4. Korelace

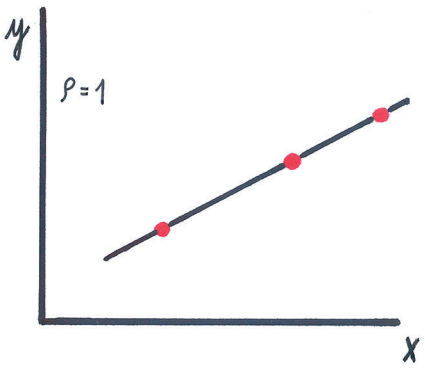
4.1 Korelace

Korelační (Pearsonův) koeficient je **míra lineárního vztahu mezi dvěma proměnnými**. Značí se obvykle *corr* (z anglického *correlation*) či ρ (ró). Existuje i Spearmanův korelační koeficient, ale vzhledem k tomu, že Pearsonův korelační koeficient se používá častěji, pracuji v této knize vždy pouze s Pearsonovým korelačním koeficientem. Korelační koeficient může nabývat hodnot **od −1 do 1**. Pozitivní koeficient znamená pozitivní vztah dvou proměnných (čím více jedné, tím více druhé) a negativní koeficient znamená negativní vztah (čím více jedné, tím méně druhé). Odhad korelačního koeficientu se nazývá výběrový korelační koeficient, značí se r a vypočítá se takto:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{s_x s_y}$$

$\bar{x}$ *je výběrový průměr* X
$\bar{y}$ *je výběrový průměr* Y
s_x *je výběrová směrodatná odchylka* X
s_y *je výběrová směrodatná odchylka* Y
n je počet pozorování
$x_1, x_2, \ldots x_n$ jsou naměřené hodnoty náhodné veličiny X
$y_1, y_2, \ldots y_n$ jsou naměřené hodnoty náhodné veličiny Y

Vzorec pro korelaci sice vypadá složitě, ale interpretace je jednoduchá – viz první odstavec této kapitoly.



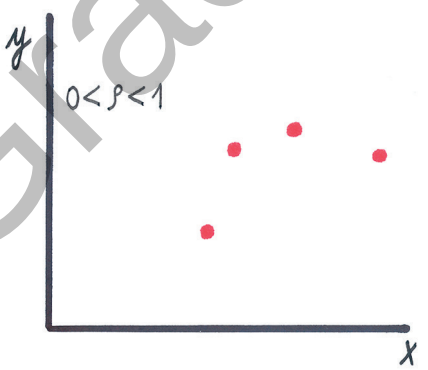
Pokud mezi dvěma proměnnými existuje **přesný pozitivní lineární vztah,** (všechny body leží na přímce s pozitivní směrnicí $y = ax + b$, $a > 0$) **pak je korelace = 1**. Říkáme, že mezi proměnnými existuje dokonalá pozitivní korelace. Tuto situaci můžeme zobrazit pomocí bodů ve dvourozměrném grafu – zde obrázek 4.1.

Obr. 4.1 *Korelace = 1.*

PŘÍKLAD 4.1

Představme si, že máme tři děti. Každé z nich vypilo určitý počet sklenic jablečného džusu. Tímto způsobem do sebe každé dítě dostalo určité množství cukru. Mezi počtem sklenic jablečného džusu a příjmem cukru existuje přesný lineární vztah: čím více sklenic, tím více cukru. Tuto situaci můžeme zobrazit graficky (obrázek 4.1). Každý červený bod zastupuje jedno dítě. Hodnota bodů na ose *x* zobrazuje počet vypitých sklenic džusu a hodnota bodů na ose y zobrazuje množství takto přijatého cukru. To, že body lze spojit rostoucí přímkou, odpovídá korelaci = 1.

Karl Pearson, podle kterého je pojmenován korelační (Pearsonův) koeficient, byl označován jako „vášnivý ateista". Vášnivost se očividně může projevovat různými způsoby.



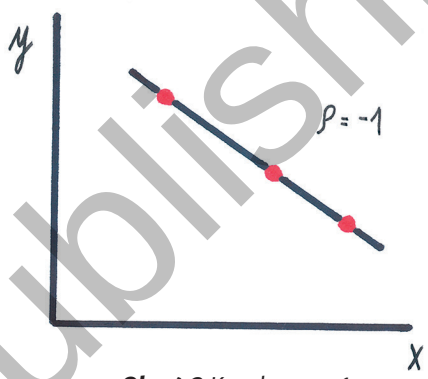
Pokud je mezi proměnnými **ne úplně přesný, ale stále pozitivní lineární vztah** (přibližně platí, že čím vyšší je *X*, tím vyšší je *Y* – viz obrázek 4.2), **korelační koeficient je v intervalu (0,1)**. Říkáme, že mezi proměnnými existuje pozitivní korelace (proměnné mezi sebou pozitivně korelují).

Obr. 4.2 *Hodnota korelace je v intervalu (0,1).*

PŘÍKLAD 4.2

Existuje pozitivní korelace ($0 < \rho < 1$) mezi stářím a počtem bílých vlasů. Ovšem není to přesná (dokonalá) korelace s hodnotou 1. Jistě totiž existují i staří lidé, kteří bílé vlasy nemají (například plešatí penzisté), a také mnoho mladých lidí, kteří mají bílých vlasů mnoho. Nejedná se o přesný lineární vztah (viz přímka na obrázku 4.1), ale o přibližný trend (viz obrázek 4.2). Na obrázku 4.2 by opět každý bod odpovídal jednomu člověku. Hodnota bodů na ose *x* by zde měřila věk a hodnota bodů na ose *y* by měřila počet bílých vlasů.

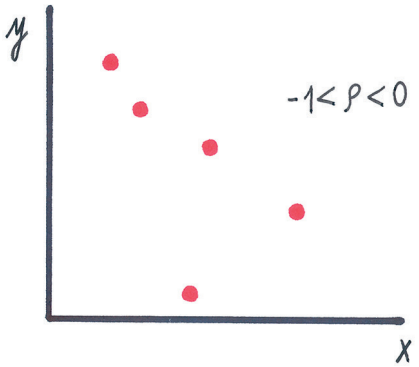
Pokud je mezi proměnnými **přesný negativní lineární vztah** (všechny body leží na přímce s negativní směrnicí $y = ax + b$, $a < 0$ – viz obrázek 4.3), **korelační koeficient je −1**. Říkáme, že mezi proměnnými existuje dokonalá negativní korelace.



Obr. 4.3 *Korelace = −1.*

PŘÍKLAD 4.3

Uvažujme situaci, kdy na stole je 100 jablek. Do místnosti přijde několik hladových Eskymáků a začnou jablka jíst. Každý eskymák sní tři jablka. Obrázek 4.3 ukazuje vztah mezi počtem příchozích eskymáků a počtem jablek, které zbydou po večeři na stole. Hodnota bodů na ose *x* ukazuje počet eskymáků a hodnota bodů na ose *y* ukazuje počet zbývajících jablek. Každý červený bod odpovídá odlišnému počtu eskymáků a tím pádem různým počtům zbývajících jablek.

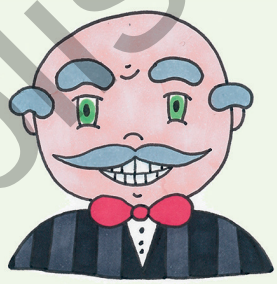


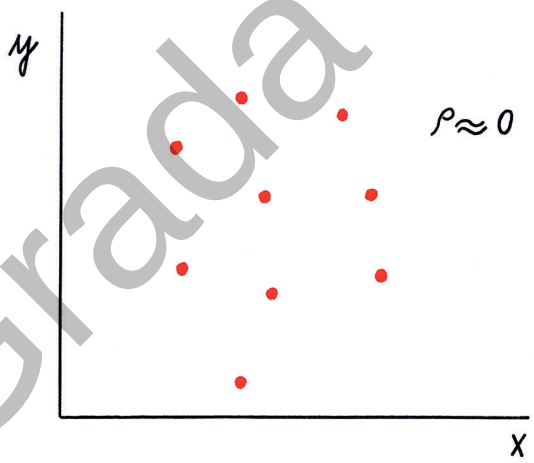
Pokud je mezi proměnnými **ne úplně přesný, ale stále negativní lineární vztah** (přibližně platí, že čím vyšší je X, tím nižší je Y – viz obrázek 4.4), **korelační koeficient je v intervalu (−1,0)**. Říkáme, že proměnné mezi sebou negativně korelují.

Obr. 4.4 *Hodnota korelace je v intervalu (−1,0).*

PŘÍKLAD 4.4

Ve věku od 20 do 100 let existuje negativní korelace mezi stářím a počtem zubů (čím více let, tím méně zubů). Ovšem není to přesná (dokonalá) korelace. Jistě totiž existují někteří staří lidé, kteří mají mnoho zubů (například penzisté mající vzornou dentální hygienu), a také jistě existují mladí lidé, kteří mají zubů málo (například někteří hokejisté). I zde korelační koeficient neukazuje přesnou závislost, ale pouze přibližný trend.





Pokud mezi proměnnými není žádný výrazný vztah (viz obrázek 4.5), **korelační koeficient je blízko** nuly. Říkáme, že proměnné spolu nekorelují.

Obr. 4.5 *Hodnota korelace je přibližně rovna nule.*

Před odletem jsem si v novinách přečetl, že pravděpodobnost, že se letadlo zřítí, je jedna ku jednomu miliónu. To mi připadalo jako příliš velká pravděpodobnost na to, abych letěl. Pak jsem se dal zase do čtení a přečetl jsem si, že pravděpodobnost, že v letadle bude klaun a letadlo se zřítí, je jedna ku jedné miliardě. Hned po přečtení jsem si šel koupit klaunský kostým i letenku.

PŘÍKLAD 4.5

Korelace mezi inteligencí (IQ) a spotřebou kávy (počet šálků denně) je blízká nule. To odpovídá bodům v obrázku 4.5. Neexistuje zde žádný vztah: IQ nemá vliv na spotřebu kávy a spotřeba kávy nemá vliv na IQ.

Korelace nabývá hodnot na intervalu <−1,1>. Korelace téměř nikdy není přesně rovna nule. Není přesně stanoveno, od jaké velikosti je korelace významně pozitivní či významně negativní. Pokud máme méně než 20 pozorování, tak korelace poukazuje na významný vztah přibližně od absolutní hodnoty vyšší než 0,5 – tzn. při $|r| > 0{,}5$. Se zvyšujícím se počtem pozorování se hraniční hodnota (hodnota, při jejímž dosažení můžeme říct, že se jedná o významný vztah) snižuje. Např. při 100 pozorováních bude korelace pravděpodobně významná již při $|r| > 0{,}3$. Pokud je pozorování málo, tak je hodnota velmi nepřesná, a proto na potvrzení významnosti potřebujeme silný důkaz (velký korelační koeficient). Naopak při velkém

množství pozorování odpovídá vypočtený koeficient skutečnému stavu více (s více daty je přesnější) a lze mu tedy dávat vyšší důležitost i při nižších hodnotách. Pro posouzení, zda je hodnota korelace významná, lze využít i speciální statistický test.

4.2 Korelace není kauzalita (vliv)

Hodnota korelace vypovídá pouze o pozorovaném vztahu. Neříká vůbec nic o kauzalitě (o vlivu).

PŘÍKLAD 4.6

Existuje pozitivní korelace mezi lidským zdravím a štěstím. Z tohoto údaje ovšem není možné určit směr kauzality vlivu. Zdraví přináší štěstí, anebo štěstí přináší zdraví? Skutečnost, že existuje pozitivní korelace mezi štěstím a zdravím, se dá formulovat dvěma způsoby: 1) zdravější lidé jsou šťastnější, 2) šťastnější lidé jsou zdravější. I když z prvního tvrzení mnoho lidí vyvozuje kauzalitu „zdraví způsobuje štěstí“ a z druhého tvrzení mnoho lidí vyvozuje kauzalitu „štěstí způsobuje zdraví“, tak obě tato tvrzení jsou ekvivalentní. Navíc žádné z nich nemluví o kauzalitě (o vlivu) – obě tvrzení popisují totožnou korelaci. Pokud bychom chtěli popsat vliv, tak bychom museli jedno z tvrzení přeformulovat – např. na „zdraví přináší štěstí“.

PŘÍKLAD 4.7

Existuje pozitivní korelace mezi chozením k psychologovi a výskytem psychických nemocí. Tj. lidé, kteří chodí k psychologovi, mají psychické problémy častěji než lidé, kteří k psychologovi nechodí. Znamená to, že chození k psychologovi způsobuje psychická onemocnění? Je sice pravda, že existuje mnoho špatných terapeutů a zařízení, která mohou psychickou nemoc způsobit (například to mohou být některá uzavřená oddělení psychiatrických léčeben), ale obecně psychologové psychické problémy zmenšují. Tedy korelace mezi chozením k psychologovi a psychickými problémy platí téměř jistě proto, že výskyt psychických těžkostí „způsobuje“ chození k psychologovi. Ovšem samotná informace, že existuje pozitivní korelace mezi chozením k psychologovi a výskytem psychických nemocí nám o kauzálním působení (o vlivu) neříká nic.

Pokud mezi dvěma proměnnými existuje korelace (korelují mezi sebou buď pozitivně, či negativně), nastává jedna z pěti situací:

1) Proměnná X (první proměnná) má vliv na proměnnou Y (druhá proměnná) a ne naopak – kauzální působení X na Y

PŘÍKLAD 4.8

Existuje pozitivní korelace mezi věkem a počtem vrásek. Věk přináší vrásky a ne naopak.

2) Proměnná Y (druhá proměnná) má vliv na proměnnou X (první proměnná) a ne naopak – kauzální působení Y na X

PŘÍKLAD 4.9

Existuje negativní korelace mezi tím, jak jsou lidé šťastní a délkou jejich každodenního dojíždění do práce (Zhu, Chen a kol., 2019). Delší dojíždění působí lidem nižší štěstí a ne naopak.

3) Proměnná X má vliv na proměnnou Y a zároveň proměnná Y má vliv na proměnnou X – kauzální působení X na Y i Y na X

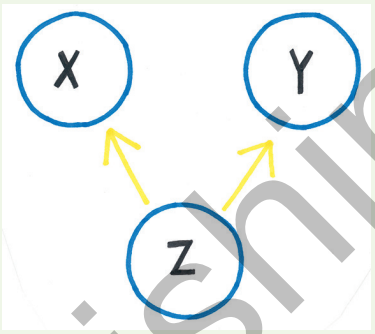
PŘÍKLAD 4.10

Zdravější lidé jsou častěji ženatí či vdaní (Yim, Park a kol., 2012). Být zdravým způsobuje, že lidé si častěji najdou partnera a ožení se s ním, a zároveň bytí v manželství přináší lidem větší zdraví (alespoň většinou).

4) Proměnná X nemá vliv na proměnnou Y, proměnná Y nemá vliv na proměnnou X a zároveň existuje třetí (skrytá) proměnná Z, která má vliv na X i Y

PŘÍKLAD 4.11

Počet bobrů na Vltavě pro poslední roky pozitivně koreluje s českým HDP. Počet bobrů na Vltavě nemá významný vliv na české HDP a české HDP nemá významný vliv na počet bobrů na Vltavě. Korelace je tu proto, že obě proměnné se zvyšují v čase (proměnná *Z*): s časem se zvyšuje, jak počet bobrů, tak i české HDP (viz obrázek 4.6). Pokud by si zde někdo pletl korelaci s kauzalitou, pak by pravděpodobně vysazoval tisíce bobrů na Vltavě s domněním, že česká ekonomika díky nim zažije obrovský růst.



Obrázek 4.6 *Proměnná X nemá vliv na proměnnou Y, proměnná Y nemá vliv na proměnnou X a zároveň existuje třetí proměnná Z, která má vliv na X i na Y.*

5) Proměnná X nemá vliv na proměnnou Y a proměnná Y nemá vliv na proměnnou X (žádná kauzalita) a korelace je jen náhodná

PŘÍKLAD 4.12

Počet obyvatel v České republice vykazuje v posledních pěti letech pozitivní korelaci s počtem banánových plantáží v Ekvádoru. Tato korelace neobsahuje žádnou kauzalitu; jedná se o pouhou statistickou náhodu. Tento případ může nastat především v případech, kdy máme málo pozorování (zde například pět ročních pozorování). Málo pozorování totiž neukazuje daný trend přesně.

Špatná interpretace korelace ve formě přisuzování kauzality může vést ke špatným implikacím a doporučením v reálném světě. To je vidět na příkladech níže.

Politik a statistik jedou vlakem Skotskem. „To je zajímavé," říká politik, „všechny skotské ovce jsou černé!" „To nemůžete říci," odporuje statistik, „nelze dělat takto uspěchané závěry. Statisticky vzato jsme právě učinili pozorování, že několik skotských zvířat vypadajících jako ovce je z jedné strany černých."

PŘÍKLAD 4.13

V rámci celého světa existuje pozitivní korelace mezi pitím supersladkých nápojů a zdravím. Způsobuje pití supersladkých nápojů lepší zdraví? Téměř jistě nikoliv. Ovšem supersladké nápoje pijí více lidé žijící v bohatších zemích. Právě bohatství a kvalitní zdravotní systém (ne supersladké nápoje) jim pak přináší lepší zdraví. Bohatství v průměru přináší vyšší konzumaci supersladkých nápojů a lepší zdraví. Korelace mezi supersladkými nápoji a zdravím platí pouze díky třetí proměnné „bohatství". Tento příklad tedy odpovídá výše popsanému typu korelace typ 4, obrázek 4.6. Neporozumění tomuto statistickému principu může vést k manipulaci. Například obsahem reklamy by mohla být informace: „Lidé, kteří pijí supersladké nápoje, jsou zdravější." To je sice pravda – korelace (vztah) existuje, ovšem mnoho lidí by tuto skutečnost pochopilo jako kauzalitu (vliv) a začali by pít supersladké nápoje pro zlepšení zdraví. Jedná se ale o manipulativní předání informace. Lidské zdraví by se při pití supersladkých nápojů naopak mohlo zhoršit. Je možné, že pití supersladkých nápojů zdraví zhoršuje, je možné, že ho zlepšuje a je také možné, že na něj nemá žádný vliv. Ovšem na základě znalosti samotné korelace (lidé, kteří pijí supersladké nápoje, jsou zdravější) nemůžeme o kauzalitě (vlivu) říct vůbec nic. Nemůžeme tedy sdělit ani žádné praktické doporučení.

Vědci často postupují tak, že po nalezení korelace začnou zkoumat, jaké vlivy v daném případě působí. Např. po nalezení korelace „supersladké nápoje – zdraví" je vhodné prozkoumat vlivy mezi těmito proměnnými. Statistické metody, které dokážou určit kauzalitu (vliv), jsou představeny v kapitole 5.

Přílišné slavení narozenin urychluje stárnutí. Statistiky ukazují, že lidé, kteří slavili mnoho narozenin, jsou starší než lidé, kteří slavili méně narozenin. Doporučujeme tedy vyhnout se jakýmkoliv oslavám. Toto platí obzvláště pro starší lidi, kteří nechtějí, aby se jim dále zvyšoval věk.

PŘÍKLAD 4.14

Lidé, kteří mají na ruce sádru, mají častěji zlomenou ruku. Způsobuje sádra lámání ruky? Ne. Naopak zlomená ruka „způsobuje“ aplikaci sádry. Nemusíme se tedy bát – když šáhneme na sádru, nezlomí nám to ruku.

PŘÍKLAD 4.15

Lidé, kteří kašlou, mají zápal plic častěji než lidé, kteří nekašlou. Způsobuje kašlání zápal plic? Znamená to, že bychom se měli snažit nikdy nekašlat a ochránit se tak před zápalem plic?

Na uvedených příkladech je jasně vidět, že korelace neznamená kauzalitu (vliv). Lidé, kteří si korelaci vysvětlují jako kauzalitu, v životě mohou dělat špatná rozhodnutí a jsou lehce manipulovatelní. Statistické metody, které dokážou určit kauzalitu (vliv), jsou představeny v kapitole 5.