Jiří Dvořák

O dětech, čápech a kauzalitě

Pokroky matematiky, fyziky a astronomie, Vol. 62 (2017), No. 4, 264–274

Persistent URL: `http://dml.cz/dmlcz/147069`

O dětech, čápech a kauzalitě

Jiří Dvořák, Nehvizdy

Abstrakt. Lidová moudrost v mnoha zemích světa spojuje přirozený přírůstek obyvatelstva s výskytem čápů. V našich podmínkách jde o čápa bílého (lat. *Ciconia ciconia*) a tradiční formulace zní „čápi nosí děti“. Se vší úctou k moudrosti našich předků nemůžeme takové tvrzení přijmout jako fakt, aniž bychom takovou vazbu potvrdili experimentálně či statistickou analýzou dostupných údajů. Právě to bude obsahem tohoto článku — provedeme analýzu počtu hnízdících párů čápů v 17 evropských zemích ve vztahu k porodnosti. Během toho se dotkneme i témat jako korelace, kauzalita a statistická významnost. Abychom nenarušovali tok textu a úvah, ponecháváme vysvětlení potřebných statistických pojmů do závěrečné části článku. Čtenář, který pojem zná, se jím nemusí zdržovat; čtenář, který vysvětlení či připomenutí potřebuje, je naopak snadno najde.

1. Čápi a děti — co na to říká statistika?

V tomto článku se pokusíme posoudit platnost lidového tvrzení, že „čápi nosí děti“. Autor, nejsa odborníkem v oblasti ornitologie, fyziologie ani antropologie, nemá jinou možnost než uchýlit se k analýze dostupných empirických údajů a pomocí jednoduchých statistických metod podat argumenty pro či proti tomuto tvrzení.

Inspirací je článek [5], který se touto otázkou zabýval a který také poskytuje potřebná data, viz tabulku 1. Jako zdroj dat o počtech čápů (v období cca 1980–1990) uvádí článek [5] „personal communication“ se členem *Royal Society for the Protection of Birds*, jako zdroj geografických a demografických údajů pak uvádí *Britannica Yearbook for 1990*.

Údaje o počtu obyvatel a porodnosti v evropských zemích jsou pečlivě sledované a nemíníme je zpochybňovat, stejně jako údaje o rozloze států. Údaje o počtu čápů jsme ověřili v odborné literatuře – souhrnné informace stejně jako odkazy na dílčí studie, odkud jsou jednotlivé údaje převzaty, poskytuje práce [7]. Hodnoty v ní uvedené jsou o něco novější, typicky z let 1993 až 1995, a proto částečně odlišné od údajů z [5]. Tyto rozdíly však nepovažujeme za podstatné a naši analýzu založíme na datech z původního článku [5].

Standardním postupem zjišťování velikosti populace čápů na daném území je počítání hnízdících párů. Obrázek 1 ukazuje hodnoty porodnosti v jednotlivých státech (v tisících dětí za rok) vykreslené právě proti počtu hnízdících párů čápů. Pohled na obrázek naznačuje přítomnost vazby mezi oběma veličinami. Povaha této vazby se dá popsat přirozeným jazykem jako „čím víc čápů, tím víc dětí“.

Jednoduchým nástrojem k vyjádření síly takového vztahu je tzv. *korelační koeficient*, který udává sílu lineárního vztahu mezi dvěma náhodnými veličinami, viz odst. 6.4. Pokud odhadneme korelační koeficient $\rho_{X,Y}$ pro uvažovanou dvojici náhod-

RNDr. Jiří Dvořák, Ph.D., Katedra pravděpodobnosti a matematické statistiky MFF UK, Sokolovská 83, 186 75 Praha 8, e-mail: `dvorak@karlin.mff.cuni.cz`

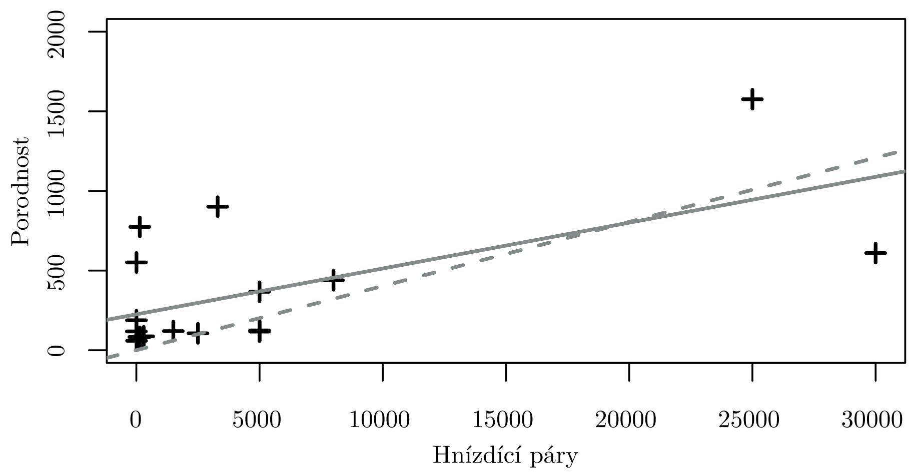
Stát	Rozloha (km^2)	Čápi (páry)	Obyvatelstvo (mil.)	Porodnost (tis./rok)
Albánie	28 750	100	3,2	83
Belgie	30 520	1	9,9	118
Bulharsko	111 000	5 000	9,0	117
Dánsko	43 100	9	5,1	59
Francie	544 000	140	56,0	774
Itálie	301 280	5	57,0	551
Maďarsko	93 000	5 000	11,0	124
Německo	357 000	3 300	78,0	901
Nizozemsko	41 900	4	15,0	188
Polsko	312 680	30 000	38,0	610
Portugalsko	92 390	1 500	10,0	120
Rakousko	83 860	300	7,6	87
Rumunsko	237 500	5 000	23,0	367
Řecko	132 000	2 500	10,0	106
Španělsko	504 750	8 000	39,0	439
Švýcarsko	41 290	150	6,7	82
Turecko	779 450	25 000	56,0	1 576

Tab. 1. Počty pozorovaných hnízdících párů čápů bílých (lat. *Ciconia ciconia*) v 17 evropských zemích a odpovídající geografické (rozloha státu) a demografické informace (počet obyvatel a počet narozených dětí za rok). Převzato z článku [5].

ných veličin (X = počet hnízdících párů čápů; Y = počet narozených dětí) na základě hodnot z tabulky 1, dostaneme hodnotu přibližně 0,62. To potvrzuje, že mezi X a Y je skutečně vazba typu „čím víc, tím víc“. Navíc je tato vazba poměrně silná — pokud by žádná vazba nebyla přítomna a náhodné veličiny byly nezávislé, byl by korelační koeficient nulový. Odhadnutá hodnota je tedy blíže extrémní hodnotě 1 (perfektní lineární závislost) než hodnotě 0 (žádná vazba mezi X a Y).

Samotný odhad korelačního koeficientu nám ale neříká, jak silný důkaz přítomnosti vazby mezi náhodnými veličinami X a Y jsme dostali. Co když jsme náhodou naměřili netypické hodnoty? Mohlo by se stát, že i v situaci, kdy žádná vazba přítomna není, obdržíme takový či ještě vyšší odhad korelačního koeficientu?

Odpověď nám může dát formální *statistický test*, viz odst. 6.5. V našem případě půjde o test nulovosti korelačního koeficientu. Nulovou hypotézou tedy je $\rho_{X,Y} = 0$, alternativní hypotézou je $\rho_{X,Y} \neq 0$. Test na hladině významnosti 0,05 (tedy 5 %) bude v našem případě zamítat nulovou hypotézu. P-hodnota testu je přibližně 0,008 a test by zamítal i na libovolné hladině vyšší než 0,008, tedy 0,8 %. Uvedenou p-hodnotu můžeme interpretovat tak, že pokud platí nulová hypotéza ($\rho_{X,Y} = 0$), je v tomto experimentu pravděpodobnost zjištění korelačního koeficientu, který bude v absolutní hodnotě větší než 0,62, pouze 0,8 %.



Obr. 1. Porodnost v jednotlivých státech (v tisících dětí za rok) vykreslená proti počtu hnízdících párů čápů. Plná čára ukazuje proloženou regresní přímku s absolutním členem. Čárkovaná čára ukazuje proloženou regresní přímku bez absolutního členu.

Korelační koeficient je tedy statisticky významně odlišný od nuly a zjištěné důkazy jsou velmi silné. Můžeme se tedy pokusit popsat vztah mezi X a Y lineárním předpisem. K tomu nám poslouží *lineární regrese*, viz odst. 6.6. Konkrétně budeme odhadovat regresní přímku, tedy přímku, která nejlépe vysvětluje Y jako lineární funkci X.

Nejprve uvažujme model bez absolutního členu: $Y = bX$. V tomto případě odhadneme parametr jako $b = 0{,}04$. Tedy jeden hnízdící pár čápů odpovídá v průměru 40 narozeným dětem ročně — připomeňme, že hodnoty veličiny Y, počty dětí narozených za rok, byly uvedeny v tisících. Tomuto odhadnutému modelu odpovídá čárkovaná čára na obrázku 1. Je však vidět, že pozorované hodnoty nevystihuje dobře.

Zaměřme se tedy na model s absolutním členem: $Y = a + bX$. V tomto modelu odhadneme parametry jako $a = 225$, $b = 0{,}03$. Tomuto modelu odpovídá plná čára na obrázku 1. Tentokrát jsou již pozorované hodnoty proloženy uspokojivě. Odhadnutý model můžeme interpretovat tak, že průměrně 225 tisíc dětí ročně se rodí bez souvislosti s čápy, nadto každý pár čápů odpovídá v průměru 30 narozeným dětem ročně. Pozorované odchylky od těchto hodnot pak považujeme za náhodné chyby a vliv případných dalších, neuvažovaných faktorů.

2. Vyhodnocení statistické analýzy

Můžeme tedy po provedení této statistické analýzy učinit závěr, že čápi nosí děti? Nemůžeme. Udělali jsme snad nějakou chybu ve výpočtu nebo v úvaze? Nikoliv, všechny výše uvedené postupy a úvahy jsou korektní. Předchozí tvrzení se zdají v rozporu — kde je tedy zakopaný pes?

Jediný problém je v tom, že jsme se snažili učinit závěry o něčem, o čem použité statistické nástroje vůbec nevypovídají. Korektním závěrem předchozí analýzy je pouze to, že *existuje statisticky významná vazba mezi počtem narozených dětí a počtem čápů.*

Použité metody mohou odhalit takovou vazbu, nejsou však schopné odlišit příčinu a následek nebo poznat situaci, kdy statisticky významná korelace neodpovídá žádné věcné vazbě. Je tedy stejně dobře možné, že „čápi nosí děti“ jako „děti nosí čápy“, případně „čápi a děti spolu nijak věcně nesouvisí“. Odhalili jsme tedy významnou korelaci, nikoliv kauzalitu. Tyto dva pojmy nesmíme zaměňovat: to, že spolu dvě veličiny souvisí (vykazují korelaci), ještě neznamená, že jedna je příčinou druhé.

3. Odbočka: korelace vs. kauzalita

Pro korelaci mezi dvěma náhodnými veličinami může být několik různých vysvětlení. Jedna z nich může skutečně ovlivňovat druhou, například za větší balení čokolády zaplatíme více peněz. Zde je kauzální vztah jasný a způsobuje kladnou korelaci.

Také mohou být obě náhodné veličiny X a Y současně ovlivňovány jinou náhodnou veličinou Z, například u žáků prvního stupně základní školy je vysoká kladná korelace mezi úrovní čtenářských dovedností a velikostí bot. Přenecháváme laskavému čtenáři k vlastní úvaze, co je oním společným vysvětlujícím faktorem. Zde jsou také kauzální vztahy jasné — Z ovlivňuje X a jsou spolu korelované, Z ovlivňuje Y a jsou spolu korelované, vliv Z způsobuje korelaci mezi X a Y, přestože mezi těmito veličinami není žádná přímá věcná vazba.

Uveďme pro ilustraci několik dalších příkladů. Všechny patří ke statistickému folklóru a uvádíme je bez reference, neboť je obtížné dohledat původní zdroj. Totéž platí pro příklad v předchozím odstavci. Je jistě pravda, že čím více hasičských jednotek je vysláno k požáru, tím větší je způsobená škoda; méně zřejmé je, proč velikost dlaně vykazuje zápornou korelaci s očekávanou délkou života dané osoby. Zde nabízíme nápovědu: ženy mají obvykle menší dlaně než muži, žijí však v průměru déle. Závěrem pak ještě dodejme, že počet utonutí v jednotlivých měsících silně koreluje s celkovými prodeji zmrzlinářských výrobků.

Poslední možností je, že mezi náhodnými veličinami není žádná věcná vazba, ať už přímá nebo zprostředkovaná. Potom hovoříme o tzv. falešné korelaci. Na tu často narazíme v situacích, kdy je k dispozici velké množství měření různých veličin a testujeme, které z nich jsou spolu významně korelované. V takovém případě provádíme velké množství dílčích testů (se zvolenou hladinou α). I pokud jsou všechny korelace ve skutečnosti nulové, je pravděpodobnost, že alespoň jeden z dílčích testů ukáže významnou korelaci, často výrazně vyšší než α. Jinými slovy, když provádíme mnoho dílčích testů, máme velkou šanci, že „něco vyjde významně“. Tuto situaci označujeme jako problém mnohonásobného porovnávání a musíme zde poznamenat, že by bylo hrubou chybou v takovém případě reportovat jen nalezené významné korelace bez informace o tom, kolik dílčích testů bylo ve skutečnosti provedeno, resp. bez provedení korekce na mnohonásobné porovnávání, viz například [4], [6].

Odhalování falešných korelací je věnován populární projekt Spurious Correlations [8], který umožňuje hledat korelace v rozsáhlé sadě veřejně dostupných údajů. Tím, že umožňuje snadno vyhledávat co nejsilnější korelace, ať už kladné či záporné, současně ilustruje problém mnohonásobného porovnávání — v takto rozsáhlé sadě údajů jednoduše musí být některé korelace velmi silné.

Pro doplnění si uveďme alespoň několik příkladů falešných korelací nalezených pomocí [8]. Počet titulů Ph.D. udělených v matematických oborech ve Spojených státech kladně koreluje s množstvím uranu uskladněným v jaderných elektrárnách tamtéž (odhadnutý korelační koeficient $> 0{,}95$), avšak záporně koreluje s roční spotřebou plnotučného mléka na jednoho obyvatele USA ($< -0{,}94$). Dále, počet filmů (za rok), ve kterých se objevil Nicolas Cage, kladně koreluje s počtem lidí (opět za rok), kteří utonuli po pádu do bazénu ($> 0{,}66$), ale negativně koreluje s počtem lidí, kteří utonuli po pádu z rybářské lodi ($< -0{,}54$).

4. Zpět k čápům a dětem

Výše jsme uvedli, že naše analýza ukazuje statisticky významnou korelaci mezi počtem narozených dětí a počtem hnízdících párů čápů, s upozorněním, že příčinné vztahy zůstávají nejasné. Ve světle diskuse o korelaci a kauzalitě, kterou jsme právě provedli, můžeme na základě dostupných dat rozhodnout, zda skutečně mezi uvažovanými veličinami existuje kauzální vztah, zda je nalezená korelace falešná nebo zda je způsobena nějakým vnějším faktorem, který ovlivňuje obě veličiny?

Při pohledu na údaje dostupné v tabulce 1 můžeme soudit, že rozloha státu i počet obyvatel mohou ovlivňovat obě zkoumané veličiny. Vyšší počet obyvatel s sebou typicky nese vyšší porodnost a může mít vliv i na počet hnízdících párů čápů. Ti totiž pro vybudování hnízda nejčastěji využívají lidmi vytvořených konstrukcí jako jsou vysoké komíny. Větší rozloha státu v evropských podmínkách obvykle znamená vyšší počet obyvatel a tedy vyšší porodnost, zároveň více prostoru umožňuje „ubytovat“ více čápů bez zvýšení konkurence o zdroje potravy a vhodná místa k budování hnízd.

Vhodným nástrojem, jak posoudit vliv takových vnějších faktorů, je tzv. *parciální korelační koeficient*, viz odst. 6.7. Ten umožňuje kvantifikovat sílu lineárního vztahu mezi dvěma náhodnými veličinami po odstranění (také lineárního) vlivu jedné či více dalších náhodných veličin.

Pokud odstraníme vliv počtu obyvatel, dostaneme odhad parciálního korelačního koeficientu přibližně 0,65; test významnosti této korelace dává p-hodnotu přibližně 0,006 a na hladině 0,05 zamítáme nulovou hypotézu o nulovosti tohoto korelačního koeficientu. Tento faktor tedy nepomohl vysvětlit přítomnost korelace mezi porodností a počtem hnízdících párů čápů.

Naopak, pokud odstraníme vliv rozlohy státu, dostaneme odhad parciálního korelačního koeficientu přibližně 0,27; test významnosti této korelace dává p-hodnotu přibližně 0,307 a na hladině 0,05 tedy nezamítáme nulovou hypotézu o nulovosti tohoto korelačního koeficientu. Tento faktor tedy vysvětlil velkou část nalezené korelace a po odstranění jeho vlivu už zbylá korelace není statisticky významná.

5. Závěr

Můžeme tedy učinit závěr, že 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.

Pozornému čtenáři jistě neunikne opatrnost tohoto závěru. Říkáme pouze, že pozorovanou korelaci je možné vysvětlit určitým způsobem. 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. Pro odhalování kauzálních vztahů bychom potřebovali provádět kontrolované experimenty, v nichž je možné cíleně měnit hodnotu jedné veličiny a sledovat změny hodnot veličiny druhé, při zachování všech ostatních vlivů beze změny.

Pro zajímavost dodejme, že odpovídající údaje pro Českou republiku jsou následující: rozloha 78 866 km^2, počet obyvatel zhruba 10,3 milionu, přibližně 96 tisíc živě narozených dětí (tyto dva údaje převzaty z Českého statistického úřadu [3]), počet hnízdících párů čápů bílých cca 800 (převzato z článku [7]). Uvedené údaje platí pro rok 1995, protože pro tento rok máme k dispozici údaje o počtu čápů.

Pokud údaje platné pro ČR přidáme k tabulce 1 a zopakujeme celou analýzu, dostaneme prakticky totožné výsledky. Chování čápů v ČR tedy odpovídá jejich chování v ostatních částech Evropy.

Závěrem už jen zdůrazněme, že je tento článek míněn zejména jako upozornění na záludnosti interpretace výsledků statistické analýzy. I při použití jednoduchých, základních statistických metod zde hrozí riziko vyvození nepodložených závěrů. Výstupem naší analýzy je tedy tvrzení, že pozorovanou silnou korelaci mezi počtem narozených dětí a počtem hnízdících párů čápů je možné vysvětlit společným vlivem rozlohy daného státu. Původní otázku pátrající po přítomnosti kauzálního vztahu, tedy zda čápi nosí děti, však musíme nechat otevřenou.

6. Glosář statistických pojmů

V této kapitole nabízíme krátké vysvětlení statistických pojmů využívaných v tomto článku. Podrobnější informace je možné najít například v [1], [2].

6.1. Střední hodnota

Buď X reálná náhodná veličina nabývající diskrétních hodnot x_i s pravděpodobnostmi p_i. Střední hodnota takové náhodné veličiny je pak definována jako

$$\mathbb{E}X = \sum_i p_i x_i$$

a udává očekávanou (průměrnou) hodnotu náhodné veličiny. Pro spojité náhodné veličiny je možné definovat střední hodnotu podobně pomocí tzv. hustoty. V tomto článku předpokládáme, že střední hodnota uvažovaných náhodných veličin je konečná.

Pokud máme k dispozici nezávislá pozorování $X_1, X_2, \ldots, X_n$ náhodné veličiny X, nabízí se přirozený odhad střední hodnoty $\mathbb{E}X$ ve tvaru

$$\widehat{\mathbb{E}X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

6.2. Rozptyl

Buď X reálná náhodná veličina nabývající diskrétních hodnot x_i s pravděpodobnostmi p_i. Střední hodnotu označíme $\mathbb{E}X$. Rozptyl náhodné veličiny X je pak definován jako

$$\operatorname{var} X = \mathbb{E}\left(X - \mathbb{E}X\right)^2 = \sum_i p_i \left(x_i - \mathbb{E}X\right)^2$$

a udává rozptýlenost hodnot náhodné veličiny kolem její střední hodnoty. Pro spojité náhodné veličiny je rozptyl také definován pomocí první rovnosti výše. V tomto článku předpokládáme, že rozptyl uvažovaných náhodných veličin je konečný.

Pokud máme k dispozici nezávislá pozorování $X_1, X_2, \ldots, X_n$ náhodné veličiny X, nabízí se přirozený odhad rozptylu $\operatorname{var} X$ ve tvaru

$$\widehat{\operatorname{var} X} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \widehat{\mathbb{E}X} \right)^2 .$$

V praxi se obvykle pracuje s odhadem rozptylu, který místo dělení počtem pozorování n používá dělení hodnotou $n-1$. Takový odhad je z určitého hlediska výhodnější, pro naše úvahy to však není podstatné.

6.3. Kovariance

Buď (X, Y) reálný náhodný vektor nabývající diskrétních hodnot (x_i, y_i) s pravděpodobnostmi p_i. Střední hodnoty náhodných veličin X a Y označíme $\mathbb{E}X$, $\mathbb{E}Y$, jejich rozptyly označíme $\operatorname{var} X$, $\operatorname{var} Y$. Kovariance náhodných veličin X a Y je pak definována jako

$$\operatorname{cov}(X, Y) = \mathbb{E}\,(X - \mathbb{E}X)\,(Y - \mathbb{E}Y) = \sum_i p_i\,(x_i - \mathbb{E}X)\,(y_i - \mathbb{E}Y)$$

a udává, jak moc jsou hodnoty X ovlivněny hodnotami Y a naopak. Pro spojité náhodné vektory je kovariance také definována pomocí první rovnosti výše.

Pokud máme k dispozici nezávislá pozorování (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ náhodného vektoru (X, Y), nabízí se přirozený odhad kovariance $\operatorname{cov}(X, Y)$ ve tvaru

$$\widehat{\operatorname{cov}(X, Y)} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \widehat{\mathbb{E}X} \right) \left(Y_i - \widehat{\mathbb{E}Y} \right).$$

6.4. Korelační koeficient

Buď (X, Y) reálný náhodný vektor. Střední hodnoty náhodných veličin X a Y označíme $\mathbb{E}X$, $\mathbb{E}Y$, jejich rozptyly označíme $\operatorname{var} X$, $\operatorname{var} Y$ a jejich kovarianci $\operatorname{cov}(X, Y)$. Korelační koeficient náhodných veličin X a Y je pak definován jako

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var} X \ \operatorname{var} Y}}$$

a udává, jak moc jsou hodnoty X ovlivněny hodnotami Y a naopak, tentokrát po normalizaci pomocí rozptylu. To umožňuje porovnávat sílu vztahu mezi dvojicemi náhodných veličin bez ohledu na případné rozdíly jejich rozptylu. Důsledkem Cauchyovy–Schwarzovy nerovnosti je, že korelační koeficient nabývá pouze reálných hodnot z intervalu $[-1, 1]$.

Korelační koeficient $\rho_{X,Y}$ udává (stejně jako kovariance) sílu lineárního vztahu mezi náhodnými veličinami X a Y. Extrémních hodnot 1 a -1 nabývá korelační koeficient právě tehdy, když je mezi náhodnými veličinami lineární vztah $Y = aX + b$ pro nějaké reálné hodnoty $a \neq 0$ a b. Pokud je v takovém případě $a > 0$, je $\rho_{X,Y} = 1$; pokud je $a < 0$, je $\rho_{X,Y} = -1$. Hodnota korelačního koeficientu 0 odpovídá nepřítomnosti lineárního vztahu. Příkladem takové situace je nezávislost veličin X a Y — v takovém případě se tyto náhodné veličiny vůbec neovlivňují.

Pokud máme k dispozici nezávislá pozorování (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$, nabízí se přirozený odhad korelačního koeficientu ve tvaru

$$\widehat{\rho}_{X,Y} = \frac{\widehat{\operatorname{cov}(X, Y)}}{\sqrt{\widehat{\operatorname{var} X}\,\widehat{\operatorname{var} Y}}}.$$

Ilustrujme nyní skutečnost, že korelační koeficient měří sílu *lineárního* vztahu mezi X a Y. Nechť pro tyto náhodné veličiny platí $Y = 2X$ a máme k dispozici nezávislá pozorování $(1, 2), (2, 4), \ldots, (10, 20)$. Odhad korelačního koeficientu podle vzorce výše má pak hodnotu 1.

Dále uvažme vztah mezi náhodnými veličinami ve tvaru $Y = X^3$. Pokud máme k dispozici nezávislá pozorování $(1, 1), (2, 8), \ldots, (10, 1\,000)$, dostaneme podle vzorce výše odhad korelačního koeficientu přibližně 0,928. Přestože je tedy mezi náhodnými veličinami pořád naprosto pevná vazba (jsou svázány pomocí deterministické funkce), nedosahuje hodnota korelačního koeficientu extrémní hodnoty 1 či -1. Je možné dokonce najít příklady, kdy pro dvě náhodné veličiny svázané deterministickou funkcí vyjde odhad korelačního koeficientu 0 — například stačí uvažovat vztah $Y = X^2$ a pozorování $(-5, 25), (-4, 16), \ldots, (5, 25)$.

6.5. Testování hypotéz

Uvažme situaci, kdy chceme rozhodnout, zda je určité tvrzení v souladu s pozorovanými daty. Toto tvrzení budeme nazývat nulová hypotéza. Dále formulujeme tzv. alternativní hypotézu, které budeme věřit, pokud ukážeme, že nulová hypotéza neplatí. Často je alternativní hypotéza doplňkem nulové hypotézy, není to však nutné.

Ilustračním příkladem nám budiž házení mincí a testování, zda je mince spravedlivá. Nulovou hypotézou zde bude „pravděpodobnost, že padne panna, je rovna $1/2$*“. Alternativní hypotézou pak bude „pravděpodobnost, že padne panna, není rovna* $1/2$*“.*

Statistický test potom provedeme tak, že z pozorovaných dat spočítáme vhodnou testovou statistiku T, přičemž vyžadujeme, aby rozdělení testové statistiky za platnosti nulové hypotézy bylo známé. Pak posoudíme, zda hodnota T spočítaná z dat je typická, nebo naopak extrémní vzhledem k tomuto referenčnímu rozdělení. Pokud je extrémní, zamítáme nulovou hypotézu ve prospěch alternativní hypotézy. Pokud je typická, nezamítáme nulovou hypotézu. Extrémní hodnoty jsou takové, které leží v takzvané zamítací oblasti. Typické hodnoty naopak leží mimo zamítací oblast.

V našem příkladu může být pozorovanými daty výsledek dvaceti nezávislých hodů zkoumanou mincí. Testovou statistikou T *pak bude celkový počet hodů, ve kterých padne panna. Za platnosti nulové hypotézy má* T *binomické rozdělení s parametry* 20 *a* $1/2$.

Pokud je nulová hypotéza platná a my ji zamítneme, udělali jsme chybu prvního druhu. Pokud je nulová hypotéza neplatná a my ji nezamítneme, udělali jsme chybu druhého druhu. V ostatních případech jsme chybu neudělali. Statistické testy jsou konstruovány tak, aby pravděpodobnost, že uděláme chybu prvního druhu, byla nejvýše rovna předem stanovené hodnotě α, například $\alpha = 0{,}05$. Tato volba pak určuje, jak bude vypadat zamítací oblast pro daný test (přesný postup se pro různé testy liší). Hodnotu α označujeme jako *hladinu testu.*

V našem příkladu je pro $\alpha = 0{,}05$ zamítací oblast rovna množině $C = \{0, 1, 2, 3, 4, 5, 16, 17, 18, 19, 20\}$. Za platnosti nulové hypotézy je totiž pravděpodobnost, že T bude mít hodnotu z C, rovna přibližně 0,041. Kdybychom navíc zahrnuli do zamítací oblasti libovolnou hodnotu mezi 6 a 15, už bychom přesáhli hodnotu 0,05 a takto postavený test by neměl zvolenou hladinu.

Možností, jak zvolit zamítací oblast, která dodrží stanovenou hladinu testu, samozřejmě může být více. My chceme zvolit takovou, která nám dá co nejmenší pravděpodobnost chyby druhého druhu.

V příkladu s házením mincí chyba druhého druhu nastává, pokud se skutečná pravděpodobnost padnutí panny liší od $1/2$ a provedený test přesto nulovou hypotézu nezamítne. Pokud se skutečná pravděpodobnost liší od $1/2$, budou typické hodnoty testové statistiky blíže k extrémním hodnotám 0 a 20 než za platnosti nulové hypotézy. Proto v zamítací oblasti chceme mít právě tyto extrémní hodnoty a jejich okolí a zvolili jsme $C = \{0, 1, 2, 3, 4, 5, 16, 17, 18, 19, 20\}$. Hladinu testu by dodržela i jiná zamítací oblast, například $\widetilde{C} = \{6\}$, její použití by však vedlo ke zbytečně vysoké pravděpodobnosti chyby druhého druhu — vždyť v testu s touto zamítací oblastí bychom nezamítli nulovou hypotézu ani v případě, že ve dvaceti hodech neuvidíme žádnou pannu.

Sílu toho, jak pozorovaná data svědčí proti nulové hypotéze, můžeme vyjádřit pomocí tzv. p-hodnoty. To je pravděpodobnost, že bychom za platnosti nulové hypotézy napozorovali taková data, která by svědčila stejně či ještě více proti nulové hypotéze než naše skutečně pozorovaná data. Jinými slovy, pokud ze skutečně pozorovaných dat vypočteme hodnotu testové statistiky T_0, udává p-hodnota pravděpodobnost, že při opakování experimentu dojdeme k hodnotě testové statistiky T_1, která bude vzhledem k referenčnímu rozdělení stejně nebo ještě více extrémní než T_0. O zamítnutí/nezamítnutí nulové hypotézy je možné rozhodnout i pomocí p-hodnoty — zamítáme, právě když je p-hodnota menší nebo rovna hladině testu α. Tento postup je ekvivalentní postupu založenému na zamítací oblasti.

Pokud při 20 nezávislých hodech mincí padne pouze dvakrát panna, je $T_0 = 2$ a množina všech hodnot stejně či více extrémních vzhledem k referenčnímu rozdělení je $D = \{0, 1, 2, 18, 19, 20\}$. Proto bude p-hodnota našeho testu přibližně 0,0004 — platí, že při opakování experimentu je za platnosti nulové hypotézy pravděpodobnost 0,0004, že hodnota testové statistiky T_1 bude ležet v D. Vzhledem k tomu, že zjištěná p-hodnota je menší než zvolená hladina $\alpha = 0{,}05$, zamítáme nulovou hypotézu o tom, že pravděpodobnost padnutí panny je $1/2$.

Při analýze dat prezentované v tomto článku jsme použili test nulovosti korelačního koeficientu. Nulovou hypotézou tedy je $\rho_{X,Y} = 0$, alternativní hypotézou je $\rho_{X,Y} \neq 0$. Testovou statistikou je

$$T = \widehat{\rho}_{X,Y} \cdot \sqrt{(n-2)/(1-\widehat{\rho}^2_{X,Y})},$$

kde $\widehat{\rho}_{X,Y}$ je odhad korelačního koeficientu (s normalizací $1/(n-1)$ místo $1/n$ v odhadu kovariance a rozptylů) a $n \geq 3$ je počet pozorování. Za platnosti nulové hypotézy a předpokladu, že náhodný vektor (X, Y) má dvourozměrné normální rozdělení s kladnými rozptyly, má testová statistika t-rozdělení s $n-2$ stupni volnosti [2, s. 94].

6.6. Lineární regrese

Pro potřeby tohoto článku se omezíme na jednoduchý případ, kdy chceme hodnoty náhodné veličiny Y vysvětlit (lineárním modelem) pomocí hodnot veličiny X. V základním přístupu předpokládáme, že hodnoty X jsou nenáhodné, resp. jsou změřeny přesně, bez náhodné chyby. Jde vlastně o modelování střední hodnoty Y v závislosti na vysvětlující proměnné X. Pokud máme k dispozici pozorování $(X_1, Y_1), \dots, (X_n, Y_n)$, je odpovídající model tvaru $Y_i = a + bX_i + e_i$, $i = 1, \dots, n$. Na pravé straně této rovnosti je jediným náhodným členem e_i, které může popisovat například chybu měření a všechny další vlivy, které nejdou (lineárně) vysvětlit pomocí hodnoty X_i. Předpokládáme, že všechna e_i mají nulovou střední hodnotu a stejný rozptyl.

Pro své pozorování tedy hledáme hodnoty parametrů a, b v uvedeném modelu. Budeme požadovat, aby tyto hodnoty minimalizovaly součet čtverců chyb, tedy $\sum_{i=1}^{n}(Y_i - a - bX_i)^2$. Odhad střední hodnoty Y_i je potom $\widehat{Y}_i = a + bX_i$. Jednotlivým chybám $Y_i - \widehat{Y}_i = Y_i - a - bX_i$, $i = 1, \dots, n$, říkáme rezidua. Jde o odhad náhodného členu e_i v uvažovaném modelu, který už nejde vysvětlit lineárně pomocí hodnoty X_i. Přímce $y = a + bx$, kde a, b jsou odhadnuté hodnoty parametrů, potom říkáme *regresní přímka*.

Je snadné upravit tento postup pro model bez absolutního členu: $Y_i = bX_i + e_i$, $i = 1, \dots, n$. Obrázek 1 ukazuje odhadnuté regresní přímky pro data z tabulky 1 pro model s absolutním členem i bez absolutního členu.

6.7. Parciální korelační koeficient

Tento nástroj je obdobou klasického korelačního koeficientu, bere však do úvahy možnost, že jsou obě náhodné veličiny X, Y ovlivněny hodnotami společné vysvětlující proměnné Z. V takovém případě můžou X a Y vykazovat silnou korelaci, která je však způsobena vlivem Z, aniž by spolu X a Y přímo souvisely.

Parciální korelační koeficient tedy slouží k posouzení síly lineárního vztahu mezi dvěma náhodnými veličinami X a Y poté, co odstraníme případný (lineární) vliv třetí proměnné Z.

Nechť máme k dispozici pozorování $(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)$. Uvažujme nejprve lineární regresi, kde vysvětlíme hodnoty X_i pomocí hodnot Z_i. Dostaneme tedy rezidua $X_i - \widehat{X}_i$ (část, kterou už není možné vysvětlit lineárně pomocí hodnot Z_i). Podobně uvažujme lineární regresi, kde vysvětlíme hodnoty Y_i pomocí hodnot Z_i; dostaneme rezidua $Y_i - \widehat{Y}_i$.

Parciální korelační koeficient je pak klasický korelační koeficient mezi rezidui $X - \widehat{X}$ a $Y - \widehat{Y}$, tedy mezi částmi, které už nelze vysvětlit lineárně pomocí hodnot Z. Na základě pozorovaných dat je možné jej odhadnout jako $\widehat{\rho}_{X',Y'}$, přičemž pokládáme $X'_i = X_i - \widehat{X}_i$, $Y'_i = Y_i - \widehat{Y}_i$, $i = 1, \dots, n$.

L i t e r a t u r a

[1] Anděl, J.: *Statistické metody*. 4. vyd., MatfyzPress, Praha, 2007.

[2] Anděl, J.: *Základy matematické statistiky*. 3. vyd., MatfyzPress, Praha, 2011.

[3] Český statistický úřad: *Česká republika od roku 1989 v číslech — 2013* [online], [cit. 20. 11. 2017]. Dostupné z: `https://www.czso.cz/csu/czso/ceska-republika-v-cislech-od-roku-1989-wau52m1y38`

[4] Hsu, J. C.: *Multiple comparisons: theory and methods*. 1st ed., Chapman and Hall/CRC, Boca Raton, 1996.

[5] Matthews, R.: *Storks deliver babies (p=0.008)*. Teaching Statistics *22 (2)* (2000), 36–38.

[6] Miller, R. G.: *Simultaneous statistical inference*. 2nd ed., Springer, New York, 1981.

[7] Van den Bossche, W., Berthold, P., Kaatz, M., Nowak, E., Querner, U.: *Eastern European white stork populations: migration studies and elaboration of conservation measures*. Scripten 66. Bundesamt für Naturschutz, Bonn, 2002.

[8] Vigen, T.: *Spurious correlations* [online], [cit. 20. 11. 2017]. Dostupné z: `http://www.tylervigen.com`