

# XI. Statistické testy o parametrech dvou výběrů



Dvouvýběrový párový a nepárový t-test  
Neparametrické alternativy t-testu

# Anotace

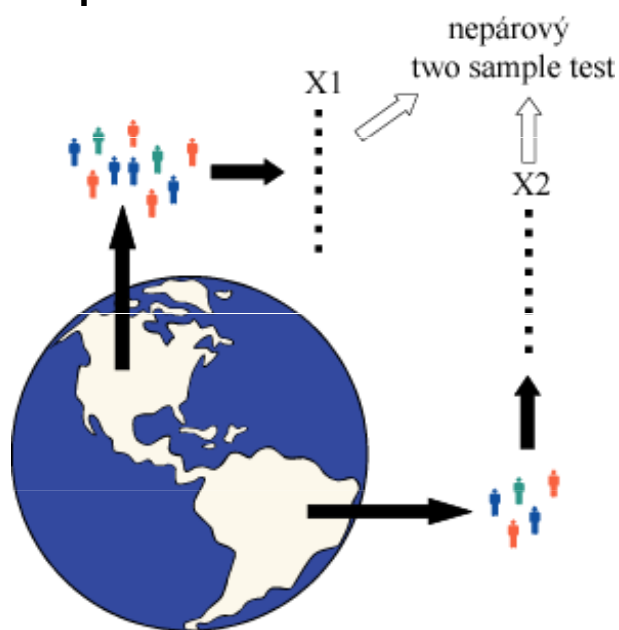


- Jedním z nejčastějších úkolů statistické analýzy dat je srovnání spojitých dat ve dvou skupinách pacientů. Na výběr je celá škála testů, výběr konkrétního testu se pak odvíjí od toho, zda je o srovnání párové nebo nepárové a zda je vhodné použít test parametrický (má předpoklady o rozložení dat) nebo neparametrický (nemá předpoklady o rozložení dat, nicméně má nižší vypovídací sílu).
- Nejznámějšími testy z této skupiny jsou tzv. t-testy používané pro srovnání průměrů dvou skupin hodnot

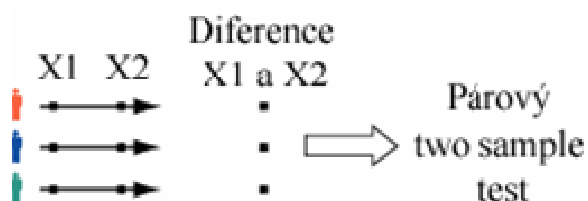
# Dvojitéstypová nepárová



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



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

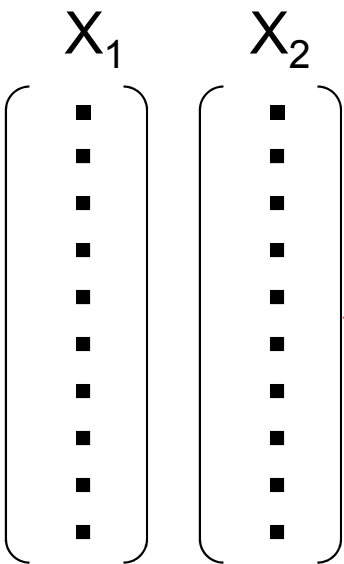


- Základním testem pro srovnání dvou závislých rozložení spojitých čísel je **párový two-sample t-test**

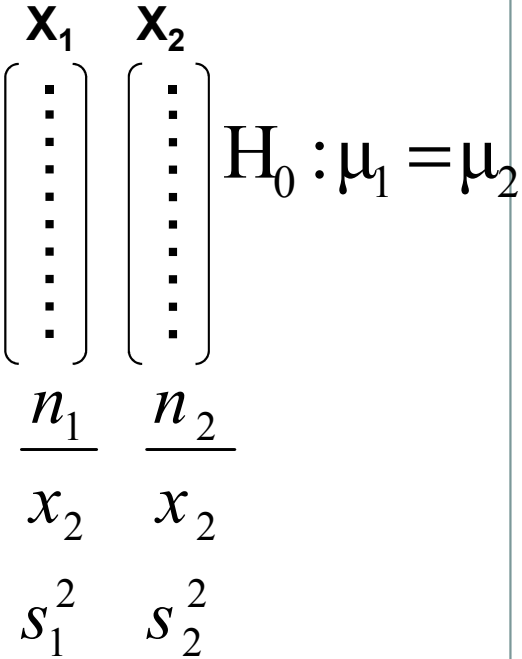
# Dvojčlenný test pro nezávislé



Data



Nezávislé uspořádání

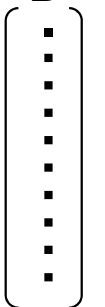


$$H_0 : \mu_1 = \mu_2$$

Párové uspořádání



$$X_1 - X_2 = D$$



$$H_0 : \bar{D} = 0$$

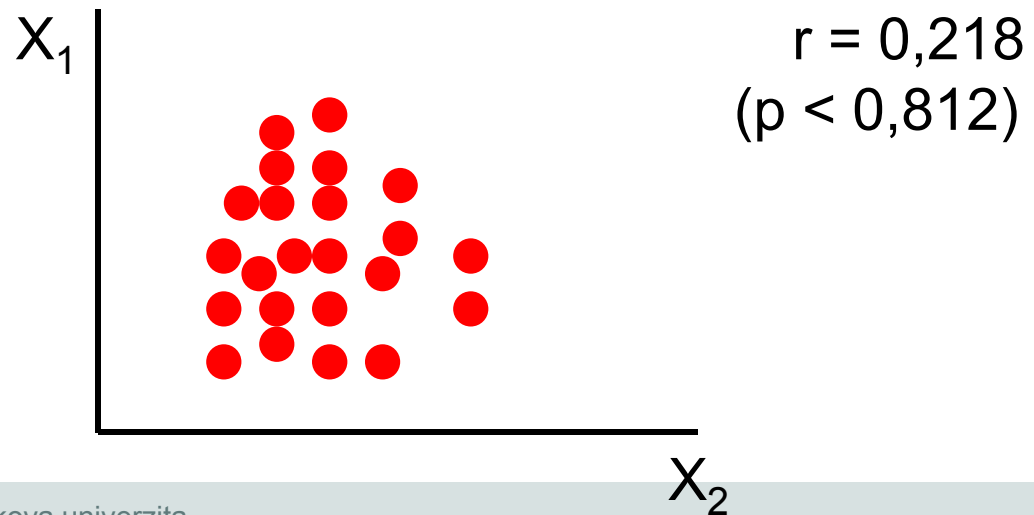
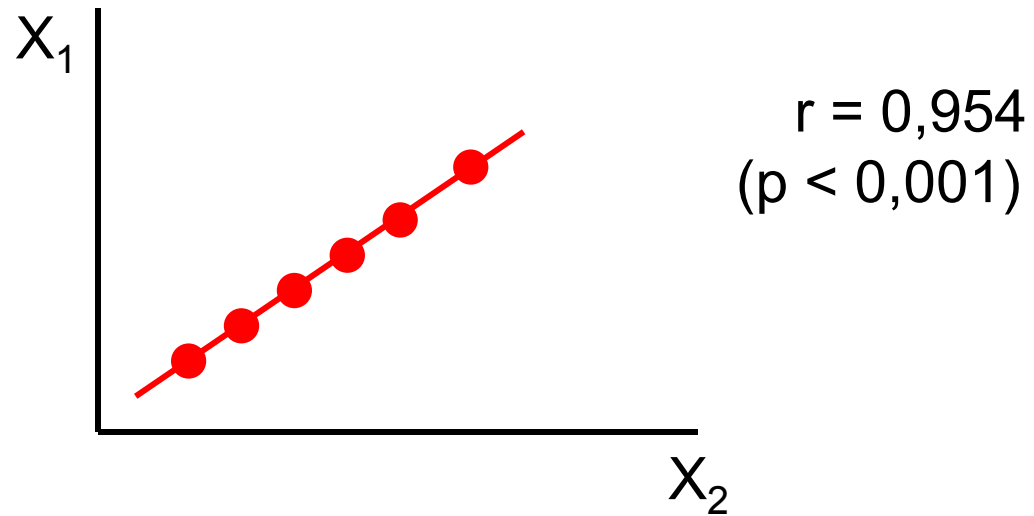
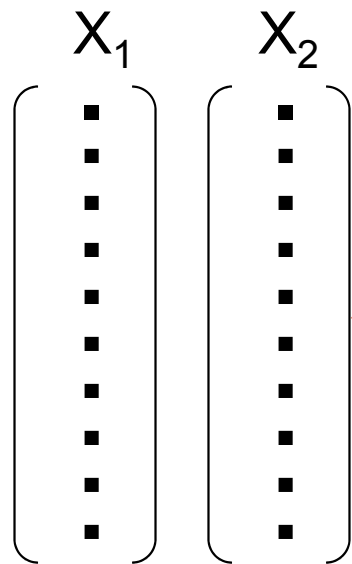
Design uspořádání zásadně ovlivňuje interpretaci parametrů

$$\frac{n}{D} s_D^2$$

$$(n = n_2 = n_1)$$

# Důležitý test párové nepárové

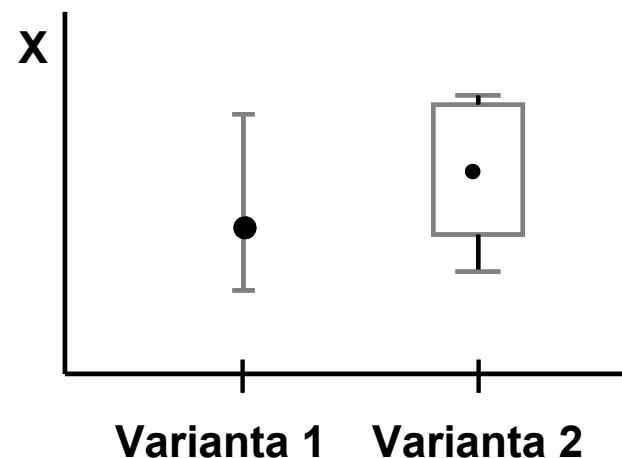
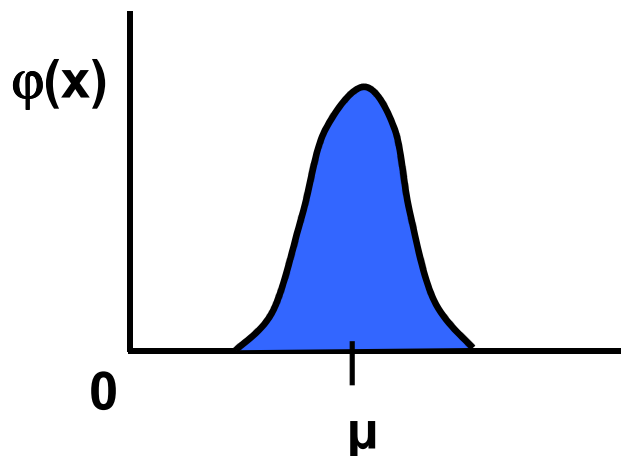
## Identifikace párovitosti (Korelace, Kovariance)



# Předpoklady nezávislého dvoustranného t-testu



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



# Nezávislostový test – výpočet I



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

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

## F-test pro srovnání rozptylů

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

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

# Nezávislostový test – výpočet I



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

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

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

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

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



# Dvojovýt-test - příklad



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

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

$$t = \frac{\text{Rozdíl}_{\text{průmě}}}{SE(\text{rozdílprůo ěrů})} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad v = n_1 + n_2 - 2$$

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

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

# Neparametrické alternativní párového t-testu



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

## Mann Whitney U-test

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

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

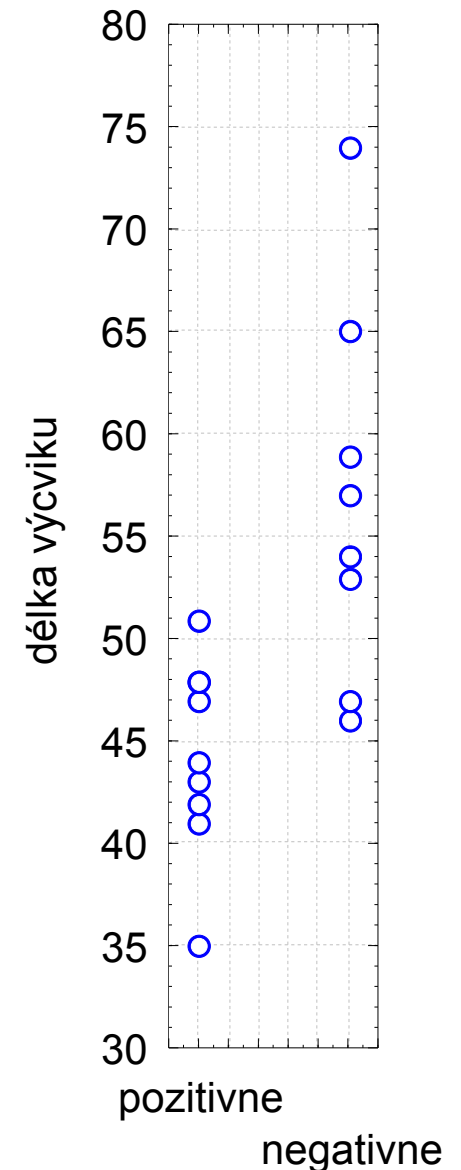
• Pro oba soubory je tedy vytvořen součet pořadí a menší z obou součtů je porovnán s kritickou hodnotou testu, pokud je tato hodnota menší než kritická hodnota testu, zamítáme nulovou hypotézu shody distribučních funkcí obou skupin.

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

# Mann – Whitney U test - příklad



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



# Párové testy – předpoklady



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

## Existuje několik možných designů experimentu, které lze srovnávat:

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

# Párový t-test



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

# Párový t-test – příklad

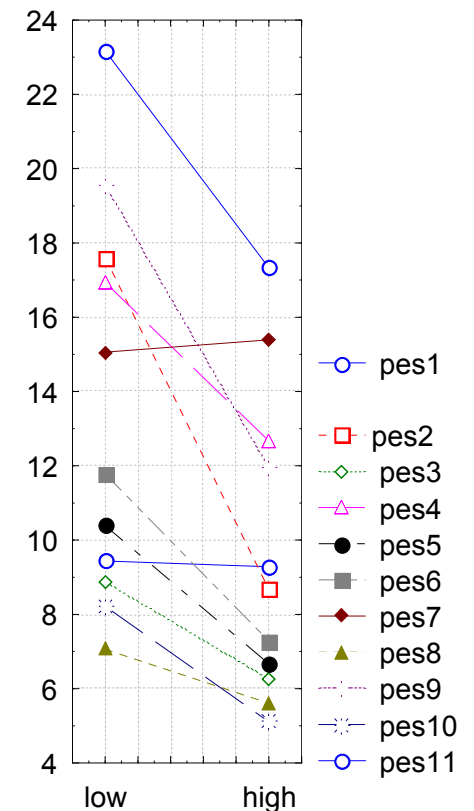


Byl prováděn pokus s dietou 11 diabetických psů, každý pes byl vystaven dvěma dietám s odlišným typem sacharidů (snadno vstřebatelné X pozvolna se rozkládající na glukózu), hodnoty krevní glukózy v průběhu jednotlivých diet mají být srovnány pro zjištění vlivu diety na hladinu krevní glukózy. Protože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

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

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

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



# Neparametrická obdoba párového t-testu



## Wilcoxon test

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

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

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

# Wilcoxonův test – příklad I

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

A.....parametr krve před podáním léku

B.....parametr krve po podání léku

$W_+$  ..... © pořadí kladných rozdílů = 51

$W_-$  ..... = 4

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

počet párů = n = 10

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



# Wilcoxonův test – příklad I



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

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

# Znaménkový test – příklad I



## Párově uspořádaný experiment pro nominální data

### I. Dva preparáty, každý na 1/2 listu

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

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

V – větší; M – menší

n = 10 listů s rozdílnými výsledky

jev → A je větší: +  $n_+ = 7$

jev → B je menší: -  $n_- = 3$

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

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

- aplikované na vzorek s antigenem

n = 10

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

n – nenulových rozdílů: 6 → A:  $n_+ = 4$

→ A:  $n_- = 2$

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

# Znaménkový test – příklady



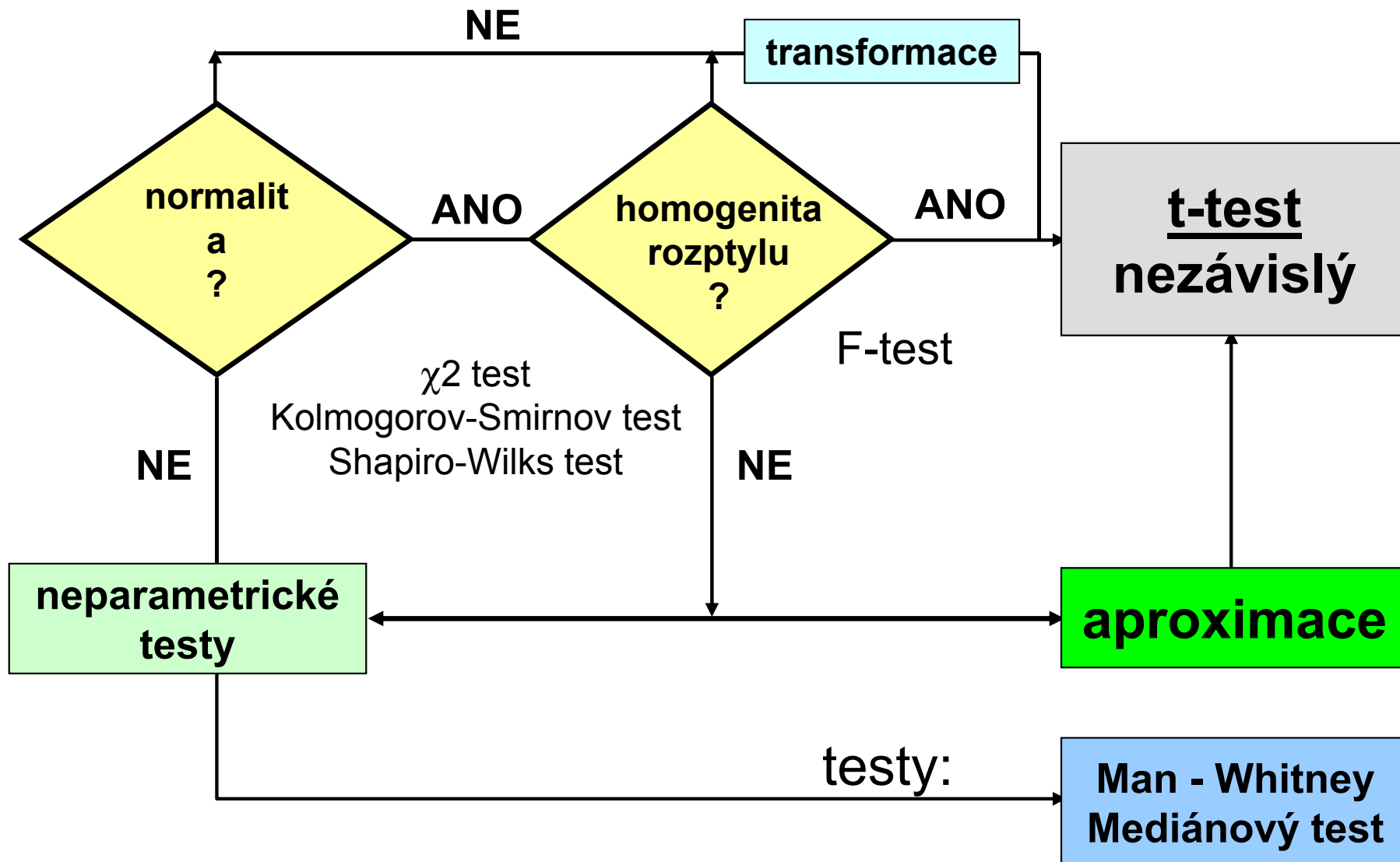
- Na konferenci veterinářů bylo předneseno, že průměrný čas konzultace je 12 minut. Následovala debata, zda je lepší použít medián nebo průměr. Jeden z nich se rozhodl ověřit teorii, že průměrná konzultace trvá 12 minut na vlastní praxi a zaznamenal si trvání svých 43 konzultací. K otestování hypotézy,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

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

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

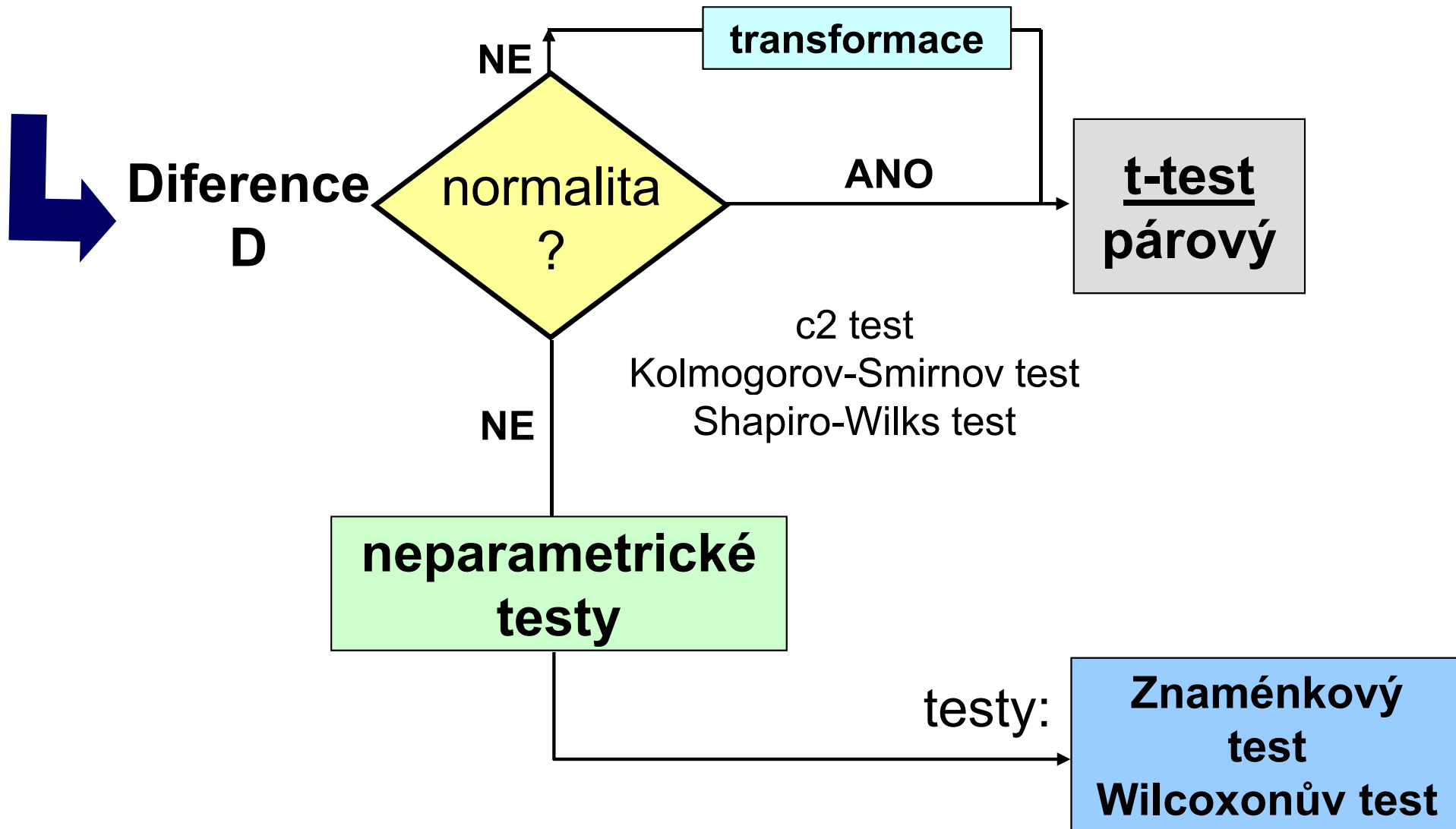
# Dvojčíslo testyschéma analýzy

## Nezávislé uspořádání



# Důležité testy schéma analýzy

## Párové uspořádání



# XII. Binomické rozložení



**Popis binomického rozložení**  
**Testování hypotéz binomicky rozložených dat**

# 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



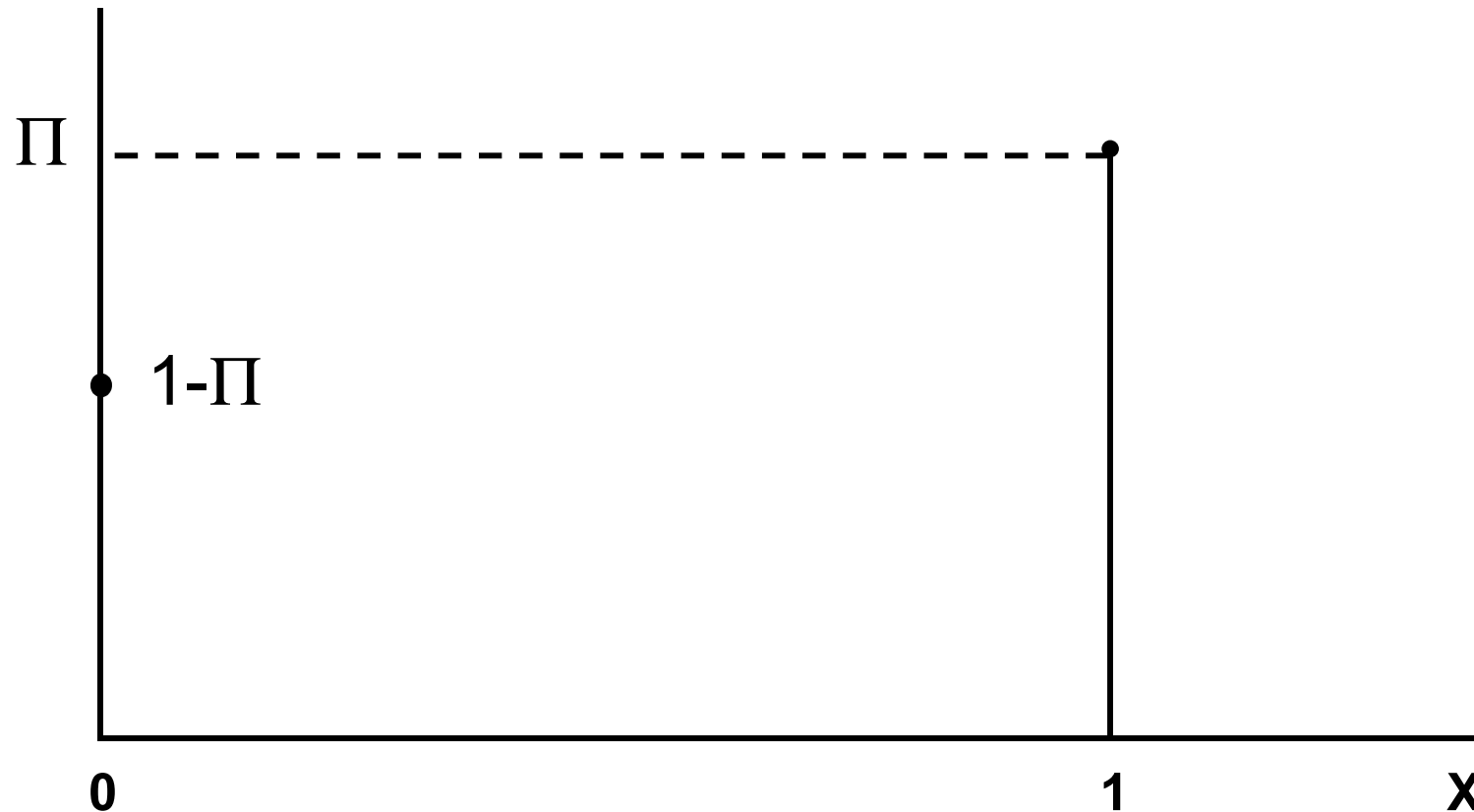
$$\Pi(x) = \Pi \text{ pro } X = 1$$

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

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



$X = 1$  .....jev





# Binomické rozložení

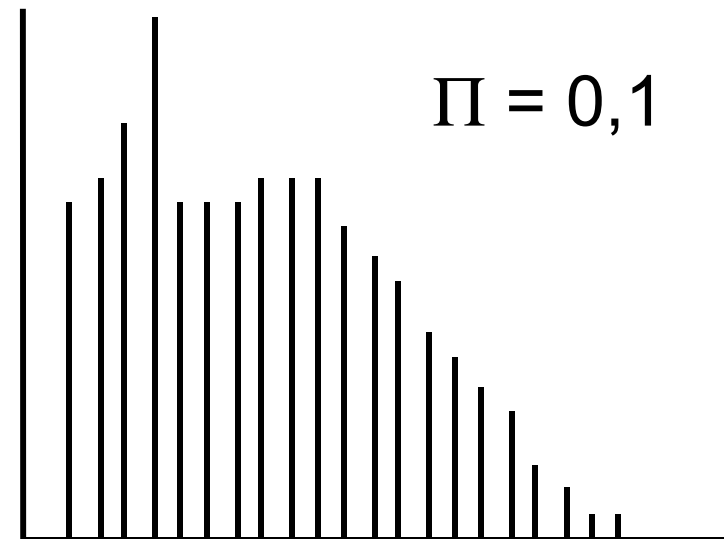
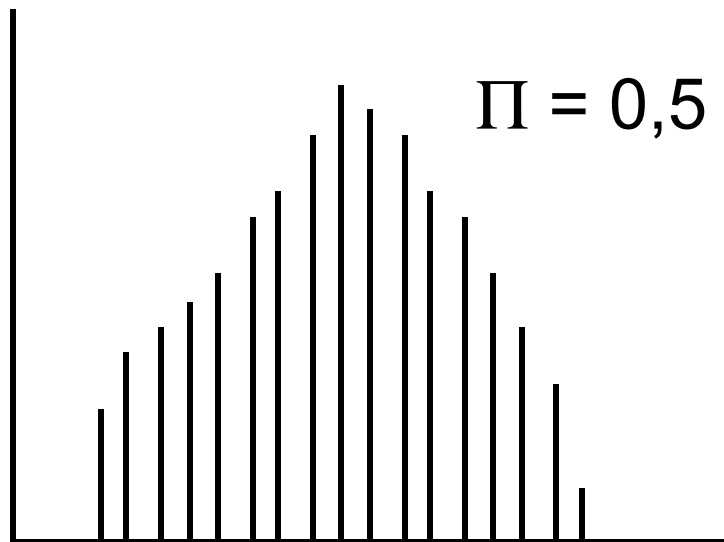


$X$  ..... celkový počet nastání jevu v  $n$  nezávislých pokusech

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

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

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



# Binomické rozložení jako model pro zkoumání výskytu šedivání u lidí

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

**X** ..... počet lidí s jistým symptomem

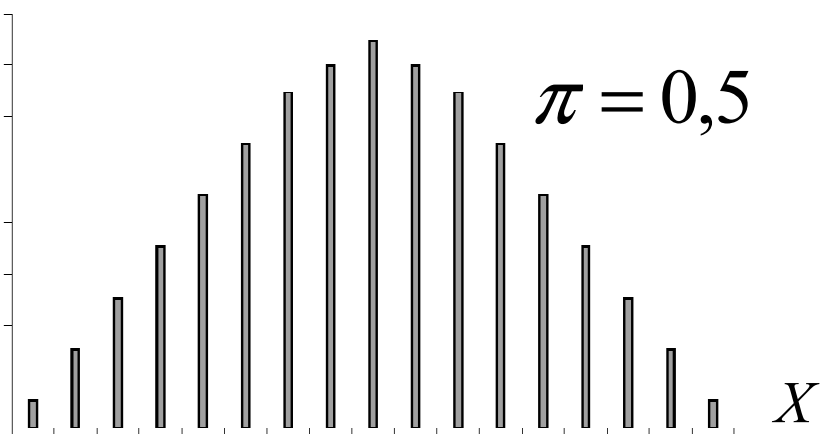
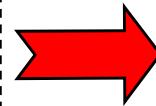
**r** znamená celkový počet nastání  
jevu v **n** nezávislých experimentech

**r** : 0 ..... n

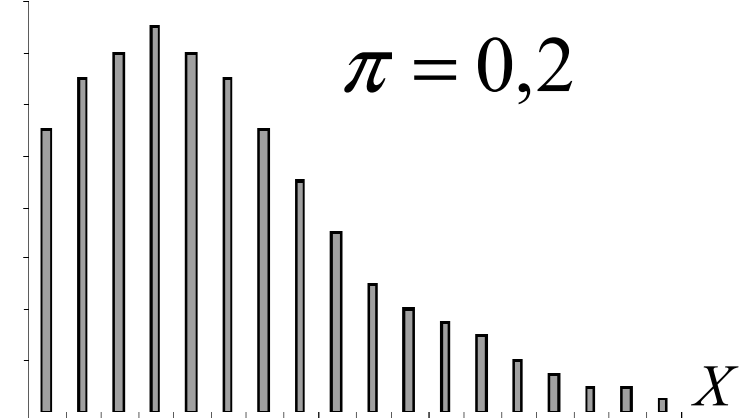
$p \sim \pi$  .. jediný parametr  
binomického rozložení

**p** .... relativní četnost nastání jevu

**p** ..... určuje tvar distribuce



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



**Binomická proměnná X**

# Binomické rozložení jako model



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

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

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

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

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

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

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

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

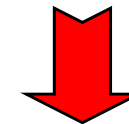
**X: Binomická proměnná**

**Střed rozložení:**

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

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

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



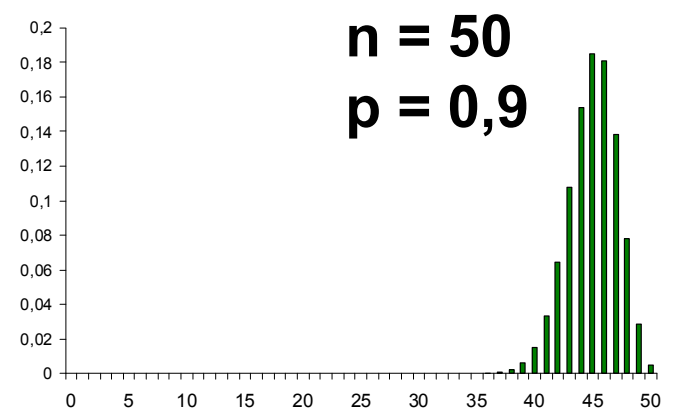
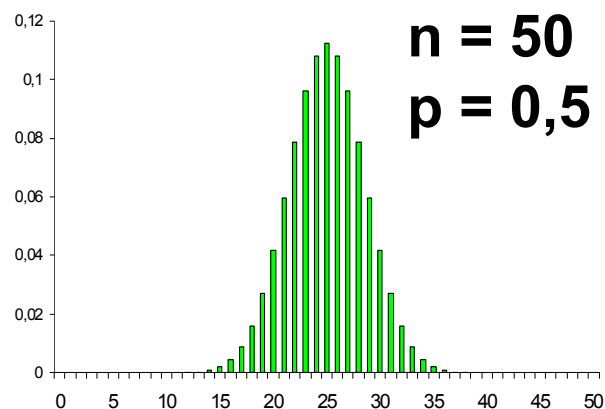
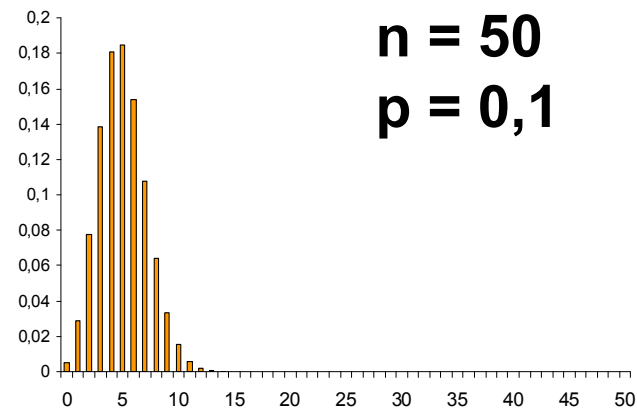
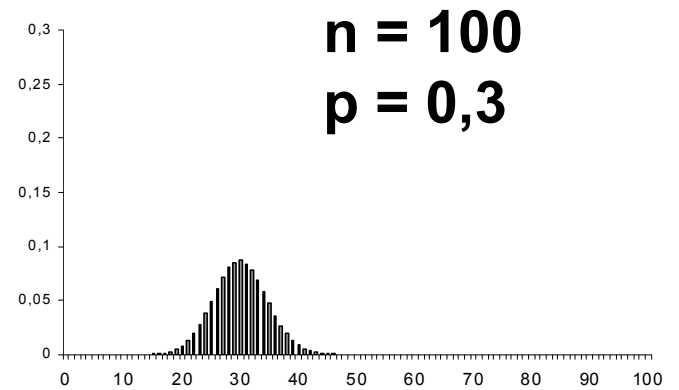
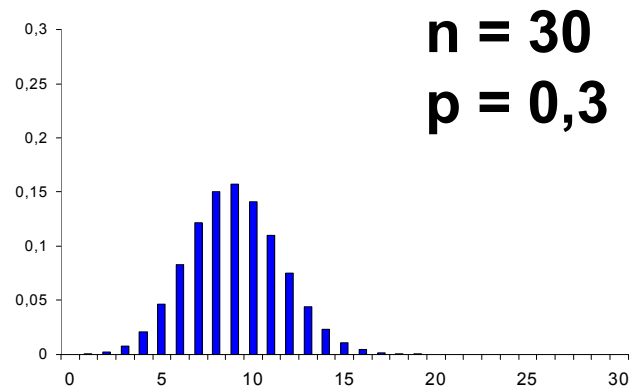
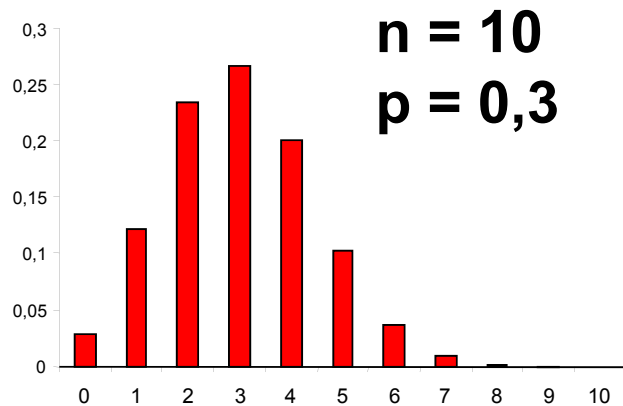
$$E(x) = n \cdot p = 20$$

**je střed rozložení  
a nejpravděpodobnější  
hodnota**

# Binomické rozložení jako model



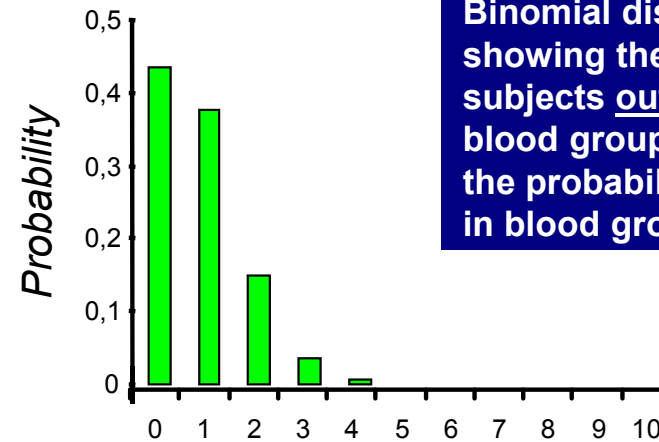
$$P(X = r) = \frac{n!}{r!(n-r)!} \cdot p^r \cdot q^{(n-r)} \quad q = 1 - p$$



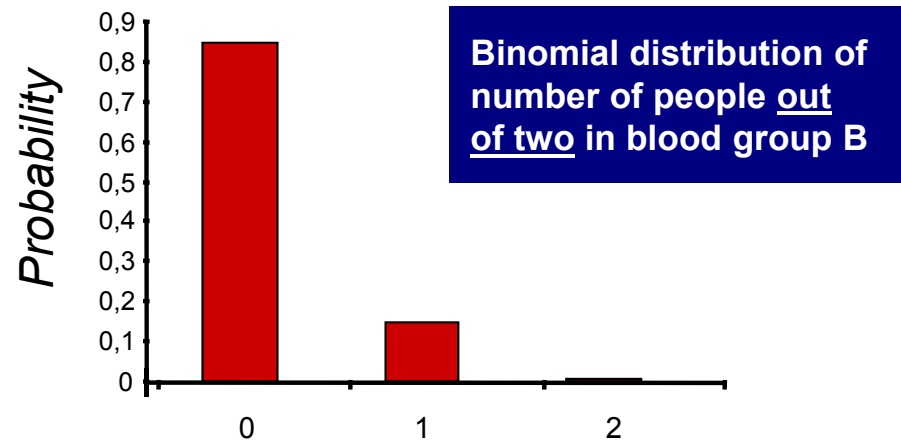
# Aplikace binomického rozložení

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

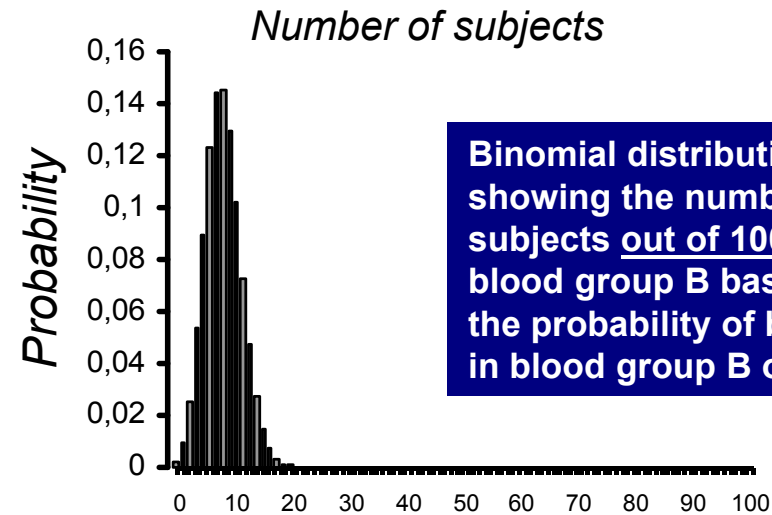
		<i>Number in blood group B</i>	<i>Probability</i>
B	B	2	0,0064
not B	B	1	0,0736
B	not B	1	0,0736
not B	not B	0	0,8464



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.

*Number: blood group B in 2 cases*

*Number of subjects*

# Aplikace binomického rozložení

**Populace:** 60% jedinců má zvýšenou hladinu cholesterolu

**Výběr:** 5 lidí

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

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

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

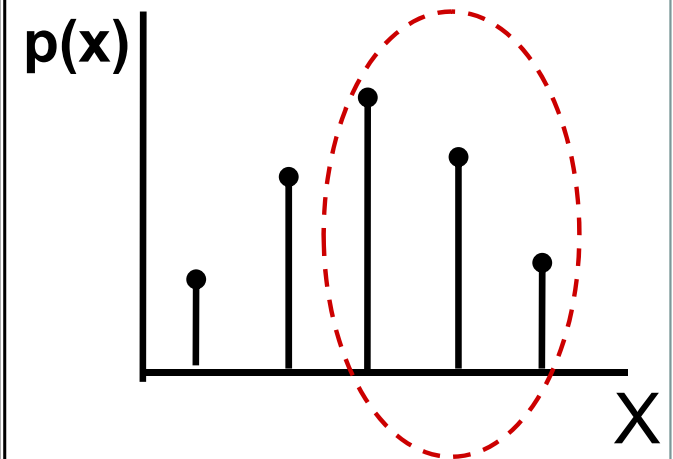
II. Jaká je P, že právě 3 lidé budou mít vyšší hladinu cholesterolu ? ~ Tzn. Výběr přesně odpovídá dané populaci ?

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

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

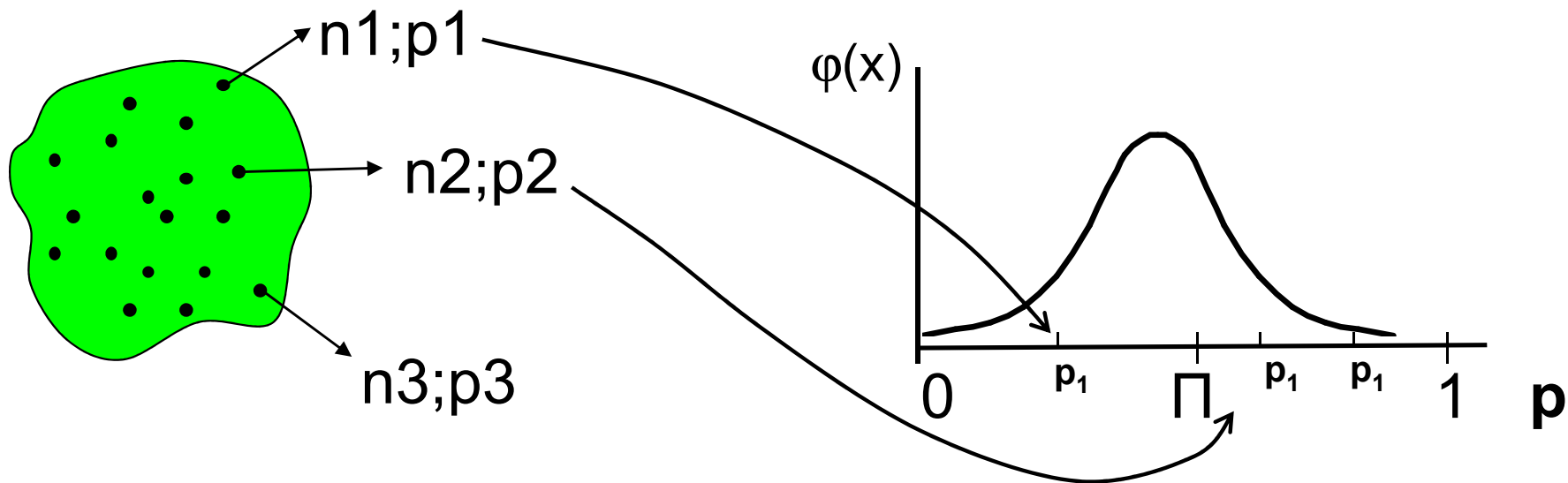
Jaká je P, že většina jedinců (tedy minimálně 3) má vyšší hladinu cholesterolu ? ~ Tzn. výběr alespoň obecně odpovídá zkoumané populaci ?

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

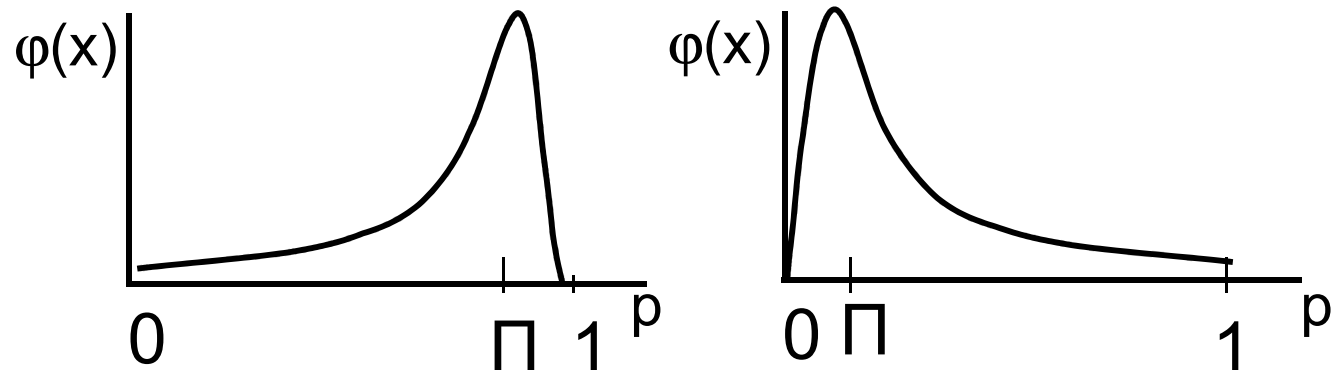


# Odhad parametru $\Pi$ binomického rozložení

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



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



# Odhad parametru $\pi$ binomického rozložení



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

1) Bodový

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

2) Intervalový – aproximace

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

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



# Odhad parametru 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 %**

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

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

$$0,504 \leq \pi \leq 0,697$$



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

# Odhad parametru $\pi$ binomického rozložení

## Intervalový odhad bez aproximací na normální rozložení

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



spodní limit intervalu

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

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



horní limit intervalu

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

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

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

# Odhad parametru $\Pi$ binomického rozložení:

## příklad I



Náhodný vzorek  $n = 200$  jedinců.

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

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

95% interval spolehlivosti = ?

### Spodní hranice

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

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

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

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

### Horní hranice

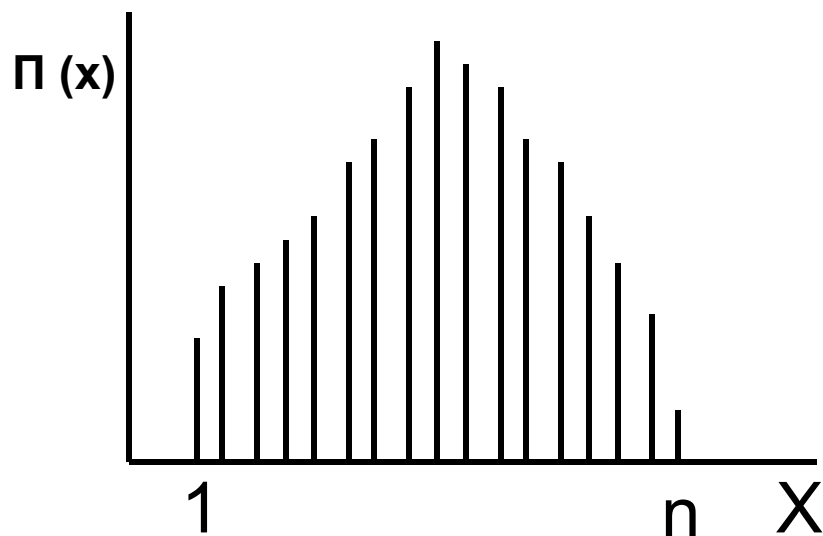
$$v'_1 = 2(r + 1) = 10$$

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

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

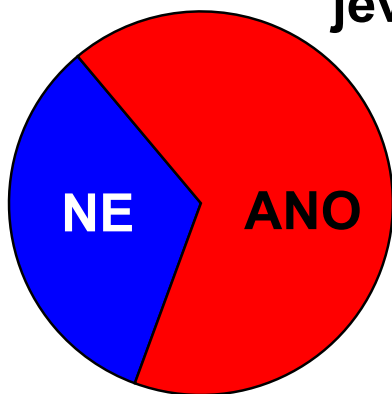
$$L_2 = \frac{(4 + 1) \cdot 2,08}{200 - 4 + (4 + 1) \cdot 2,08} = \underline{\underline{0,051}}$$

# Binomické rozložení v datech: vizualizace

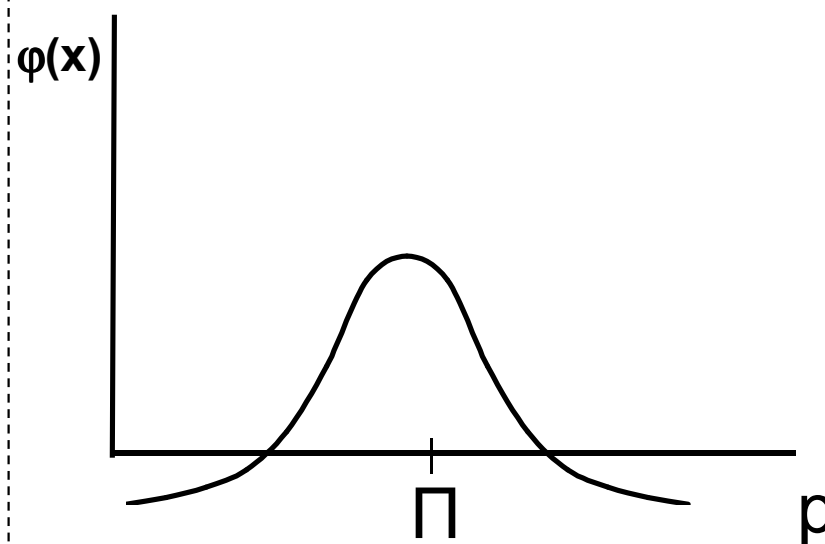
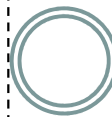


Pravděpodobnost výskytu hodnot  $X$

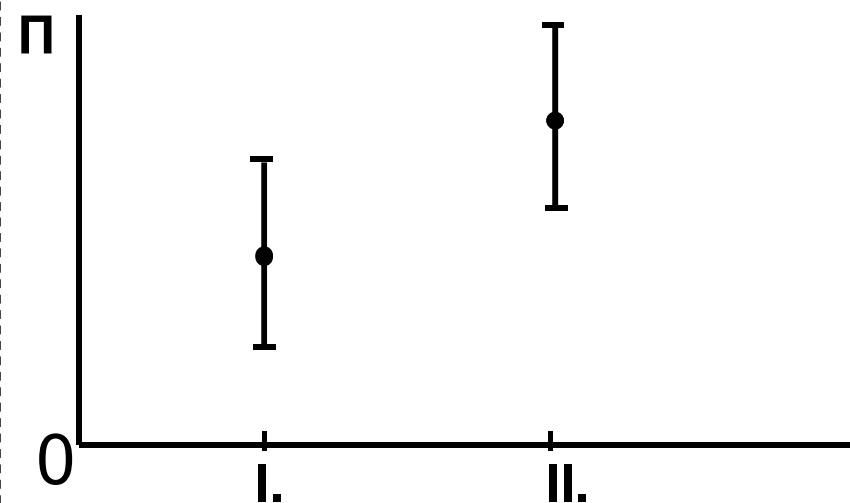
$n$  opakování      jev ANO  
jev NE



Binární podstata původních hodnot



Modelové rozložení odhadovaného parametru



Interval spolehlivosti pro  $\Pi$

# Statistické testování binomických dat

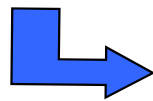


I.

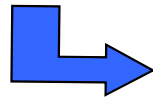
Liší se odhad  $\hat{p}$  od předpokládané hodnoty  $P$  ?

II.

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



- závislé odhady -



- nezávislé odhady -

III.

Je výskyt kategorií dvou jevů nezávislý ?

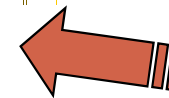
IV.

Hodnocení relativního rizika z výskytu určitého jevu v rámci skupiny lidí

# Jednovýběrový binomický test

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

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



Korekce na  
kontinuitu

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

# Test $\pi ? p$



- ✓ Stromy s pozmeněným tvarem koruny  
 $n = 9\,000$  jedinců  
 $r = 2\,250$  změněných jedinců

?

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

?

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

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

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

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

# Test $\pi ? p$

## 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)	0,0002 4	0,0029 3	0,0161 1	0,0537 1	0,1208 5	0,1933 5	0,2255 9	0,1933 6	0,1208 5	0,0537 1	0,0161 1	0,0029 3	0,0002 4

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

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

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

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



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



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

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

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

# Dvojitý binomický test ( $p_1 ? p_2$ )

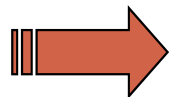
*Tento příklad je ukázkou testování rozdílů mezi dvěma binomickými populacemi (tedy srovnání dvou odhadů parametru  $p$ ).*

✓ 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	$\hat{p} = 0,510$	$\hat{q} = 0,490$

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

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

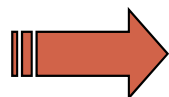


**Nezamítáme  $H_0$ :  $0,10 < P < 0,20$**

**S korekcí na kontinuitu:**

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

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



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

# Příklad I



**a)** 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  $H_0: p = \pi$ , hladina významnosti  $\alpha = 0,05$

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

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

**b)** Jaká je pravděpodobnost, že rodina se třemi dětmi bude mít 2 (3) chlapce?

Podrobně analyzujte problém a použijte obecného definičního vztahu pro binomické rozložení.

$$n = 3$$

$$r = 2$$

$p=0,5$  (stejná pravděpodobnost narození chlapce jako narození dívky)

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

$$P(2) = \binom{3}{2} \cdot 0,5^2 \cdot 0,5^{(1)} = \frac{3!}{2!(1)!} \cdot 0,5^2 \cdot 0,5^{(1)} = 0,375$$

**pravděpodobnost narození 2 chlapců v rodině se třemi dětmi je 0,375**

$$r = 3 \text{ platí } P(3) = \binom{3}{3} \cdot 0,5^3 \cdot 0,5^0 = 1 \cdot 0,5^3 \cdot 0,5^0 = 0,125$$

**pravděpodobnost narození 3 chlapců v rodině se třemi dětmi je 0,125**

# Příklad I



Předpokládá se, že lidé trpící určitou krevní chorobou mají abnormální jeden z chromozómů. S cílem odhadnout podíl takto postižených chromozómů bylo studováno 5 buněk od každého ze 120 pacientů a byl zjišťován počet buněk s postiženým chromozómem (tento počet = sledovaný jev =  $r$ ). Výsledky jsou uvedeny v následující tabulce. Odhadněte podíl postižených chromozómů u populace nemocných lidí.

(četnost jev)	0	1	2	3	4	5	celkem
f(poč.pacientů)	6	31	42	29	10	2	120

Pro odhad  $p$  se používá vztah 
$$\hat{p} = \frac{\sum_{i=1}^k f_i X_i}{\sum_{i=1}^k f_i} / n$$

$X_i$	$f_i$	$X_i f_i$
0	6	0
1	31	31
2	42	84
3	29	87
4	10	40
5	2	10

$$\sum_{i=1}^k f_i X_i = 252$$

$$\sum_{i=1}^k f_i = 120$$

$$n = 5$$



$$\hat{p} = \frac{252/120}{5} = 0,42$$

pravděpodobnost výskytu postiženého chromozómu