Může se to zdát očividné, nicméně stojí za to zdůraznit, že užitečnost dat závisí na tom, že se vůbec podařilo sesbírat data správná a že jsme je sesbírali bez zkreslení či překroucení. Každá z těchto podmínek je náchylná na rizika spjatá s temnými daty. Vlastně je takových možných rizik tolik, že jen vyčerpávající vypočítání těch nejpodstatnějších je zhola nemožné. Nicméně i pouhé širší povědomí o situacích, na které je třeba dát pozor, je pro práci s temnými daty nesmírně prospěšné. Tato kapitola zkoumá, která data se máme snažit sbírat, a v té příští se podíváme na to, jak toho lze dobře dosáhnout, obojí z perspektivy rizik spjatých s temnými daty.

Různé definice a měření toho, co měřit nemáme

Významný druh temných dat vzniká z toho, že užíváme nepatřičných definic – nebo z toho, že, jak se říká, nevíme, o čem mluvíme. Podívejme se na pár příkladů.

Imigrace

Ankety navrhujeme s ohledem na cílové otázky, avšak administrativní data mohla být shromážděna z důvodu zcela odlišného. Znamená to, že administrativní data nemusejí být vhodná pro to, aby podala odpověď na otázku, která nás zajímá. Například ve Velké Británii nedávno došlo ke sporům ohledně přesnosti Statistik dlouhodobé mezinárodní migrace (LTIM). Národní statistický úřad (ONS) udal číslo 257 000, založené na Mezinárodním průzkumu cestování (IPS), pro počet lidí, kteří se za rok do září 2015 přistěhovali do Británie z Evropské unie. Avšak počet žadatelů

o národní číslo pojištěnce (NIN) dosahoval mezi obyvateli EU v tomto období až 655 000. NIN jsou osobní čísla pro lidi pracující ve Velké Británii, jež mají zaručit, že budou náležitě zaznamenány platby daní a odvody do národního pojištění (např. zdravotního a sociálního), takže tento nesoulad vypadá přinejmenším podivně. Jako by čísla ONS byla úplně mimo. Britský politik Nigel Farage k tomu uvedl: „Chtějí nás opít rohlíkem. Čísla NIN jsou jednoduchým a jasným odrazem skutečných počtů lidí v této zemi, jelikož bez nich nemůžete ani legálně pracovat, ani se hlásit o podporu.“[8]

Měření IPS, které pokrývá všechny hlavní vzdušné, mořské a tunelové vstupní brány do Velké Británie, běží bez ustání od roku 1961. Rok co rok provádí mezi 700 000 až 800 000 pohovorů. Třebaže toto číslo představuje pouhý zlomek lidí cestujících z/do Velké Británie, lze podle odpovědí sestavit odhad celkového počtu migrantů. Jde však pouze o odhad, takže je s ním nevyhnutně spjata určitá nejistota. ONS udal měřítko vlastní nejistoty coby ± 23 000, tedy v intervalu 234 000 až 280 000 s tím, že má jistotu 95 procent, že se právě v něm skrývá správné číslo. Třebaže je značná, zcela jistě tato nejistota nevysvětluje rozdíl mezi tímto číslem a číslem NIN.

A tak se tedy ONS pustil do důkladného průzkumu rozporu mezi tímto odhadem a číslem NIN.[9] Zjistilo se, že jeho hlavní příčinou je krátkodobá migrace („přistěhovalci, kteří v zemi … pobývají 1 až 12 měsíců“). Dlouhodobí migranti v zemi pobývají 12 měsíců a víc. Krátkodobí migranti se sice mohou ucházet o práci a číslo NIN, avšak hlavní počty zde představují čísla spjatá s LTIM (dlouhodobými migranty). ONS se dokonce nechal slyšet, že „rozdíly v definicích mezi těmito daty jsou podstatné a žádný celkový součet, který by prostě „přičetl“ či „odečetl“ různé prvky registrací NIN tak, aby odpovídaly definicím LTIM, není možné poskytnout… Údaje z registrací NIN tedy nejsou dobrým vodítkem pro LTIM…“ Jednoduše řečeno zde byla administrativní data zacílena na operaci, pro niž byla sebrána, a pro další účely se příliš nehodila. Nepatřičné či nevhodné definice v podstatě zastiňují data, o něž se zajímáme. Dokládají tak *TD-typu 8: Definice dat* a klíčové je při nich pamatovat na to, že to, zda jsou data temná, či nikoli, závisí na tom, co chceme zjistit.

Kriminalita

Další příklad temných dat vznikajících z rozdílů v definicích dokládají statistiky kriminality. Na národní úrovni pocházejí tyto statistiky ze dvou hlavních a poněkud odlišných zdrojů: ze statistik Úřadu pro průzkum

kriminality v Anglii a Walesu (CSE&W) a z Policejních záznamů o zločinnosti (PRC). CSE&W je rovnocenný s Národním úřadem pro oběti zločinu ve Spojených státech. Založili ho v roce 1982 (coby Britský úřad pro oběti zločinu) tak, že se lidí vyptali na zkušenosti se zločinem za uplynulý rok. Data PCR se sbírají od 43 policejních sborů Anglie a Walesu i od britské dopravní policie a analyzuje je Národní statistický úřad.

Odlišný způsob sběru informací má okamžité důsledky pro temná data. Už z povahy věci vyplývá, že jelikož CSE&W pořádá mezi lidmi anketu ohledně toho, jakou kriminalitu sami zažili coby oběti, nenahlašuje tím pádem údaje o vraždách či vlastnictví omamných látek. Také nezahrnuje lidi žijící ve skupinovém bydlení, např. v domovech důchodců či na studentských kolejích, a vynechává zločiny proti obchodním organizacím či veřejným institucím. Jak vidíme, máme zde vysoký potenciál pro vznik temných dat, třebaže toto riziko jasně vyplývá z určité definice toho, co daná anketa pokrývá.

I z PRC statistik plynou temná data, třebaže mají vůči těm od CSE&W doplňkový odstín. Už z definice nezahrnují PRC statistiky zločinů nenahlášených policii, třeba proto, že měla oběť dojem, že s nimi policie stejně nic nezmůže. Na tom záleží, jelikož se odhaduje, že k nahlášení dochází jen zhruba u čtyř z každých deseti zločinů, byť se toto číslo výrazně liší podle druhu zločinu. Navíc se v policejních statistikách objevují jen ty zločiny, jež spadají do kategorie „zločinů s ohlašovací povinností," tedy takových, které nemůže soudit porota (a pár dalších). Další komplikace vznikají z mechanismů zpětných vazeb (*TD-typ 11: Zpětná vazba a gaming*). Například počet zločinů souvisejících s držením drog bude záviset na rozsahu policejní činnosti a rozsah policejní činnosti bude zase záviset na tom, jaké je domnělé držení drog, a to zase bude záviset na počtu zločinů souvisejících s držením drog v minulosti.

Užívání odlišných definic vysvětluje rozpory mezi úrovněmi zločinnosti, které nahlásily tyto dva zdroje. Například v roce 1997 zaznamenaly PRC 4,6 milionu zločinů, zatímco CSE&W jich odhadoval na 16,5 milionu. Rozdíly také vysvětlují zmatení mezi mediálními odborníky i běžnými čtenáři nad tím, že podle PRC kriminalita v letech 1997–2003 narostla (z 4,6 milionu na 5,5 milionu případů), zatímco podle Úřadu pro průzkum kriminality klesla (z 16,5 milionu na 12,4 milionu).[3] Narůstá tedy kriminalita, nebo naopak klesá? Sami nejspíš uhodnete, kterou stranu se rozhodla zvýraznit média.

Lékařství

Imigrace a kriminalita jsou pouze dvě z nepřeberného množství oblastí, v nichž mohou definice vyvolat temná data, pokud nezohledníme případy či typy, které bychom do sesbíraných dat zahrnout měli. Občas mohou být následky překvapivé. Temná data související s definicí mohou například vysvětlit, proč na komplikace spojené s Alzheimerovou chorobou umírá víc lidí než v minulosti.

Alzheimerova choroba je nejběžnější formou demence. Je progresivní, přičemž raná stadia obvykle zahrnují drobné výpadky paměti a stadia pozdější se vyznačují zmateností, neschopností chápat, co se děje, a změnami osobnosti. Má se za to, že postihuje nějakých 50 milionů lidí po celém světě, ovšem tento počet narůstá a dle předpovědi bude v roce 2030 dosahovat až hodnoty 75 milionů. Temná data nám tento nárůst mohou vysvětlit přinejmenším dvěma způsoby.

Zaprvé na ni nikdo nikdy nezemřel před rokem 1901, kdy německý psychiatr Alois Alzheimer popsal první případ choroby, kterou po něm posléze pojmenovali. Navíc byla diagnóza původně vyhrazena pro lidi ve věku 45–65 let s příznaky demence. Až později, v poslední čtvrtině 20. století, se věkové vymezení rozšířilo. Je nasnadě, že takovým rozšiřováním definic se promění i množství lidí diagnostikovaných s chorobou. Data dříve pokládaná za irelevantní se nám zviditelní.

Druhé vysvětlení pomocí temných dat pro to, že na komplikace spojené s Alzheimerovou chorobou umírá víc lidí než v minulosti, se může zdát paradoxní: dochází k tomu vlivem pokroku lékařské vědy. Díky rozvinuté medicíně se lidé, kteří by zemřeli zamlada, dožívají dost vysokého věku na to, aby podléhali dlouhodobým degenerativním chorobám, jakou je Alzheimerova choroba. Tím vzniká celá řada zajímavých problémů včetně otázky, zda prodlužování života nutně přináší všeobecný prospěch.

Skutečnost, že počet diagnóz autistické poruchy se od roku 2000 ve Spojených státech zdvojnásobil, lze také vysvětlit pomocí *TD-typu 8: Definice dat*.[4] V kapitole 2 jsme viděli, že jedním z důvodů pro tento nárůst je klam dostupnosti – u této poruchy došlo ke zvýšení povědomí. Dalším velice podstatným důvodem pro tento nárůst je to, že došlo ke změnám v samotné formální *definici* a *diagnóze* autismu. Zejména jde o to, že byť byl autismus zahrnut do *Diagnostické a statistické příručky duševních poruch* už v roce 1980, jeho diagnóza se roku 1987 a pak znovu roku 1994 změnila tak, že její kritéria se značně rozvolnila. Uvolněte diagnostická kritéria tak, že je snazší je splnit, a bude to znamenat, že jim vyhoví víc lidí.

Navíc se v roce 1991 americké ministerstvo školství rozhodlo, že diagnóza autismus pro dítě znamená zvláštní pedagogické služby, a roku 2006 americká akademie pediatrů doporučila, aby se autismus u dětí zkoumal i během obvyklých pediatrických prohlídek. Změníte-li to, jak se data užívají, nepřekvapí vás, když se změní také chování při sběru těchto dat – jev zpětné vazby tohoto druhu prozkoumáme v kapitole 5. Týž jev dokládají také důsledky spuštění (v Anglii v únoru 2009) kampaně pro celonárodní boj proti demenci spolu s národní strategií proti demenci se záměrem zlepšit její diagnostiku a kvalitu péče. Asi nepřekvapí, že v důsledku došlo u diagnózy demence v porovnání s hodnotami z roku 2009 k nárůstu odhadem o čtyři procenta roku 2010 a o 12 procent roku 2011.[5]

Ekonomie

Obecně vzato je očividné, že změny v čase mohou vést ke změnám v povaze dat, která sbíráme. Nejenže to může ztížit retrospektivní srovnávání, ale také to může vést k obvinění z nepoctivosti. Jako zřejmý příklad lze uvést definice nezaměstnanosti: změňte definici a výkonnost vlády může najednou vypadat mnohem lépe.

Dalším příkladem v ekonomii je měření inflace. Definice inflace se zakládají na tom, že zaznamenávají ceny definovaného souboru zboží a služeb (kterému se říká „koš“ zboží a služeb – není to ovšem skutečný koš) a zjišťuje se, jak se průměrné ceny mění v průběhu času. Máme tu však různé komplikace – a všechny závisejí na temných datech *TD-typu 8: Definice dat*. Jednou z nich je otázka, jak vypočítat průměr, protože statistici mají několik různých způsobů, jak to provést – např. průměr aritmetický, geometrický, harmonický atd. V poslední době přešlo Spojené království od indexu založeného na aritmetickém průměru k indexu založenému na průměru geometrickém, čímž se sladilo s většinou ostatních zemí. Použitím jiné definice se na věci díváte z jiného úhlu pohledu, takže přirozeně vidíte – a nevidíte – jiné aspekty daných dat.

Kromě důsledků vyplývajících ze změny vzorce vznikají temná data v inflačních indexech také zásadnějším způsobem: je třeba rozhodnout, jaké položky do koše zahrnout a jak získat jejich ceny. Obecně platí, jak ukázaly předchozí příklady, že si musíme být vědomi rizika temných dat pokaždé, když se výběr provádí během sběru dat. Zde nám otázka toho, co do košíku vložit, může způsobit problémy, protože společnost se mění a smyslem inflačních indexů je odrážet určitým způsobem životní náklady. Poslední fráze je záměrně nejednoznačná („určitým způsobem“), protože

různé indexy měří různé aspekty zkušeností s inflací. Některé měří, jak změny cen ovlivňují jednotlivce, jiné zase, jak ovlivňují větší ekonomiku, a tak dále. V každém případě je důležité, aby koš položek byl *relevantní* – to znamená, aby se skládal ze zboží a služeb, které lidé skutečně nakupují. Tento problém lze jasně znázornit porovnáním toho, co mohlo být zahrnuto do koše cenového indexu před dvěma stoletími, s tím, co by do něj šlo vložit dnes. Před dvěma sty let by důležitou položkou byly jistě svíčky, dnes však výraznou část spotřebního koše lidí obecně nepředstavují. Dnes tvoří výrazné výdaje mobilní telefony a automobily. To znamená, že máme jmenovitý seznam položek, jež by v zásadě mohly být zahrnuty do košíku, ale nebudeme do něj chtít zahrnout všechny. Ačkoli se při určování, které položky do koše přesně zahrnout a zaznamenat jejich cenu, dlouze a pilně přemýšlí, je nasnadě, že se zde otevírá prostor pro nejednoznačnost, do nějž se může vloudit také libovůle.

Co se týká druhého bodu, tedy jak získat ceny položek v koši, tradičně se to dělalo prováděním průzkumů a vysíláním týmů výzkumníků do obchodů a na trhy, aby zde zaznamenávali ceny zboží. Americký statistický úřad práce provádí průzkum ve 23 000 obchodech a každý měsíc zjišťuje ceny přibližně 80 000 položek spotřebního zboží, u kterých provede celkový součet a získá tak index spotřebitelských cen. Podobně se postupuje i v jiných zemích.

Možná jste si všimli, že tento tradiční přístup ke zjišťování cen zboží zcela opomíjí nakupování online. Vzhledem k tomu, že tyto nákupy nyní představují přibližně 17 procent maloobchodních tržeb ve Spojeném království[6] a téměř 10 procent maloobchodních tržeb v USA[7], vypadá to, že velký počet relevantních cen nemusí vůbec do indexu přispívat. (Měl bych dodat, že jde o údaje „v době psaní této knihy“, protože trendy jsou vzestupné a strmé). Z tohoto důvodu mnohé země vyvíjejí opatření založená na internetovém sběru cen. Tato opatření se nesnaží přesně kopírovat tradiční opatření, protože koše se budou lišit. Příklad uvidíme v kapitole 10.

Společnost se neustále mění, v současnosti možná víc než kdykoli v minulosti, protože počítač a s ním spojené technologie jako sledování, dolování dat, umělá inteligence, automatizované transakce a web mají stále větší vliv. Takto rychlé tempo změn má důležité obecné důsledky pro datovou analýzu z hlediska temných dat, protože prognózy týkající se budoucnosti se nutně zakládají na tom, co se stalo v minulosti. Z technického hlediska se posloupnosti dat v čase poměrně přirozeně označují jako *časové řady* dat. Rychlost změn metod datového sběru a technologií znamená, že potřebné časové řady často nesahají příliš daleko do minulosti.

Nové typy dat jsou nutně krátkodobé, takže data budou k dispozici pouze z relativně krátkého období bezprostřední minulosti. Za touto hranicí leží temnota.

Všechno změřit nelze

Datové soubory jsou vždy konečné. To je jistě pravda, pokud jde o počet případů – konečný počet lidí v populaci nebo konečný počet případů, kdy se něco měří. Ale platí to i z hlediska toho, *co* se měří nebo *jaké údaje* se o předmětech zájmu shromažďují. Pokud studujeme lidi, můžeme zjišťovat jejich věk, váhu, výšku, kvalifikaci, oblíbené jídlo, příjem a řadu dalších věcí. Vždy však zůstane nespočet dalších charakteristik, které jsme nezjistili. Tyto další charakteristiky jsou nevyhnutelnými temnými daty, z nichž vyplývají důsledky.

Kauzalita

Když zkoumání populace naznačilo spojitost mezi rakovinou plic a kouřením, upozornil přední statistik Ronald Fisher na to, že to nemusí nutně znamenat, že kouření rakovinu způsobuje. Kromě jiných možností poznamenal, že rakovina plic i náchylnost ke kouření jsou možná způsobeny nějakým jiným faktorem – například genetickým –, který obojí podporuje. To by byl klasický případ *TD-typu 5*: *Když chybí to podstatné* – nějaké jiné, neměřené proměnné, která způsobuje obojí a vyvolává tak mezi nimi korelaci, i když ani jedna druhou nezpůsobuje. To také ukazuje, jak obtížné může být temná data odhalit.

Ve skutečnosti jsme se s takovou situací setkali na samém počátku knihy. Hned v první kapitole jsem se zmínil, že u dětí v prvních letech školní docházky koreluje výška se slovní zásobou. Kdybyste tedy provedli průzkum měřící výšku a testující slovní zásobu u vzorku dětí ve věku od 5 do 10 let, zjistili byste, že vyšší děti mají v průměru širší slovní zásobu než děti menší. Z toho byste mohli vyvodit závěr, že čím víc slov děti naučíte, tím budou vyšší. A skutečně, kdybyste takový průzkum provedli a změřili počáteční výšky skupiny dětí, poté je vystavili intenzivní výuce nových slov a nakonec znovu změřili jejich výšku na konci roku, přišli byste na to, že opravdu vyrostly.

Ale čtenář jistě ví, oč tu běží. Zatímco výška a slovní zásoba těchto dětí jistě korelují, není to proto, že by mezi nimi existovala příčinná souvislost. Obě totiž souvisejí s třetí proměnnou, kterou nás možná nenapadlo

v našem průzkumu měřit, a sice s věkem dětí. Věk byl v této studii temnou datovou proměnnou a jeho nezměření mohlo vyvolat velmi zavádějící pochopení toho, co nám data vlastně ukazují.

Tato situace se liší od těch, kdy u některých osob (nebo obecněji objektů) chybějí hodnoty některých atributů v záznamu, a liší se také od situací, kdy u některých osob (nebo objektů) nedojde k záznamu *jakýchkoli* atributů. Nyní hodnoty určitého atributu nebo určitých atributů chybí u všech případů v databázi. Všechny záznamy pro takovou proměnnou by byly zaznamenány jako prázdné nebo „NA“ (nedostupné), pokud by vůbec došlo k záznamu proměnné. Například jsme v průzkumu možná nedopatřením zapomněli uvést otázku, kolik je respondentům let, takže nemáme informace o věku u nikoho. Nebo jsme si možná mysleli, že věk nebude relevantní, takže nás vůbec nenapadlo tuto otázku zahrnout. Nic z toho není vůbec přitažené za vlasy: pokud jsou průzkumy příliš dlouhé, bude to mít nepříznivý dopad na počet odpovědí, takže je třeba pečlivě vybírat, jaké otázky chceme zahrnout.

Paradox!

Občas mohou data *TD-typu 5: Když chybí to podstatné*, tedy taková, z nichž nám chybějí celé proměnné či důležité vlastnosti, vést k matoucím důsledkům.

Tragédii Titaniku zná každý – šlo o nepotopitelný parník, který se potopil. Podrobné zkoumání míry přežití cestujících a posádky však odhaluje něco zvláštního.[8] Jak ukazuje tabulka 2a, bylo na lodi 908 členů posádky, z nichž přežilo jen 212, tedy 23,3 procenta. A mezi 627 pasažéry třetí třídy, kteří cestovali nejhlouběji v lodi a pro které bylo nejtěžší dostat se ven, přežilo jen 151, tj. 24,1 procenta. Přestože mezi mírou přežití těchto dvou skupin není velký rozdíl, ukazuje se, že cestující měli o něco vyšší pravděpodobnost přežití než posádka.

Nyní se však podívejme zvlášť na míru přežití mužů a žen, kterou ukazuje tabulka 2b. Nejprve co se týká mužů. Mezi posádkou bylo 885 mužů, z nichž 192 přežilo, což představuje 21,7 %. A mezi cestujícími třetí třídy bylo 462 mužů a přežilo jich 75, což je 16,2 %, takže mužská posádka měla vyšší míru přežití než muži cestující ve třetí třídě.

Zadruhé co se týká žen. Mezi posádkou bylo 23 žen a 20 z nich přežilo, což představuje 87,0 %. A mezi cestujícími třetí třídy bylo 165 žen a přežilo jich 76, což je 46,1 %, takže ženská posádka měla vyšší míru přežití než ženy cestující třetí třídou.

Tab. 2 Poměrné zastoupení členů posádky a cestujících třetí třídou, kteří přežili potopení Titaniku: a) celkem; b) muži a ženy zvlášť

a)

Posádka	**Cestující třetí třídou**
212/908 = **23,3 %**	151/627 = **24,1 %**

b)

	Posádka	**Cestující třetí třídou**
muži	192/885 = **21,7 %**	75/462 = **16,2 %**
ženy	20/23 = **87,0 %**	76/165 = **46,1 %**

Co tedy zjišťujeme? Že u mužů a žen zvlášť měla posádka vyšší míru přežití než cestující třetí třídy, avšak celkově měla posádka nižší míru přežití než cestující třetí třídy.

To není žádný trik, čísla hovoří jasně. Zdá se to však téměř paradoxní – a tento jev se opravdu také často nazývá *Simpsonův paradox* podle Edwarda H. Simpsona, který jej popsal ve své práci z roku 1951 (ačkoli jiní tento jev popsali nejméně o půlstoletí dříve).

Důsledky jsou dost možná závažné. Kdybychom nezaznamenávali pohlaví osob na lodi – kdyby tyto údaje chyběly –, docela rádi bychom uvedli, že výsledky naší analýzy jsou takové, že cestující třetí třídy měli větší pravděpodobnost přežití než posádka. To by však bylo zavádějící, kdybychom se zajímali o muže – protože u nich jsou výsledky opačné. A stejně tak by to bylo zavádějící, kdybychom se zajímali o ženy. To znamená, že závěr by byl zavádějící, kdybychom se zajímali o kohokoli, neboť každý z cestujících byl buď mužem, nebo ženou.

Proč k této situaci dochází, prozkoumáme za chvíli, už teď je ale zcela zřejmé, že její potenciální důsledky jsou ohromující. Neomezené množství vlastností lidí, kteří se plavili na Titaniku, nikdy zaznamenáno nebylo. Pokud by některá z nich mohla mít za následek převrácení našich závěrů, jejich vynechání a chybějící údaje by mohly být velmi zavádějící. To by nemuselo v případě Titaniku tolik vadit, protože v takovém případě pouze popisujeme historické údaje, ale vezměme si následující příklad.

Předpokládejme, že provádíme klinickou studii toho druhu, o které jsme hovořili v předchozí kapitole, a porovnáváme lék A s lékem B. Abychom mohli léky porovnat, podáváme lék A jedné skupině lidí a lék B druhé skupině. Obě skupiny obsahují lidi různého věku, které pro pohodlí nazveme „mladší“ a „starší“, to podle toho, zda jsou řekněme mladší,

anebo starší 40 let. Pro konkretizaci budeme předpokládat, že skupina, která dostává lék A, sestává z 10 mladších a 90 starších, zatímco ve skupině, která dostává B, je 90 mladších a 10 starších.

Tab. 3 Průměrné výsledky léků A a B: a) pro mladší a starší účastníky zvlášť; b) celkově

(a)

	Průměrný výsledek	
	lék A	lék B
mladší	8	6
starší	4	2

(b)

Průměrné skóre	
lék A	lék B
4,4	5,6

Nyní se podívejme na výsledky, kde předpokládáme, že čím vyšší skóre, tím je léčba účinnější. Tyto (hypotetické) výsledky jsou uvedeny v tabulce 3.

Předpokládejme, že zjistíme, že průměrné skóre mladších ve skupině A je 8 a průměrné skóre mladších ve skupině B je 6, jak ukazuje tabulka 3a. To nám říká, že lék A je pro mladší účinnější, protože 8 je větší než 6.

Podobně u starších předpokládejme, že průměrné skóre ve skupině A je 4 a průměrné skóre ve skupině B je 2, jak je uvedeno v druhém řádku tabulky 3a. Pro starší je lék A účinnější než lék B.

Přestože průměrné skóre u starších je nižší než u mladších, ať už je léčba podána jakýmkoli způsobem, je zřejmé, že jak pro mladší, tak pro starší je lék A účinnější než lék B. Měli bychom doporučit, aby byl předepsán lék A.

Ale jak je to celkově? Celkové průměrné skóre všech osob A je $(8 \times 10 + 4 \times 90)/100 = 4{,}4$, zatímco celkové průměrné skóre všech osob, které dostávají lék B, je $(6 \times 90 + 2 \times 10)/100 = 5{,}6$. Tyto výsledky jsou uvedeny v tabulce 3b. Celkově platí, že pokud zanedbáme věk pacientů, lék B má vyšší skóre než lék A.

To znamená, že pokud bychom nezaznamenávali věk pacientů – pokud by údaje chyběly –, došli bychom k závěru, že B je lepší než A, přestože

A je lepší než B pro mladší i pro starší, tedy přestože A je lepší než B *pro všechny*.

Prvotní reakcí zřejmě bude, že bychom měli při sběru dat zaznamenávat věk. To zní samozřejmě velmi dobře, ale opět existuje nespočet dalších proměnných, které bychom mohli také zaznamenávat a z nichž každá by mohla mít stejně podivný inverzní efekt. A všechny možné proměnné zaznamenávat nemůžeme. Některé z nich nevyhnutelně tvoří temná data.

Klíč k záhadě spočívá ve způsobu výpočtu těchto celkových průměrů. V příkladu se zkouškou léků je ve skupině A mnohem víc starších než mladších, zatímco ve skupině B je tomu naopak. To vychyluje celkové průměry směrem dolů – 8 je větší než 6 a 4 je větší než 2, avšak pokud přisoudíte čtyřem dostatečnou váhu, když počítáte průměr 8 a 4, a šesti dáte dostatečnou váhu, když průměrujete 6 a 2, situace se obrátí.

Nyní tedy vidíme, v čem tkvěl problém – v rozdílu mezi podíly mladších v obou skupinách. Ve skupině, která dostávala lék A, bylo 10 procent mladších, zatímco skupina dostávající lék B měla 90 procent mladších. Kdyby obě skupiny měly stejný podíl mladších, problém by nevznikl. Vzhledem k tomu, že studie s léky je experimentem, v němž máme vládu nad tím, kolik pacientů jednotlivé léky dostane, mohli bychom problém odstranit tím, že vyrovnáme podíly mladších tak, aby byly v každé skupině stejné.

Tato metoda tedy funguje, pokud ovládáme, kdo vstupuje do které skupiny. Ale na Titaniku žádná taková kontrola nebyla – cestující byli cestující a posádka byla posádka. Kdo z nich je kdo, to jsme si vybrat nemohli. Následuje další příklad, v němž nemáme vládu nad tím, kdo je v které skupině.

Ve studii z roku 1991, která se zabývala vlivem rasy na tresty smrti v případech odsouzení za vraždu na Floridě, bylo k trestu smrti odsouzeno 53 ze 483 obžalovaných bělochů a 15 ze 191 obžalovaných Afroameričanů;[9] to znamená, že k trestu smrti byl odsouzen větší podíl bělochů (11,0 procenta) než Afroameričanů (7,9 procenta), jak ukazuje tabulka 4a.

Pokud však nyní vezmeme v úvahu rasu *oběti* i obžalovaného, obdržíme poněkud odlišný výsledek – a opět se objeví záhadný obrázek.