

O dětech, čápech a kauzalitě

Jiří Dvořák

Katedra pravděpodobnosti a matematické statistiky, MFF UK

Mikulášský den ČStS
7. 12. 2021

Lidová moudrost (hlasovací otázka)

Jaký je Váš názor na začátku přednášky?

Lidová moudrost (hlasovací otázka)

Jaký je Váš názor na začátku přednášky?

- A) Čápi nosí děti,
- B) děti nosí čápy,
- C) čápi a děti spolu souvisí jinak,
- D) čápi a děti spolu nesouvisí.

Jak rozhodnout?

O vztahu dětí a čápů může zasvěceně hovořit

- ornitolog
- neonatolog
- pediatr
- ...

Jak rozhodnout?

O vztahu dětí a čápů může zasvěceně hovořit

- ornitolog
- neonatolog
- pediatr
- ...
- ...
- ...
- statistik

Empirické údaje

Stát	Čápi [páry]	Porodnost [tis./rok]
Albánie	100	83
Belgie	1	118
Bulharsko	5000	117
Dánsko	9	59
Francie	140	774
Itálie	5	551
Maďarsko	5000	124
Německo	3300	901
Nizozemsko	4	188
Polsko	30 000	610
Portugalsko	1500	120
Rakousko	300	87
Rumunsko	5000	367
Řecko	2500	106
Španělsko	8000	439
Švýcarsko	150	82
Turecko	25 000	1576

Údaje převzaty z článku

Matthews, R. (2000) *Storks Deliver Babies* ($p=0.008$). *Teaching Statistics* 22 (2), 36–38.

Údaje převzaty z článku

Matthews, R. (2000) *Storks Deliver Babies* ($p=0.008$). *Teaching Statistics* 22 (2), 36–38.

Podrobně ověřeny, viz článek

Dvořák, J. (2017) *O dětech, čápech a kauzalitě*. *Pokroky matematiky, fyziky a astronomie* 62 (4), 264–274.

Údaje převzaty z článku

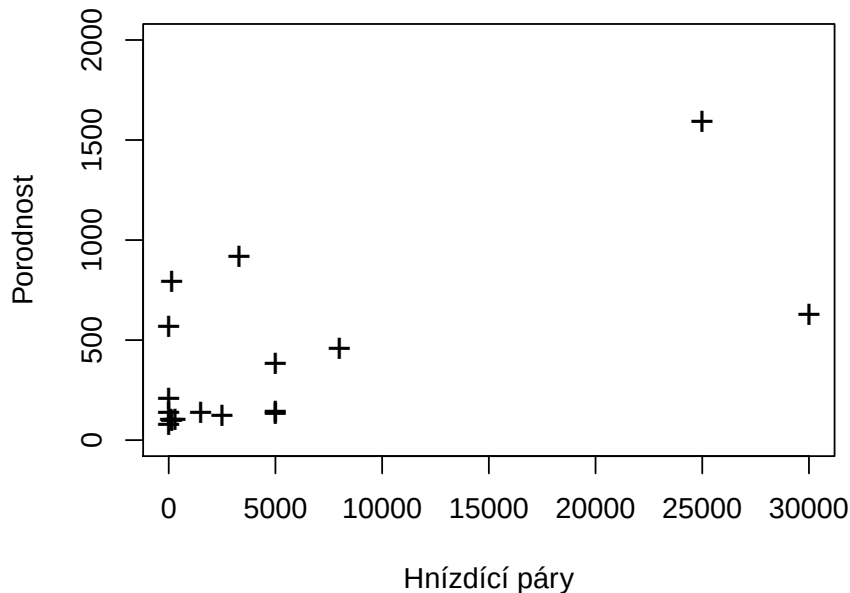
Matthews, R. (2000) *Storks Deliver Babies* ($p=0.008$). *Teaching Statistics* 22 (2), 36–38.

Podrobně ověřeny, viz článek

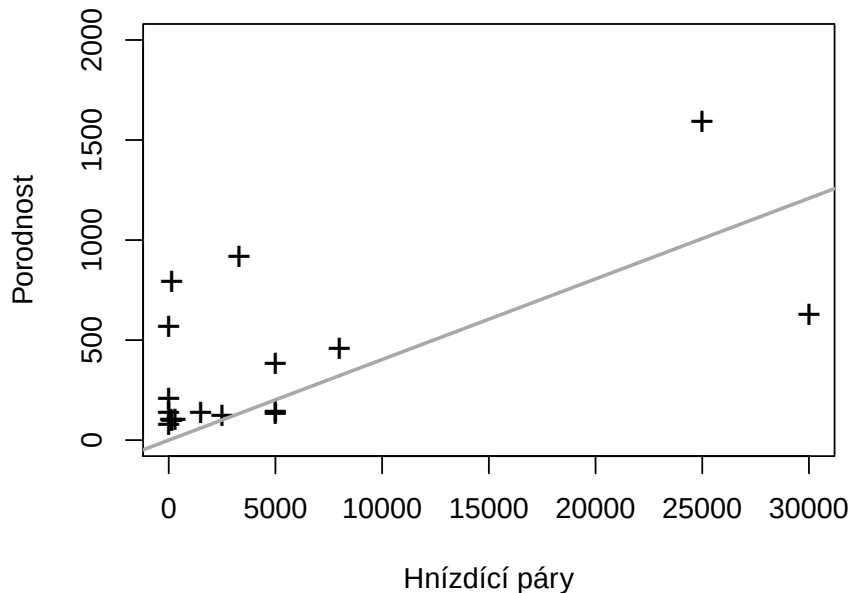
Dvořák, J. (2017) *O dětech, čápech a kauzalitě*. *Pokroky matematiky, fyziky a astronomie* 62 (4), 264–274.

Doplnění údajů pro ČR (cca 96 000 novorozeňat ročně, cca 800 hnízdících párů čápů) výsledky analýzy zásadně neovlivní.

Empirické údaje



Proložíme přímku

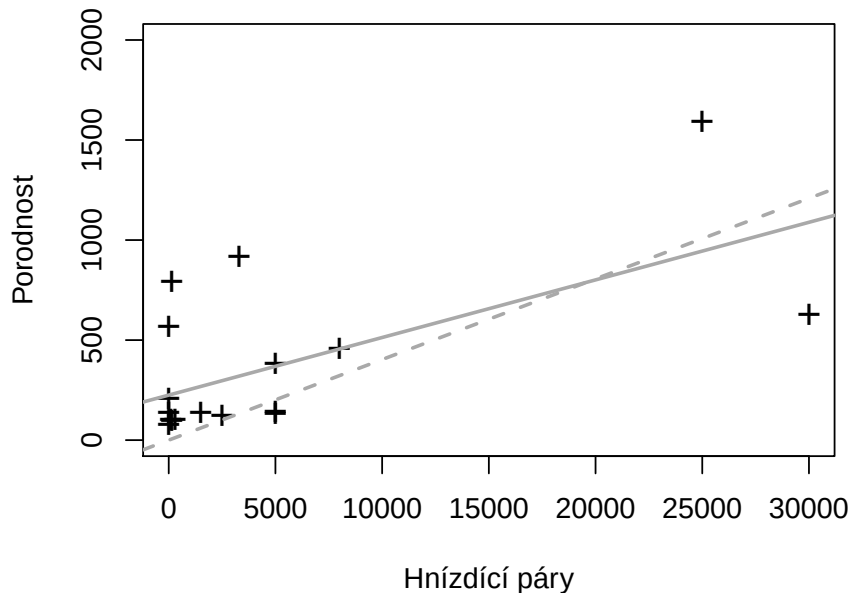


Proložená přímka nevystihuje data dobře, zejména blízko 0.

Odhadnutá hodnota je $\hat{b} \doteq 0,04$.

1 pár čápů $\sim$ 40 dětí ročně (porodnost vyjádřena v tisících).

Proložíme přímku II



Interpretace II

Nová přímka vystihuje data lépe.

Odhadnuté hodnoty jsou $\hat{a} \doteq 225$, $\hat{b} \doteq 0,03$.

1 pár čápů $\sim$ 30 dětí ročně (porodnost vyjádřena v tisících),
225 000 dětí ročně i bez čápů.

Nová přímka vystihuje data lépe.

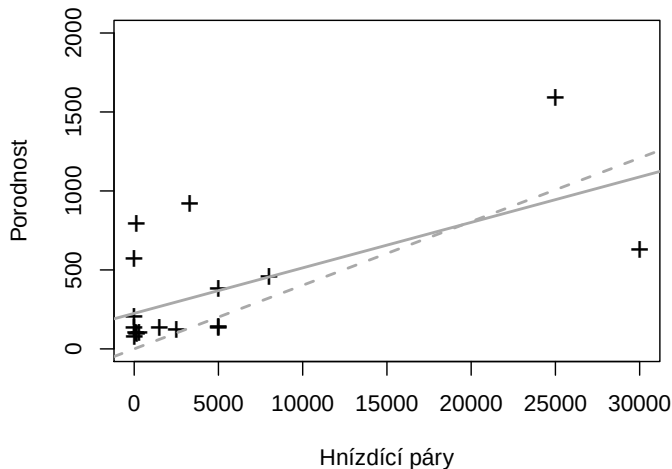
Odhadnuté hodnoty jsou $\hat{a} \doteq 225, \hat{b} \doteq 0,03$.

1 pár čápů $\sim$ 30 dětí ročně (porodnost vyjádřena v tisících),
225 000 dětí ročně i bez čápů.

Odchyłky od vzorce: náhodné chyby, vliv případných dalších faktorů.

Pokud čápi s dětmi nesouvisí, očekáváme vodorovnou přímku s $\hat{b} \doteq 0$.

Statistická významnost



$\hat{\rho} \doteq 0,62$ a *statisticky významně* se liší od 0 s *p*-hodnotou 0,008.

Ne každá (statisticky) významná korelace je (věcně) důležitá!

Změny hodnot jedné veličiny nemusí způsobovat/vysvětlovat změny hodnot druhé veličiny, nemusí být přítomen kauzální vztah.

Počet titulů Ph.D. udělených v matematických oborech ve Spojených státech:

- kladně koreluje s množstvím uranu uskladněným v jaderných elektrárnách tamtéž ($\hat{\rho} > 0,95$),¹
- záporně koreluje s roční spotřebou plnotučného mléka na jednoho obyvatele USA ($\hat{\rho} < -0,94$).²

¹http://tylervigen.com/view_correlation?id=1100

²http://tylervigen.com/view_correlation?id=1103

Počet filmů (za rok), ve kterých se objevil Nicolas Cage:

- kladně koreluje s počtem lidí (opět za rok), kteří utonuli po pádu do bazénu ($\hat{\rho} > 0,66$),³
- negativně koreluje s počtem lidí, kteří utonuli po pádu z rybářské lodi ($\hat{\rho} < -0,54$).⁴

³http://tylervigen.com/view_correlation?id=359

⁴http://tylervigen.com/view_correlation?id=10037

Varovné příklady II.

- Čím více hasičů dorazí k požáru, tím větší je celková škoda.

- Čím více hasičů dorazí k požáru, tím větší je celková škoda.
- U žáků prvního stupně ZŠ je kladná korelace mezi úrovní čtenářských dovedností a velikostí bot.

Varovné příklady II.

- Čím více hasičů dorazí k požáru, tím větší je celková škoda.
- U žáků prvního stupně ZŠ je kladná korelace mezi úrovní čtenářských dovedností a velikostí bot.
- Velikost dlaně je záporně korelovaná s očekávanou délkou života dané osoby.

Varovné příklady I. a II. mají rozdílnou povahu:

- I. falešné korelace (náhoda, vybíráme z obrovského množství dat),
- II. vliv jiné veličiny.

Pozorovaná korelace může být způsobena:

- přímým kauzálním vztahem,
- vlivem jiné veličiny (zprostředkovaný vztah),
- náhodou (nepřítomnost kauzálního vztahu).

Pozorovaná korelace může být způsobena:

- přímým kauzálním vztahem,
- vlivem jiné veličiny (zprostředkovaný vztah),
- náhodou (nepřítomnost kauzálního vztahu).

Hodnota korelačního koeficientu nedokáže prokázat (ne)přítomnost kauzálního vztahu.

Čápi a děti (hlasovací otázka)

Jak vysvětlit silnou kladnou korelaci mezi počtem čápů a počtem novorozenců?

- A) přímým kauzálním vztahem,
- B) vlivem jiné veličiny (*zprostředkovaný vztah*),
- C) náhodou (*nepřítomnost kauzálního vztahu*),
- D) žádná z předchozích odpovědí není správná.

Empirické údaje

Stát	Rozloha [km ²]	Čápi [páry]	Porodnost [tis./rok]
Albánie	28 750	100	83
Belgie	30 520	1	118
Bulharsko	111 000	5000	117
Dánsko	43 100	9	59
Francie	544 000	140	774
Itálie	301 280	5	551
Maďarsko	93 000	5000	124
Německo	357 000	3300	901
Nizozemsko	41 900	4	188
Polsko	312 680	30 000	610
Portugalsko	92 390	1500	120
Rakousko	83 860	300	87
Rumunsko	237 500	5000	367
Řecko	132 000	2500	106
Španělsko	504 750	8000	439
Švýcarsko	41 290	150	82
Turecko	779 450	25 000	1576

Závěr analýzy

Nalezenou korelaci mezi porodností a počtem hnízdících párů čápů *je možné vysvětlit* vlivem dalšího faktoru, který působí na obě tyto veličiny. Tímto faktorem je rozloha daného státu.

Závěr analýzy

Nalezenou korelaci mezi porodností a počtem hnízdících párů čápů *je možné vysvětlit* vlivem dalšího faktoru, který působí na obě tyto veličiny. Tímto faktorem je rozloha daného státu.

Závěr je velmi opatrný. Neříkáme, že to tak určitě je, neříkáme, že korelace není pouze falešná, neříkáme, že není přítomen přímý kauzální vztah.

Použité metody nám tak silnou informaci nedávají, pouze jsme našli možný způsob, jak korelaci vysvětlit.

Závěr analýzy

Nalezenou korelaci mezi porodností a počtem hnízdících párů čápů *je možné vysvětlit* vlivem dalšího faktoru, který působí na obě tyto veličiny. Tímto faktorem je rozloha daného státu.

Závěr je velmi opatrný. Neříkáme, že to tak určitě je, neříkáme, že korelace není pouze falešná, neříkáme, že není přítomen přímý kauzální vztah.

Použité metody nám tak silnou informaci nedávají, pouze jsme našli možný způsob, jak korelaci vysvětlit.

O tom, zda skutečně čápi nosí děti (o kauzálním vztahu), jsme se bohužel nedozvěděli nic.