Tohle nebylo příliš těžké, že ano? Viděli jsme, v jaké formě se může souvislost projevit v jedné skupině proměnných. Ale souvislost může mít ještě jiné formy. Zkusme si ještě jiný příklad.

A teď se podívejme na jinou tabulku. Sloupce v této tabulce představují proměnnou X, která má kategorie A, B, C, a D. Řádky reprezentují proměnnou Y. Ta má kategorie J, K, L a M. Všechna čísla jsou jenom v jejích tmavých polích. Je nějaká souvislost mezi proměnnými v tabulce 8.2.?

Tabulka 8.2.

Proměnná X

Proměnná Y

	A	B	C	D
J				
K				
L				
M				

Dr. Watson se horlivě hlásí:
To je jednoduché. Tady není žádná pořádná souvislost. Na diagonále není doslovně nic. Ani na jedné z nich!

Dr. Watson byl stvořen tak, aby se mýlil, tedy za jeho omyly můžeme my. Tenhle je docela typický. Tahle souvislost má prostě jinou formu. Ale řešení je opět docela prosté. Řekněme, že proměnná X v tabulce 8.2. představuje nějakou územní dimenzi, že kategorie A až D

reprezentují volební obvody. Proměnná Y reprezentuje politické strany, pro které by respondent mohl hlasovat. A teď se podívejme znovu na distribuci dat v tabulce. Vidíme, že **všichni** respondenti z obvodu A preferují politickou stranu K, že všichni z obvodu B hlasují pro stranu M atd. V každém sloupci jsou pozorování nahromaděna do jediného pole a pozice tohoto pole je pro každý sloupec jiná, jedinečná pouze pro tento sloupec. Když známe hodnotu X, odhadneme hodnotu Y bez jakéhokoliv omylu. Tedy naše fiktivní data představují jinou formu perfektní, deterministické souvislosti.

Ve skutečnosti by taková souvislost měla jinou formu. V jakékoliv společnosti s minimální úrovní demokracie by - doufejme - prázdná pole neexistovala. Pokud bychom zjistili, že proporce respondentů hlasujících pro určitou politickou stranu v určitém obvodu se podstatně liší od proporcí v jiných obvodech, pak by se zvýšila pravděpodobnost správného odhadu volby politických stran na základě znalosti volebních obvodů. Celá řada důležitých statistických operací je v podstatě **založena na srovnání nalezené distribuce pozorování do polí tabulky s takovou distribucí, jakou bychom obdrželi, kdyby byla pozorování zařazena do polí tabulky náhodně.**

A teď se podívejme na ještě jinou formu, ve které může být souvislost vyjádřena. Chtěli bychom řešit následující důležitý problém: kdo mezi studenty sociologie konzumuje více piva, muži nebo ženy? Následující čísla se týkají **průměrného** počtu půllitrů vypitých během jednoho týdne:

Tabulka 8.3.

	Aritmetický průměr:
muži	8
ženy	2

Můžeme na základě těchto dat uzavřít, že existuje souvislost mezi proměnnými "pohlaví" a "spotřeba piva"? Zdravý rozum a Dr. Watson navrhují, že ano: muži v našem vzorku pijí pivo čtyřikrát více než ženy. Zdá se, že opravdu existuje souvislost mezi proměnnými "pohlaví" a "spotřeba piva". Ale taková data jako je aritmetický průměr nebo údaje vyjádřené

v procentech představují velice podstatnou **redukci informace**. Zamyslete se třeba nad následující pohádku:

Pohádka pro odrostlejší děti 19.

O pohřešovaném kuřeti

Tuhle historku jste už asi slyšeli. Nevím, jaký je její původ, ale vypráví se po univerzitách a výzkumných ústavech celého světa. Je to v podstatě citát z výzkumné zprávy:

"Po aplikaci preparátu B se 33.3% kuřat uzdravilo, 33.3% uhynulo a o zbývajících 33.3% nejsme schopni poskytnout uspokojující informaci. Dosud se nám nepodařilo to třetí kuře chytit."

Morálka téhle pohádky je pro náš problém docela jasná: více bychom věřili průměru, který by byl vypočítán na vzorku 500 pozorování, než průměru vypočítaném pro vzorek pěti jedinců. Vzpomínáte, co jsme si řekli v kapitole 4. o intervalu spolehlivosti?

To ale ještě není všechno. Aritmetický průměr, stejně jako jiné podobné reprezentace středních hodnot redukuje informaci o mnoha jedincích do jednoho jediného údaje, a to je pěkně silná redukce, při které můžeme ztratit důležitý kus informace: Studujeme opět konzumaci piva ve dvou populacích. Pro obě populace jsme obdrželi zcela shodný průměr: 8 piv za týden. Můžeme tedy navrhnout, že jsou obě populace vzhledem ke konzumaci piva shodné? Pro spolehlivý závěr potřebujeme vědět, jak dobře průměr popisuje původní data. Uveďme si dva extrémní příklady, ilustrující původní data, ze kterých byl průměr vypočítán. Abychom ušetřili místo, předstírejme, že oba vzorky sestávaly jen z pěti jedinců:

Tabulka 8.3.

JEDINEC	Populace A: počet piv:	Populace B: počet piv
Jedinec 1.	8	0
Jedinec 2.	8	0
Jedinec 3.	8	0
Jedinec 4.	8	0
Jedinec 5.	8	40
Součet: Aritmetický průměr:	40 8	40 8

Je zřejmé, že průměr 8 reprezentuje skupinu A perfektně. Ale skupina B, to je docela jiná záležitost. To je vlastně skupina abstinentů, do které se vloudil jediný pivní hrdina, který nese obtížné břemeno: udržet průměrnou konzumaci piva na úrovni srovnatelné se skupinou A.

Je nesporné, že rozdíl mezi dvěma průměry signalizuje přítomnost souvislosti mezi proměnnou, podle které byli jedinci rozděleni do dvou subpopulací, a proměnnou popsanou jako průměr. Problém je jenom v tom, jak zjistit, že ten rozdíl mezi dvěma průměry je dostatečně významný. Teď už víme, že nestačí vzít v úvahu jen velikost vzorku, ale i to, jak je populace homogenní. Za chvíli se seznámíme s konceptem směrodatné chyby, která měří homogennost populace, ale hlavně, její diskuse nám umožní pochopit jiný důležitý koncept: statistickou významnost.

Ale ještě, než se podíváme, jak se souvislost opravdu měří, podívejme se, proč může mít souvislost tak mnoho různých tváří.

8.2. Statistika je třídní...

Vlastně třídní není ani tak statistika, ale proměnné, které můžeme, a jak uvidíte, musíme klasifikovat do několika skupin, které jsou vzájemně v hierarchickém vztahu. Je to důležité proto, že pro každou tu třídu proměnných můžeme použít jenom určitý soubor statistických operací. Skutečný statistik by vám předložil poněkud složitější třídění znaků a probral by i jiné principy pro klasifikaci proměnných, ale pro naši diskusi nám postačí podívat se na tři základní rodiny: **nominální, pořadové a intervalové proměnné.** (Zatajil jsem vám ještě jednu skupinu proměnných: alternativní znaky. Schoval jsem si to jako příjemné překvapení. Tak se, prosím, tvařte potěšeně, až je uvedu ve výkladu regresní analýzy.)

Nominální proměnné:

Říká se jim také kvalitativní proměnné. Jejich kategorie jsou pouhá jména a nedává mnoho smyslu se ptát, zda určitá kategorie je vyšší nebo nižší než jiná. Příkladem nominální proměnné je třeba respondentovo pohlaví, jeho barva vlasů, rodiště. To jsou proletáři mezi proměnnými. Řadu statistických operací, které můžeme používat pro ordinální a intervalové proměnné, nemůžeme zde uplatnit.

Pořadové proměnné

Říká se jim také ordinální proměnné. U těchto proměnných mohou být jejich kategorie seřazeny do nějaké hierarchie. Můžeme se smysluplně ptát, zda sledovaná vlastnost je u určitého jedince vyšší (nižší, silnější, lepší atd.) než u jiného respondenta. Nevíme však, o kolik je větší. Víme kupř., že stříbrná medaile je lepší než bronzová, ale ne tak dobrá, jako zlatá. Otázka, kolikrát je stříbrná medaile lepší než bronzová, však nedává smysl.

Intervalové proměnné

Ty mají takové kategorie, že nejen dává smysl se ptát, zda určitá kategorie je vyšší než jiná, ale také otázka, kolikrát je vyšší, je zde smysluplná. Příjem, věk, počet dětí jsou typickými ukázkami tohoto typu proměnných. Intervalové proměnné jsou aristokracií mezi ostatními proměnnými. Statistika s nimi může provádět taková kouzla, která nejsou dovolena pro nižší úrovně měření. Bohužel, intervalových proměnných není ve světě sociálního výzkumu mnoho.

Vidíme tedy, že by nemělo být obvykle příliš obtížné rozhodnout, do které skupiny určitá proměnná náleží. Stačí na ni aplikovat obě zmíněné kritické otázky a zamyslet se, zda jejich aplikace dává nějaký rozumný smysl.Následující shrnutí by nám mělo tento proces ulehčit.

Kritické otázky:

(A) Je určitá kategorie proměnné větší (menší) než jiná kategorie?
(B) Kolikrát je větší (menší)?

Jsou tyto otázky smysluplné?

A	B	
ne	ne	nominální proměnná
ano	ne	pořadová proměnná
ano	ano	intervalová proměnná

A opět slyšíme dr. Watsona brumlat někde v pozadí, k čemu je to všechno vůbec dobré. Důvod, proč potřebujeme vědět, do jaké skupiny určitá proměnná patří, je opravdu vážný: pro každou ze tří skupin proměnných můžeme použít jen určitý soubor statistických operací. Jako máme nominální, pořadové a intervalové proměnné, máme také nominální, pořadové a intervalové statistické operace. Jenže ty mají zajímavou hierarchii:

> **Nominální** statistické operace nedovedou mnoho z toho, co dovedou operace vyššího řádu. Ale mají jednu příjemnou vlastnost: můžeme je aplikovat na nominální, právě tak jako na pořadové nebo intervalové proměnné.

Pořadové operace dokáží více než nominální, ale zdaleka ne tolik, co intervalové. Můžeme je aplikovat jen na ordinální a intervalové proměnné, ne však na nominální.

Intervalové statistické operace dokáží daleko více, než obě předchozí. Můžeme je však aplikovat výhradně jen na proměnné intervalového charakteru.

A zde je tato hierarchie vyjádřena v tabulce:

Proměnné	Nominální operace:	Pořadové operace:	Intervalové operace:
nominální	ANO	NE	NE
pořadové	ANO	ANO	NE
intervalové	ANO	ANO	ANO

Smysl našeho výkladu snad pochopíme lépe na následujícím příkladu, který nepatří do oblasti měření souvislosti mezi znaky, ale do oblasti popisné statistiky. Často je pro nás výhodné vyjádřit informaci o vzorku nebo o celé populaci v co nejjednodušší formě. Chceme kupř. říci něco o počtu dětí v rodině v Praze. Publikovat seznam všech rodin s počtem dětí by poskytlo velmi úplnou informaci, ale bylo by to dosti nepohodlné, nepřehledné, a z mnoha důvodů i prakticky nemožné. Proto se obvykle spokojíme s informací o průměrném počtu dětí. **Aritmetický průměr** je **intervalový popis střední hodnoty**. Můžeme jej tedy použít jenom pro popis intervalových dat, jako počet dětí, příjem, věk apod. Ale zjistit, jaká je průměrná barva očí studentů sociologie by byl z hlediska statistiky docela absurdní úkol. Pro proměnné na různé úrovni měření používáme odpovídající indikátory centrální tendence:

intervalová data	**aritmetický průměr**
pořadová data	**medián**
nominální data	**modus**

Aritmetický průměr, ten umíme všichni vypočítat: prostě sečteme pozorované hodnoty a vydělíme je počtem sledovaných jedinců.

Medián je ta hodnota, která je právě v prostředku všech pozorování, která jsme seřadili podle jejich velikosti. Seřadíme třeba děti ve třídě podle velikosti a velikost dítěte, které je právě uprostřed řady, reprezentuje medián.

Modus je prostě kategorie s nejvyšší četností. Zjistíme-li třeba, že studenty mají nejčastěji modré oči, "modrá" bude modus.

A teď se podívejme, zda opravdu platí to, co jsme si řekli o aplikovatelnosti různého typu statistik. Nominální měřítko, modus, by měl být a opravdu je aplikovatelný samozřejmě na nominální, ale i na pořadová data. Ale je aplikovatelný kupř. i jako charakteristika intervalové proměnné, jako kupř. počet dětí v rodině. Mohli bychom kupř. roztřídit rodiny v našem vzorku do kategorií podle počtu dětí. Kategorie, ve které jsme nalezli nejvyšší počet pozorování, t.zv. modální kategorie, může být pak použita jako charakteristika dané populace.

Někdy může být dokonce výhodné použít statistiky nižší úrovně. Jsou totiž méně citlivé k extrémním hodnotám. Podívejte se znovu na tabulku 8.4. a sledujete data pro populaci B. Aritmetický průměr už známe, ale 8 není právě přesvědčivou reprezentací této populace. Jaký bude medián? Jedinec číslo tři je právě v prostředku, hodnota sledované proměnné se rovná nule. To je, alespoň intuitivně, lepší reprezentace vzorku. Mezi odpověďmi na naši otázku se nejčastěji objevuje nula. Tedy modus se rovná nule. Prostě, medián a modus nebyly ovlivněny atypicky vysokou hodnotou odpovědi jedince číslo 5. Jistě jste si všimli, že pro zcela homogenní populaci a ze stejné tabulky aritmetický průměr, medián a modus mají stejnou hodnotu: osm.

Nejčastěji je však výhodnější zvolit "nejvyšší" typ statistické operace z těch, které smíme použít; tyto operace prostě dovedou mnohem více, než operace "nižší". Už víme, že střední hodnoty charakterizují vzorek tím lépe, čím je tento vzorek homogennější. Pracujeme-li s intervalovými proměnnými, můžeme popsat homogenitu vzorku docela chytrým způsobem. Nabízí se nám tu dva koncepty: **rozptyl** (variabilita) a jeho mnohem rafinovanější příbuzná, **směrodatná odchylka**. Rozptyl nám poskytne informaci, jak se pozorování v průměru liší od průměru. Ale...

Dr. Watson nás přerušuje:
Já už vím, jako to udělat: Nejdřív vypočítám aritmetický průměr, pro každého jedince odečtu pozorovanou hodnotu od toho průměru. Pak sečtu všechny ty odchylky dohromady a výsledek vydělím a dostanu...

A teď musíme přerušit my: "Nulu, pokaždé nulu, doktore!" Když by průměr vypočítán správně, součet negativních odchylek musí být přesně stejně veliký, jako součet pozitivních odchylek. Ale s výjimkou jediného kroku měl dr. Watson pravdu. Ukažme si teď, jak se to skutečně dělá. Protože jsem vám slíbil, že nebudu používat (skoro) žádné vzorečky, budeme si prostě povídat, jak to uděláme:

Návod k výpočtu směrodatné odchylky
(speciálně pro Dr. Watsona)

Co uděláme:	Co to znamená
1. Pozorovanou hodnotu pro každého jedince odečteme od vypočítaného průměru.	Vypočítali jsme soubor odchylek pro celý vzorek.
2. Odchylku vypočítanou pro každého jedince umocníme.	Negativní číslo násobené samo sebou ním dá vždy pozitivní hodnotu. Tím překonáme problém, na který narazil dr. Watson. Součet neumocněných odchylek by nám musel pokaždé dát nulu.
3. Umocněné odchylky sečteme.	Součet představuje souhrnnou velikost (umocněných) odchylek.

4. Součet vydělíme počtem jedinců ve vzorku.	Často potřebujeme porovnat homogenitu souborů různé velikosti. Abychom kontrolovali rozdíly ve velikosti vzorků, vypočítáme, kolik z celkové sumy čtvercových odchylek připadá v průměru na každého jedince. Výsledek tohoto kroku je **rozptyl**, neboli **variabilita** sledovaného souboru.
5. Výsledek dělení odmocníme.	Rozptyl není měřen v jednotkách, ve kterých bylo pozorování prováděno, ale v mocninách těchto jednotek. Původně jsme pro každého jedince odchylky od průměru umocnili. Nyní kumulovaný průměrný výsledek opět odmocníme a dostaneme **směrodatnou odchylku.**

Cvičení 8.3.

Zkuste vypočítat rozptyl a směrodatnou odchylku pro populaci A a B z naší tabulky 8.4.

Směrodatná odchylka je svým způsobem magické číslo. Většina jevů v přírodě (bohužel ne tak docela automaticky ve společenských vědách) má tak zvané **normální rozložení**: Na stromě je nejméně hodně malých lístků. S přibývající velikostí stromových lístků jejich frekvence přirůstá a dosáhne maxima u listů střední velikosti. Když velikost lístků překročí průměrnou hodnotu, jejich četnost ubývá a opět, podobně jako tomu bylo s nejmenšími lístky, nejméně bude opět těch největších stromových listů. Podobnou distribuci - i když ne tak soustavně - objevíme i u řady sociálních jevů, jako výše příjmu, počet dětí v rodině, léta školního vzdělání atd. Můžeme si to snadno graficky znázornit. Na vodorovnou osu naneseme hodnoty, které může studovaná proměnná nabývat. Na kolmou osu pak naneseme množství případů (pozorování, jedinců) kteří mají danou hodnotu proměnné. Pak spojíme nalezené průsečíky křivkou. Máme-li hodně pozorování, dostaneme ladnou křivku zvonovitého tvaru.

Má-li studovaná proměnná alespoň přibližně takové normální rozdělení, směrodatná odchylka může začít dělat své divy. Magické operace začínají asi takto. Nejdříve odečteme standardní odchylku od průměru. Pak ji opět přičteme k průměru. Mezi těmito dvěma hodnotami bude vždycky přibližně **68% ze všech pozorování**. Nezáleží na tom, má-li naše křivka tvar profilu hory Řípu, nebo je velice plochá, nebo činí příkře do výše ve formě jakéhosi falického symbolu, v rozsahu definovaném průměrem ∓ směrodatná odchylka bude vždy 68% pozorování. Když od průměru odečteme a přičteme místo jedné směrodatné odchylky **dvě**, v rozmezí definovaném těmito novými hodnotami bude **95%** pozorování.

A teď si asi říkáte "No, a co z toho?" Kupodivu hodně. Právě tato vlastnost směrodatné odchylky nám umožní dělat některá zajímavá kouzla. Přirozeně, směrodatná odchylka měří homogennost souboru. Umožní nám definovat, jak dobře vypočítaný průměr charakterizuje populaci. Nechá nás formulovat kupř. tvrzení tohoto typu: "Měsíční příjem osob našeho vzorku byl 1.480,- Kčs průměrně. Můžeme předpokládat, že průměrný příjem populace spadá s 95% pravděpodobností do oblasti mezi 1.410,- a 1.550,- Kčs." (Vzpomínáte? O podobných operacích jsme již hovořili v naší kapitole 5. v souvislosti s problémy výběrové chyby.) To je příklad důležité role, kterou směrodatná odchylka hraje pro definování **statistické významnosti** našich výsledků. Díky směrodatné odchylce jsme třeba schopni říci, že existuje jen pětiprocentní pravděpodobnost, že rozdíly v průměrné konzumaci alkoholu ve dvou zkoumaných skupinách jsou náhodné a že pro zbývajících 95% procent můžeme doufat, že rozdíly jsou skutečně funkcí nějaké vlastnosti (třeba pohlaví), podle které byly zkoumané osoby zatříděny do obou skupin.

Tolik tedy o zdánlivě tak jednoduché věci, jako je aritmetický průměr. Jednoduché, ale představující intervalovou statistickou operaci, a to nám umožní aplikovat na ní takové účinné triky jako je měření rozptylu, nebo směrodatnou odchylku. Medián a modus nám to tak lehce neumožní.

V oblasti měření souvislostí jsou rozdíly mezi jednotlivými úrovněmi statistických operací ještě markantnější. Právě z tohoto důvodu je výhodné používat měření na intervalové úrovni co nejčastěji a musíme o tom začít ve výzkumném procesu přemýšlet velmi brzy .

Uveďme si alespoň jeden příklad. Jaká proměnná je "vzdělání"? Nominální, pořadová, nebo intervalová? Jediná odpověď kterou můžeme navrhnout, je: přijde na to, jak jsme vzdělání definovali v naší operační definici. Velice často je vzdělání popsáno v kategoriích jako "neukončené základní vzdělání", "ukončené základní vzdělání" atd. Tedy nejčastěji bude vzdělání patři mezi **pořadové proměnné**. Pro některé účely může být výhodná definice vzdělání jako **nominální proměnné**; kupř. když chceme studovat vliv určitého typu vzdělání na kariéru bývalých studentů: jak se srovnává deset let po dokončení školy plat absolventů průmyslovky s platem absolventů gymnázia, platem absolventů techniky a - nedej Bože - s příjmem studentů s titulem bakaláře v sociologii. Svého času se mnohé doktorské práce studentů v USA zabývaly vlivem prestiže university na budoucí status studentů. Porovnávaly se kupř. platy absolventů nejprestižnějších universit (Big Ten a Ivy League) s platy absolventů jiných univerzit. Mimochodem, často se ukázalo, že zjištěné rozdíly v platech mizí, kontrolujeme-li sociální status studentových rodičů. (Pamatujete si ještě koncept nepravé korelace?)

Hodláme-li použít proměnnou "vzdělání" jako element respondentova sociálního statusu (ale i pro mnohé jiné účely), měli bychom vážně uvažovat o takové operační definici vzdělání, která by nám umožnila jednat se vzděláním jako s intervalovou proměnnou. Je to jednoduché. Proč nepopsat vzdělání jako počet úspěšně dokončených let formálního školního vzdělání? Pro účely mezinárodního srovnávání je to naprosto nezbytné. Čemu v našem vzdělávacím systému odpovídá absolvování "lycea" v Itálii? Čemu odpovídá dokončené středoškolské vzdělání získané v Kanadě? V Ontariu je to 12 nebo 13 let, ve většině jiných provincií 11 nebo 12 let. (A jako by situace nebyla již tak dost komplikovaná, toto středoškolské vzdělání je v Kanadě povinné.) Ale i pro každý jednoduchý docela domácí výzkum je velice výhodné mít vzdělání definováno v letech. Zejména proto, že intervalové proměnné nám, mimo jiné, umožní odpověď na řadu docela zajímavých problémů, jako kupř. "Jak by se typicky zvýšil příjem jedince v závislosti na faktu, že jeho vzdělání vzrostlo o jeden rok, ale všechny ostatní studované faktory, ovlivňující příjem by zůstaly nezměněny?" Nebo: "Co ovlivňuje dosaženou výši jedince více: jeho pohlaví, jeho věk, povolání jeho rodičů?" V naší deváté kapitole uvidíme, že to není tak jednoduché, a že na úrovni intervalového měření mohou mít tyto operace mnohem jednodušší a přehlednější logiku než obdobné operace na nominálních či ordinálních datech.