10. SEMINÁŘ

INDUKTIVNÍ STATISTIKA

3. HODNOCENÍ ZÁVISLOSTÍ

HODNOCENÍ ZÁVISLOSTÍ

KVANTITATIVNÍ VELIČINY

- Východiskem pro korelační a regresní analýzu je **bodový graf.**

KVALITATIVNÍ VELIČINY

- Vychází se z **kombinační (kontingenční) tabulky**, která je výsledkem třídění druhého stupně

- K měření stupně (síly) závislosti se používají různé míry (ukazatele) závislosti.
  - jejich použití je vázáno na různé podmínky
  - různá kvalita

Závislost kvantitativních znaků

- Z. funkční – určité hodnotě jedné veličiny odpovídá jediná hodnota veličiny druhé (př.O=4a)
- Z. statistická – jedné hodnotě znaku prvního odpovídá **celé rozložení hodnot** znaku druhého

Závislost – pojem komplexní

různá – intenzita (**síla**, stupeň, těsnost)

směr (přímý, nepřímý)

typ (tvar, druh)

BODOVÝ GRAF

- východisko pro posouzení vztahu mezi dvěma kvantitativními veličinami
- orientační informace o typu, směru a síle závislosti

- každou dvojici údajů znázorníme bodem, jehož souřadnice jsou rovny naměřeným hodnotám

  Osa x: nezávisle proměnná

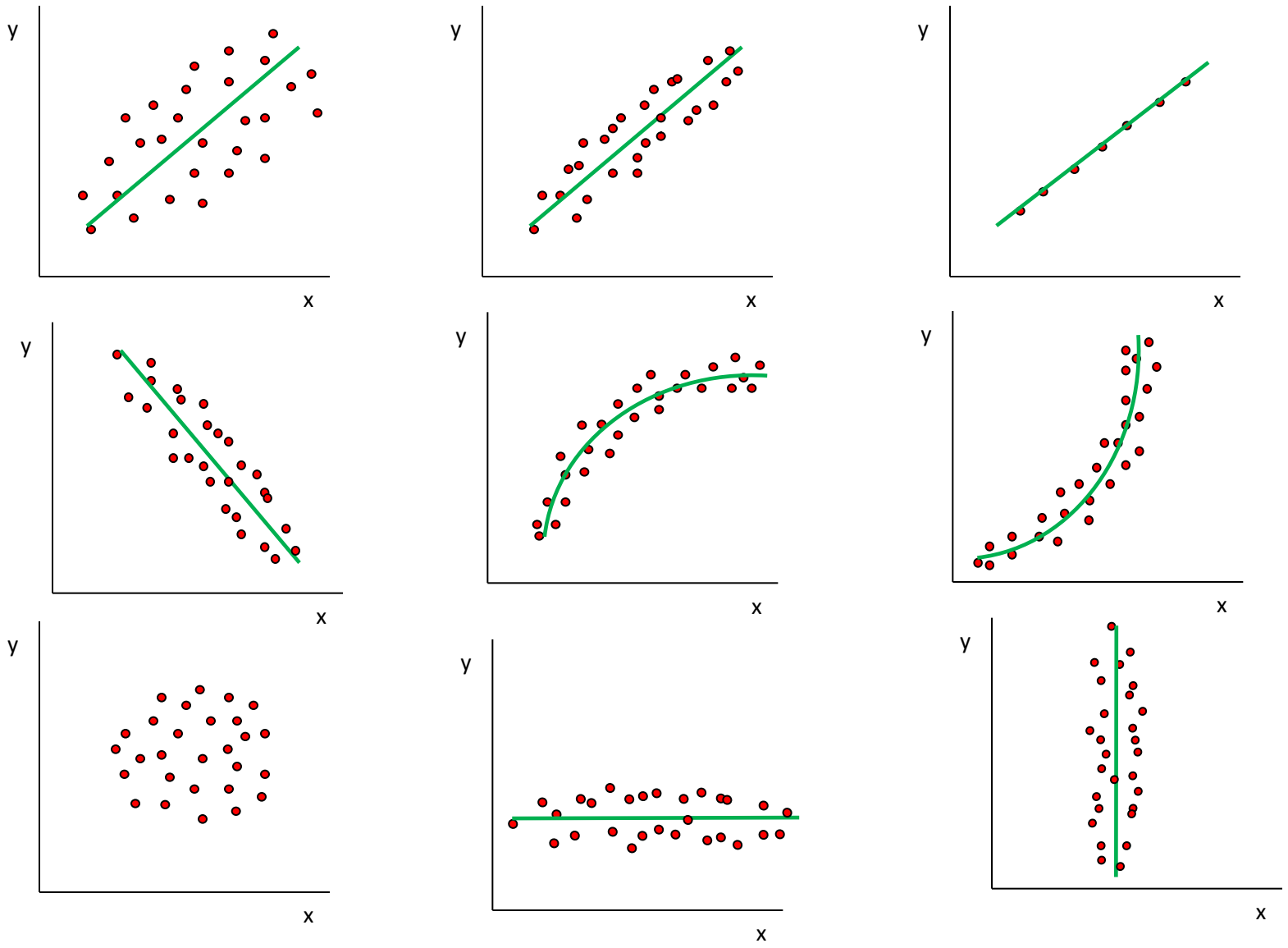
  Osa y: závisle proměnná

- zakreslenými body prokládáme čáru (přímku, křivku)

- **typ** závislosti (tvar křivky) - lineární
  - logaritmická
  - exponenciální
  - parabolická

(typ z. – křivka proložená empirickými body, vyjadřuje se matematickou fcí = **regresní křivka /regresní fce/**→ ***regresní analýza***

BODOVÝ GRAF



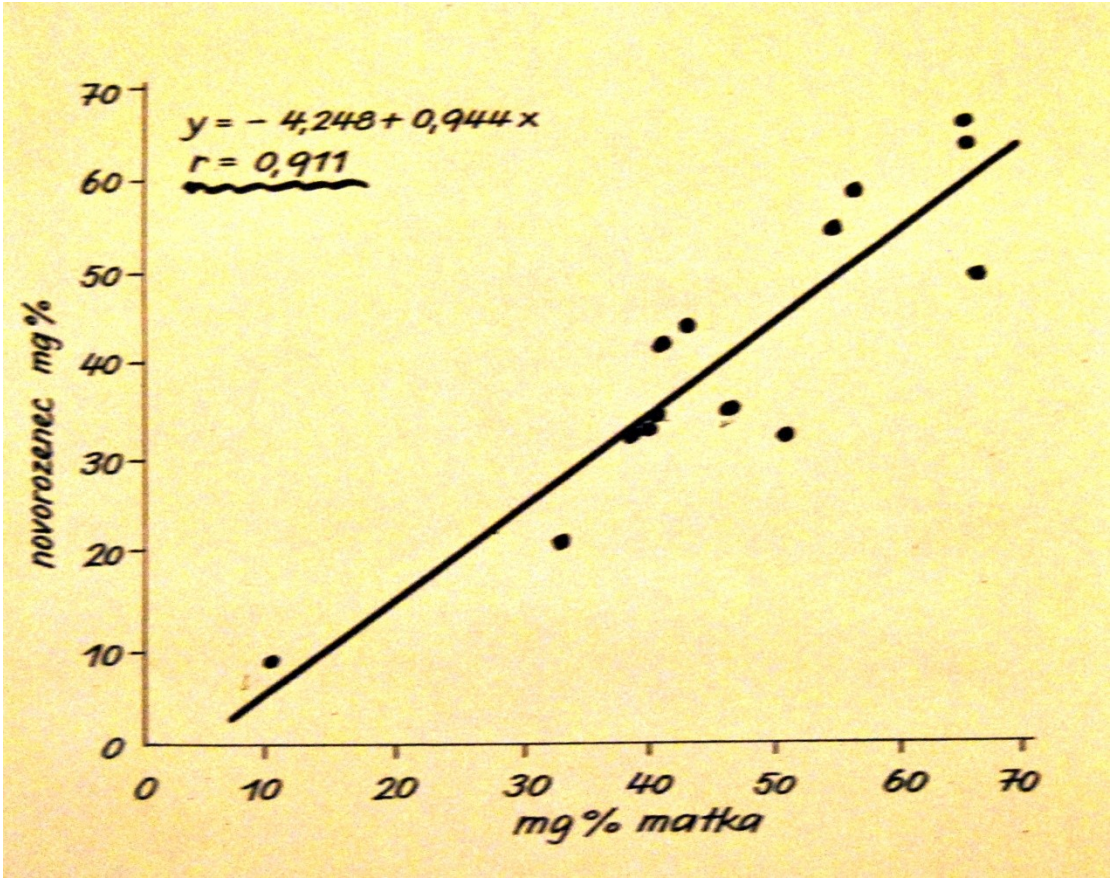
Příklad

Posuďte vztah mezi obsahem kyseliny mléčné v krvi matky a novorozence těsně po porodu (mg/100ml).

matka	**novorozenec**
x	**y**
39,0	31,8
6,5	34,5
41,1	33,7
43,0	43,0
33,5	21,0
11,2 *	9,0 *
40,2	32,6
50,9	32,0
66,5 *	48,7
54,7	48,2
66,4	62,4
64,7	64,7 *
56,8	6,8
40,9	40,9

Bodový graf

Závislost mezi obsahem kyseliny mléčné u novorozence a matky těsně po porodu.



HODNOCENÍ ZÁVISLOSTI KVANTITATIVNÍCH VELIČIN

LINEÁRNÍ ZÁVISLOST

Nejužívanější mírou korelace je PEARSONŮV KORELAČNÍ KOEFICIENT

NELINEÁRNÍ ZÁVISLOST

Např. SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE

LINEÁRNÍ ZÁVISLOST

• Hodnocením stupně lineární závislosti se zabývá **KORELAČNÍ ANALÝZA**

• Nejužívanější mírou korelace je **PEARSONŮV KORELAČNÍ KOEFICIENT - hodnotí těsnost (sílu) lineární vazby**

Označuje se **r** … pro výběrový soubor (výběrová charakteristika)

ρ… pro základní soubor (parametr)

Podmínka pro použití:

- lineární závislost (odhadujeme z bodového grafu)
- dvojrozměrné normální rozdělení
- obě sledované veličiny musí být náhodné

LINEÁRNÍ ZÁVISLOST - vlastnosti korelačního koeficientu

r(**ρ**) nabývá hodnot od **- 1 do 1**

Z tohoto intervalu mají hodnoty -1, 0 a 1 zvláštní význam:

r(ρ) = -1 funkční nepřímá závislost

r(ρ) = 0 neexistuje lineární závislost

r(ρ) = 1 přímá funkční závislost

Nabývá záporných hodnot pro nepřímou statist.z, a kladných hodnot pro přímou statist. závislost

Hodnocení **r**: Čím více se hodnota **r**(**ρ**) blíží **1**, tím je větší těsnost vztahu.

Pearsonův koeficient korelace je nejlepší mírou korelace, proto tam, kde je to možné, transformujeme nelineární vztah na lineární.

LINEÁRNÍ ZÁVISLOST

- Z údajů o výběrovém souboru vypočítáme VÝBĚROVÝ KORELAČNÍ KOEFICIENT **r**.

$$r = \frac{\sum(x_i - m_x)(y_i - m_y)}{n \cdot s_x \cdot s_y}$$

- **r** je výběrová charakteristika a proto je zatížena náhodnou chybou SE:

$$SE_r = \frac{1 - \rho^2}{\sqrt{n - 1}}$$

- **r** je nejlepším bodovým odhadem neznámého parametru **ρ**
- Pozor při intervalovém odhadu – pokud **ρ ≠ 0** nemá **r** normální rozdělení, je třeba provést logaritmickou transformaci

TEST HYPOTÉZY O NULOVÉM KORELAČNÍM KOEFICIENTU

- Jde o zjištění významnosti **r**, tj. zda je zjištěná závislost dílem náhody nebo zda skutečně existuje i v základním souboru.

  H_0 - veličiny jsou nezávislé, tj. **$r(\rho) = 0$**
  H_A - veličiny jsou závislé, tj. **$r(\rho) \neq 0$**

- Statistická hypotéza zjišťuje, zda se **r** významně liší od nuly – k tomu lze využít:

  a) pro n ≤ 50: kritické hodnoty Pearsonova r

  Absolutní hodnota r se porovná s kritickými hodnotami Pearsonova korelačního koeficientu:

  - je-li $|r| < k.h.$, pak nezamítáme H_0
  - je-li $|r| \geq k.h.$, pak zamítáme H_0

  b) pro n > 50: u-test $u = r \cdot \sqrt{n-1}$

TEST HYPOTÉZY O NULOVÉM KORELAČNÍM KOEFICIENTU

Pro soubory (pro n > 50) → **u-test**

1) $H_0 \equiv \rho = 0$ → veličiny jsou nezávislé

   $H_A \equiv \rho \neq 0$ → veličiny jsou závislé

2) **Testovací charakteristika u** má za platnosti Ho **normální rozdělení**

$$u = r \cdot \sqrt{n - 1}$$

3) **|u|** → kritické hodnoty NR: 1,96; 2,58

TEST HYPOTÉZY O NULOVÉM KORELAČNÍM KOEFICIENTU

Interpretace:

$|u| > 2{,}58$ … zamítám Ho na 1% HV

$|u| > 1{,}96$ … zamítám Ho na 5% HV

a přijímám H_A: hodnocené veličiny jsou závislé

$|u| < 1{,}96$ …Ho nezamítám →

závislost se nepodařilo prokázat

Pozn. Pro malé výběry ($n < 50$) není možno použít u- test →

tabulka kritických hodnot korelačního koeficientu (skr. str. 28).

Kritické hodnoty k.k. r_k představují pro daný rozsah výběru a zvolenou hladinu významnosti nejmenší hodnotu výběrového k.k. **r,** pro níž se už Ho zamítá

Příklad

V souboru 225 jednoletých brněnských chlapců byl sledován vztah mezi tělesnou délkou a hmotností. Výpočtem jsme zjistili **r = 0,648**.

Zhodnoťte závislost pomocí u-testu.

Řešení

r = 0,648, n = 225

u = r . √ n – 1= 0,648.√225-1= **9,7**

u = 9,7

|u|(9,7) >2,58 → zamítáme Ho na 1% HV,

přijímáme H_A, která říká že tělesná výška a hmotnost u 1-letých chlapců jsou veličiny závislé, riziko, že se mýlíme je menší než 1%

Příklad:

Zhodnoťte závislost obsahu kyseliny mléčné v krvi matky a novorozence těsně po porodu (viz naměřené hodnoty v úvodu).

Zhodnoťte statistickou významnost korelačního koeficientu r.

r = 0,911 **n = 14**

Tab. 2. Kritické hodnoty Pearsonova korelačního koeficientu
(pro rozsah výběru n a hladinu významnosti 0,05 a 0,01)

n	0,05	0,01	n	0,05	0,01	n	0,05	0,01
3	0,9969	0,9999	14	0,5324	0,6614	25	0,3961	0,5052
4	0,9500	0,9900	15	0,5140	0,6411	30	0,3610	0,4629
5	0,8783	0,9587	16	0,4973	0,6226	35	0,3338	0,4296
6	0,8114	0,9172	17	0,4822	0,6055	40	0,3120	0,4026
7	0,7545	0,8745	18	0,4683	0,5897	45	0,2940	0,3801
8	0,7067	0,8343	19	0,4555	0,5751	50	0,2787	0,3610
9	0,6664	0,7977	20	0,4438	0,5614	60	0,2542	0,3301
10	0,6319	0,7646	21	0,4329	0,5487	70	0,2352	0,3060
11	0,6021	0,7348	22	0,4227	0,5368	80	0,2199	0,2864
12	0,5760	0,7079	23	0,4132	0,5256	90	0,2072	0,2702
13	0,5529	0,6835	24	0,4044	0,5151	100	0,1966	0,2565

Příklad - Řešení

$r = 0{,}911$ $n = 14$

$H_0 \equiv \rho = 0$ veličiny jsou nezávislé

Kritické hodnoty (viz tabulka)

$r_{0,05}(14) = 0{,}5324$

$r_{0,01}(14) = 0{,}6614$

$r > 0{,}6614 \Rightarrow$ **H_0 zamítáme** na 1% HV

Prokázali jsme závislost mezi kyselinou mléčnou u matek a novorozenců

TEST HYPOTÉZY O NULOVÉM KORELAČNÍM KOEFICIENTU

Příklad:

Zhodnoťte významnost korelace mezi podílem dětí s nízkou porodní hmotností a kojeneckou úmrtností

(posuďte statistickou významnost korelačního koeficientu r.)

a) v souboru 14 okresů JmK, když r = 0,429

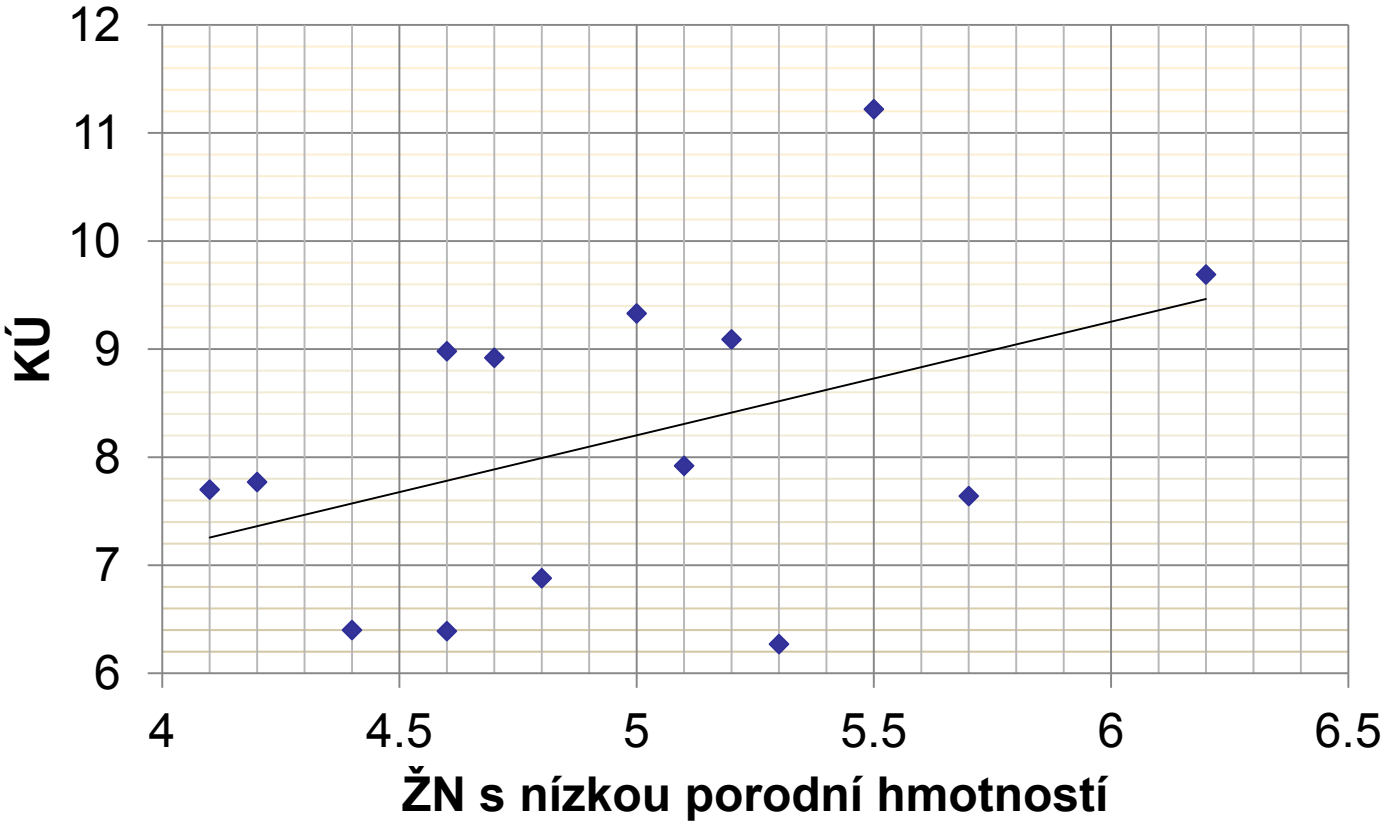
b) v souboru 76 okresů ČR, když r = 0,471

a) Kritické hodnoty Pearsonova korelačního koeficientu

b) u-test, $u = r \cdot \sqrt{n-1}$, kritické hodnoty normálního rozdělení

Okresy JmK	Por. hmotnost do 2500g na 100 ŽN (nezávis. prom. X)	KÚ (závis. prom. Y)
Blansko	4,10	7,70
Brno – město	6,20	9,69
Brno – venkov	5,00	9,33
Břeclav	4,40	6,40
Hodonín	4,20	7,77
Jihlava	4,60	8,98
Kroměříž	5,30	6,27
Prostějov	5,50	11,22
Třebíč	4,80	6,88
Uherské Hradiště	4,70	8,92
Vyškov	5,20	9,09
Zlín	5,10	7,92
Znojmo	5,70	7,64
Žďár nad Sázavou	4,60	6,39

Bodový graf



Řešení:

Ad **a)** 14 okresů na 5 % HV

$n < 50 \Rightarrow$ vypočítáme Pearsonův koeficient a porovnáme jej s kritickou hodnotou (dle tabulky)

r = 0,429

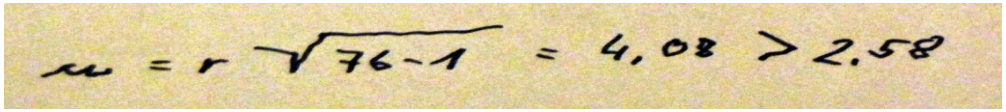
r < 0,5324 => **H0 nezamítáme**

Nepodařilo se prokázat závislost mezi kojeneckou úmrtností a podílem dětí s nízkou porodní hmotností.

Ad **b**) 76 okresů

$n > 50 \Rightarrow$ vypočítáme **testovací charakteristiku u**, kterou porovnáme s kritickými hodnotami normálního (Gaussova) rozdělení

r = 0,471



=> **zamítáme Ho na 1% HV,**

prokázali jsme existenci vztahu mezi KÚ a podílem dětí s nízkou PH.

KOEFICIENT DETERMINACE

- V případě stat. významné závislosti můžeme počítat tzv. **KOEFICIENT DETERMINACE: r^2**
- Nabývá hodnot od 0 do 1; vyjádříme-li ho v %, udává, **kolik % variability závislé veličiny Y lze vysvětlit změnami v nezávislé veličině X.**
- **$100 . r^2$** udává procento variability náhodné veličiny **Y**, která připadá na vrub lineární závislosti veličiny **Y** na veličině **X.**

KOEFICIENT DETERMINACE r^2

Příklad:

Jestliže těsnost vztahu mezi hmotností a tělesnou délkou jednoletých chlapců vyjadřuje korelační koeficient **r = 0,648**, pak **42%** celkové variability hmotnosti jednoletých chlapců připadá na vrub závislosti hmotnosti na délce. Znamená to, že variabilita vah jednoletých chlapců určité délky by byla o **42%** nižší než variabilita celková (pro chlapce všech délek).

REGRESNÍ ANALÝZA (rovnice regresní přímky)

Zjistíme-li významnou lineární závislost, je užitečné vyjádřit ji pomocí regresní přímky ve tvaru: **y = a + bx**

y hodnota závislé veličiny
x hodnota nezávislé veličiny
a regresní koeficient, udává posun přímky na ose y
b regresní koeficient, úhel přímky s osou x (sklon přímky)

$$\mathbf{b} = r\,(s_y/s_x)$$

$$\mathbf{a} = m_y - b.\,m_x$$

Přímka se používá k PREDIKCI jedné veličiny pomocí druhé, tzn. zjišťujeme jaká bude hodnota y, pro určenou hodnotu x.

REGRESNÍ ANALÝZA (rovnice regresní přímky)

Příklad:

Určete rovnici regresní přímky pro laktát

1) Regresní koeficienty **a, b**

b = 0,911 x (14,94/14,90) = **0,945**

a = 39,95 – (0,945 x 46,81) = **- 4,285**

2) Rovnice regresní přímky

y = - 4,285 + (0,945 . x)

(y lze vypočítat pro libovolně zvolené x)

Z této rovnice lze vypočítat prům. obsah laktátu v krvi novorozence **(y)**

pro libovolně zvolené množství laktátu v krvi matky **(x)**

- např.: při hladině laktátu v krvi matky **30 mg (x)**, lze u jejího novorozence očekávat prům. hodnotu laktátu **24,06 mg (y)**

REGRESNÍ ANALÝZA

Příklad:

V souboru 76 okresů ČR byla zjištěna závislost mezi podílem dětí s nízkou porodní hmotností (X) a kojeneckou úmrtností (Y), kterou lze vyjádřit rovnicí: $y = 4{,}139 + 0{,}942x$. Vypočítejte, jaká by byla kojenecká úmrtnost v okrese, kde na 100 živě narozených připadá 7 dětí s nízkou porodní hmotností.

76 okresů ČR

x = podíl dětí s nízkou p.h. na 100 ŽN (v %)

y = KÚ

Rovnice regresní přímky vypočítaná z dat o výběrovém souboru se chová jako náhodná veličina a je zatížená náhodnou výběrovou chybou SE.

Pro odhad regresní přímky - tzv. **PÁS SPOLEHLIVOSTI**
(CI pro každý bod přímky)

Řešení příkladu

KÚ = 4,139 + (0,942 x 7) = **10,73 ‰**

Interpretace: V okrese, kde podíl dětí s nízkou p.h. na 100 živě narozených je 7%, lze očekávat hodnotu KÚ **10,73 ‰.**

NELINEÁRNÍ ZÁVISLOST

Pro hodnocení nelineární závislosti používáme:

1) Transformace – (např. transformace inverzní, logaritmická)

pomocí vhodné funkce se nelineární závislost převede na lineární, tzv. se „zlinearizuje“

2) Pořadový korelační koeficient

nejsou vázány na žádné předpoklady, lze užít u jakékoliv závislosti

(**Spearmanův**, Kendallův)

NELINEÁRNÍ ZÁVISLOST

SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE

*Východiskem je **rozdíl pořadí u každé dvojice**, bez ohledu na znaménko*

Postup:

- Nejprve seřadíme všechny hodnoty veličiny **X** dle velikosti a označíme je pořadovými čísly.
- Pak seřadíme všechny hodnoty veličiny **Y** dle velikosti a označíme je pořadovými čísly.
- Pro každou dvojici hodnot **x,y** stanovíme jejich rozdíl **d.**
- **Spearmanův koeficient pořadové korelace** vypočítáme ze vztahu:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

n = počet dvojic údajů

d = rozdíl pořadí u každé dvojice

NELINEÁRNÍ ZÁVISLOST

r_s nabývá hodnot od -1 do 1, opět platí, že když:

$r_s = 0$, jde o nezávislost

$r_s = 1$, jde o přímou funkční závislost

$r_s = -1$, jde o nepřímou funkční závislost

Hodnocení r_s: Čím více se hodnota **$r_s(\rho_s)$** blíží **± 1**, tím je větší těsnost vztahu.

TEST VÝZNAMNOSTI

Absolutní hodnota r_s se porovná s kritickými hodnotami Spearmanova koeficientu pořadové korelace:

- je-li $|r_s| < \text{k. h.}$, pak nezamítáme H_0
- je-li $|r_s| \geq \text{k. h.}$, pak zamítáme H_0

SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE

matka x	pořadí hodnot x	novorozenec y	pořadí h. y	rozdíl pořadí
39,0	3	31,8	3	0
6,5	8	34,5	7	1
41,1	6	33,7	6	0
43,0	7	43,0	9	2
33,5	2	21,0	2	0
11,2 *	1	9,0 *	1	0
40,2	4	32,6	5	1
50,9	9	32,0	4	5
66,5 *	14	48,7	11	3
54,7	10	48,2	10	0
66,4	13	62,4	13	0
64,7	12	64,7 *	14	2
56,8	11	6,8	12	1
40,9	5	40,9	8	3

SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE - výpočet

$\sum d_i^2 = 54$

$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} = \underline{0,881}$

pro $n = 14$ $\quad r_{0,05} = 0,5341 \quad r_{0,01} = 0,6747$

$\underline{H_0 (\rho_s = 0)}$ zamítáme na 1% HV

$\Rightarrow$ veličiny jsou ZÁVISLÉ !

Kritické hodnoty

Tab. 2. Kritické hodnoty Pearsonova korelačního koeficientu
(pro rozsah výběru n a hladinu významnosti 0,05 a 0,01)

n	0,05	0,01	n	0,05	0,01	n	0,05	0,01
3	0,9969	0,9999	14	0,5324	0,6614	25	0,3961	0,5052
4	0,9500	0,9900	15	0,5140	0,6411	30	0,3610	0,4629
5	0,8783	0,9587	16	0,4973	0,6226	35	0,3338	0,4296
6	0,8114	0,9172	17	0,4822	0,6055	40	0,3120	0,4026
7	0,7545	0,8745	18	0,4683	0,5897	45	0,2940	0,3801
8	0,7067	0,8343	19	0,4555	0,5751	50	0,2787	0,3610
9	0,6664	0,7977	20	0,4438	0,5614	60	0,2542	0,3301
10	0,6319	0,7646	21	0,4329	0,5487	70	0,2352	0,3060
11	0,6021	0,7348	22	0,4227	0,5368	80	0,2199	0,2864
12	0,5760	0,7079	23	0,4132	0,5256	90	0,2072	0,2702
13	0,5529	0,6835	24	0,4044	0,5151	100	0,1966	0,2565

Tab. 3. Kritické hodnoty Spearmanova koeficientu pořadové korelace
(pro rozsah výběru n a hladinu významnosti 0,05 a 0,01)

N	0,05	0,01	n	0,05	0,01	n	0,05	0,01
			11	0,6091	0,7545	21	0,4351	0,5545
			12	0,5804	0,7273	22	0,4241	0,5426
			13	0,5549	0,6978	23	0,4150	0,5306
			14	0,5341	0,6747	24	0,4061	0,5200
5	0,9000	-	15	0,5179	0,6536	25	0,3977	0,5100
6	0,8286	0,9429	16	0,5000	0,6324	26	0,3894	0,5002
7	0,7450	0,8929	17	0,4853	0,6152	27	0,3822	0,4915
8	0,6905	0,8571	18	0,4716	0,5975	28	0,3749	0,4828
9	0,6833	0,8167	19	0,4579	0,5825	29	0,3685	0,4744
10	0,6364	0,7818	20	0,4451	0,5684	30	0,3620	0,4665

Tab. 4. Kritické hodnoty rozdělení χ^2 pro počet stupňů volnosti *f* a hladinu významnosti 0,05 a 0,01

f	0,05	0,01	*f*	0,05	0,01
1	3,84	6,63	11	19,68	24,73
2	5,99	9,21	12	21,03	26,22
3	7,81	11,35	13	22,36	27,69
4	9,49	13,28	14	23,69	29,14
5	11,07	15,09	15	25,00	30,58
6	12,59	16,81	16	26,30	32,00
7	14,07	18,48	17	27,59	33,31
8	15,51	20,09	18	28,87	34,81
9	16,92	21,67	19	30,14	36,19
10	18,31	23,21	20	31,41	37,57

HODNOCENÍ ZÁVISLOSTI KVALITATIVNÍCH ZNAKŮ

Kvalitativní (kategoriální) veličiny – lze popsat slovním určením

a) ***alternativní*** - 2 obměny (např. zdravý- nemocný, očkovaný- neočkovaný, HIV pozit. – HIV negat.,)

b) ***množné*** – více než 2 obměny (např. rodinný stav, diagnoza..)

!!! Exaktní definice obměny každé veličiny !!!

Východisko pro hodnocení závislosti kvalit.veličin (X,Y) tvoří **roztřídění** podle obou hledisek do **kombinační (kontingenční) tabulky** (určitý počet řádků a sloupců – podle počtu obměn každé veličiny – **r** řádků, **k** sloupců)

HODNOCENÍ ZÁVISLOSTI KVALITATIVNÍCH ZNAKŮ

- Východiskem je **kontingenční tabulka:**

KUŘÁCTVÍ	VZDĚLÁNÍ ZŠ	SŠ	VŠ	CELKEM
Nekuřák	269	74	46	389
Slabý kuřák	213	44	17	274
Silný kuřák	197	50	14	261
CELKEM	679	168	77	924

- Je založeno na **srovnání empirických a teoretických četností.**
- **Empirická četnost (E)** – rozdělení lidí podle kuřáctví a vzdělání jak bylo skutečně zjištěno ve výběrovém souboru.
- **Teoretická četnost (T)** – jaké by bylo rozdělení lidí ve výběrovém souboru podle kuřáctví a vzdělání, kdyby šlo o jevy nezávislé.

Závislost dvou alternativních znaků

- Závislost rizikového činitele (jevu A -kouření cigaret) a určité nemoci Ca plic- jevu B)

- 4 zákl. možnosti: a - osoba kouří – onemocněla
  b - osoba kouří – neonemocněla
  c – osoba nekouří – onemocněla
  d – osoba nekouří – neonemocněla

Schema čtyřpolní (2x2) kontingenční tabulky
(model vztahu jevu A a B)

Osoby	S nemocí	Bez nemoci	∑
Kouří	a	b	a+b
Nekouří	c	d	c+d
∑	a+c	b+d	a+b+c+d

Závislost dvou alternativních znaků

Osoby	S nemocí	Bez nemoci	∑
Kouří	a	b	a+b
Nekouří	c	d	c+d
∑	a+c	b+d	a+b+c+d

- **Kdy usuzuji na závislost jevu A a B?**

1. Kuřáci onemocněli Ca plic častěji než nekuřáci
   a/a+b > c/c+d

2. Mezi nemocnými je větší počet kuřáků než mezi zdravými
   a/a+c > b/b+d

(Statistickou významnost rozdílů uvedených pravděpodobností lze posoudit např. u- testem pro srovnání pravděpodobností).

Hodnocení závislosti kvalitativních znaků

- **založeno na srovnání empirických a teoretických četností**

  posoudím, zda se velikost skupin a,b,c,d (tzv. **zjištěné** – **empirické** četnosti) statisticky významně liší od tzv. **očekávaných** (**teoretických**) četností – a0, b0, c0, d0.

  Výpočet Oi : předpokládáme nezávislost znaků A a B (tj. Ho – testovaná hypotéza) při daných okrajových četnostech (a+b, c+d, a+c, b+d).

- *Pravidlo pro násobení pravděpodobností dvou nezávislých jevů* – pokud A,B jsou 2 nezávislé jevy, pak pravděpodobnost současného výskytu obou jevů je dána součinem pravděpodobností **$P(A \cap B) = P(A) \cdot P(B)$**

(**Očekávané četnosti vypočítáme** tak, že mezi sebou vynásobíme příslušné okrajové četnosti a získaný součin dělíme celkovým počtem statistických jednotek).

Hodnocení závislosti kvalitativních znaků

- východiskem je kontingenční tabulka – zjištěné (empirické) četnosti Ei

	Nemoc B	Nemoc B	
	ano	ne	CELKEM
Rizik. činitel A ano	24 (a)	56 (b)	80
Rizik.činitel A ne	16 (c)	104 (d)	120
CELKEM	40	160	200

- **Empirická(zjištěná) četnost - Ei** – rozdělení lidí podle přítomnosti nemoci(B) a kuřáckého zvyku(A), jak bylo skutečně zjištěno ve výběrovém souboru
- **Teoretická (očekávaná) četnost - Oi**– jaké by bylo rozdělení lidí ve výběrovém souboru podle kuřáckého zvyku a přítomnosti nemoci, kdyby šlo o jevy nezávislé (za platnosti Ho o nezávislosti jevů)

Hodnocení závislosti kvalitativních znaků (tabulka zjištěných četností)

	Nemoc B	Nemoc B	
	ano	ne	CELKEM
Rizik. činitel A ano	24 (a)	56 (b)	80
Rizik.činitel A ne	16 (c)	104 (d)	120
CELKEM	40	160	200

- Výpočet očekávaných četností:

ao = (a+b).(a+c)/n = (80.40)/200=**16**
bo = (a+b).(b+d)/n = (80.160)/200 = **64**
co = (a+c).(c+d)/n = (40.120)/200=**24**
do = (b+d).(c+d)/n = (160.120/200= **96**

Hodnocení závislosti kvalitativních znaků

- založeno na srovnání zjištěných a očekávaných četností pomocí

  chí-kvadrát testu (χ^2) = test dobré shody

- Očekávané četnosti sledovaných skupin

  Teoretická (očekávaná) četnost - Oi– jaké by bylo rozdělení lidí ve výběrovém souboru podle kuřáckého zvyku a přítomnosti nemoci, kdyby šlo o jevy nezávislé (za platnosti Ho o nezávislosti znaků A a B)

	Nemoc B	Nemoc B	
	ano	ne	CELKEM
Rizik. činitel A ano	16 (a_0)	64 (b_0)	80
Rizik. činitel A ne	24 (c_0)	94 (d_0)	120
CELKEM	40	160	200

TEST HYPOTÉZY O NEZÁVISLOSTI

- Platnost Ho o nezávislosti znaků A a B ověříme pomocí testovací charakteristiky χ^2 (chí kvadrát), která má za platnosti Ho rozložení χ^2 s počtem stupňů volnosti f = 1.

  $f = (r-1) .(k-1)$

1. STANOVENÍ HYPOTÉZ

- H_0 – zkoumané jevy jsou nezávislé
- H_A - zkoumané jevy jsou závislé

2. HLADINA VÝZNAMNOSTI

$\alpha = 5\%$ nebo $\alpha = 1\%$

3. VÝBĚR TESTU

- chí-kvadrát test (χ^2)

TEST HYPOTÉZY O NEZÁVISLOSTI

4. VÝPOČET TESTOVACÍ CHARAKTERISTIKY CHÍ – KVADRÁT

a) Pro každé políčko vypočítáme teoretickou četnost

b) Pro každé políčko vypočítáme rozdíl mezi empirickou (E) a teoretickou četností (T) podle vzorečku:

$$\frac{(E-T)^2}{T}$$

c) Součet vypočítaných rozdílů je hodnota chí-kvadrátu:

$$\chi^2 = \sum \frac{(E-T)^2}{T}$$

5. PODMÍNKY PRO POUŽITÍ TESTU

Všechny **teoretické četnosti** musí být **větší než 5**.

TEST HYPOTÉZY O NEZÁVISLOSTI

6. SROVNÁNÍ S KRITICKÝMI HODNOTAMI

- **Testovací charakteristiku chí-kvadrát (χ^2)** srovnáme s příslušnými **kritickými hodnotami** chí-kvadrát rozdělení:
  - kritické hodnoty určujeme z tabulek podle zvolené hladiny významnosti a tzv. stupňů volnosti:

$$f = (ř - 1)(s - 1)$$

7. ZAMÍTNEME NEBO NEZAMÍTNEME NULOVOU HYPOTÉZU

$\chi^2 < \text{k. h.}$, nezamítáme H_0

$\chi^2 \geq \text{k. h.}$, zamítáme H_0

8. INTERPRETACE VÝSLEDKU

Chí-kvadrát – informuje **pouze o stat. významnosti** zjištěné asociace, ne o těsnosti vztahu! Čím větší je hodnota χ^2, tím menší je riziko chyby při zamítnutí H_0.

Hodnocení závislosti kvalitativních znaků

Lze předpokládat tyto výsledky testu:

a) mezi empirickými a teoretickými četnostmi **není statisticky významný rozdíl**, zjištěné rozdíly nejsou natolik velké, aby nemohly být způsobeny **náhodou**

(*kritická hodnota je větší než vypočítaná*)

b) mezi empirickými a teoretickými četnostmi je statisticky **významný** **rozdíl**, zjištěné rozdíly jsou natolik velké, že nemohou být způsobeny náhodou

E_i	O_i	$E_i - O_i$	$(E_i - O_i)^2$	$\frac{(E_i - O_i)^2}{O_i}$	
24	16	8	64	4,000	$(\frac{64}{16}$
56	64	-8	64	1,000	$(\frac{64}{64})$
16	24	-8	64	2,667	$(\frac{64}{24})$
104	96	8	64	0,667	$(\frac{64}{96})$

$$\sum = \underline{8{,}334} = \chi^2$$

Výsledek a interpretace příkladu

Tab. 4. Kritické hodnoty rozdělení χ^2 pro počet stupňů volnosti *f* a hladinu významnosti 0,05 a 0,01

f	0,05	0,01	*f*	0,05	0,01
1	3,84	6,63	11	19,68	24,73
2	5,99	9,21	12	21,03	26,22
3	7,81	11,35	13	22,36	27,69
4	9,49	13,28	14	23,69	29,14
5	11,07	15,09	15	25,00	30,58
6	12,59	16,81	16	26,30	32,00
7	14,07	18,48	17	27,59	33,31
8	15,51	20,09	18	28,87	34,81
9	16,92	21,67	19	30,14	36,19
10	18,31	23,21	20	31,41	37,57

- $\chi^2 = 8{,}33$
- f = (2-1). (2-1) = 1
- 5% k.h. = 3,84, 1% k.h.= 6,63
- 8,33 > 6,63 →

Ho zamítáme na 1% HV,

přijímáme H_A o závislosti jevů A a B (závislost mezi kouřením a Ca plic).

Závislost dvou alternativních znaků

- Pro čtyřpolní tabulku (typu 2x2) můžeme hodnotu veličiny χ^2 počítat jednodušeji (lze použít, jsou-li všechny očekávané četnosti >5):

$$\chi^2 = \frac{(a.d - b.c)^2 . n}{(a+b)(c+d)(a+c)(b+d)} =$$

$$= \frac{(24 \cdot 104 - 16 \cdot 56)^2 \cdot 200}{80 \cdot 120 \cdot 40 \cdot 160} = \mathbf{8,33}$$

Testovací charakteristika χ^2 *(chí kvadrát)*

- aproximativní test, slouží pouze k zamítnutí či nezamítnutí Ho o nezávislosti

  (neříká nic o směru ani síle z.)
- důkaz objektivní existence závislosti,

  ne příčinnosti!
- jsou-li rozdíly mezi teor. a emp. četnostmi příliš velké, je malá pst, že se mýlíme, když Ho zamítáme a preferujeme H_A **(veličiny jsou závislé)**
- jsou-li rozdíly mezi četnostmi malé ($\rightarrow 0$), Ho nezamítáme, asociaci se nepodařilo prokázat

Příklad k samostatnému řešení

V porodnici byla provedena retrospektivní studie, která hodnotila vztah mezi způsobem výživy (A1,A2) novorozenců a výskytem novorozeneckého ikteru u 210 novorozenců.

A1 = nový způsob výživy – mléko se podávalo častěji, v kratších intervalech, **A2 = starý způsob výživy**, méně časté podávání.

Úkol:

Testem χ^2 posuďte statistickou významnost závislost dvou hodnocených jevů, výsledek interpretujte.

způsob výživy	výskyt ikteru +	-	součet
A_1	61	49	110
A_2	85	15	100
součet	146	64	210

Tabulka teoretických četností
(za předpokladu platnosti Ho)

způsob výživy	výskyt ikteru +	-	součet
A_1	76	34	110
A_2	70	30	100
součet	146	64	210

Pomocná tabulka pro výpočet hodnoty χ^2

E_i	O_i	$E_i - O_i$	$(E_i - O_i)^2$	$\frac{(E_i - O_i)^2}{O_i}$
61	76	−15	225	2,96
49	34	15	225	6,61
85	70	15	225	3,21
15	30	−15	225	7,5
			$\Sigma =$	20,28 = χ^2

- k.h. (0,05) = 3,84 f = (2 -1) . (2-1) = 1
- k.h. (0,01) = 6,63
- **χ^2 (20,28) > 6,63** → zamítáme Ho o nezávislosti na 1% HV, přijímáme HA, **prokázali jsme statisticky významnou závislost mezi způsobem výživy novorozence a výskytem novorozeneckého ikteru.**

Kritické hodnoty

Tab. 2. Kritické hodnoty Pearsonova korelačního koeficientu
(pro rozsah výběru n a hladinu významnosti 0,05 a 0,01)

n	0,05	0,01	n	0,05	0,01	n	0,05	0,01
3	0,9969	0,9999	14	0,5324	0,6614	25	0,3961	0,5052
4	0,9500	0,9900	15	0,5140	0,6411	30	0,3610	0,4629
5	0,8783	0,9587	16	0,4973	0,6226	35	0,3338	0,4296
6	0,8114	0,9172	17	0,4822	0,6055	40	0,3120	0,4026
7	0,7545	0,8745	18	0,4683	0,5897	45	0,2940	0,3801
8	0,7067	0,8343	19	0,4555	0,5751	50	0,2787	0,3610
9	0,6664	0,7977	20	0,4438	0,5614	60	0,2542	0,3301
10	0,6319	0,7646	21	0,4329	0,5487	70	0,2352	0,3060
11	0,6021	0,7348	22	0,4227	0,5368	80	0,2199	0,2864
12	0,5760	0,7079	23	0,4132	0,5256	90	0,2072	0,2702
13	0,5529	0,6835	24	0,4044	0,5151	100	0,1966	0,2565

Tab. 3. Kritické hodnoty Spearmanova koeficientu pořadové korelace
(pro rozsah výběru n a hladinu významnosti 0,05 a 0,01)

N	0,05	0,01	n	0,05	0,01	n	0,05	0,01
			11	0,6091	0,7545	21	0,4351	0,5545
			12	0,5804	0,7273	22	0,4241	0,5426
			13	0,5549	0,6978	23	0,4150	0,5306
			14	0,5341	0,6747	24	0,4061	0,5200
5	0,9000	-	15	0,5179	0,6536	25	0,3977	0,5100
6	0,8286	0,9429	16	0,5000	0,6324	26	0,3894	0,5002
7	0,7450	0,8929	17	0,4853	0,6152	27	0,3822	0,4915
8	0,6905	0,8571	18	0,4716	0,5975	28	0,3749	0,4828
9	0,6833	0,8167	19	0,4579	0,5825	29	0,3685	0,4744
10	0,6364	0,7818	20	0,4451	0,5684	30	0,3620	0,4665

Tab. 4. Kritické hodnoty rozdělení χ^2 pro počet stupňů volnosti *f* a hladinu významnosti 0,05 a 0,01

f	0,05	0,01	*f*	0,05	0,01
1	3,84	6,63	11	19,68	24,73
2	5,99	9,21	12	21,03	26,22
3	7,81	11,35	13	22,36	27,69
4	9,49	13,28	14	23,69	29,14
5	11,07	15,09	15	25,00	30,58
6	12,59	16,81	16	26,30	32,00
7	14,07	18,48	17	27,59	33,31
8	15,51	20,09	18	28,87	34,81
9	16,92	21,67	19	30,14	36,19
10	18,31	23,21	20	31,41	37,57

Děkuji za pozornost,

a přeji **hodně úspěchů**
u zápočtového testu a
hlavně u **zkoušky** !!!

Závislost kvalitativních znaků množných

- Př. : Bylo vyšetřeno 152 pacientů s cílem zjistit, zda existuje **závislost určitého typu zhoubného nádoru na jeho lokalizaci.**

  (- typ nádoru označen I,II,III

  - lokalizace nádoru A,B,C)

  Nádory byly roztříděny podle typu zhoubného nádoru a podle jeho lokalizace.

  Výsledkem třídění je kontingenční tab. 3x3

Tab.- výsledky vyšetření 152 pacientů (**zjištěné** tj. **empirické četnosti**)

		Typ nádoru			
		I	II	III	Σ
LOKALIZACE	A	28	10	6	44
	B	22	4	4	30
	C	35	25	18	78
Σ		85	39	28	152

Tab.- **očekávané četnosti**

		Typ nádoru			
		I	II	III	Σ
LOKALIZACE	A	24,6	11,3	8,1	44,0
	B	16,8	7,7	5,5	30,0
	C	43,6	20,0	14,4	78,0
Σ		85,0	39,0	28,0	152,0

Pomocná tabulka pro výpočet χ^2

e_{ij}	o_{ij}	$e_{ij} - o_{ij}$	$(e_{ij} - o_{ij})^2$	$\frac{(e_{ij} - o_{ij})^2}{o_{ij}}$
28	24,6	3,4	11,56	0,470
10	11,3	−1,3	1,69	0,150
6	8,1	−2,1	4,41	0,544
22	16,8	5,2	27,04	1,610
4	7,7	−3,7	13,69	1,778
4	5,5	−1,5	2,25	0,409
35	43,6	−8,6	73,96	1,696
25	20,0	5,0	25,00	1,250
18	14,4	3,6	12,96	0,900

$$\chi^2 = \sum_i \frac{(e_{ij} - o_{ij})^2}{o_{ij}} = \underline{\underline{8,807}}$$

Závěr a interpretace

Vypočítanou hodnotu testovací charakteristiky χ^2 (**8,807**) porovnáme s kritickou hodnotou pro 4 stupně volnosti / f= (3-1)(3-1)= 4/ a pro hladinu významnosti 0,05

$chí^2_{0,95}$ (4) = 9,49

8,807 < 9,49 → hypotézu o nezávislosti nezamítáme, závislost mezi typem zhoubného nádoru a jeho lokalizací se nepodařilo prokázat.