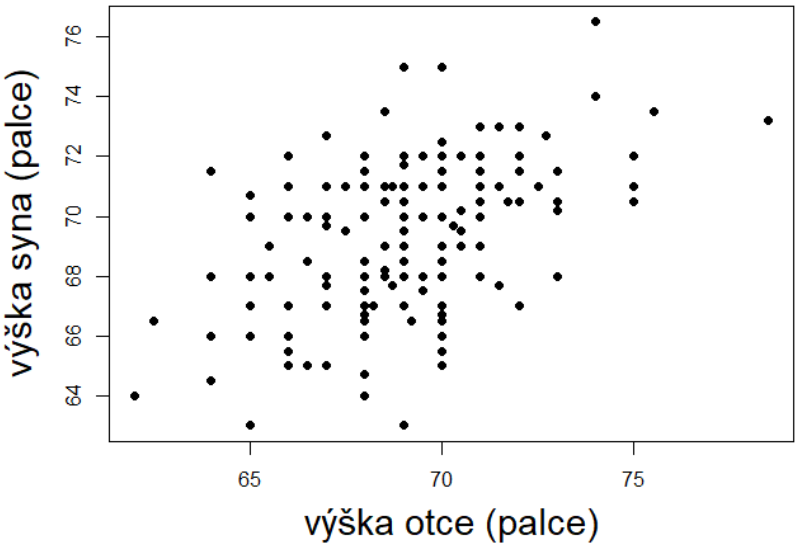
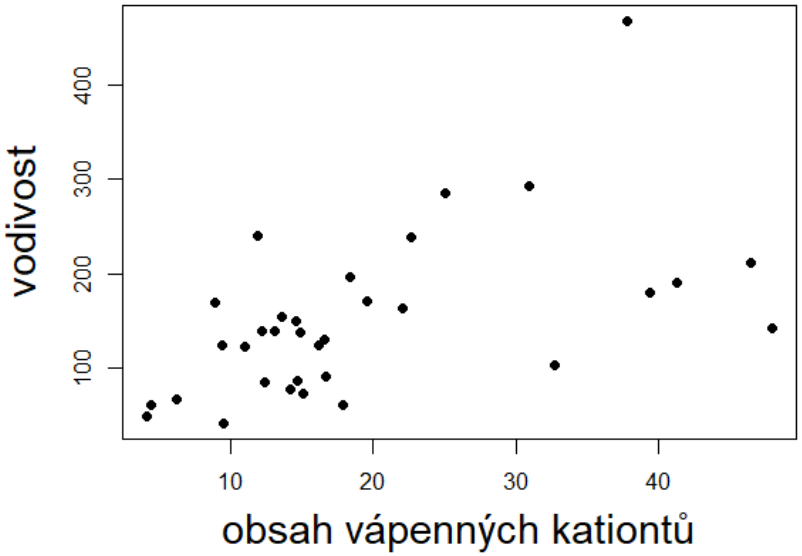
Příkladová data

Výška otce	Výška syna
175	178
177	173
188	188
173	173
163	164
163	168
178	169
…	…



Vodivost vody	Ca ionty
164	22.081
155	13.600
467	37.800
171	19.600
67	6.280
78	14.237



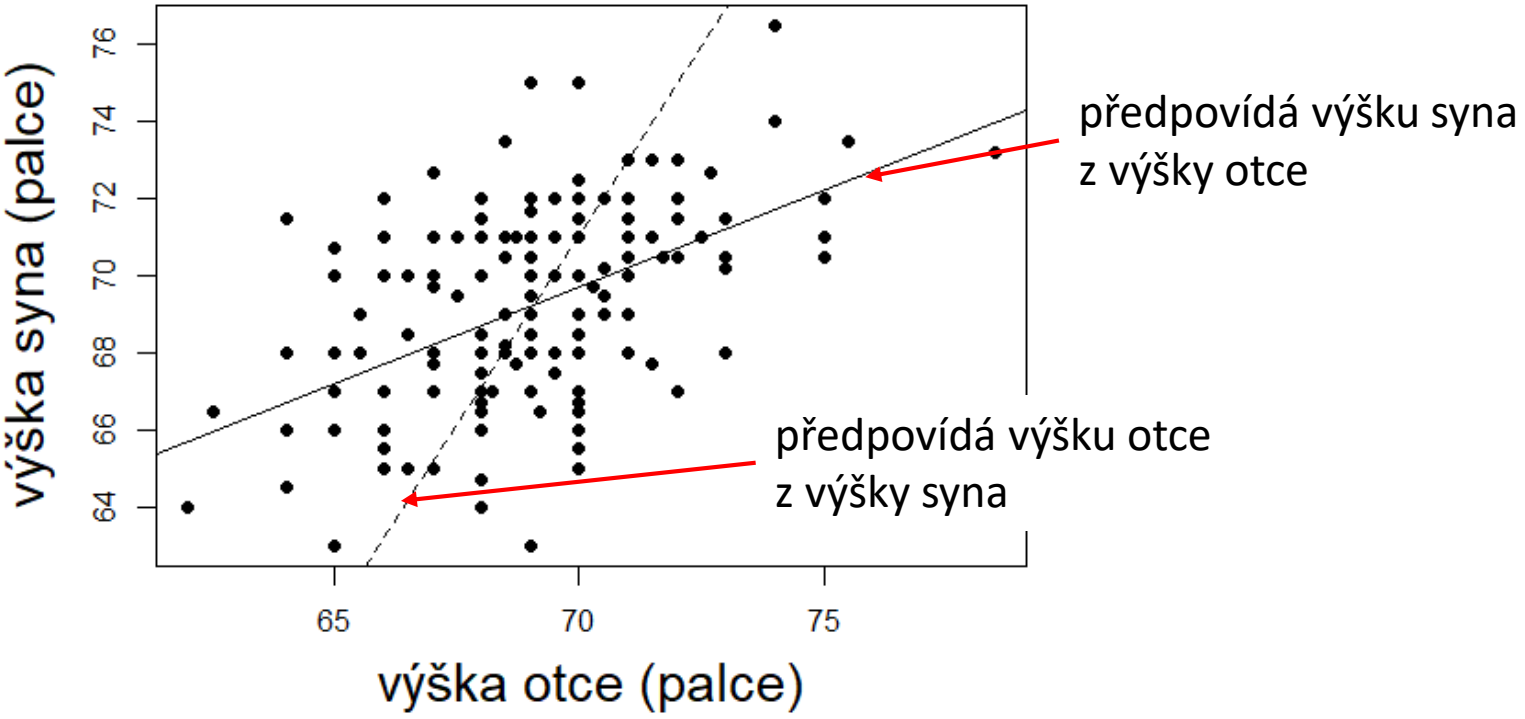
Analýza vztahu dvou kvantitativních proměnných

Dva přístupy, pohledy: **korelace a regrese**.

KORELACE popisuje sílu vzájemné závislosti.

REGRESE pomocí jedné proměnné popisuje hodnoty druhé proměnné

Příklad: výšky otce a syna (data GaltonSyn)



Korelace – závislost dvou kvantitativních proměnných [correlation]

- Data převážně spojitá, proto kvantitativní.
- Popisujeme sílu závislosti dvou proměnných

Opakování: $\boldsymbol{X}$ a $\boldsymbol{Y}$ náhodné veličiny

- Variance $\sigma_X^2 = var(X) = E\left[(X - \mu_X)^2\right]$
- kovariance = sdružená variance pro dvojici náhodných veličin ($\boldsymbol{X}$, $\boldsymbol{Y}$)
- Značení: $\sigma_{XY} = cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$
- Jsou-li $\boldsymbol{X}$ a $\boldsymbol{Y}$ nezávislé náhodné veličiny, potom $cov(X, Y) = 0$.

Proto je kovariance ukazatelem (populační) závislosti či nezávislosti mezi $\boldsymbol{X}$ a $\boldsymbol{Y}$. Hodnota kovariance se však mění s jednotkami měřené veličiny, proto používáme upravenou bezrozměrnou charakteristiku:

(populační) korelační koeficient:

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sqrt{var(X)\, var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = E\left[\frac{X - \mu_X}{\sigma_X} \cdot \frac{Y - \mu_Y}{\sigma_Y}\right]$$

normované tvary => bezrozměrnost

Populační korelační koeficient [(population) correlation coefficient]

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = E\left[\frac{X-\mu_X}{\sigma_X} \cdot \frac{Y-\mu_Y}{\sigma_Y}\right]$$

Vlastnosti:

- Vyjadřuje **sílu (těsnost) lineární závislosti** mezi dvěma náhodnými veličinami.
- Je **bezrozměrný**, nezávislý na použitém měřítku, jednotkách měření.
- Hodnota koeficientu **nezávisí na pořadí** náhodných veličin, protože platí, že
  $cov(X,Y) = cov(Y,X) \rightarrow \rho_{X,Y} = \rho_{Y,X}$
- Korelační koeficient nabývá hodnot $\boldsymbol{\rho_{X,Y} \in \langle -1, 1 \rangle}$ včetně.
  - Když $\boldsymbol{\rho_{X,Y} = 1}$, leží všechny dvojice (X, Y) na přímce, která roste => nejsilnější pozitivní závislost.
  - Když $\boldsymbol{\rho_{X,Y} = -1}$, leží všechny dvojice (X, Y) na přímce, která klesá => nejsilnější negativní závislost.
  - Když $\boldsymbol{\rho_{X,Y} = 0}$, očekáváme velmi slabou závislost. Nicméně existují i takové kombinace hodnot, se kterými veličiny závislé získají nulovou kovarianci.
  - Ale pro veličiny ***X*** a ***Y*** **nezávislé** vždy platí $\boldsymbol{cov(X,Y) = 0}$, a proto je $\boldsymbol{\rho_{X,Y} = 0}$. Informaci zužitkujeme dále.

Výběrový korelační koeficient

Předchozí úvahy platily pro teoretické populační charakteristiky.

Pracujeme-li s **výběry *X* a *Y***, skutečnou hodnotu $\rho_{X,Y}$ neznáme.

Výběr má vypadat takto: dvojice (X_i, Y_i) k sobě nějakým způsobem patří (např. výška otce a syna, různé rozměry stejného jedince, charakteristiky jednoho vzorku), dvojice jsou ale mezi sebou nezávislé. ***n*** znamená počet dvojic ve výběru.

Potom výběrové rozptyly jsou:

$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

A výběrová kovariance:

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Z těchto odhadů vychází parametrický odhad korelačního koeficientu: $\boldsymbol{r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}}$

Výběrový Pearsonův (lineární) korelační koeficient

[Pearson's product-moment correlation coefficient]

Předpoklad: ***X*** a ***Y*** pocházejí z normálního rozdělení

Odhad koeficientu:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_X}\right)\left(\frac{Y_i - \bar{Y}}{S_Y}\right)$$

výběrová obdoba normovaného tvaru: z-skóry

Jsou-li ***X*** a ***Y*** nezávislé, je $\rho_{XY} = 0$. Ale výběrový koeficient je jen odhad a tudíž náhodná veličina, která se pohybuje kolem skutečné hodnoty 0.

Proto chceme testovat hypotézu H_0: $\rho_{XY} = 0$, z čehož odvozujeme, že veličiny ***X*** a ***Y*** jsou nezávislé. Fakt, že i závislé veličiny mohou mít $\rho_{XY} = 0$, se vejde do prsti. α.

Ověříme předpoklad o normálním rozdělení ***X***, ***Y*** . (Správně také ***X*** + ***Y*** .)

Testová statistika: $T = \frac{r_{XY}}{\sqrt{1-r_{XY}^2}}\sqrt{n-2} \sim t_{n-2}$

Kritérium pro oboustrannou hypotézu: $|T| \geq t_{n-2}(1 - {}^{\alpha}/_{2})$

Pearsonův korelační koeficient – poznámky:

- Pearsonův odhad korelačního koeficientu je vychýlený, $\boldsymbol{E}r = \rho - \frac{1-\rho^2}{n} + o\left(\frac{1}{n}\right)$
- Předpoklad pro užití *t*-testu je, že $\boldsymbol{r_{XY}}$ má normální rozdělení. Toto je splněno, jen pokud platí nulová hypotéza, že $\boldsymbol{\rho_{XY}} = 0$. Potom $r_{XY} \dot{\sim}_{H_0} N\left(0, \frac{1}{n-1}\right)$.
- Když $\boldsymbol{\rho_{XY}} \neq 0$ (zamítám H_0), potom $\boldsymbol{r_{XY}}$ nemá normální rozdělení. Pomůže Fisherova z-transformace: $Z = \frac{1}{2} \cdot ln\frac{1+r}{1-r}$. Potom $\boldsymbol{E}Z \doteq \frac{1}{2} \cdot ln\frac{1+\rho}{1-\rho}$, $\mathbf{var}\, Z \doteq \frac{\mathbf{1}}{\boldsymbol{n-3}}$ a rozdělení $\boldsymbol{Z}$ se velmi rychle blíží normálnímu rozdělení.
- Toto využiju k sestrojení konfidenčního intervalu pro odhad nenulového $\boldsymbol{\rho}$:

  Testuji H_0: $\rho_{XY} = \rho_0$, kde $\rho_0 \in (-1, 1)$. H_1: $\rho_{XY} \neq \rho_0$.

  Spočtu Z a přepočítám $\rho_0 \rightarrow \zeta_0 = \frac{1}{2} \cdot ln\frac{1+\rho_0}{1-\rho_0}$ [čti: *dzéta*].

  Test. statistika: $U = \frac{Z-\zeta_0}{\sqrt{\frac{1}{n-3}}} \dot{\sim} N(0,1) \rightarrow P\left(z\left(\frac{\alpha}{2}\right) < \sqrt{n-3}\left(Z - \frac{1}{2} \cdot ln\frac{1+\rho_0}{1-\rho_0}\right) < z\left(1-\frac{\alpha}{2}\right)\right) \doteq 1-\alpha$

  Odtud posléze $\underbrace{\frac{D-1}{D+1} < \rho < \frac{H-1}{H+1}}_{\textit{konfidenční interval}}$, kde $D = exp\left\{2Z - \frac{2 \cdot z(1-\alpha/2)}{\sqrt{n-3}}\right\}$

  $H = exp\left\{2Z - \frac{2 \cdot z(\alpha/2)}{\sqrt{n-3}}\right\}$

Pearsonův korelační koeficient – poznámky:

- Existuje také test pro porovnání několika nezávislých korelačních koeficientů a následně mnohonásobné porovnání dvojic těchto koeficientů.
- Korelační matice: měřím několik charakteristik na jednom subjektu, ve výběru mám n subjektů. Pro popis těsnosti vzájemné závislosti všech dvojic měřených charakteristik jsou všechny korelační koeficienty uspořádané je do matice.

	Otec	Matka	Syn
Otec	1	0.12	0.50
Matka	0.12	1	0.29
Syn	0.50	0.29	1

- Parciální korelace: vyjadřují vzájemnou závislost dvou proměnných za předpokladu, že třetí proměnná (nebo více proměnných) se nemění. Lze také říci, že je to korelační koeficient po „odfiltrování“ vlivu třetí proměnné. Další formulace říká, že je to korelace podmíněná hodnotami třetí proměnné, tzv. parciální korelační koeficient prvního řádu. Souvisí s regresními koeficienty mnohonásobného regresního modelu. Více Lepš & Šmilauer, str. 299.

Neparametrická varianta:

Výběrový Spearmanův korelační koeficient

Předpoklad: ***X*** i ***Y*** mají nějaké spojité rozdělení

Určím pořadí: $X_1, X_2, \dots, X_n \rightarrow R_1, R_2, \dots, R_n$

$Y_1, Y_2, \dots, Y_n \rightarrow Q_1, Q_2, \dots, Q_n$

R_i a Q_i dosadím místo X_i a Y_i úpravou získáme následující vzorec

Výpočet: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$

Opět $r_S \in \langle -1, 1 \rangle$.

Hypotézu H_0: $\rho_{XY} = 0$, tedy že ***X*** a ***Y*** jsou nezávislé náh. veličiny, můžeme otestovat dvojím způsobem:

a) nejsou shody v pořadí: přesné hodnoty v tabulkách (např. R až do ***n*** < 1290)

b) jsou shody v pořadí nebo velká ***n***: aproximace normálním nebo *t*-rozdělením

→ $|r_S|\sqrt{n-1} \dot{\sim} N(0,1) \rightarrow$ pro $|r_S|\sqrt{n-1} \geq z(1-{}^{\alpha}/_{2})$ zamítám H_0

→ $T = \frac{r_S}{\sqrt{1-r_S^2}}\sqrt{n-2} \dot{\sim} t_{n-2}$

Spearmanův korelační koeficient - poznámky

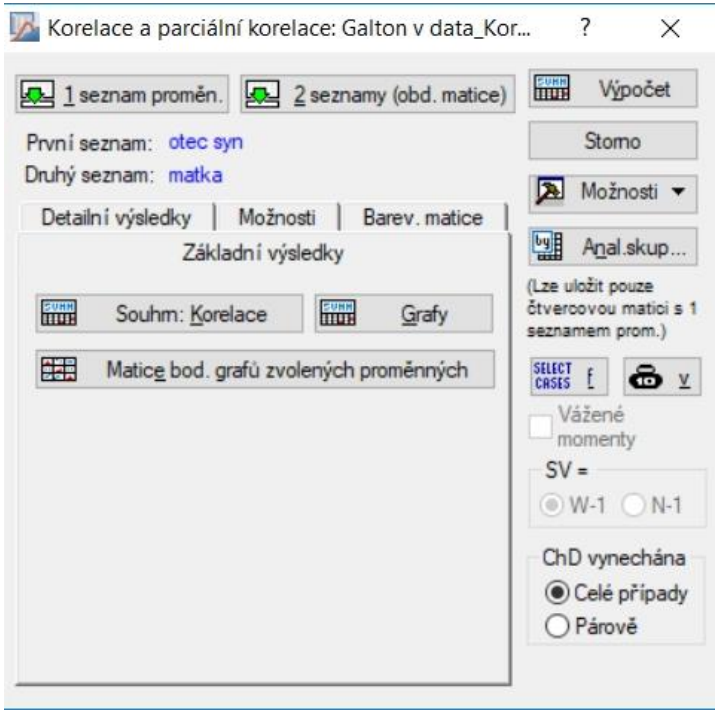
Výhody Spearmanova korelačního koeficientu

- Je citlivý na jakoukoli monotónní závislost, nejen na lineární
- Je méně citlivý k výskytu odlehlých hodnot

Příklad EMISE

Korelační koeficient – výpočet v softwaru

STAT: Statistiky → Základní statistiky → Korelační matice



R: `cor.test(x, y, method = „s“)`

metody: Pearson (default)

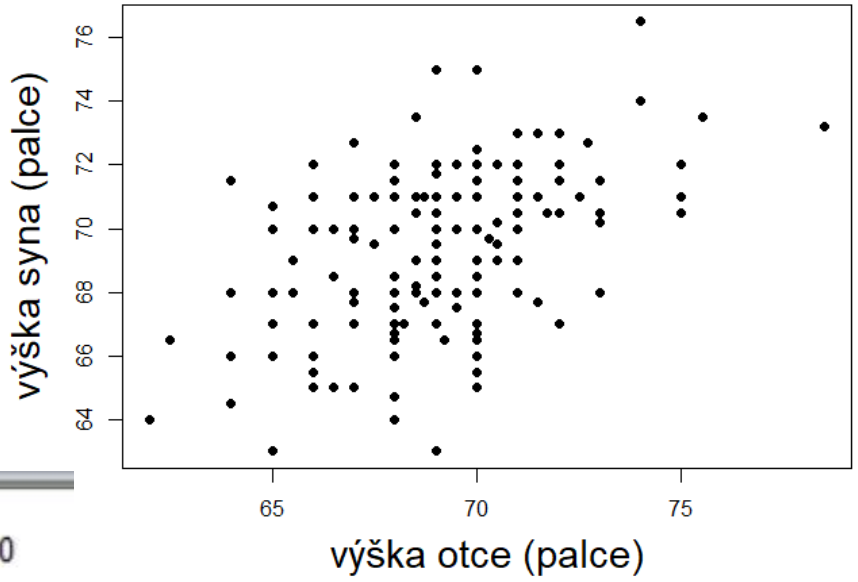
Spearman , Kendall

Korelace – příklady:

data: GaltonSyn

Výšky obou rodičů a náhodně vybraného syna v palcích.
Je nějaká závislost mezi výškou otce a syna?

(záložka *Možnosti: Formát …* = první volba)



Korelace (Galton v data_KoreRegre)
Označ. korelace jsou významné na hlad. p < ,05000
N=173 (Celé případy vynechány u ChD)

Proměnná	Průměry	Sm.odch.	otec	syn
otec	69,09075	2,542825	1,000000	0,504980
syn	69,26416	2,522593	0,504980	1,000000

výběrový Pearsonův korelační koeficient

(záložka *Možnosti: Formát …* = druhá volba)

Korelace (Galton v data_KoreRegre)
Označ. korelace jsou významné na hlad. p < ,05000
N=173 (Celé případy vynechány u ChD)

Proměnná	otec	syn
otec	1,0000	,5050
	p= ---	p=,000
syn	,5050	1,0000
	p=,000	p= ---

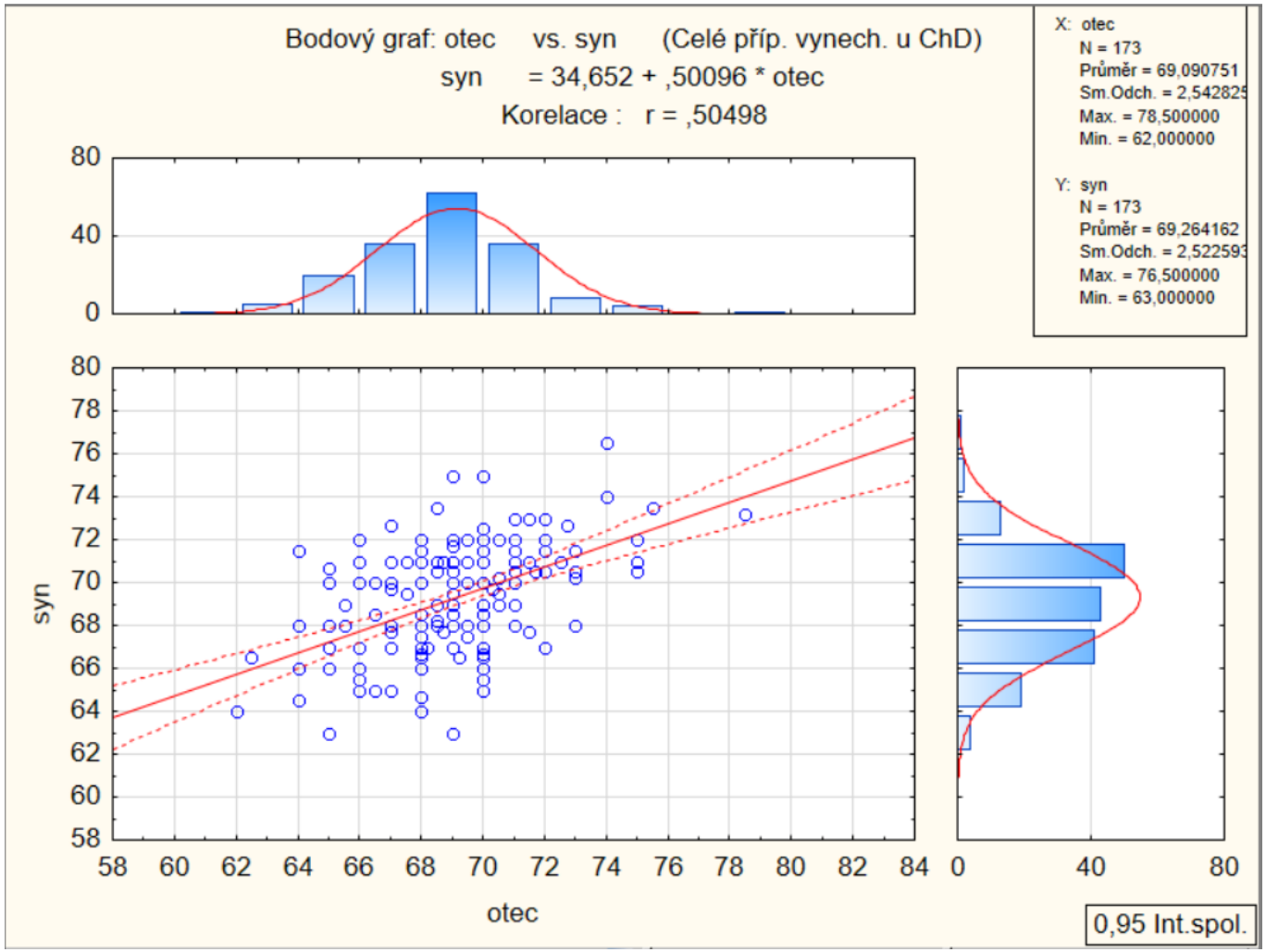
p-hodnota testu hypotézy, že ρ = 0.

Hodnoty t-statistiky ve třetí volbě *Formátu zobrazení*

Korelace – příklady:

data: GaltonSyn

Zde mohu posoudit normalitu obou proměnných.
Také je zobrazena regresní přímka a její konfidenční interval.
V popisu grafu je vypsána rovnice regresní přímky.

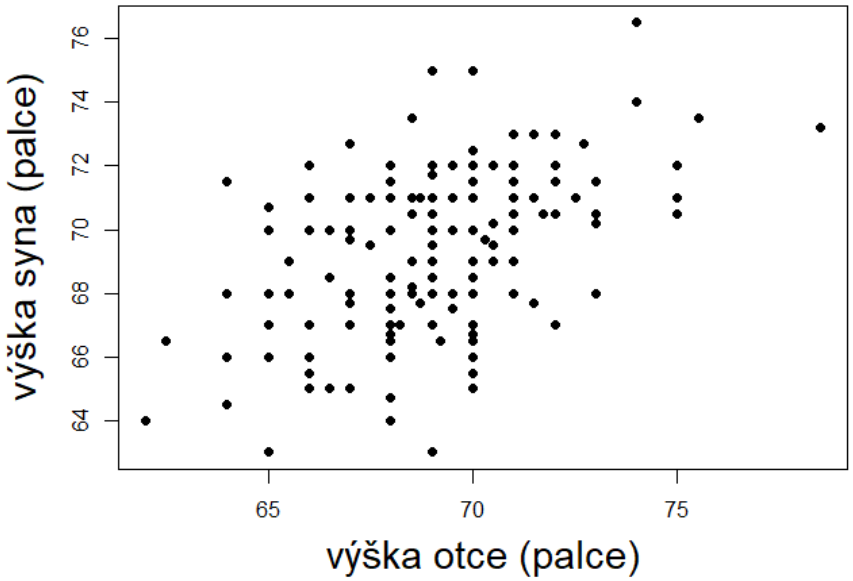


Korelace – příklady:

data: GaltonSyn

R:

```
> cor.test(otec, syn)
        Pearson's product-moment correlation
data:  otec and syn
t = 7.6506, df = 171, p-value = 1.392e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3847696 0.6083457
sample estimates:
      cor
0.5049801
```



Korelace – příklady: neparametricky

data: Potoky

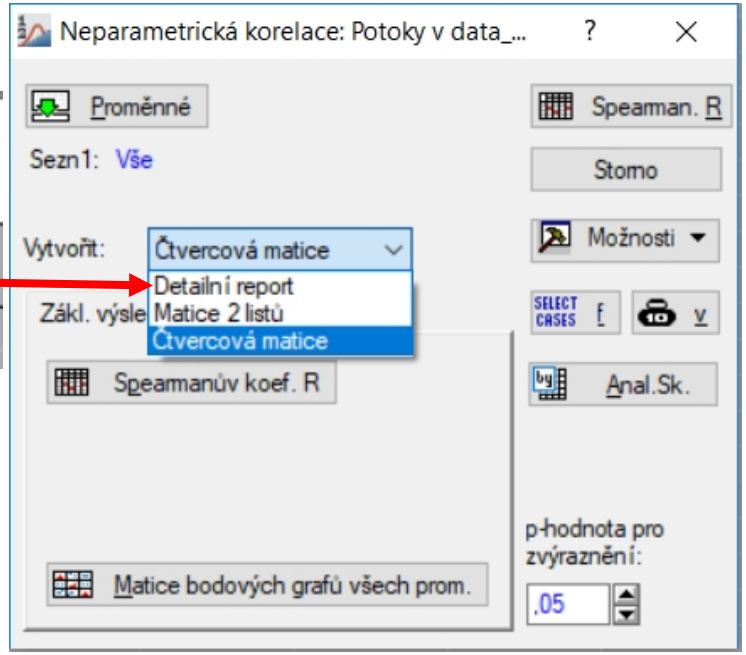
Ze 33 šumavských potoků máme údaje o vodivosti a o obsahu vápenných iontů. Očekáváme pozitivní závislost, ptáme se však, jak je silná. (To nás opravňuje použít jednostrannou alternativu při testování.)

Při kontrole histogramů vidíme, že obsah vápenných iontů má výrazně šikmé rozdělení. Použijeme proto Spearmanův korelační koeficient:

Statistiky → Neparametrické statistiky → Korelace

Spearmanovy korelace (Potoky v data_KoreRegre)
ChD vynechány párově
Označ. korelace jsou významné na hl. p <,05000

Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	p-hodn.
Ca & Conduct	33	0,584106	4,006724	0,000359



Na proměnnou ***Ca*** jsme také mohli použít logaritmickou transformaci a pokračovat na Pearsonův korelační koeficient.

Korelace – příklady:

data: Potoky

R:

```
> cor.test(Conduct, Ca, method = "spearman")
        Spearman's rank correlation rho
data:  Conduct and Ca
S = 2488.7, p-value = 0.0003585
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5841063
```