

# **Statistické metody a zpracování dat**

## **VII. Korelační a regresní počet**

*Petr Dobrovolný*

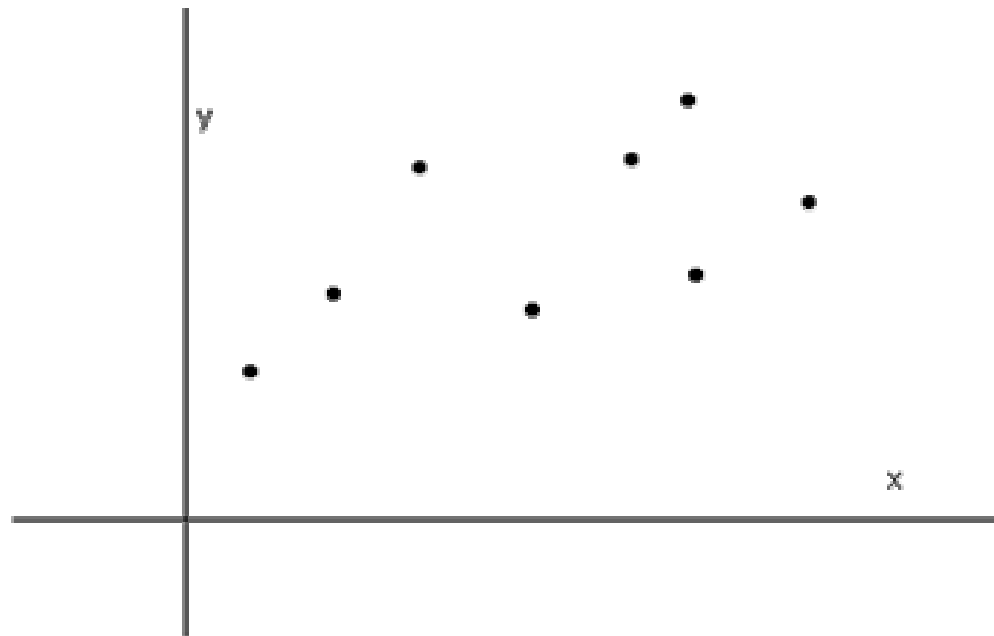
# K čemu to je dobré?

## Analýza závislostí

- V řadě geografických disciplín studujeme jevy, u kterých vyšetřujeme ne jednu jejich vlastnost (znak), ale znaků několik.
- Tyto znaky mohou být navzájem závislé.
- Cílem této části statistiky je vyšetřovat, do jaké míry spolu dva či více statistických znaků souvisí.
- Do jaké míry změna hodnoty jednoho znaku podmiňuje změnu hodnot znaku jiného.

# Příklady použití

*Př. Vztah mezi teplotou vzduchu a nadmořskou výškou, mezi množstvím srážek a velikostí odtoku, mezi výnosy a hodnotami několika meteorologických prvků, mezi počtem dojíždějících a vzdáleností od centra dojížd'ky, ...*



# Analýza závislostí

- Předmětem statistické analýzy v tomto případě bude stanovení **síly závislosti** a **druhu závislosti**
- Analýzou síly závislosti statistických znaků se zabývá **korelační počet**
- Analýzou druhu závislosti statistických znaků se zabývá **regresní počet**
- Budeme tedy pracovat s dvourozměrnými soubory
- **Korelační i regresní počet** však lze využít i pro studium vícerozměrných souborů, pro studium znaků kvantitativních i kvalitativních.

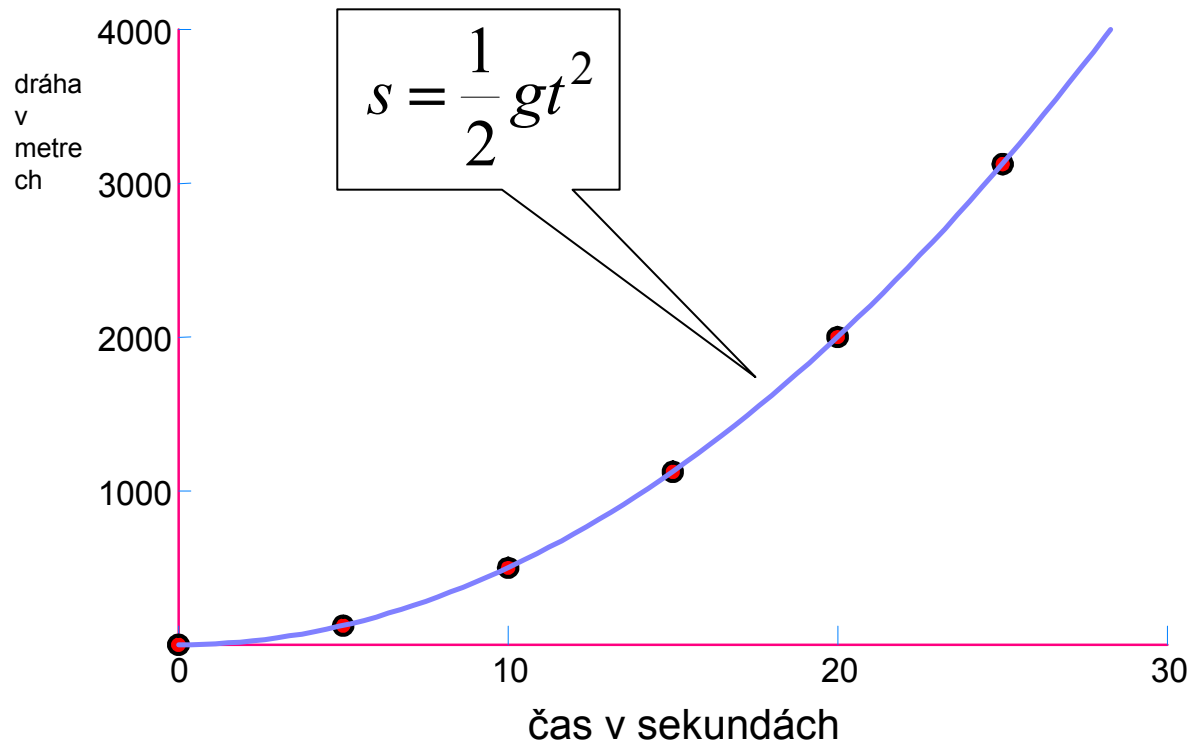
# Druhy závislostí

- **Vztahy jednostranné:** Změna statistického znaku jednoho souboru náhodné veličiny - tzv. **nezávisle** proměnné (x) podmiňuje změnu statistického znaku souboru druhé náhodné veličiny - tzv. **závisle** proměnné (y).
- V tomto případě jde o vztahy příčiny a následku
- **Vztahy vzájemné:** Nelze rozlišit mezi souborem závisle a nezávisle proměnné (např. vztah hodnot teploty vzduchu na dvou sousedních stanicích)
- V geografii – tzv. **prostorová autokorelace**

Druhy závislostí:

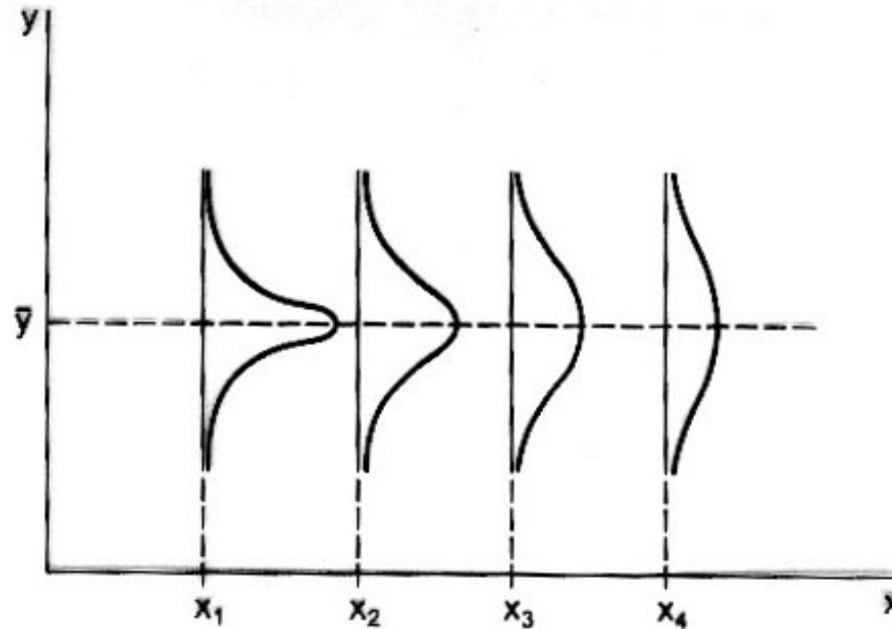
- Závislost funkční
- Závislost statistická
- Závislost korelační

# Závislost funkční



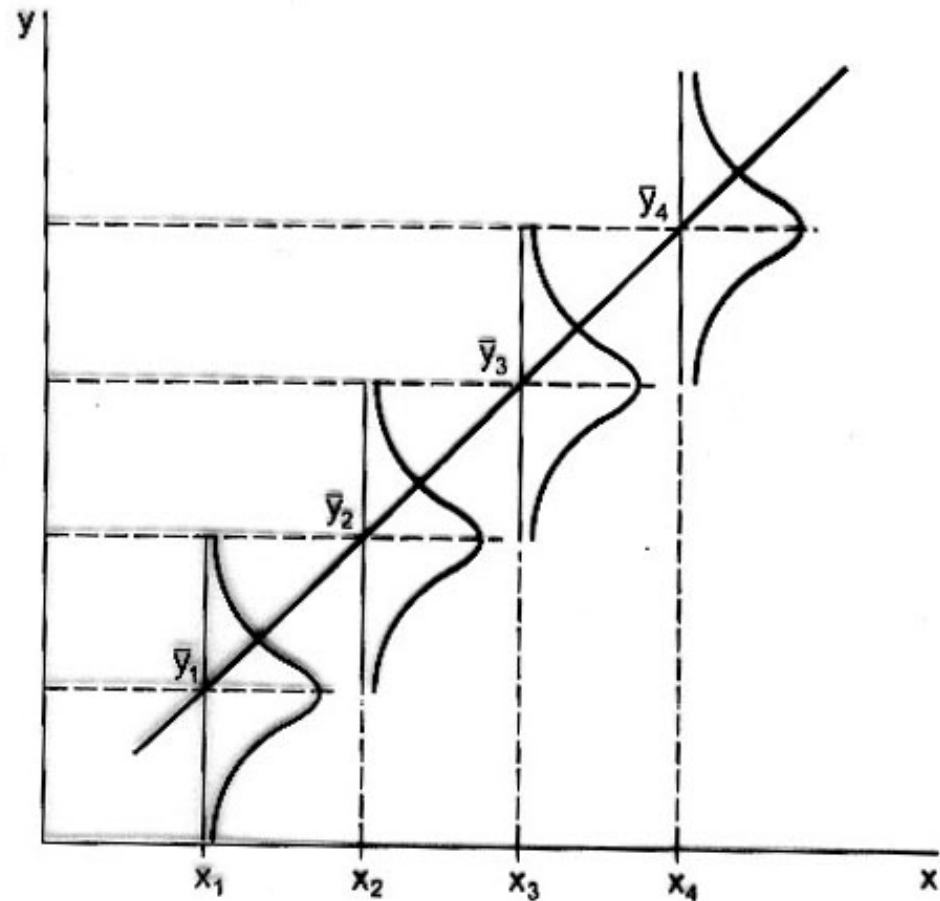
Každé hodnotě znaku nezávisle proměnné náhodné veličiny  $x$  odpovídá vždy pouze jediná určitá hodnota závisle proměnné veličiny  $y$

# Závislost statistická



- Každé hodnotě znaku nezávisle proměnné náhodné veličiny  $x$  odpovídá více hodnot závisle proměnné veličiny  $y$ ,
- Hodnoty  $y$  mají své rozdělení
- Při změně znaku nezávisle proměnné  $x$  mění podmíněná rozdělení relativních četností závisle proměnné  $y$

# Závislost korelační



Se změnou hodnoty znaku nezávisle proměnné  $x$  se mění podmíněná rozdělení relativních četností hodnoty znaku závisle proměnné  $y$  tak, že změna  $x$  podmiňuje změnu průměru  $\bar{y}$  souborů hodnot  $y$ , odpovídajících daným hodnotám  $x$ .



# Určení těsnosti korelační závislosti

- Úkolem korelačního počtu je vyjádřit tendenci změn hodnoty znaku závisle proměnné při změně hodnoty znaku nezávisle proměnné **matematickou funkcí**
- Tato funkce představuje tzv. **regresní čáru** a vyjadřuje, jaká hodnota znaku závisle proměnné odpovídá s největší pravděpodobností určité hodnotě znaku nezávisle proměnné.
- Odhad regresní závislosti je tím přesnější, čím větší je **těsnost korelační závislosti**.
- Určení těsnosti korelační závislosti je prvním krokem analýzy.

# Charakteristiky korelační závislosti

Máme dva výběrové soubory náhodných veličin  $X$ ,  $Y$ . Proměnlivost hodnot znaku obou výběrů můžeme vyjádřit odchylkami  $d_{xi}$  a  $d_{yi}$  prvků od jejich průměrů:

$$d_{xi} = x_i - \bar{x} \quad d_{yi} = y_i - \bar{y}$$

Vzájemnou proměnlivost obou výběrových souborů charakterizuje součin odchylek :

$$(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Suma součinů odchylek vydělaná rozsahem výběrů  $n$  určuje tzv. **kovarianci** výběrových souborů  $s_{xy}$  – tedy první společnou charakteristiku proměnlivosti obou souborů:

$$s_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n - 1}$$

# Charakteristiky korelační závislosti

- Kovariance je obdobou rozptylu
- Omezenost - je mírou **absolutní** – nelze jí použít k porovnání těsnosti vztahu dvou či více dvojic výběrových souborů.

**Relativní míra** – kovariance dělená součinem směrodatných odchylek  $s_x$  a  $s_y$  obou výběrů - **korelační koeficient**  $r_{xy}$ :

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{1}{n-1} \sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2 \cdot \frac{1}{n-1} \sum (y_i - \bar{y})^2}}$$

# Charakteristiky korelační závislosti

Úpravou výše uvedeného vztahu lze **korelační koeficient**  $r_{xy}$  vypočítat také podle následujícího vzorce:

$$r_{xy} = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2] \cdot [n \sum y^2 - (\sum y)^2]}}$$

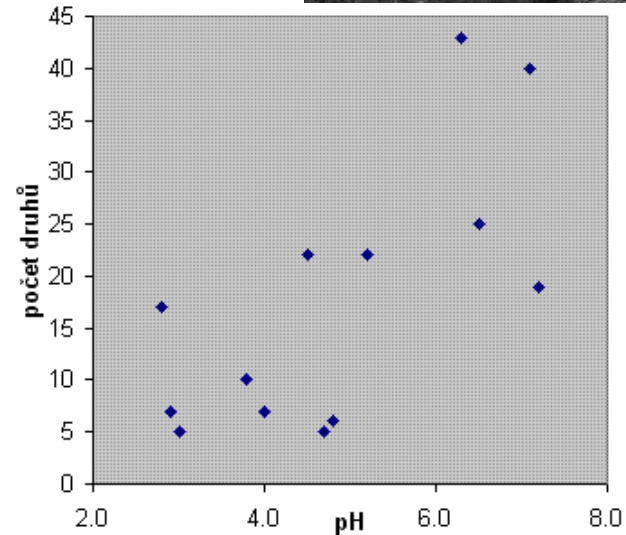
*(vzorec je uveden pouze pro názornost výpočtu v následujícím příkladě)*

# Příklad

Jaká je závislost mezi pH půdy na výsypkách a počtem rostlinných druhů?



pH počet druhů				
x	y	x <sup>2</sup>	y <sup>2</sup>	xy
2.8	17	7.8	289	47.6
2.9	7	8.4	49	20.3
3.8	10	14.4	100	38.0
4.5	22	20.3	484	99.0
7.1	40	50.4	1600	284.0
6.5	25	42.3	625	162.5
3.0	5	9.0	25	15.0
4.7	5	22.1	25	23.5
5.2	22	27.0	484	114.4
4.0	7	16.0	49	28.0
4.8	6	23.0	36	28.8
6.3	43	39.7	1849	270.9
7.2	19	51.8	361	136.8



$$r_{xy} = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2] \cdot [n \sum y^2 - (\sum y)^2]}}$$

$$r_{xy} = \frac{13 \cdot 1268,8 - 62,8 \cdot 228}{\sqrt{[13 \cdot 332,3 - 3943,84] \cdot [13 \cdot 5976 - 51984]}}$$

$$r_{xy} = 0,700$$

$$\sum x = 62,8 \quad \sum y = 228 \quad \sum xy = 1268,8$$

$$\sum x^2 = 332,3 \quad \sum y^2 = 59676$$

$$(\sum x)^2 = 3943,84 \quad (\sum y)^2 = 51984$$

# Příklad - pokračování



**Interpretace:** ze statistických tabulek zjistíme:

Hodnotě  $r_{xy} = 0,700$  přísluší pro  $\nu = n - 2 = 11$

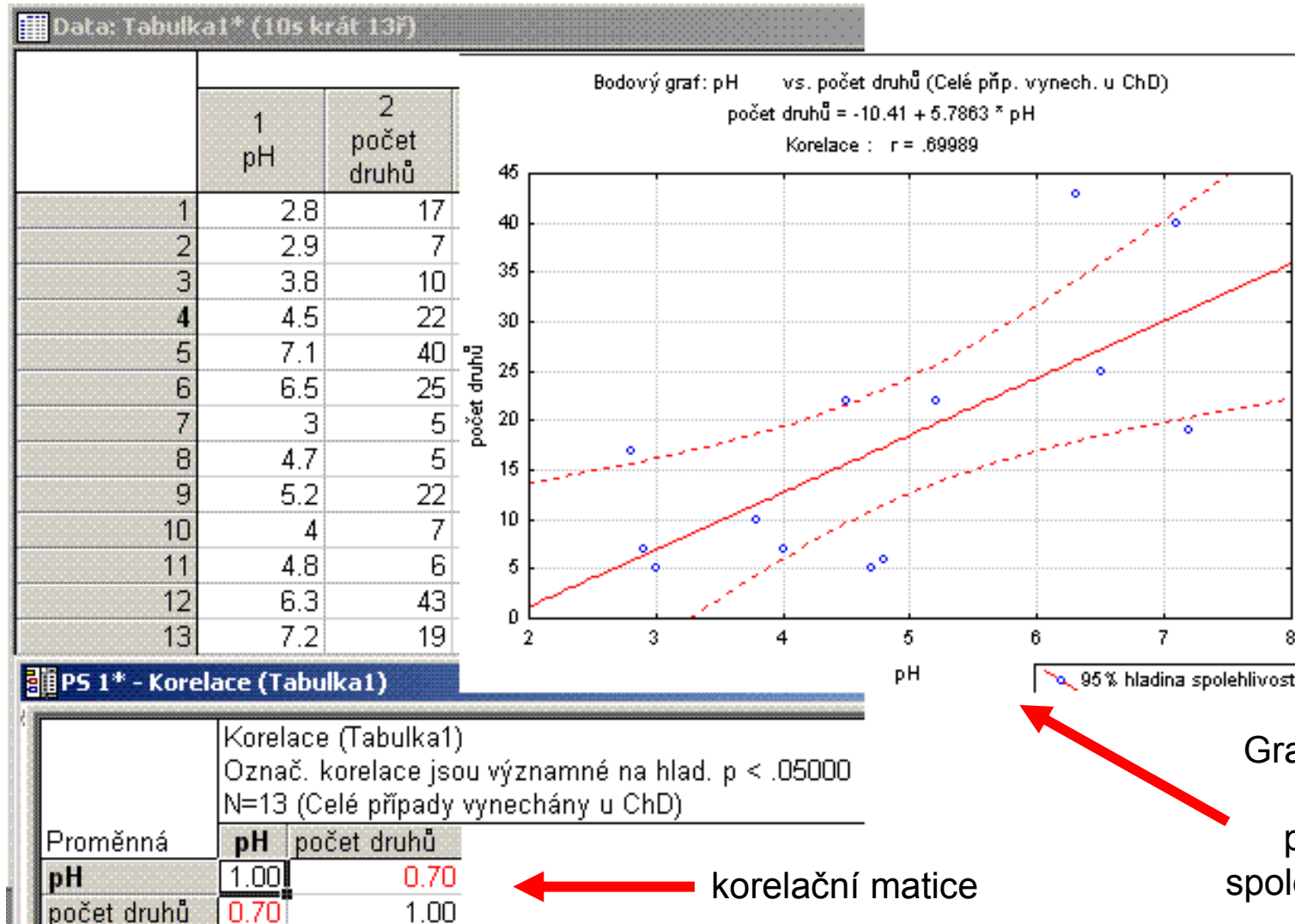
na hladině významnosti  $\alpha = 0,05$  kritická hodnota  $r_{krit} = 0,553$

**Závěr:** prokázali jsme statisticky významný vztah mezi pH a množstvím rostlinných druhů rostoucích na výsypkách.

# Příklad

## Řešení v programu Statistica:

*Statistika – Základní statistiky/tabulky – Korelační matice*



# Příklad

## Korelační matice – $r_{xy}$ mezi dvojicemi více proměnných

Statistika – Základní statistiky/tabulky – Korelační matice

The screenshot displays the SPSS interface with a data table, a dialog box for calculating correlation matrices, and a resulting correlation matrix table.

**Data: matice\_mes\_orig\* (15s krát 108r)**

	1	2	3	4	5	6	7
	mě sí c	sezona	√Prom	sumaPAHs	sumaPCB	sumaHCH	sumaDDT
1	Led-1997	z	I	180.228	0.2238	0.05875	0.0204
2	Úno-1997	z	II	20	0.065	0.0568	0.016
3	Bře-1997	j	III	25.1675	0.1405	0.06	0.06325
4	Dub-1997	j	IV	16.445	0.2615	0.0506	0.041
5	Kvě-1997	j	V	10.872	0.176	0.06025	0.08
6	Čen-1997	I	VI	4.9475	0.23175	0.03525	0.0465
7	Čec-1997	I	VII	4.646	0.294	0.0404	0.0626
8	Srp-1997	I	VIII	5.095	0.21325	0.03625	0.07975
9	Zář-1997	p	IX	5.5425	0.1465	0.05725	0.08425
10	Říj-1997	p	X	32.3	0.0902	0.05825	0.044
11	Lis-1997	p	XI	56.06	0.13075	0.0426	0.041

**Korelace a parciální korelace: matice\_mes\_orig**

1 seznam proměn. 2 seznamy (obd. matice) Souhrn

První seznam: sumaPAHs-SSV  
Druhý seznam: žádné

Základ | Detaily | Možnosti

Výpočet: Korelační matice!

Matice bod. grafů zvolených proměnných

Vážené momenty  
SV =  W-1  N-1  
 ChD vynechána  
 Celé případy  
 Párově

**PS 1\* - Korelace (matice\_mes\_orig)**

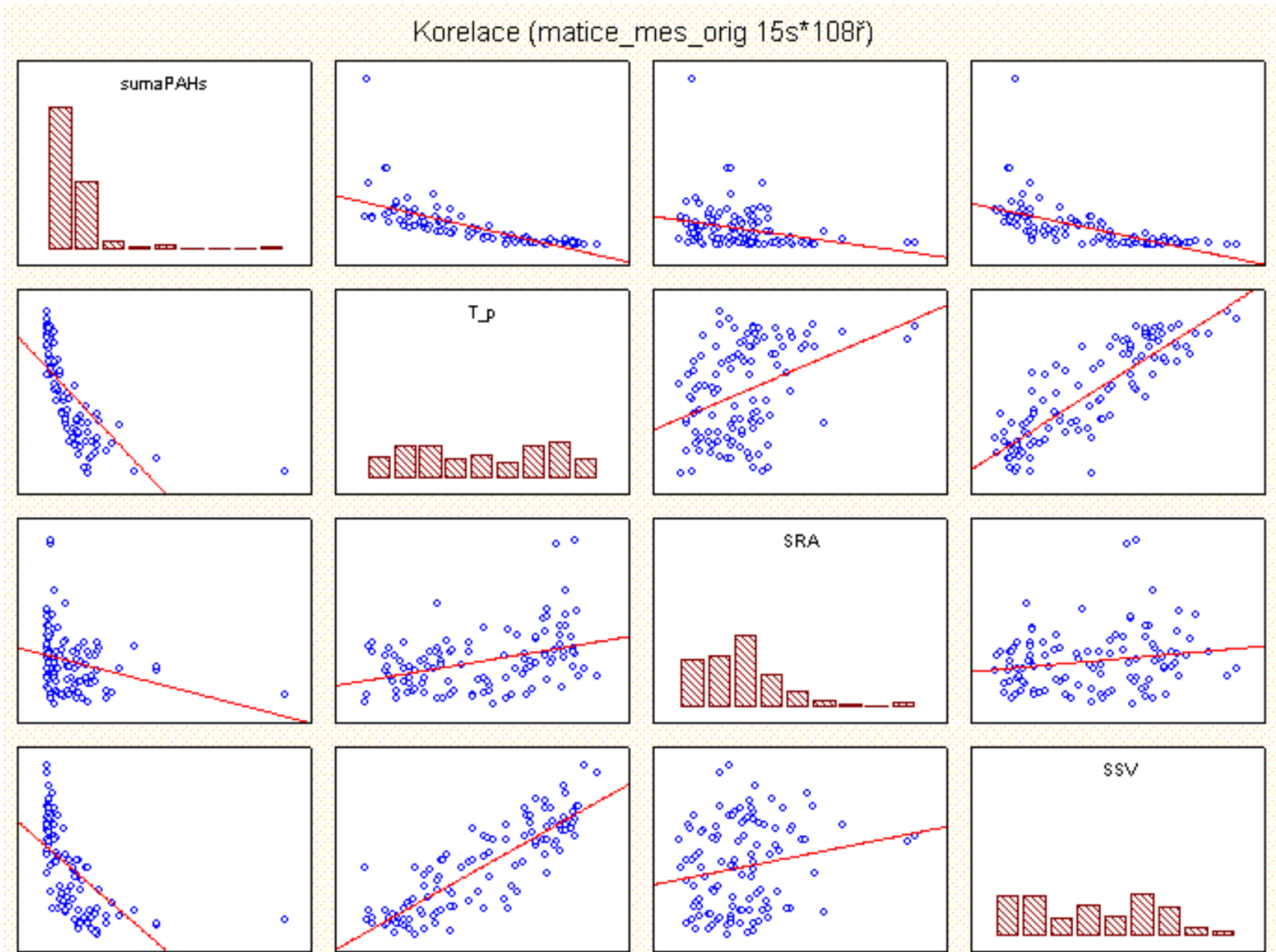
Korelace (matice\_mes\_orig)  
Označ. korelace jsou významné na hlad. p < .05000  
N=42 (Celé případy vynechány u ChD)

Proměnná	sumaPAHs	sumaPCB	sumaHCH	sumaDDT	T_p	SRA	T_max	T_min	E_p	F_p	SCE_p	SSV
sumaPAHs	1.00	0.01	0.21	-0.34	-0.65	-0.33	-0.65	-0.65	-0.58	0.05	0.65	-0.55
sumaPCB	0.01	1.00	-0.27	0.40	0.27	0.22	0.27	0.25	0.24	-0.03	-0.07	0.37
sumaHCH	0.21	-0.27	1.00	-0.20	-0.38	-0.37	-0.37	-0.37	-0.40	0.21	0.17	-0.30
sumaDDT	-0.34	0.40	-0.20	1.00	0.49	0.06	0.51	0.48	0.42	-0.26	-0.42	0.50
T_p	-0.65	0.27	-0.38	0.49	1.00	0.36	1.00	0.99	0.97	-0.66	-0.56	0.88
SRA	-0.33	0.22	-0.37	0.06	0.36	1.00	0.34	0.40	0.46	0.04	-0.09	0.17
T_max	-0.65	0.27	-0.37	0.51	1.00	0.34	1.00	0.99	0.95	-0.66	-0.56	0.90
T_min	-0.65	0.25	-0.37	0.48	0.99	0.40	0.99	1.00	0.98	-0.65	-0.59	0.83
E_p	-0.58	0.24	-0.40	0.42	0.97	0.46	0.95	0.98	1.00	-0.67	-0.48	0.77
F_p	0.05	-0.03	0.21	-0.26	-0.66	0.04	-0.66	-0.65	-0.67	1.00	0.28	-0.55
SCE_p	0.65	-0.07	0.17	-0.42	-0.56	-0.09	-0.56	-0.59	-0.48	0.28	1.00	-0.44
SSV	-0.55	0.37	-0.30	0.50	0.88	0.17	0.90	0.83	0.77	-0.55	-0.44	1.00



# Příklad

*Statistika – Základní statistiky/tabulky – Korelační matice – Matice bodových grafů*



# Koeficient determinace

- Koeficient korelace se často ve výpočtech doplňuje hodnotou koeficientu determinace ( $r^2_{xy}$ ).
- Jeho hodnota kolísá v intervalu 0 až 1
- Vynásoben 100 udává v procentech tu část rozptylu závisle proměnné  $y$ , která je vysvětlena (podmíněna) změnami hodnot nezávisle proměnné  $x$ .

---

V našem případě:

$$r_{xy} = 0,700 \quad \longrightarrow \quad r^2_{xy} = 0,49 = 49\%$$

**Interpretace:** Změna počtu druhů rostlin na výsypkách je z 49 % podmíněna změnami pH půdy na kterých tyto rostliny rostou.

# Podmínky použitelnosti $r_{xy}$

Výpočet  $r_{xy}$  se opírá o rozptyl a směrodatnou odchylku

Jeho použití tedy předpokládá splnění tří následujících podmínek:

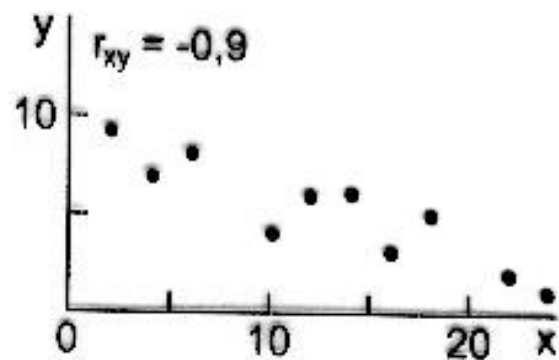
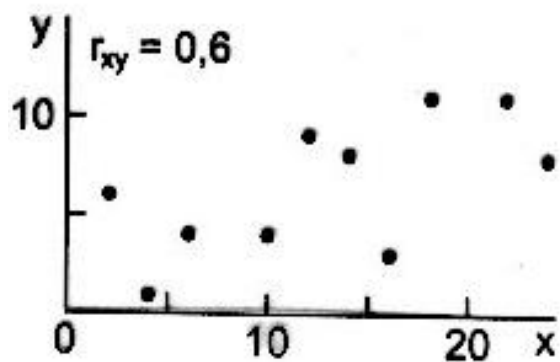
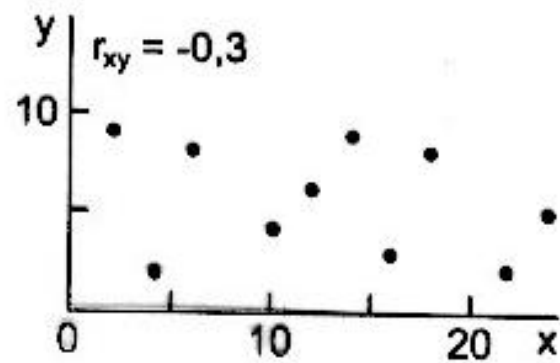
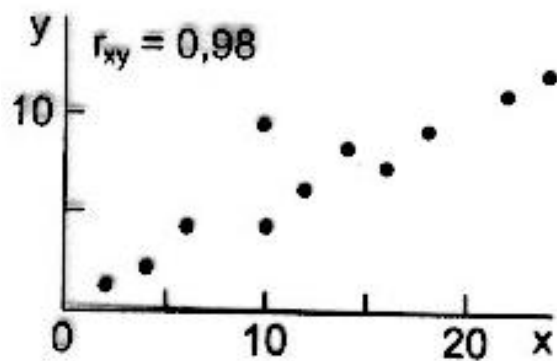
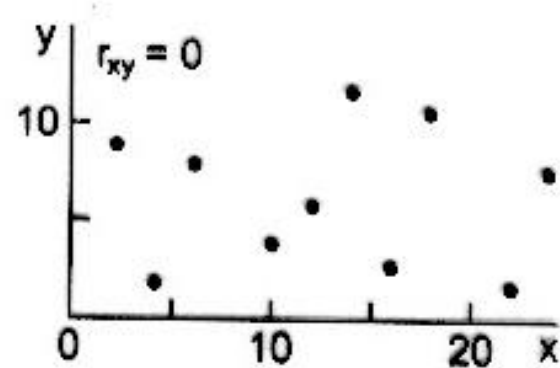
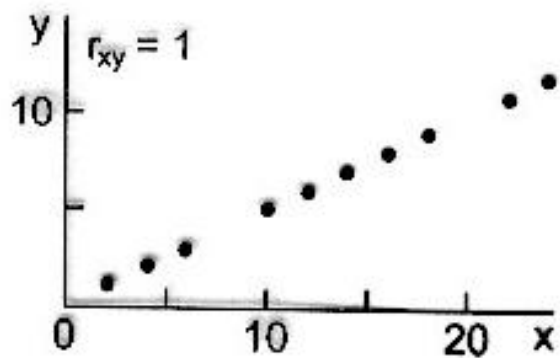
- normální rozdělení použitých výběrů
- dvojrozměrnost normálního rozdělení (každé hodnotě znaku veličiny  $x$  odpovídá soubor hodnot znaku  $y$ , který má normální rozdělení a naopak)
- linearita vztahu hodnot  $x$  a  $y$  (regresní čára je přímka)

Hodnota  $r_{xy}$  nás informuje o druhu a těsnosti závislosti

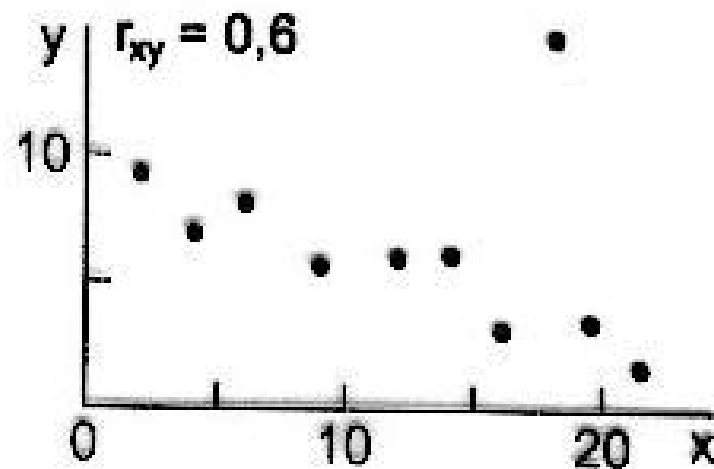
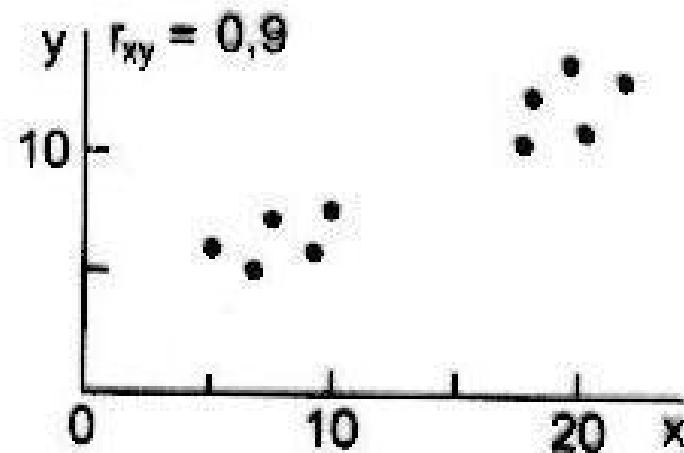
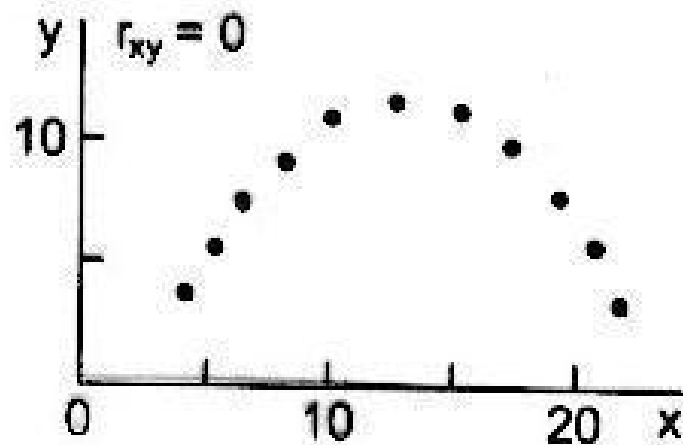
Dokonalá korelační závislost přímá  $r_{xy} = 1$

Dokonalá korelační závislost nepřímá  $r_{xy} = -1$

# Graf korelačního pole pro různá $r_{xy}$



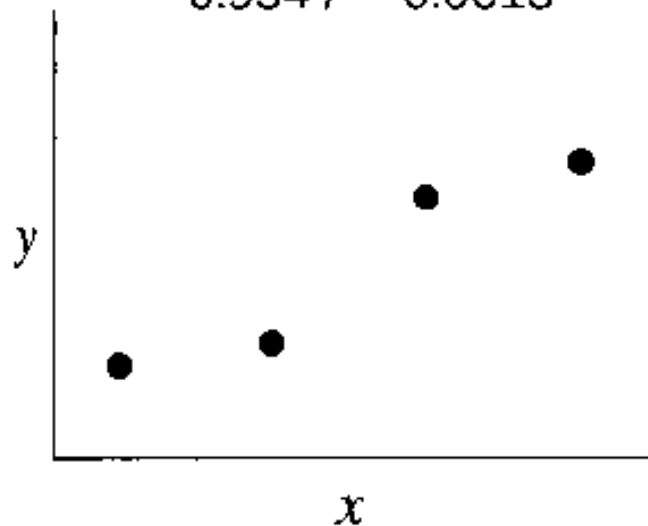
# Graf korelačního pole pro různá $r_{xy}$ ???



# Hodnocení významnosti koeficientu korelace

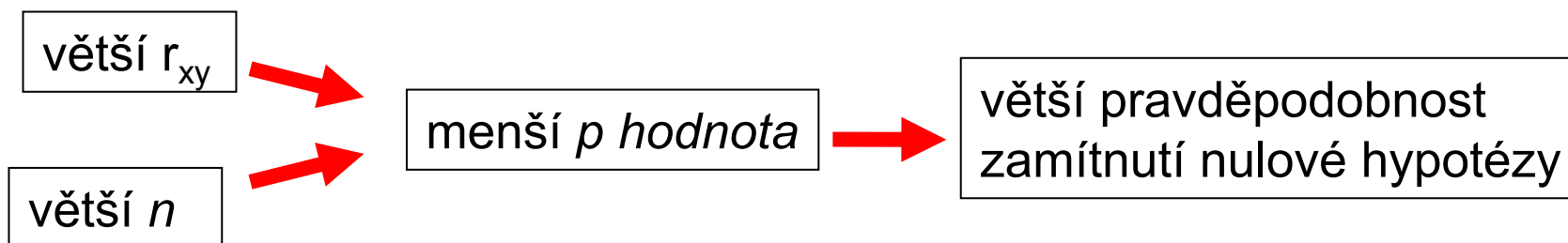
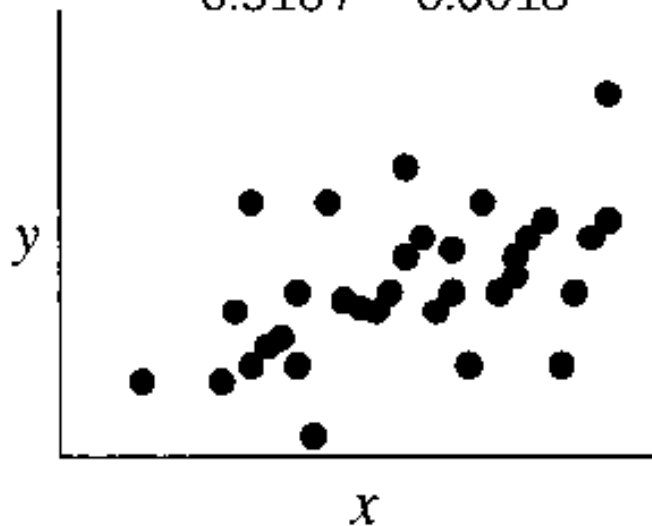
(a) Small sample size ( $n=4$ )

Correlation coefficient  
= 0.954  $P = 0.0613$



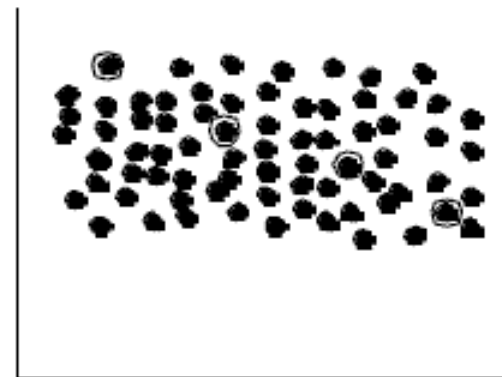
(b) Large sample size ( $n=33$ )

Correlation coefficient  
= 0.516  $P = 0.0018$



# Hodnocení významnosti koeficientu korelace

- Významnost  $r_{xy}$  závisí na povaze řešeného problému
- Jeho hodnota je mírou relativní a posouzení těsnosti je do značné míry subjektivní.



Významnost  $r_{xy}$  lze též zjistit objektivně – testováním:

Ze dvou základních jednorozměrných souborů lze provést sérii dvojic výběrů, které mají koeficienty korelace  $r_{xy}$ .

Soubor těchto výběrových koeficientů korelace má při velkých výběrech a při hodnotě korelačního koeficientu základního souboru ( $\rho$ ) blízké nule tzv. normální rozdělení.

Jeho průměr je  $\bar{r}_{xy} = \rho$  a směrodatná odchylka  $s_r$  se vypočte podle vztahu:

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

# Hodnocení významnosti koeficientu korelace

Při testování  $r_{xy}$  vycházíme z nulové hypotézy, která je  $\rho = 0$  (tedy mezi dvěma základními soubory nepředpokládáme žádný korelační vztah).

Testovací kritérium se vypočte podle vztahu:

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n - 2}$$

Přísluší mu  $t$ -rozdělení s  $\nu = n - 2$  stupni volnosti.

S určitou pravděpodobností - tedy na určité hladině významnosti předpokládáme, že hodnota  $t$  nepřekročí kritickou hodnotu  $t_p$  (při správnosti nulové hypotézy).

V opačném případě zamítáme nulovou hypotézu – mezi výběry náhodných veličin vztah existuje.



# Koeficient pořadové korelace (Spearmanův) ( $r_s$ )

Používá se k určení závislosti **kvalitativních znaků**.

Každé hodnotě  $x_i$  a  $y_i$  přiřadíme pořadové číslo  $px_i$  a  $py_i$  podle velikosti hodnot  $x_i$  a  $y_i$ .

Určíme rozdíly  $D_i$  dvojic pořadových čísel odpovídajících si hodnot.

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

# Koeficient pořadové korelace - příklad



**Příklad:** Kvantifikujte vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů (na m<sup>2</sup>).

Zjištěná data		Pořadová čísla		Diference	
Počet roků	Počet druhů	Počet roků	Počet druhů	D	D <sup>2</sup>
1	2	1	1	0	0
2	3	2	2	0	0
3	5	3	4	-1	1
4	4	4	3	-1	1
8	7	5	6,5	-1,5	2,25
10	6	6	5	1	1
> 10	7	7	6,5	0,5	0,25

$$\sum D_i^2 = 5,5$$

$$r_s = 1 - \frac{6 \sum D_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \times 5,5}{7 \times (49 - 1)} = 0,902$$

V tabulkách vyhledáme pro  $n=7$  a  $\alpha=0,05$  kritickou hodnotu:

$$r_{krit} = 0,786$$

**Závěr:** Existuje statisticky významný vztah mezi dobou, po kterou jsou pole ponechána ladem a počtem rostlinných druhů, které se na nich vyskytují.

# Koeficient pořadové korelace

Řešení v programu Statistica:



*Statistika – Neparametrická statistika – Korelace (Spearman, Kendallovo Tau, Gama)*

The screenshot shows the 'Neparametrické korelace: Tabulka6' dialog box in Statistica. The 'Data: Tabulka6' window shows the following data:

	1 Počet roků	2 Počet druhů
	1	2
	2	3
	3	5
	4	4
	8	7
	10	6
	99	7

The dialog box shows the following settings:

- Proměnné: List 1: žádné
- Vytvořit: Čtvercová matice
- Zákl. výsledky: Spearmanovo R, Gama, Kendallovo tau
- úroveň p na zvýraznění: .05

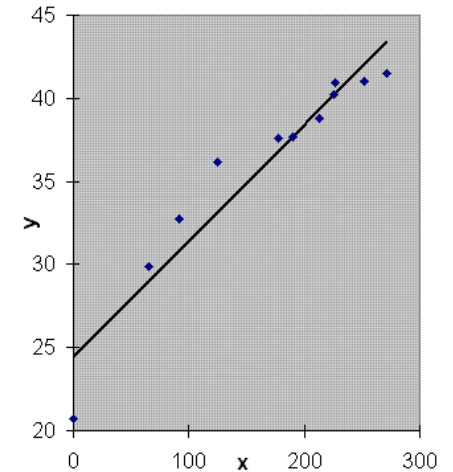
The output window shows the following results:

Spearmanovy korelace (Tabulka6)  
ChD vynechány párově  
Označ. korelace jsou významné na hl. p <.05000

Proměnná	Počet roků	Počet druhů
Počet roků	1.000000	0.900937
Počet druhů	0.900937	1.000000

# Nelineární závislost

V případě, kdy regresní čára není přímka, ale je vyjádřena složitější matematickou funkcí, se jako míry korelační závislosti používá tzv. korelační poměr ( $\eta_{yx}$ ).



Prvky výběru závisle proměnné  $y_i$  rozdělíme podle hodnot nezávisle proměnné  $x_i$  do skupin označených  $y_j$  a pro každou skupinu vypočteme průměr  $\bar{y}_j$ . Korelační poměr se vypočte podle vztahu:

$$\eta_{yx} = \sqrt{\frac{\sum (\bar{y}_j - \bar{y}) \cdot n_j}{\sum (y_i - \bar{y})^2}} = \sqrt{\frac{\sum (\bar{y}_j n_j - n\bar{y})^2}{\sum y_i^2 - n\bar{y}^2}}$$

V uvedeném vzorci je  $n_j$  četnost v  $y_j$ . Při výpočtu **záleží** na tom, kterou proměnou zvolíme za závislou a kterou za nezávislou.

Porovnání hodnot korelačního koeficientu a korelačního poměru lze použít jako kritéria linearit vztahu.

Pokud se hodnoty přibližně rovnají, jedná se o závislost lineární, pokud je  $r_{xy}$  výrazně větší, jde o závislost nelineární.

# Koeficient mnohonásobné korelace ( $r_{xyz}$ )

Vztah dvou proměnných je často ovlivněn dalšími proměnnými.

Používá se pro hodnocení korelační závislosti tří nebo více výběrů náhodných veličin.

Při jeho určení se vychází z jednotlivých korelačních koeficientů pro dva výběry ( $r_{xy}$ ,  $r_{xz}$ ,  $r_{yz}$ ) a jejich hodnoty se dosazují do vzorce pro  $r_{xyz}$ :

$$r_{xyz} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xy} \cdot r_{xz} \cdot r_{yz}}{1 - r_{xy}^2}}$$

Příklad – viz. vícerozměrná regrese

## Dílčí (parciální) korelace:

Řeší otázku vlivu jedné nebo více nezávisle proměnných na závisle proměnnou při **vyloučení vlivu** zbývajících nezávisle proměnných, u nichž předpokládáme konstantní hodnotu.

Jedná se o zvláštní případ mnohonásobné korelace, kdy další proměnné považujeme za „**rušivé**“ (např. věk, počet obyvatel sídla, ...).

Hodnota koeficientu dílčí korelace  $r_{xy \cdot z}$  se vypočte podle vztahu:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2) \cdot (1 - r_{yz}^2)}}$$

Tečkou v indexu se označuje nezávisle proměnná, jejíž hodnotu považujeme za konstantní.

# Parciální korelace

Příklad (viz. Brázdil a kol., 1995, str. 129, cvič. 8.3)

Způsob zadání proměnných (korelace mezi y a z při vyloučení vlivu x)

The screenshot shows the SPSS interface. On the left is a data table with columns '1 mesic', '2 x', '3 y', and '4 z'. The rows represent months from XI to X. On the right is the 'Korelace a parciální korelace: monohonas\_reg' dialog box. The 'První seznam' (First list) contains 'y-z' and the 'Druhý seznam' (Second list) contains 'x'. The 'Parciální korelace' button is circled in red. Other options include 'Výpočet: Korelační matice', 'Matice', '2D bod. grafy', '3D bod. grafy', 'Matice bod. grafů', 'Kateg. bod. grafy', 'Povrch. grafy', and '3D histogramy'. There are also checkboxes for 'Vážené momenty' and 'ChD vynechána'.

PS 2\* - Parciální korelace (monohonas\_reg)

Parciální korelace (monohonas\_reg)  
Označ. korelace jsou významné na hlad.  $p < ,05000$   
N=12 (Celé případy vynechány u ChD)

Proměnná	y	z			
y	1,00	0,99			
z	0,99	1,00			

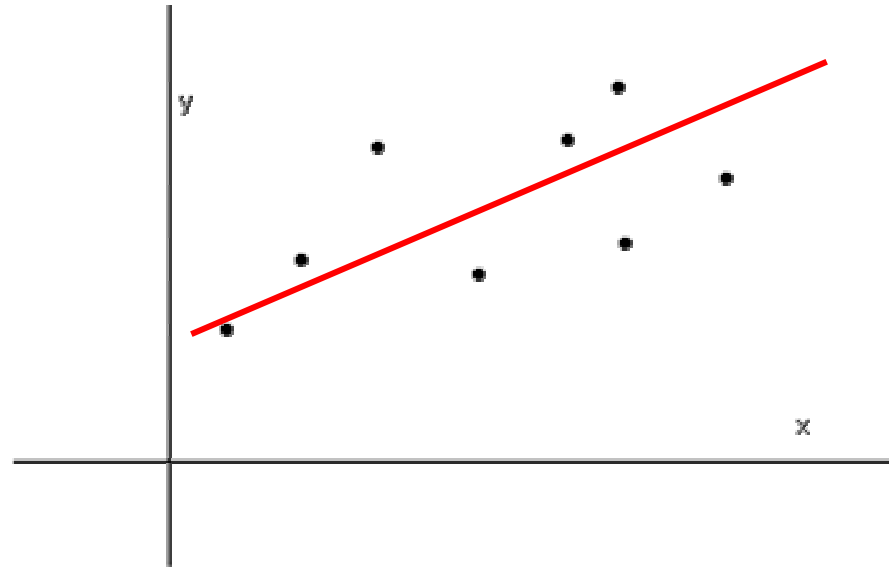
# Poznámky k aplikaci korelačního počtu:

Použití korelačního počtu je nevhodné např. v těchto případech:

- Korelace je způsobena formálními vztahy mezi veličinami (hodnoty  $x$  a  $y$  se doplňují do 100%)
- Korelace je způsobena nehomogenitou studovaného materiálu (obsahuje tzv. subpopulace – viz. obr. bodového grafu)
- Korelace je výsledkem působení třetí veličiny (korelace mezi počtem lékařů a počtem nemocných, ...)



# Regresní analýza



Úkolem regresní analýzy je sestavit **vztah (model)** závislosti mezi závisle a nezávisle proměnnou.

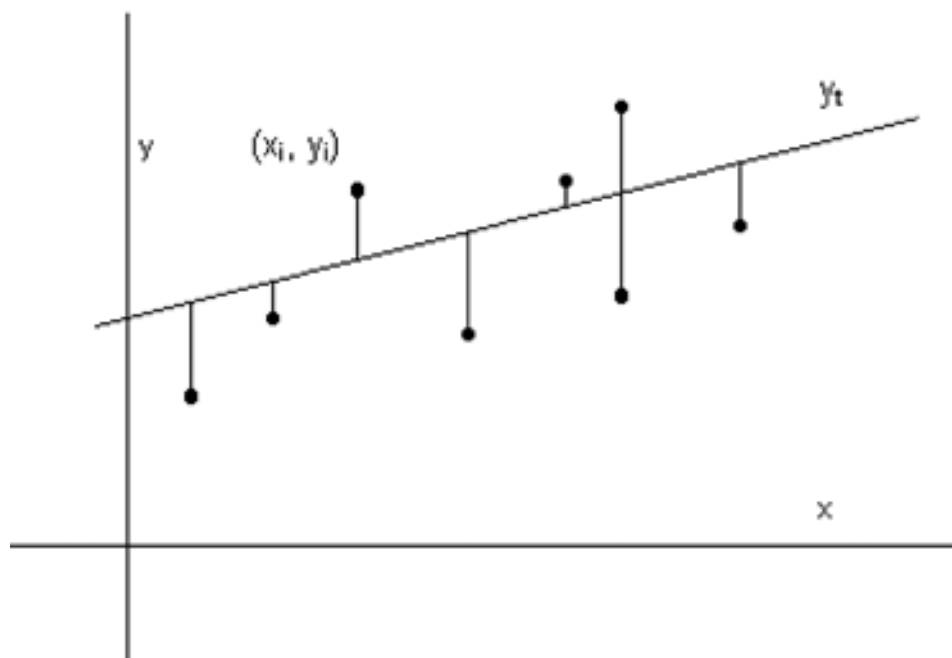
**Regresní analýza řeší :**

- odhady neznámých parametrů regresní funkce
- testování hypotéz o těchto parametrech
- ověřování předpokladů regresního modelu

# Určení lineární regresní závislosti

Nejjednodušším případem regresní závislosti je případ, kdy regresní funkce je přímkou. Rovnice regresní přímky má tvar:

$$y' = a + bx$$

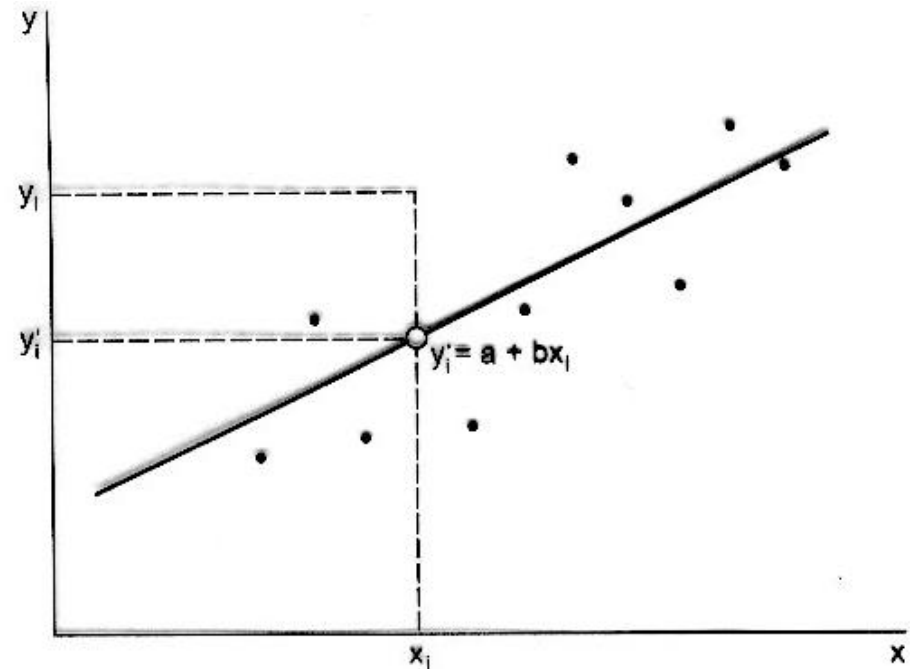


Symbol  $y'$  se používá pro označení **nejpravděpodobnější teoretické hodnoty**  $y$  odpovídající danému  $x$ , která leží na regresní přímce a která se odlišuje od konkrétních hodnot  $y_i$ , které se nacházejí mimo ni.

# MNČ

Průběh regresní přímky je určen tzv. **metodou nejmenších čtverců**, kdy musí být splněna podmínka takového průběhu přímky, při kterém je součet čtverců vzdálenosti všech bodů pole od přímky minimální, tedy platí:

$$\sum (y_i - y_i')^2 = \min$$



Výpočet vertikální vzdálenosti bodů korelačního pole od regresní přímky se provádí podle uvedeného obrázku. Z něho je zřejmé, že pro vzdálenost konkrétní hodnoty závisle proměnné  $y_i$  od bodu regresní přímky  $y_i'$  musí platit:

$$y_i - y_i' = y_i - (a + bx_i) = y_i - a - bx_i$$

Součet čtverců svislých vzdáleností  $y_i$  od regresní přímky je potom:

$$\sum (y_i - y_i')^2 = \sum (y_i - a - bx_i)^2 = A$$

# MNČ

Pro MNČ musí platit

$$A = \sum (y_i - a - bx_i)^2 = \min$$

Následnými úpravami lze obdržet vztahy pro výpočet koeficientů regresní přímky  $a, b$

$$b = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad a = \bar{y} - b \bar{x}$$

**Koeficient  $b$**  (angl. slope) se označuje jako koeficient regrese a je směrnicí regresní přímky (tangentou úhlu, který přímka  $a$  svírá s osou  $x$ ). Je-li  $b > 0$ , mluvíme o regresi pozitivní, je-li  $b < 0$  o regresi negativní.

# Výpočet koeficientů regresní přímky

Vzorec pro výpočet koeficientu  $b$  lze zjednodušit pomocí vztahů pro kovarianci  $s_{xy}$  a směrodatnou odchylku  $s_x$ , tedy:

$$b = \frac{s_{xy}}{s_x^2}$$

Hodnota **koeficientu  $a$**  (angl. intercept) představuje y-ovou souřadnici průsečíku regresní přímky s osou y (tedy při  $x=0$ ).

Dosazením výrazu pro koeficient  $a = \bar{y} - b\bar{x}$  do rovnice přímky  $y' = a + bx$  dostaneme:

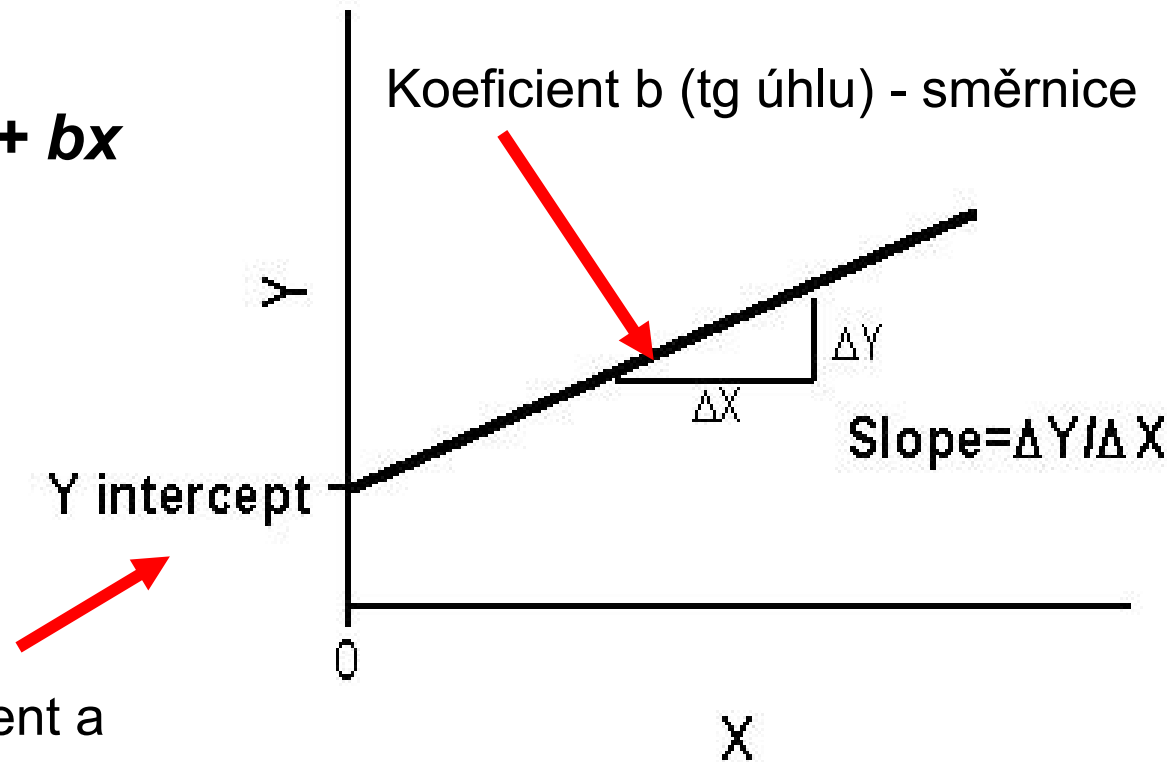
$$y' = bx + \bar{y} - b\bar{x}$$

$$y' - \bar{y} = b(x - \bar{x})$$

Tohoto vztahu lze využít pro konstrukci regresní přímky – pro dvě zvolená  $x_1$ ,  $x_2$  vypočteme  $y_1$  a  $y_2$  a souřadnice obou bodů vyneseme do korelačního diagramu. Regresní přímka vznikne proložením oběma body.

# Koeficienty lineární regrese závislosti

$$y' = a + bx$$



Koeficient a

Hranice (EXCEL)

Abs. člen (Statistica)

Koeficienty (parametry) jsou bodovými odhady !



# Intervaly a pásy spolehlivosti lineární regresní závislosti

- Konstrukci regresní přímky provádíme na základě výběrových souborů.
- Proto se její rovnice může u různých výběrů ze stejných základních souborů lišit.
- Z tohoto důvodu je potřebné doplnit průběh regresní přímky také tzv. **intervaly spolehlivosti**.
- Výpočtem intervalů spolehlivosti určíme pro vybraná  $x$  interval, v němž se mohou s určitou pravděpodobností vyskytovat hodnoty  $y$  s tím, že jejich nejreprezentativnější hodnota je  $y'$ .

# Intervaly a pásy spolehlivosti

Nejprve je zapotřebí zvolit hladinu spolehlivosti – tedy pravděpodobnost, s níž očekáváme výskyt hodnot  $y$  v určených mezích  $1-p$  ( $p=0,05$  či  $0,01$ ). Poloviční šířka intervalu spolehlivosti  $l$  je dána výrazem:

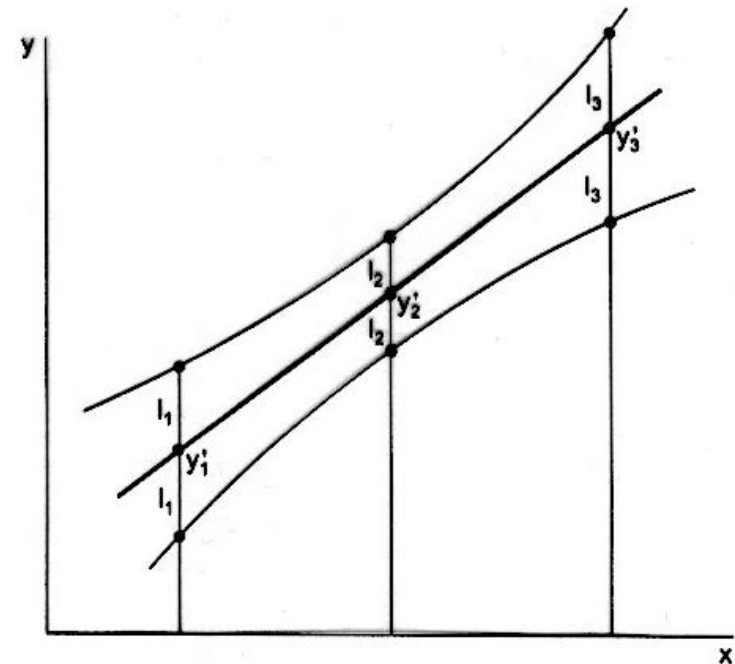
$$l = t_{1-p} \cdot \frac{h\sqrt{A}}{\sqrt{n-2}} \qquad h = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}}$$

Hodnota  $t_p$  je kritická hodnota rozdělení pro  $n-2$  stupňů volnosti a hladinu významnosti  $p$ . Meze intervalů spolehlivosti určíme pomocí hodnot  $y'$  z rovnice  $y' - \bar{y} = b(x - \bar{x})$

horní mez:  $y' + l$

dolní mez:  $y' - l$

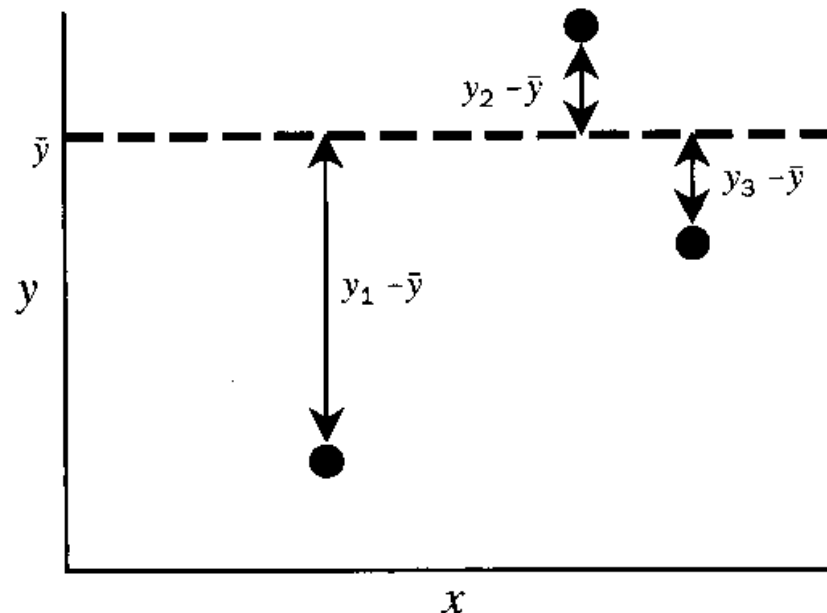
Pásy spolehlivosti vzniknou spojením krajních bodů intervalů spolehlivosti.





# Testování významnosti regresní závislosti

- K testování významnosti zjištěné regresní závislosti lze využít **t-testu**, kterým lze zjistit, zda se směrnice významně liší od nuly
- Nejčastěji se k testování používá **analýzy rozptylu (ANOVA)**.
- **Princip:** Zjistíme celkovou proměnlivost hodnot  $y$  a následně vypočteme, z jaké části je tato celková variabilita objasněna proměnlivostí v hodnotách  $x$ .



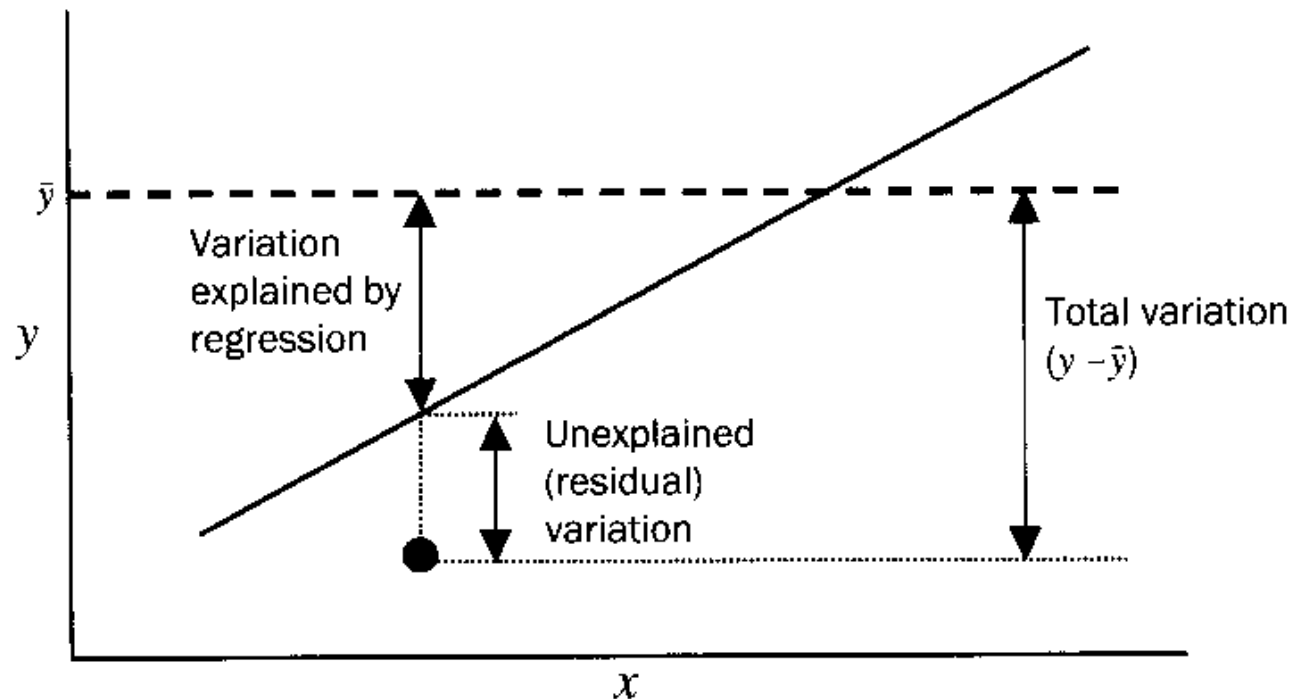
$SS_{total}$  - **celková variabilita**: celková suma čtverců: od každé hodnoty  $y$  odečteme průměr, výsledek povýšíme na druhou a sečteme pro všechna  $y$ .

# Testování významnosti regresní závislosti

Celkovou variabilitu  $SS_{total}$  lze rozdělit na dvě části:

$SS_{regrese}$  - variabilitu **vysvětlenou** regresní čarou

$SS_{reziduální}$  – zbytková variabilita **nevysvětlená** regresním modelem



$$SS_{reziduální} = SS_{total} - SS_{regrese}$$

# Testování významnosti regresní závislosti

Tabulka ANOVA

Variabilita	stupně volnosti (df)	Suma čtverců (SS)	Průměr sumy čtverců (MS)	F hodnota	p hodnota
Regresní	1	$SS_{regres}$	$\frac{SS_{regres}}{1}$	$\frac{MS_{regres}}{MS_{resid}}$	F hodnota pro df = 1 a n-2
Reziduální	n-2	$SS_{rezidual}$	$\frac{SS_{rezidual}}{n-2}$		
Celková	n-1	$SS_{total}$			

Koeficient determinace regresní závislosti:

$$r^2 = \frac{SS_{regres}}{SS_{total}}$$

# Příklad regresní analýzy v EXCELU

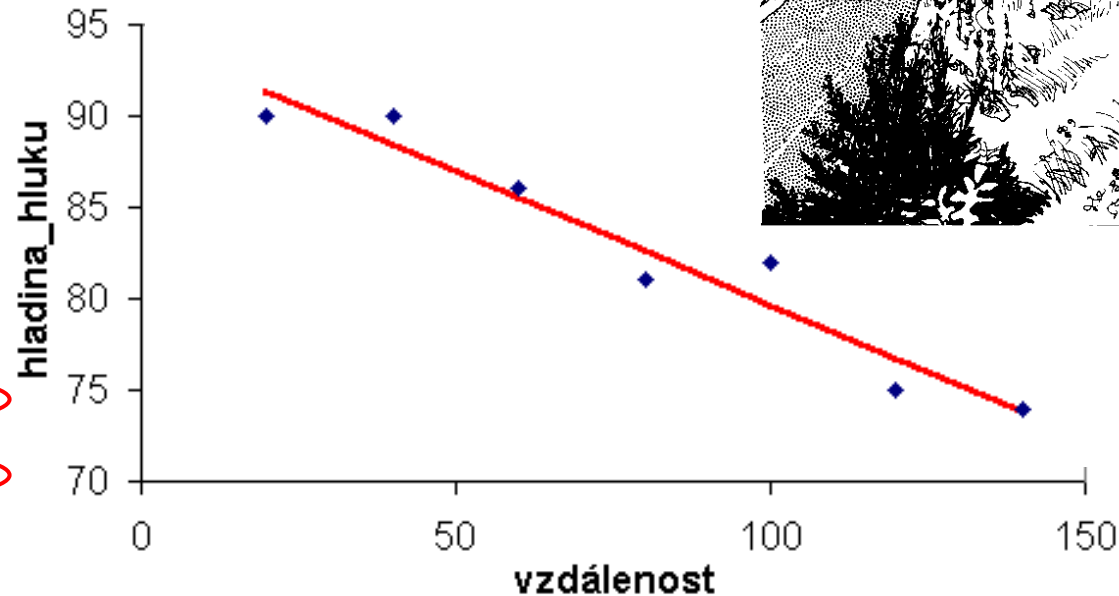
vzdálenost	hladina hluku
20	90
40	90
60	86
80	81
100	82
120	75
140	74

Zjistěte, jak souvisí hladina hluku se vzdáleností od komunikace.



VÝSLEDEK

Regresní statistika	
Násobné R	0,969074891
Hodnota spolehlivosti R	0,939106145
Nastavená hodnota spolehlivosti R	0,926927374
Chyba stř. hodnoty	1,764733893
Pozorování	7



ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	1	240,1428571	240,1429	77,11009	0,000317686
Rezidua	5	15,57142857	3,114286		
Celkem	6	255,7142857			

95% int. odhad hladiny hluku ve vzdálenosti 0 metrů  
 95% int. odhad poklesu hl. hluku na každý metr

	Koeficienty	Chyba stř. hodnoty	t stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	94,2857	1,4915	63,2165	0,0000	90,4518	98,1197	90,4518	98,1197
vzdálenost	-0,1464	0,0167	-8,7812	0,0003	-0,1893	-0,1036	-0,1893	-0,1036

$$y' = 94,2857 - 0,1464x$$

pokles hl. hluku na každý metr

Existuje signifikantní pokles hladiny hluku se vzdáleností od komunikace. Lineární regresní model vysvětluje 93,9 % variability hodnot hladiny hluku.

# Řešení v programu Statistica



1) Statistika – Vícerozměrná regrese

(zvolení závisle a nezávisle proměnné)

The screenshot shows the 'Výsledky - vícerozměrná regrese: Tabulka11' window. On the left, a data table is visible with two columns: '1 vzdálenost' and '2 hladina hluku'. The main window displays the following regression statistics:

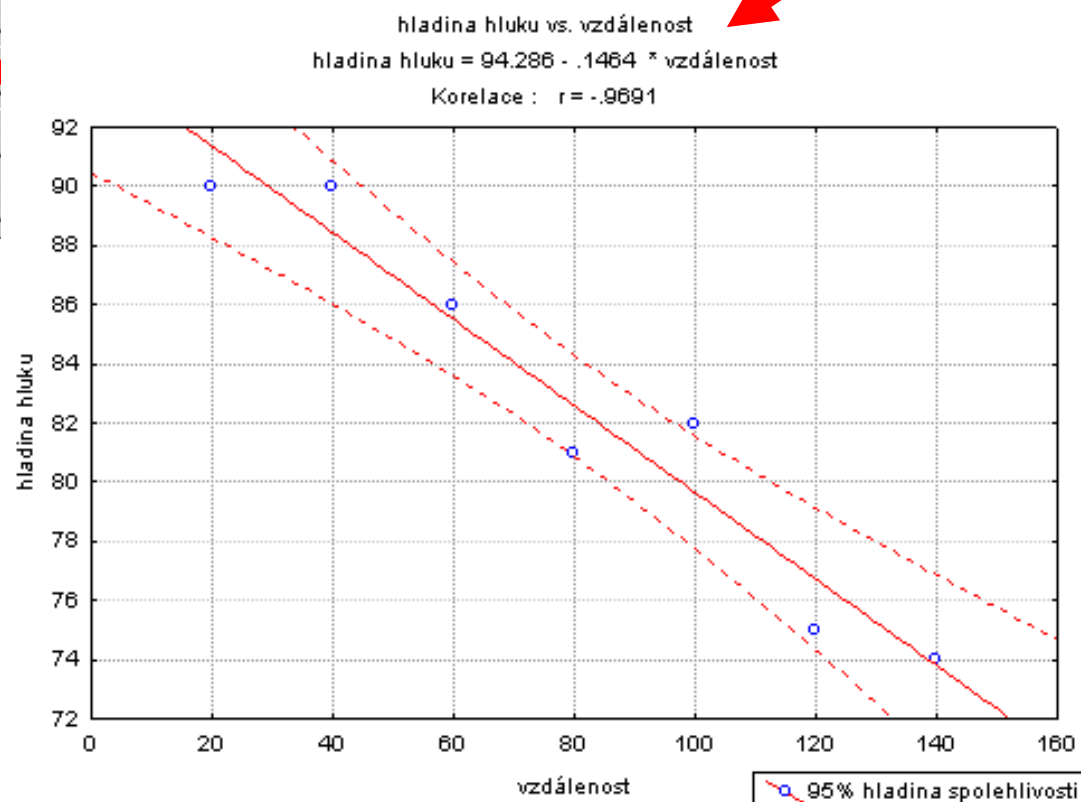
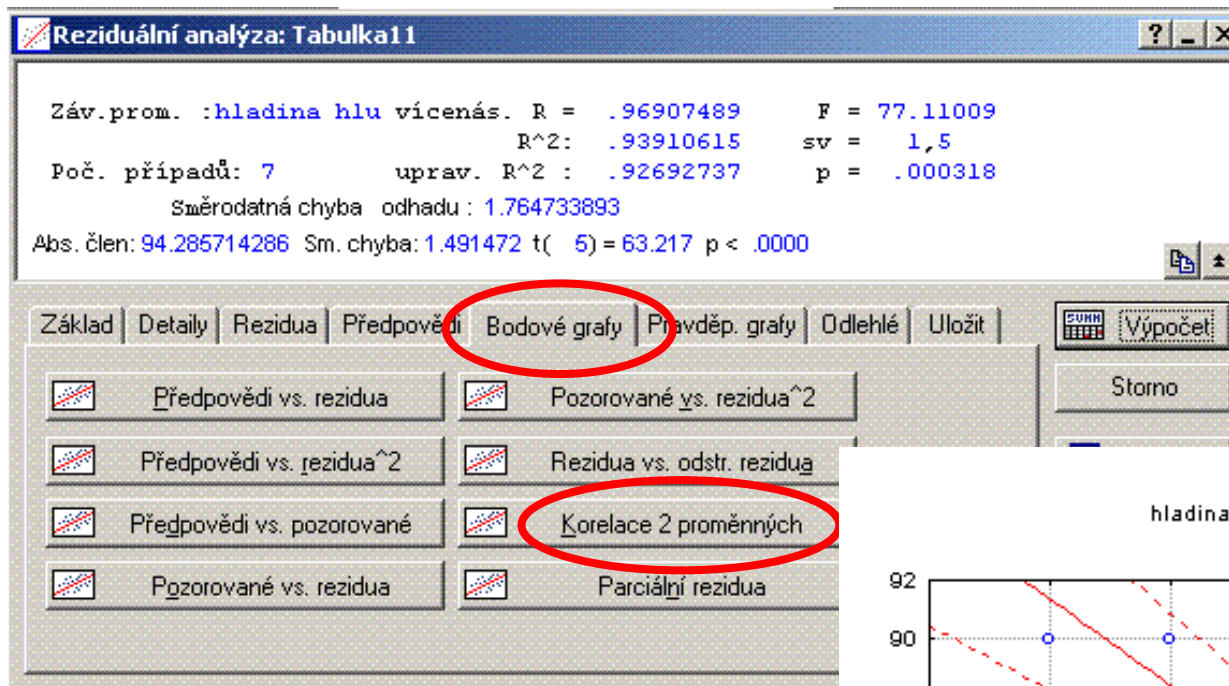
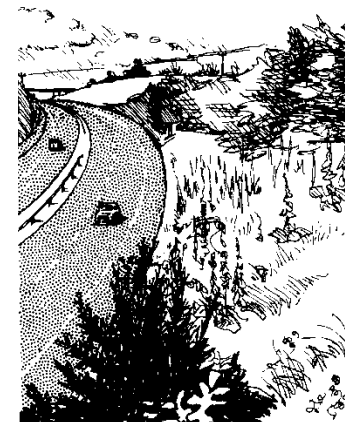
Statistika	Hodnota
Záv.prom. : hladina hluku	vícenás. R = .96907489
Poč. případů: 7	R <sup>2</sup> = .93910615
Směrodatná chyba odhadu: 1.764733893	sv = 1,5
Abs. člen: 94.285714286	uprav. R <sup>2</sup> = .92692737
Sm. chyba: 1.491472	p = .000318
t( 5) = 63.217	p = .0000

Below the statistics, the text 'vzdálenost beta=-.97' is displayed. At the bottom, the 'Alfa na zvýraznění efektů' is set to .05. The 'OK' button is circled in red.

1	2	
zdálenost	hladina hluku	
1	20	90
2	40	90
3	60	86
4	80	81
5	100	82
6	120	75
7	140	74

# Řešení v programu Statistica

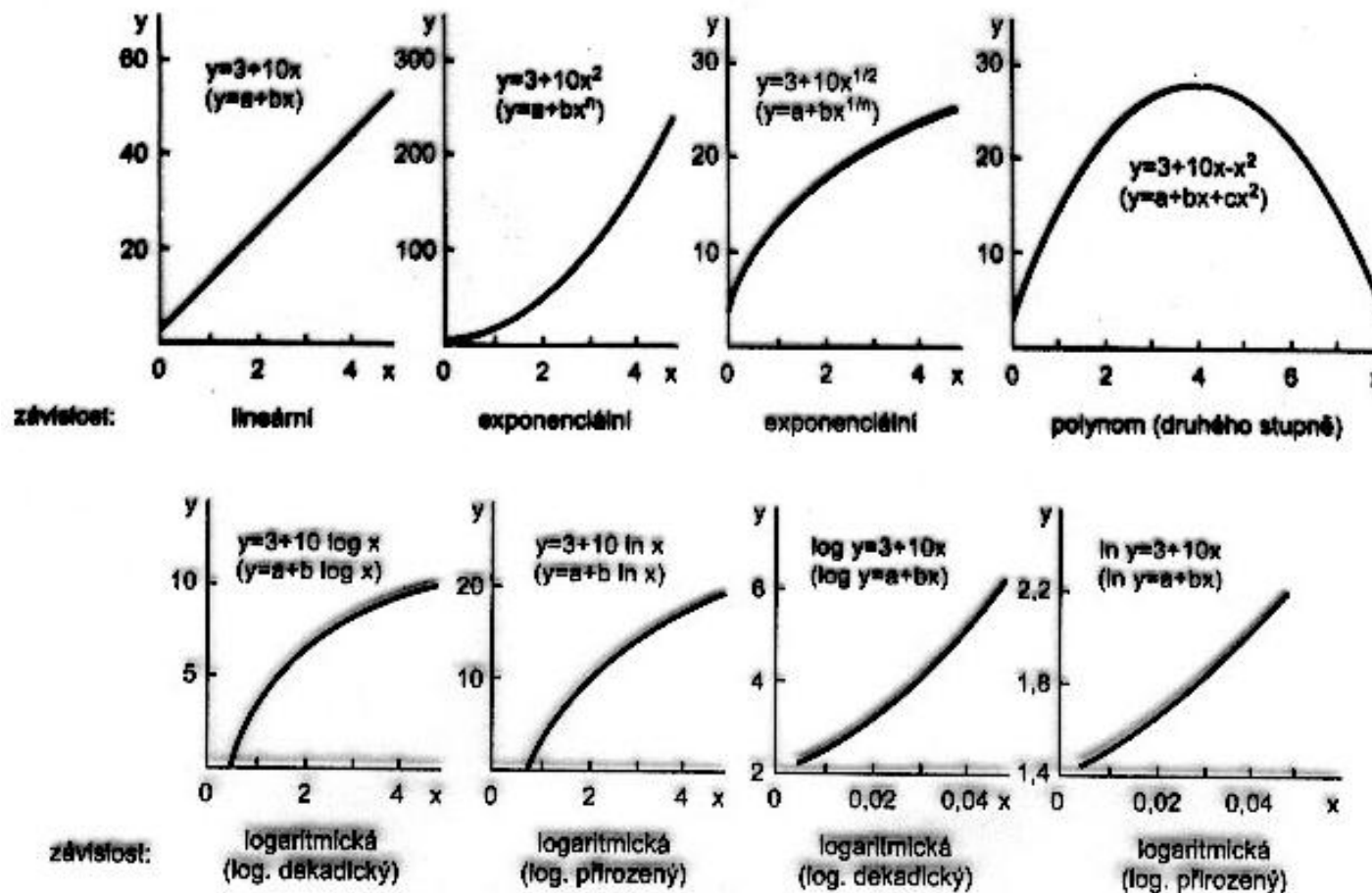
## 2) OK – Bodové grafy – Korelace 2 proměnných



# Další typy regresních funkcí

Regresní vztah dvou proměnných často nelze vhodně vyjádřit přímkou – jiné typy funkcí.

Může mít tvar např. logaritmických či exponenciálních funkcí a nebo je vztah vyjádřen rovnicí polynomu m-tého stupně.



# Další typy regresních funkcí

Volbu vhodné funkce, která by nejlépe vystihovala povahu studované závislosti provádíme na základě výpočtu **směrodatné chyby** aritmetického průměru  $s_{\bar{y}}$  (viz. – Odhady parametrů a intervaly spolehlivosti).

Určení hodnoty směrodatné chyby aritmetického průměru spočívá v určení sumy čtverců odchylek  $A$  konkrétních hodnot  $y_i$  závisle proměnné od teoretických hodnot  $y'_i$  tedy:

$$\sum (y_i - y'_i)^2 = \sum (y_i - a - bx_i)^2 = A$$

$$s_{\bar{y}} = \sqrt{\frac{A}{n-p}} = \sqrt{\frac{\sum (y_i - y'_i)^2}{n-p}}$$

kde  $p$  je počet parametrů použitého modelu.

Povaze studované závislosti vyhovuje nejlépe ta z uvažovaných funkcí, která má hodnotu směrodatné chyby minimální.

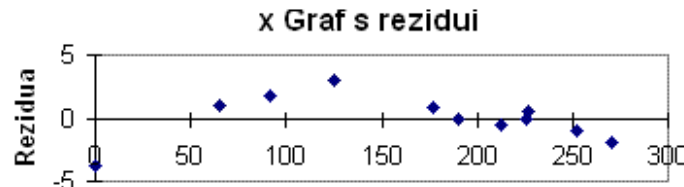
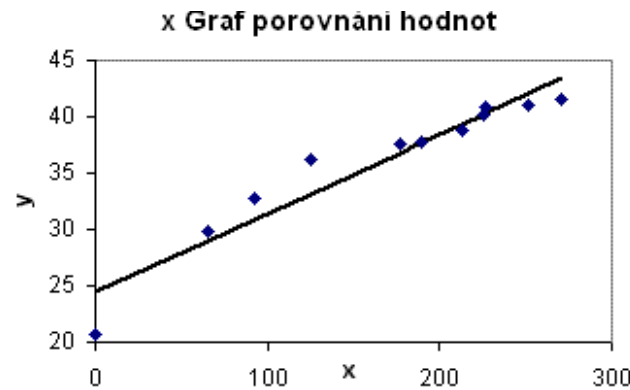
Konkrétní balíky statistických programů obsahují obvykle řadu nástrojů pro zvolení vhodné regresní závislosti.



# Regresní závislost není přímka

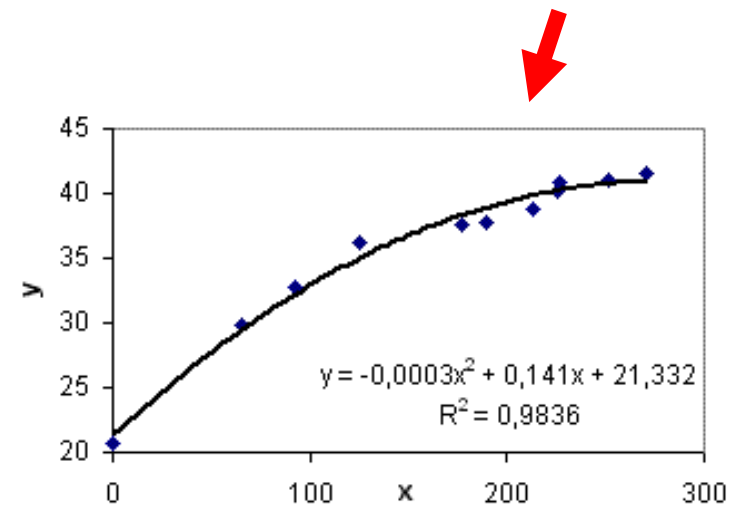
Příklad (viz. Brázdil a kol., 1995, str. 139, cvič. 8.5)

Stanice	x	y
Ivančice	0	20,7
Sen. Mlýn	65	29,9
Jakub	92	32,7
Senorady	125	36,2
Zastávka	177	37,6
Kratochvilka	190	37,7
Příbram	213	38,8
Bučín	226	40,2
Hlína	227	40,9
Ketkovice	252	41
Vys. Popovice	271	41,5



nevhodný model

vhodný model



## VÝSLEDEK

### Regresní statistika

Násobné R	0,956724
Hodnota spc	0,915321
Nastavená h	0,905912
Chyba stř. h	1,921192
Pozorování	11

koeficient determinace

směrodatná chyba odhadu

### ANOVA

	Rozdíl	SS	MS	F	Významnost F
Regrese	1	359,0703	359,0703	97,28322	4,01237E-06
Rezidua	9	33,21881	3,690979		
Celkem	10	392,2891			

	Koeficienty	chyba stř. hod.	t stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 95,0%	Horní 95,0%
Hranice	24,43416	1,31782	18,54135	1,77E-08	21,45303879	27,41527	21,45304	27,41527
x	0,069872	0,007084	9,863225	4,01E-06	0,053846464	0,085897	0,053846	0,085897

# Hledání vhodného regresního modelu

Lze postupovat dvěma způsoby:

1. Volba vhodného modelu na základě praktické zkušenosti či teoretických předpokladů
  2. Posouzením bodového grafu a interpretací nástrojů regresní analýzy
- Podle ad 2) je nejvhodnější model takový, který prochází všem vyšetřovaným bodům nejbližší.
  - Protože však vycházíme z výběrového souboru bodů, je třeba brát ohled na ad 1) !!!

Způsoby hodnocení vhodnosti regresního modelu

- analýza reziduálních hodnot
- výpočet směrodatné chyby odhadu ( $s_e$ )
- výpočet koeficientu determinace ( $r^2_{xy}$ ).

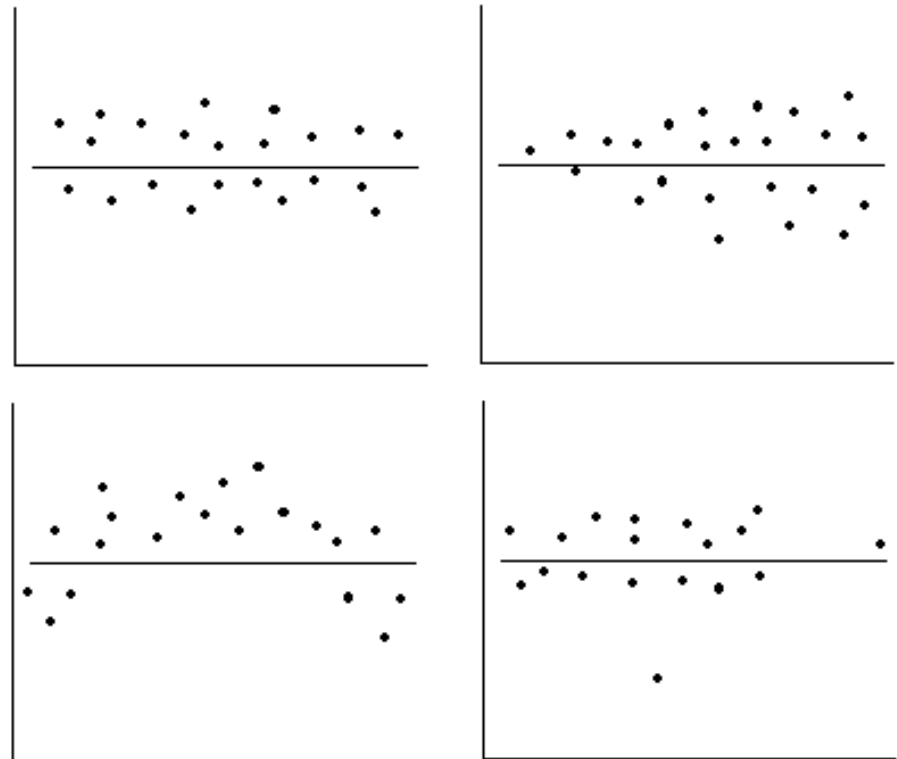
# Hledání vhodného regresního modelu

## Analýza reziduálních hodnot

**Rezidua** jsou vzdálenosti skutečných hodnot  $y_i$  od modelem odhadnutých hodnot  $y_j`$

Zvolený regresní model považujeme za vhodný, pokud reziduální hodnoty splňují všechny následující podmínky:

- rezidua jsou ***náhodná a nezávislá***
- mají ***normální rozdělení*** s nulovým průměrem a konstantním rozptylem
- rozptyl reziduí je ***konstantní***.



# Hledání vhodného regresního modelu

**Směrodatná chyba odhadu** – je vyjádřením směrodatné odchylky resp. rozptylu reziduálních hodnot a vhodnou mírou pro posouzení vhodnosti použité regresní závislosti

$$s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - y'_i)^2}{n - 2}}$$

Čím je hodnota reziduálního rozptylu nižší, tím je model vhodnější.

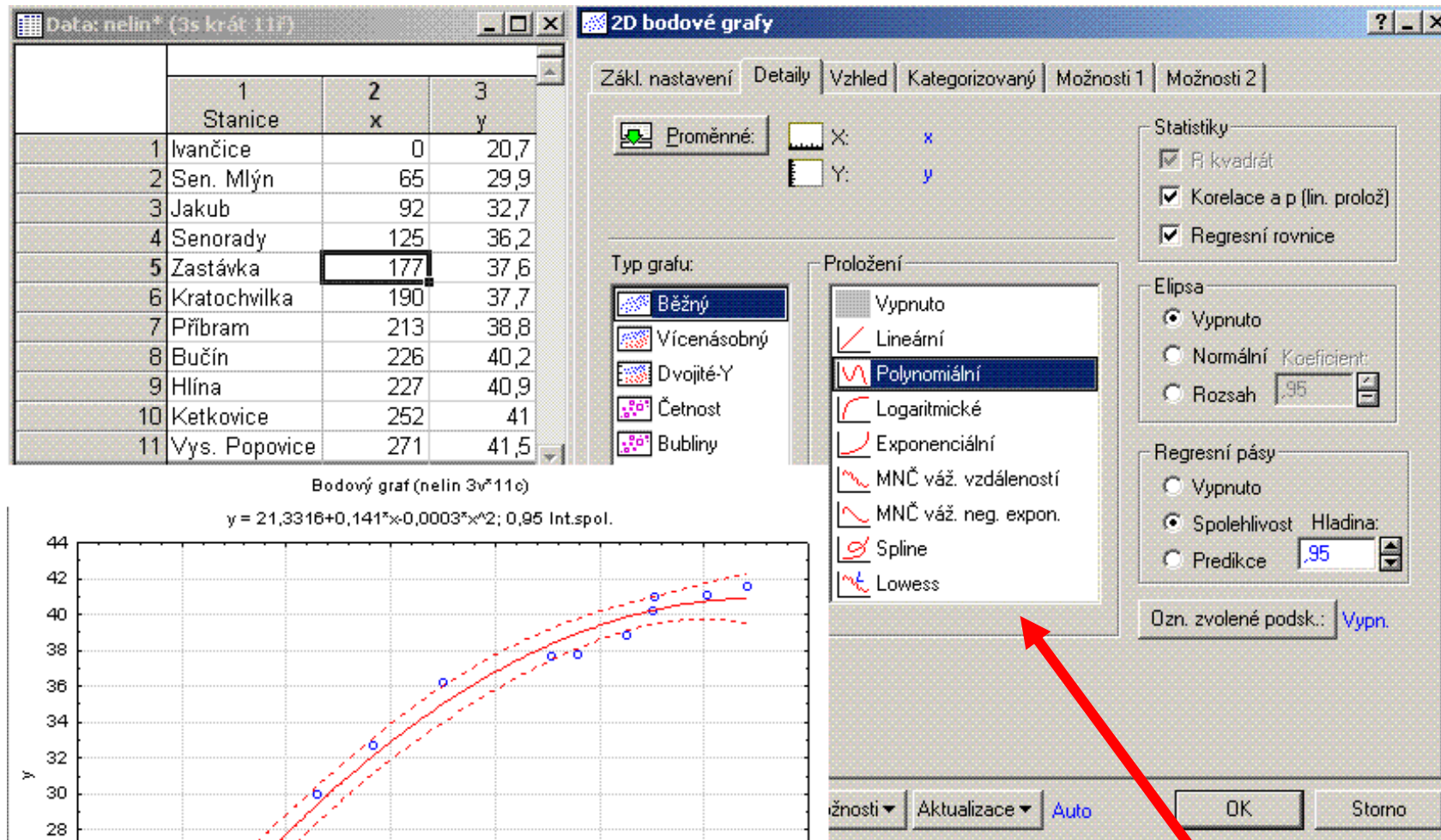
Koeficient determinace ( $r^2_{xy}$ ) – viz. Korelační počet

$$r^2 = \frac{SS_{regres}}{SS_{total}}$$

Čím je hodnota koeficientu determinace větší, tím je model vhodnější.

# Hledání vhodného regresního modelu

## Grafy – Bodové grafy



x:y:  $r = 0,9567$ ;  $p = 0,000004$ ;  $y = 24,4342 + 0,0699 \cdot x$

# Vícerozměrná regrese

Popisuje závislost více proměnných z nichž více je příčinami (vysvětlující proměnné) a jen jedna je důsledek (vysvětlovaná proměnná).

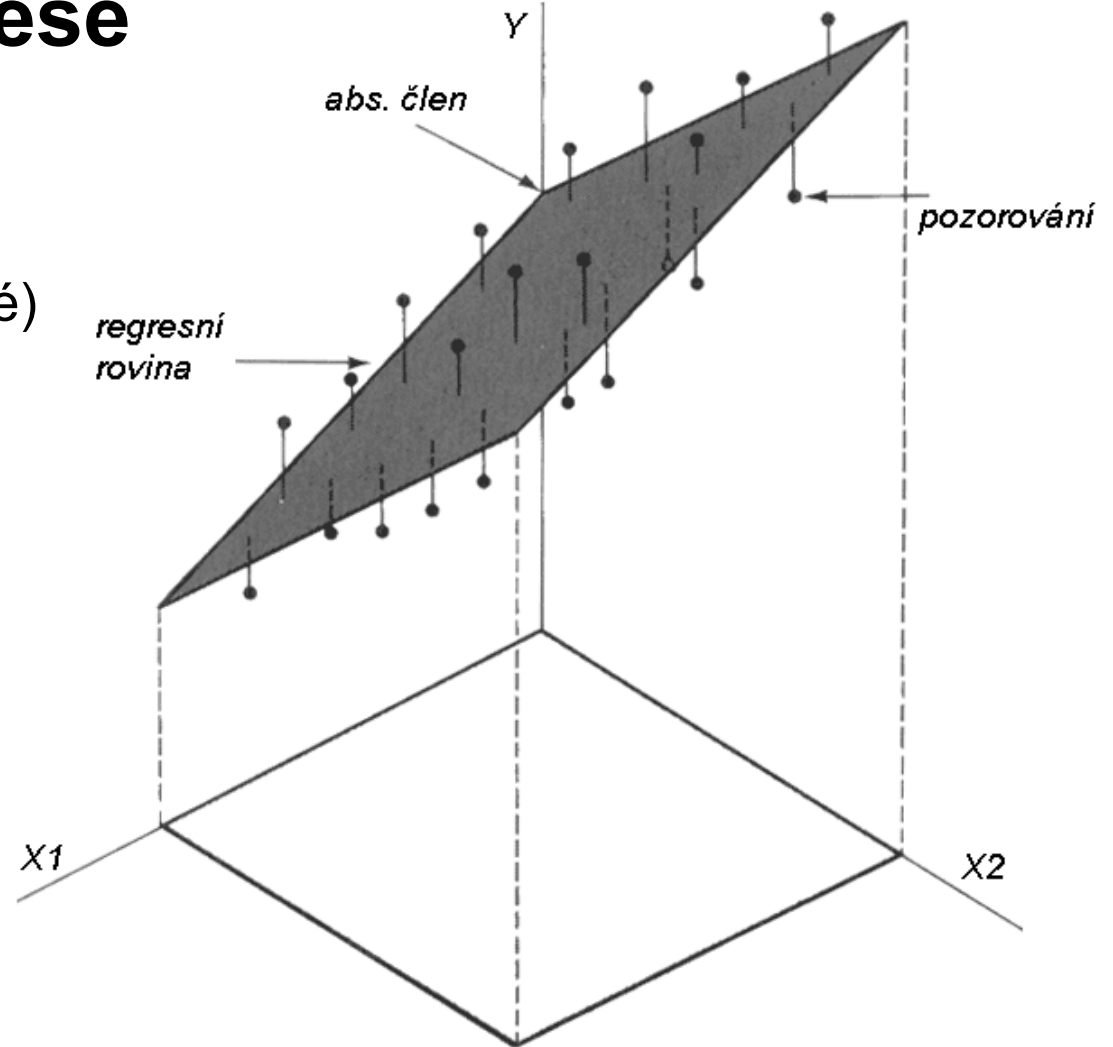
Jsou-li dvě vysvětlující proměnné regresní model je rovina

Odhad parametrů se provádí MNČ

Výstupy a interpretace jsou „**obdobné**“ jako u modelů jednorozměrné regrese

Např:

$$\text{Úhrn\_srážek} = 345,6 + 0,45 * \text{zem\_délka} + 1,23 * \text{nadm\_výška}$$



$$y' = a + b_1x_1 + b_2x_2 + \dots$$

# Vícerozměrná regrese

## Statistika – vícerozměrná regrese

(data viz. Brázdil a kol., 1995, str. 129, cvič. 8.3)

The screenshot displays the SPSS interface with three windows:

- Data: monohonas...**: A data grid with columns '1 mesic', '2 x1', '3 x2', and '4 y'. The data points are:
 

1 mesic	2 x1	3 x2	4 y
XI	25	155	200
XII	45	930	1092
I	34	383	463
II	192	1443	1789
III	136	1069	1258
IV	218	1460	1718
V	221	1208	1635
VII	201	1325	1670
VII	228	491	829
VIII	158	785	1018
IX	64	186	271
X	75	222	321
- Výsledky - vícerozměrná regrese: monohonas\_reg**: Summary statistics for the regression model.
 

Výsledky- vícerozm. regrese

Záv.prom. : y      vícenás. R = ,99721494      F = 804,5077  
 R^2 = ,99443763      sv = 2,9  
 Poč. případů: 12      uprav. R^2 = ,99320155      p = ,000000  
 Směrodatná chyba odhad: 49,781772065  
 Abs. člen: 11,028912950      Sm. chyba: 30,21982      t( 9) = ,36496      p = ,7236

x1 beta=,165      x2 beta=,874
- PS 1\* - Korelace (monohonas\_reg)**: Correlation matrix.
 

Proměnná	x1	x2	y
x1	1,000000	0,707573	0,783184
x2	0,707573	1,000000	0,990370
y	0,783184	0,990370	1,000000
- PS 1\* - Výsledky regrese se závislou proměnnou : y (monohonas\_reg)**: Detailed regression results.
 

Výsledky regrese se závislou proměnnou : y (monohonas\_reg)  
 R= ,99721494 R^2= ,99443763 Upravené R^2= ,99320155  
 F(2,9)=804,51 p<,00000 Směrod. chyba odhadu : 49,782

	Beta	Sm.chyba beta	B	Sm.chyba B	t(9)	Úroveň p
N=12						
Abs.člen			11,02891	30,21982	0,36496	0,723573
x1	0,165068	0,035181	1,24873	0,26614	4,69193	0,001133
x2	0,873572	0,035181	1,04975	0,04228	24,83066	0,000000

Punkva – pod Sk. Mlýnem (y)  
model odtoku:

$$y = 11,0289 + 0,165x_1 + 0,874x_2$$