10. SEMINÁŘ

INDUKTIVNÍ STATISTIKA

3. HODNOCENÍ ZÁVISLOSTÍ

HODNOCENÍ ZÁVISLOSTÍ

KVALITATIVNÍ VELIČINY

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

KVANTITATIVNÍ VELIČINY

- Východiskem pro korelační a regresní analýzu je bodový graf.

- K měření stupně (síly) závislosti se používají různé míry (ukazatele) závislosti.
  - Jejich použití je vázáno na splnění určitých podmínek.

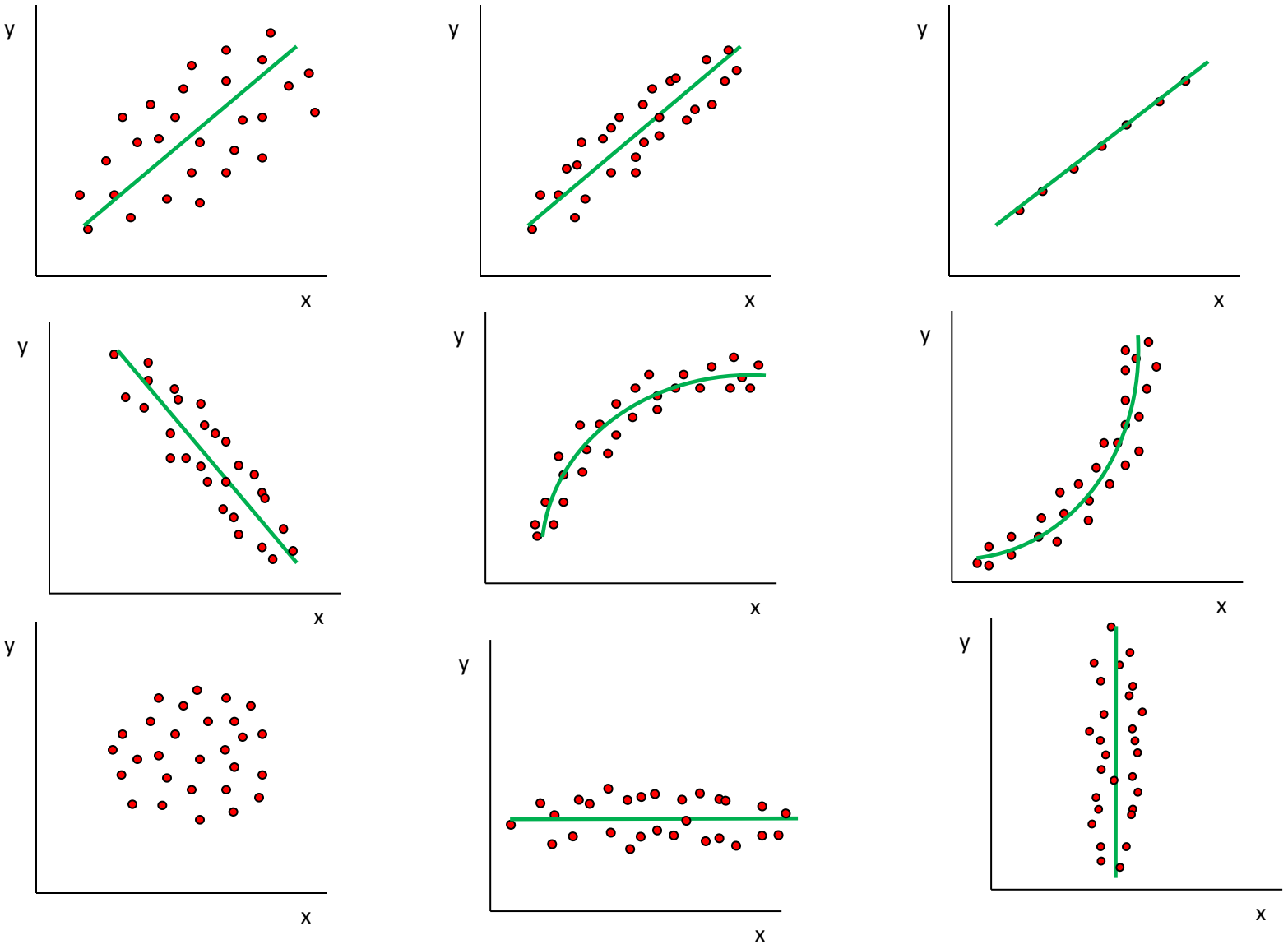
Okresy JmK	Počet dětí s nízkou por. hmotností (do 2500g) na 100 ŽN (nezávis. prom. X)	KÚ (závis. prom. Y)
Blansko	4,10	7,70
Brno – město	6,20	9,69
Brno – venkov	5,00	9,33
Břeclav	4,40	6,40
Hodonín	4,20	7,77
Jihlava	4,60	8,98
Kroměříž	5,30	6,27
Prostějov	5,50	11,22
Třebíč	4,80	6,88
Uherské Hradiště	4,70	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

- Body jsou dány dvojicí hodnot pro každou statistickou jednotku
  - Osa x: nezávisle proměnná
  - Osa y: závisle proměnná
- Zakreslenými body prokládáme čáru (přímku, křivku)

- **Typ** závislosti (funkce)
- **Směr** závislosti (přímá, nepřímá)
- **Těsnost** závislosti (rozptyl bodů)

BODOVÝ GRAF



HODNOCENÍ ZÁVISLOSTI KVANTITATIVNÍCH VELIČIN

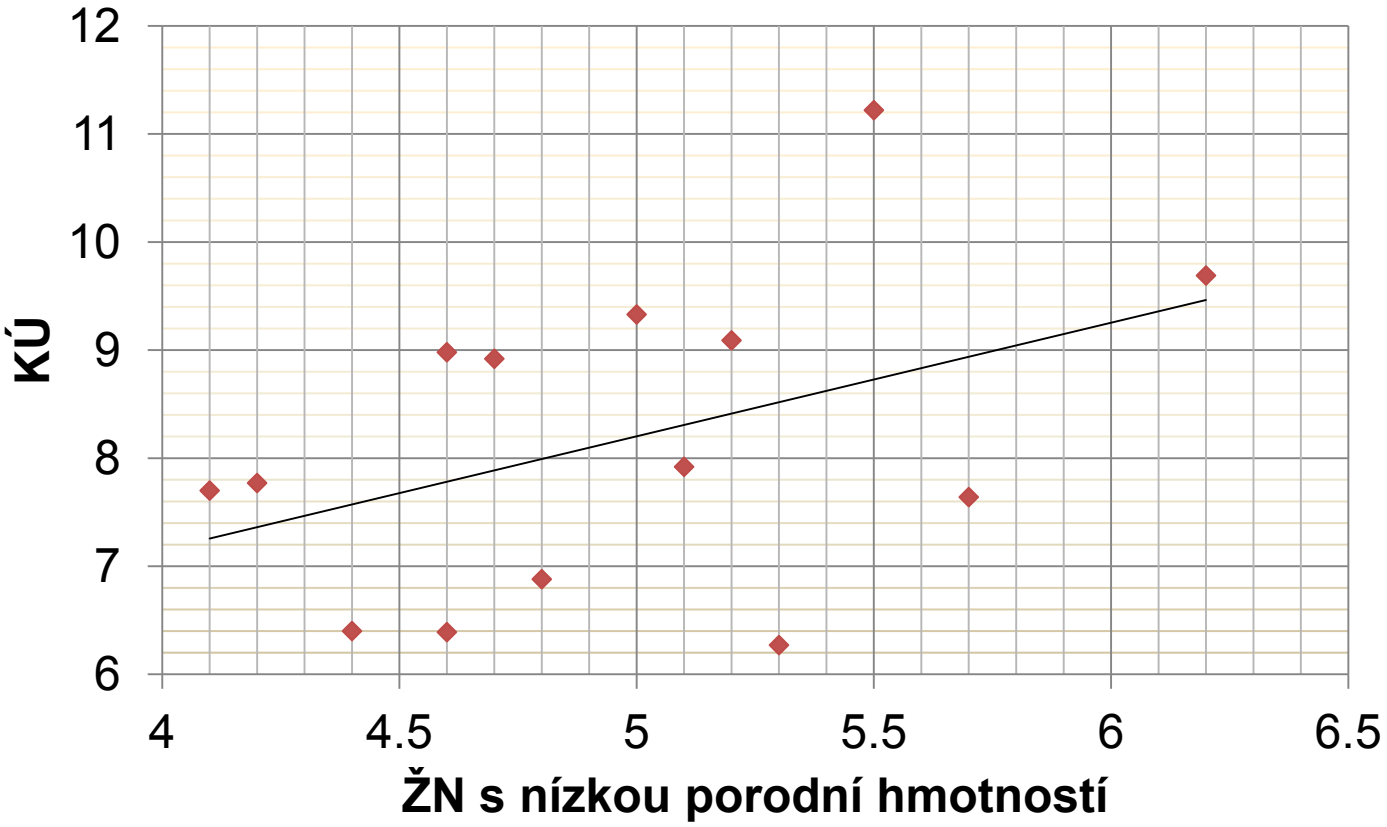
LINEÁRNÍ ZÁVISLOST

Nejužívanější mírou korelace je **PEARSONŮV KORELAČNÍ KOEFICIENT**

NELINEÁRNÍ ZÁVISLOST

Např. **SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE**

Bodový graf



LINEÁRNÍ ZÁVISLOST

Nejužívanější mírou korelace je **PEARSONŮV KORELAČNÍ KOEFICIENT**

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

LINEÁRNÍ ZÁVISLOST

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

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 statistické významnosti **r**

  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

Příklad:

Zhodnoťte významnost korelace mezi podílem dětí s nízkou porodní hmotností a kojeneckou úmrtností a) v souboru 14 okresů, když r = 0,429 a b) v souboru 72 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

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.**

Vypočítejte, z kolika % jsou rozdíly v KÚ mezi okresy ČR způsobeny rozdíly v podílu dětí s nízkou por. hmotností.

REGRESNÍ ANALÝZA

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

y hodnota závislé veličiny
x hodnota nezávislé veličiny
a regresní koeficient, udává posun po ose y
b regresní koeficient, úhel přímky s osou x

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

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.

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 slouží tzv. **PÁS SPOLEHLIVOSTI** (CI pro každý bod přímky).

NELINEÁRNÍ ZÁVISLOST

SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE

- 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:

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

Okresy JmK	Por. hmotnost do 2500g na 100 ŽN	Pořadí Podle PH	KÚ	Pořadí Podle KÚ	Rozdíl pořadí
Blansko	4,10	1	7,70	6	- 5
Brno – město	6,20	14	9,69	13	1
Brno – venkov	5,00	8	9,33	12	- 4
Břeclav	4,40	3	6,40	3	0
Hodonín	4,20	2	7,77	7	- 5
Jihlava	4,60	4,5	8,98	10	- 5,5
Kroměříž	5,30	11	6,27	1	10
Prostějov	5,50	12	11,22	14	- 2
Třebíč	4,80	7	6,88	4	3
Uherské Hradiště	4,70	6	8,92	9	- 3
Vyškov	5,20	10	9,09	11	- 1
Zlín	5,10	9	7,92	8	1
Znojmo	5,70	13	7,64	5	8
Žďár nad Sázavou	4,60	4,5	6,39	2	2,5

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

HODNOCENÍ ZÁVISLOSTI KVALITATIVNÍCH ZNAKŮ

- Východiskem je **kontingenční tabulka:**

KUŘÁCTVÍ	VZDĚLÁNÍ ZŠ	SŠ	VŠ	CELKEM
Nekuřák	269	74	46	389
Kuřák	410	94	31	535
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.

TEST HYPOTÉZY O NEZÁVISLOSTI

1. STANOVENÍ HYPOTÉZ

- $\mathbf{H_0}$ – mezi empirickými a teoretickými četnostmi není rozdíl
- $\mathbf{H_A}$ - mezi empirickými a teoretickými četnostmi je rozdíl

2. HLADINA VÝZNAMNOSTI

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

3. VÝBĚR TESTU

- chí-kvadrát test (χ^2)

TEST HYPOTÉZY O NEZÁVISLOSTI

4. PODMÍNKY PRO POUŽITÍ TESTU

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

5. VÝPOČET TESTOVACÍ CHARAKTERISTIKY CHÍ - KVADRÁT

a) Pro každé políčko tabulky vypočítáme teoretickou četnost.

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

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

c) Součet vypočítaných rozdílů je hodnota chí-kvadrátu:

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

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. ZAMÍTÁME NEBO NEZAMÍTÁME 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.

ZÁVISLOST KVALITATIVNÍCH ZNAKŮ

- Vyhodnoťte statistickou významnost závislosti mezi kouřením a vzděláním na 5% hladině významnosti.

KUŘÁCTVÍ	VZDĚLÁNÍ ZŠ	SŠ	VŠ	CELKEM
Nekuřák	269	74	46	389
Kuřák	410	94	31	535
CELKEM	679	168	77	924

- Pro každé políčko tabulky této šestipolní vypočítejte teoretickou četnost
- Pro každé políčko vypočítejte rozdíl mezi empirickou a teoretickou četností: $\frac{(E-T)^2}{T}$
- Proveďte součet dílčích výpočtů pro jednotlivá políčka: $\chi^2 = \sum \frac{(E-T)^2}{T}$
- Srovnejte hodnotu χ^2 s příslušnými kritickými hodnotami v tabulce.