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

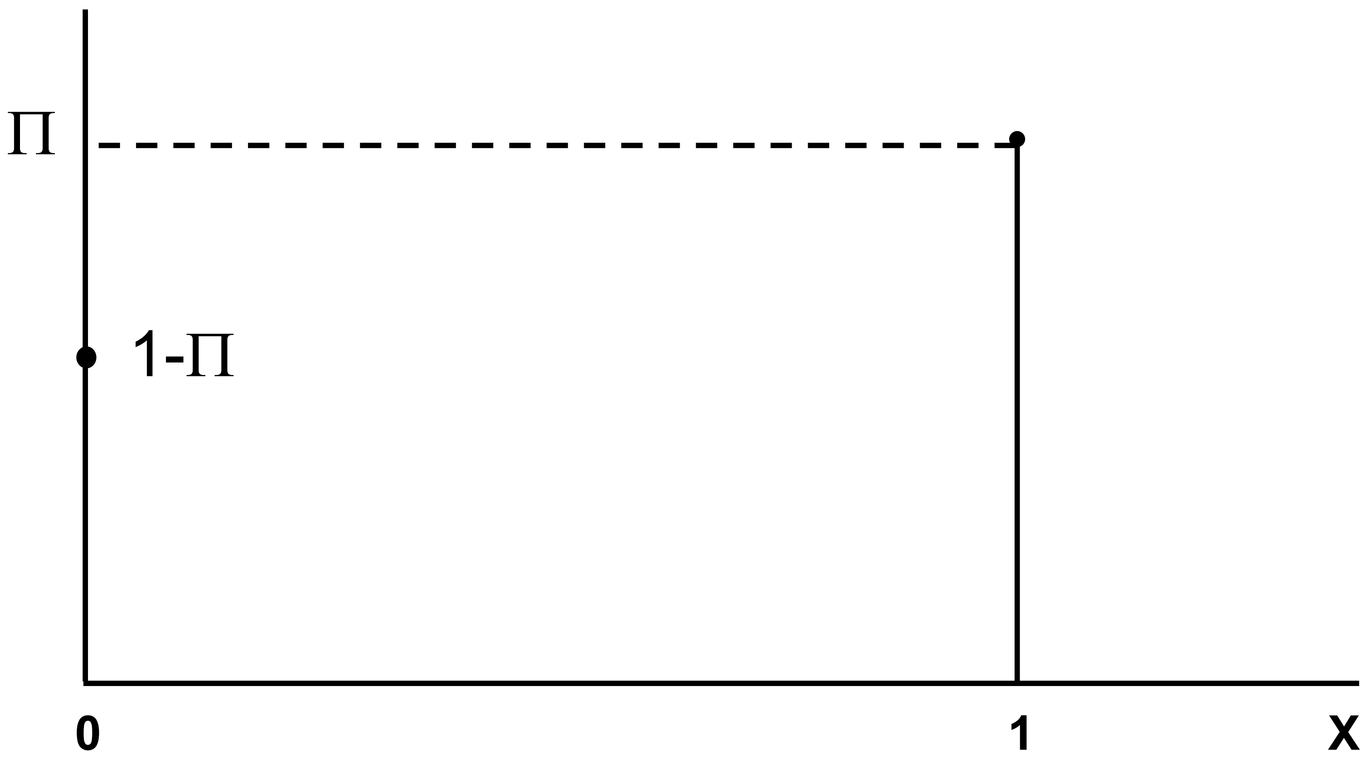
- Analýzy na binomickém rozložení
- Analýzy na Poissonově rozložení
- Analýza kontingenčních tabulek

Binomické rozložení

Alternativní rozložení

$\Pi(x) = \Pi$ pro $X = 1$
$\Pi(x) = 1 - \Pi$ pro $X = 0$
$\Pi(x) = 0$ jinak

$X = 1$ ......jev

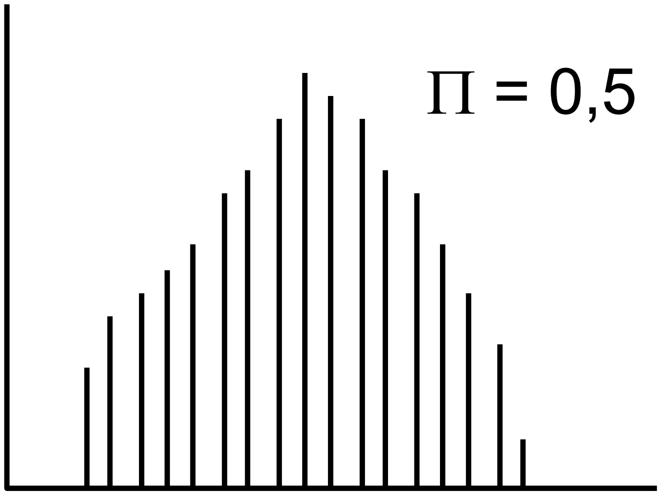


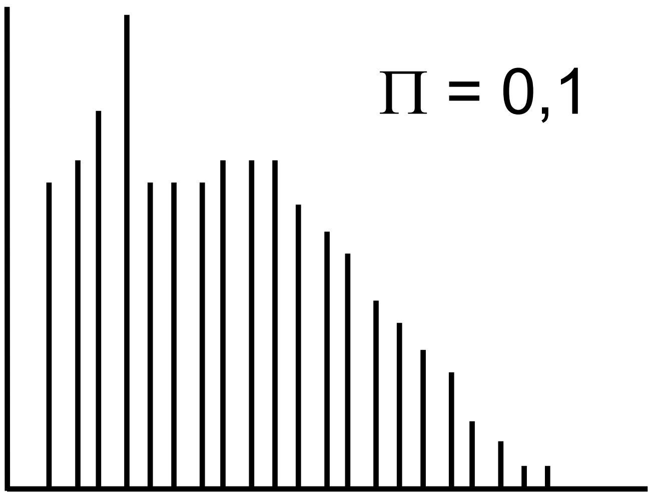
Binomické rozložení

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

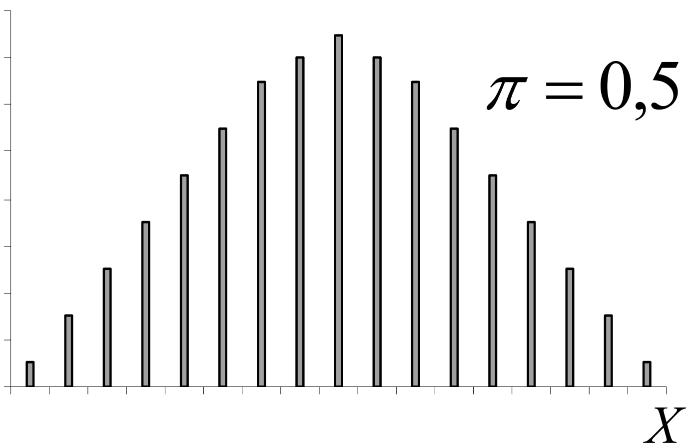
r : **0 …… n**

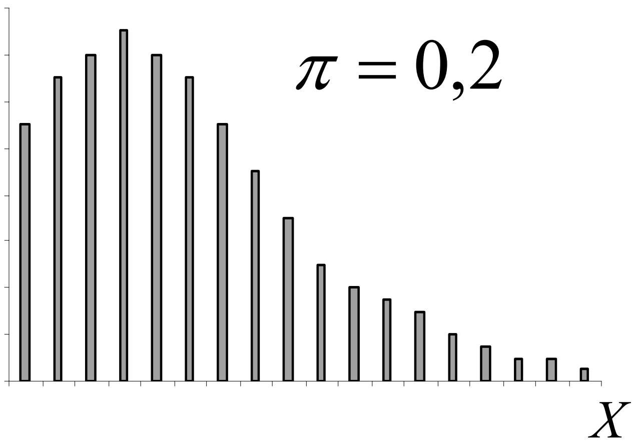
p ~ π .. jediný parametr binomického rozložení

p .... relativní četnost nastání jevu

p ……… určuje tvar distribuce

$$\mathbf{p} = \mathbf{r}/\mathbf{n}$$





Binomická proměnná X

Binomické rozložení jako model

Jev: narození chlapce Π = 0,5
n : rodina s 5 dětmi
r: 0,1,2,3,4,5 chlapců

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

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:

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

Rozptyl:

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

Příklad: n = 100 respondentů
r = 20 má symptom

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

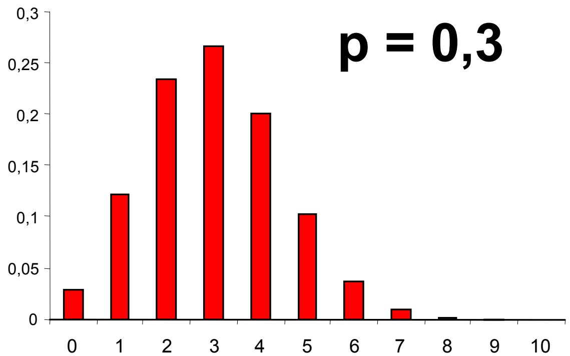
je střed rozložení
a nejpravděpodobnější
hodnota

Binomické rozložení jako model

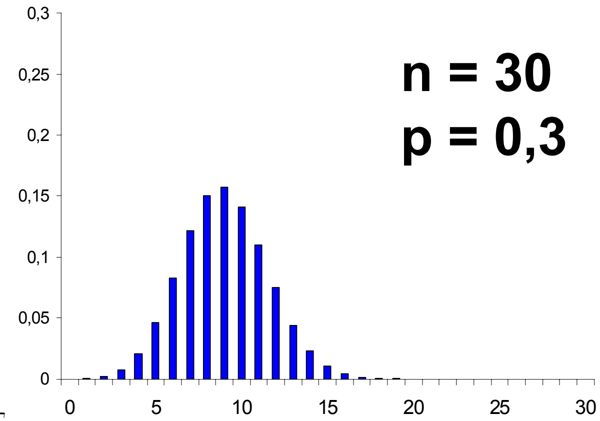
$$P(x = r) = \frac{n!}{r!\,(n-r)!} \cdot p^{r} \cdot q^{(n-r)}$$

$q = 1 - p$

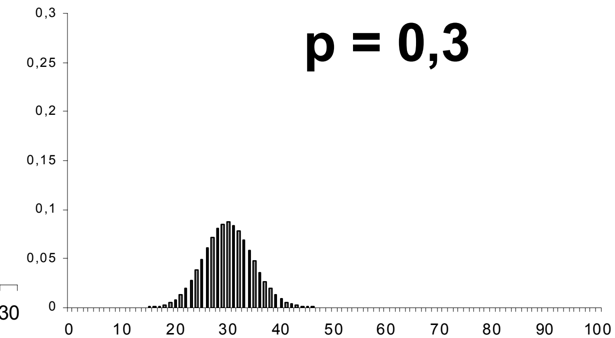
n = 10
p = 0,3



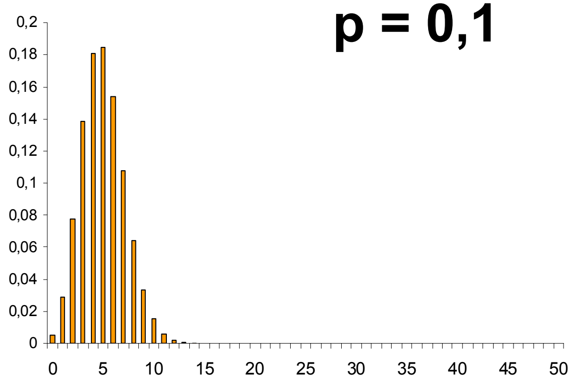
n = 30
p = 0,3



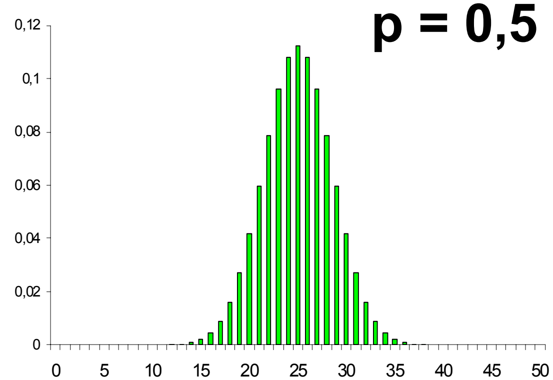
n = 100
p = 0,3



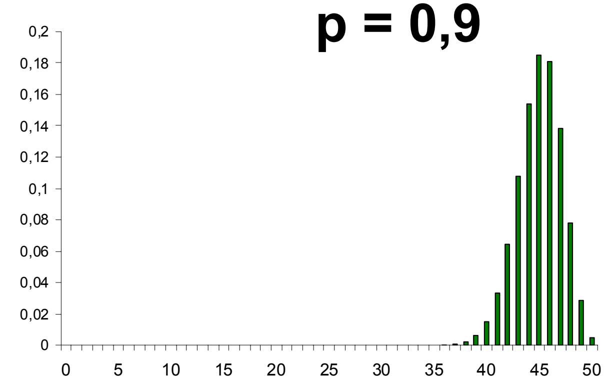
n = 50
p = 0,1



n = 50
p = 0,5



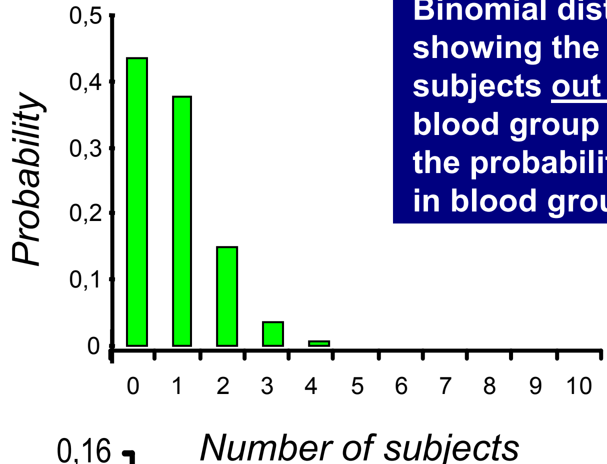
n = 50
p = 0,9



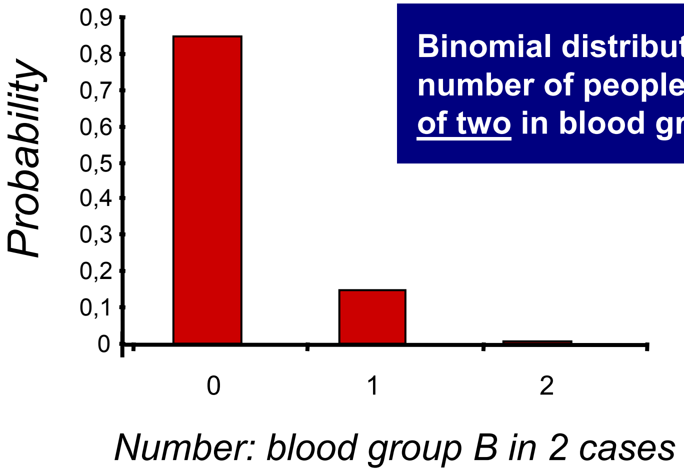
Aplikace binomického rozložení

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

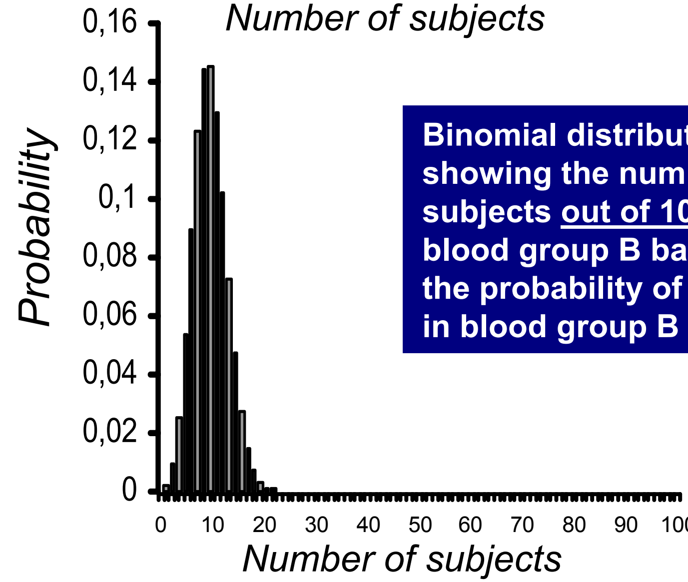
		Number in blood group B	Probability
B	**B**	**2**	**0,0064**
not B	**B**	**1**	**0,0736**
B	**not B**	**1**	**0,0736**
not B	**not B**	**0**	**0,8464**



Binomial distribution showing the number of subjects out of ten in blood group B based on the probability of being in in blood group B of 0,08.



Binomial distribution of number of people out of two in blood group B



Binomial distribution showing the number of subjects out of 100 in blood group B based on the probability of being in in blood group B of 0,08.

Aplikace binomického rozložení

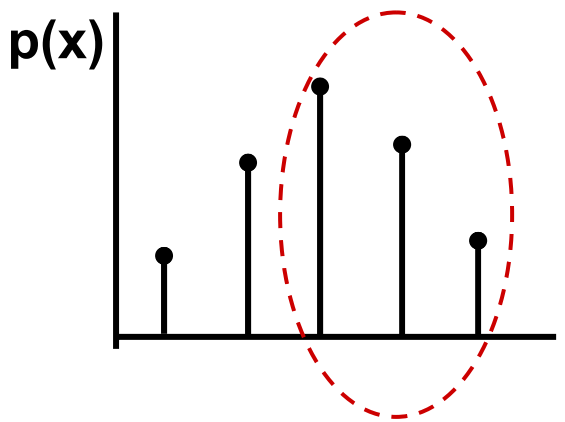
Populace: 60% jedinců má zvýšenou hladinu cholesterolu
Výběr: 5 lidí

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. p = 5 . 0,6 = 3 lidé ~ E(x)
n . p (1-p) = 1,2 ~ D(x)

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

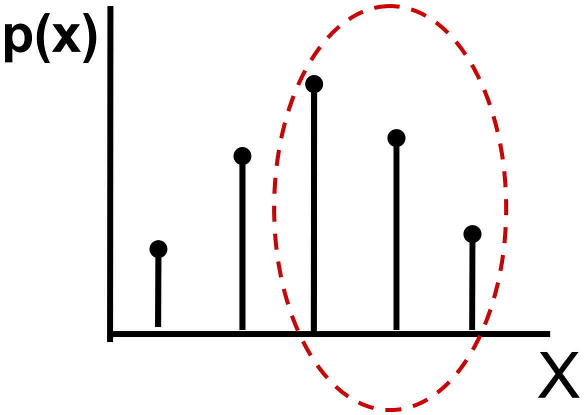
P(3) = ?

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

P(3) = 35%

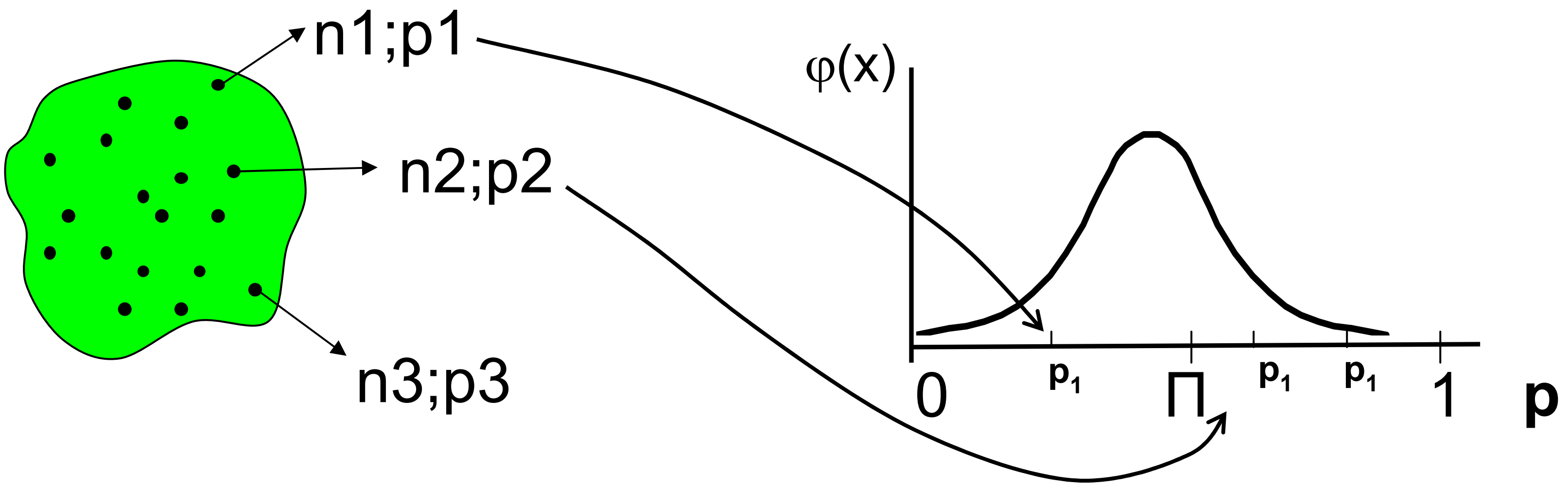
Jaká je P, že většina jedinců (tedy minimálně 3) má vyšší hladinu cholesterolu ? ~ Tzn. výběr alespoň obecně odpovídá zkoumané populaci ?

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

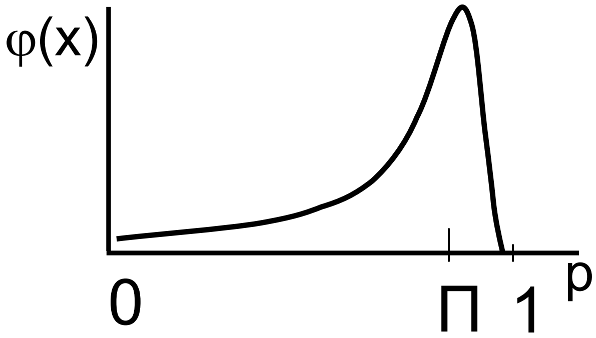


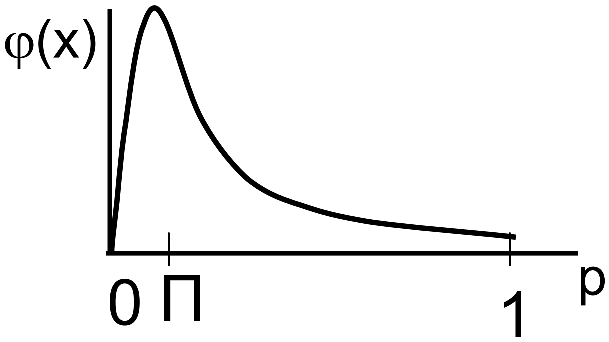
Odhad parametru Π binomického rozložení

Při vícenásobném odhadu se parametr Π chová jako normálně rozložen



U malých nebo velkých hodnot p (Π) je však předpoklad normality omezen





Odhad parametru Π binomického rozložení

I. vztahy

$$\pi \approx \hat{p}\,; \quad \hat{p} = {}^{r}\!/\!_{n}$$

1) Bodový

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

2) Intervalový – aproximace

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

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

Odhad parametru Π 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

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

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

$$\Downarrow$$

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

Odhad parametru p binomického rozložení

Intervalový odhad bez aproximací na normální rozložení - I. Vztahy

$$L_1 = \frac{r}{r + (n - r + 1) \cdot F_{\alpha/2}^{(\nu_1; \nu_2)}}$$

spodní limit intervalu

$$\nu_1 = 2(n - r + 1); \quad \nu_2 = 2r$$

$$L_2 = \frac{(r + 1) \cdot F_{\alpha/2}^{(\nu_1'; \nu_2')}}{n - r + (r + 1) \cdot F_{\alpha/2}^{(\nu_1'; \nu_2')}}$$

horní limit intervalu

$$\nu_1' = 2(r + 1) = \nu_2 + 2$$

$$\nu_2' = 2(n - r) = \nu_1 - 2$$

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

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

$$\nu_1 = 2(n - r + 1) = 2(200 - 4 + 1) = 394$$

$$\nu_2 = 2r = 2 \cdot 4 = 8$$

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

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

Horní hranice

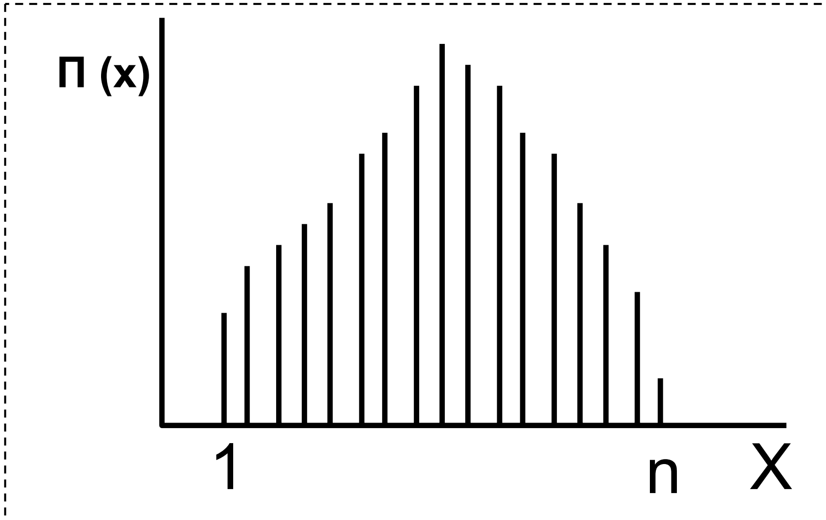
$$\nu_1' = 2(r + 1) = 10$$

$$\nu_2' = 2(n - r) = 2(200 - 4) = 392$$

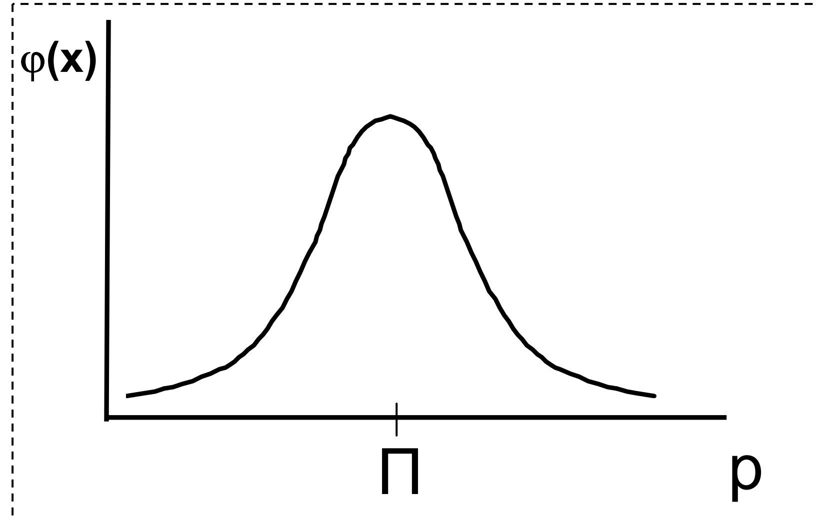
$$F_{1-\alpha/2}^{(10;392)} = \underline{\underline{2,08}}$$

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

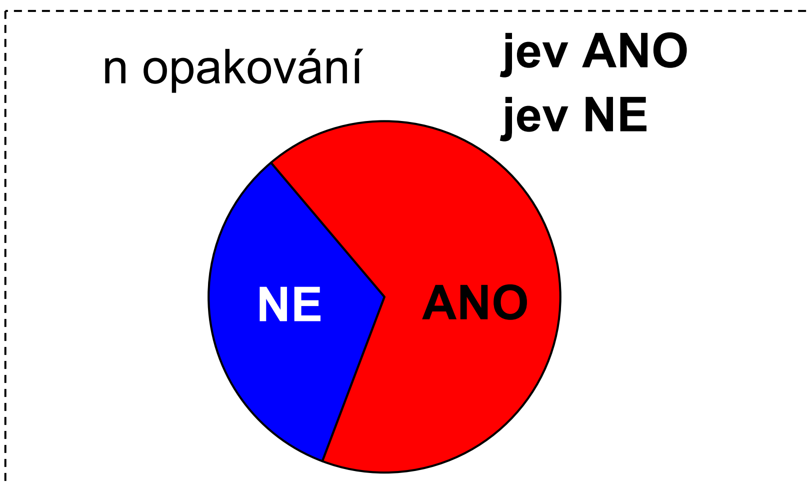
Binomické rozložení v datech - shrnutí



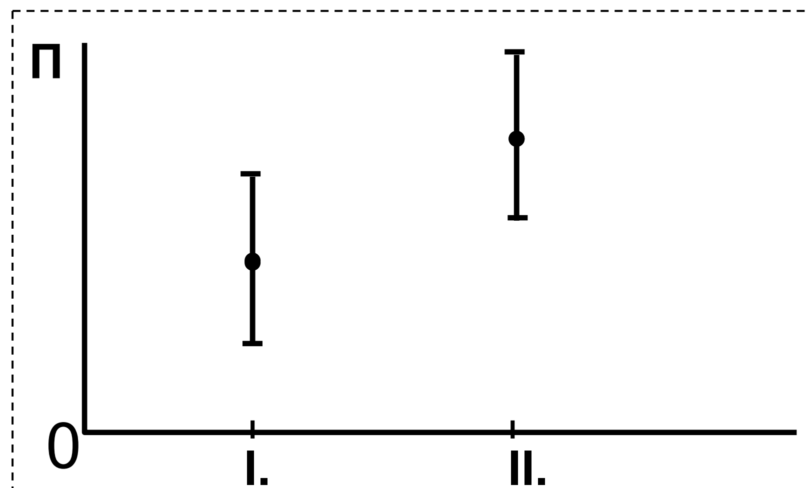
Pravděpodobnost výskytu hodnot X



Modelové rozložení odhadovaného parametru



Binární podstata původních hodnot

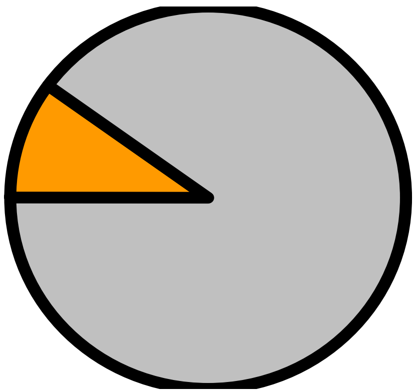


Interval spolehlivosti pro Π

„Two sample binomial test“

Analýza binárních nebo kategoriálních dat I.

Binární proměnná (1 / 0)

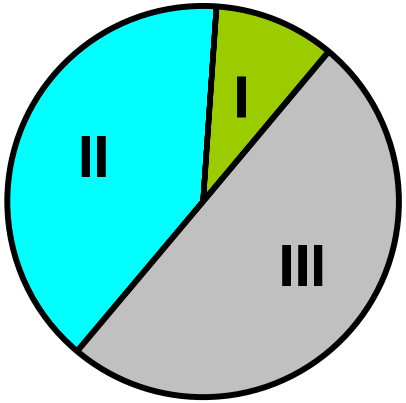


1: 10	
0: 90	
n: 100	

$p_1 = 0{,}1$

$p_0 = 0{,}9$

Kategoriální proměnná



I: 10	
II: 40	
III: 50	
n: 100	

$p_I = 0{,}1$

$p_{II} = 0{,}4$

$p_{III} = 0{,}5$

Analýza binárních nebo kategoriálních dat II.

I. **Liší se odhad p od předpokládané hodnoty P ?**

II. **Liší se dva nebo více odhadů 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
$p \le \Pi$	$p > \Pi$	z	$z > z_{1-\alpha}$
$p \ge \Pi$	$p < \Pi$	z	$z < z_{\alpha}$
$p = \Pi$	$p \ne \Pi$	z	$\lvert z \rvert > z_{1-\alpha/2}$

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

Korekce na kontinuitu

H_0	H_A	Testová statistika	Interval spolehlivosti
$p \le \Pi$	$p > \Pi$	$L_1 = \frac{(r+1)F_{\alpha, \nu_1', \nu_2'}}{n - r + (r+1)F_{\alpha, \nu_1', \nu_2'}}$	$p = r/n > L_1$
$p \ge \Pi$	$p < \Pi$	$L_2 = \frac{r}{r + (n - r + 1)F_{\alpha, \nu_1', \nu_2'}}$	$p < L_2$
$p = \Pi$	$p \ne \Pi$	$L_1; L_2\ (F_{\alpha/2}; F_{1-\alpha/2})$	$p < L_2$ v $p > L_1$

Test p ? π

✓ Stromy s pozměněným tvarem koruny

n = 9 000 jedinců
r = 2 250 změněných jedinců

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

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

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

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

P << 0,1

95 % Interval spolehlivosti … p: (0,241; 0,258)

Test p ? π

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
P(r)	0,00024	0,00293	0,01611	0,05371	0,12085	0,19335	0,22559	0,19336	0,12085	0,05371	0,01611	0,00293	0,00024

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

H_0: p = 0,5 je tedy značně nepravděpodobná

b) Pozorované $\hat{p} = {}^{10}\!/_{12} = 0{,}833$ **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 (p1 ? p2)

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

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

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

Dvouvýběrový binomický test (p1 ? p2)

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
Treated	15	9	24	$\hat{p}_1 = 0{,}625$	$\hat{q}_1 = 0{,}375$
Not Treated	10	15	25	$\hat{p}_2 = 0{,}400$	$\hat{q}_2 = 0{,}600$
Total	25	24	49	$\bar{p} = 0{,}510$	$\bar{q} = 0{,}490$

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

Z0,05(2) = t0,05(2) = 1,96

Nezamítáme H_0: 0,10 < P < 0,20

S korekcí na kontinuitu:

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

Z0,05(2) = t0,05(2) = 1,96

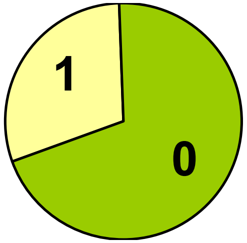
Nezamítáme H_0: 0,10 < P < 0,20

Analýza kontingenčních tabulek

Test dobré shody - základní teorie

Binomické jevy (1/0)

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



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 ?

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

Tabulková hodnota: $\chi^2_{(0{,}95)}(\nu = 1) = \underline{\underline{3{,}84}}$ $\qquad (0{,}95 = 1 - \alpha)$

Rozdíl je vysoce statisticky významný (p << 0,001]

Kontingenční tabulky - H_0 : Nezávislost dvou jevů A a B

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

Kontingenční tabulka 2 x 2

B ↓ / A →	+	-	Podíl (+)
+	a	b	$\frac{a}{(a+b)}$ **p_1**
-	c	d	$\frac{c}{(c+d)}$ **p_2**
Podíl (+)	$\frac{a}{(a+c)}$	$\frac{b}{(b+d)}$	

N = a + b + c + d

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

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

Očekávané četnosti:

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

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

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

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

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

p …… počet parametrů = 2

$$\chi^2_c = \sum \sum \frac{\left(\left|f_{ij} - F_{ij}\right| - 0,5\right)^2}{F_{ij}}$$

2 x 2 kontingenční tabulka - příklad (α = 0,05)

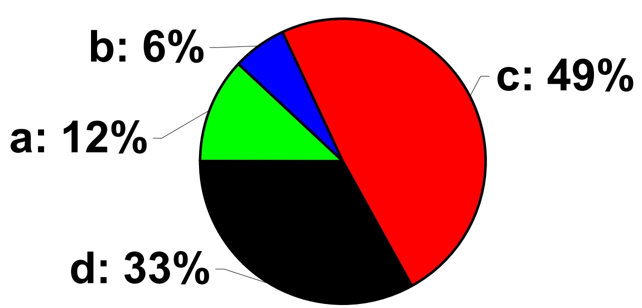
gen \ †	Ano	Ne	Σ
Ano	20	82	102
Ne	10	54	64
Σ	30	136	166

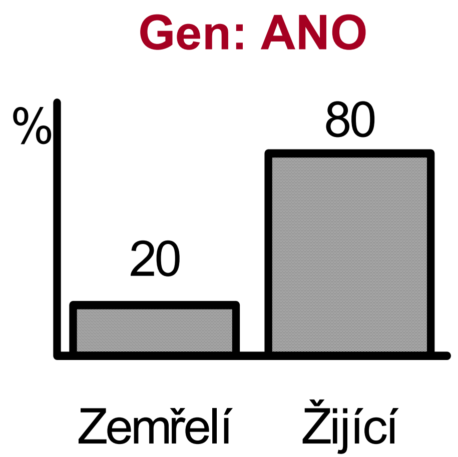
F_A = 102 * 30 / 166 = 18,43
F_B = 102 * 136 / 166 = 83,57
F_C = 11,57
F_D = 52,43

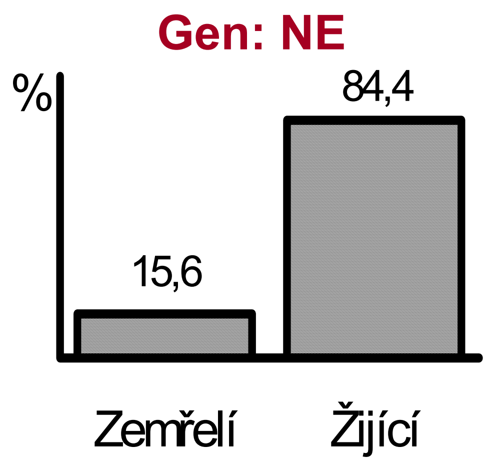
$$\chi^2_{(1)} = \frac{(20-18{,}43)^2}{18{,}43} + \frac{(82-83{,}57)^2}{83{,}57} + \frac{(10-11{,}57)^2}{11{,}57} + \frac{(54-52{,}43)^2}{52{,}43} = 0{,}423$$

$$0{,}423 < \chi^{2\,(1)}_{0{,}95} = 3{,}84$$

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.	Σ
ANO	a	b	c	d	číslo 1
NE	e	f	g	h	
Σ	číslo2				

Stupně volnosti:
(R-1) * (C-1) = 1 * 3 = 3

$$F_a = \frac{\check{c}\acute{\imath}slo\,1 \cdot \check{c}\acute{\imath}slo\,2}{N}$$

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

Očekávané četnosti:

$$p_a = \frac{a}{a+e} \qquad p_b = \frac{b}{b+f} \qquad p_c = \frac{c}{c+g} \qquad p_d = \frac{d}{d+h}$$

Kontingenční tabulky

Příklad 1

? **Ověřte na datech z pokusu se 100 květinkami 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		
	Žlutá	Červená	n
$f_{poz.}$	84	16	100
$f_{oček.}$	75	25	

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

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 roztříděno 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
$f_{poz.}$	152	39	53	6	250
$f_{oček.}$	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
$f_{poz.}$	152	39	53	244
$f_{oček.}$	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
$f_{poz.}$	6	244	25
$f_{oček}$	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 Ch - Ch**
72 Ch – H
65 H - 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.

Σ 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?

Σ 121 párů 1 : 1
očekávané četnosti = 60,5 : 60,5

$$\chi^2_{(1)} = 0,669 \qquad \frac{H - H}{Ch - Ch}$$

Σ 193 párů 1 : 1
očekávané četnosti = 96,5 : 96,5

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

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

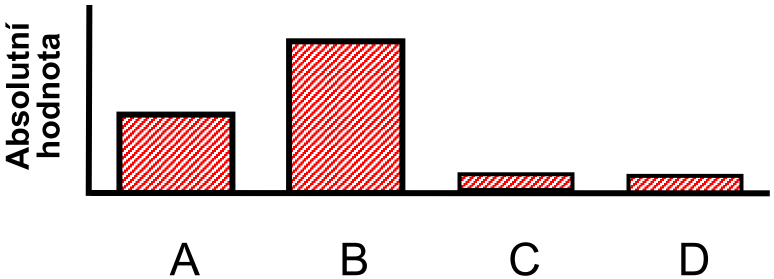
H_0: shoda f_i a F_i **a = 0,05** F_A: 64,13 F_B: 96,19 F_C: 192,39 F_D: 128,27

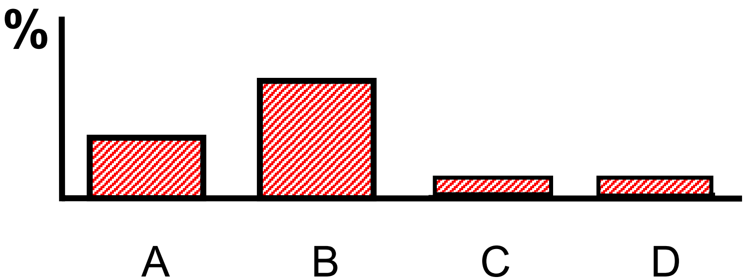
$$\chi^2_{(3)} = \frac{(32 - 64{,}13)^2}{64{,}13} + \dots\dots\dots + \frac{(116 - 128{,}27)^2}{128{,}27} = \underline{\underline{49{,}06}}$$

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

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: Π = 0,6
Absolutní četnost jevu označena r_i

Sledovalo s autorů z s zemí:

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

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

Test homogenity binomických rozložení

$$\chi^2_{S-1} = \frac{\left(\sum r_i\, p_i - \overline{p}\sum r_i\right)}{\overline{p}\,(1-\overline{p})}$$

Po možném sloučení s výběrů

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

Test shody reálného r $\left(\sum r_i\right)$ 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.

χ2 test - příklad složitější kontingenční tabulky I.

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

	Caffeine consumption (mg/day)				
Marital status	**0**	**1 - 150**	**151 - 300**	**> 300**	**Total**
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

	Caffeine consumption (mg/day)				
Marital status	**0**	**1 - 150**	**151 - 300**	**> 300**	**Total**
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 %)

χ2 test - příklad složitější kontingenční tabulky II.

Expected frequencies

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

Contributions of each cell

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

$\chi 2$ 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
Celkem	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
Celkem	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 sestrojit 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 směsného 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 (**r**) x kontrola (**k**)	2	5,64
A, B skupina u pacientů s vředy x kombinovaný vzorek (**r + k**)	1	0,68
0, A, B skupina u pacientů s **s** vředy x kombinovaný vzorek (**r + k**)	1	34,29
Celkem	**4**	**40,61**

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 rozborem 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- deliquents	Total
Spectacle wearers	**Yes**	1	5	6
	No	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

(I)

0	6
9	1

(II)

1	5
8	2

(III)

2	4
7	3

(IV)

3	3
6	4

(V)

4	2
5	5

(VI)

5	1
4	6

(VII)

6	0
3	7

Probability associated with each set of frequencies

	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
				Total	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

$$\} \sum = 102$$

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

H_0 nezamítnuta

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

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

OBECNĚ

		Skupina 1	Skupina 2
Znak	ANO	a	b
	NE	c	d

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

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

✓ **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

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

IS: ln RR - $Z_{1-\alpha/2}$. SE (ln RR)
ln RR + $Z_{1-\alpha/2}$. 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
Znak	ANO	a	b
	NE	c	d
	odds	**a/c**	**b/d**

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

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

PŘÍKLAD

		Vady chrupu: ANO	Vady chrupu: 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

OBECNĚ

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

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

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

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

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

Poissonovo rozložení

Poissonovo rozložení

Celkový počet jevů v **n** nezávislých pokusech

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

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

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

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

$$P(X = 1) = e^{-\mu} \cdot \mu^{2}$$

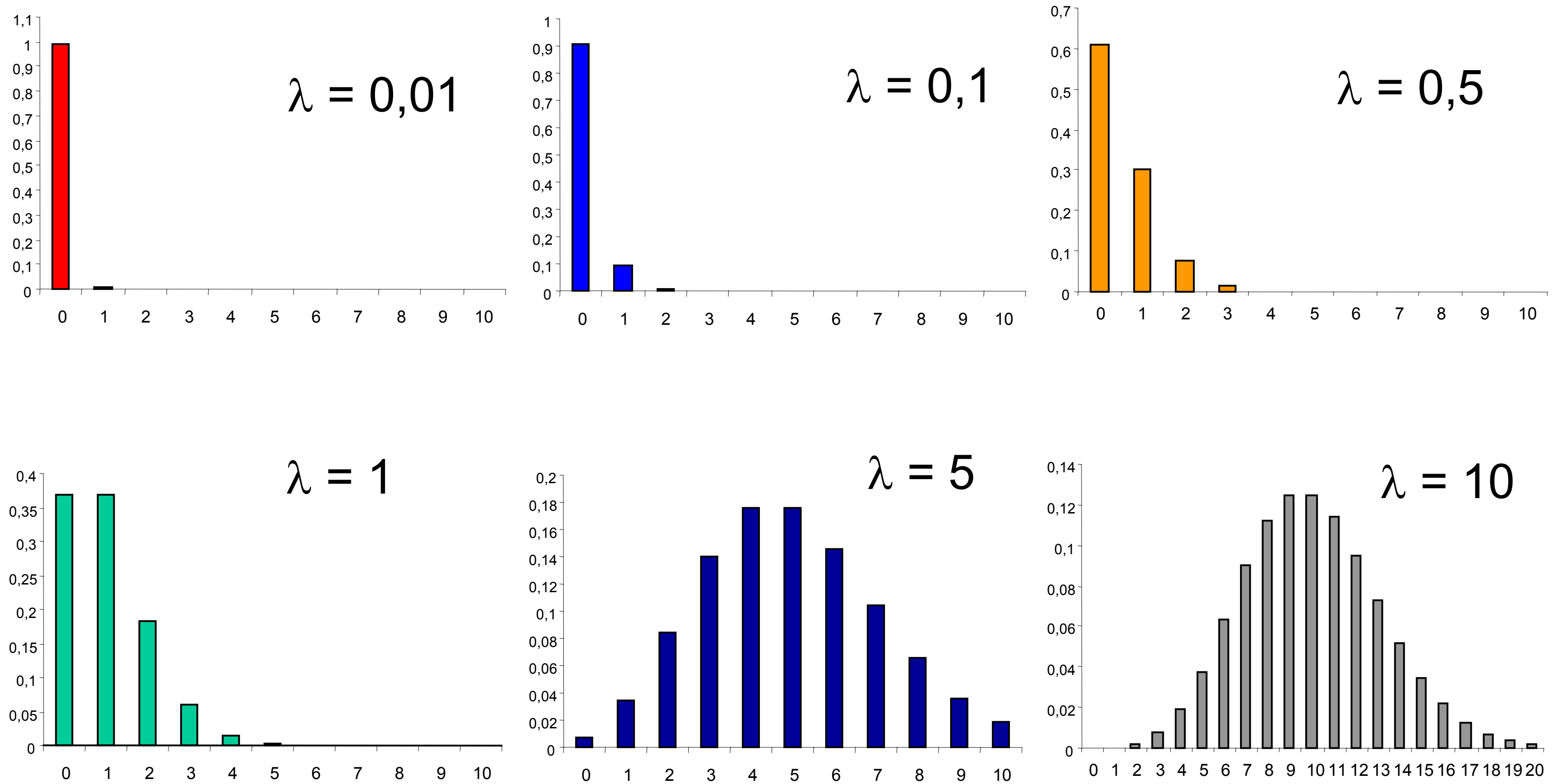
$$P(X = 2) = \frac{e^{-\mu} \cdot \mu^{2}}{2}$$

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

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

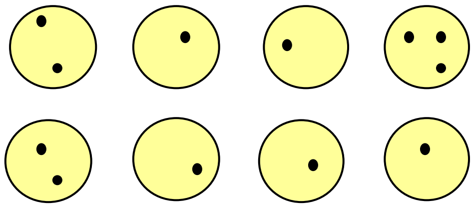
Poissonovo rozložení jako model

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

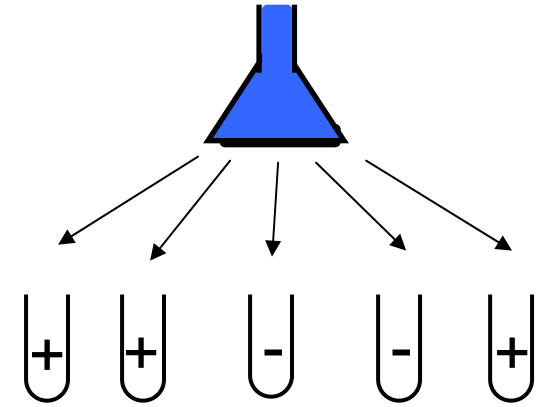


Poissonovo rozložení v přírodě existuje

Mutace bakterií na inkubačních miskách

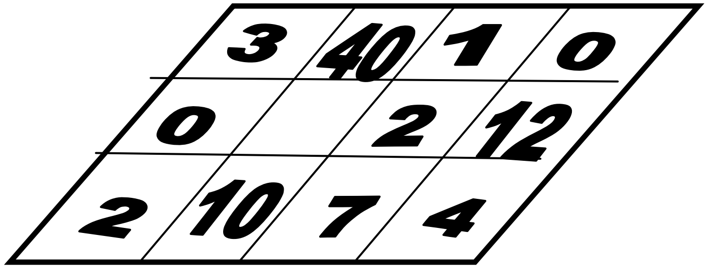


Orientační stanovení jevu
(při produkci plynu bakteriemi)



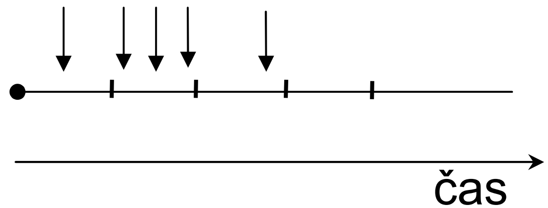
The most probable number technique

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



Výskyt jevu v čase

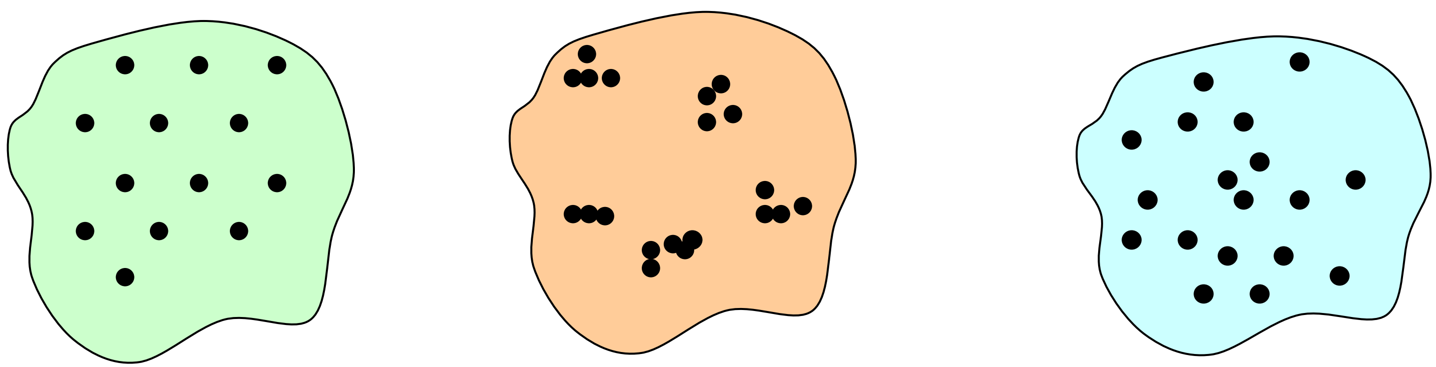
(srdeční arytmie v určitých časových intervalech)



Poissonovo rozložení jako model pro náhodný výskyt jevů

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

$\sigma^2 < \mu$ $\sigma^2 > \mu$ $\sigma^2 = \mu$



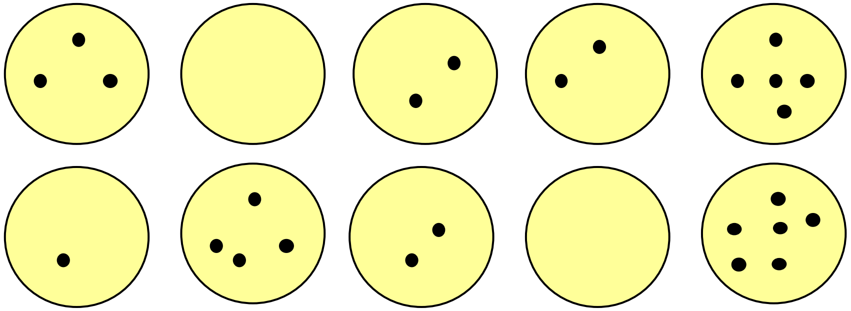
Uniform Clustered Random

Poisson

Pokud je λ spíše větší (~ 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)
λ..........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 λ Poissonova rozložení a jaká je jeho interpretace ?

Jaký je interval 95% spolehlivosti pro parametr λ ?

- **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 λ (bez aproximace na normální rozložení)

Spodní hranice IS

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

Horní hranice IS

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

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

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

i	Počet intervalů s právě **i** zaznamenanými částicemi $\mathbf{s_t}$	teoretické četnosti $\mathbf{np_i}$	$\frac{(s_i - np_i)^2}{np_i}$
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{ξ ≥ 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 aeras 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 Obs	Full moon days Exp	New moon days Obs	New moon days 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 %	Crimes on new moon days in India Frequency	Deaths per day in Montreal hospital %	Deaths per day in Montreal hospital Frequency	Expected distribution Poisson (0,51)
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 křepelek na dané ploše

n = 8 000 "pod lokalit"
r = 28

$$\left.\right\} \hat{p} = 0{,}0035$$

Nechť je srovnávací soubor
(předchozí průzkum)

$$p_o = 0{,}0020$$

$$p_o \cdot 8\,000 = 16 = \mu = \lambda$$

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

1) Vzít data jako pocházející z populace:

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

2) $P(r \geq 28) = ?$ $[0{,}00411]$ $\Big\} < 0{,}05 \Rightarrow H_o$ zamítnuta

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

$p > p_o$, **aby r = 28 bylo pravděpodobnější**