Analýza dat pro Neurovědy

RNDr. Eva Janoušová

doc. RNDr. Ladislav Dušek, Dr.

Jaro 2013

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. Relativní riziko („relative risk“) a poměr šancí („odds ratio“)
3. Binomické rozdělení
4. Poissonovo rozdělení

1. Analýza kontingenčních tabulek

Kontingenční tabulka

- Frekvenční sumarizace dvou binárních, nominálních nebo ordinálních proměnných.
- Obecně: **R x C kontingenční tabulka** (R – počet kategorií jedné proměnné, C – počet kategorií druhé proměnné).
- Speciální případ: 2 × 2 tabulka = čtyřpolní tabulka.

- Př.: 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 – hypotézy

- Kontingenční tabulky umožňují testování různých hypotéz:
- **Nezávislost** (Pearsonův chí-kvadrát test)
  - Jeden výběr, dvě charakteristiky – obdoba nepárového uspořádání
  - Př.: pacienti s AD – pohlaví × vzdělání (VŠ, SŠ, ZŠ)
- **Shoda struktury** (Pearsonův chí-kvadrát test)
  - Více výběrů, jedna charakteristika – obdoba nepárového uspořádání
  - Př.: pacienti s AD v několika nemocnicích × věková struktura
- **Symetrie** (McNemarův test)
  - Jeden výběr, opakovaně jedna charakteristika – obdoba párového uspořádání
  - Př.: 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:

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

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

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

- 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 α, když $C^2 \geq c^2_{(1-a)}(r-1)(c-1)$

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

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

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

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

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

$C^2 = 69{,}4 > c^2_{(0,95)}(3 - 1)(4 - 1) = c^2_{(0,95)}(6) = 12{,}6$ → **zamítáme H_0** o nezávislosti → 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.

Čtyřpolní tabulky

- Nejjednodušší možná kontingenčí tabulka, kdy obě sledované veličiny mají pouze dvě kategorie.

- **Příklad**: Sumarizace vztahu pohlaví a kategorizovaného MMSE skóre (MMSE skóre v normě (tzn. MMSE ≥ 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

Asociace ve čtyřpolní tabulce

- **Můžeme rozhodovat o závislosti/nezávislosti dvou sledovaných veličin**.
- **Můžeme rozhodovat i o míře (těsnosti) této závislosti** – relativní riziko, poměr šancí.

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

- Při rozhodování o nezávislosti můžeme použít **Pearsonův chí-kvadrát test**, ale pro malá n je standardem v klinických analýzách tzv. **Fisherův exaktní test** („Fisher exact test“).

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	NÚ II: 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 ≥ 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: 80 % očekávaných četností větších než 5 – u čtyřpolní tabulky to znamená 100 %.

- **Nedodržení předpokladů pro Pearsonův chí-kvadrát test může stejně jako u t-testu a analýzy rozptylu vést k nesmyslným závěrům!**

- Situace s malými n_{ij} a tedy i e_{ij} jsou ale v medicíně i biologii velmi časté – **Fisherův exaktní test je klíčový pro hodnocení čtyřpolních tabulek**.

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
Pearson Chi-square	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		
	$Y = 1$	$Y = 2$	Celkem
$X = 1$	a	b	$a + b$
$X = 2$	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

- 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.
- **Řešení:**

rozdílné výsledky

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

Statistic	Statistics: mmse_kat(2) x mmse24_kat(2) (Data_neuro_vycistena3)		
	Chi-square	df	p
Pearson Chi-square	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. 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		
	Do 25 let	25 a více let	Celkem
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.**

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		
	Experimentální	Kontrolní	Celkem
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		
	Do 25 let	25 a více let	Celkem
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

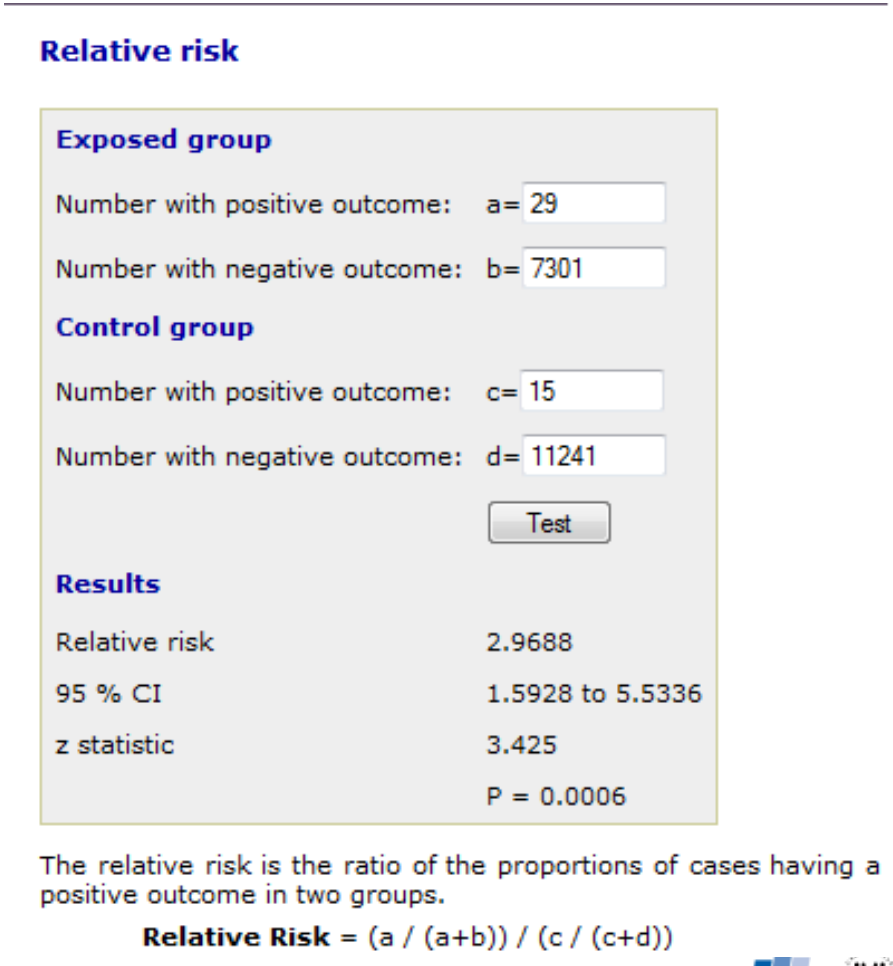
$$RR = \frac{P_1}{P_0} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{29}{29+7301}}{\frac{15}{15+11241}} = 2{,}97$$

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

Relativní riziko

- Výpočet pomocí webového kalkulátoru (http://www.medcalc.org/calc/relative_risk.php):

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



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**

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

Sledovaný jev	Skupina		
	Experimentální	Kontrolní	Celkem
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		
	Do 25 let	25 a více let	Celkem
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

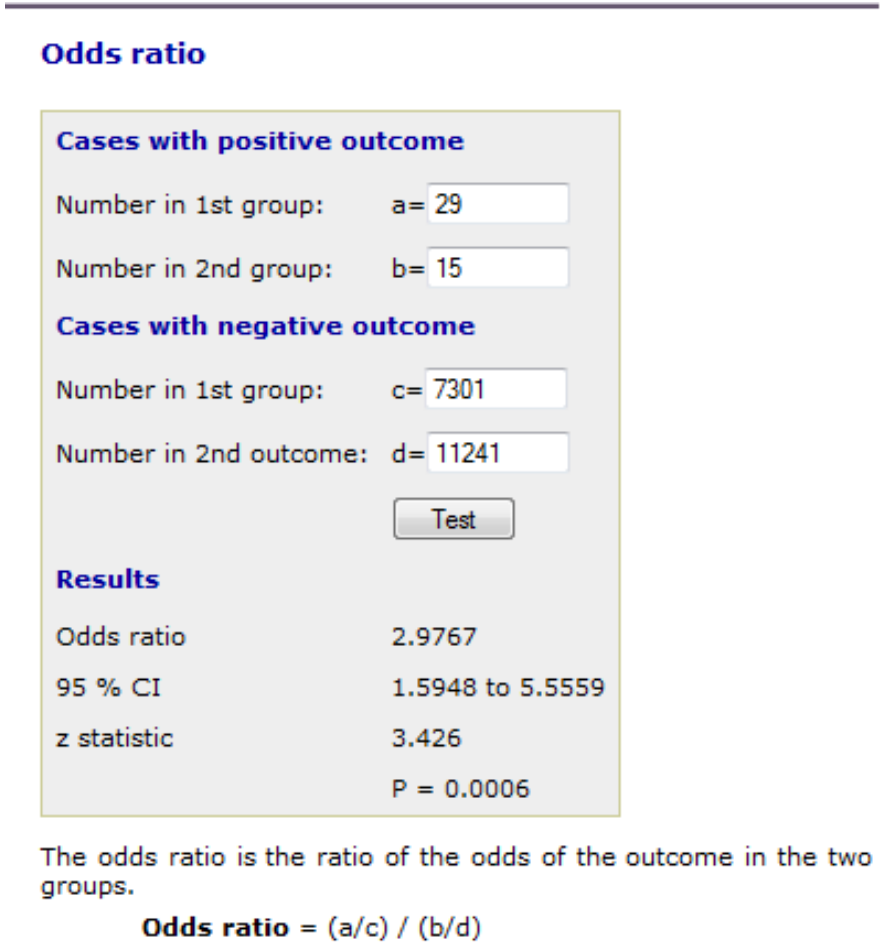
$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{29}{7301}}{\frac{15}{11241}} = 2{,}98$$

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

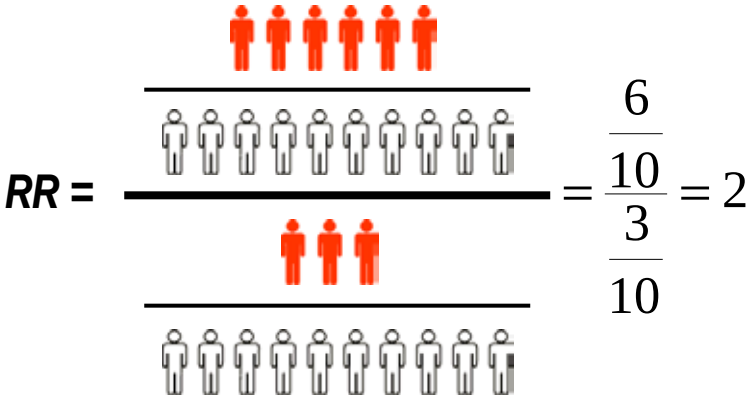
Poměr šancí

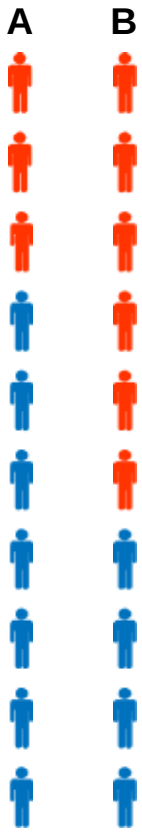
- Výpočet pomocí webového kalkulátoru (http://www.medcalc.org/calc/odds_ratio.php):

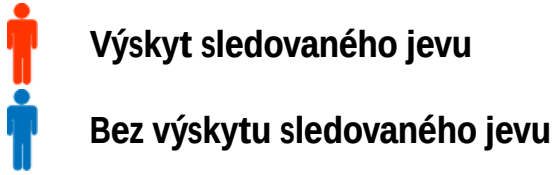
SIDS	Věk matky		
	Do 25 let	25 a více let	Celkem
Ano	29	15	44
Ne	7301	11241	18542
Celkem	7330	11256	18586

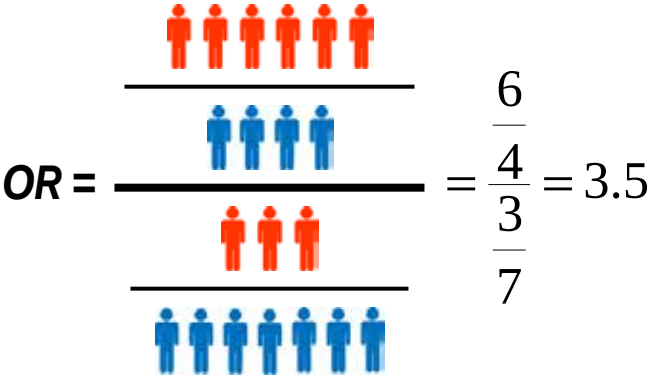


Grafické srovnání *RR* a *OR*









Úkol 3.

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

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

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

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

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

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

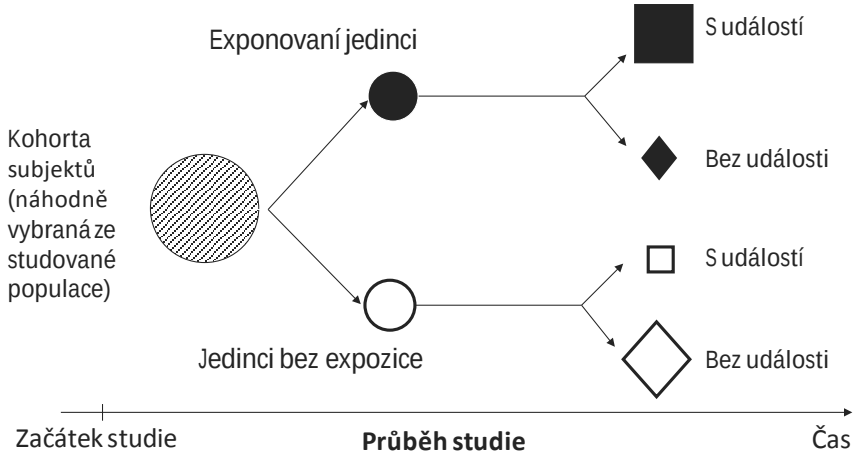
Výhody a nevýhody *RR* a *OR*

- Nevýhoda *OR*:
  - obtížná interpretace.

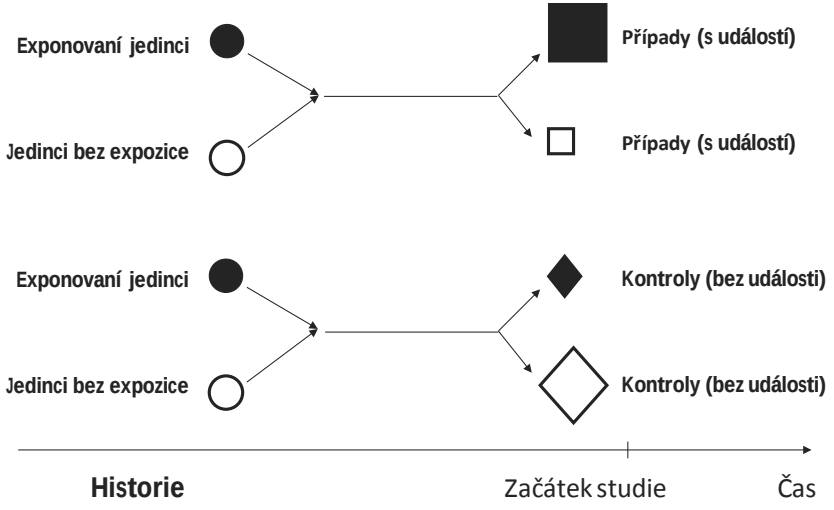
- Výhoda i nevýhoda *RR*:
  - nezajímá ho samotná pravděpodobnost výskytu jevu, ale pouze jejich podíl → korektní použití *RR* je však pouze v případě, že pravděpodobnost výskytu jevu v kontrolní skupině je reprezentativní (není ovlivněna výběrem sledovaných subjektů).

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

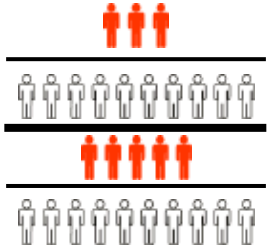
- Pomocí RR i OR můžeme srovnat pravděpodobnosti výskytu sledovaného jevu ve dvou různých skupinách:

- **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

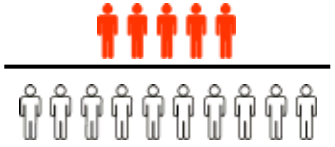
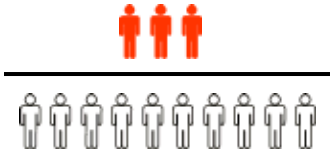
Další způsoby vyjádření rozdílu rizika

- **Relativní redukce rizika** (RRR)

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

- **Absolutní redukce rizika** (ARR)

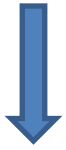
Bez léčby **S léčbou**

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

Další způsoby vyjádření rozdílu rizika

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

ARR = 20% ⟶ Pro snížení počtu událostí o 20 je třeba léčit 100 pacientů.



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 %.

→ **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 %.

→ **NNT =** $\frac{1}{0{,}006} = \frac{100}{0{,}6} = 166{,}7$

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

3. 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,π)
- Parametry:

  n ... počet nezávislých pokusů

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

  p = r/n ... pravděpodobnost nastání sledované události (p ~π)
- 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 neuropsychologického 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

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

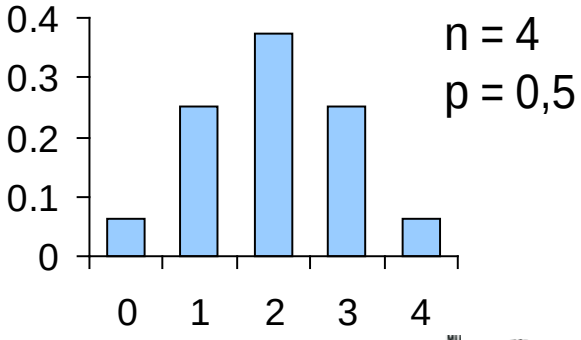
$$P(X = 0) = \frac{4!}{0!\,4!} \times 0{,}5^0 \times (1 - 0{,}5)^4 = 0{,}0625$$

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

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

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

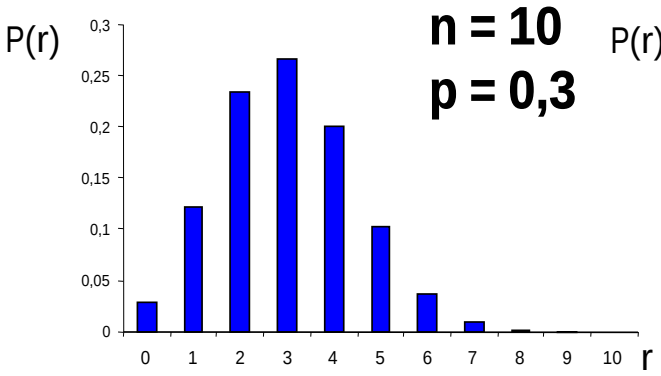
$$P(X = 4) = 0{,}0625$$

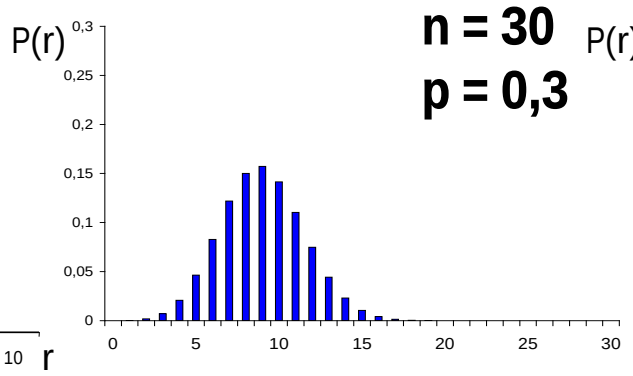


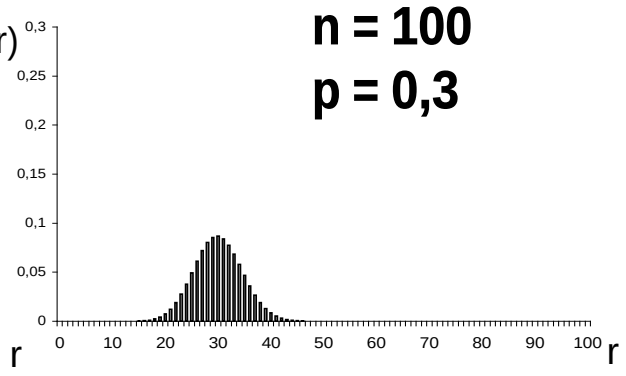
Nejpravděpodobnější počet chlapců – střední hodnota: $E(X) = n \cdot p = 4 \cdot 0{,}5 = 2$

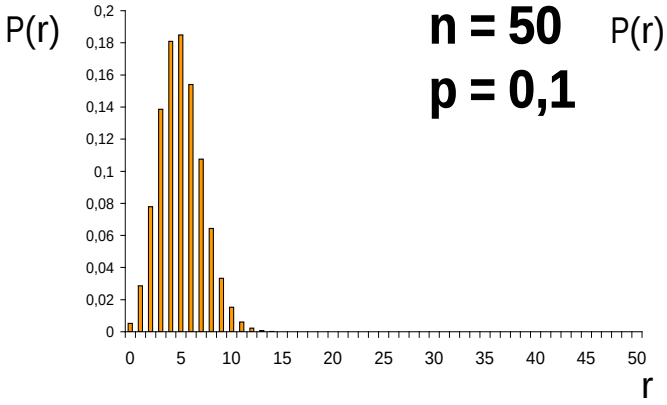
Binomické rozdělení – tvar pro různé *n* a *p*

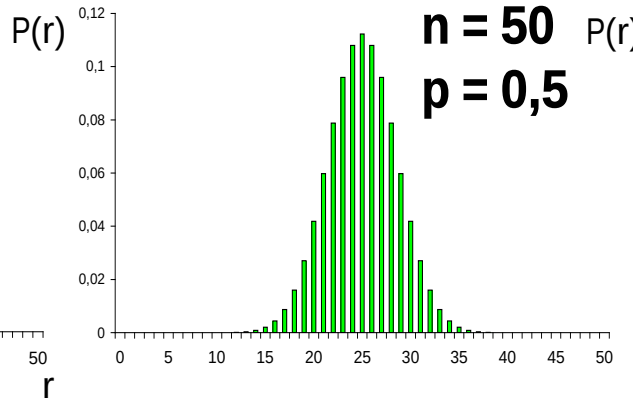
- Čím vícekrát opakujeme experiment, tím menší relativní podíl připadá na jednotlivé hodnoty X, neboť všechny dohromady musí dát součet 1 (100%).
- Rozdělení s *p*=0,5 je symetrické kolem středu osy *x*, menší či větší *p* posouvá střed rozdělení směrem k limitním hodnotám (tedy hodnotám 0 či *n*).

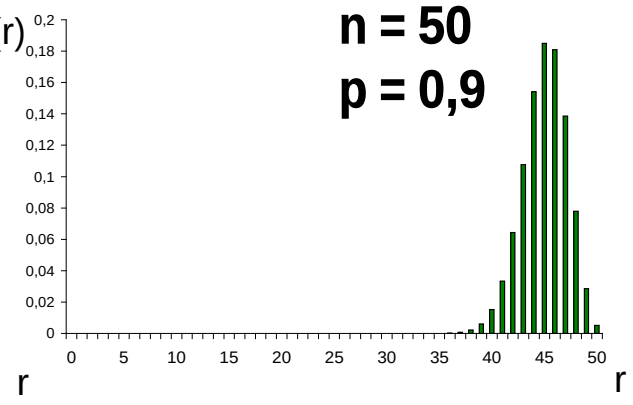






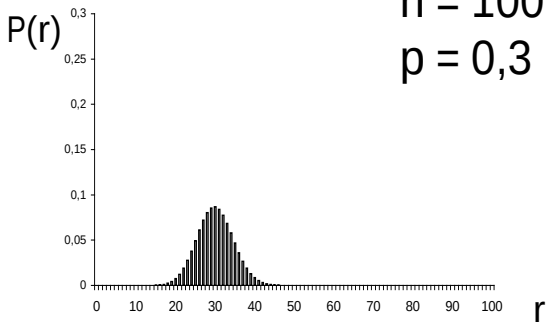






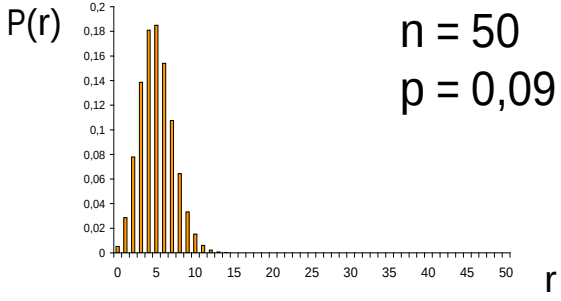
Binomické rozložení – speciální případy

- 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 → aproximace na normální rozdělení.



n = 100
p = 0,3

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



n = 50
p = 0,09

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:

$$p \approx p; \quad p = r/n$$ (bodový odhad parametru π)

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

- Řešení:

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

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

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

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

- 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 α – IS širší pro malé α (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:

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

... kde:

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

Horní hranice IS:

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

... kde:

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

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

Statistické testování binomických dat

1. **Liší se odhad *p* od předpokládané (referenční) hodnoty π?**
   (Např. liší se procento pacientů s nežádoucími účinky léčby od předpokládaného procenta?)
   → **jednovýběrový binomický test** (tzn. test pro podíl u jednoho výběru)

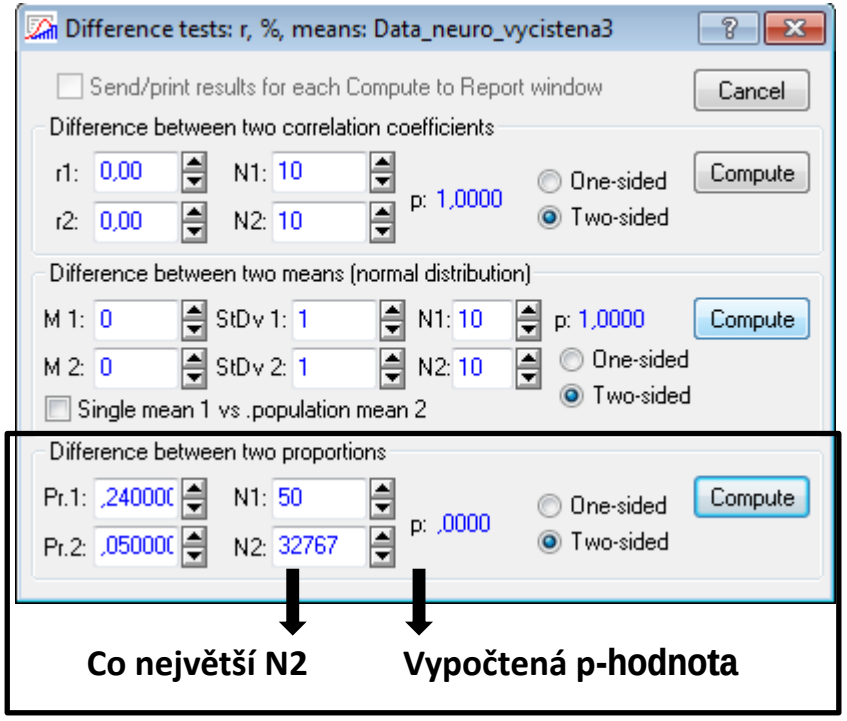
2. **Liší se *p* ve dvou souborech?**
   (Např. liší se podíl pacientů s nežádoucími účinky léčby podle typu léčby?)
   → **dvouvýběrový binomický test** (tzn. test pro podíl u dvou výběrů)

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: $H_0 : p = p$ a $H_1 : p \neq p$

- **Řešení:**
- π = 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.

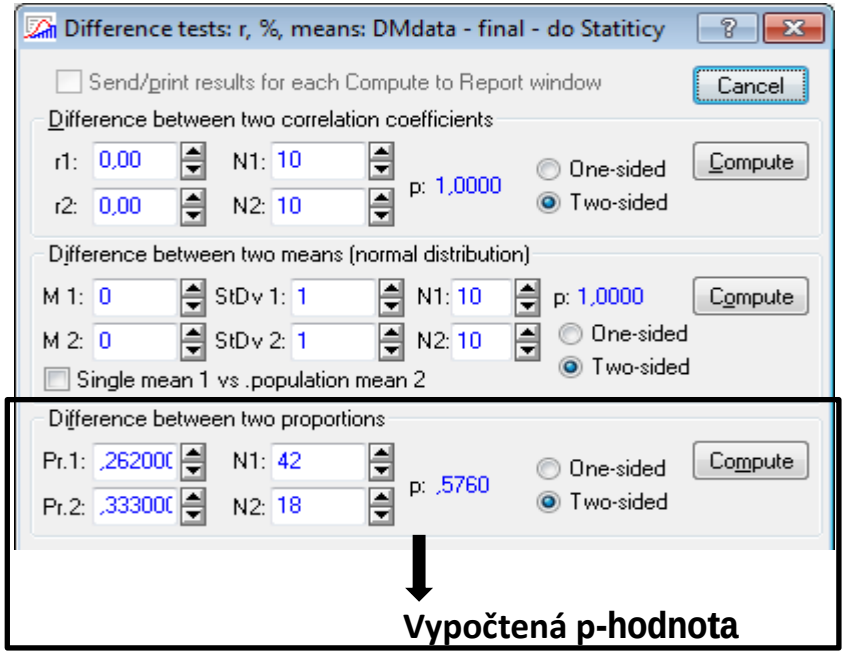


Dvouvýběrový binomický test

- **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.



4. Poissonovo rozdělení

Poissonovo rozdělení

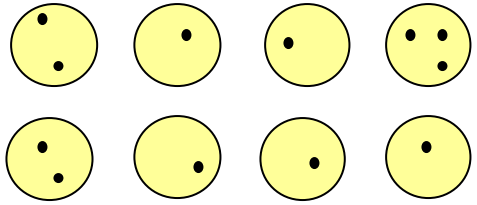
- Diskrétní rozdělení, které **popisuje počet výskytů sledované události na danou jednotku** (času, plochy, objemu), když se tyto události vyskytují vzájemně nezávisle s konstantní intenzitou (parametr λ).
- Značení: Po(λ)
- Jedná se o zobecnění binomického rozdělení pro $n \to \infty$ a $p \to 0$
 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

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

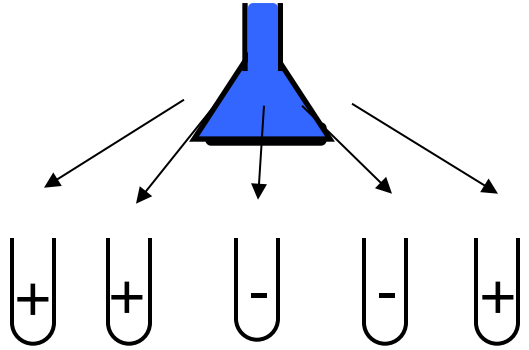
- **Střední hodnota:** EX = λ (λ vyjadřuje střední počet jevů na jednu experimentální jednotku)
- **Rozptyl:** DX = λ
- Příklady: počet krvinek v poli mikroskopu, počet pooperačních komplikací během určitého časového intervalu po výkonu, počet pacientů, kteří přišli do ordinace během jedné hodiny, počet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

Poissonovo rozdělení – příklady

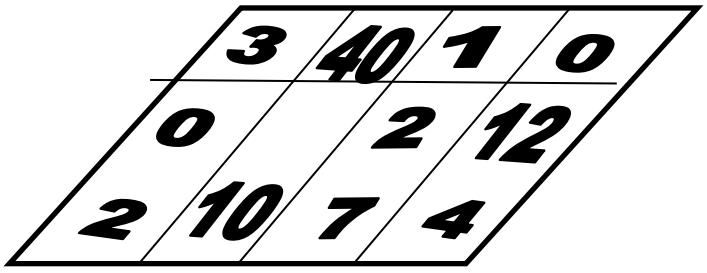
Výskyt jevu na experimentální jednotku
(mutace bakterií na inkubačních miskách)



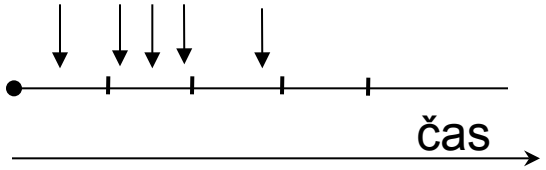
Orientační stanovení jevu
(např. produkce plynu bakteriemi)



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

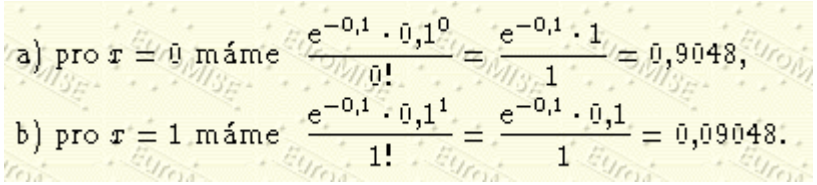


Výskyt jevu v čase
(vyzáření částice v určitých časových intervalech)



Poissonovo rozdělení – příklad

- **Příklad:** Předpokládejme, že v určité populaci krys se vyskytuje albín s pravděpodobností π=0,001, ostatní krysy 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 π=0,001. Předpokládaný počet albínů ve výběru o rozsahu *n* je λ=n*π (průměr binomické náhodné veličiny), tj. v našem příkladu λ=n*π=100*0,001=0,1. Počet albínů označme *x*. Potom:

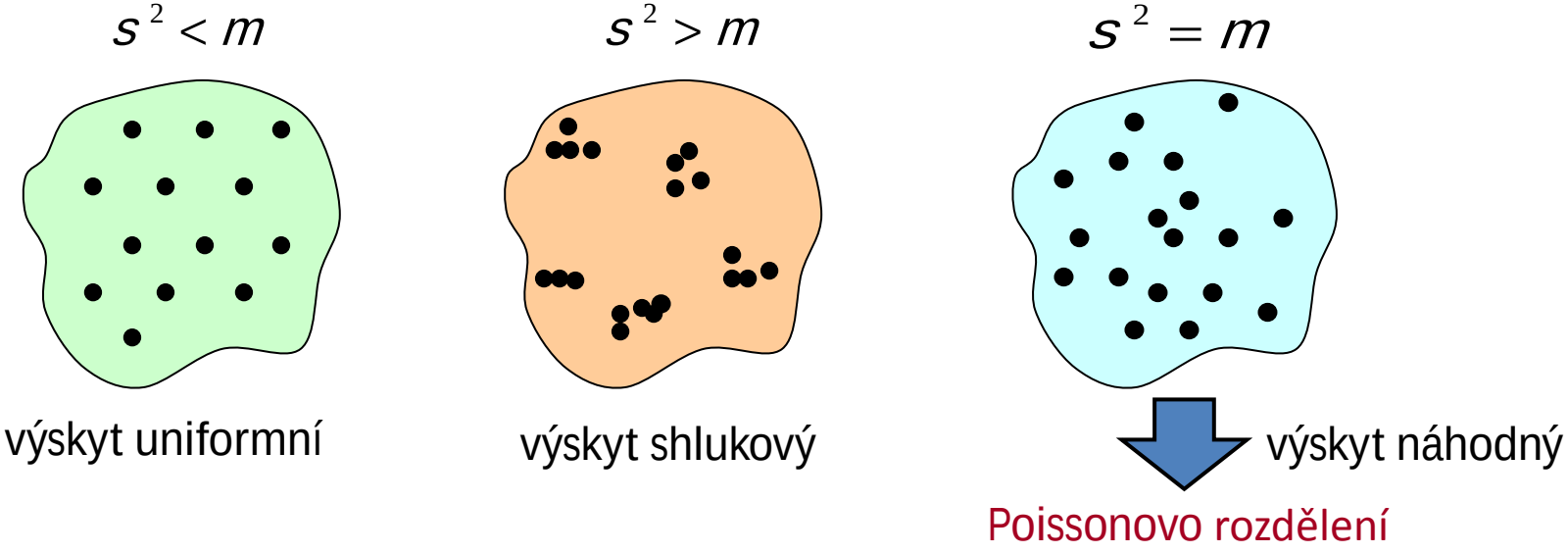


- Jak je vidět, pravděpodobnost, že ve vzorku 100 krys nebude žádný albín, je desetkrát vyšší než pravděpodobnost, že ve vzorku bude právě jeden albín. Pravděpodobnosti výskytu dvou a více albínů jsou již velmi malé.

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

Poissonovo rozdělení – předpoklady

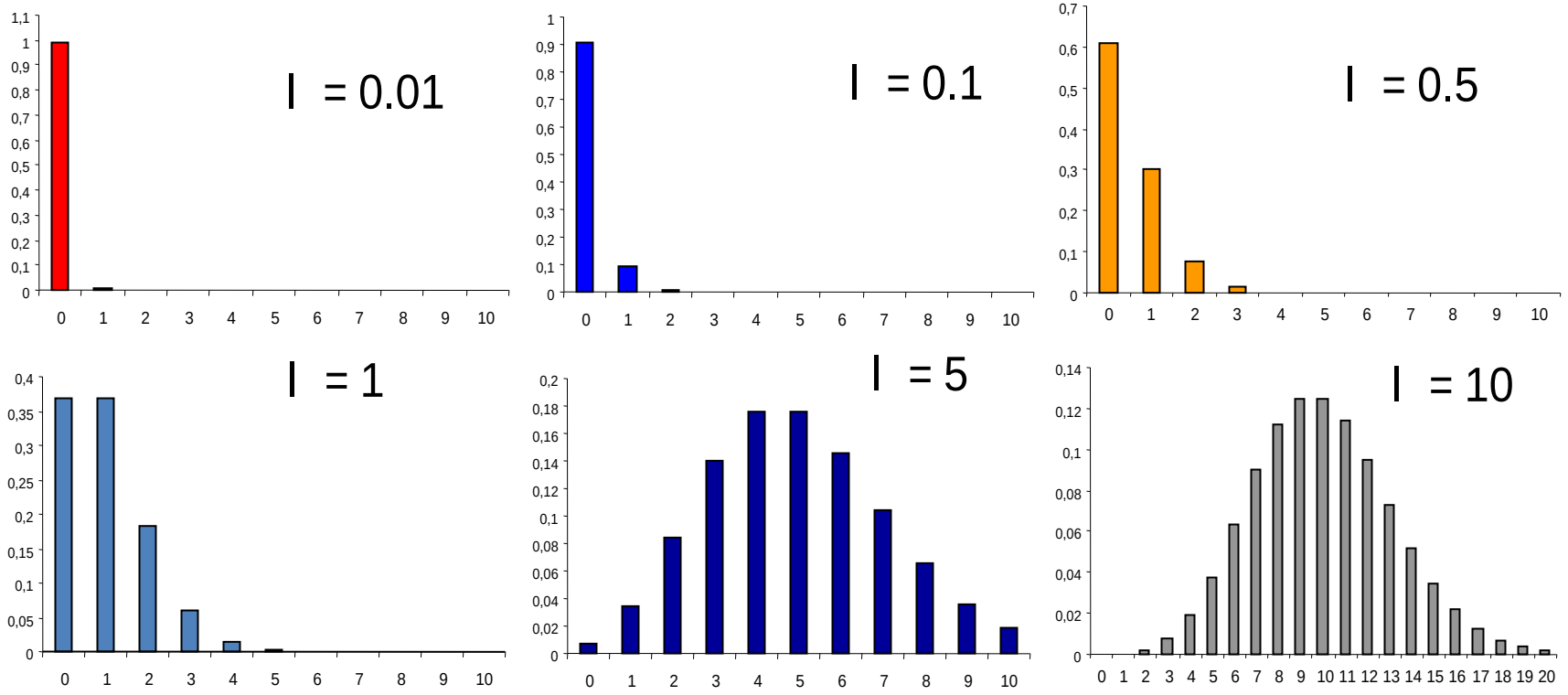
- výskyt jevu je zcela náhodný (tedy náhodný v čase nebo prostoru podle typu situace)



- 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, prostorou jednotku apod. je pravděpodobnost výskytu stejná

Poissonovo rozdělení – tvar pro různé λ

- Čím větší je λ, tím více se tvar Poissonova rozdělení blíží normálnímu rozdělení.



Poissonovo rozdělení – intervaly spolehlivosti - příklad

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

- Vzorečky:

  $\lambda \approx \bar{x}$ (bodový odhad parametru λ)

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

- Řešení:

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

$$150 - 1{,}96 \times \sqrt{\frac{150}{10}} \leq \lambda \leq 150 + 1{,}96 \times \sqrt{\frac{150}{10}}$$

$$150 - 1{,}96 \times 3{,}873 \leq \lambda \leq 150 + 1{,}96 \times 3{,}873$$

$$142 \leq \lambda \leq 158$$

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

Poissonovo rozdělení – interval spolehlivosti

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

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

Dolní hranice IS:

$$D = \frac{c^2_{a/2}(n_1)}{2}$$

... kde:

$$n_1 = 2r$$

Horní hranice IS:

$$H = \frac{c^2_{1-a/2}(n_2)}{2}$$

... kde:

$$n_2 = n_1 + 2 = 2r + 2$$

Poděkování…

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