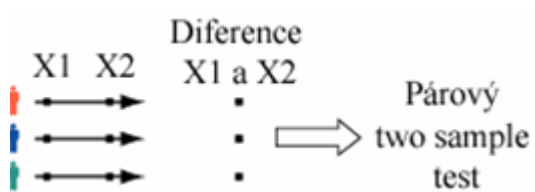


# Two sample testy

- Při použití two sample testů srovnáváme spolu dvě rozložení. Jejich základním dělením je podle designu experimentu na testy párové a nepárové.



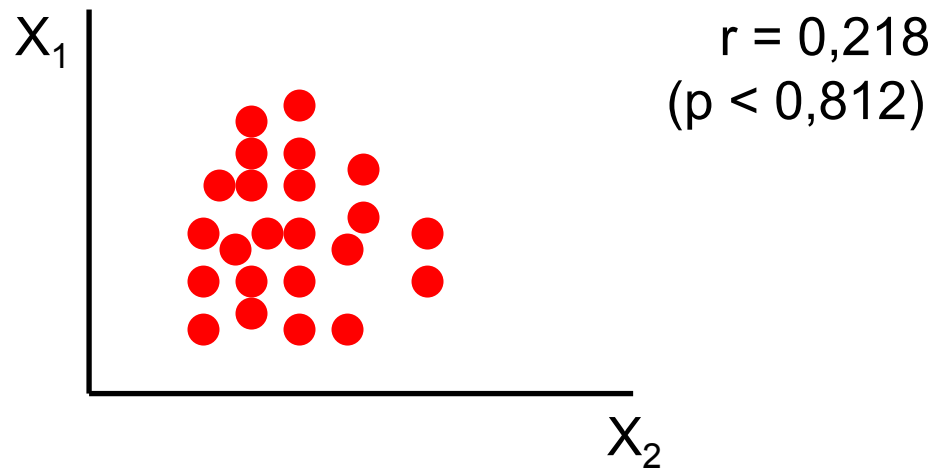
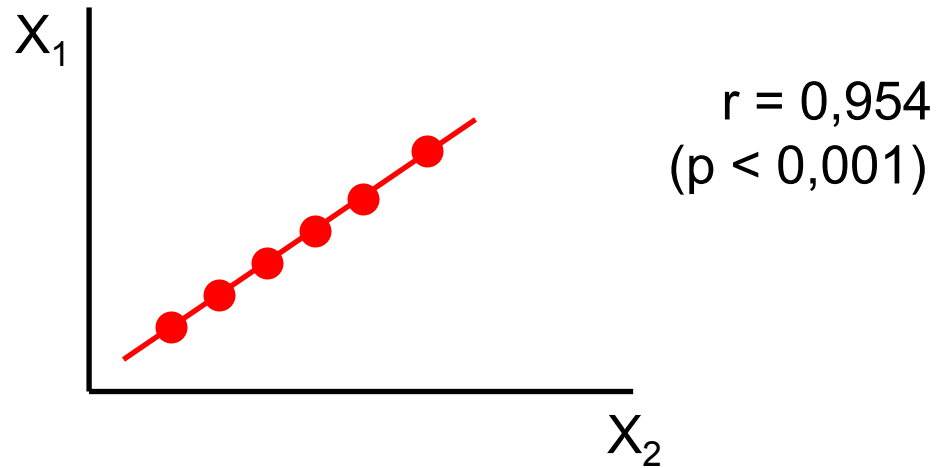
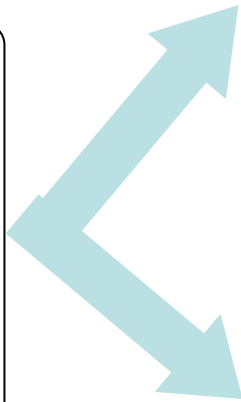
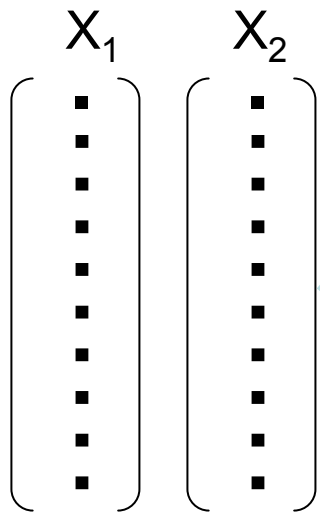
- Základním testem pro srovnání dvou nezávislých rozložení spojitých čísel je **nepárový two-sample t-test**
- Základním testem pro srovnání dvou závislých rozložení spojitých čísel je **párový two-sample t-test**





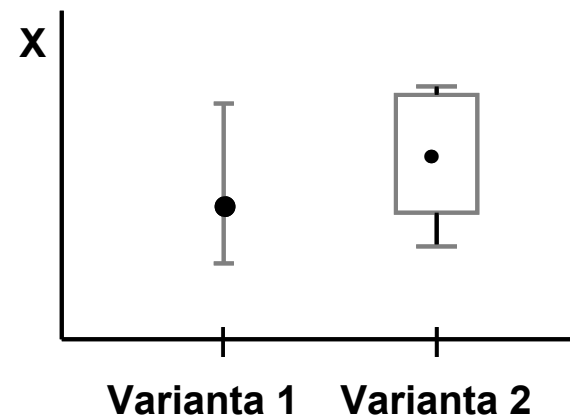
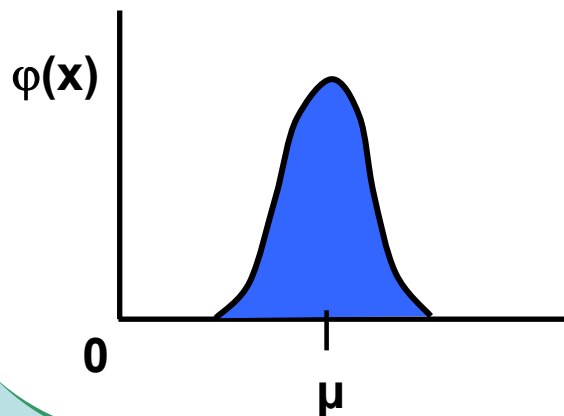
# Srovnání dvou pokusných variant – obecné schéma zapojených testů II.

Identifikace párovitosti (Korelace, Kovariance)



# Předpoklady nepárového two sample t-testu

- Náhodný výběr subjektů jednotlivých skupin z jejich cílových populací
- Nezávislost obou srovnávaných vzorků
- Přibližně normální rozložení proměnné ve vzorcích, drobné odchylky od normality ovšem nejsou kritické, test je robustní proti drobným odchylkám od tohoto předpokladu, normalita může být testována testy normality
- Rozptyl v obou vzorcích by měl být přibližně shodný (homoscedastic). Tento předpoklad je testován několika možnými testy – Levenův test nebo F-test.
- Vždy je vhodné prohlédnout histogramy proměnné v jednotlivých vzorcích pro okometrické srovnání a ověření předpokladů normality a homogenity rozptylu – nenahradí statistické testy, ale poskytne prvotní představu.



# Nepárový two sample t-test – výpočet I

1. nulová hypotéza: průměry obou skupin jsou shodné, alternativní hypotéza je, že nejsou shodné, two tailed test
2. prohlédnout průběh dat, průměr, medián apod. pro zjištění odchylek od normality a nehomogenita rozptylu, provést F –test

$H_0$	$H_A$	Testová statistika
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F = \frac{s_1^2}{s_2^2}$
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F = \frac{s_2^2}{s_1^2}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{\max(s_1^2; s_2^2)}{\min(s_1^2; s_2^2)}$

## F-test pro srovnání dvou výběrových rozptylů

- Používá se pro srovnání rozptylu dvou skupin hodnot, často za účelem ověření homogenity rozptylu těchto skupin dat.

- V případě ověření homogenity je testována hypotéza shody rozptylů (two tailed); v případě shodných rozptylů je vše v pořádku a je možné pokračovat ve výpočtu t-testu, v opačném případě není vhodné test počítat.

## Nepárový two sample t-test – výpočet II

3. Výpočet testové statistiky (stupně volnosti jsou  $\nu = n_1 + n_2 - 2$ ):

$$t = \frac{\text{Rozdíl průměů}}{SE(\text{rozdílprůměů})} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \text{vážený odhad rozptylu}$$

4. výsledné t srovnáme s tabulární hodnotou t pro dané stupně volnosti a  $\alpha$  (obvykle  $\alpha=0,05$ )
5. Lze spočítat interval spolehlivosti pro rozdíl průměů (např. 95%), počet stupňů volnosti a  $s^2$  odpovídají předchozím vzorcům

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975} SE(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{0,975} \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$



# Test homogeneity rozptylů: Two sample F test

$H_0$	$H_A$	F
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F_1 = \frac{s_1^2}{s_2^2}$
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F_2 = \frac{s_2^2}{s_1^2}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F_3 = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$
$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$		
$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2}(v_1; v_2)} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2}(v_2; v_1)$		

## Two sample testing – nezávislý t-test

$$H_0 = \bar{X}_1 = \bar{X}_2$$

$$1) \sigma_1^2 = \sigma_2^2 \Rightarrow s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} = \frac{SS_1 + SS_2}{\nu_1 + \nu_2}$$

$$2) s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p \cdot \sqrt{\frac{n_2 + n_1}{n_1 \cdot n_2}}$$

$$3) t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

$$\mu_1 - \mu_2 : \bar{X}_1 - \bar{X}_2 \pm t_{1-\alpha/2}^{(n_1+n_2-2)} \cdot s_{\bar{X}_1 - \bar{X}_2}$$

---

Pokud  $\sigma_1^2 \neq \sigma_2^2$  : nelze vyjádřit  $s_p^2$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Two sample t-test - příklad

**Průměrná hmotnost ovcí v čase páření byla srovnávána pro kontrolní skupinu a skupinu krmenou zvýšenou dávkou potravy. Kontrolní skupina obsahuje 30 ovcí, skupina se zvýšeným příjmem potravy pak 24 ovcí.**

- Vlastní experiment byl prováděn tak, že na začátku máme 54 ovcí (ideálně stejného plemene, stejně staré atd.), které náhodně rozdělíme do dvou skupin (náhodné rozdělování objektů do pokusných skupin je objektem celého specializovaného odvětví statistiky nazývaného randomizace). Poté co experiment proběhne, musíme nejprve ověřit teoretický předpoklad pro využití nepárového t-testu. Pro obě proměnné jsou vykresleny grafy (můžeme též spočítat základní popisnou statistiku), na kterých můžeme posoudit normalitu a homogenitu rozptylu, kromě okometrického pohledu můžeme pro ověření normality použít testy normality, pro ověření homogenity rozptylu pak F-test
- Pokud platí všechny předpoklady Two sample nepárového t-testu, můžeme spočítat testovou charakteristiku, výsledné  $t$  je 2,43 s 52 stupni volnosti, podle tabulek je  $t_{0,975(52)} = 2,01$ , tedy  $t > t_{0,975(52)}$  a nulovou hypotézu můžeme zamítnout, skutečná pravděpodobnost je pak 0,018. Rozdíl mezi skupinami je 1,59 kg ve prospěch skupiny s lepší výživou.

$$t = \frac{\text{Rozdíl} - \text{průměrně}}{SE(\text{rozdílprůěrů})} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \nu = n_1 + n_2 - 2$$

- Pro rozdíl mezi oběma soubory jsou spočítány 95% konfidenční intervaly jako  $1,59 \pm 2,01 * (0,655)$  kg, což odpovídá rozsahu 0,28 až 2,91 kg. To, že konfidenční interval nezahrnuje 0 je dalším potvrzením, že mezi skupinami je významný rozdíl – jde o další způsob testování významnosti rozdílů mezi skupinami dat – nulovou hypotézu o tom, že rozdíl průměrů dvou skupin dat je roven nějaké hodnotě zamítáme v případě, kdy 95% konfidenční interval rozdílu nezahrnuje tuto hodnotu (v tomto případě 0).

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975} SE(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{0,975} \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

# Neparametrické alternativy nepárového t-testu

X1	X2	ALL	Rank ALL	X1 rank	X2 rank
27	25	25	5	6	5
35	29	29	7,5	11	7,5
38	31	31	9	13	9
37	23	23	4	12	4
39	18	18	2	14	2
29	17	17	1	7,5	1
41	32	32	10	15	10
	19	19	3		3
		27	6		
		35	11		
		38	13		
		37	12		
		39	14		
		29	7,5		
		41	15		

## Mann Whitney U-test

•Stejně jako řada jiných neparametrických testů počítá i tento test s pořadím dat v souborech namísto s originálními daty. Jde o neparametrickou obdobu nepárového t-testu a z těchto neparametrických testů má nejvyšší sílu testu (95% párového t-testu).

•V případě Mann-Whitney testu jsou nejprve čísla obou souborů sloučena a je vytvořeno jejich pořadí v tomto sloučeném souboru, pak jsou hodnoty vráceny do původních souborů a nadále se pracuje již jen s jejich pořadím.

•Pro oba soubory je tedy vytvořen součet pořadí a menší z obou součtů je porovnán s kritickou hodnotou testu, pokud je tato hodnota menší než kritická hodnota testu, zamítáme nulovou hypotézu shody distribučních funkcí obou skupin.

•Podobným způsobem je počítán i **Wilcoxon rank sum test** (pozor, existuje ještě Wilcoxonův párový test!!!)

# Man – Whitney test

**\*Změna počtu buněk po aplikaci preparátu: A**

27<sup>6</sup>, 35<sup>11</sup>, 38<sup>13</sup>, 37<sup>12</sup>, 39<sup>14</sup>, 29<sup>7,5</sup>, 41<sup>15</sup>

**\*Kontrolní skupina: B**

25<sup>5</sup>, 29<sup>7,5</sup>, 31<sup>9</sup>, 23<sup>4</sup>, 18<sup>2</sup>, 17<sup>1</sup>, 32<sup>10</sup>, 19<sup>3</sup>

**R<sub>A</sub>** ..... součet pořadí pro skupinu **A = 78,5**

**R<sub>B</sub>** ..... = **41,5**

$$U_A = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_A = 7 \cdot 8 + \frac{7 \cdot 8}{2} - R_A = 5,5$$

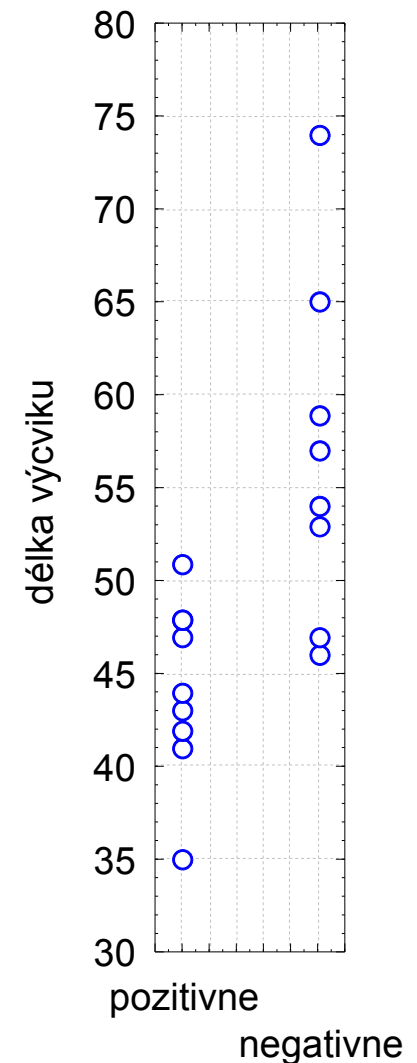
$$U_A + U_B = n_1 \cdot n_2 \Rightarrow U_B = 50,5$$

$$\min(U_A; U_B) = 5,5 \quad [n_1 = 7; n_2 = 8]$$

Pokud je **min(U<sub>A</sub>; U<sub>B</sub>)** menší než kritická hodnota testu, pak zamítáme hypotézu shody distribučních funkcí obou skupin

# Mann – Whitney test - příklad

- 17 štěňat bylo trénováno v chození na záchod metodou pozitivního posilování (pochvala, když jde na záchod venku) nebo negativního (trest, když jde na záchod doma). Jako parametr bylo měřeno, za kolik dní je štěně vycvičeno.
- nulová hypotéza je, že není rozdíl v metodách tréninku, tedy, že oběma metodami je štěně vycvičeno za stejnou dobu.
- po srovnání rozložení + malý počet hodnot je vhodné použít neparametrický test
- je vytvořeno pořadí sloučených hodnot
- pořadí hodnot v jednotlivých skupinách dat je sečteno a menší ze součtů je použit pro srovnání s kritickou hodnotou testu
- výsledkem testu je  $p < \alpha$ , nulovou hypotézu tedy zamítáme a výsledkem testu je, že pozitivní působení při výcviku štěňat dává lepší výsledky



# Párové two sample testy – předpoklady

- Skupiny dat jsou spojeny přes objekt měření, příkladem může být měření parametrů pacienta před léčbou a po léčbě (nemusí jít přímo o stejný objekt, dalším příkladem mohou být např. krysy ze stejné linie).
- Oba soubory musí mít shodný počet hodnot, protože všechna měření v jednom souboru musí být spárována s měřením v druhém souboru. Při vlastním výpočtu se potom počítá se změnou hodnot (diferencí) subjektů v obou souborech.
- Před párovým testem je vhodné ověřit si zda existuje vazba mezi oběma skupinami – vynesení do grafu, korelace.

## **Existuje několik možných designů experimentu, stručně lze sumarizovat:**

1. pokus je párový a jako párový se projeví
2. párové provedení pokusu – párově se neprojeví
  - možná párovost není
  - špatně provedený pokus – malé n, velká variabilita, špatný výběr jedinců
3. čekali jsme nezávislé a jsou
4. čekali jsem nezávislé a nejsou
  - vazba
  - náhoda

# Párový two sample t-test

- Tento test nemá žádné předpoklady o rozložení vstupních dat, protože je počítán až na základě jejich diferencí.
- Tyto difference by měly být normálně rozloženy a otázkou v párovém t-testu je, zda se průměrná hodnota diferencí rovná nějakému číslu, typicky jde o srovnání s nulou jako důkaz neexistence změny mezi oběma spárovanými skupinami.
- V podstatě jde o one sample t-test, kde místo rozdílu průměru vzorku a cílové populace je uveden průměr diferencí a srovnávané číslo (0 v případě otázky, zda není rozdíl mezi vzorky).
- Pro srovnání s 0 (testovou statistikou je t rozložení): 
$$t = \frac{\bar{D}}{s} \sqrt{n} \quad \nu = n - 1$$
- Někdy je obtížné rozhodnout, zda jde nebo nejde o párové uspořádání, párový test by měl být použit pouze v případě, že můžeme potvrdit vazbu (korelace, vynesení do grafu), jedním z důvodů proč toto ověřovat je fakt, že v případě párového t-testu není nutné brát ohled na variabilitu původních dvou souborů, tento předpoklad však platí pouze v případě vazby mezi proměnnými. Výpočet obou typů testů se vlastně liší v použité s, jednou jde o s diferencí, v druhém případě o složený odhad rozptylu obou souborů.
- Zda je párové uspořádání efektivnější lze určit na základě:
  - Síly vazby
  - Je-li  $s_D$  výrazně menší než  $s_{x_1-x_2}$
- Závislost je možné rozepsat pomocí vzorce: 
$$s_D^2 \cong \sigma_{x_1}^2 + \sigma_{x_2}^2 - 2Cov(x_1; x_2)$$
- v případě  $Cov=0$ , tedy v případě neexistence vazby pak  $s_D^2$  odpovídá součtu původních rozptylů, tedy přibližně  $S_{x_1-x_2}$ .



# Two sample testing: paired design

$$H_0 : \mu_d = 0 \quad t = \frac{\bar{D}}{S_{\bar{D}}} \cong \frac{\bar{D} - \mu_d}{S_{\bar{D}}}$$

$$\mu_d : \bar{D} \pm t_{1-\alpha/2}^{(n-1)} S_{\bar{D}}$$

Paired Independent

$$\sigma_D^2 ? \approx \sigma_{\bar{X}_1 - \bar{X}_2}^2$$

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 - 2Cov(\bar{X}_1; \bar{X}_2)$$

Mathematically:

$$\sigma_D \sim 2\sigma$$

$$\sigma_D^2 \sim 2\sigma^2$$

1. Evaluate experiment as paired and as independent

$$\sigma_D^2 \sim 2\sigma^2 \quad S_p^2$$

$$2. \quad 2\hat{\sigma}^2 = 2S_p^2 - \frac{2S_p^2 - S_D^2}{(2n-1)}$$

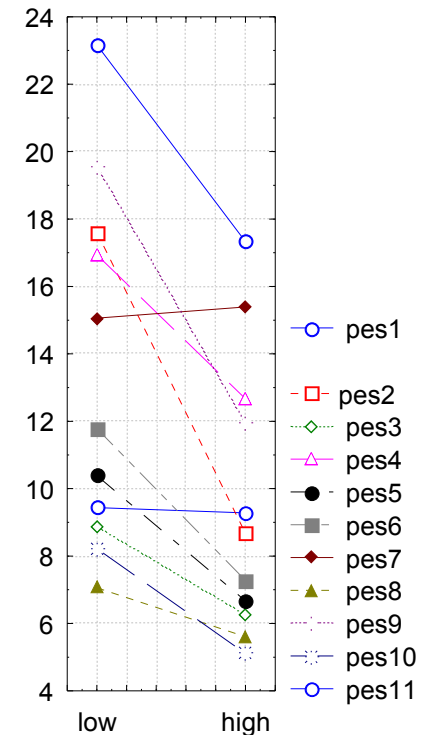
# Párový two sample t-test – příklad

Byl prováděn pokus s dietou 11 diabetických psů, každý pes byl vystaven dvěma dietám s odlišným typem sacharidů (snadno vstřebatelné X pozvolna se rozkládající na glukózu), hodnoty krevní glukózy v průběhu jednotlivých diet mají být srovnány pro zjištění vlivu diety na hladinu krevní glukózy. Protože každý pes absolvoval obě diety, jde o párové uspořádání, kdy výsledky hodnoty v obou pokusech jsou spojeny přes pokusné zvíře.

1. Nulová hypotéza zní, že skutečný průměrný rozdíl mezi oběma dietami je 0, alternativní hypotéza zní, že to není 0.
2. Pro každého psa je spočítán rozdíl mezi jeho hladinou glukózy při obou dietách a měly by být ověřeny předpoklady pro one sample t-test – tedy alespoň přibližně normální rozložení.
3. Je spočítána testová charakteristika, výpočet vlastně probíhá jako one-sample t-test, kde je zjišťována významnost průměru diferencí obou souborů jako rozdíl mezi touto hodnotou a nulou (nula je hodnota, kterou by průměrná diference měla nabývat, pokud platí nulová hypotéza).  $T=4.37$  s 10 stupni volnosti, skutečná hodnota  $p=0,0014$  a tedy na hladině  $p=0,05$  můžeme nulovou hypotézu zamítnout

$$t = \frac{\text{rozdíl}_\text{průměru}_\text{vzorku}_\text{a}_\text{populace}}{SE(\text{průměru})} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

4. Závěrem můžeme říci, že nulová hypotéza neexistence rozdílu mezi oběma dietami byla zamítnuta, což znamená, že high-fibre dieta má významný vliv na snížení hladiny krevní glukózy.



# Paired?

A one-tailed t test for the hypotheses  $H_0: \mu \geq 0$  and  $H_A: \mu < 0$

Máme hodnoty hmotnostních změn u lidí, seřazené po užívání drog, které mají za následek ztrátu hmotnosti. Každá změna hmotnosti (v kg) je hmotnost po mínus hmotnost před užitím drogy.

$$n = 12$$

$$\bar{X} = -0,61 \text{ kg}$$

$$s^2 = 0,4008 \text{ kg}^2$$

$$s_{\bar{x}} = \sqrt{\frac{0,4008 \text{ kg}^2}{12}} = 0,18 \text{ kg}$$

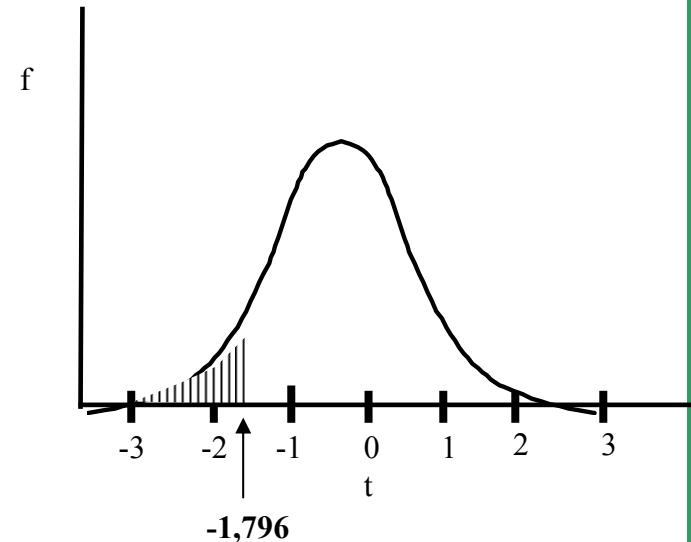
$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}} = \frac{-0,61 \text{ kg}}{0,18 \text{ kg}} = -3,389$$

$$v = n - 1 = 11$$

$$t_{0,05(1),11} = 1,796$$

**Když**  $t \leq -1,796$  , **zamítáme**  $H_0$ .

$$0,0025 < P(t \leq -3,389) < 0,005$$



Rozložení t pro  $v=11$ , ukazuje kritickou oblast rozložení (šrafováno) pro one-tailed test s  $\alpha=0,05$ . (Kritická hodnota t je  $-1,796$ .)

0,2  
-0,5  
-1,3  
-1,6  
-0,7  
0,4  
-0,1  
0,0  
-0,6  
-1,1  
-1,2  
-0,8

# Neparametrická obdoba párového t-testu

## Wilcoxon test

- Jsou vytvořeny difference mezi soubory, je vytvořeno jejich pořadí bez ohledu na znaménko a poté je sečteno pořadí kladných a pořadí záporných rozdílů. Menší z těchto dvou hodnot je srovnána s kritickou hodnotou testu a pokud je menší než kritická hodnota testu, pak zamítáme hypotézu shody obou souborů hodnot. Pro test existuje aproximace na normální rozložení, ale pouze pro velká  $n > 25$ .

$$t = \frac{\text{Menší\_suma\_diferencí} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Před zásahem	Po zásahu	Změna	Absolutní pořadí
6	2	4	10
2,5	3	-0,5	1,5
6,3	5	1,3	6
8,1	9	-0,9	5
1,5	2	-0,5	1,5
3,4	4	-0,6	3
2,5	1	1,5	8
1,11	2	-0,89	4
2,6	4	-1,4	7
1	3	-2	9

# Wilcoxonův test – příklad I

člověk	A	B	diference	pořadí
1	142	138	4	4,5
2	140	136	4	4,5
3	144	147	-3	3
4	144	139	5	7
5	142	143	-1	1
6	146	141	5	7
7	149	143	6	9,5
8	150	145	5	7
9	142	136	6	9,5
10	148	146	2	2

A.....parametr krve před podáním léku

B.....parametr krve po podání léku

$W_+$  .....  $\Sigma$  pořadí kladných rozdílů = 51

$W_-$  ..... = 4

$W = \min(W_+; W_-) = 4$

.....  
počet párů =  $n = 10$

Pokud je  $W$  menší než kritická hodnota testu, pak zamítáme hypotézu shody distribučních funkcí obou skupin.

## Wilcoxonův test – příklad II

Byla testována nová dieta pro laboratorní krysy, při pokusu byl zjišťován její vliv na různých liniích krys, bylo proto zvoleno párové uspořádání kdy krysy v obou dietách jsou spojeny přes svoji linii, tj. na začátku byly dvojice krys stejné linie, jedna z nich byla náhodně přiřazena k dietě, druhá z dvojice pak do druhé diety.

1. nulová hypotéza je, že váha krys není ovlivněna použitou dietou, alternativní, že ovlivnění dietou existuje
2. spočítáme difference – tyto difference jsou nenormální a proto je vhodné využít neparametrický test
3. Spočítáme sumu pořadí kladných a záporných diferencí, zde je menší suma záporných diferencí – 31
4. výsledkem výpočtu je  $p > 0,05$  a tedy nemáme dostatečné důkazy pro zamítnutí nulové hypotézy, nelze říci, že by nová dieta byla efektivnější než stará
5. pro doplnění výsledků je vhodné zjistit také skutečnou velikost rozdílu hmotností ve skupinách, např. ve formě mediánu

# Znaménkový test – příklady I

## Párově uspořádaný experiment pro nominální data

### I. Dva preparáty, každý na ½ listu

- sledovaná veličina: počet skvrn (hodnoceno pouze jako rozdíl)

	Počet skvrn									
A	V	V	M	V	V	M	M	V	V	V
B	M	M	V	M	M	V	V	M	M	M

V – větší; M – menší

n = 10 listů s rozdílnými výsledky

jev → A je větší: +  $n_+ = 7$

jev → B je menší: -  $n_- = 3$

$$\min(n_+; n_-) = 3$$

### II. dvě protilátky z různých zdrojů (A;B)

- aplikované na vzorek s antigenem

n = 10

A	+	+	-	+	-	+	-	+	+	-
B	-	-	+	-	+	+	-	-	+	-

n – nenulových rozdílů: 6

$$A: n_+ = 4$$

$$A: n_- = 2$$

$$\min(n_+; n_-) = 2$$

## Znaménkový test – příklady II

- Na konferenci veterinářů bylo předneseno, že průměrný čas konzultace je 12 minut. Následovala debata, zda je lepší použít medián nebo průměr. Jeden z nich se rozhodl ověřit teorii, že průměrná konzultace trvá 12 minut na vlastní praxi a zaznamenal si trvání svých 43 konzultací. K otestování hypotézy, že podíl konzultací kratších a delších než 12 minut použil znaménkový test.

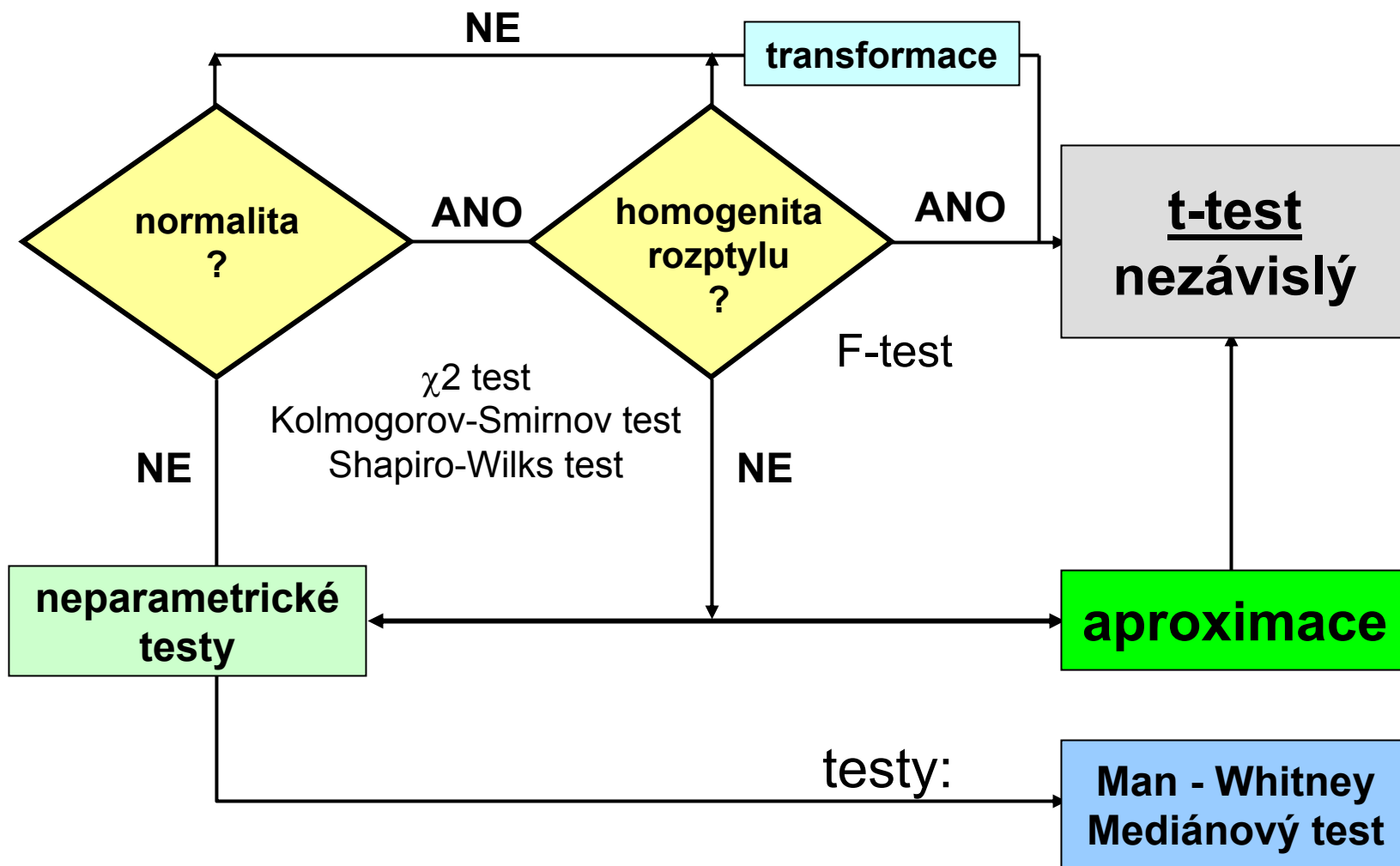
Délka konzultace	Počet
<12	22
12	6
>12	15
Celkem	43

Další výpočet probíhá obdobně jako v případě klasického znaménkového testu na diferencích dvou skupin dat.



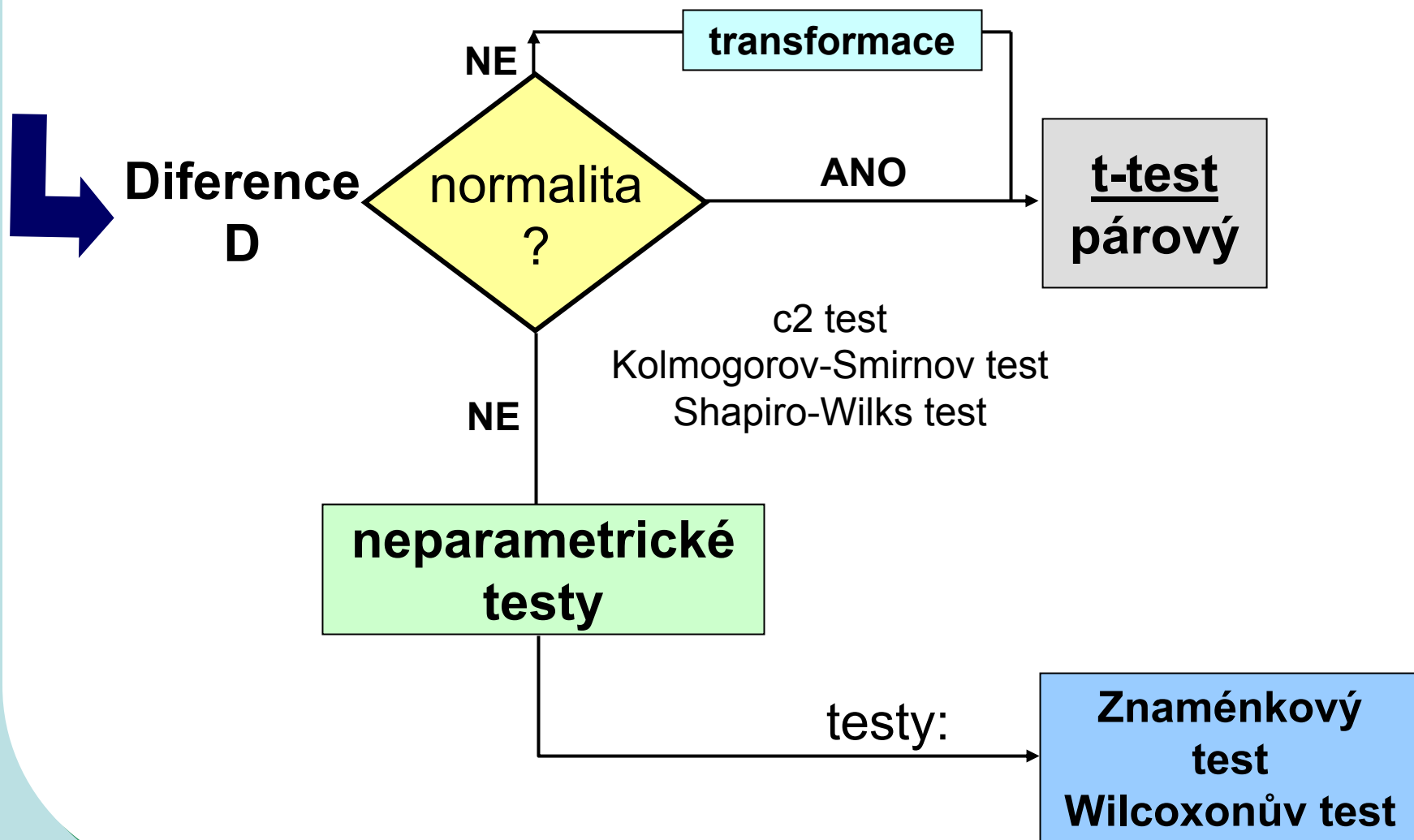
# Srovnání dvou pokusných zásahů - obecné schéma zapojených testů III

## Nezávislé uspořádání



# Srovnání dvou pokusných zásahů - obecné schéma zapojených testů IV

## Párové uspořádání



# Testování – typ dat

---

- Spojitá čísla
    - T test, Mann-Whitney test, Wilcoxon test, Znaménkový test atd.
- 

- Binární data?
- Kategoriální data?
  - Výše zmíněné testy nelze použít
  - Základní přístupy testování lze ovšem použít i na tato data
    - Nulová a alternativní hypotéza
    - One sample a two sample testy

- Analýzy na binomickém rozložení
- Analýzy na Poissonově rozložení
- Analýza kontingenčních tabulek



## 13. Binomické rozložení



# Alternativní rozložení

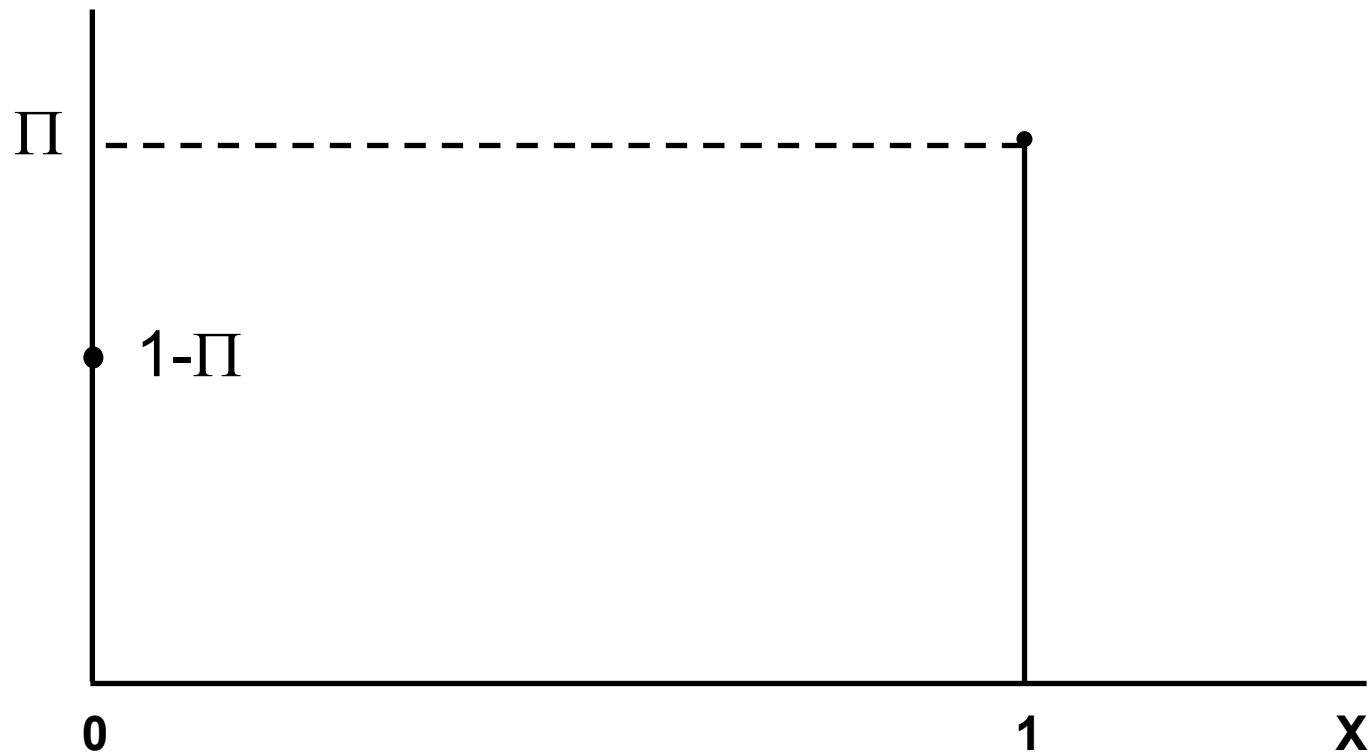
$$\Pi(x) = \Pi \text{ pro } X = 1$$

$$\Pi(x) = 1 - \Pi \text{ pro } X = 0$$

$$\Pi(x) = 0 \text{ jinak}$$



$X = 1$  .....jev



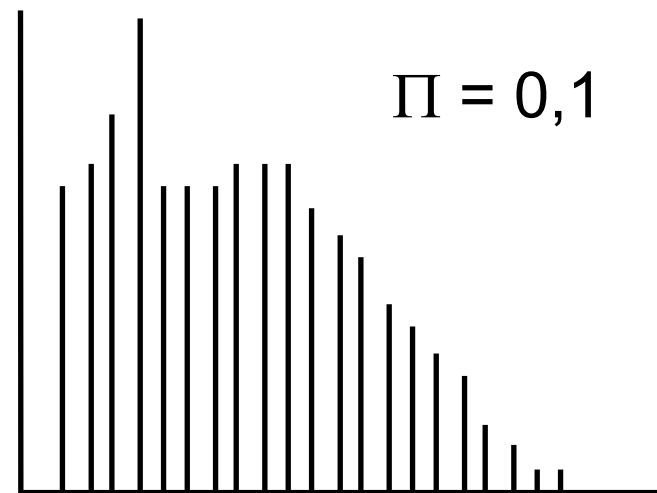
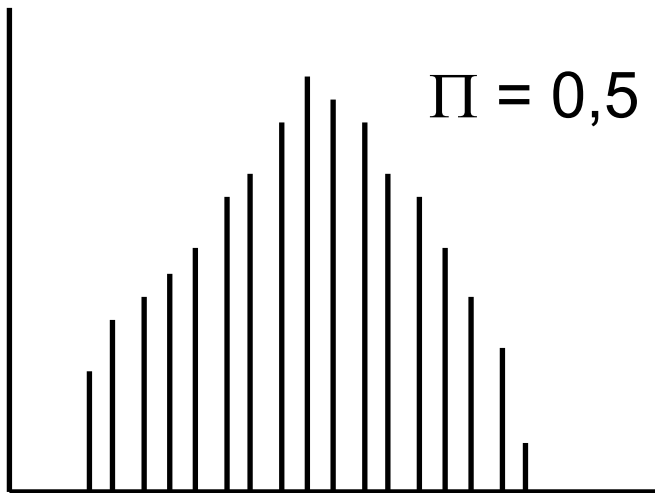
# Binomické rozložení

$X$  ..... celkový počet nastání jevu v  $n$  nezávislých pokusech

$$E(x) = n \cdot \Pi$$

$$D(x) = n \cdot \Pi (1 - \Pi)$$

$\Pi \sim p$   **jediný parametr distribuce určuje tvar distribuce**



# Binomické rozložení

## - model pro zkoumání výskytu sledovaného jevu

$n$  ..... počet nezávislých opakování (dotazů)

$X$  ..... počet lidí s jistým symptomem

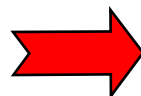
$r$  znamená celkový počet nastání jevu v  $n$  nezávislých experimentech

$r : 0 \dots n$

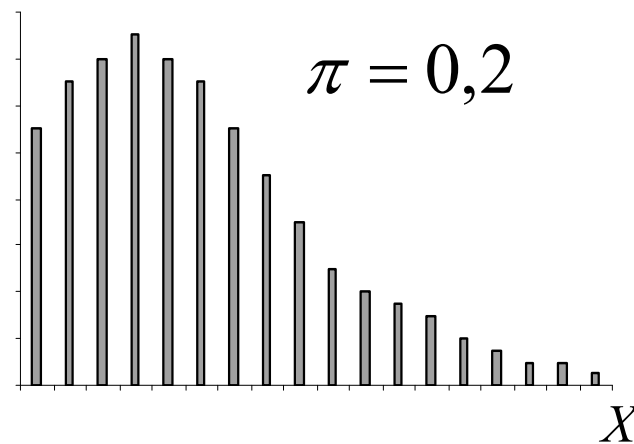
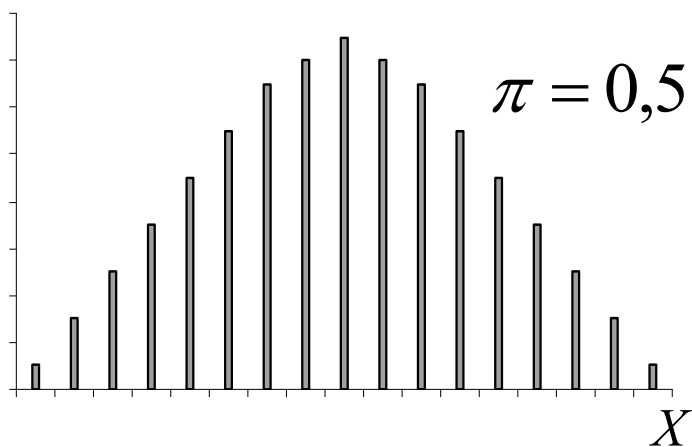
$p \sim \pi$  .. jediný parametr binomického rozložení

$p$  .... relativní četnost nastání jevu

$p$  ..... určuje tvar distribuce



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



**Binomická proměnná  $X$**

# Binomické rozložení jako model

**Jev:** narození chlapce  $\Pi = 0,5$   
**n :** rodina s 5 dětmi  
**r:** 0,1,2,3,4,5 chlapců

$$P(r) = \binom{n}{r} \cdot p^r \cdot (1-p)^{(n-r)} = \frac{n!}{r!(n-r)!} \cdot p^r \cdot q^{(n-r)}$$

$$r = 0: \frac{5!}{(0! 5!)} \cdot (0,5)^0 \cdot (0,5)^5 = 0,031$$

$$r = 1: \frac{5!}{(1! 4!)} \cdot (0,5)^1 \cdot (0,5)^4 = 0,15625$$

$$r = 2: P(r) = 0,3125$$

$$r = 3: P(r) = 0,3125$$

$$r = 4: P(r) = 0,15625$$

$$r = 5: P(r) = 0,031$$

**X:** Binomická proměnná

**Střed rozložení:**

**Rozptyl:**  $E(x) = n \cdot p$

$$D(x) = n \cdot p \cdot (1 - p)$$

**Příklad:**  $n = 100$  respondentů  
 $r = 20$  má symptom



$$E(x) = n \cdot p = 20$$

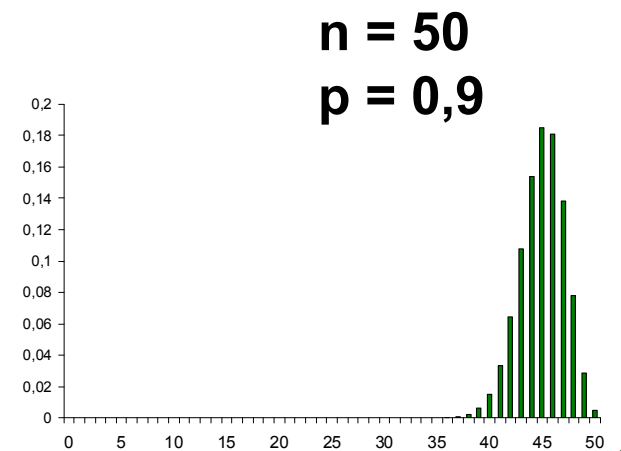
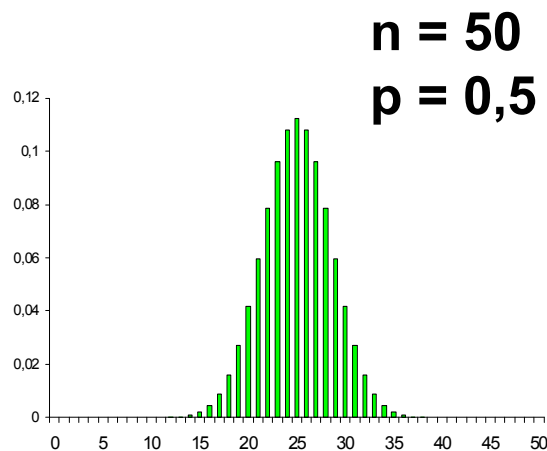
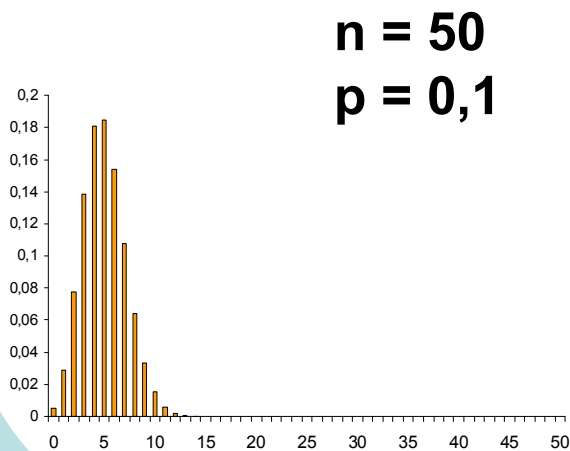
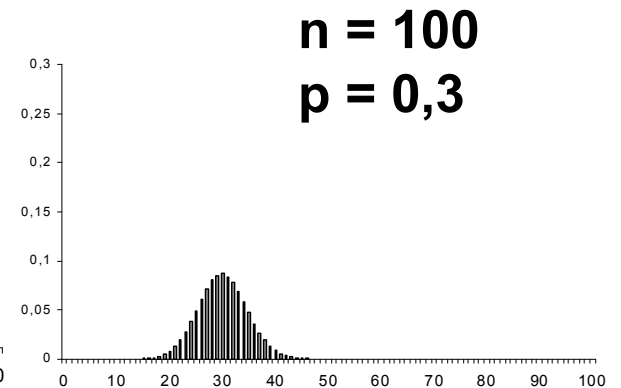
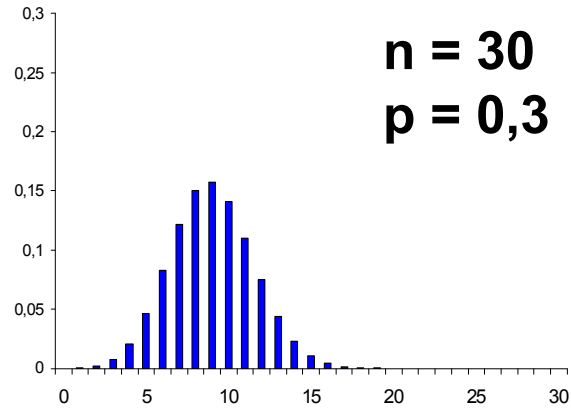
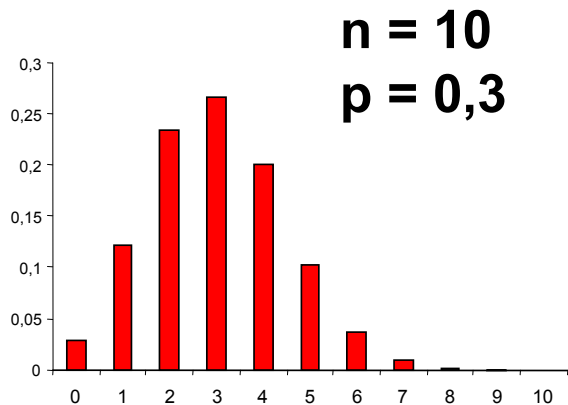
**je střed rozložení  
a nejpravděpodobnější  
hodnota**



# Binomické rozložení jako model

$$P(x = r) = \frac{n!}{r!(n-r)!} \cdot p^r \cdot q^{(n-r)}$$

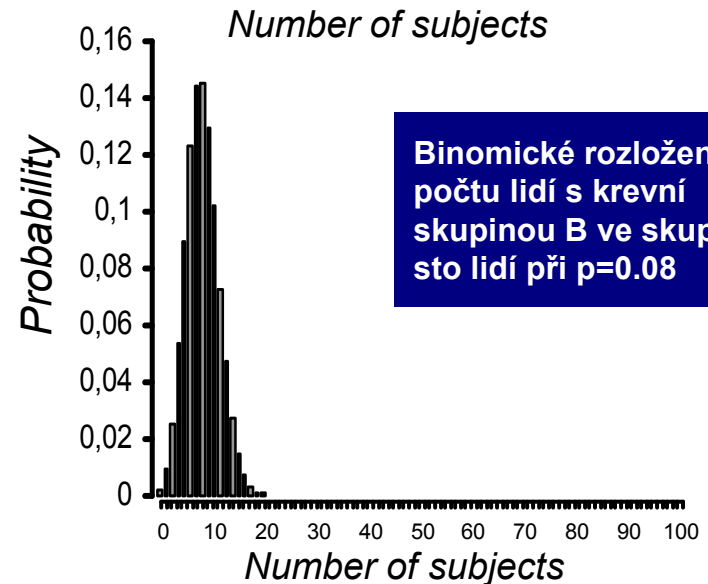
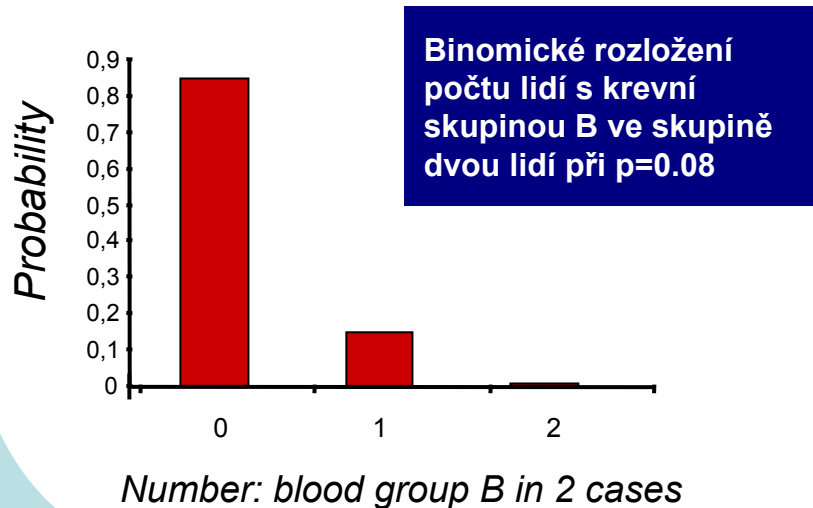
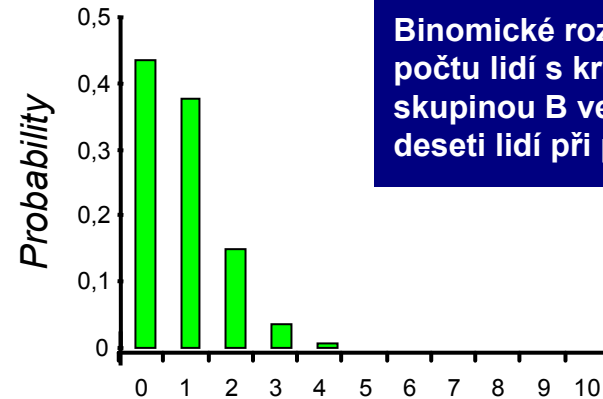
$$q = 1 - p$$



# Aplikace binomického rozložení

Výskyt krevní skupiny B v určité populaci:  $p = 0,08$

			Počet s krevní skupinou B	Pravděpodobnost
B	B	2		0,0064
not B	B	1		0,0736
B	not B	1		0,0736
not B	not B	0		0,8464



# Aplikace binomického rozložení

*Populace: 60% jedinců má zvýšenou hladinu cholesterolu*

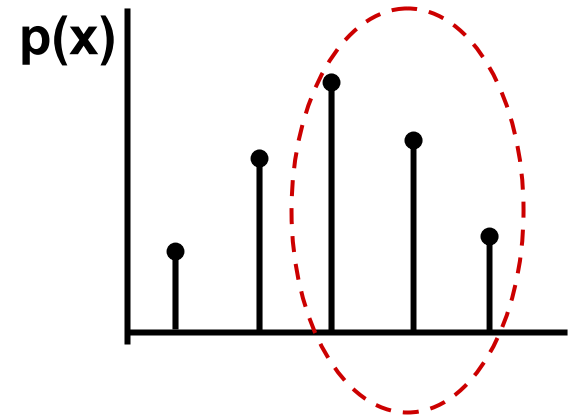
*Výběr: 5 lidí*

## I. Kolik lidí má ve výběru vyšší hladinu cholesterolu ?

II. Jaká je  $P$ , že právě 3 lidé budou mít vyšší hladinu cholesterolu ? ~ Tzn. Výběr přesně odpovídá dané populaci ?

$P(3) = ?$

Jaká je  $P$ , že většina jedinců (tedy minimálně 3) má vyšší hladinu cholesterolu ? ~ Tzn. výběr alespoň obecně odpovídá zkoumané populaci ?



# Aplikace binomického rozložení

*Populace: 60% jedinců má zvýšenou hladinu cholesterolu*

*Výběr: 5 lidí*

## I. Kolik lidí má ve výběru vyšší hladinu cholesterolu ?

$$n \cdot p = 5 \cdot 0,6 = 3 \text{ lidé} \sim E(x)$$

$$n \cdot p(1-p) = 1,2 \sim D(x)$$

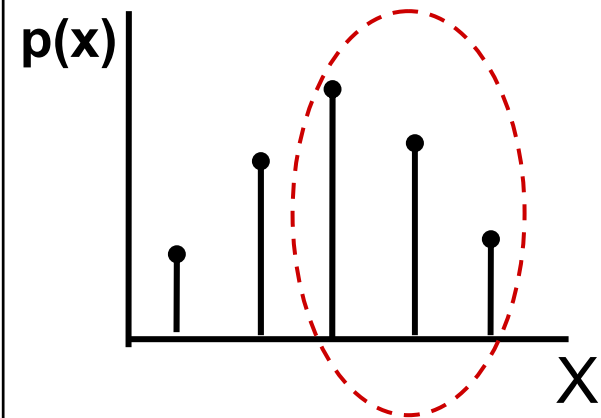
II. Jaká je P, že právě 3 lidé budou mít vyšší hladinu cholesterolu ? ~ Tzn. Výběr přesně odpovídá dané populaci ?

$$P(3) = ? \quad P_{(3)} = \frac{5!}{3!(5-3)!} \cdot (0,6)^3 \cdot (0,4)^2 = 0,346$$

$$P(3) = 35\%$$

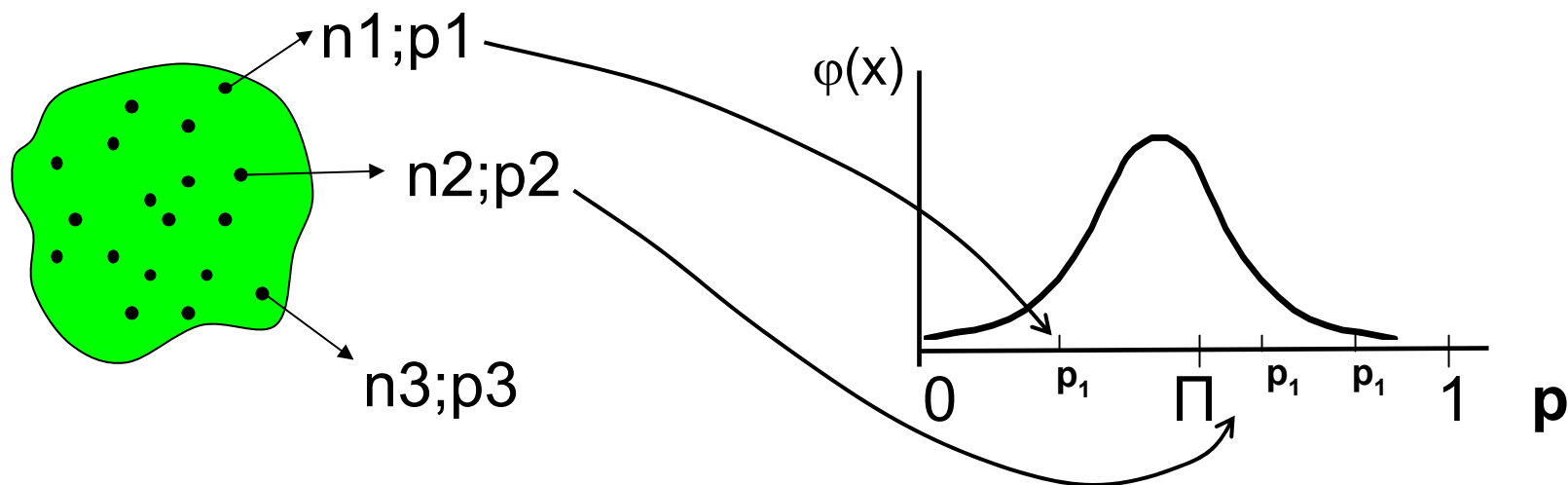
Jaká je P, že většina jedinců (tedy minimálně 3) má vyšší hladinu cholesterolu ? ~ Tzn. výběr alespoň obecně odpovídá zkoumané populaci ?

$$P(X > 3) = P(3) + P(4) + P(5) = 0,346 + 0,259 + 0,078 = 68 \%$$

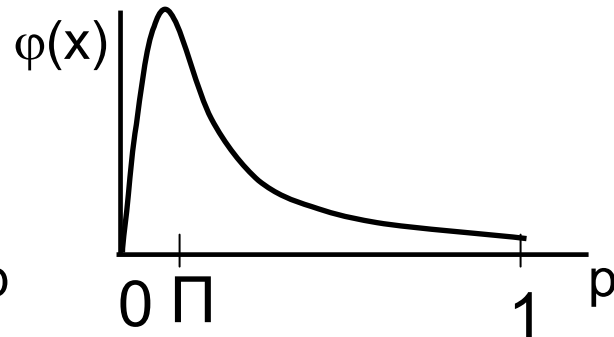
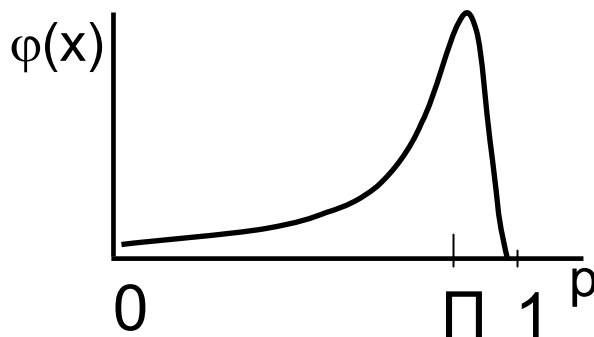


# Odhad parametru $\Pi$ binomického rozložení

*Při vícenásobném odhadu se parametr  $\Pi$  chová jako normálně rozložěn*



U malých nebo velkých hodnot  $p$  ( $\Pi$ ) je však předpoklad normality omezen



# Odhad parametru $\Pi$ binomického rozložení

## I. vztahy

$$\pi \approx \hat{p}; \quad \hat{p} = r/n$$

1) Bodový

$$\hat{p}; \quad s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}$$

2) Intervalový – aproximace

$$\hat{p} - Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \leq \pi \leq \hat{p} + Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

$$\pi: \hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}}$$

# Odhad parametru $\pi$ binomického rozložení

## *II. aproximace*

**X: % jedinců s daným znakem**

**n = 100 jedinců**

**r = 60;  $\hat{p} = 0,6$**

**$s_{\hat{p}} = 0,049$**

**Interval spolehlivosti : 95 %**

**$Z_{0,975} = 1,96$**

$$0,6 - 1,96 \cdot 0,049 \leq \pi \leq 0,6 + 1,96 \cdot 0,049$$

$$\boxed{0,504 \leq \pi \leq 0,697}$$



$$P(0,504 \leq \pi \leq 0,697) \geq 0,95$$

# Odhad parametru $p$ binomického rozložení

## Intervalový odhad bez aproximací na normální rozložení - I. Vztahy

$$L_1 = \frac{r}{r + (n - r + 1) \cdot F_{\alpha/2}^{(v_1; v_2)}}$$



spodní limit intervalu

$$v_1 = 2(n - r + 1); \quad v_2 = 2r$$

$$L_2 = \frac{(r + 1) \cdot F_{\alpha/2}^{(v'_1; v'_2)}}{n - r + (r + 1) \cdot F_{\alpha/2}^{(v'_1; v'_2)}}$$



horní limit intervalu

$$v'_1 = 2(r + 1) = v_2 + 2$$

$$v'_2 = 2(n - r) = v_1 - 2$$

$$P(L_1 \leq \pi \leq L_2) \geq 1 - \alpha$$



# Odhad parametru p binomického rozložení

## Intervalový odhad bez aproximací na normální rozložení - II. Příklad:

Náhodný vzorek  $n = 200$  jedinců.

Zjištěno pouze  $r = 4$  jedinci bez určitého znaku.

$$\hat{p} = \frac{4}{200} = \underline{\underline{0,02}}$$

95% interval spolehlivosti = ?

### Spodní hranice

$$v_1 = 2(n - r + 1) = 2(200 - 4 + 1) = 394$$

$$v_2 = 2r = 2 \cdot 4 = 8$$

$$F_{1-\alpha/2}^{(394;8)} = \underline{\underline{3,67}}$$

$$L_1 = \frac{4}{4 + (200 - 4 + 1) \cdot 3,67} = \underline{\underline{0,0055}}$$

### Horní hranice

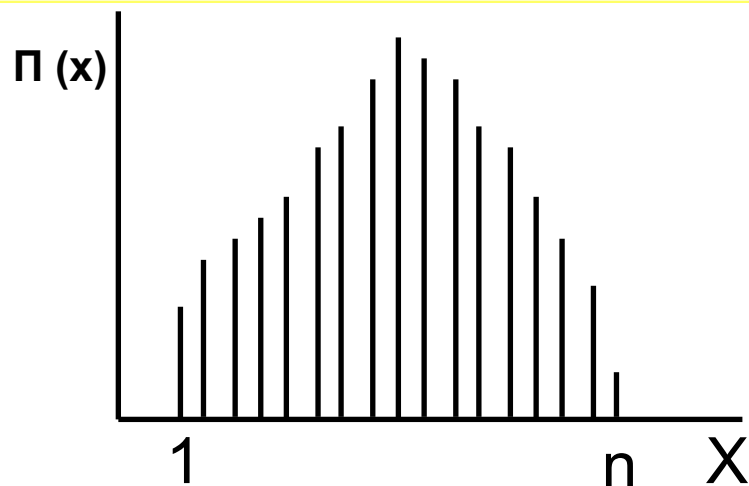
$$v'_1 = 2(r + 1) = 10$$

$$v'_2 = 2(n - r) = 2(200 - 4) = 392$$

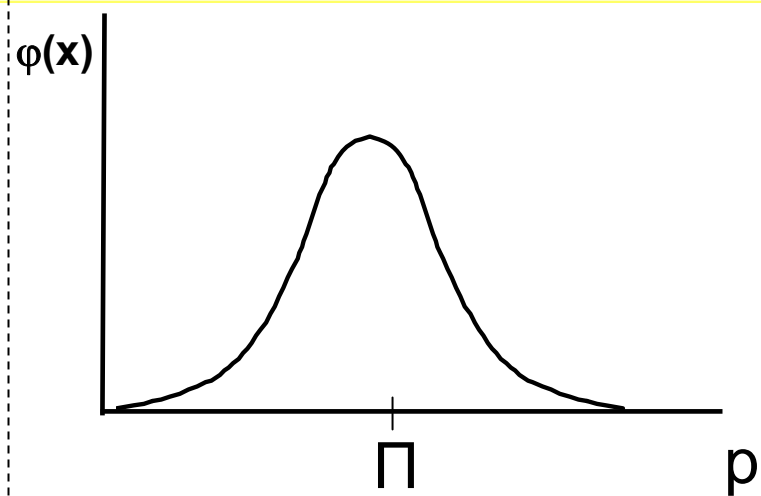
$$F_{1-\alpha/2}^{(10;392)} = \underline{\underline{2,08}}$$

$$L_2 = \frac{(4 + 1) \cdot 2,08}{200 - 4 + (4 + 1) \cdot 2,08} = \underline{\underline{0,051}}$$

# Binomické rozložení v datech - shrnutí

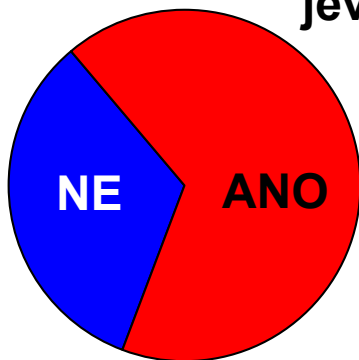


Pravděpodobnost výskytu hodnot  $X$

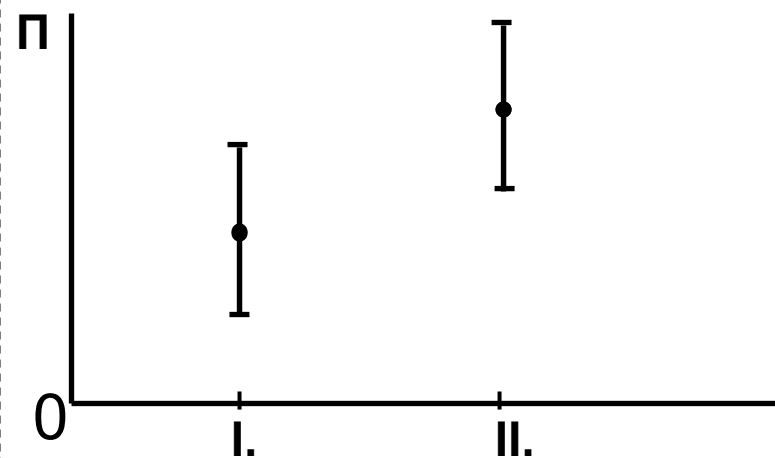


Modelové rozložení odhadovaného parametru

$n$  opakování      jev ANO  
jev NE



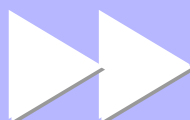
Binární podstata původních hodnot



Interval spolehlivosti pro  $\Pi$

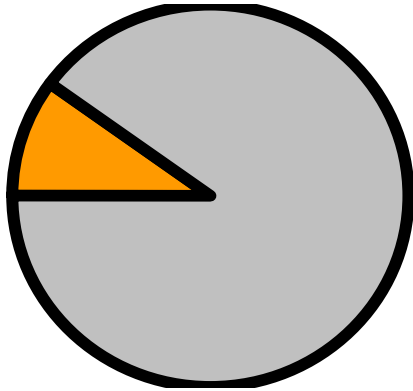


## 14. Binomický test



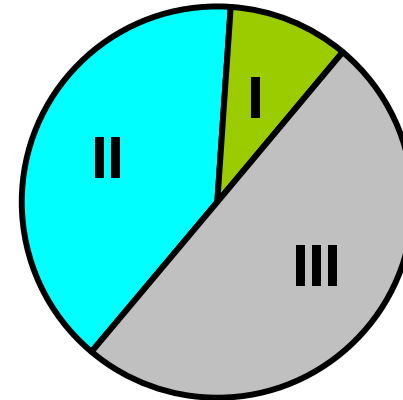
# Analýza binárních nebo kategoriálních dat I.

Binární proměnná  
(1 / 0)



<u>1</u> : 10	}	$p_1 = 0,1$ $p_0 = 0,9$
<u>0</u> : 90		
<u>n</u> : 100		

Kategoriální  
proměnná



<u>I</u> : 10	}	$p_I = 0,1$ $p_{II} = 0,4$ $p_{III} = 0,5$
<u>II</u> : 40		
<u>III</u> : 50		
<u>n</u> : 100		

# Analýza binárních nebo kategoriálních dat II.

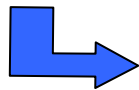
---

I.

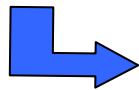
Liší se odhad  $\underline{p}$  od předpokládané hodnoty  $P$  ?

II.

Liší se dva nebo více odhadů  $\underline{p}$  ?



- závislé odhady -



- nezávislé odhady -

III.

Je výskyt kategorií dvou jevů nezávislý ?

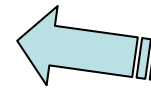
IV.

Hodnocení relativního rizika z výskytu určitého jevu v rámci skupiny lidí

# Jednovýběrový binomický test (One sample binomial test)

$H_0$	$H_A$	Testová statistika	Interval spolehlivosti
$p \leq \Pi$	$p > \Pi$	$z$	$z > z_{1-\alpha}$
$p \geq \Pi$	$p < \Pi$	$z$	$z < z_{\alpha}$
$p = \Pi$	$p \neq \Pi$	$z$	$ z  > z_{1-\alpha/2}$

$$Z = \frac{n \cdot \hat{p} - n \cdot \pi}{\sqrt{n \cdot \hat{p}(1 - \hat{p})}} \cong \frac{|n \cdot \hat{p} - n \cdot \pi| - 0,5}{\sqrt{n \cdot \hat{p}(1 - \hat{p})}}$$



**Korekce na  
kontinuitu**

$H_0$	$H_A$	Testová statistika	Interval spolehlivosti
$p \leq \Pi$	$p > \Pi$	$L_1 = \frac{(r + 1) F_{\alpha, v_1', v_2'}}{n - r + (r + 1) F_{\alpha, v_1', v_2'}}$	$p = r / n > L_1$
$p \geq \Pi$	$p < \Pi$	$L_2 = \frac{r}{r + (n - r + 1) F_{\alpha, v_1', v_2'}}$	$p < L_2$
$p = \Pi$	$p \neq \Pi$	$L_1; L_2 (F_{\alpha/2}; F_{1-\alpha/2})$	$p < L_2 \vee p > L_1$

## Test $p ? \pi$

✓ Stromy s pozmeněným tvarem koruny

$n = 9\ 000$  jedinců

$r = 2\ 250$  změněných jedinců

**?** Jak je pravděpodobná změna u až 1/3 jedinců? **?**

$$Z = \frac{n \cdot p - n \cdot \pi}{\sqrt{p(1-p) \cdot n}} = \frac{2250 - 3000}{\sqrt{0,25 \cdot 0,75 \cdot 9000}} = \underline{\underline{-18,26}}$$

$$\alpha = 5\ %; \quad Z_{1-\alpha/2} = 1,96; \quad Z_{1-\alpha} = 1,645$$

$Z > Z_{1-\alpha/2}$  .....zamítáme  $H_0: p = 0,3$

$$P \ll 0,1$$

**95 % Interval spolehlivosti ...  $p: (0,241; 0,258)$**

# Test $p ? \pi$

## Příklad testu bez aproximace na normální rozložení

- ✓ 12 jedinců bylo zkoumáno pro výskyt určitého znaku, 10 jedinců znak nemělo
- ? Jak hodně se tento výsledek liší od výsledku 6 - 6: tedy od situace, kdy polovina jedinců znak má?

### a) Využití distribuční funkce

r	0	1	2	3	4	5	6	7	8	9	10	11	12
<b>P(r)</b>	0,0002 4	0,0029 3	0,0161 1	0,0537 1	0,1208 5	0,1933 5	0,2255 9	0,1933 6	0,1208 5	0,0537 1	0,0161 1	0,0029 3	0,0002 4

$$P(r \geq 10) = 0,01611 + 0,00393 + 0,00024 = 0,01928$$

**$H_0: p = 0,5$  je tedy značně nepravděpodobná**

b) Pozorované  $\hat{p} = \frac{10}{12} = 0,833$  překročilo horní limit 95 % intervalu spolehlivosti pro p:

$$p = 0,5 : L_2 = \frac{(6 + 1) \cdot 2,64}{12 - 6 + (6 + 1) \cdot 2,64} = \underline{\underline{0,755}}$$



## Dvouvýběrový binomický test ( $p_1 \neq p_2$ )

$$Z = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\bar{p}(1 - \bar{p})}{n_1} + \frac{\bar{p}(1 - \bar{p})}{n_2}}}$$

$$\bar{p} = \frac{n_1 \cdot \bar{p}_1 + n_2 \cdot \bar{p}_2}{n_1 + n_2}$$

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{(1-\alpha/2)} \cdot \sqrt{\frac{\bar{p}(1 - \bar{p})}{n_1} + \frac{\bar{p}(1 - \bar{p})}{n_2}}$$

# Dvouvýběrový binomický test ( $p_1 \neq p_2$ )

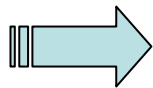
*Tento příklad je původní ukázkou testování rozdílů mezi dvěma binomickými populacemi (tedy srovnání dvou odhadů parametru  $p$ ).*

✓ Celkem 49 pokusných myší bylo použito k testování toxického preparátu během dvouměsíční kultivace. Následující tabulka obsahuje původní data zároveň s testem nulové hypotézy: Podíl přežívajících jedinců je u zasažené populace stejný.

	Alive	Dead	Total	Proportion alive	Proportion dead
<b>Treated</b>	15	9	24	$\hat{p}_1 = 0,625$	$\hat{q}_1 = 0,375$
<b>Not Treated</b>	10	15	25	$\hat{p}_2 = 0,400$	$\hat{q}_2 = 0,600$
<b>Total</b>	25	24	49	$\hat{p} = 0,510$	$\hat{q} = 0,490$

$$Z = \frac{0,625 - 0,400}{\sqrt{\frac{(0,510)(0,490)}{24} + \frac{(0,510)(0,490)}{25}}} = \frac{0,225}{\sqrt{0,010413 + 0,009996}} = 1,573$$

$Z_{0,05(2)} = t_{0,05(2)} = 1,96$

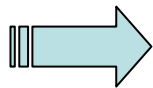


**Nezamítáme  $H_0$ :  $0,10 < P < 0,20$**

**S korekcí na kontinuitu:**

$$Z = \frac{\frac{15 - 0,5}{24} - \frac{10 + 0,5}{25}}{0,143} = \frac{0,604 - 0,420}{0,143} = 1,287$$

$Z_{0,05(2)} = t_{0,05(2)} = 1,96$



**Nezamítáme  $H_0$ :  $0,10 < P < 0,20$**



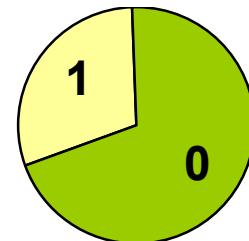
## 15. Analýza kontingenčních tabulek



# Test dobré shody - základní teorie

## Binomické jevy (1/0)

$$\chi_{(1)}^2 = \underbrace{\frac{\left[ \begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\text{očekávaná četnost}}}_{\text{I. jev 1}} + \underbrace{\frac{\left[ \begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\text{očekávaná četnost}}}_{\text{II. jev 2}}$$



### Příklad

✓ 10 000 lidí hází mincí → rub: 4 000 případů (R)  
líc: 6 000 případů (L)

? Lze výsledek považovat za statisticky významně odlišný (nebo neodlišný) od očekávaného poměru R : L = 1 : 1 ?

$$\chi_{(1)}^2 = \frac{(4000 - 5000)^2}{5000} + \frac{(6000 - 5000)^2}{5000} = 400$$

Tabulková hodnota:  $\chi_{(0,95)}^2 (v = 1) = \underline{\underline{3,84}}$  (0,95 = 1 -  $\alpha$ )



**Rozdíl je vysoce statisticky významný (p << 0,001)**

# Kontingenční tabulky - $H_0$ : Nezávislost dvou jevů A a B

$$P(A \cap B) = P(A) \cdot P(B)$$

$\begin{array}{c} \rightarrow \\ \downarrow \end{array} \begin{array}{c} B \\ A \end{array}$	+	-	Podíl (+)
+	a	b	$\frac{a}{(a+b)}$ $\mathbf{p_1}$
-	c	d	$\frac{c}{(c+d)}$ $\mathbf{p_2}$
Podíl (+)	$\frac{a}{(a+c)}$	$\frac{b}{(b+d)}$	

$$N = a + b + c + d$$

$$P(B^+) = \frac{(a+b)}{N}$$

$$P(B^-) = \frac{(c+d)}{N}$$

**Kontingenční  
tabulka  
2 x 2**

**Očekávané četnosti:**

$$F_{(A)} = \frac{(a+b)(a+c)}{N}$$

$$F_{(C)} = \frac{(a+c)(d+c)}{N}$$

$$F_{(B)} = \frac{(a+b)(b+d)}{N}$$

$$F_{(D)} = \frac{(b+d)(c+d)}{N}$$

$$\chi_{\nu=1}^2 = \sum_{i=1}^4 \frac{(f_i - F_i)^2}{F_i}$$

$$\nu = 1 = s - p - 1$$

**p ..... počet parametrů = 2**

$$P_{(A)}; P_{(B)}$$

$$\chi_c^2 = \sum \sum \frac{(|f_{ij} - F_{ij}| - 0,5)^2}{F_{ij}}$$

## 2 x 2 kontingenční tabulka - příklad ( $\alpha = 0,05$ )

gen \ †	Ano	Ne	$\Sigma$
Ano	20	82	102
Ne	10	54	64
$\Sigma$	30	136	166

$$F_A = 102 * 30 / 166 = 18,43$$

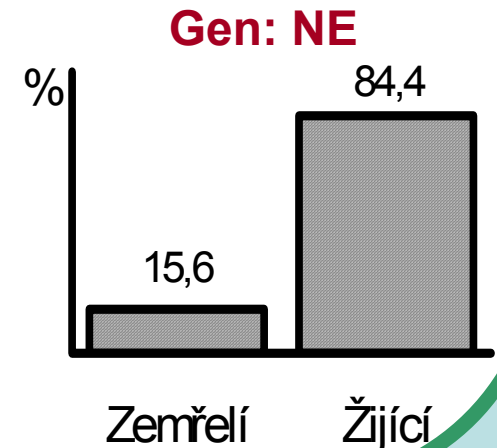
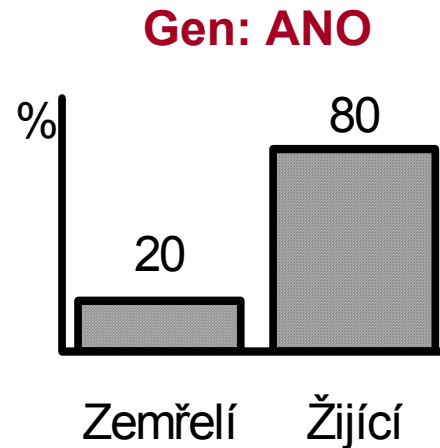
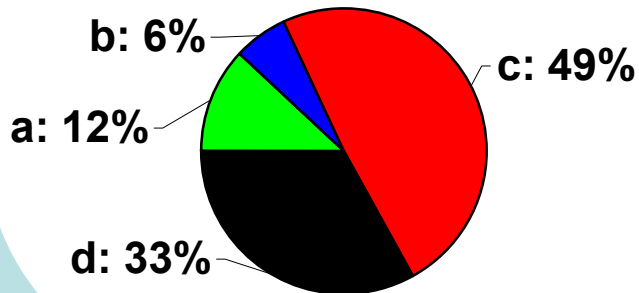
$$F_B = 102 * 136 / 166 = 83,57$$

$$F_C = 11,57$$

$$F_D = 52,43$$

$$\chi^2_{(1)} = \frac{(20-18,43)^2}{18,43} + \frac{(82-83,57)^2}{83,57} + \frac{(10-11,57)^2}{11,57} + \frac{(54-52,43)^2}{52,43} = 0,423 \quad 0,423 < \chi^2_{0,95}^{(1)} = 3,84$$

### Kontingenční tabulka v obrázku



# R x C kontingenční tabulka

Výběr: N lidí ze sociologického průzkumu (delikventi)

Jev **A**: Původ z rozvrácených rodin

Jev **B**: Stupeň zločinnosti I < II < III < IV

A \ B	I.	II.	III.	IV.	$\Sigma$
ANO	a	b	c	d	číslo 1
NE	e	f	g	h	
$\Sigma$	číslo2				

Stupně volnosti:

$$(R-1) * (C-1) = 1 * 3 = 3$$

$$F_a = \frac{\text{číslo 1} \cdot \text{číslo 2}}{N}$$

Tabulky:  $\chi^2_{(1-\alpha)}^{(v)}$

## Očekávané četnosti:

$$p_a = \frac{a}{a+e}$$

$$p_b = \frac{b}{b+f}$$

$$p_c = \frac{c}{c+g}$$

$$p_d = \frac{d}{d+h}$$

# Kontingenční tabulky

## Příklad 1



Ověřte na datech z pokusu se 100 květinami určitého druhu, že barva květů se geneticky štěpí v poměru žlutá : červená = 3 : 1.



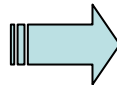
$H_0$ : Pozorovaná frekvence pro jednotlivé barvy květů jsou vzorkem populace mající poměr mezi žlutými a červenými květy 3 : 1.

Součet frekvencí u obou barev květů ( $f_i$ ) se rovná 100 a pozorované frekvence u kategorií barvy budou srovnány s očekávanými frekvencemi (uvedeny v závorkách):

	Kategorie barvy		n
	Žlutá	Červená	
$f_{\text{poz.}}$	84	16	100
$f_{\text{oček.}}$	75	25	

$$\chi^2 = \sum \frac{(f_{\text{poz.}} - f_{\text{oč.}})^2}{f_{\text{oč.}}} = \frac{(84 - 75)^2}{75} + \frac{(16 - 25)^2}{25} = 4,320$$

**St. volnosti =  $n = k - 1 = 1$**



**Zamítáme hypotézu shody srovnávaných četností**

Při testování  $H_0$  jsme použili matematický zápis ( $0,025 < P < 0,05$ ). Z tabulek  $\chi^2$  rozložení vidíme, že pravděpodobnost překročení hranice 2,706 je 0,1 (10 %), což může být stručně zapsáno jako  $P(\chi^2 \geq 2,706) = 0,10$ .

Dále lze zjistit pro  $P(\chi^2 \geq 3,841) = 0,05$ . V řešené úloze jsme dospěli k hodnotě testové statistiky  $\chi^2 = 4,320$ . Pro tento případ lze tedy psát  $0,025 < P(\chi^2 \geq 4,320) < 0,05$ ; a jednodušeji  $0,025 < P < 0,05$ . Jde v podstatě o přibližné určení hranic chyby 1. druhu.



# Kontingenční tabulky

## Příklad 2

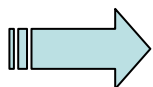
Tento příklad je rozšířením problému z příkladu 1 na srovnání pozorovaných a očekávaných frekvencí pro více kategorií sledovaného znaku:

✓ Celkem bylo zkoumáno 250 semen určitého druhu rostliny a roztríděno do následujících kategorií: žluté/hladké; žluté/vrásčité; zelené/hladké; zelené/vrásčité. Předpokládaný poměr výskytu těchto kategorií v populaci je 9 : 3 : 3 : 1. Následující tabulka obsahuje původní data z pozorování a dále postup při testování  $H_0$ .

	žluté/hladké	žluté/vrásčité	zelené/hladké	zelené/vrásčité	n
f <sub>poz.</sub>	152	39	53	6	250
f <sub>oček.</sub>	140,6250	46,8750	46,8750	15,6250	

$$\nu = k - 1 = 3$$

$$\chi^2 = \frac{11,3750^2}{140,6250} + \frac{7,8750^2}{46,8750} + \frac{6,1250^2}{46,8750} + \frac{9,6250^2}{15,6250} = 8,972$$



**Zamítáme hypotézu shody pozorovaných četností s očekávanými**

# Testy dobré shody - příklad

## Příklad 3

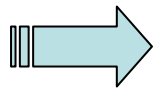
Složitější příklady řešené srovnáváním frekvencí je možné rozdělit na testování dílčích hypotéz:

- ✓ Předpokládejme, že chceme pro data z předchozí úlohy testovat hypotézu existence štěpného poměru 9 : 3 : 3 pro první tři kategorie semen:

	žluté/hladké	žluté/vrásčité	zelené/hladké	n
<b>f<sub>poz.</sub></b>	152	39	53	244
<b>f<sub>oček.</sub></b>	146,400	48,800	48,800	

$n = k - 1 = 2$

$$\chi^2 = \frac{5,600^2}{146,40} + \frac{9,800^2}{48,80} + \frac{4,200^2}{48,80} = 2,544$$



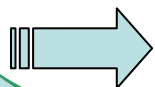
**Nezamítáme hypotézu shody pozorovaných četností s očekávanými.**

- ✓ Nyní otestujeme hypotézu štěpného poměru kategorií zelené/vrásčité:ostatní typy = 1:15

	zelené/vrásčité	ostatní	n
<b>f<sub>poz.</sub></b>	6	244	25
<b>f<sub>oček.</sub></b>	15,625	234,375	

$n = k - 1 = 1$

$$\chi^2 = \frac{9,625^2}{15,625} + \frac{9,625^2}{234,375} = 6,324$$



**Zamítáme hypotézu shody pozorovaných četností s očekávanými.**

# Test dobré shody pro více kategorií – využití aditivity testu

## Příklad

✓ U 193 párů dvojčat byly zjištěny následující poměry pohlaví:  $56 \text{ Ch} - \text{Ch}$   
 $72 \text{ Ch} - \text{H}$   
 $65 \text{ H} - \text{H}$

? *Za předpokladu, že narození chlapečka má stejnou pravděpodobnost jako narození holčičky, lze očekávat poměry pro výše uvedené skupiny = 0,25 : 0,5 : 0,25. Ověřte tento předpoklad na uvedeném vzorku populace.*

$\Sigma$  193 párů  $1/4 : 1/2 : 1/4$   
očekávané četnosti = 48,25 : 96,50 : 48,25

$$\chi_{(2)}^2 = 13,28$$

Proč lze v předchozím případě očekávat zamítnutí  $H_0$ ?

Testujte následující hypotézy:

1) Jsou relativní počty párů se shodným pohlavím ve shodě s očekávanými četnostmi? (ignorujte Ch – H páry)

2) Je relativní četnost kombinace Ch - Ch a H - H párů oproti párům s rozdílným pohlavím ve shodě s očekávanými četnostmi?

$\Sigma$  121 párů  $1 : 1$   
očekávané četnosti = 60,5 : 60,5

$$\chi_{(1)}^2 = 0,669$$

$$\frac{H - H}{Ch - Ch}$$

$\Sigma$  193 párů  $1 : 1$   
očekávané četnosti = 96,5 : 96,5

$$\chi_{(1)}^2 = 12,44$$

# Test dobré shody - příklad

Města - zatížení exhalacemi - třídy (A > B > C > D)

Svět: A : B : C : D = 2 : 3 : 6 : 4

Konkrétní země (n = 184 měst): A : B : C : D = 32 : 151 : 182 : 116

$H_0$ : shoda  $f_i$  a  $F_i$      $\alpha = 0,05$

$F_A$ : 64,13

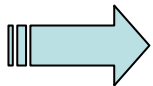
$F_C$ : 192,39

$F_B$ : 96,19

$F_D$ : 128,27

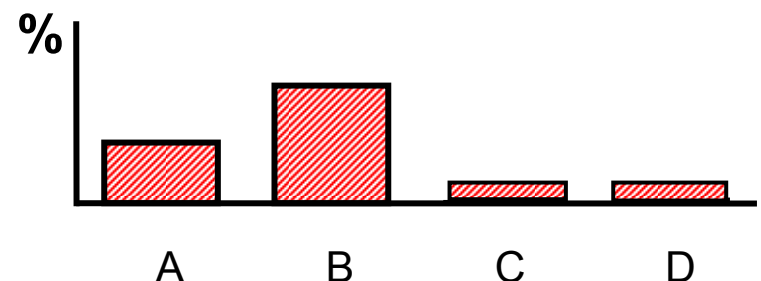
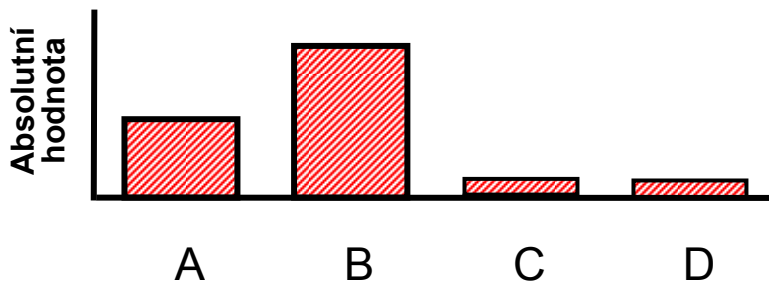
$$\chi^2_{(3)} = \frac{(32 - 64,13)^2}{64,13} + \dots + \frac{(116 - 128,27)^2}{128,27} = \underline{\underline{49,06}}$$

Tabulky :  $\chi^2_{1-\alpha}^{(v)} = \chi^2_{0,95}^{(3)} = 7,81$



**Zamítáme hypotézu shody pozorovaných četností s očekávanými.**

**Příspěvek kategorií A, B, C, D k celkové hodnotě  $\chi^2$**



# Test homogenity více binomických rozložení



Jev: Úmrtnost na leukemii

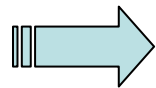
Předpoklad:  $\Pi = 0,6$

Absolutní četnost jevu označena  $r_i$

*Sledovalo s autorů z s zemí:*

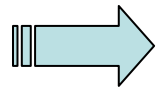
Autor	$n_i$	$r_i$	$p_i$
1			
2			
⋮			
⋮			
⋮			
s	$\sum n_i = N$		

$$\bar{p} = \frac{\sum p_i}{S}$$



Test homogenity binomických rozložení

$$\chi_{S-1}^2 = \frac{(\sum r_i p_i - \bar{p} \sum r_i)}{\bar{p} (1 - \bar{p})}$$



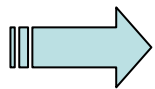
Po možném sloučení s výběrů

$$\chi_{(1)}^2 = \frac{\left( \left| \sum r_i - N \cdot \Pi \right| - \frac{1}{2} \right)^2}{N \cdot \Pi \cdot (1 - \Pi)}$$

Test shody reálného  $r$  ( $\sum r_i$ ) a  $n \cdot \Pi$

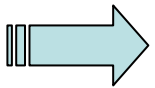
# Příklad analýzy homogenity binomických četností

Pomocí  $\chi^2$  rozložení lze rovněž posuzovat homogenitu většího množství nezávislých pokusů testujících tutéž hypotézu.



Bylo provedeno 6 nezávislých výběrů z populace mladých mužů, kteří v dětství onemocněli těžkým zánětem mozkových blan.

$H_0$ : V této populaci se vyskytují praváci a leváci v poměru 1 : 1.



Nalezněte v literatuře příslušné vztahy pro testování homogenity všech šesti výběrových populací a na základě výsledků tohoto testu rozhodněte o dalším postupu.

Následující tabulka obsahuje původní data a výsledek testování (v závorkách jsou uvedeny očekávané četnosti):

Vzorek	Praváci	Leváci	n	$\chi^2$	St. volnosti
1	3 (7)	11 (7)	14	4,5714	1
2	4 (8)	12 (8)	16	4,000	1
3	15 (10)	5 (10)	20	5,000	1
4	14 (9)	14 (9)	18	5,5556	1
5	13 (8,5)	4 (8,5)	17	4,7647	1
6	17 (11)	5 (11)	22	6,5455	1

$$\chi^2_{heterogenita} = 30,2036$$

$$\nu = s - 1 = 5$$

$$P < 0,001$$

Jednoduchým testováním lze zjistit, že všechny testy pro jednotlivé výběry jsou významné, což znamená, že ani v jednom případě nebyla potvrzena shoda očekávaných a pozorovaných četností. Test homogenity štěpného poměru v zkoumaných populacích rovněž vedl k zamítnutí možnosti sloučit jednotlivé výběry a posuzovat je jako celek (kromě testovaného poměru 1 : 1 neexistuje tedy v datech žádný jiný jednotný štěpný poměr mezi oběma vlastnostmi).

V případě, že by tento test neprokázal odchylky mezi jednotlivými výběrovými populacemi, bylo by možné jednotlivé odběry sloučit a posuzovat jako homogenní vzorek.

## $\chi^2$ test - příklad složitější kontingenční tabulky I.

*Caffeine consumption and marital status in antenatal patients (from Martin and Bracken, 1987)*

Marital status	Caffeine consumption (mg/day)				Total
	0	1 - 150	151 - 300	> 300	
Married	652	1537	598	242	3029
Divorced, separed or widowed	36	46	38	21	141
Single	218	327	106	67	718
Total	906	1910	742	330	3888

*Caffeine consumption and marital status data*

Marital status	Caffeine consumption (mg/day)				Total
	0	1 - 150	151 - 300	> 300	
Married	22 %	51 %	20 %	8 %	3029 (100 %)
Divorced, separed or widowed	26 %	33 %	27 %	15 %	141 (100 %)
Single	30 %	46 %	15 %	9 %	718 (100 %)
Total	23 %	49 %	19 %	8 %	3888 (100 %)

# $\chi^2$ test - příklad složitější kontingenční tabulky II.

## *Expected frequencies*

<b>Marital status</b>	<b>Caffeine consumption (mg/day)</b>				<b>Total</b>
	<b>0</b>	<b>1 - 150</b>	<b>151 - 300</b>	<b>&gt; 300</b>	
Married	705,8	1488	578,1	257,1	3029
Divorced, separated or widowed	32,9	69,3	26,9	12,0	141
Single	167,3	352,7	137	60,9	718
<b>Total</b>	<b>906</b>	<b>1910</b>	<b>742</b>	<b>330</b>	<b>3888</b>

## *Contributions of each cell*

<b>Marital status</b>	<b>Caffeine consumption (mg/day)</b>				<b>Total</b>
	<b>0</b>	<b>1 - 150</b>	<b>151 - 300</b>	<b>&gt; 300</b>	
Married	4,11	1,61	0,69	0,89	7,30
Divorced, separated or widowed	0,30	7,82	4,57	6,82	19,51
Single	15,36	1,88	7,02	0,60	24,86
<b>Total</b>	<b>19,77</b>	<b>11,31</b>	<b>12,28</b>	<b>8,31</b>	<b>51,66</b>



## $\chi^2$ test - příklad frakcionace složitější kontingenční tabulky I.

Cílem rozsáhlejšího průzkumu populace bylo prozkoumat vztah mezi dvěma typy chorob a krevními skupinami u lidí. Konkrétní data jsou uvedena v tabulce:

Krevní skupina	Žaludeční vředy	Rakovina žaludku	Kontrola	Celkem
0	983	383	2892	4258
A	679	416	2625	3720
B	134	84	570	788
<b>Celkem</b>	1796	883	6087	8766

? Vypočítejte testovou charakteristiku pro tuto kontingenční tabulku a otestujte nulovou hypotézu nezávislosti jevů ( $\chi^2 = 40,54$ ; 4 st. volnosti) ?

# $\chi^2$ test - příklad frakcionace složitější kontingenční tabulky II.

K podrobnějšímu průzkumu složitějších tabulek výrazně napomáhá přepis původní tabulky do podoby procentického zastoupení kategorií:

Krevní skupina	Žaludeční vředy	Rakovina žaludku	Kontrola
0	983	383	2892
A	679	416	2625
B	134	84	570
<b>Celkem</b>	1796	883	6087

Z této tabulky je patrné:

- 1.** Jsou jenom malé rozdíly v distribuci krevních skupin u kontroly a u skupiny nemocných rakovinou žaludku.
- 2.** Pacienti s vředy mají mnohem častěji krevní skupinu 0.

*Na základě těchto poznatků je možné sestavit menší kontingenční tabulku, která otestuje hypotézu o shodné distribuci krevních skupin pro nemocné rakovinou a pro zdravé lidi.*

**? Sestavte tuto tabulku a otestujte nulovou hypotézu. ?**  
( $\chi^2 = 5,64$  (2 st. v.), P je přibližně rovna 0,06)

## $\chi^2$ test - příklad frakcionace složitější kontingenční tabulky III.

- Z tohoto dílčího testu vyplývá možnost sloučení skupiny nemocných rakovinou a zdravých lidí neboť se vzhledem k distribuci krevních skupin chovají jako homogenní populace. Dalším logickým krokem v podrobné analýze je testování shody relativních četností výskytu krevních skupin A a B mezi kombinovaným vzorkem (sloučená skupina s rakovinou a kontrola) a mezi vzorkem lidí nemocných žaludečními vředy - tzn. nyní neuvažujeme krevní skupinu 0. Výsledkem tohoto testu je  $\chi^2 = 0,68$  (1 st. vol.);  $P > 0,7$ . Vzorky pro krevní skupiny A a B lze tedy sloučit do smíšeného vzorku A + B.
- Nyní otestujeme shodu relativních četností výskytu skupiny 0 oproti A + B, a to mezi kombinovanou populací (kontrola + nemocní rakovinou) a mezi vzorkem nemocných vředařů ( $\chi^2 = 34,29$ ; 1 st. vol.). Lze tedy shrnout, že vysoká hodnota původního  $\chi^2$  se 4 st. volnosti byla způsobena zvýšenou četností lidí s krevní skupinou 0 mezi nemocnými žaludečními vředy.

## $\chi^2$ test - příklad frakcionace složitější kontingenční tabulky IV.

Průběh hodnocení lze shrnout do tabulky:

Srovnání	St. volnosti	$\chi^2$
0, A, B skupina u pacientů s rakovinou ( <b>r</b> ) x kontrola ( <b>k</b> )	2	5,64
A, B skupina u pacientů s vředy x kombinovaný vzorek ( <b>r + k</b> )	1	0,68
0, A, B skupina u pacientů s <b>s</b> vředy x kombinovaný vzorek ( <b>r + k</b> )	1	34,29
<b>Celkem</b>	<b>4</b>	<b>40,61</b>

Celkový součet testových statistik  $\chi^2$  (40,61) odpovídá přibližně původní hodnotě  $\chi^2$  (40,54). Což platí i o stupních volnosti (4). Tato skutečnost potvrzuje, že jsme detailním rozbořem vyčerpali informační obsah původní kontingenční tabulky a kromě popsané závislosti (zvýšený výskyt krevní skupiny 0 u lidí s žaludečními vředy) jsou jednotlivé kategorie zkoumaných jevů zcela nezávislé.

# Kontingenční tabulka 2 x 2: Řešení při nedostatečné velikosti vzorku

Yates' corection

Fisher's exact test



$H_0$ : Nezávislost jevů

Test analyzuje všechny možné 2 x 2 tabulky, které dávají stejnou sumu řádků a sloupců jako tabulka zdrojová.

Algoritmus každé tabulce přiřazuje pravděpodobnost, že taková situace nastane, je-li  $H_0$  pravdivá.

*Spectacle wearing among juvenile delinquents and non-delinquents who failed a vision test (Weindling et al., 1986)*

	Juvenile delinquents	Non- delinquents	Total
Spectacle wearers	1	5	6
	8	2	10
Total	9	7	16

# Kontingenční tabulka 2 x 2: Řešení při nedostatečné velikosti vzorku

All tables of frequencies which have the same row and column totals

Probability associated with each set of frequencies

(I)	0	6
(II)	1	5
(III)	2	4
(IV)	3	3
	9	1
	8	2
	7	3
	6	4

(V)	4	2
(VI)	5	1
(VII)	6	0
	5	5
	4	6
	3	7

	a	b	c	d	P
( I )	0	6	9	1	0,00087
( II )	1	5	8	2	0,02360
( III )	2	4	7	3	0,15734
( IV )	3	3	6	4	0,36713
( V )	4	2	5	5	0,33042
( VI )	5	1	4	6	0,11014
( VII )	6	0	3	7	0,01049
	<b>Total</b>				0,99999

## 2 x 2 frekvenční tabulka pro párové uspořádání (Mc Nemar's test - matched variables)

Příklad: Srovnání 2 metod stanovení antigenu v krvi (antigen vždy přítomen)

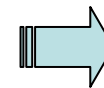


$H_0$ : metoda 1 = metoda 2

Metoda 1	Metoda 2	Frekvence
úspěch	úspěch	202
úspěch	neúspěch	60
neúspěch	úspěch	42
neúspěch	neúspěch	10

}  $\Sigma = 102$

$$\chi^2_{(c)} = \frac{(|60 - 42| - 1)^2}{102} = 2,83$$



$H_0$  nezamítnuta

Tabulky :  $\chi^2_{1-\alpha} (v=1) = 3,84$

# Aplikace analýzy 2 x 2 tabulky pro hodnocení rizika

## I. Prospektivní studie - odhad relativního rizika

Jedinci jsou sledováni prospektivně, zda se vyskytne nějaká vlastnost.

**VÝBĚR JE DÁN SLOUPCEM**

### OBEČNĚ

		Skupina 1	Skupina 2
Znak	ANO	a	b
	NE	c	d

$$\text{Riziko: } \frac{a}{(a+c)} \quad \frac{b}{(b+d)}$$

$$RR = \frac{\frac{a}{(a+c)}}{\frac{b}{(b+d)}}$$



$$H_0: RR = 1$$

### PŘÍKLAD

		Retardace plodu	
		Symetrická	Asymetrická
Agar skore > 7	ANO	2	33
	NE	14	58

$$2/16=0,13$$

$$33/91=0,36$$

$$RR = \frac{2/16}{33/91} = 0,345$$

Riziko u "symetrické skupiny" je asi 35 % rizika u asymetrické skupiny

$$SE (\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}$$

$$\text{IS: } \ln RR - Z_{1-\alpha/2} \cdot SE (\ln RR) \\ \ln RR + Z_{1-\alpha/2} \cdot SE (\ln RR)$$



# Aplikace analýzy 2 x 2 tabulky pro hodnocení rizika

## II. Retrospektivní studie - "ODDS RATIO"

Zcela zásadně odlišný přístup od retrospektivní studie  
**VÝBĚR JE DÁN VLASTNOSTÍ - ŘÁDKEM**

Není tedy možné analyzovat relativní riziko, protože přípravou řádků můžeme měnit velikost kontrol.

### OBECNĚ

		Skupina 1	Skupina 2
		a	b
Znak	ANO	a	b
	NE	c	d

odds      a/c      b/d

$$\text{Odds ratio} : \frac{a/c}{b/d}$$

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

### PŘÍKLAD

		Vady chrupu	
		ANO	NE
Plavání týdně	> 6h	32	118
	≥ 6h	17	127

$$OR = (32 / 17) / (118 / 127) = 2,026$$

$$\ln (OR) = 0,706$$

$$SE (\ln (OR)) = 0,326$$

# Srovnání dvou relativních četností u párově uspořádaného pokusu (pair - matched groups)

## Situace: Skupiny nejsou nezávislé

### OBEČNĚ

Výskyt jevu		Počet párů
Skupina 1	Skupina 2	
+	+	a
+	-	b
-	+	c
-	-	d

$$\left. \begin{aligned} p_1 &= (a+b)/n \\ p_2 &= (a+c)/n \end{aligned} \right\} ++ \text{ páry}$$

$$p_1 - p_2 = \frac{b-c}{n}$$

$$SE(p_1 - p_2) = \frac{1}{n} \cdot \sqrt{b+c - \frac{(b-c)^2}{n}}$$

$$Z = (b-c) / \sqrt{b+c}$$

### PŘÍKLAD

Potíže se spaním		Frekvence
Drogy	Kontrola	
+	+	4
+	-	3
-	+	9
-	-	16

$$p_D = 7/32 \quad p_K = 13/32$$

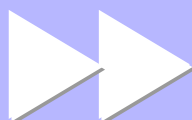
$$p_K - p_D = (13 - 7) / 32 = \underline{\underline{0,1875}}$$

$$SE(p_K - p_D) = \underline{\underline{0,113}}$$

$$Z = \frac{3-9}{\sqrt{3+9}} = \underline{\underline{-1,73}} \quad (p = 0,08)$$

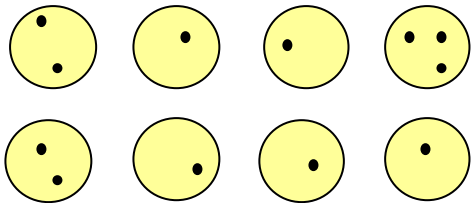


## 16. Poissonovo rozložení



# Poissonovo rozložení v přírodě existuje

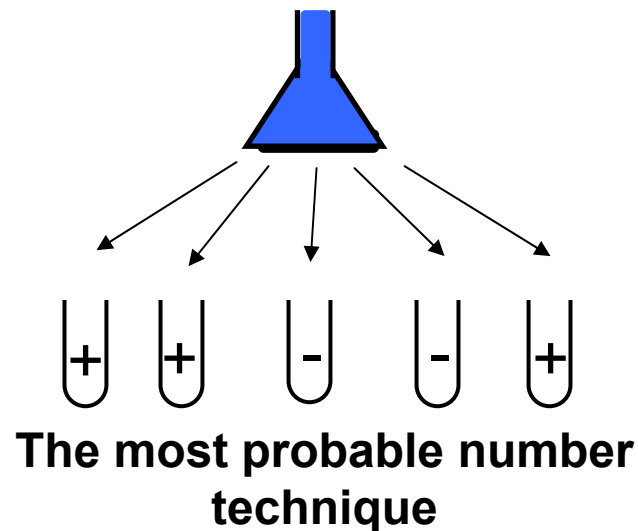
Mutace bakterií na inkubačních miskách



Výskyt jevu v prostoru  
(počet žížal na určité plochu pole)

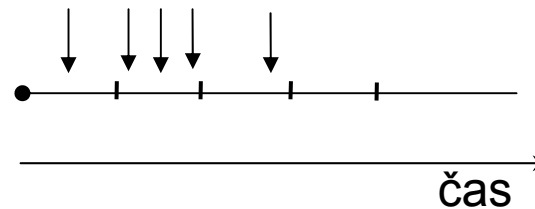
3	40	1	0
0		2	12
2	10	7	4

Orientační stanovení jevu  
(při produkci plynu bakteriemi)



Výskyt jevu v čase

(srdeční arytmie v určitých časových intervalech)




# Poissonovo rozložení


Celkový počet jevů v  $n$  nezávislých pokusech


$$\left. \begin{array}{l} E(x) = n p \\ D(x) = n p \end{array} \right\} E(x) = D(x)$$


$$P(r) = \frac{e^{-\mu} \cdot \mu^r}{r!} = e^{-\lambda} \cdot \frac{\lambda^r}{r!}$$


$\mu = \lambda =$  průměrný počet jevů z  $n$  pokusů


$$P(X = 0) = e^{-\mu}$$


$$P(X = 1) = e^{-\mu} \cdot \mu^1$$

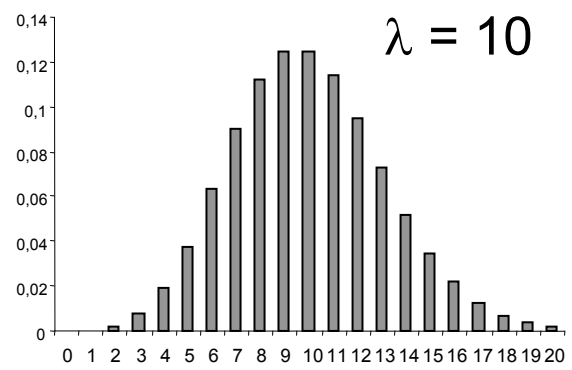
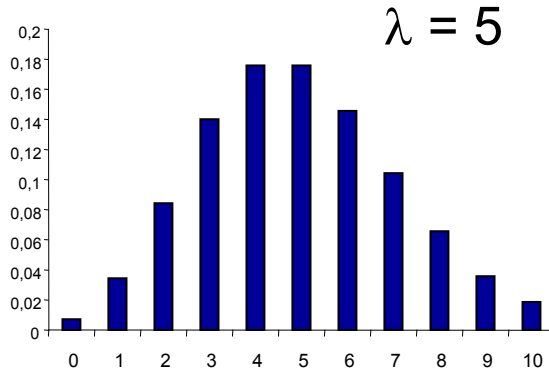
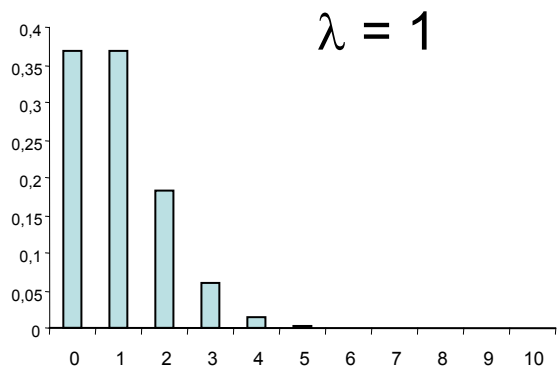
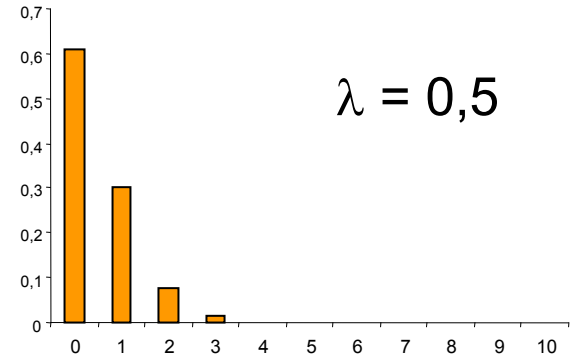
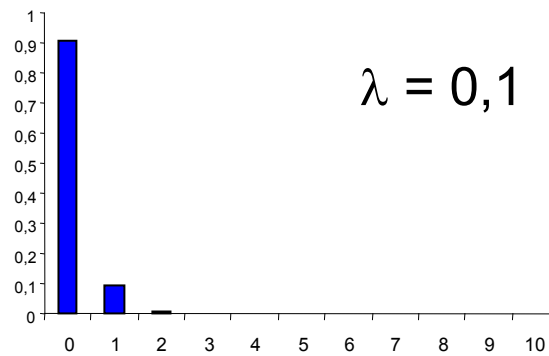
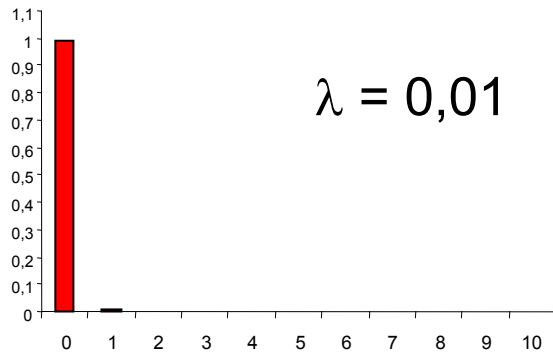

$$P(X = 2) = \frac{e^{-\mu} \cdot \mu^2}{2}$$


$$P(X = 3) = \frac{e^{-\mu} \cdot \mu^3}{(3)(2)}$$


$$P(X = 4) = \frac{e^{-\mu} \cdot \mu^4}{(4)(3)(2)}$$

# Poissonovo rozložení jako model

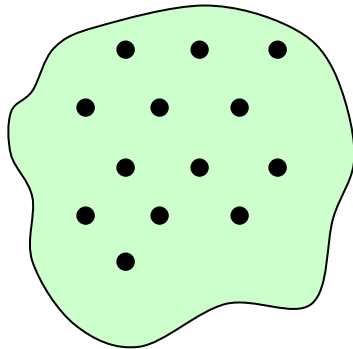
$$P(x = r) = e^{-\lambda} \cdot \frac{\lambda^r}{r!}$$



# Poissonovo rozložení jako model pro náhodný výskyt jevů

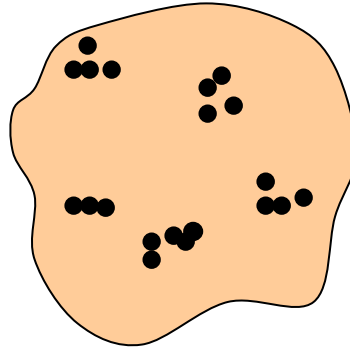
Předpoklad: náhodná distribuce jevu mezi studovanými objekty (příp. v čase, v prostoru).

$$\sigma^2 < \mu$$



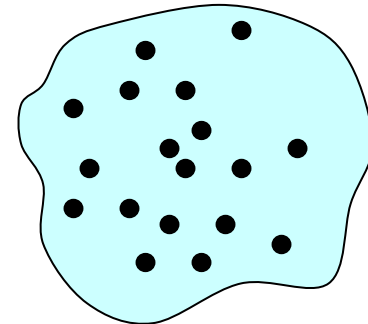
Uniform

$$\sigma^2 > \mu$$

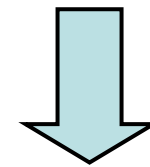


Clustered

$$\sigma^2 = \mu$$



Random



**Poisson**

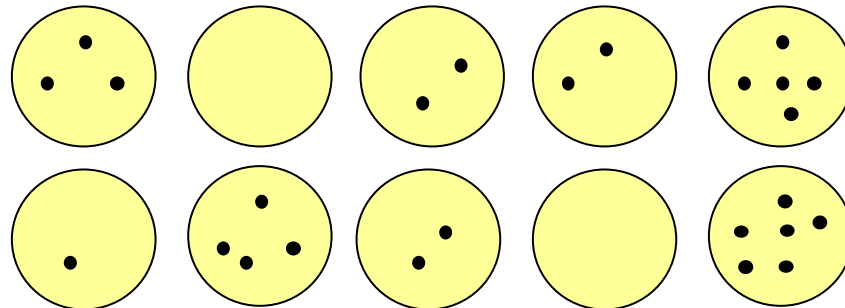
Pokud je  $\lambda$  spíše větší ( $\sim 5 - 10$ ), pak Poisson odpovídá spíše binomickému až normálnímu rozložení.

# Formální prezentace Poissonova rozložení

Př: pokus.....10 000 bakterií na misce  
n = 10 misek

Jev: mutace (r=25)

$\lambda$ .....průměrný počet mutantů na  
jednu misku



$$r=25$$

$$\bar{x} \approx \lambda = 25/10 = 2,5$$

95 % IS:

$$\bar{x} - Z_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}} \leq \lambda \leq \bar{x} + Z_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}}$$

$$2,5 - 1,96 \cdot \sqrt{0,25} \leq \lambda \leq 2,5 + 1,96 \cdot \sqrt{0,25}$$

$$1,52 \leq \lambda \leq 3,48$$



# Poissonova náhodná proměnná

Při měření počtu krvinek změněných určitou chorobou (relativně vzácné) je pozorován zředěný vzorek krve pod mikroskopem v komůrce rozdělené na stejně velká pole. Sledovaná veličina, udávající počet krvinek v  $i$ -tém poli může být považována za rozdělenou podle Poissonova rozložení:

$n = 169$  = počet nezávislých pozorování proměnné

$r = 10$  = počet pozorovaných krvinek

Jaká je hodnota parametru  $\lambda$  Poissonova rozložení a jaká je jeho interpretace ?

Jaký je interval 95% spolehlivosti pro parametr  $\lambda$  ?

Pokud bychom sledovali celkový počet červených krvinek (opět v  $n = 169$  nezávislých políčkách), bylo by i tuto proměnnou možno považovat za rozloženou podle Poissonova rozložení ? Uvažujte celkový počet pozorovaných krvinek jako 2013.

**Výpočet intervalu spolehlivosti pro  $\lambda$  (bez aproximace na normální rozložení)**

Spodní hranice IS

$$L_1 = \frac{\chi^2_{1-\alpha/2} (f_1 = 2r)}{2}$$

Horní hranice IS

$$L_2 = \frac{\chi^2_{\alpha/2} (f_2 = f_1 + 2)}{2}$$

# Poissonova náhodná proměnná

Konstantní zářič:  $n = 2608$  časových intervalů (každý 7,5 s)

$i$ : počet částic v intervalu ( $x$ )

$s_i$ : pozorovaná četnost intervalů  $s$  i částicemi

$$P(x = i) = \frac{\lambda^i \cdot e^{-\lambda}}{i!} \sim p_i$$

$i$	Počet intervalů s právě $i$ zaznamenanými částicemi $s_t$	teoretické četnosti $np_i$	$\frac{(s_i - np_i)^2}{np_i}$
0	57	54,399	0,1244
1	203	210,523	0,2688
2	383	407,361	1,4568
3	525	525,496	0,0005
4	532	508,418	1,0938
5	408	393,515	0,5332
6	273	253,817	1,4498
7	139	140,325	0,0125
8	45	67,882	7,7132
9	27	29,189	0,1642
10	10	17,075 (= $P\{\xi \geq 10\}$ )	0,0677
11	4		
12	2		
13	0		
	$n = 2608$	2608,00	12,8849

**Poissonova proměnná:**

\* Výborný model pro experimenty, v nichž je během časového průběhu zjišťován počet výskytu určitého jevu

# Aplikace Poissonova rozložení

Number of crimes per day in three areas of India during 1978 to 1982 (Thrakur and Sharma, 1984) showing observed frequencies (Obs) and expected frequencies using the Poisson distribution (Exp)

Number of crimes	Full moon days		New moon days	
	Obs	Exp	Obs	Exp
	40	45,2	114	112,8
	64	63,1	56	56,4
	56	44,3	11	14,1
	19	20,7	4	2,4
	1	7,1	0	0,3
	2	2,0	0	0
	0	0,5	0	0
	0	0,1	0	0
	0	0	0	0
	1	0	0	0
	183	183	186	186
		1,40		0,50
		1,16		0,75

Comparison of distributions of crimes on the new moon days (Thrakur and Sharma, 1984) and number of deaths in a Montreal hospital in 1971 (Zweig and Csank, 1978)

n	Crimes on new moon days in India		Deaths per day in Montreal hospital		Expected distribution Poisson (0,51)
	%	Frequency	%	Frequency	
0	61,3	114	60,3	220	60,0
1	30,1	56	31,0	113	30,6
2	5,9	11	6,3	23	7,8
3	2,2	4	2,2	8	1,3
4+	0,5	1	0,3	1	0,2
Total	100	186	100,0	365	99,9%
Mean		0,505		0,512	
SD		0,752		0,736	

# Poisson distribution: one - sample test

$$P_{(r)} = \frac{(e^{-\lambda} \cdot \lambda^r)}{r!}$$

Př: Počet hnízd křepelk na dané ploše

$$\left. \begin{array}{l} n = 8\,000 \quad \text{"pod lokalit"} \\ r = 28 \end{array} \right\} \hat{p} = 0,0035$$

Nechť je srovnávací soubor  
(předchozí průzkum)

$$p_o = 0,0020$$

$$\underline{p_o \cdot 8\,000 = 16 = \mu = \lambda}$$

$$H_o : p \leq p_o \sim \mu \leq 16 \quad ?$$

1) Vztít data jako pocházející z populace:

$$P(r = 28) = \frac{e^{-16} \cdot 16^{28}}{28!} = 0,00192 \dots\dots\dots$$

$$2) \left. \begin{array}{l} P(r \geq 28) = ? \\ [0,00411] \end{array} \right\} < 0,05 \Rightarrow H_o \text{ zamítnuta} \dots\dots\dots$$



**r = 28** je příliš velké pro populaci s  $p_o$



$\underline{p > p_o}$ , aby r = 28 bylo pravděpodobnější

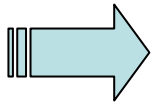
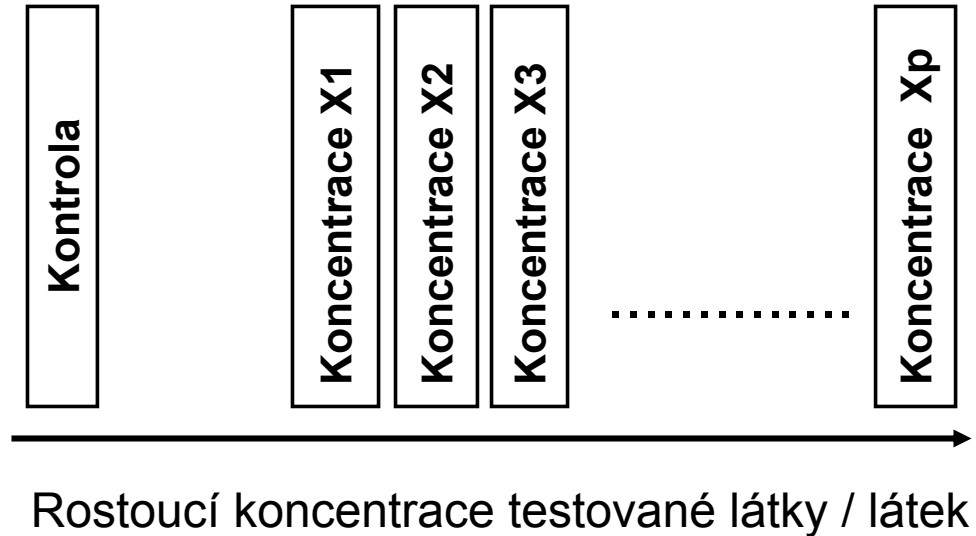


# 17. ANOVA – Analýza rozptylu

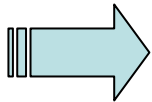


# Analýza rozptylu - ANOVA

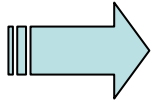
Základní technika  
sloužící  
k posouzení rozdílů  
mezi více úrovněmi  
pokusného zásahu



Celkově významné změny v reakci biologického systému



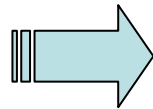
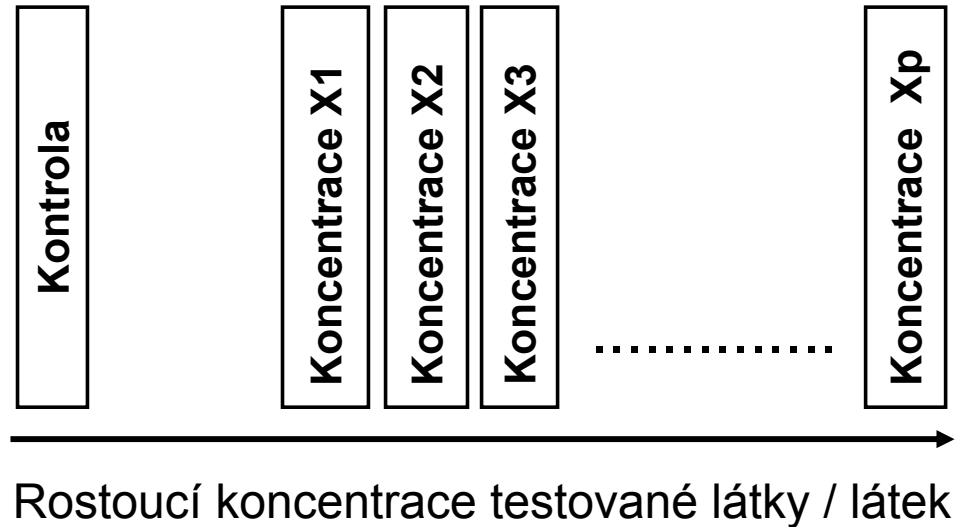
Vzájemné rozdíly účinku jednotlivých dávek



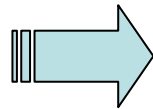
Rozdíly účinku dávek od kontroly

# Analýza rozptylu - ANOVA

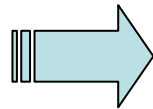
Významné kroky  
analýzy, vedoucí k  
efektivnímu srovnání  
variant



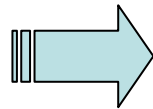
Splnění předpokladů analýzy  
Transformace dat



Relevantnost kontroly  
(vliv vlastní aplikace látek)



Vhodnost modelu ANOVA pro účely testu



Vlastní srovnání variant  
Minimalizace chyb při ověřování hypotéz

# Analýza rozptylu - ANOVA

***SPLNĚNÍ PŘEDPOKLADŮ ANOVA JE NEZBYTNOU PODMÍNKOU  
POUŽITÍ TÉTO TECHNIKY***

**ANOVA**  
**= parametrická**  
**analýza dat**

1. Předpoklad nezávislosti  
opakování experimentu

2. Homogenita  
rozptylu v rámci  
pokusných variant

3. Normalita rozložení  
v rámci pokusných  
variant

**ALTERNATIVOU JSOU NEPARAMETRICKÉ METODY**



# Analýza rozptylu - ANOVA

## *Předpoklady analýzy rozptylu jsou nezbytné pro dosažení síly testu*

• **Symetrické rozložení hodnot a normalita odchylek** od hodnoceného modelu ANOVA. Velkou část dat lze adekvátně normalizovat použitím logaritmické transformace. Předpoklad lognormální transformace může pochopitelně být teoreticky vyloučen u mnoha datových souborů obsahujících diskrétní parametry, kde je indikována vhodnost jiného typu transformace. U asymetricky rozložených a u diskrétních dat je nutné využít neparametrické alternativy analýzy rozptylu.

• **Statistická nezávislost reziduí** vyhodnocovaného modelu ANOVA. Pokud odhad a posouzení korelačních vztahů mezi pokusnými variantami není přímo předmětem výzkumu, lze jejich vliv na vyhodnocení odstranit znáhodněním dat v rámci pokusných variant - tedy změnou pořadí v náhodné. Rozsah vlivu těchto autokorelačních vztahů musí být ovšem primárně omezen správností experimentálního uspořádání.

• **Homogenita rozptylu** je nutným předpokladem pro smysluplnost vzájemných srovnání pokusných variant. U testů toxicity by splnění tohoto předpokladu mělo být ověřováno (Bartlettův test), neboť vážné rozdíly (až řádové) v jednotkách testovaného parametru mohou nastat v důsledku inhibice dávkami látky. Nehomogenita rozptylu je často ve vztahu k nenormalitě (asymetrii) dat a lze ji odstranit vhodnou normalizující transformací.

• **Aditivita** jako předpoklad týkající se složitějších experimentálních uspořádání. Exaktní otestování aditivity více pokusných faktorů je procedura poměrně náročná na experimentální design vyvážený co do počtu opakování. Je rovněž obtížné testovat interakci na nestandardních datech, neboť případná transformace může změnit charakter odchylek původních dat od hodnoceného modelu ANOVA.

# Analýza rozptylu - ANOVA

## *Omezení aplikace ANOVA lze řešit*

• **Chybějící data.** Vážným problémem jsou chybějící údaje o celé skupině kombinací testovaných látek, například u faktoriálních pokusů, kdy je znemožněno hodnocení experimentu jako celku.

• **Různé počty opakování** Jde o typický jev pro experimentální datové soubory. Při různých počtech opakování v experimentálních variantách jsou testy ANOVA citlivější na nenormalitu dat. Pokud jsou počty opakování zcela odlišné (až na řádové rozdíly), je nutno použít neparametrické techniky nebo analýzu rozptylu nevyvážených pokusů.

• **Odlehlé hodnoty.** Ojedinelé odlehlé hodnoty musí být před parametrickou analýzou rozptylu vyloučeny.

• **Nedostatek nezávislosti mezi rezidui modelu.** Jde o závažný nedostatek, zkreslující výsledek F-testu. Velmi často je tato skutečnost důsledkem špatného provedení nebo naplánování experimentu.

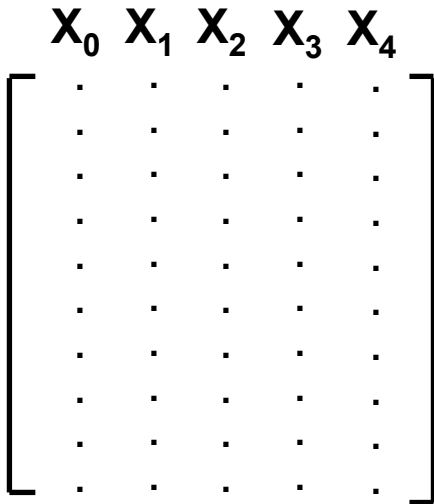
• **Nehomogenita rozptylu.** Velmi častý nedostatek experimentálních dat, často související s nenormalitou rozložení nebo s odlehlými hodnotami.

• **Nenormalita dat.** I v tomto případě lze situaci upravit vyloučením odlehlých hodnot nebo normalizující transformací.

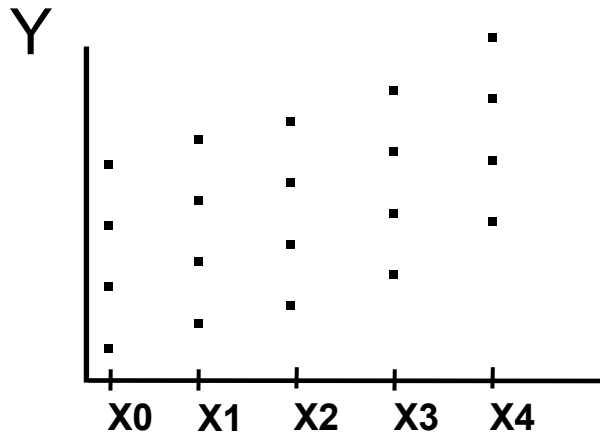
• **Neadditivita kombinovaného vlivu více pokusných zásahů.** Tuto situaci lze testovat jednak speciálními testy aditivity nebo přímo F testem kontrolujícím významnost vlivu interakce pokusných zásahů. Při významné interakci je nutné prozkoumat především její charakter ve vhodném experimentálním uspořádání.

# Modely analýzy rozptylu

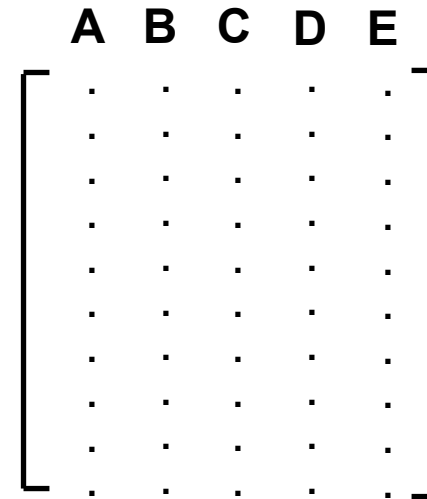
## Model I. Pevný model



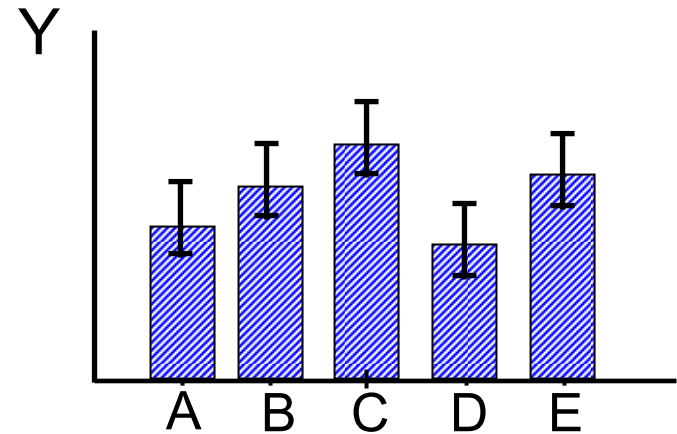
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$



## Model II. Náhodný model



$$y_{ij} = \mu + A_i + \varepsilon_{ij}$$



# ANOVA – základní výpočet

- Základním principem ANOVY je porovnání rozptylu připadajícího na:
  - Rozdělení dat do skupin (tzv. effect, variance between groups)
  - Variabilitu objektů uvnitř skupin (tzv. error, variance within groups), předpokládá se, že jde o náhodnou variabilitu (=error)

## 1. Variabilita mezi skupinami

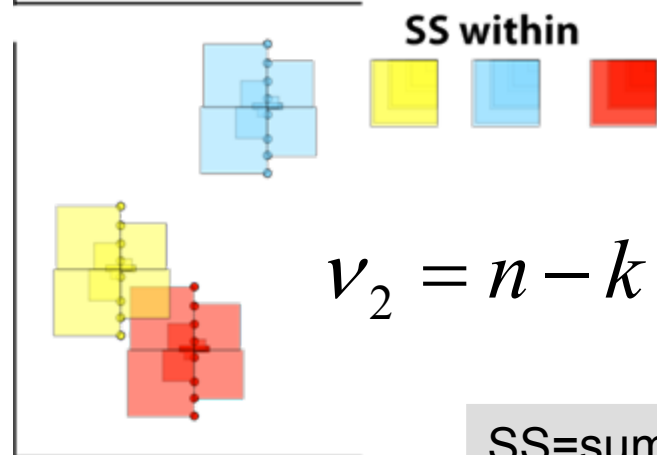
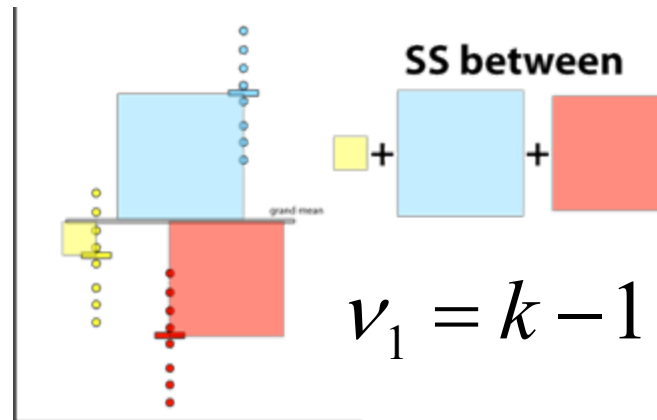
Rozptyl je počítán pro celkový průměr (tzv. grand mean) a průměry v jednotlivých skupinách dat

Stupně volnosti jsou odvozeny od počtu skupin (= počet skupin -1)

## 2. Variabilita uvnitř skupin

Rozptyl je počítán pro průměry jednotlivých skupin a objekty uvnitř příslušných, celková variabilita je pak sečtena pro všechny skupiny

Stupně volnosti jsou odvozeny od počtu hodnot (= počet hodnot - počet skupin)



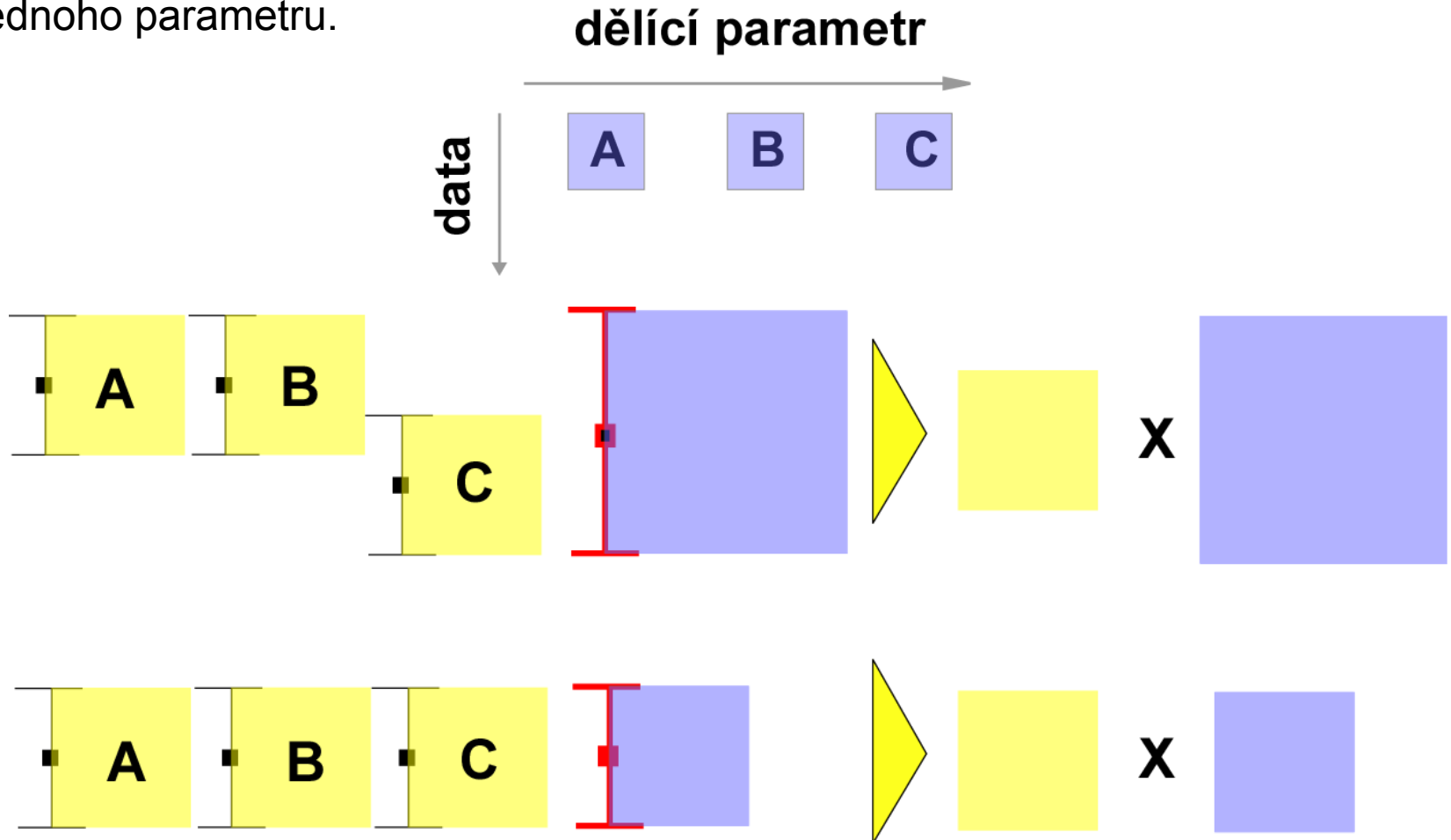
$$F = \frac{\text{between\_groups}}{\text{within\_groups}}$$

Výsledný poměr (F) porovnáme s tabulkami F rozložení pro  $v_1$  a  $v_2$  stupňů volnosti

SS=sum of squares

# Jednoduchý ANOVA design

Nejjednodušším případem ANOVA designu je rozdělení na skupiny podle jednoho parametru.



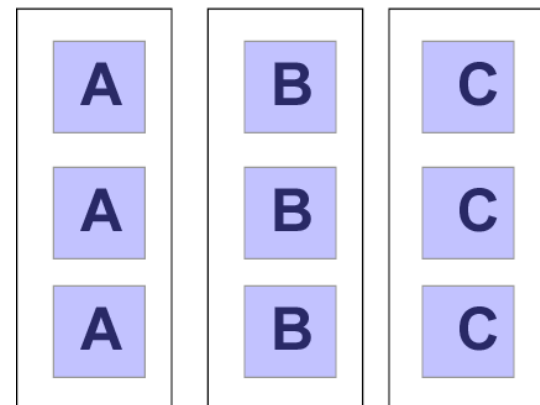
# Nested ANOVA

- Rozdělení skupin na náhodné podskupiny (např. opakování experimentu)
- Cílem je zjistit, zda data v jedné skupině nejsou pouhou náhodou
- Nejprve je testována shoda podskupin v hlavních skupinách,
  - pokud jsou shodné, je vše v pořádku
  - pokud nejsou, stále lze zjišťovat, zda se variabilita uvnitř hlavních skupin liší od celkové variability

## jednoduchá ANOVA



## nested ANOVA

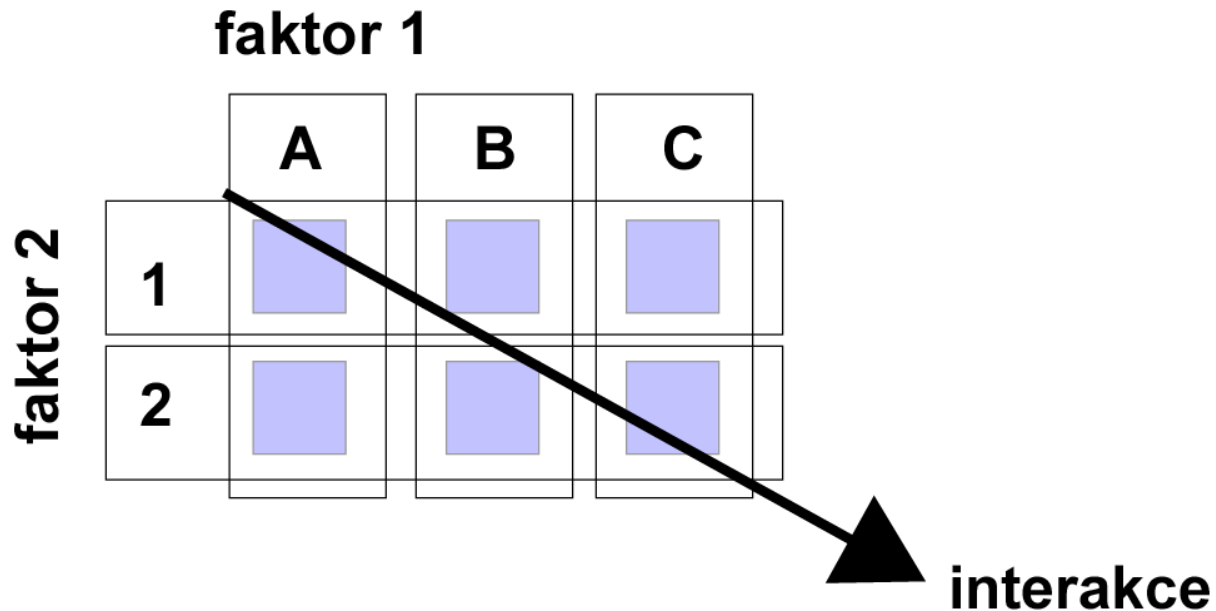


# Two way ANOVA

Pro rozdělení do kategorií je zde více parametrů

Na rozdíl od nested ANOVY nejde o náhodná opakování experimentu, ale o řízené zásahy (např. vliv pH a koncentrace  $O_2$ )

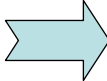
Kromě vlivu hlavních faktorů se uplatňuje i jejich interakce



# Modely analýzy rozptylu - základní výstup

*Základním výstupem analýzy rozptylu je  
Tabulka ANOVA - frakcionace komponent rozptylu*

Zdroj rozptylu	St. v.	SS	MS	F
Pok. zásah (mezi skupinami)	$a - 1$	$SS_B$	$SS_B / (a - 1)$	$MS_B / MS_E$
Uvnitř skupin	$N - a$	$SS_E$	$SS_E / (N - a)$	
Celkem	$N - 1$	$SS_T$		

$SS_B / SS_T$   Kvantifikovaný podíl rozdílu mezi pokusnými zásahy na celkovém rozptylu

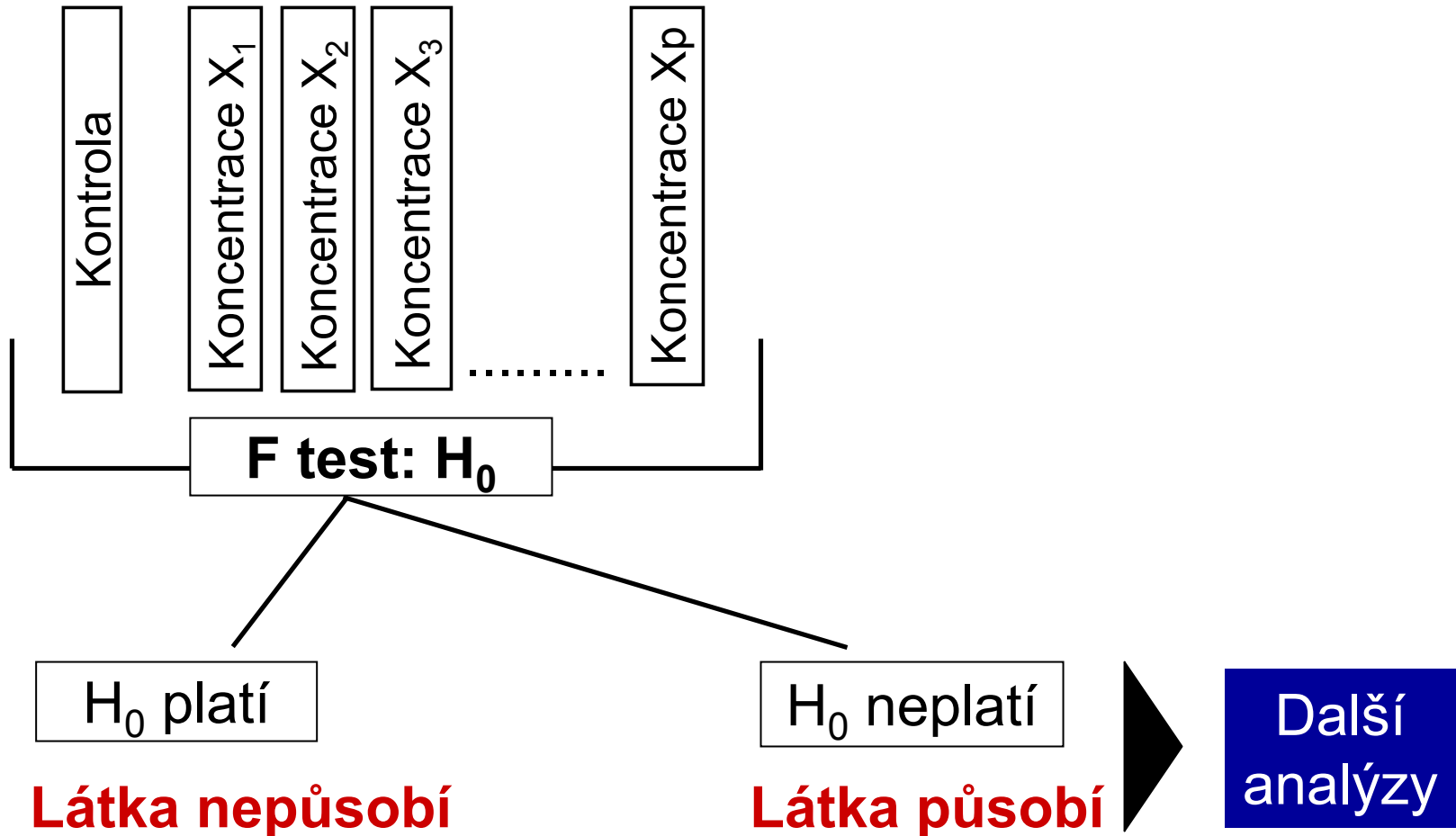
$MS_B / MS_T$   Statistická významnost rozdílu



# Analýza rozptylu - obecný F test

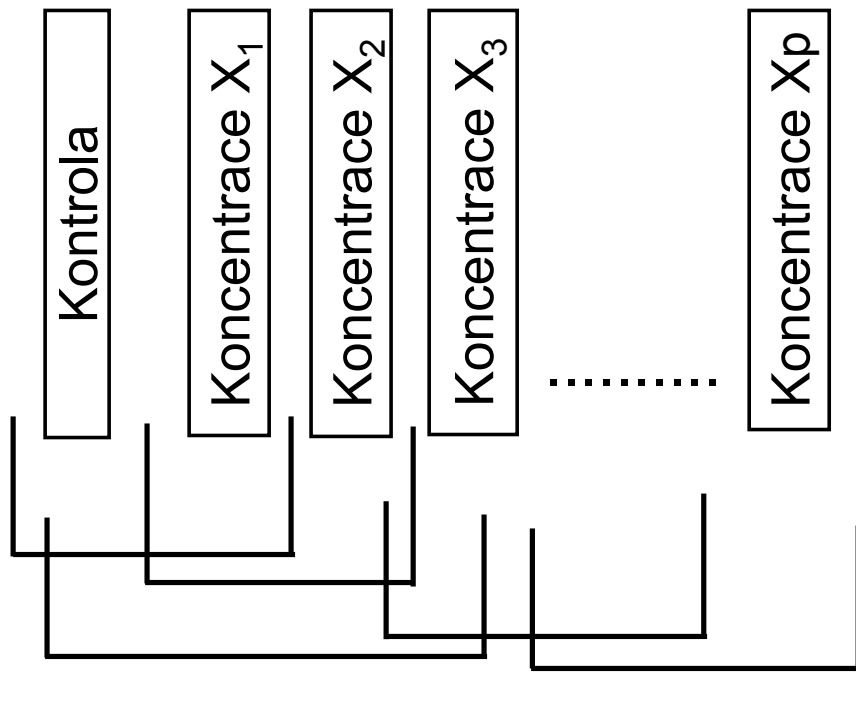
obecný F test

$$H_0: m_1 = m_2 = m_3 = \dots = m_p$$



# Analýza rozptylu - Testy kontrastů

ANOVA:  $H_0$  zamítnuta  
Testy kontrastů



Plánované

Neplánované

Pro srovnání variant s kontrolou

Testování kontrastů  
"Multiple range testy"

Parametrické

Neparametrické

# Příklad: Anova - One way

Dávka rostlinného stimulantu (0, 4, 8, 12 mg/l)

A = 4 ; n = 8

## I. ANOVA

Bartlett's test: P = 0,9847

K-S test: P = 0,482 - 0,6525 pro jednotlivé kategorie

Source	D. f.	SS	MS	F
Between Groups	3	305,8	101,9	8,56
Within Groups	28	322,2	11,9	
Total (corr.)	31	638,0		

## II. Multiple Range Test

NKS -test

Level	Average	Homogenous Groups
0	34,8	x
4	41,4	x
12	41,8	x
8	52,6	x

# Příklad: Anova - One way

I. Zásah: 4 klinická stadia virové choroby (napadá kr. buňky)

*Sledovaná veličina: aktivita enzymu v těchto krevních buňkách*

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

n = 3  
MODEL = ?

	I	II	III	IV
	22,8	16,4	11,2	14,2
	19,4	17,8	18,2	10,1
	12,5	19,1	15,8	12,8
$\Sigma$	65,7	53,3	45,2	37,1
průměr	21,9	17,8	15,1	12,4

II.

Source	D.f.	MS	F	P
Between groups	3	49,6	8,39	0,0075
Within groups	8	5,9		
Total (corr.)	11	-		

III. Komponenta rozptylu:

$$\sigma_A^2 \sim S_A^2 = \frac{MS_A - MS_e}{n} = \frac{49,6 - 5,9}{3} = 14,57$$

$$S_A^2 = 2,5 \cdot S_e^2$$

IV.

$$\rho_I \sim r_I = \frac{S_A^2}{S_A^2 + S_e^2} = 0,7142$$

# Srovnání variant v testech

## *Srovnávání variant po celkovém testu ANOVA*



Mnoho existujících algoritmů není vhodných pro konkrétní případ

Day and Quin  
Ecological Monographs, 1989

Test	Využití	Poznámka
Dunnett Williams	Srovnání s kontrolou	Ex. i modifikace pro různá n.
ANOVA testy (F)	Orthogonální kontrasty	Plánovaná srovnání
Ryan Q test	Jednoduché kontrasty	Vyhodnocen jako nejlepší test

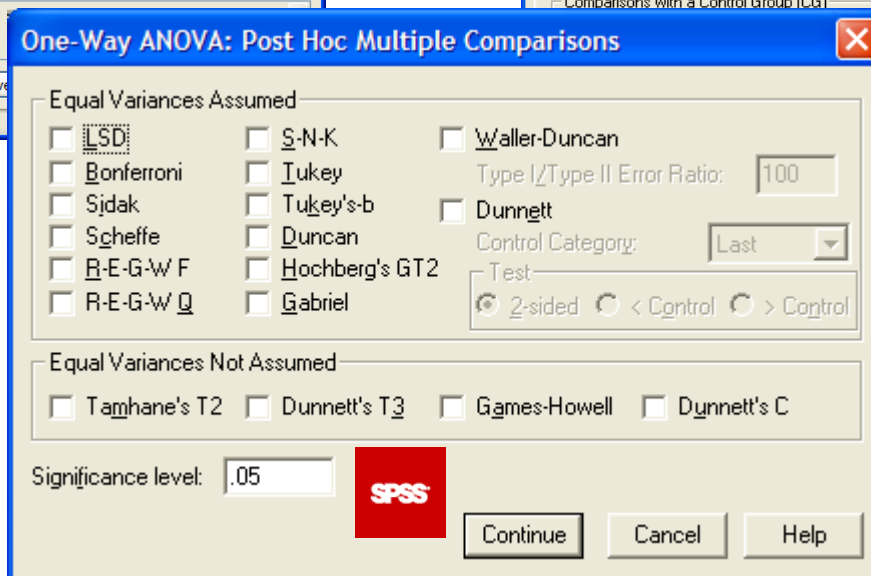
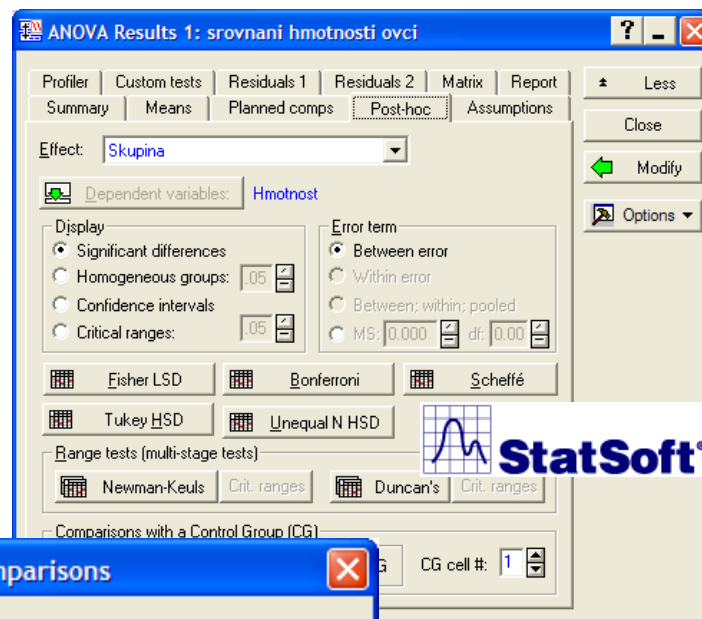
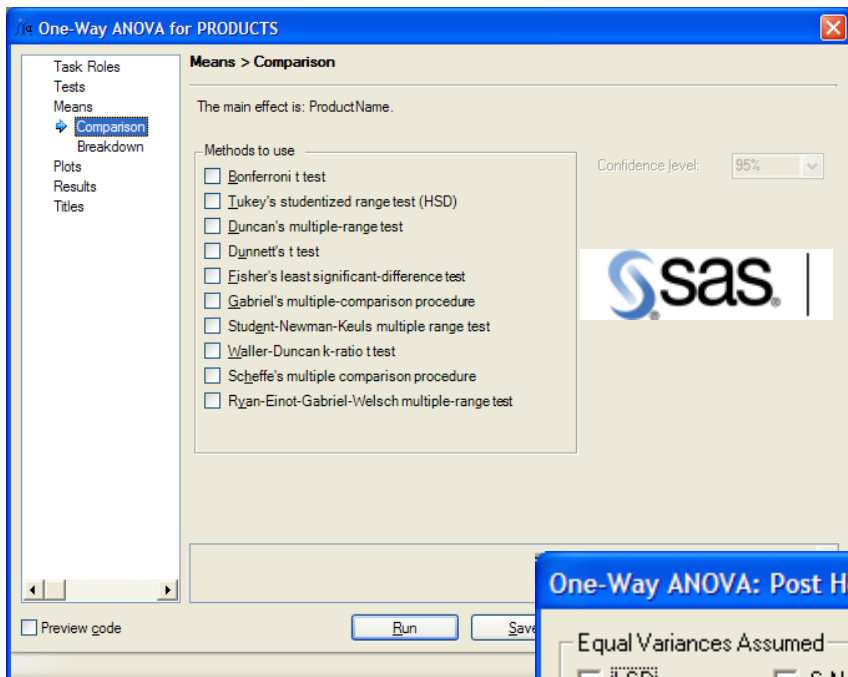
### Testy pro jednoduché kontrasty

Scheffe	Tukey	LSD
Bonferroni	Dunn-Sidák	Kramer

### Testy nevhodné

Duncan	Student - Newmann-Keuls	Waller-Duncan k ratio
--------	-------------------------	-----------------------

# Řada post-hoc testů v různých SW



# Hypotetické příklady - Multiple Range Tests

	<u>Level</u>	<u>Homogenous Group</u>
15	1	x
18	2	xx
22	3	xx
26	4	x
38	5	x

	<u>Level</u>	<u>Homogenous Group</u>
15	1	x
22	2	x
24	3	xx
29	4	x
30	5	x

	<u>Level</u>	<u>Homogenous Group</u>
15	1	x
18	2	xx
22	3	x
29	4	x
36	5	x



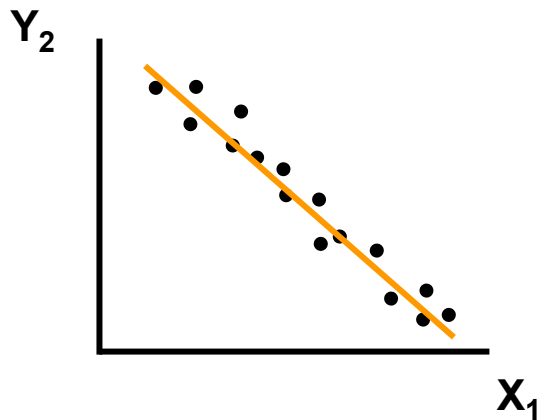
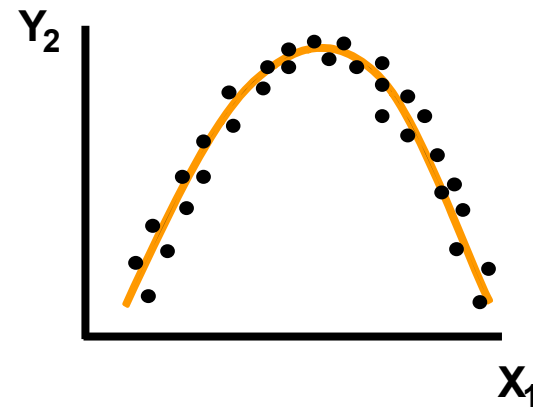
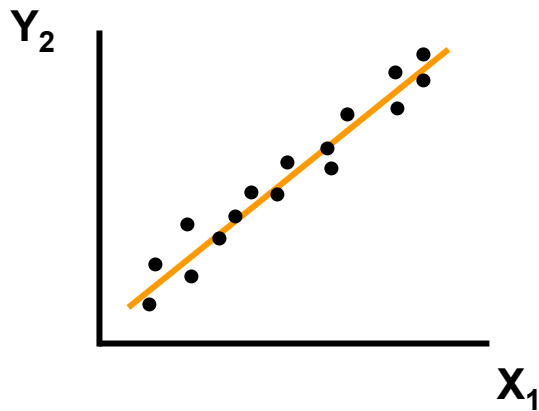
## 18. Korelace a regrese





# Základy korelační analýzy - I.

## Korelace - vztah (závislost) dvou znaků (parametrů)



$x_2 \backslash x_1$	<b>ANO</b>	<b>NE</b>
<b>ANO</b>	a	b
<b>NE</b>	c	d

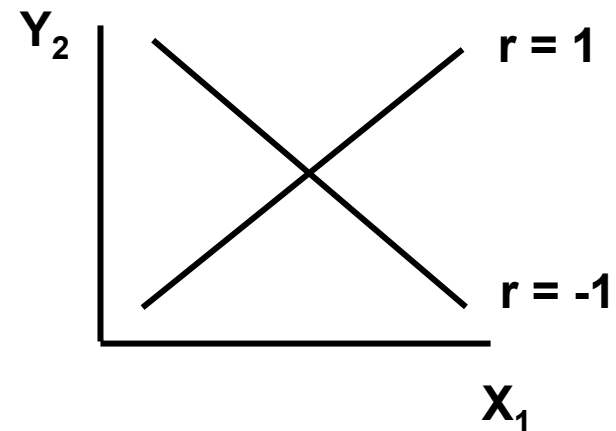
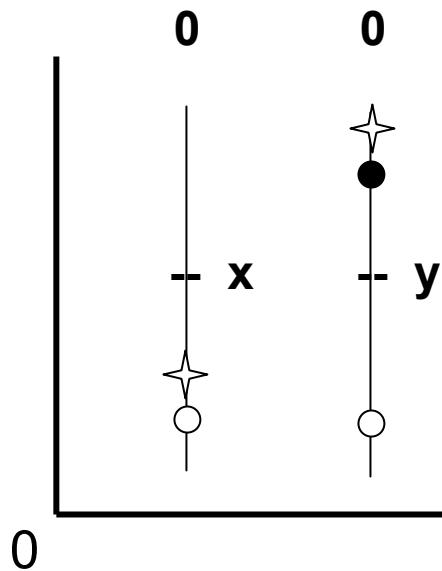
# Základy korelační analýzy - II.

## Parametrické míry korelace

Kovariance

$$\text{Cov}(x, y) = E(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Pearsonův  
koeficient korelace



## Základy korelační analýzy - III.

<b>P<sub>I</sub> (zem)</b>	10	14	15	32	40	20	16	50
<b>P<sub>I</sub> (rostl.)</b>	19	22	26	41	35	32	25	40

$$I = 1, \dots, n; n = 8; v = 6$$

$$r = \frac{\text{Cov}(x, y)}{S_x \cdot S_y} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sqrt{\left[ \sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \right] \left[ \sum y_i^2 - \frac{1}{n} (\sum y_i)^2 \right]}} = 0,7176$$

I.  $H_0: \rho = \phi: \alpha = 0,05$

tab :  $r(v=6) = 0,7076$

II.  $H_0: \rho = \phi$

$$t = \left[ \frac{r}{\sqrt{1-r^2}} \right] \cdot \sqrt{n-2} \quad v = n-2$$

$$\left. \begin{array}{l} t = \frac{0,7176}{0,6965} \cdot \sqrt{6} = 2,524 \\ \text{tab} : t_{0,975}^{(n-2)} = 2,447 \end{array} \right\} P \leq 0,05$$

# Základy korelační analýzy - IV.

## Srovnání dvou korelačních koeficientů (r)

1.  $n_1 = 1258$   
 $r_1 = 0,682$

2.  $n_2 = 462$   
 $r_2 = 0,402$

Krevní tlak x koncentrace kysl. radikálů

$$Z_i = 1.1513 \cdot \log \frac{(1 + r_i)}{(1 - r_i)}$$

$$Z_1 = 0,833$$

$$Z_2 = 0,426$$

**Test:**  $H_0: \rho_1 = \rho_2$  ;  $\alpha = 0,05$

$$Z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0,407}{0,0545} = 7,461$$

tabulky :  $Z_{0,975} = 1,96$

**7,461 >> 1,96 => P << 0,01**

# Základy korelační analýzy - V. Neparametrická korelace (rs)

<b>P<sub>I</sub> v půdě</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>6</b>	<b>7</b>	<b>5</b>	<b>4</b>	<b>8</b>
<b>P<sub>I</sub> v rostl.</b>	<b>1</b>	<b>2</b>	<b>4</b>	<b>8</b>	<b>6</b>	<b>5</b>	<b>3</b>	<b>7</b>
<b>d<sub>I</sub></b>	<b>0</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>-1</b>	<b>0</b>	<b>-1</b>	<b>-1</b>

$$i = 1, \dots, n; \quad n = 8 \Rightarrow v = 6$$

$$r_s = 1 - \frac{6 \cdot \sum di^2}{n(n^2 - 1)} = 0,9048$$

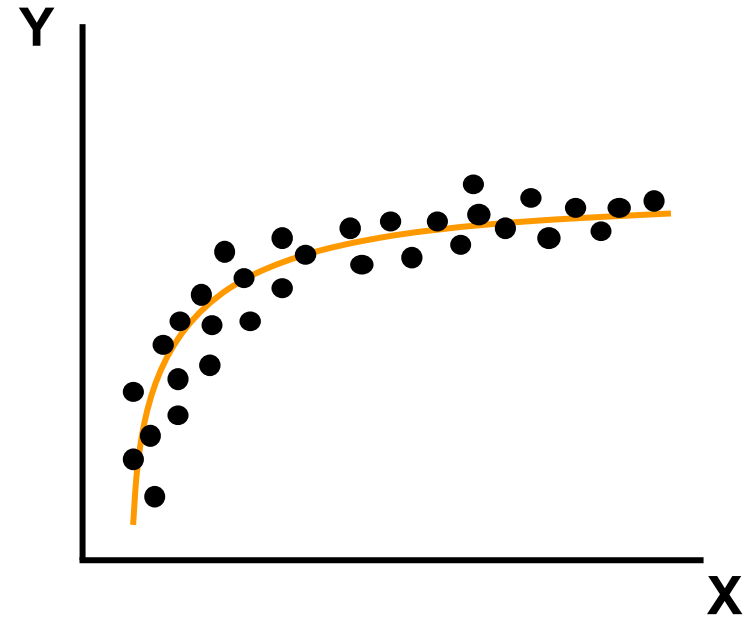
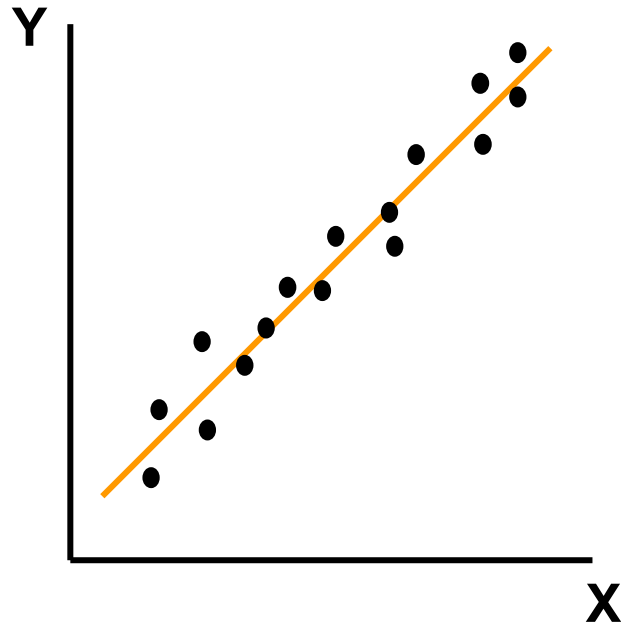
$$\text{tab : } r_s(v = 6) = 0,89$$

<b>Pacient č.</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
<b>Lékař 1</b>	<b>4</b>	<b>1</b>	<b>6</b>	<b>5</b>	<b>3</b>	<b>2</b>	<b>7</b>
<b>Lékař 2</b>	<b>4</b>	<b>2</b>	<b>5</b>	<b>6</b>	<b>1</b>	<b>3</b>	<b>7</b>
<b>d<sub>I</sub></b>	<b>0</b>	<b>6 - 8</b>	<b>1</b>	<b>-1</b>	<b>2</b>	<b>-1</b>	<b>0</b>

$$r_s = 1 - \frac{6 \cdot 8}{7(49 - 1)} = 0,857$$

**P = 0,358**

# Korelace v grafech I.



Vztahy velmi často implikují funkční vztah mezi Y a X.

$$Y = a + b \cdot X$$

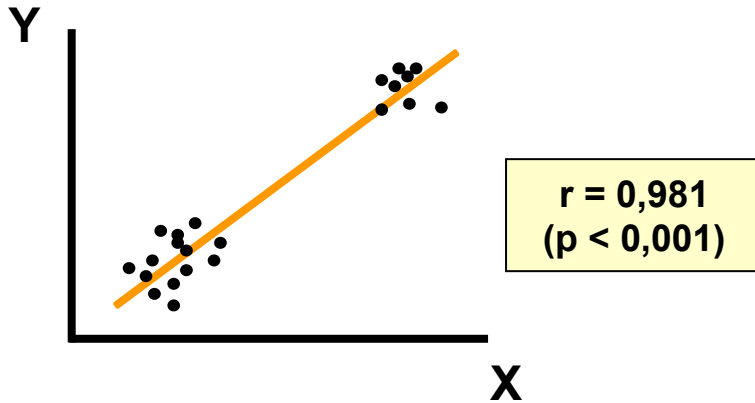
$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3$$

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2$$

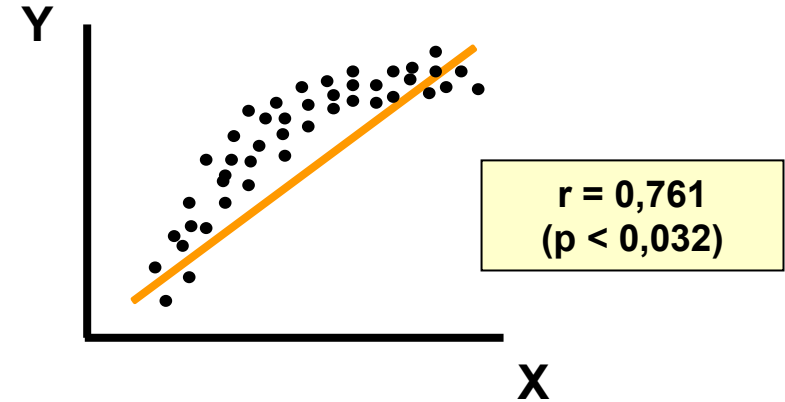
$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_1 \cdot X_2$$

# Korelace v grafech II.

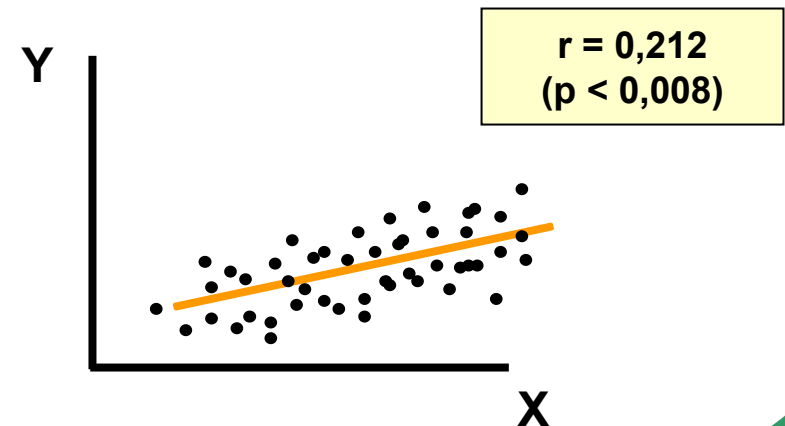
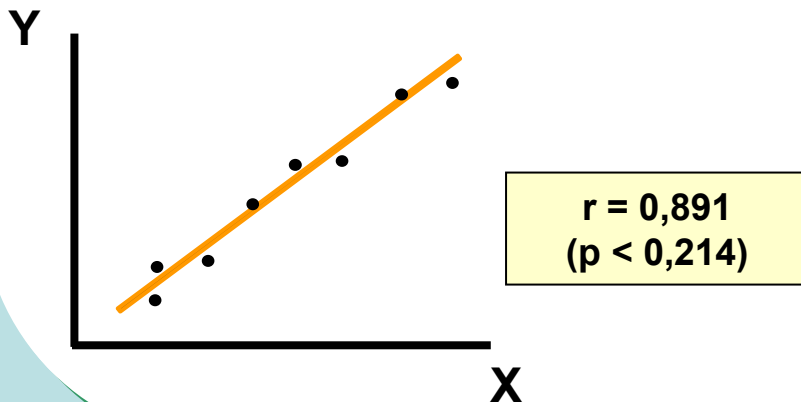
## Problém rozložení hodnot



## Problém typu modelu



## Problém velikosti vzorku



# Základy regresní analýzy

Regrese - funkční vztah dvou nebo více proměnných

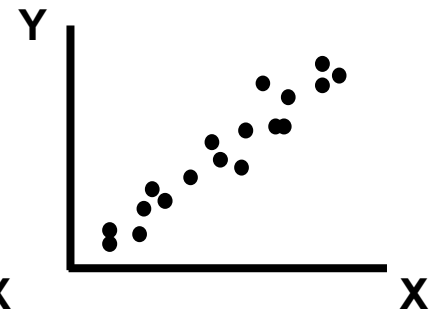
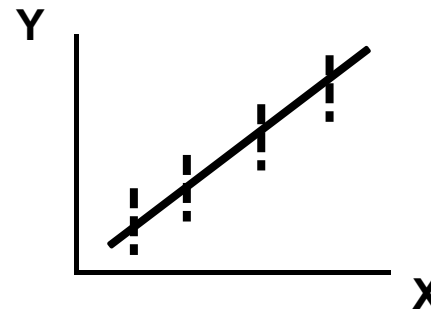
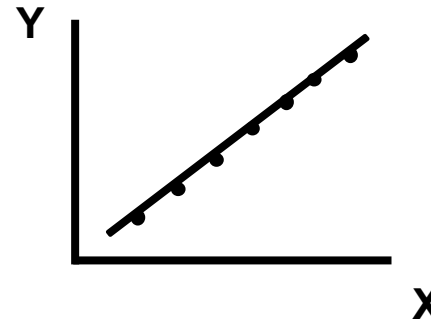
Jednorozměrná  
 $y = f(x)$

Vícerozměrná  
 $y = f(x_1, x_2, x_3, \dots, x_p)$

Deterministický

Regresní, stochastický

Vztah x, y



Pro každé x existuje pravděpodobnostní rozložení y



# Příklady lineární nebo "linearizovatelné" regrese

- I. **Y ..... koncentrace antigenů**  
**X ..... čas**

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \approx Y = \beta_0 + \beta_1 (\text{čas}) + \beta_2 (\text{čas})^2$$

$$\beta_0 : 0,014 \quad P = 0,328$$

$$\beta_1 : 0,182 \quad P = 0,000$$

$$\beta_2 : 0,089 \quad P = 0,001$$

- II. **Y ..... koncentrace O<sub>2</sub> ve vodě**  
**X ..... koncentrace org. C ve vodě**

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

- III.
- $Y = \exp(a + b \cdot x)$  ..... exponenciální
- $Y = a \cdot x^b$  ..... multiplikativní
- $\frac{1}{Y} = a + b \cdot x$  ..... reciproční

# Regresní analýza přímky - "Simple regression"

$$Y = a + b \cdot x + e \quad \approx \quad \alpha + \beta \cdot X + \varepsilon$$

$y$   $\left\{ \begin{array}{l} \alpha \approx a \text{ (intercept): } a = \bar{y} - b \cdot \bar{x} \\ \beta \cdot X \approx b \cdot x \text{ (sklon; slope)} \\ \varepsilon \approx e \text{ - náhodná složka : } N(0; \sigma_e^2) = N(0; \sigma_y^2 x) \end{array} \right.$

Komponenty  
tvořící  $y$  se  
sčítají

$\varepsilon$  - náhodná složka modelu přímky = rezidua přímky

$$\sigma_e^2 \left( \sigma_{y \cdot x}^2 \right) \Rightarrow \text{rozptyl reziduí}$$

# Základní regresní analýzy: model přímky v datech

$$\begin{matrix} 1 \\ \vdots \\ n \end{matrix} \quad \mathbf{x} \quad \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

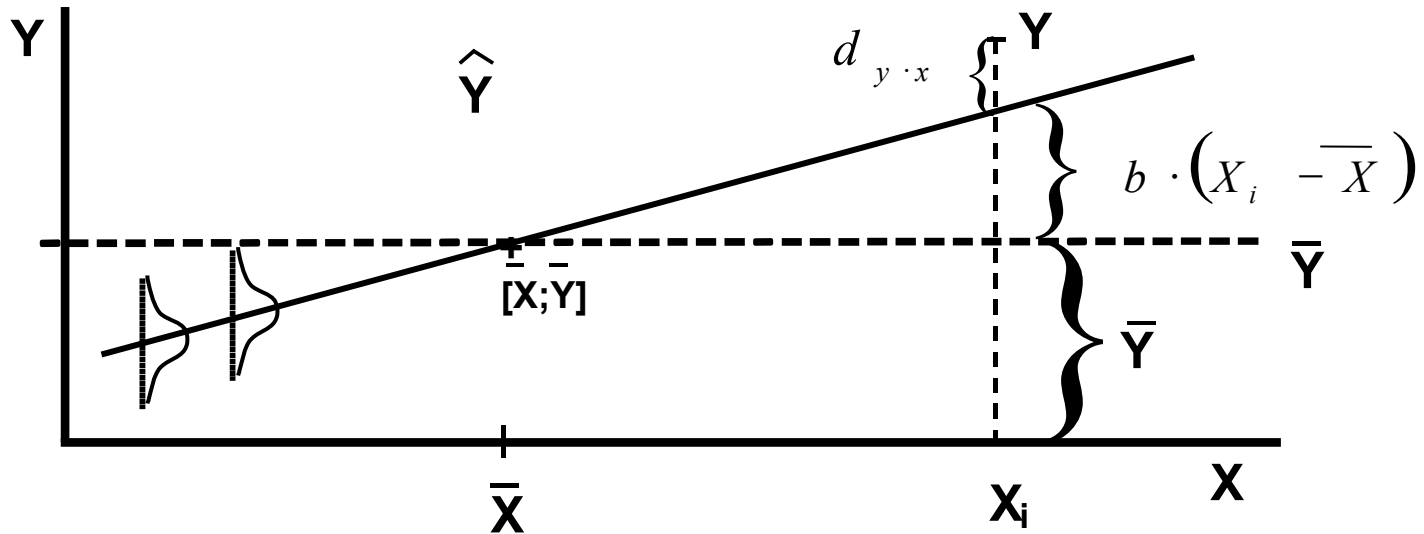
$$\begin{matrix} 1 \\ \vdots \\ n \end{matrix} \quad \mathbf{y} \quad \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$



$$\begin{matrix} 1 \\ \vdots \\ n \end{matrix} \quad \hat{\mathbf{y}} \quad \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} = a + b \cdot \begin{matrix} \mathbf{x} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix} \quad \longrightarrow \quad \begin{matrix} \mathbf{y} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix} - \begin{matrix} \hat{\mathbf{y}} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix} = \begin{matrix} \mathbf{e} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix}$$



# Základní regresní analýzy: model přímky v datech



$$d_{y \cdot x} = y - \hat{y} \quad \boxed{d_{y \cdot x} = y - \bar{y} - b(X_i - \bar{X})} \quad \hat{y} = \bar{y} + b(X_i - \bar{X})$$

**Smysl proložení přímky**  
minimalizace odchylek

$$d_{y \cdot x}^2 \rightarrow \sum [y - \hat{\alpha} - \hat{\beta}(X_i - \bar{X})]$$

## Metoda nejmenších čtverců

- 1) X: Pevná, nestochastická proměnná
- 2) Rozložení hodnot y pro každé x je normální
- 3) Rozložení hodnot y pro každé x má stejný rozptyl
- 4) Rezidua jsou navzájem nezávislá a mají normální rozložení:  $N(0; \sigma_e^2)$

# Odhady parametrů pro lineární regresi

I.  $b \sim \beta : b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$        $S_b^2 \sim \sigma_\beta^2 : \frac{1}{\sum (X_i - \bar{X})^2} \cdot S_{y \cdot x}^2$

$S_{y \cdot x}^2 =$  mean squared deviation from regression

$S_{y \cdot x} =$  sample standard deviation from regression

$$S_{y \cdot x}^2 = \frac{\sum d_{y \cdot x}^2}{n-2} = \frac{\sum Y_i^2 - \frac{\sum Y_i^2}{n} - b^2 \cdot \sum (X_i - \bar{X})^2}{n-2}$$

II.  $a \sim \alpha : a = \bar{Y} - b \cdot \bar{X}$        $S_a^2 \sim \sigma_\alpha^2$        $S_\alpha^2 = \left[ \frac{1}{n} + \frac{\bar{X}^2}{\sum X^2} \right] \cdot S_{y \cdot x}^2$

intercept

III.  $\hat{Y}$  : modelová hodnota

$$\hat{Y}_i = a - b \cdot X_i \quad S_{\hat{y}_i} = (S_{y \cdot x}) \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum X^2}}$$

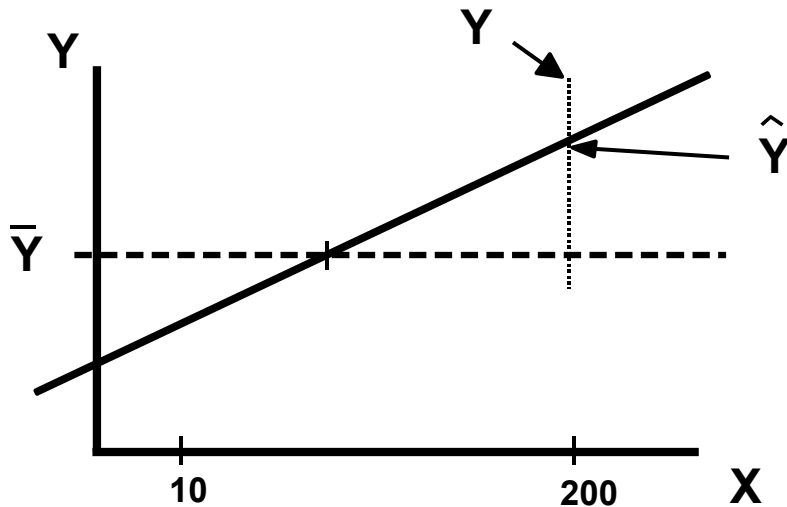
# Smysl lineární regrese

**X: Množství spáleného odpadu (tuny)**

**Y: Koncentrace kovu ve vzduchu(ng/m<sup>3</sup>)**

Platí: X = 0; 10; 100; 150; 200; 250; 300 tun

Model: Y = a + b · X



Výsledek:  $\hat{Y} = 14 + 0,123 \cdot X$ ;  $\hat{Y} \rightarrow \left[ \frac{\text{ng kov}}{m^3} \right]$



Např. : Skutečná data pro X = 200 t:

$Y_i = 16; 25; 41; 28; 31; 20 \Rightarrow Y_i = 26.8$

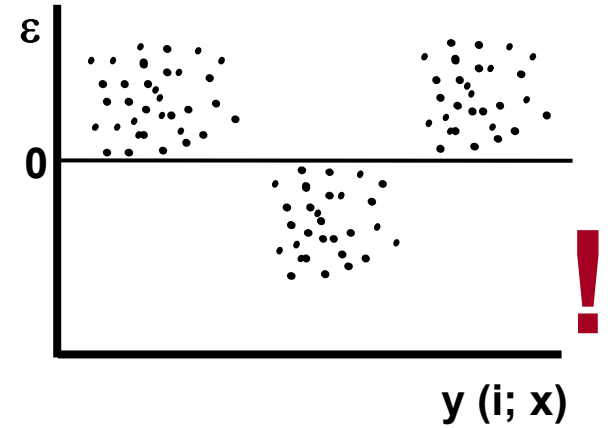
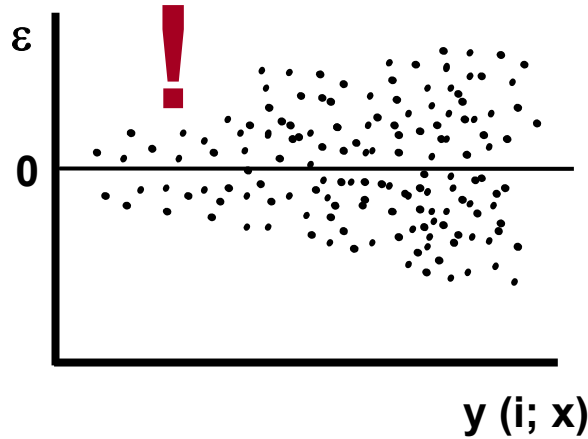
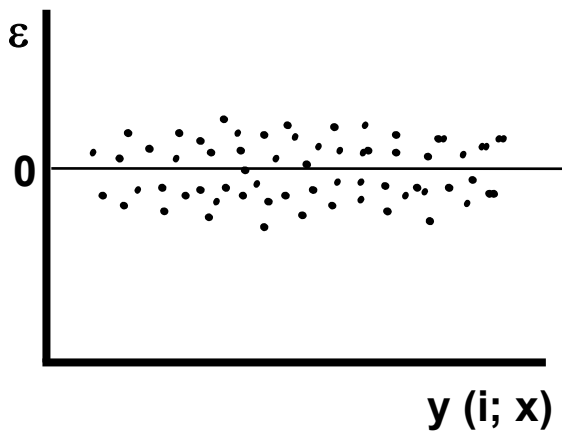
$$\left. \begin{aligned} \hat{Y} &= \bar{Y} + b \cdot (X - \bar{X}) \\ \hat{Y} &= a + b \cdot X \end{aligned} \right\} a = \bar{Y} - b \cdot \bar{X}$$

Odhadnuto z modelu pro X = 200 t:

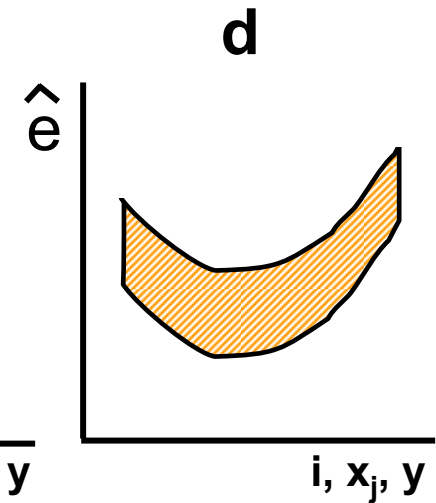
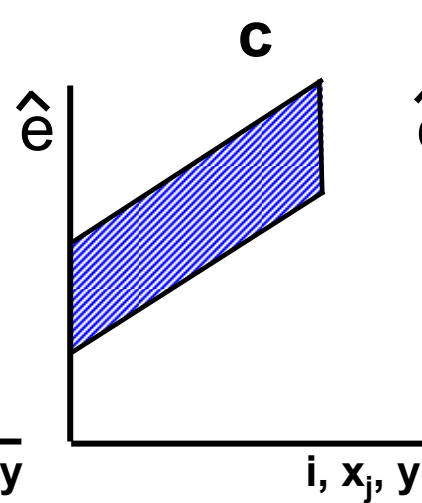
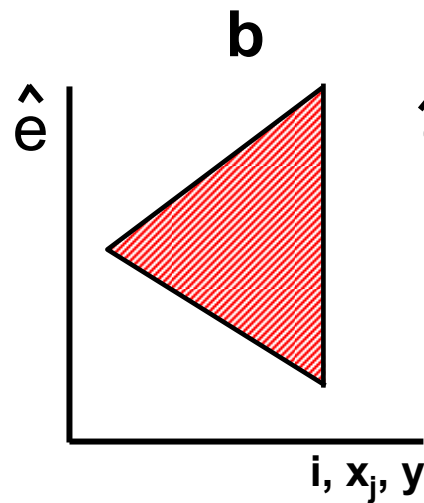
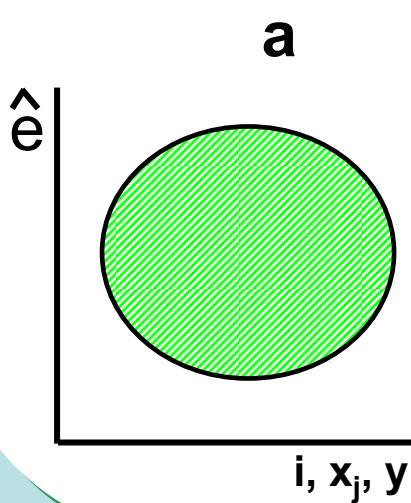
$$\hat{Y} = 14 + 0,123 \cdot 200 = 38,6$$

# Regresní analýza v grafech

## 3) Grafy residuí modelů (příklady)



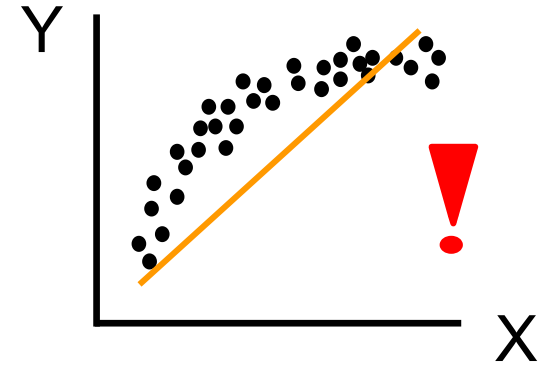
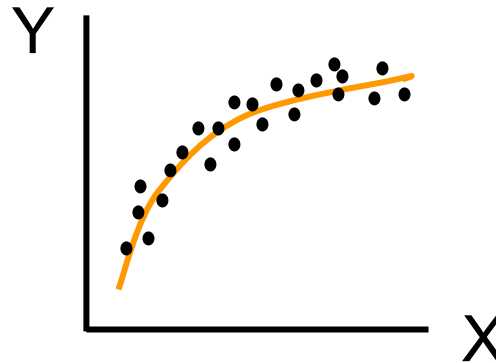
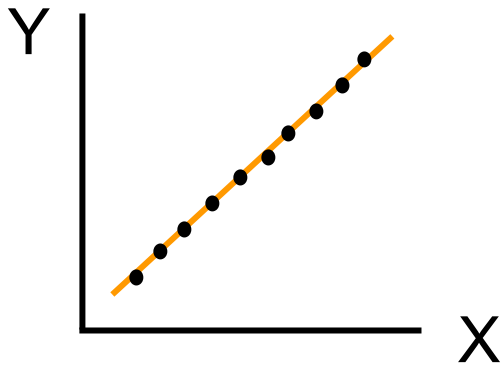
## Obecné tvary residuí modelů (schéma)



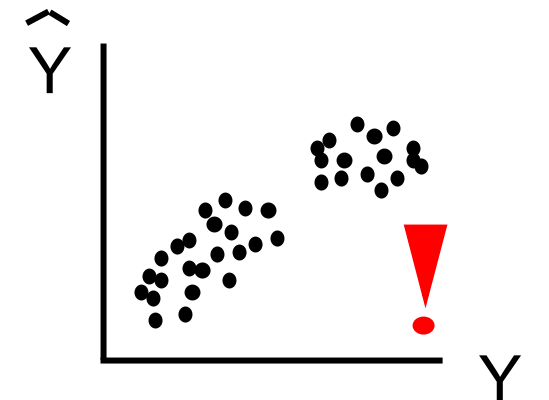
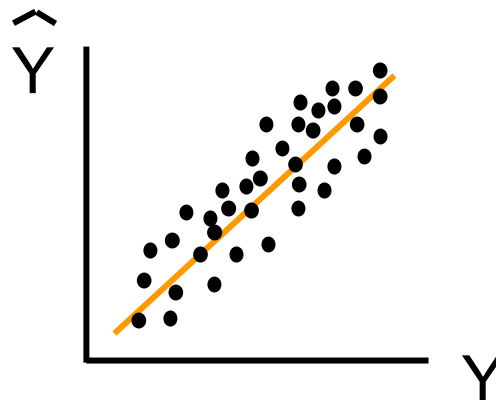
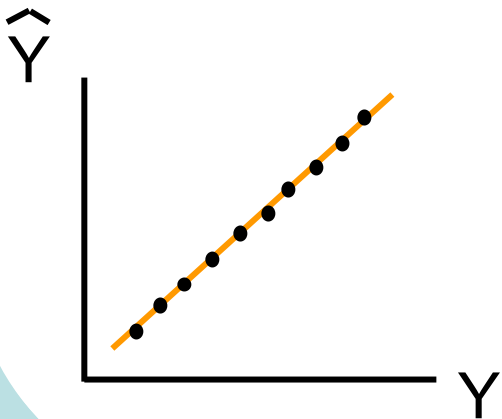


# Regresní analýza v grafech

## 1) Y vs. X



## 2) Y vs. $\hat{Y}$



# Lineární regrese - příklad

**X: Koncentrace drogy: 0; 2; 6; 8; 10; 12; 15 mg/ml krve**

**Y: Koncentrace volných metabolitů**

**Pro každé X: 3 opakování Y**

**Model:  $Y = a + b \cdot x \rightarrow Y = 0,11 + 0,092 \cdot X$**

$$t_{0,975}^{(v=19)} = 2,093$$

$$\text{I. } \left. \begin{array}{l} H_0 : \beta = 0; \alpha = 0,05 \\ b = 0,092 ; s_b = 0,023 \end{array} \right\} t = \frac{b}{S_b} = 4,00$$

$$\beta : b \pm t_{1-\alpha/2}^{(n-2)} \cdot S_b$$

**P < 0,01**

$$P(0,044 \leq \beta \leq 0,140) = 0,95$$

$$\text{II. } \left. \begin{array}{l} H_0 : \alpha = 0; \alpha = 0,05 \\ a = 0,11; s_a = 0,029 \end{array} \right\} t = \frac{a}{S_a} = 3,793$$

$$t_{0,975}^{(v=19)} = 2,093$$

$$\alpha : \alpha \pm t_{1-\alpha/2}^{(n-2)} \cdot S_a$$

$$P(0,049 \leq \alpha \leq 0,171) = 0,95$$



# Analýza rozptylu jako nástroj analýzy regresních modelů - příklad na modelu přímky

3) → Celková ANOVA  $\begin{cases} SS_B/SS_T & \text{(variance ratio)} \\ MS_B/MS_E = F \end{cases}$

4) Analýza rozptylu regresního modelu (zde přímky)

Zdroj rozptylu	st.v.	SS	MS	F
Model (přímka)	1	$SS_{MOD}$	$MS_{MOD}$	$MS_{MOD} / MS_R$
Residuum	$na - 2$	$SS_R$	$MS_R$	
celkem	$na - 1$	$SS_T$		

$(SS_{MOD}/SS_T) \cdot 100 = \% \text{ rozptylu } Y$   
 "vyčerpaného"  
 přímkou = koeficient determinace ( $R^2$ )

# Lineární regrese - příklad

X: konc.Cd: 1,2,3,4,5,6 ng/ml

Y: absorb: 0,23; 0,49; 0,72; 0,90; 1,16; 1,39

**b=0,228**

**$S_b=4,99 \cdot 10^{-3}$**

**P = 0,000**

**a=0,016**

**$S_a=0,019$**

**P = 0,457**

r = 0,999

$R_2 = 99,81\%$

St. Error of est: 0,021

## ANOVA

Source	D.f.	SS	MS	F	P
Model	1	0,912	0,912	2086,3	0
Residual	4	0,0017	0,000425		
Total ( c )	5	0,9138			

$$s^2_{y.x} = 4,25 \cdot 10^{-4}$$

$$s^2_y = 0,18275$$