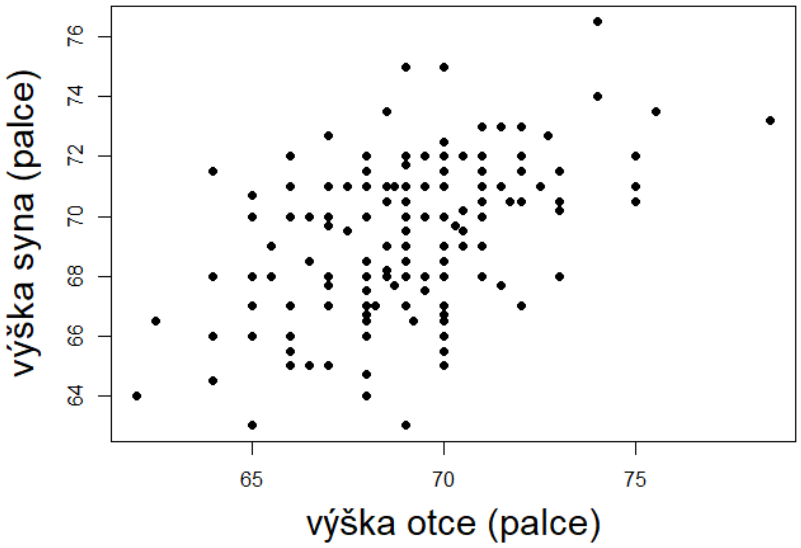
# Přehled základních metod:

| | **Kvalitativní data, kategorie** (skupiny, faktory, …) | **Kvantitativní data** (délka, hmotnost, množství, …) |
|---|---|---|
| **Kvantitativní data** (délka, hmotnost, množství, …) | t-test, Wilcoxon, ANOVA | Korelace, regrese |
| **Kvalitativní data, kategorie** (skupiny, faktory, …) | Kontingenční tabulky | Logistická analýza |
| | Kvalitativní data, kategorie, skupiny | Kvantitativní data (délka, hmotnost, množství, …) |

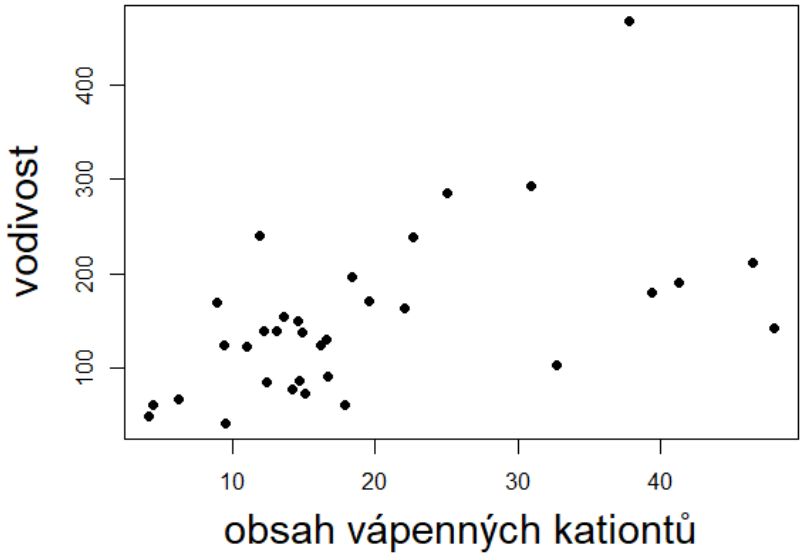
# Příkladová data

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




**data Galton**

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

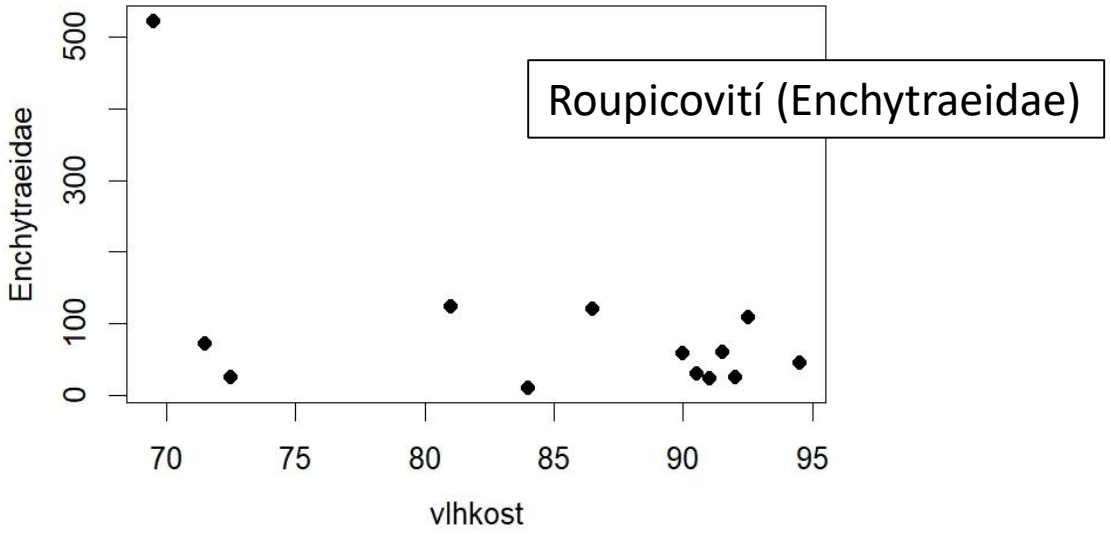


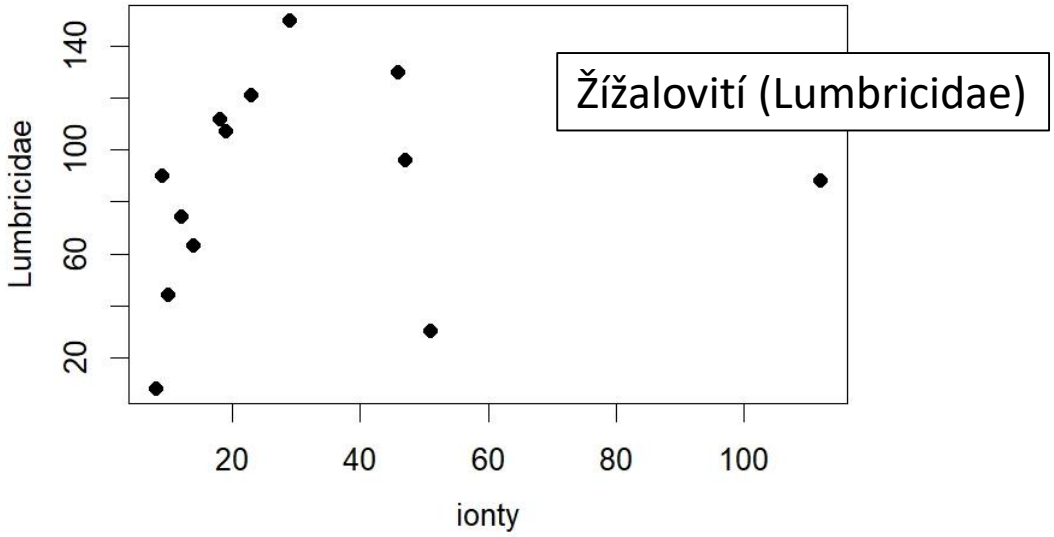

**data Potoky**

# Příkladová data Opaskovci

| Vlhkost půdy | počet jedinců |
|---|---|
| 71 | 71 |
| 72 | 25 |
| 81 | 123 |
| 84 | 10 |
| 86 | 120 |
| 90 | 58 |
| … | … |

| Ca ionty | počet jedinců |
|---|---|
| 12 | 74 |
| 47 | 96 |
| 18 | 112 |
| 46 | 130 |
| 8 | 8 |
| … | … |



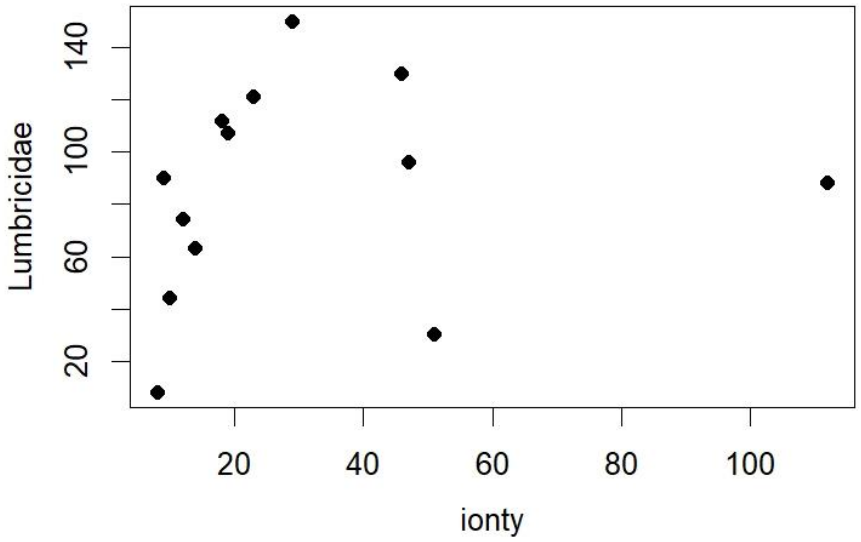



# Nezávislost náhodných veličin [independence]

Situace: veličiny ***X*** a ***Y*** naměřené na stejném subjektu, subjektů ***n***.
(např. počet žížal a množství iontů *Ca* v jednom vzorku, 20 nezávislých vzorků)

***X*** a ***Y*** jsou nezávislé, když pro podobné hodnoty ***X*** dostávám zcela různé hodnoty ***Y.***

Příklad pro spojitá data:
vzorky s podobným obsahem iontů *Ca* měly zcela různé počty žížal.



# Nezávislost náhodných veličin

Příklad pro kategoriální data:

parazitace ryb rodem *Diplostomum* a příslušnost ryby k vodnímu typu.

Máme 100 ryb a tyto údaje: příslušnost k vodnímu typu a přítomnost parazita.
Je výskyt parazita *Diplostomum* závislý na vodním prostředí?

Počet ryb:

| rameno | řeka | stojatá | štěrkovna |
|---|---|---|---|
| 25 | 25 | 25 | 25 |

| Diplostomum | |
|---|---|
| ano | ne |
| 25 | 75 |

| | rameno | řeka | stojatá | štěrkovna |
|---|---|---|---|---|
| Diplo NE | 24 | 23 | 8 | 20 |
| Diplo ANO | 1 | 2 | 17 | 5 |

Nezávislost znamená, že

P(ryba ve stojatá) * P(ryba parazitovaná) = P(ryba ve stojatá & parazitovaná)

25/100 * 25/100 = 17/100

0.25*0.25 = 0.0625 ≠ 0.17

Dá se očekávat, že výskyt parazita na vodním prostředí bude ZÁVISLÝ. Testujeme…

# Nezávislost náhodných veličin

Statisticky: chování veličiny ***X*** je **nezávislé** na chování veličiny ***Y***

Pozná se to na pravděpodobnostech, že ***X*** bude mít nějakou hodnotu
a zároveň ***Y*** bude mít nějakou hodnotu.

Pro nezávislé diskrétní veličiny platí, že

$$\boldsymbol{P\left(X = x_i^*, Y = y_j^*\right) = P\left(X = x_i^*\right) \cdot P\left(Y = y_j^*\right)} \ \ldots \text{ pro každé } \boldsymbol{x_i^*} \text{ a } \boldsymbol{y_j^*}$$

a pro nezávislé spojité veličiny:

$$\boldsymbol{f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)} \ \ldots \text{pro každé } \boldsymbol{x} \text{ a } \boldsymbol{y}$$

# Kovariance veličin X a Y [covariance]

= sdružená variance pro dvojici náhodných veličin ***X*** a ***Y***

popisuje vzájemnou závislost či nezávislost ***X*** a ***Y***

Značení a teoretický výpočet (= populační kovariance):

$$\sigma_{XY} = cov(X,Y) = E(X - EX)(Y - EY) = E(X - \mu_X)(Y - \mu_Y)$$

Všimněte si, že $cov(X,X) = E(X - EX)(X - EX) = E(X - EX)^2 = var\ X$

Jsou-li ***X***, ***Y*** nezávislé, potom platí, že $\sigma_{XY} = cov(X,Y) = 0$.

**!** Neplatí naopak:

i když je $cov(X,Y) = 0$, přesto ***X*** a ***Y*** mohou být závislé!

# Korelační koeficient X a Y [coefficient of correlation]

= kovariance normovaných tvarů dvou náhodných veličin ***X***, ***Y***

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

*Čti: ró xy*

- měří sílu lineární závislosti ***X*** a ***Y***
- normování zajistí bezrozměrnost, můžeme proto srovnávat sílu závislosti mezi veličinami o různých průměrech a rozptylech
- nezávislé X, Y mají $\sigma_{XY} = 0$, tedy také $\rho_{XY} = 0$. (Naopak to neplatí!)
- svých extrémních hodnot $\pm 1$ nabývá tehdy, když $Y = \alpha \pm \beta X$ (lineární vztah)

Odvození horní rovnosti:

$$\rho_{XY} = cov\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) = E\left[\left(\frac{X-\mu_X}{\sigma_X} - 0\right) \cdot \left(\frac{Y-\mu_Y}{\sigma_Y} - 0\right)\right] =$$

$$= \frac{1}{\sigma_X \sigma_Y} E[(X-\mu_X)(Y-\mu_Y)] = \frac{cov(X,Y)}{sd(X)sd(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

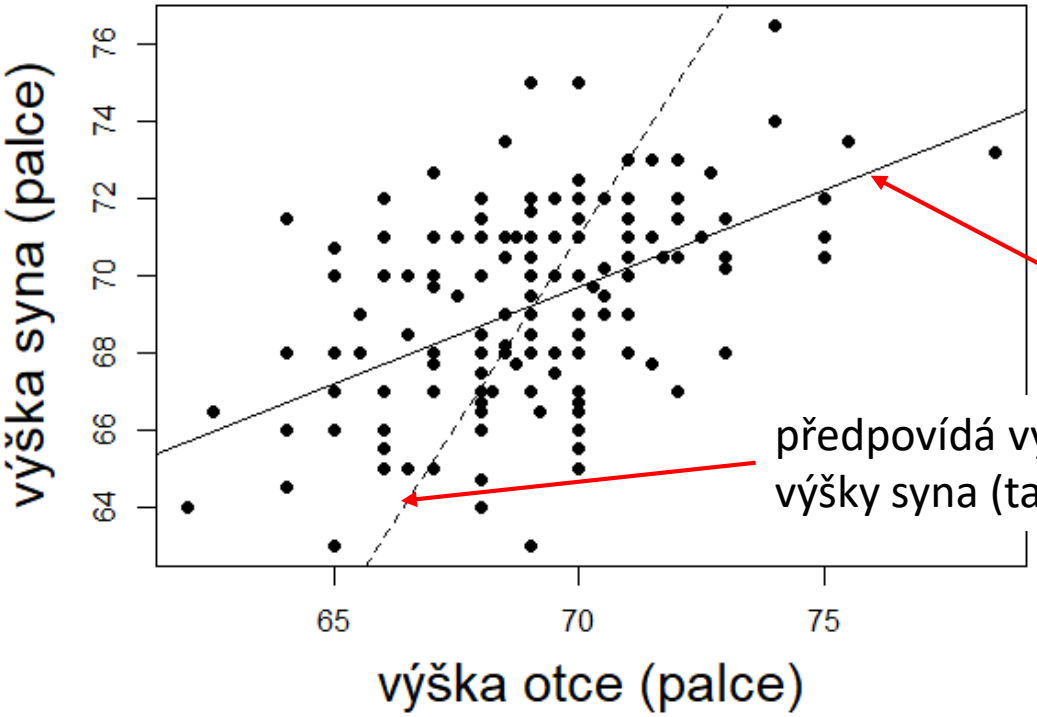
# Analýza vztahu dvou kvantitativních proměnných

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

**KORELACE** popisuje sílu vzájemné závislosti = **korelační koeficient.**

**REGRESE** pomocí jedné proměnné popisuje hodnoty druhé proměnné

Příklad: výšky otce a syna (data GaltonSyn)




**Korelace žádnou čáru neprodukuje !!**

**Regrese:** předpovídá výšku syna z výšky otce

předpovídá výšku otce z výšky syna (taky regrese)

# Korelace – síla závislosti dvou kvantitativních proměnných

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

## Kovariance → korelační koeficient

Kovariance je ukazatelem (populační) závislosti či nezávislosti mezi ***X*** a ***Y***.
Hodnota kovariance se však mění s jednotkami měřené veličiny, proto používáme upravenou bezrozměrnou charakteristiku:

**(populační) korelační koeficient**:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

# 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 jednotkách měření**, na použitém měřítku.
- Hodnota koeficientu **nezávisí na pořadí** náhodných veličin,

  protože platí $cov(X,Y) = cov(Y,X) \rightarrow \rho_{X,Y} = \rho_{Y,X}$
- 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 nezávislost.

# 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:

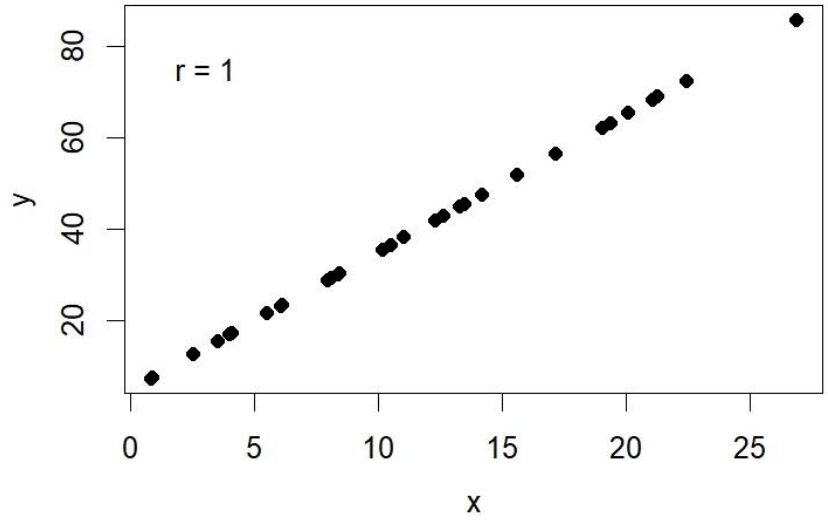
$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

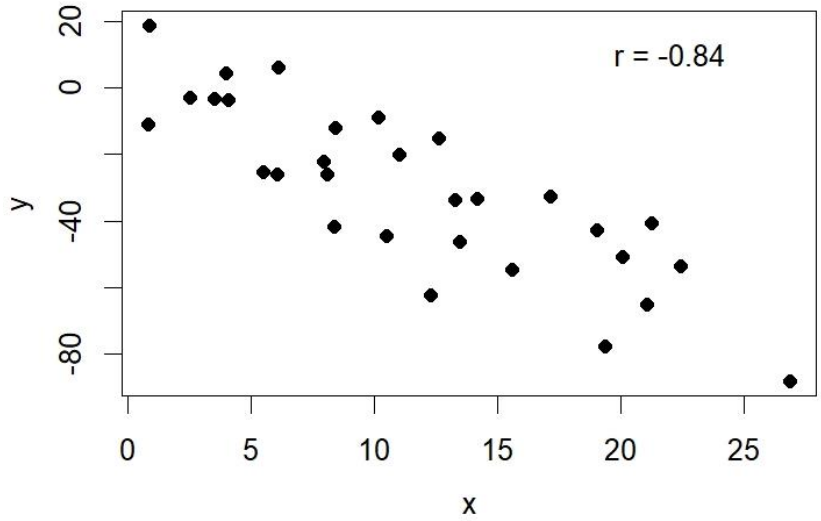
A výběrová kovariance:

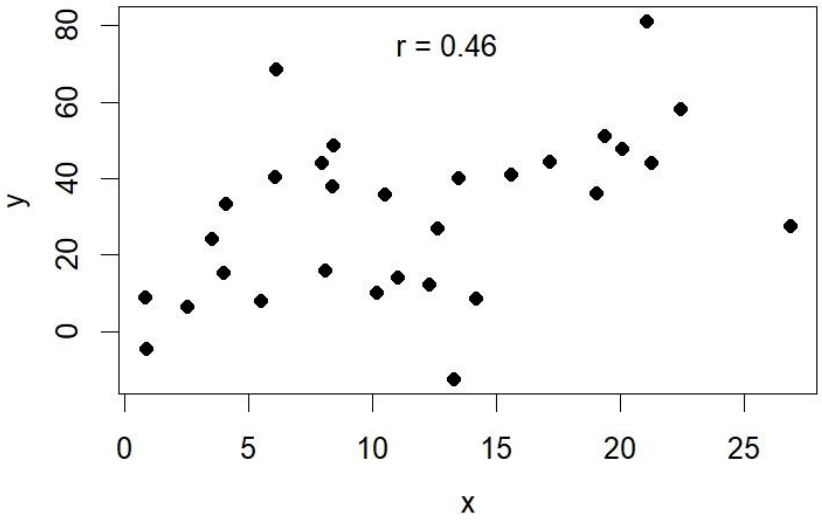
$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

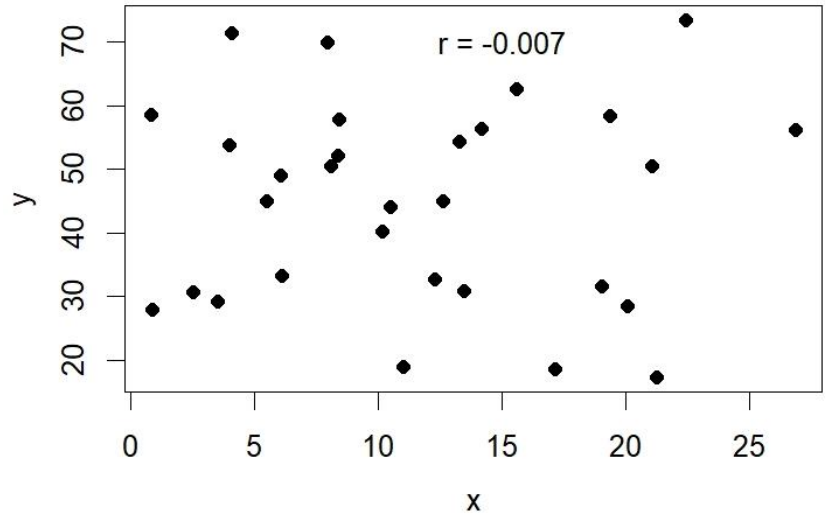
Z těchto odhadů vychází parametrický odhad korelačního koeficientu: $r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$

# Výběrový korelační koeficient



# Parametricky: 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

**R:**
```
cor(x, y, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

```
cor(Galton$otec, Galton$syn)   ## ! nezná parametr „data=…"
[1] 0.5057

cor(Galton)        ## spočítá tzv. korelační matici
       otec  matka    syn
otec  1.000 0.117  0.506
matka 0.117 1.000  0.291
syn   0.506 0.291  1.000
```

## Parametrický test

## výběrový Pearsonův korelační koeficient

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

Jsou-li ***X*** a ***Y*** nezávislé, má být populační $\boldsymbol{\rho_{XY} = 0}$.

Ale výběrový koeficient je jen odhad a tudíž se chová jako náhodná veličina, která se pohybuje kolem skutečné hodnoty 0.

Proto chceme testovat hypotézu $\boldsymbol{H_0: \rho_{XY} = 0}$.
Když $\boldsymbol{H_0}$ zamítáme, prokazujeme, že mezi hodnotami veličin ***X*** a ***Y*** je závislost, vztah.
Když $\boldsymbol{H_0}$ nezamítáme, tak data vyhovují hypotéze, že hodnoty ***X*** a ***Y*** jsou nezávislé.
Fakt, že i závislé veličiny mohou mít $\rho_{XY} = 0$, se vejde do pravděpodobnosti $\boldsymbol{\alpha}$.

**R:**
```
cor.test(x, y,
      alternative = c("two.sided", "less", "greater"),
      method = c("pearson", "kendall", "spearman"),
      exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

cor.test(formula, data, subset)
```

# Parametrický test

Ověříme předpoklad o normálním rozdělení $X$ a $Y$.

Testujeme hypotézu $H_0: \rho_{XY} = 0$.

Testová statistika: $t = \frac{r_{XY} - 0}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2} \sim t_{n-2}$

```
cor.test(otec, syn)

        Person's product-moment correlation

data:  otec and syn
t = 7.6657, df = 171, p-value = 1.275e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.3856165 0.6089717
sample estimates: cor = 0.5057206
```

Zamítám hypotézu o nezávislosti výšky otce a syna.

# Pearsonův korelační koeficient – poznámky:

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

  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 $\underbrace{\frac{D-1}{D+1} < \rho < \frac{H-1}{H+1}}_{\text{konfidenční interval}}$, 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:

- 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 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: $\boldsymbol{r_{S\,(X,Y)} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2}$

Opět $\boldsymbol{r_S \in \langle -1, 1 \rangle}$.

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

a) nejsou shody v pořadí: dostanu přesné hodnoty (*exact p-value*)

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

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

Nevýhoda: Rková metoda nepočítá konfidenční interval.

**Příklad Opaskovci**

```
> cor.test(Enchytraeidae, vlhkost, method = "s")

        Spearman's rank correlation rho

data:  Enchytraeidae and vlhkost
S = 470.15, p-value = 0.3337
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.2916096

Warning message:
In cor.test.default(Enchytraeidae, vlhkost, method = "s") :
  Cannot compute exact p-value with ties
```

Nezamítám hypotézu o nezávislosti počtu roupicovitých na vlhkosti půdy (v rozmezí 70 – 90 % vlhkosti).

# Transformace dat – přiblížení dat normálnímu rozdělení

## Logaritmická transformace

$$Y_i = \log(X_i)$$

Kdy použít? Když mají faktory spíše násobný účinek (týká se analýzy rozptylu)

Když mají data výrazně pozitivně šikmé rozdělení.

Typické pro koncentrace látek, pro hmotnosti i rozměry, také počty jedinců, zvláště při shlukovitém rozmístění.

Omezení: data nesmí obsahovat *záporná čísla*.

Také *nuly jsou problém*, ale zde pomůže přidat posunutí: $Y_{ijt} = \log(X_{ijt} + \boldsymbol{c})$, kde $\boldsymbol{c}$ odpovídá nejmenším nenulovým hodnotám v datech, případně $\boldsymbol{c} = 1$.

Převod na aditivitu:

$$\log(\boldsymbol{\mu} * \boldsymbol{\alpha_i} * \boldsymbol{\beta_j} * \boldsymbol{E_{ijt}}) = \log(\boldsymbol{\mu}) + \log(\boldsymbol{\alpha_i}) + \log(\boldsymbol{\beta_j}) + \log(\boldsymbol{E_{ijt}})$$

Při použití posunutí **+c** platí aditivita jen přibližně.

Pro výsledný tvar rozdělení hodnot je jedno, zda použijeme přirozený *LN* nebo dekadický *LOG10* či jiný. Z praktického hlediska je výhodný dekadický logaritmus, protože *log*(10) = 1, *log*(100) = 2, atd.

# Transformace dat: Logaritmická transformace

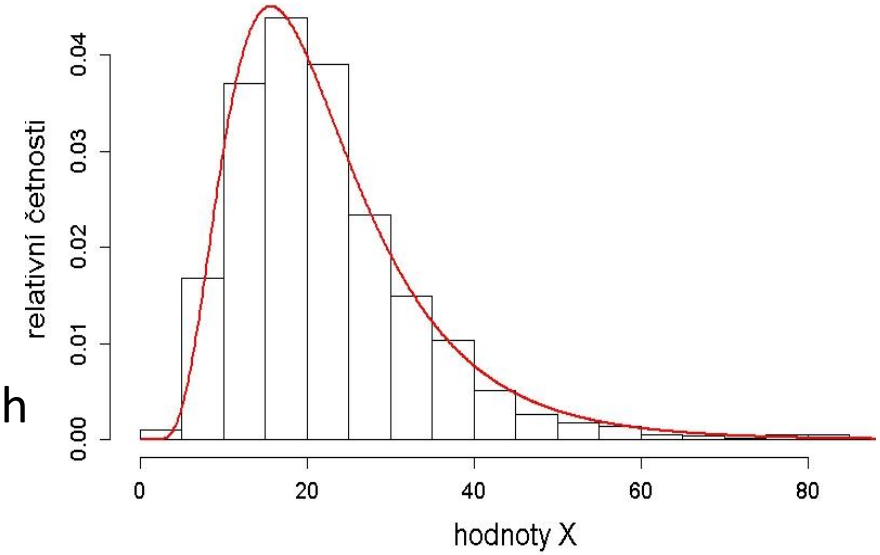
$$Y_i = \log(X_i)$$

**R:** `log(x)` … přirozený logaritmus
`log10(x)` … dekadický logaritmus

Připomenutí: když $\log(X) \sim N(\mu, \sigma^2)$, pak $X \sim LN(\mu, \sigma^2)$, logaritmicko-normální r.

Tři v jednom:

- odstraní násobný vztah faktorů
- odstraní závislost $\mu$ a $\sigma$
- přiblíží rozdělení dat normálnímu rozd.




Průměr a konfidenční interval:

Mohu použít hodnoty ze zlogaritmovaných výpočtů, které zpátky „odlogaritmuji“
$e^Y \rightarrow$ původní škála X hodnot
Průměr ale bude trochu vychýlený a konfidenční interval asymetrický (což je v pořádku, původní rozdělení dat je také asymetrické).

# Transformace dat

## Arcsinová transformace, také angulární

$$Y_i = arcsin\sqrt{p_i}$$

R: `asin(X)`

Pro data o podílech či procentech.

Doporučeno pro $p < 0.3$ a $p > 0.7$.

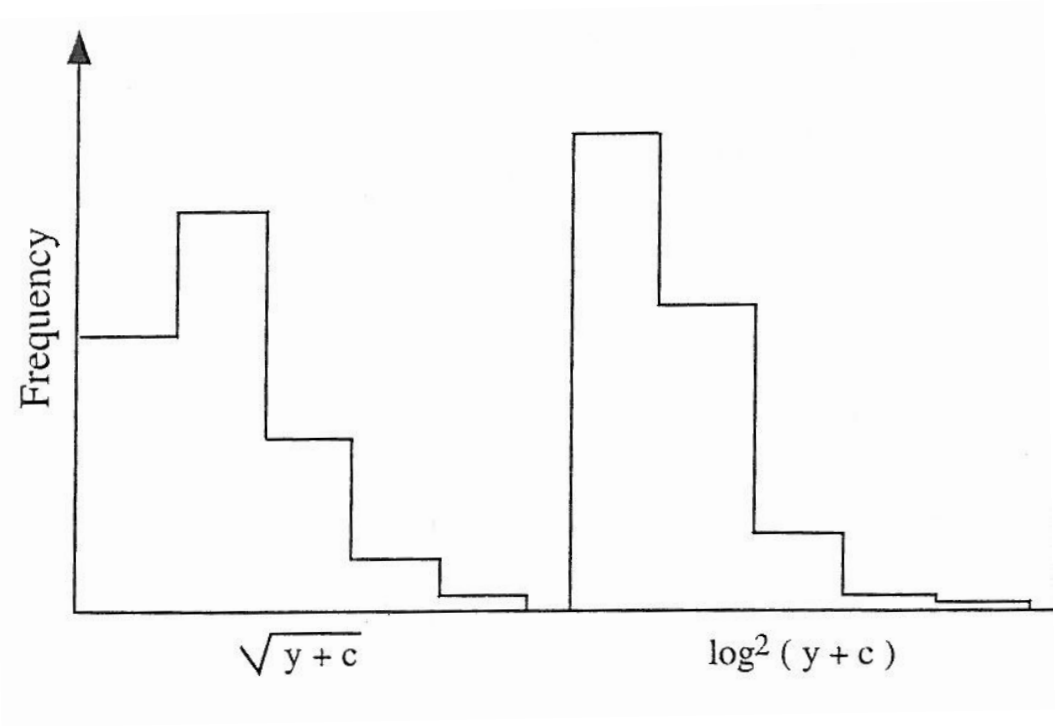
Jsou-li hodnoty $p \in \langle 0.3,\ 0.7 \rangle$, není třeba transformovat, rozdělení bude dostatečné blízké normálnímu rozdělení.

# Transformace dat

## Odmocninová transformace

$\boldsymbol{Y = \sqrt{X}}$, také $\boldsymbol{Y = \sqrt{X + 0.5}}$ pokud data obsahují nuly

Typicky pro data odpovídající Poissonovu rozdělení (funguje i log transform.), tedy počty jedinců v objemové či časové jednotce, pokud jsou rozmístěni náhodně.



# Transformace – příklad:

data: Potoky