XII. Binomické rozložení

Popis binomického rozložení

Testování hypotéz binomicky rozložených dat

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek

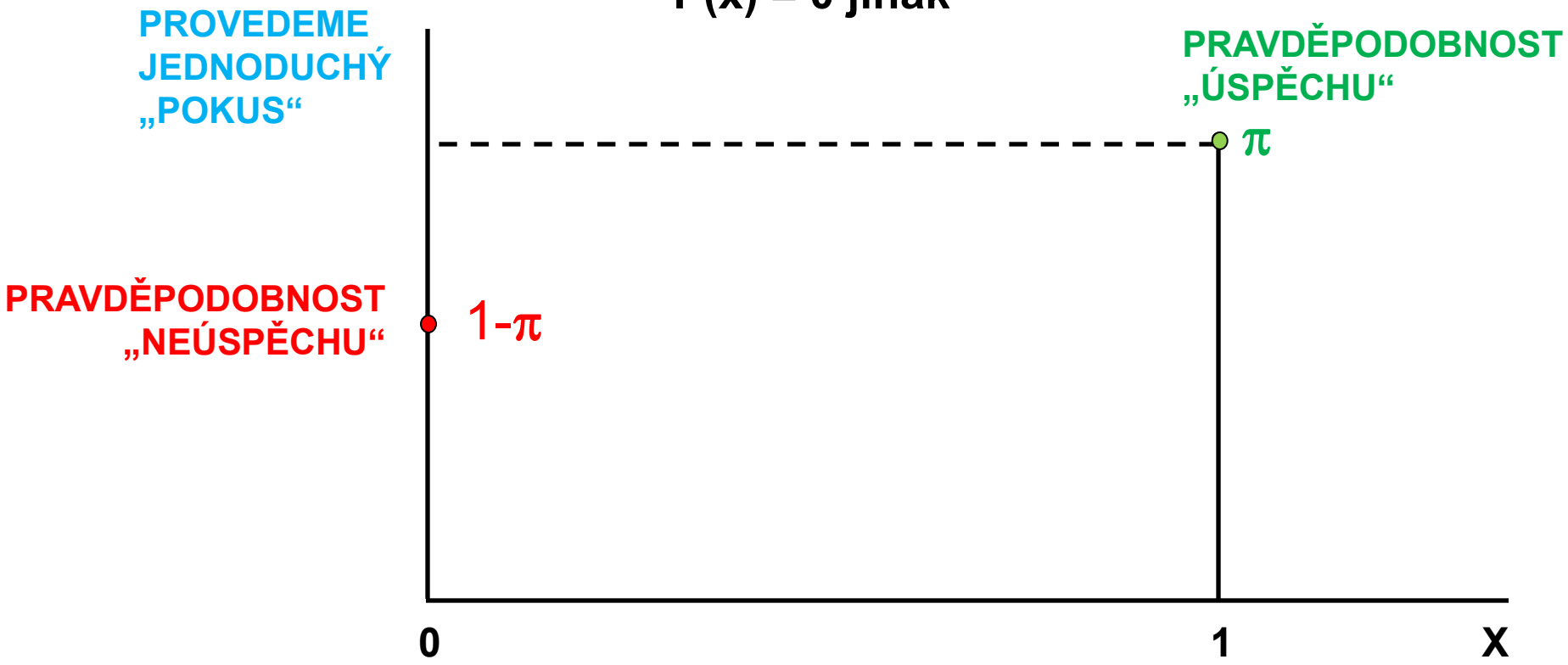
Anotace

- Kromě spojitých dat se setkáváme také s daty **kategoriálními**, jejichž nejjednodušším případem jsou data **binární**. Binární data jsou popsána **binomickým rozložením**, od chování binomického rozložení je odvozena **popisná statistika binárních dat** (procento výskytu jevu), její **interval spolehlivosti** a **binomické testy** pro srovnání procentuálního výskytů jevů v různých skupinách.

Alternativní rozložení

PRAVDĚPODOBNOSTNÍ FUNKCE
***DISKRÉTNÍHO* ROZDĚLENÍ**

$P(x) = \pi$ pro $x = 1$
$P(x) = 1 - \pi$ pro $x = 0$
$P(x) = 0$ jinak



Binomické rozložení

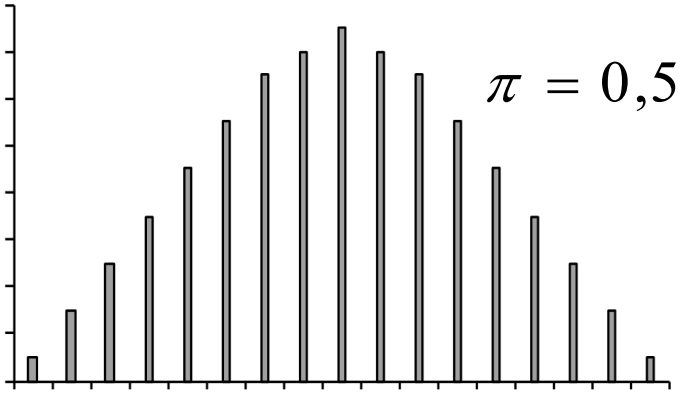
X ..... celkový počet nastání jevu v **n** nezávislých pokusech

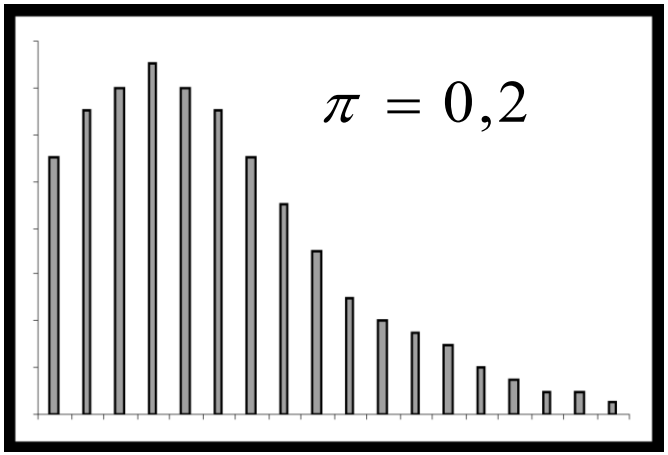
SOUČET ALTERNATIVNÍCH ROZDĚLENÍ

$E(X) = n \cdot \pi$

$D(X) = n \cdot \pi (1-\pi)$

π → **jediný parametr distribuce určuje tvar distribuce**





Binomické rozložení jako model pro zkoumání výskytu sledovaného jevu

n ..... **počet nezávislých opakování** experimentu

r ..... znamená celkový **počet nastání jevu** v **n** nezávislých experimentech

r : **0 …… n**

π .. jediný **parametr** binomického rozložení

X: Binomická proměnná

Střed rozložení:

$$\mathrm{E}(x) = n \cdot \pi$$

Rozptyl:

$$\mathrm{D}(x) = n \cdot \pi \cdot (1 - \pi)$$

p .... **odhad** parametru **π**

$$p = r/n$$

Binomické rozložení jako model

BINOMICKÁ VĚTA

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

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

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

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

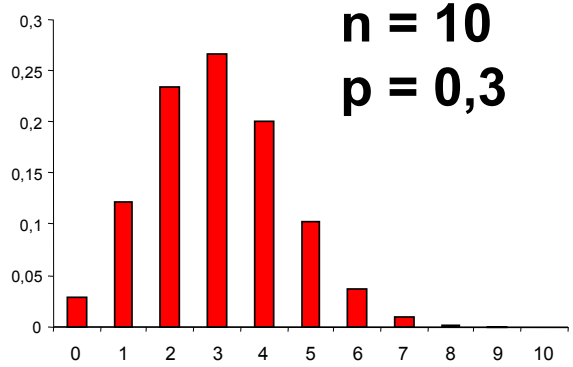
BINOMICKÝ KOEFICIENT

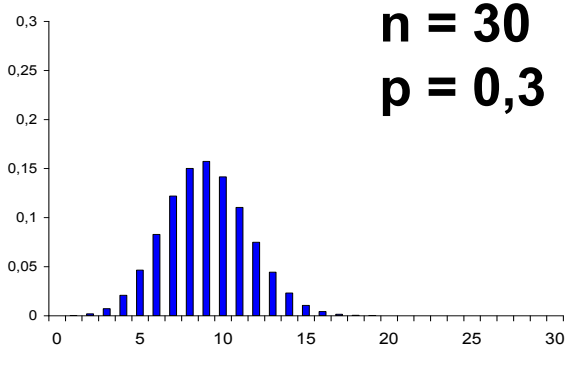
počet r-členných kombinací z n objektů

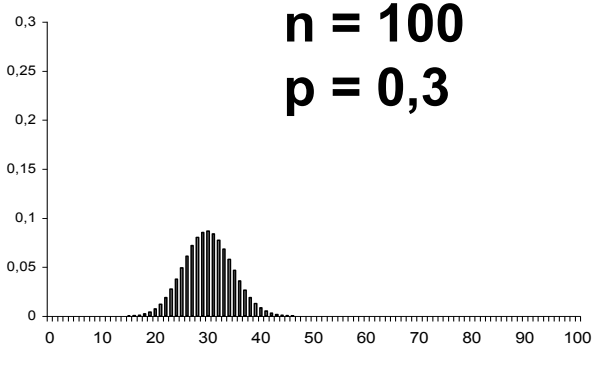
Binomické rozložení jako model

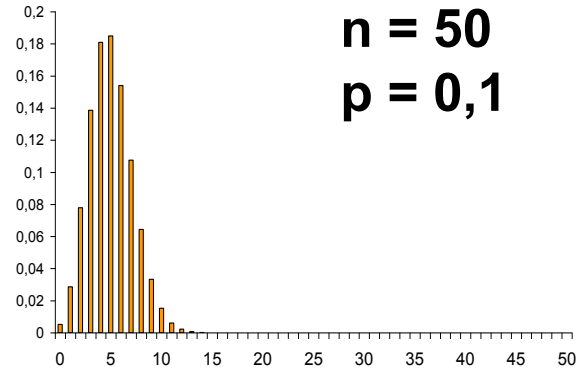
$$P(x = r) = \frac{n!}{r!(n-r)!} \cdot p^{r} \cdot q^{(n-r)}$$

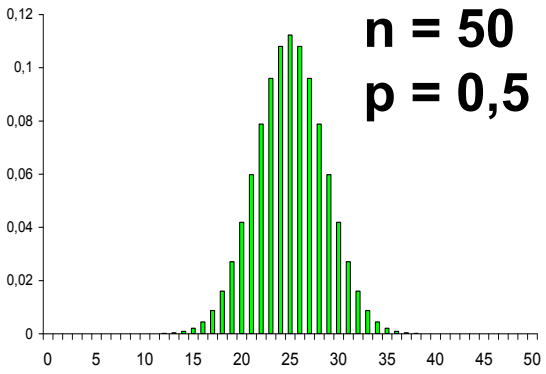
$q = 1 - p$

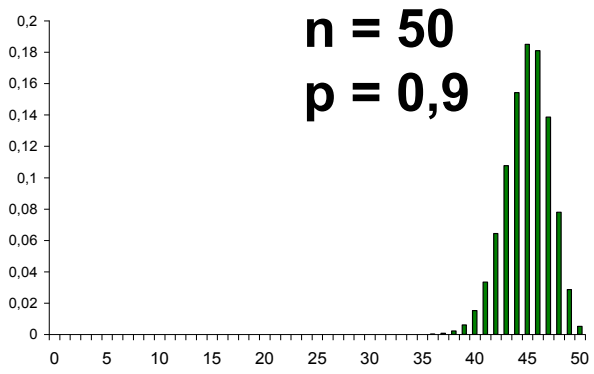








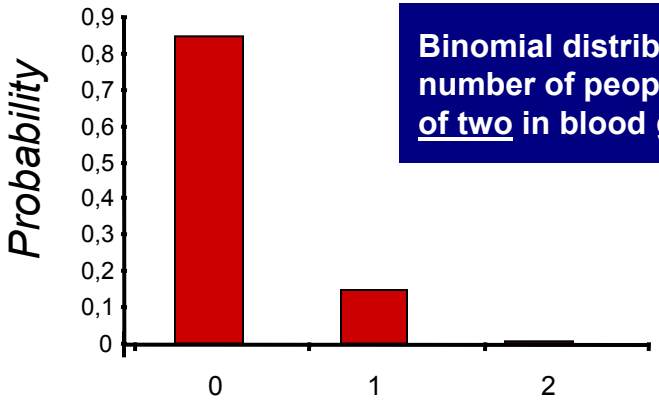




Aplikace binomického rozložení

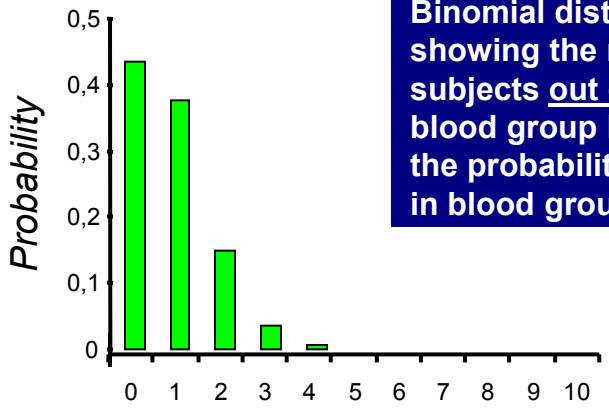
Výskyt krevní skupiny B v určité populaci: p = 0,08

		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



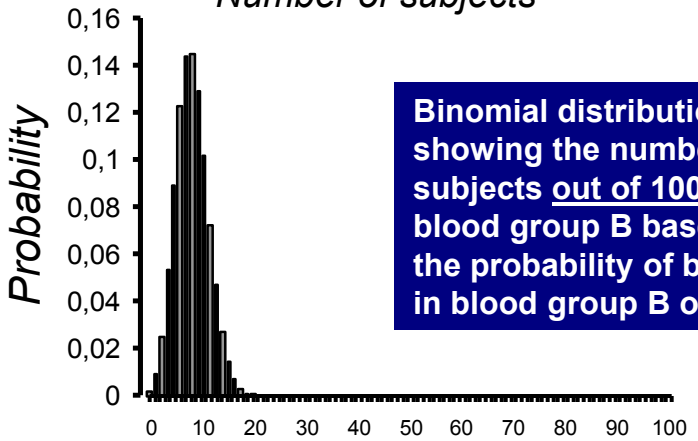
Number: blood group B in 2 cases

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



Number of subjects

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.



Number of subjects

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 očekáváme ve výběru s vyšší hladinu cholesterolu ?

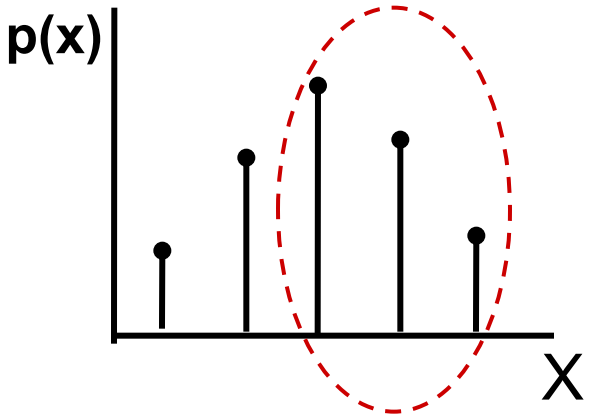
n. p = 5 . 0,6 = 3 lidé ~ E(x)
n . p (1-p) = 1,2 ~ D(x)

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%

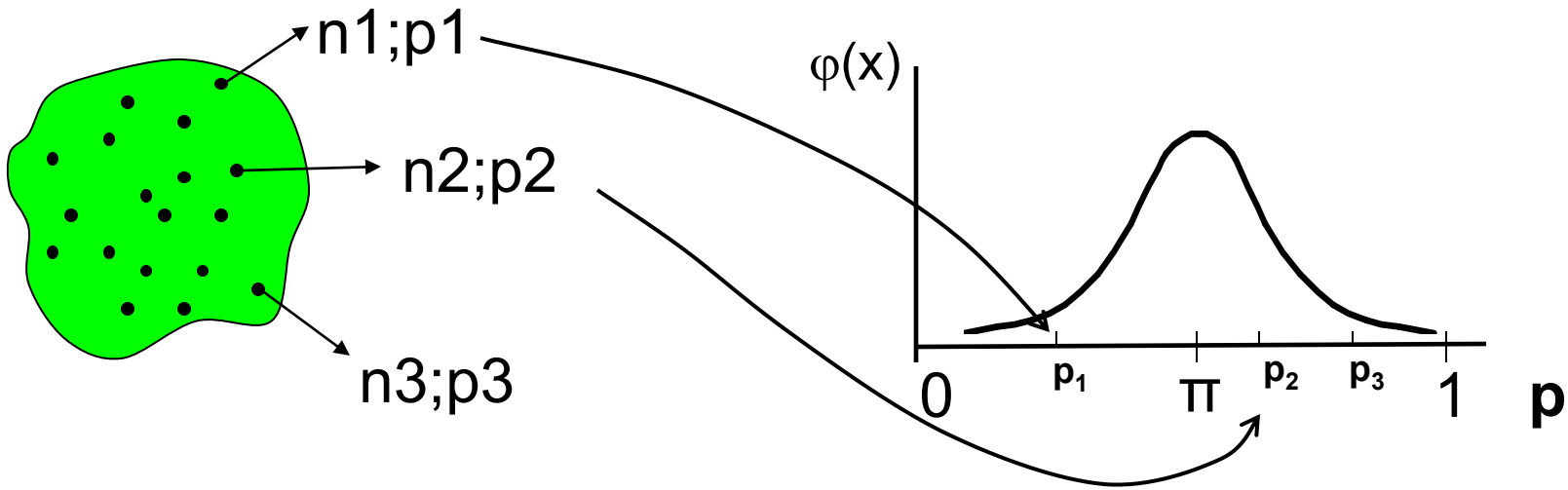


III. Jaká je P, že většina jedinců (tedy minimálně 3) má vyšší hladinu cholesterolu ? ~ Tzn. výběr alespoň obecně odpovídá zkoumané populaci ?

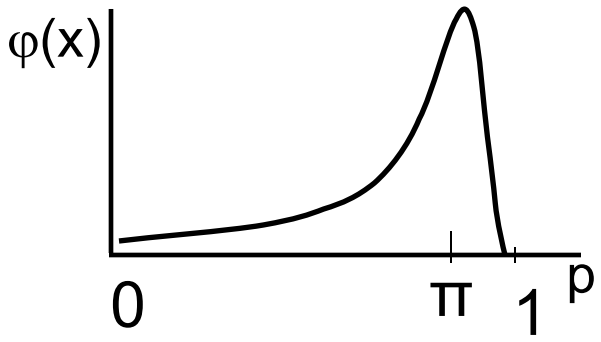
P(X > 3) = P(3) + P(4) + P (5) = 0,346 + 0,259 + 0,078 = 68 %

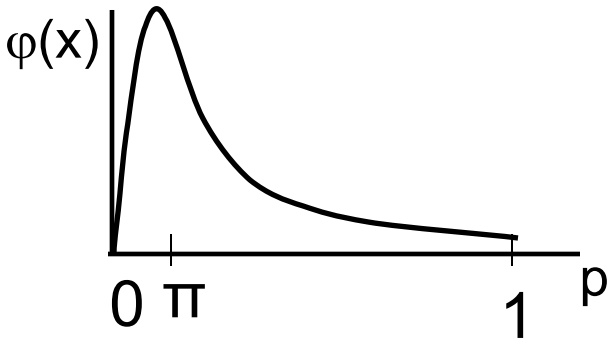
Odhad parametru π binomického rozložení

Při vícenásobném odhadu se odhad parametru π chová jako normálně rozložen



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





Odhad parametru π binomického rozložení
NORMÁLNÍ APROXIMACE

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

1) Bodový

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

2) Intervalový – aproximace

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

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

Odhad parametru π binomického rozložení: příklad I

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 π binomického rozložení

Intervalový odhad bez aproximací na normální rozložení

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

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

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

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

$$P(L_1 \le \pi \le L_2) \ge 1 - \alpha$$

Odhad parametru π binomického rozložení: příklad II

Náhodný vzorek n = 200 jedinců.
Zjištěno pouze r = 4 jedinci bez určitého znaku.

$$\hat{p} = {}^{4}\!/_{200} = \underline{\underline{0,02}}$$

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

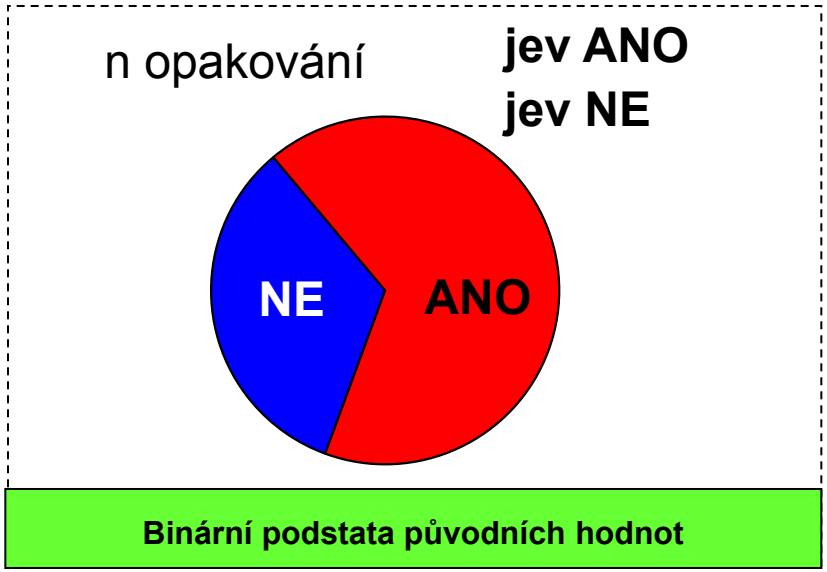
$$\nu_1' = 2(r + 1) = 10$$

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

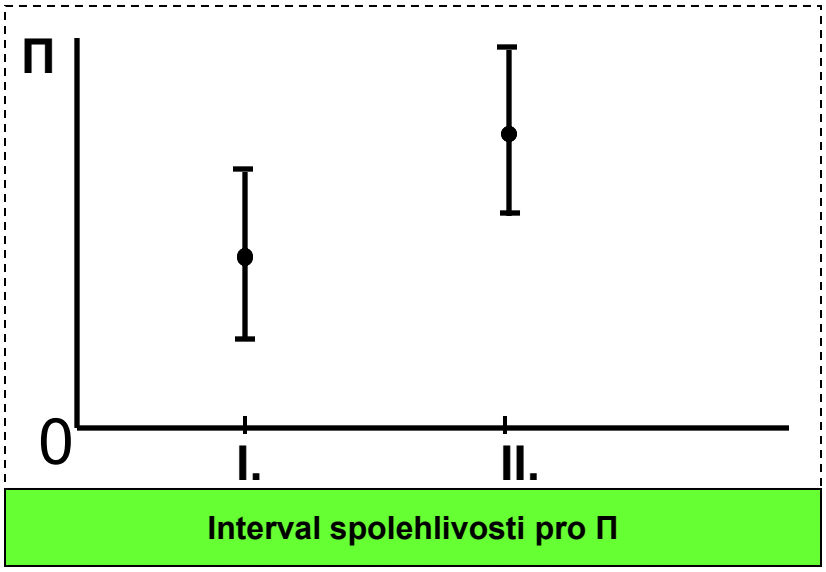
$$F_{1-\alpha/2}^{(10;392)} = \underline{\underline{2,08}}$$

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

Binomické rozložení v datech: vizualizace



Binární podstata původních hodnot



Interval spolehlivosti pro Π

Statistické testování binomických dat

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

jednovýběrový test

II. **Liší se dva nebo více odhadů p ?**

- závislé odhady - dvouvýběrový test

- nezávislé odhady -

III. **Je výskyt kategorií dvou jevů nezávislý ?**

IV. **Hodnocení relativního rizika z výskytu určitého jevu v rámci skupiny lidí**

Jednovýběrový binomický test

H_0	H_A	Testová statistika	Kritický obor
$p \leq \Pi$	$p > \Pi$	z	$z > z_{1-\alpha}$
$p \geq \Pi$	$p < \Pi$	z	$z < z_{\alpha}$
$p = \Pi$	$p \neq \Pi$	z	$\lvert z \rvert > z_{1-\alpha/2}$

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

Korekce na kontinuitu

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

Test π ? p: PŘÍKLAD 1

*Příklad testu **s aproximací** na normální rozložení*

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

$Z < -Z_{1-\alpha/2}$ ……zamítáme H_0: p < 0,01

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

Test π ? p: PŘÍKLAD 2

*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?

H_0	H_A
$p = 0,5$	$p > 0,5$

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:**

Kvantil Fischerova rozdělení
$F_{1-\alpha,\ 14,\ 12} = 2,64$

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

Test π ? p: PŘÍKLAD 3

Pravděpodobnost narození chlapce je asi 1/2. Máte zhodnotit výsledky průzkumu populace, která žije v silně poškozeném životním prostředí. Průzkum se týká 1000 náhodně vybraných rodin a zjištěný podíl narozených chlapců je 0.41.
Jaké jsou vaše závěry o této populaci?
Jak se váš odhad zpřesní, když použijete vzorek n = 10 000 rodin při zachování odhadu p = 0.41?

Použijeme **jednovýběrový binomický test s nulovou hypotézou H_0: p=π**, hladina významnosti α=0,05

testová statistika $Z = \frac{n \cdot \hat{p} - n \cdot \pi}{\sqrt{n \cdot \hat{p}(1-\hat{p})}} = \frac{1000 \cdot 0{,}41 - 1000 \cdot 0{,}5}{\sqrt{1000 \cdot 0{,}41 \cdot 0{,}59}} = -5{,}79$ a příslušný kvantil $Z_{1-\frac{\alpha}{2}} = Z_{0{,}975} = 1{,}96$

protože $|Z| > Z_{0{,}975}$ **NULOVOU HYPOTÉZU ZAMÍTÁME. Chlapci se ve zkoumavé populaci nerodí s pravděpodobností 0,5.**

interval spolehlivosti π : $\hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0{,}4 \pm Z_{0{,}975} \cdot 0{,}046 = 0{,}41 \pm 1{,}96 \cdot 0{,}016 = 0{,}41 \pm 0{,}03$

pokud použijeme n=10 000, bude int. spolehlivosti užší π : $\hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0{,}41 \pm 1{,}96 \cdot 0{,}005 = 0{,}41 \pm 0{,}01$

Dvouvýběrový binomický test (p1 ? p2)

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

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

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

Dvouvýběrový binomický test (p1 ? p2)

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

✓ Celkem 49 pokusných myší bylo použito k testování léčivého preparátu během dvouměsíční terapie. Následující tabulka obsahuje původní data zároveň s testem **nulové hypotézy: Podíl přežívajících jedinců je u léčené 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	$\hat{p} = 0{,}510$	$\hat{q} = 0{,}490$

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

➡ **Nezamítáme H_0: p = 0,116**

Kvantil standardizovaného normálního rozdělení = KRITICKÁ HODNOTA TESTU
$Z_{0,05}(2)$ = 1,96

S korekcí na spojitost:

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

➡ **Nezamítáme H_0: p = 0,198**

Korekce na spojitost, vhodná u malých vzorků