

# Analýza dat pro Neurovědy



RNDr. Eva Janoušová  
doc. RNDr. Ladislav Dušek, Dr.

Jaro 2014

# Blok 5

Jak analyzovat kategoriální a binární data I.

# Typy dat - opakování

- **Kvalitativní (kategoriální) data:**

- Binární data



- Nominální data



- Ordinální data



- **Kvantitativní data:**

- Intervalová data



- Poměrová data



# Osnova

---

1. Analýza kontingenčních tabulek
2. Binomické testy
3. Relativní riziko („relative risk“) a poměr šancí („odds ratio“)
4. Binomické rozdělení
5. Poissonovo rozdělení

# 1. Analýza kontingenčních tabulek

# Kontingenční tabulka

- Frekvenční sumarizace dvou binárních, nominálních nebo ordinálních proměnných.
- Obecně:  $R \times C$  **kontingenční tabulka** ( $R$  – počet kategorií jedné proměnné,  $C$  – počet kategorií druhé proměnné).
- Speciální případ:  $2 \times 2$  tabulka = čtyřpolní tabulka.
- Příklad: Sumarizace vyšetřených osob podle typu onemocnění a věkových kategorií.

Typ onemocnění	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	1	7	176	46	230
MCI	13	85	201	107	406
AD	9	34	90	64	197
Celkem	23	126	467	217	833

# Kontingenční tabulky – absolutní četnosti, řádková, sloupcová a celková procenta

Kontingenční tabulka absolutních četností

Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	1	7	176	46	230
MCI	13	85	201	107	406
AD	9	34	90	64	197
Celkem	23	126	467	217	833

Kontingenční tabulka řádkových procent

Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	0,4	3,0	76,5	20,0	100,0
MCI	3,2	20,9	49,5	26,4	100,0
AD	4,6	17,3	45,7	32,5	100,0
Celkem	2,8	15,1	56,1	26,1	100,0

Kontingenční tabulka sloupcových procent

Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	4,3	5,6	37,7	21,2	27,6
MCI	56,5	67,5	43,0	49,3	48,7
AD	39,1	27,0	19,3	29,5	23,6
Celkem	100,0	100,0	100,0	100,0	100,0

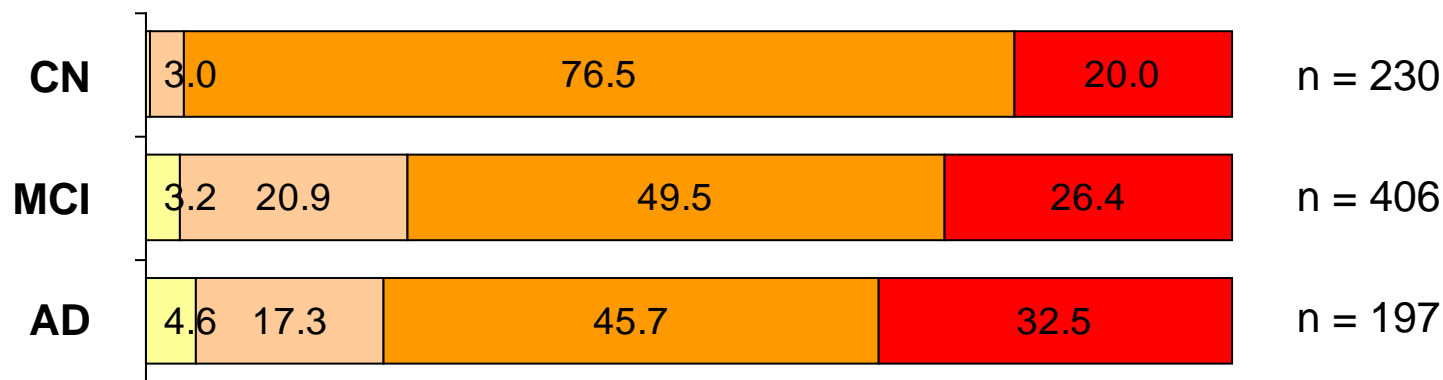
Kontingenční tabulka celkových procent

Skupina	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	0,1	0,8	21,1	5,5	27,6
MCI	1,6	10,2	24,1	12,8	48,7
AD	1,1	4,1	10,8	7,7	23,6
Celkem	2,8	15,1	56,1	26,1	100,0

# Kontingenční tabulky – ukázka finálního popisu a vizualizace

Skupina	věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	1 (0,4%)	7 (3,0%)	176 (76,5%)	46 (20,0%)	230 (100,0%)
MCI	13 (3,2%)	85 (20,9%)	201 (49,5%)	107 (26,4%)	406 (100,0%)
AD	9 (4,6%)	34 (17,3%)	90 (45,7%)	64 (32,5%)	197 (100,0%)
Celkem	23 (2,8%)	126 (15,1%)	467 (56,1%)	217 (26,1%)	833 (100,0%)

Skupina:



Věk:  <60 let  60-70 let  70-80 let  ≥80 let



# Čtyřpolní tabulky

- Nejjednodušší možná kontingenční tabulka, kdy obě sledované veličiny mají pouze dvě kategorie.
- **Příklad:** Sumarizace vztahu pohlaví a kategorizovaného MMSE skóre (MMSE skóre v normě (tzn.  $MMSE \geq 25$ ) a pod normou ( $MMSE < 25$ )) u pacientů s Alzheimerovou chorobou.

2-Way Summary Table: Observed Frequencies (Data\_neuro\_vycistena3)  
Marked cells have counts > 10  
Include condition: v3=3

Gender_rek	mmse_kat v norme	mmse_kat pod normou	Row Totals				
M	36	66	102				
F	31	64	95				
Totals	67	130	197				

# Kontingenční tabulky – hypotézy

- Kontingenční tabulky umožňují testování různých hypotéz:
- **Nezávislost a shoda struktury** (Pearsonův chí-kvadrát test, Fisherův exaktní test)
  - Jeden výběr, dvě charakteristiky nebo více výběrů, jedna charakteristika – obdoba nepárového uspořádání
  - Příklad: pacienti s AD – pohlaví × vzdělání (VŠ, SŠ, ZŠ); pacienti s AD v několika nemocnicích × věková struktura
- **Symetrie** (McNemarův test)
  - Jeden výběr, opakovaně jedna charakteristika – obdoba párového uspořádání
  - Příklad: MMSE v normě a pod normou na začátku studie a dva roky po zahájení studie

# Pearsonův chí-kvadrát test

- Založen na myšlence srovnání pozorovaných a očekávaných četností kategorií dvou proměnných.
- Pozorované četnosti jednotlivých kategorií první proměnné a druhé proměnné nám vyjadřují  $n_{ij}$ .
- Očekávané četnosti jednotlivých kategorií lze vypočítat pomocí:

$$e_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

( $n_{i.}$  je součet hodnot v řádku,  
 $n_{.j}$  je součet hodnot ve sloupci)

- Výpočet testové statistiky:

$$C^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- Nulovou hypotézu o nezávislosti dvou kategoriálních proměnných zamítáme na hladině významnosti  $\alpha$ , když  $C^2 \geq c_{(1-\alpha)}^2(r-1)(c-1)$

Typ onemocnění	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	$n_{11}$	$n_{12}$	$n_{13}$	$n_{14}$	$n_{1.}$
MCI	$n_{21}$	$n_{22}$	$n_{23}$	$n_{24}$	$n_{2.}$
AD	$n_{31}$	$n_{32}$	$n_{33}$	$n_{34}$	$n_{3.}$
Celkem	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n$

# Pearsonův chí-kvadrát test

**Příklad:** Chceme zjistit, jestli existuje vztah mezi typem onemocnění a věkovými kategoriemi v našem souboru.

**Postup:**

Tabulka pozorovaných četností:

Typ onemocnění	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	1	7	176	46	230
MCI	13	85	201	107	406
AD	9	34	90	64	197
Celkem	23	126	467	217	833

Tabulka očekávaných četností:

Typ onemocnění	Věk				Celkem
	<60 let	60-70 let	70-80 let	≥80 let	
CN	6,4	34,8	128,9	59,9	230
MCI	11,2	61,4	227,6	105,8	406
AD	5,4	29,8	110,4	51,3	197
Celkem	23	126	467	217	833

$$e_{11} = \frac{23 \times 230}{833} = 6,4$$

$$e_{21} = \frac{23 \times 406}{833} = 11,2$$

$$e_{12} = \frac{126 \times 230}{833} = 34,8 \dots$$

Testová statistika: 
$$C^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(1 - 6,4)^2}{6,4} + \frac{(7 - 34,8)^2}{34,8} + \dots = 69,4$$

$C^2 = 69,4 \stackrel{3}{>} c_{(0,95)}^2(3 - 1)(4 - 1) = c_{(0,95)}^2(6) = 12,6 \rightarrow$  zamítáme  $H_0$  o nezávislosti  $\rightarrow$  Vztah mezi typem onemocnění a věkovými kategoriemi je statisticky významný.

# Předpoklady Pearsonova chí-kvadrát testu

- Nezávislost jednotlivých pozorování
- Alespoň 80 % buněk musí mít očekávanou četnost ( $e_{ij}$ ) větší než 5
- 100 % buněk musí mít očekávanou četnost ( $e_{ij}$ ) větší než 2
  
- Může nám pomoci slučování kategorií, ale můžeme slučovat jen slučitelné kategorie!

# Úkol 1.

---

- **Zadání:** Vhodně kategorizujte výšku a zjistěte, zda existuje vztah kategorizované výšky a pohlaví.

# Fisherův exaktní test

- Určen pro čtyřpolní tabulky, je vhodný i pro tabulky s malými četnostmi – pro ty, které nesplňují předpoklad Pearsonova chí-kvadrát testu.
- Založen na výpočtu „přesné“ p-hodnoty (pravděpodobnosti, s jakou bychom dostali stejný nebo ještě extrémnější výsledek při zachování součtu řádků i sloupců v tabulce).

- **Příklad:** Chceme ověřit vztah dvou typů nežádoucích účinků, které jsou sumarizovány následující tabulkou:

		NÚ II	
		ano	ne
NÚ I	ano	2	3
	ne	6	4

- **Postup:** Všechny varianty tabulky při zachování součtu řádků a sloupců:

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

Pravděpodobnosti výskytu jednotlivých tabulek:

0,007

0,093

0,326

0,392

0,163

0,019

Oboustranná p-hodnota (sečtení pravděpodobností stejných nebo menších než je pravděpodobnost pozorované varianty):

$$p = 0,326 + 0,093 + 0,007 + 0,163 + 0,019 = 0,608$$

# Fisherův exaktní test

- **Příklad:** Chceme ověřit vztah pohlaví a kategorizovaného MMSE skóre (MMSE skóre v normě (tzn.  $MMSE \geq 25$ ) a pod normou ( $MMSE < 25$ )) u pacientů s Alzheimerovou chorobou.
- **Řešení:**

2-Way Summary Table: Observed Frequencies (Data\_neuro\_vycistena3)  
Marked cells have counts > 10  
Include condition: v3=3

Gender_rek	mmse_kat v norme	mmse_kat pod normou	Row Totals
M	36	66	102
F	31	64	95
Totals	67	130	197

Statistics: Gender\_rek(2) x mmse\_kat(2) (Data\_neuro\_vycistena3)  
Include condition: v3=3

Statistic	Chi-square	df	p
Pearson Chi-square	,1553688	df=1	p=,69346
M-L Chi-square	,1554686	df=1	p=,69336
Yates Chi-square	,0593807	df=1	p=,80748
Fisher exact, one-tailed			p=,40401
two-tailed			p=,76397
McNemar Chi-square (A/D)	7,290000	df=1	p=,00693
(B/C)	11,91753	df=1	p=,00056



# Fisherův x Pearsonův test

- Pearsonův chí-kvadrát test lze použít na jakoukoliv kontingenční tabulku, ALE je nutné hlídat předpoklady: 100% očekávaných četností větších než 2 a 80 % očekávaných četností větších než 5 – u čtyřpolní tabulky to znamená, že všechny očekávané četnosti musí být větší než 5.
- **Nedodržení předpokladů pro Pearsonův chí-kvadrát test může stejně jako u t-testu a analýzy rozptylu vést k nesmyslným závěrům!**
- **Pro hodnocení čtyřpolních tabulek je Fisherův exaktní test standardem v klinických analýzách.**

# Analýza kontingenčních tabulek na webu

---

- 2x2 tabulky: <http://graphpad.com/quickcalcs/contingency1/>
- 2x3 tabulky: <http://www.vassarstats.net/fisher2x3.html>
- 2x5 (nebo menší) tabulky:  
<http://www.quantitativeskills.com/sisa/statistics/fiveby2.htm>

# Úkol 2.

- **Zadání:** Zjistěte, zda existuje vztah mezi typem onemocnění (AD a MCI) a kategorizovaného MMSE skóre (pod normou a v normě) u žen.
- **Řešení:**

2-Way Summary Table: Observed Frequencies (Data_neuro_vycistena3)			
Marked cells have counts > 10			
Include condition: v5="F"			
Exclude condition: v3=1			
Group_3kat	mmse_kat v norme	mmse_kat pod normou	Row Totals
MCI	128	18	146
AD	31	64	95
Totals	159	82	241

Statistics: Group_3kat(2) x mmse_kat(2) (Data_neuro_vycistena3)			
Include condition: v5="F"			
Exclude condition: v3=1			
Statistic	Chi-square	df	p
<b>Pearson Chi-square</b>	77,66653	df=1	p=0,0000
M-L Chi-square	80,02626	df=1	p=0,0000
Yates Chi-square	75,23401	df=1	p=0,0000
Fisher exact, one-tailed			p=,00000
two-tailed			p=,00000
McNemar Chi-square (A/D)	20,67188	df=1	p=,00001
(B/C)	2,938776	df=1	p=,08648

# McNemarův test

- Je to **obdoba párového testu** (test symetrie pro kontingenční tabulku).
- Testová statistika pro čtyřpolní tabulku:

$$C^2 = \frac{(b - c)^2}{b + c}$$

Veličina X	Veličina Y		Celkem
	Y = 1	Y = 2	
X = 1	a	b	a + b
X = 2	c	d	c + d
Celkem	a + c	b + d	n

rozdílné výsledky

- Zaměřuje se pouze na pozorování, u kterých jsme při opakovaném měření zaznamenali rozdílné výsledky – za platnosti  $H_0$  by jejich četnosti (označeny b a c) měly být stejné.
- Testová statistika pro obecnou čtvercovou kontingenční tabulku:

$$C^2 = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

# McNemarův test

- **Příklad:** Zjistěte, zda se liší kategorizované MMSE skóre při vstupu do studie a dva roky po zahájení studie.

rozdílné výsledky

- **Řešení:**

2-Way Summary Table: Observed Frequencies (Data_neuro_vycistena3) Marked cells have counts > 10			
mmse_kat: =iif(v11<25;1;0)	mmse24_kat 0	mmse24_kat 1	Row Totals
v norme	280	102	382
pod normou	13	71	84
Totals	293	173	466

Statistics: mmse_kat(2) x mmse24_kat(2) (Data_neuro_vycistena3)			
Statistic	Chi-square	df	p
<b>Pearson Chi-square</b>	98,62901	df=1	p=0,0000
M-L Chi-square	99,04826	df=1	p=0,0000
Yates Chi-square	96,16740	df=1	p=0,0000
Fisher exact, one-tailed			p=0,0000
two-tailed			p=,00000
McNemar Chi-square (A/D)	123,2593	df=1	p=0,0000
(B/C)	67,33913	df=1	p=,00000

## 2. Binomické testy

# Binomické testy

Pokud máme spočítané podíly pacientů s výskytem sledovaného jevu, můžeme k testování použít i binomické testy:

1. **Liší se podíl ( $p$ ) pacientů s výskytem sledovaného jevu od předpokládané (referenční) hodnoty ( $\pi$ )?**  
(Např. liší se procento pacientů s nežádoucími účinky léčby od předpokládaného procenta?)  
→ **jednovýběrový binomický test** (tzn. test pro podíl u jednoho výběru)
2. **Liší se podíly pacientů s výskytem sledovaného jevu ve dvou souborech?**  
(Např. liší se podíl pacientů s nežádoucími účinky léčby podle typu léčby?)  
→ **dvouvýběrový binomický test** (tzn. test pro podíl u dvou výběrů)

# Jednovýběrový binomický test

- **Příklad:** Mezi 50 pacienty s Alzheimerovou chorobou je 12 pacientů s MMSE skóre nižším než daná hranice. Ověřte, zda podíl pacientů s nižším skóre je stejný jako v běžné populaci.
- Tzn. hypotézy budou mít tvar:  $H_0 : p = p_0$  a  $H_1 : p \neq p_0$

- **Řešení:**

- $\pi = 0,05$  (v populaci – hranice skóre jsou dělána tak, aby 5% populace bylo nižší než hranice)
- $p = 12/50 = 0,24$
- **Závěr:**  
Podíl pacientů s nižším MMSE skóre je statisticky významně odlišný od podílu v běžné populaci.

Difference tests: r, %, means: Data\_neuro\_vycistena3

Send/print results for each Compute to Report window

Cancel

Difference between two correlation coefficients

r1: 0,00 N1: 10

r2: 0,00 N2: 10 p: 1,0000

One-sided

Two-sided

Compute

Difference between two means (normal distribution)

M 1: 0 StDv 1: 1 N1: 10 p: 1,0000

M 2: 0 StDv 2: 1 N2: 10

One-sided

Two-sided

Single mean 1 vs .population mean 2

Compute

Difference between two proportions

Pr.1: 0,240000 N1: 50

Pr.2: 0,050000 N2: 32767 p: 0,0000

One-sided

Two-sided

Compute

Co největší N2

Vypočtená p-hodnota



# Dvouvýběrový binomický test

- **Příklad:** Mezi 42 pacienty s Alzheimerovou chorobou (AD) je 11 pacientů s MMSE skóre nižším než daná hranice. Mezi 18 pacienty s mírnou kognitivní poruchou (MCI) je 6 pacientů s MMSE skóre nižším než daná hranice. Ověřte, zda se podíly pacientů s nižším skóre u pacientů s AD a MCI liší.
- Tzn. hypotézy budou mít tvar:  $H_0 : p_1 = p_2$  a  $H_1 : p_1 \neq p_2$

- **Řešení:**

- $p_1 = 11/42 = 0,262$
- $p_2 = 6/18 = 0,333$

- **Závěr:**

Neprokázali jsme, že by se podíl subjektů s nižším MMSE skóre lišil u pacientů s AD a MCI.

Difference tests: r, %, means: DMdata - final - do Statistic

Send/print results for each Compute to Report window

Difference between two correlation coefficients

r1: 0,00 N1: 10 p: 1,0000  One-sided  Two-sided

r2: 0,00 N2: 10

Difference between two means (normal distribution)

M 1: 0 StDv 1: 1 N1: 10 p: 1,0000  One-sided  Two-sided

M 2: 0 StDv 2: 1 N2: 10

Single mean 1 vs .population mean 2

Difference between two proportions

Pr.1: ,262000 N1: 42 p: ,5760  One-sided  Two-sided

Pr.2: ,333000 N2: 18

**Vypočtená p-hodnota**

### 3. Relativní riziko („relative risk“) a poměr šancí („odds ratio“)

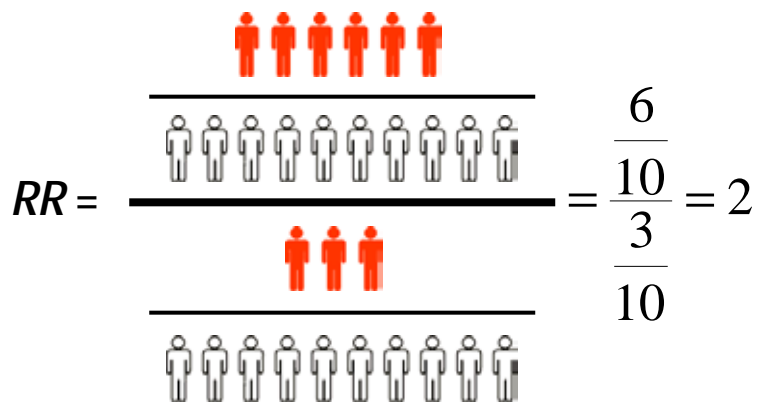
# Motivace



- Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS). Výsledky dány v tabulce:

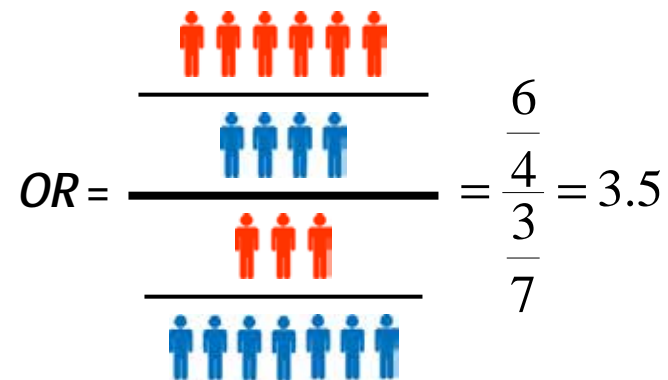
SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

- Pomocí Pearsonova chí-kvadrát nebo Fisherova exaktního testu můžeme rozhodovat o závislosti/nezávislosti dvou sledovaných veličin. Testy ale neumožňují tento vztah kvantifikovat.
- Má-li to smysl a chceme-li kvantifikovat (rozhodovat o těsnosti této závislosti) můžeme použít tzv. **relativní riziko** a **poměr šancí**.

# Grafické srovnání *RR* a *OR*



 Výskyt sledovaného jevu  
 Bez výskytu sledovaného jevu




# Relativní riziko („Relative Risk“)

- Výpočet relativního rizika (RR) umožňuje srovnat pravděpodobnosti výskytu sledovaného jevu ve dvou různých skupinách.
- 1. skupina – experimentální nebo skupina s expozicí určitému faktoru
- 2. skupina – kontrolní nebo skupina bez expozice

$$RR = \frac{\text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}{\text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}} = \frac{P_1}{P_0}$$

Sledovaný jev	Skupina		Celkem
	Experimentální	Kontrolní	
Ano	$a$	$b$	$a + b$
Ne	$c$	$d$	$c + d$
Celkem	$a + c$	$b + d$	$n$


$$RR = \frac{P_1}{P_0} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

# Relativní riziko

- **Příklad:** Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS). Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

$$RR = \frac{P_1}{P_0} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{29}{29+7301}}{\frac{15}{15+11241}} = 2,97$$



Riziko výskytu SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičích ve vyšším věku.

# Relativní riziko

**Příklad:** Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS). Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

**Závěr:** Riziko výskytu SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičích ve vyšším věku.

**Řešení** pomocí webového kalkulátoru:  
([http://www.medcalc.org/calc/relative\\_risk.php](http://www.medcalc.org/calc/relative_risk.php)):

## Relative risk

**Exposed group**

Number with positive outcome: a = 29

Number with negative outcome: b = 7301

**Control group**

Number with positive outcome: c = 15

Number with negative outcome: d = 11241

**Results**

Relative risk: **2.9688**

95 % CI: 1.5928 to 5.5336

z statistic: 3.425

P = 0.0006

The relative risk is the ratio of the proportions of cases having a positive outcome in two groups.

$$\text{Relative Risk} = (a / (a+b)) / (c / (c+d))$$

# Poměr šancí („Odds ratio“)

- Poměr šancí (OR) je další charakteristikou, která umožňuje srovnat výskyt sledovaného jevu ve dvou různých skupinách.
- 1. skupina – experimentální nebo skupina s expozicí určitému faktoru
- 2. skupina – kontrolní nebo skupina bez expozice

$$OR = \frac{\frac{\text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}{1 - \text{Pravděpodobnost výskytu jevu v 1. skupině (experimentální)}}}{\frac{\text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}}{1 - \text{Pravděpodobnost výskytu jevu ve 2. skupině (kontrolní)}}} = \frac{O_1}{O_0} = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}}$$

Sledovaný jev	Skupina		Celkem
	Experimentální	Kontrolní	
Ano	$a$	$b$	$a + b$
Ne	$c$	$d$	$c + d$
Celkem	$a + c$	$b + d$	$n$



$$OR = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}}$$



# Poměr šancí

- Příklad:** Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS). Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

$$OR = \frac{\frac{P_1}{1 - P_1}}{\frac{P_0}{1 - P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{29}{7301}}{\frac{15}{11241}} = 2,98$$



„Šance“ na výskyt SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičích ve vyšším věku.

# Poměr šancí

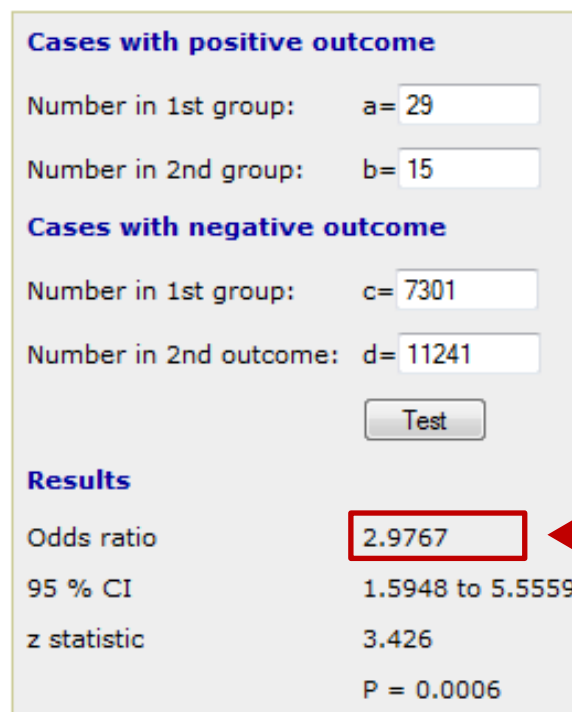
**Příklad:** Sledujeme souvislost věku matky a výskytu náhlého úmrtí kojence (SIDS). Výsledky dány v tabulce:

SIDS	Věk matky		Celkem
	Do 25 let	25 a více let	
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

**Závěr:** „Šance“ na výskyt SIDS u dětí matek ve věku do 25 je téměř třikrát vyšší než u dětí matek rodičích ve vyšším věku.

**Řešení** pomocí webového kalkulátoru:  
([http://www.medcalc.org/calc/odds\\_ratio.php](http://www.medcalc.org/calc/odds_ratio.php)):

## Odds ratio



The screenshot shows an online odds ratio calculator interface. It is divided into several sections: 'Cases with positive outcome', 'Cases with negative outcome', and 'Results'. In the 'Cases with positive outcome' section, 'Number in 1st group' is set to a=29 and 'Number in 2nd group' is set to b=15. In the 'Cases with negative outcome' section, 'Number in 1st group' is set to c=7301 and 'Number in 2nd outcome' is set to d=11241. A 'Test' button is visible below these inputs. The 'Results' section displays the calculated Odds ratio as 2.9767, which is highlighted with a red box and a red arrow pointing to it from the right. Other results shown include a 95% CI of 1.5948 to 5.5559, a z statistic of 3.426, and a P value of 0.0006.

The odds ratio is the ratio of the odds of the outcome in the two groups.

$$\text{Odds ratio} = (a/c) / (b/d)$$

# Úkol 3.

- Zadání:** Sledujeme výskyt nežádoucích účinků u mužů a u žen (viz tabulka). Vypočtěte relativní riziko a poměr šancí.

Nežádoucí účinky	Pohlaví		Celkem
	Muž	Žena	
Ano	34	19	53
Ne	16	31	47
Celkem	50	50	100

$$RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{34}{34+16}}{\frac{19}{19+31}} = 1,79$$

Riziko výskytu nežádoucích účinků u mužů je téměř 1,8-krát vyšší než u žen.

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{\frac{34}{16}}{\frac{19}{31}} = 3,47$$

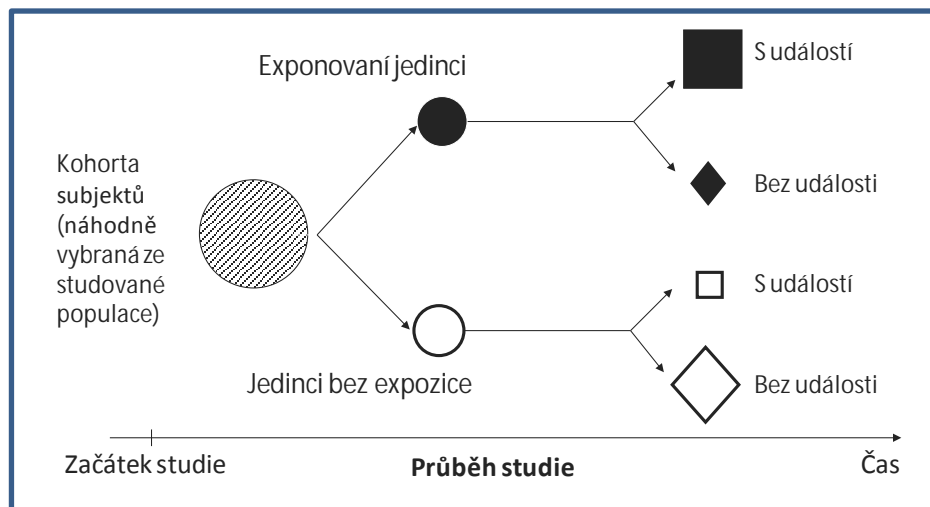
„Šance“ na výskyt nežádoucích účinků u mužů je téměř 3,5-krát vyšší než u žen.

# Výhody a nevýhody *RR* a *OR*

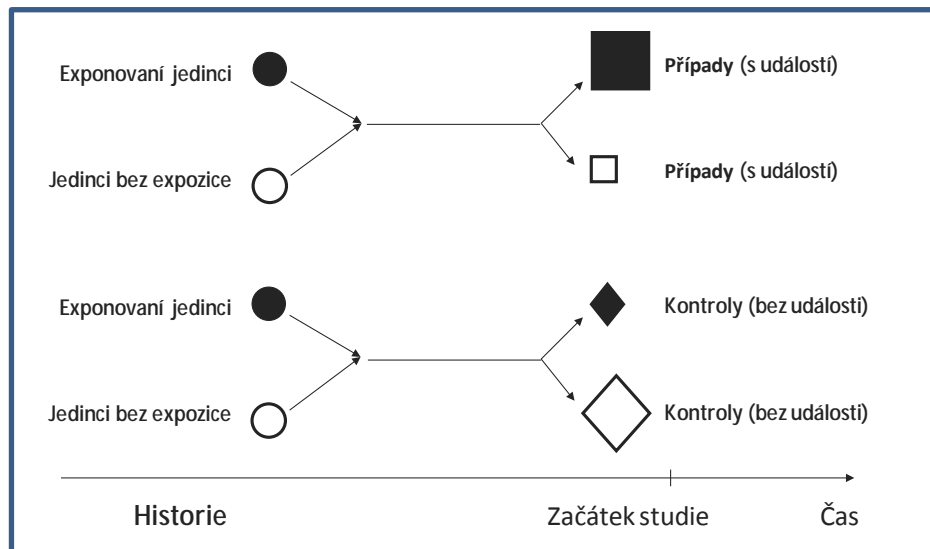
- Nevýhoda *OR*:
  - obtížná interpretace.
- Výhoda i nevýhoda *RR*:
  - nezajímá ho samotná pravděpodobnost výskytu jevu, ale pouze jejich podíl → korektní použití *RR* je však pouze v případě, že pravděpodobnost výskytu jevu v kontrolní skupině je reprezentativní (není ovlivněna výběrem sledovaných subjektů).

# Prospektivní a retrospektivní studie

- **Prospektivní studie**
- U některých subjektů je rizikový faktor přítomen a u jiných ne → sledujeme v čase, zda se vyskytne událost.



- **Retrospektivní studie**
- U některých subjektů se událost vyskytla a u jiných ne → zpětně hodnotíme, zda se liší s ohledem na nějaký rizikový faktor.



# Použití *RR* a *OR*

- **Prospektivní studie** – u některých subjektů je rizikový faktor přítomen a u jiných ne → sledujeme, zda se vyskytne událost.
- Zjištěná pravděpodobnost výskytu události v kontrolní skupině je reprezentativní, neboť prospektivně zařazujeme všechny pacienty  
→ **korektní použití *RR*.**
- **Retrospektivní studie** – u některých subjektů se událost vyskytla a u jiných ne → zpětně hodnotíme, zda se liší s ohledem na nějaký rizikový faktor.
- Zjištěná pravděpodobnost výskytu události v kontrolní skupině není reprezentativní, neboť ji ovlivňujeme zpětným výběrem skupin subjektů.  
→ **nekorektní použití *RR*.**  
→ **korektní použití *OR*.**

# Srovnávané skupiny

- Pomocí RR i OR můžeme srovnat pravděpodobnosti výskytu sledovaného jevu ve dvou různých skupinách:
- **1. skupina s pravděpodobností výskytu události  $P_1$ :**
  - experimentální skupina – např. léčená novou léčbou
  - riziková skupina – např. hypertonici
  - skupina s expozicí určitému faktoru – např. horníci
- **2. skupina s pravděpodobností výskytu události  $P_0$ :**
  - kontrolní skupina
  - skupina bez expozice

# Další způsoby vyjádření rozdílu rizika

- Relativní redukce rizika (RRR)

$$\text{RRR} = 1 - \text{RR} = 1 - \frac{\frac{\text{3}}{\text{10}}}{\frac{\text{5}}{\text{10}}} = 1 - \frac{\text{10}}{\text{5}} = 1 - 0.6 = 40\%$$

- Absolutní redukce rizika (ARR)

$$\text{ARR} = \frac{\text{5}}{\text{10}} - \frac{\text{3}}{\text{10}} = 0.2 = 20\%$$



# Další způsoby vyjádření rozdílu rizika

- Počet pacientů, které je potřeba léčit, abychom zabránili výskytu jedné události – „number needed to treat“ (NNT).

ARR = 20%     $\longrightarrow$     Pro snížení počtu událostí o 20 je třeba léčit 100 pacientů.



$$\text{NNT} = \frac{1}{0,2} = \frac{100}{20} = 5$$

**NNT = Pro snížení počtu událostí o 1 je třeba léčit 5 pacientů.**

# Absolutní vs. relativní četnost

- Vyjádření výsledků v relativní formě (procento) má často příjemnou interpretaci, ale může být zavádějící.
- Relativní vyjádření účinnosti by mělo být vždy doprovázeno absolutním vyjádřením účinnosti.
- **Příklad:** Srovnání účinnosti léčiva ve smyslu prevence CMP u kardiaků.  
Studie 1: Výskyt CMP ve skupině A je 12 %, ve skupině B je 20 %.  
Relativní změna v účinnosti = 40 %; absolutní změna = 8 %.  
Studie 2: Výskyt CMP ve skupině A je 0,9 %, ve skupině B je 1,5 %.  
Relativní změna v účinnosti = 40 %; absolutní změna = 0,6 %.
- Výsledkem je rozdílný přínos léčby při stejné relativní účinnosti.

# NNT a absolutní vs. relativní četnost

- Srovnání účinnosti léčiva ve smyslu prevence CMP u kardiaků.

Studie 1: Výskyt CMP ve skupině A je 12 %, ve skupině B je 20 %.  
Relativní změna v účinnosti = 40 %; absolutní změna = 8 %.

➔ 
$$\text{NNT} = \frac{1}{0,08} = \frac{100}{8} = 12,5$$

NNT = Pro snížení počtu událostí o 1 je třeba léčit 13 pacientů.

Studie 2: Výskyt CMP ve skupině A je 0,9 %, ve skupině B je 1,5 %.  
Relativní změna v účinnosti = 40 %; absolutní změna = 0,6 %.

➔ 
$$\text{NNT} = \frac{1}{0,006} = \frac{100}{0,6} = 166,7$$

NNT = Pro snížení počtu událostí o 1 je třeba léčit 167 pacientů.

# 4. Binomické rozdělení

# Typy dat - opakování

- **Kvalitativní (kategoriální) data:**

- Binární data



- Nominální data



- Ordinální data



- **Kvantitativní data:**

- Intervalová data



- Poměrová data



# Motivace

---

- Nejjednodušším případem kategoriálních dat jsou data binární.
- Binární data jsou popsána binomickým rozložením.
- Od chování binomického rozložení je odvozena:
  - popisná statistika binárních dat (procento výskytu jevu)
  - interval spolehlivosti pro binární data
  - binomické testy pro srovnání procentuálního výskytů jevů v různých skupinách

# Binomické rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události** (ve formě nastala/nenastala) v sérii  $n$  **nezávislých pokusech**, kdy v každém pokusu je **stejná pravděpodobnost výskytu** této události.

- Značení:  $Bi(n, \pi)$

- Parametry:

$n$  ... počet nezávislých pokusů

$r$  ... počet, kolikrát nastala sledovaná událost ( $r = 0 \dots n$ )

$p = r/n$  ... pravděpodobnost nastání sledované události ( $p \sim \pi$ )

- Pravděpodobnost, že sledovaná událost nastane  $r$ -krát, lze vypočítat:

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

- **Střední hodnota:**  $EX = n \cdot p$

- **Rozptyl:**  $DX = n \cdot p \cdot (1 - p)$

- Příklady: výskyt nežádoucích účinků léku u léčených pacientů, počet zemřelých pacientů mezi léčenými pacienty, počet pacientů s výsledkem neuropsycholog. testu pod normou

# Binomické rozdělení – příklad

- **Př. Pravděpodobnost narození chlapce je 0,5. Jaká je pravděpodobnost toho, že mezi čtyřmi dětmi v rodině je 0, 1,... až 4 chlapců. Vypočítejte i jaký je nejpravděpodobnější počet chlapců v této rodině.**
- **Řešení:**  $n = 4$  (4 děti v rodině)  
 $r = 0, 1, 2, 3, 4$  chlapců

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

$$P(X = 0) = \frac{4!}{0!4!} \times 0,5^0 \times (1 - 0,5)^4 = 0,0625$$

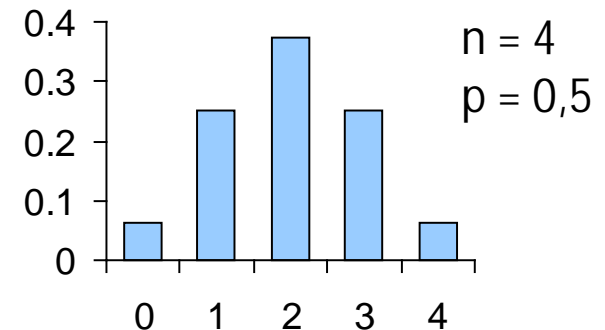
$$P(X = 1) = \frac{4!}{1!3!} \times 0,5^1 \times (1 - 0,5)^3 = 0,2500$$

$$P(X = 2) = \frac{4!}{2!2!} \times 0,5^2 \times (1 - 0,5)^2 = 0,3750$$

$$P(X = 3) = 0,2500$$

$$P(X = 4) = 0,0625$$

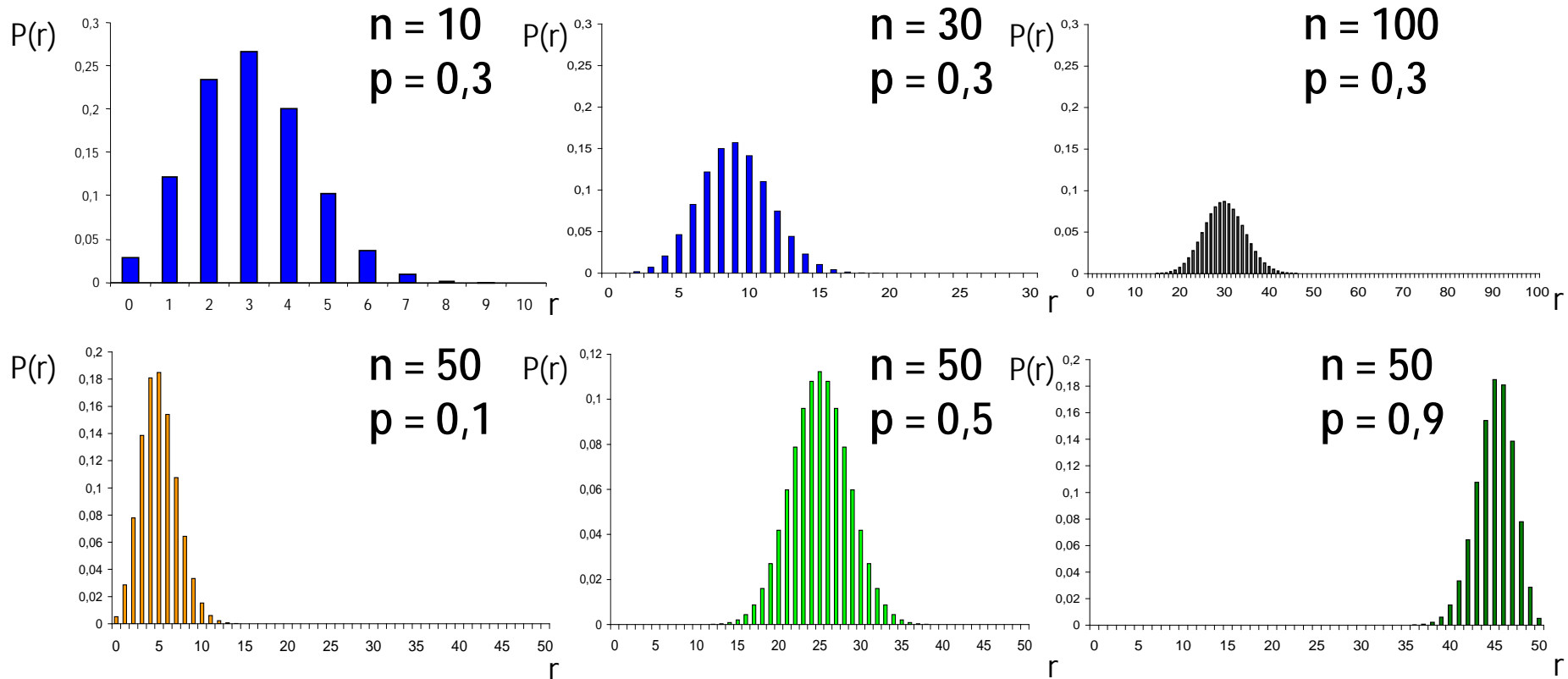
Nejpravděpodobnější počet chlapců – střední hodnota:  $E(X) = n \cdot p = 4 \cdot 0,5 = 2$





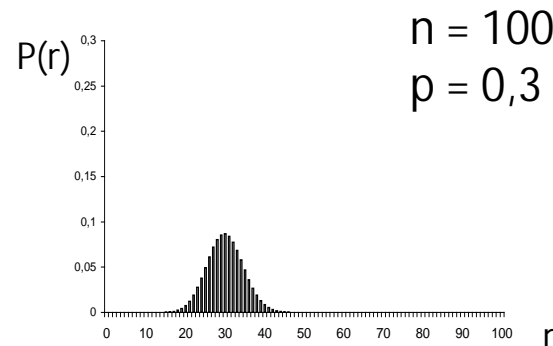
# Binomické rozdělení – tvar pro různé $n$ a $p$

- Čím vícekrát opakujeme experiment, tím menší relativní podíl připadá na jednotlivé hodnoty  $X$ , neboť všechny dohromady musí dát součet 1 (100%).
- Rozdělení s  $p=0,5$  je symetrické kolem středu osy  $x$ , menší či větší  $p$  posouvá střed rozdělení směrem k limitním hodnotám (tedy hodnotám 0 či  $n$ ).

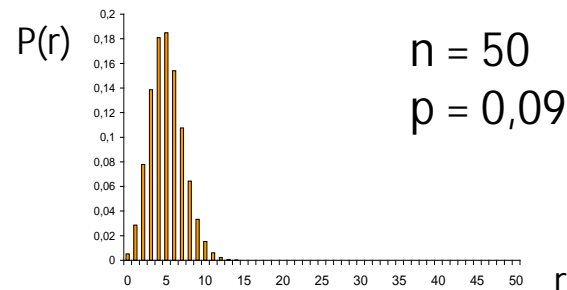


# Binomické rozložení – speciální případy

- Pokud  $n=1$ , jde o tzv. alternativní rozdělení a daná událost buď nenastane nebo nastane jednou.
- Pokud náhodný experiment opakujeme mnohokrát ( $n$  je velké), rozdělení se začne podobat spojitému rozdělení  $\rightarrow$  aproximace na normální rozdělení.



- Aproximace normálním rozdělením však nebude platit pro velmi nízké a velmi vysoké hodnoty  $p \rightarrow$  u nízkých hodnot  $p$  aproximace na Poissonovo rozdělení (pro  $n > 30$  a  $p < 0,1$ ).



# Binomické rozdělení - interval spolehlivosti - příklad

- Př. Sledování výskytu nežádoucích účinků u  $n = 100$  pacientů se schizofrenií léčených daným přípravkem. Nežádoucí účinky se vyskytly u 60 jedinců. Odhadněte pravděpodobnost výskytu nežádoucích účinků a tento odhad doplňte o 95% interval spolehlivosti.

- Vzorečky:

$$\hat{p} \approx p; \quad \hat{p} = r/n \quad (\text{bodový odhad parametru } \pi)$$

$$\hat{p} - Z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \leq p \leq \hat{p} + Z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \quad (\text{interval spolehlivosti pro } \pi)$$

- Řešení:

$$\hat{p} = 60/100 = 0,6$$

$$0,6 - 1,96 \times \sqrt{\frac{0,6 \times (1 - 0,6)}{100 - 1}} \leq p \leq 0,6 + 1,96 \times \sqrt{\frac{0,6 \times (1 - 0,6)}{100 - 1}}$$

$$0,6 - 1,96 \times 0,049 \leq p \leq 0,6 + 1,96 \times 0,049$$

$$0,503 \leq p \leq 0,697$$

- Pravděpodobnost výskytu nežádoucích účinků je 0,6 (0,503; 0,697).

# Binomické rozdělení – interval spolehlivosti

- **Ovlivnění šířky intervalu spolehlivosti (IS):**  $p \pm Z_{1-\alpha/2} \times \sqrt{\frac{p(1-p)}{n-1}}$ 
  - hodnotou  $p$  – IS bude nejširší pro  $p = 0,5$
  - hodnotou  $n$  – IS širší při malém  $n$  než při velkém
  - hodnotou  $\alpha$  – IS širší pro malé  $\alpha$  (hladinu spolehlivosti) – tzn. 99% IS bude širší než 95% IS
- **Interval spolehlivosti bez aproximace na normální rozdělení** (pokud hodnoty  $p$  jsou velmi nízké nebo velmi vysoké):

Dolní hranice IS:

$$D = \frac{r}{r + (n - r + 1) \times F_{\frac{\alpha}{2}}^{(n_1; n_2)}}$$

... kde:

$$n_1 = 2(n - r + 1); \quad n_2 = 2r$$

Horní hranice IS:

$$H = \frac{(r + 1) \times F_{\frac{\alpha}{2}}^{(n_1; n_2)}}{n - r + (r + 1) \times F_{\frac{\alpha}{2}}^{(n_1; n_2)}}$$

... kde:

$$n_1 = 2(r + 1) = n_2 + 2$$

$$n_2 = 2(n - r) = n_1 - 2$$

# 5. Poissonovo rozdělení

# Poissonovo rozdělení

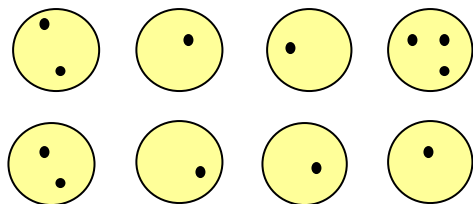
- Diskrétní rozdělení, které **popisuje počet výskytů sledované události na danou jednotku** (času, plochy, objemu), když se tyto události vyskytují vzájemně nezávisle s konstantní intenzitou (parametr  $\lambda$ ).
- Značení:  $Po(\lambda)$
- Jedná se o zobecnění binomického rozdělení pro  $n \in \mathbb{N}$  a  $p \in [0, 1]$  (aproximace je funkční již při  $n > 30$ ,  $p < 0,1$ ):  $Bi(n, p) \approx Po(n \times p)$   
Pravděpodobnost, že sledovaná událost nastane  $r$ -krát, lze vypočítat:

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

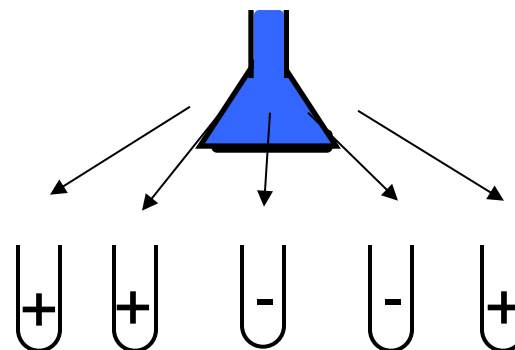
- **Střední hodnota:**  $EX = \lambda$  ( $\lambda$  vyjadřuje střední počet jevů na jednu experimentální jednotku)
- **Rozptyl:**  $DX = \lambda$
- **Příklady:** počet krvinek v poli mikroskopu, počet pooperačních komplikací během určitého časového intervalu po výkonu, počet pacientů, kteří přišli do ordinace během jedné hodiny, počet částic, které vyzáří zářič za danou časovou jednotku

# Poissonovo rozdělení – příklady

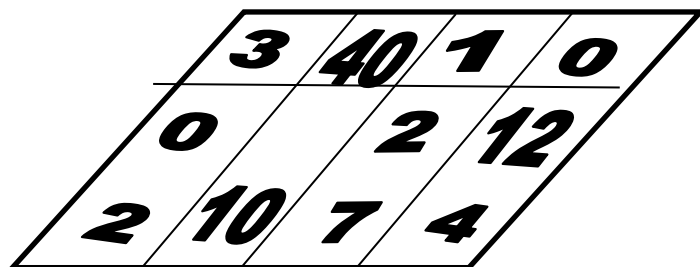
**Výskyt jevu na experimentální jednotku**  
(mutace bakterií na inkubačních miskách)



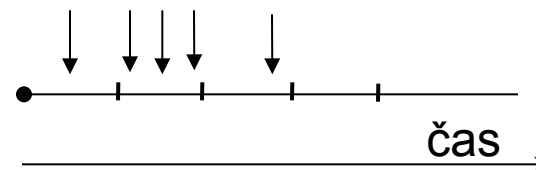
**Orientační stanovení jevu**  
(např. produkce plynu bakteriemi)



**Výskyt jevu v prostoru**  
(počet buněk v sčítacím poli preparátu)



**Výskyt jevu v čase**  
(vyzáření částice v určitých časových intervalech)



# Poissonovo rozdělení – příklad

- **Příklad:** Předpokládejme, že v určité populaci krys se vyskytuje albín s pravděpodobností  $\pi=0,001$ , ostatní krys jsou normálně pigmentované. Ve vzorku 100 krys náhodně vybraných z této populace určete pravděpodobnost, že vzorek a) neobsahuje albína, b) obsahuje právě jednoho albína.
- **Řešení:** Pravděpodobnost výskytu albína je  $\pi=0,001$ . Předpokládaný počet albínů ve výběru o rozsahu  $n$  je  $\lambda=n \cdot \pi$  (průměr binomické náhodné veličiny), tj. v našem příkladu  $\lambda=n \cdot \pi=100 \cdot 0,001=0,1$ . Počet albínů označme  $x$ . Potom:

$$\begin{aligned} \text{a) pro } x = 0 \text{ máme } & \frac{e^{-0,1} \cdot 0,1^0}{0!} = \frac{e^{-0,1} \cdot 1}{1} = 0,9048, \\ \text{b) pro } x = 1 \text{ máme } & \frac{e^{-0,1} \cdot 0,1^1}{1!} = \frac{e^{-0,1} \cdot 0,1}{1} = 0,09048. \end{aligned}$$

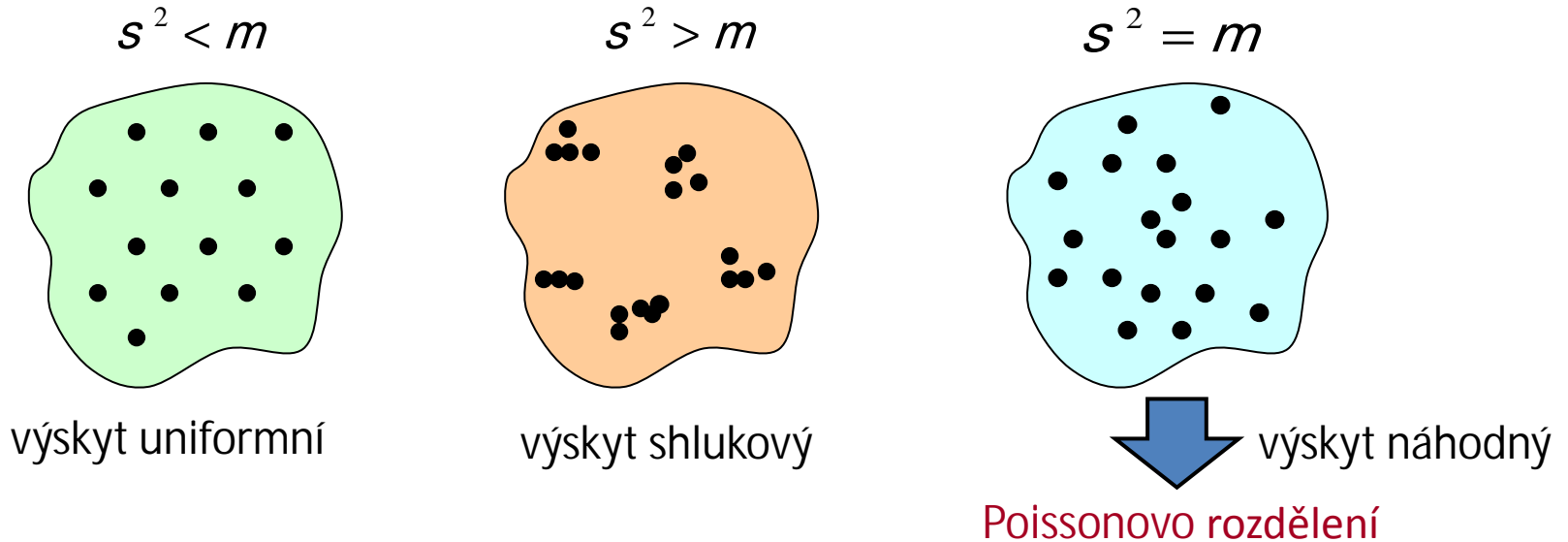
- Jak je vidět, pravděpodobnost, že ve vzorku 100 krys nebude žádný albín, je desetkrát vyšší než pravděpodobnost, že ve vzorku bude právě jeden albín. Pravděpodobnosti výskytu dvou a více albínů jsou již velmi malé.

Převzato z: Zvárová, J. (2001) *Základy statistiky pro biomedicínské obory*. Praha: Karolinum.



# Poissonovo rozdělení – předpoklady

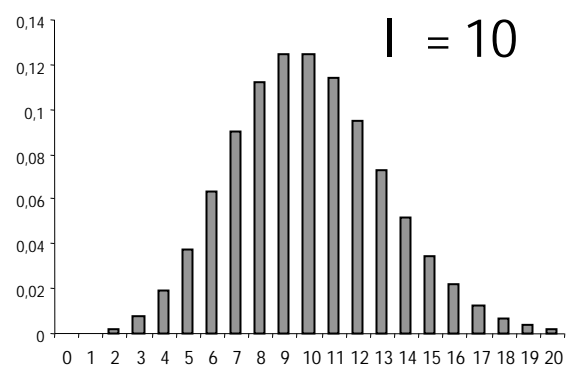
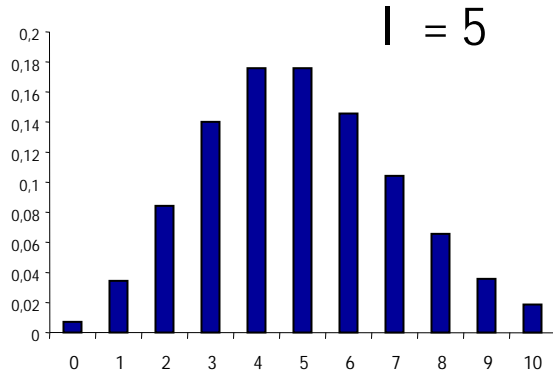
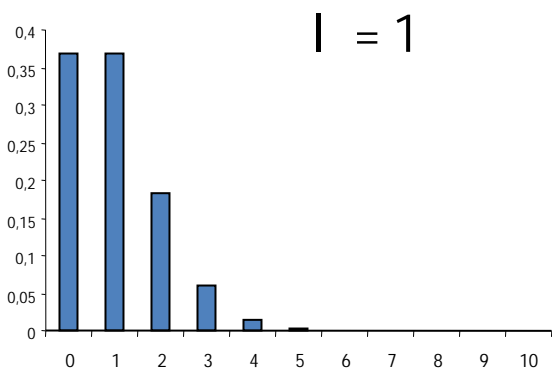
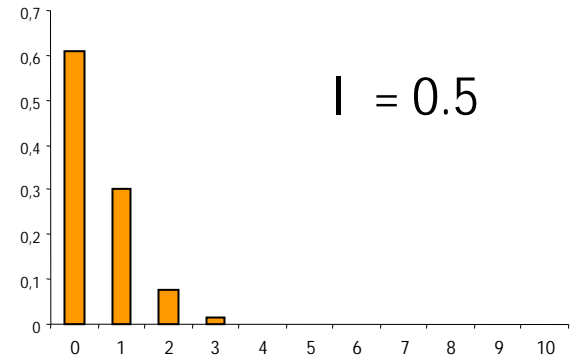
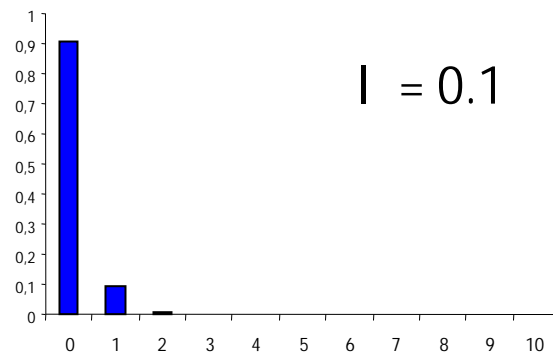
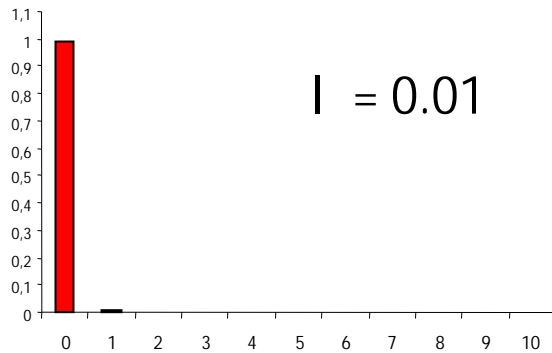
- výskyt jevu je zcela náhodný (tedy náhodný v čase nebo prostoru podle typu situace)



- výskyt jevu v konkrétní experimentální jednotce nijak nezávisí na tom, co se stalo v jiných jednotkách
- není možné, aby 2 nebo více jevů nastaly současně, přesně ve stejném místě prostoru nebo ve stejném časovém okamžiku
- pro každý dílčí časový okamžik, prostoru jednotku apod. je pravděpodobnost výskytu stejná

# Poissonovo rozdělení – tvar pro různé $\lambda$

- Čím větší je  $\lambda$ , tím více se tvar Poissonova rozdělení blíží normálnímu rozdělení.



# Poissonovo rozdělení – intervaly spolehlivosti - příklad

- **Př. Za 10 hodin vyzářil zářič 1500 částic. Spočtete průměrný počet vyzářených částic za hodinu a tento odhad průměrného počtu částic doplňte o 95% interval spolehlivosti.**

- Vzorečky:

$l \gg \bar{x}$  (bodový odhad parametru  $\lambda$ )

$$\bar{x} - Z_{1-\alpha/2} \times \sqrt{\frac{\bar{x}}{n}} \text{ £ } / \text{ £ } \bar{x} + Z_{1-\alpha/2} \times \sqrt{\frac{\bar{x}}{n}} \quad (\text{interval spolehlivosti pro } \lambda)$$

- Řešení:

$$\bar{x} = 1500 / 10 = 150$$

$$150 - 1,96 \times \sqrt{\frac{150}{10}} \text{ £ } / \text{ £ } 150 + 1,96 \times \sqrt{\frac{150}{10}}$$

$$150 - 1,96 \times 3,873 \text{ £ } / \text{ £ } 150 + 1,96 \times 3,873$$

$$142 \text{ £ } / \text{ £ } 158$$

- Průměrný počet částic vyzářených za hodinu je 150 (142;158).

# Poissonovo rozdělení – interval spolehlivosti

- **Ovlivnění šířky intervalu spolehlivosti (IS):**  $\bar{x} \pm Z_{1-\alpha/2} \times \sqrt{\frac{\bar{x}}{n}}$ 
  - hodnotou  $\lambda$  – IS širší při velkém  $\lambda$
  - hodnotou  $n$  – IS širší při malém  $n$  než při velkém
  - hodnotou  $\alpha$  – IS širší pro malé  $\alpha$  (hladinu spolehlivosti) – tzn. 99% IS bude širší než 95% IS

- **Interval spolehlivosti bez aproximace na normální rozdělení:**

Dolní hranice IS:

$$D = \frac{C_{\alpha/2}^2(n_1)}{2}$$

... kde:  
 $n_1 = 2r$

Horní hranice IS:

$$H = \frac{C_{1-\alpha/2}^2(n_2)}{2}$$

... kde:  
 $n_2 = n_1 + 2 = 2r + 2$

# Poděkování...

Příprava výukových materiálů předmětu „DSAN01 Analýza dat pro Neurovědy “ byla finančně podporována prostředky projektu FRVŠ č. 942/2013 „Inovace materiálů pro interaktivní výuku a samostudium předmětu Analýza dat pro Neurovědy“

