

## XIII

Kde lžou data,  
když čísla sedí?

*Po této kapitole se přestaneš ptát, co data říkají, a začneš se ptát, jak vznikla: kdo ve vzorku chybí, co bylo podmínkou vstupu — a kudy do bezchybně spočítaných čísel potichu vstoupil výběr.*

*Existují tři druhy lži: lži, zatracené lži a statistika.*

— Mark Twain — s připsáním Benjaminu Disraelimu, u něhož výrok nikdy nikdo nenašel, „*There are three kinds of lies: lies, damned lies, and statistics.*“  
*Chapters from My Autobiography, 1907*

*I tenhle epigraf je cvičení z kapitoly: Twain výrok „připsal“ Disraelimu, v jehož díle ani dopisech ho nikdo nenašel; nejstarší doložené výskyty jsou z Anglie devadesátých let 19. století, autor neznámý. Citát přežil díky lepšímu příběhu, ne lepšímu prameni — přežívá to, co se dobře vypráví.*

## Díry, které tam nebyly

Ten obrázek jsi skoro určitě viděl. Silueta bombardéru posetá červenými tečkami: trup a křídla prostřílená jako řešeto, motory a kabina čisté. K tomu poučný příběh — píše se rok 1943, armáda chce přidat pancíř na místa s nejvíce zásahy a geniální statistik Abraham Wald otáčí otázku: díváte se jen na letadla, která se vrátila; pancéřujte tam, kde díry nejsou, protože zásah do těch míst znamená, že se letadlo do vaší statistiky nedostalo. Nejsdílenější statistická anekdota internetu.

A teď to, co ti mem neřekne. Ten obrázek nenakreslil Wald. Vznikl v roce 2016 pro Wikipedii — editor s přezdívkou McGeddon ho překreslil podle náčrtu, který si designér Cameron Moll kolem roku 2005 vymyslel do prezentací: vzal siluetu civilního letadla a posel ji tečkami víceméně náhodně, aby měl čím ilustrovat myšlenku. Wald žádný diagram s tečkami nezanechal. A slavná scéna, v níž statistik před generály otáčí otázku? Ve Waldových memorandech není. Do světa ji pustil až memoár W. Allena Wallise, šéfa skupiny, sepsaný o sedmatřicet let později — a rozletěla se s knihou Jordana Ellenberga *How Not to Be Wrong* (2014).

Skutečný příběh je tišší a silnější. Rok 1943, Statistical Research Group na Kolumbijské univerzitě: hrstka statistiků počítá pro armádu všechno od zaměřovačů po přejímací testy munice — Wallis později vzpomínal, že se tam sešla nejpozoruhodnější

statistická sestava, jaká kdy seděla v jedné místnosti; u vedlejších stolů mladí Milton Friedman a Frederick Mosteller. Abraham Wald — matematik, který pět let předtím utekl z Vídně před anšlusem — dostane otázku zranitelnosti letadel. Jenže data jsou děravá stejným způsobem jako letadla: zásahy lze spočítat jen na strojích, které se vrátily. Wald nenapsal bonmot; napsal metodu. Osm technických memorand „A Method of Estimating Plane Vulnerability Based on Damage of Survivors“ — postup, jak z rozložení děr na přeživších dopočítat pravděpodobnost, že zásah do daného místa letadlo srazí. Jinými slovy: jak spočítat letadla, která nikdo nespočítal. Z rovnic pak plyne přesně to, co intuice neviděla — málo děr na motorech vrátivších se strojů neznamená, že se do motorů netrefuje; znamená to, že zásah do motoru se domů nevrací.

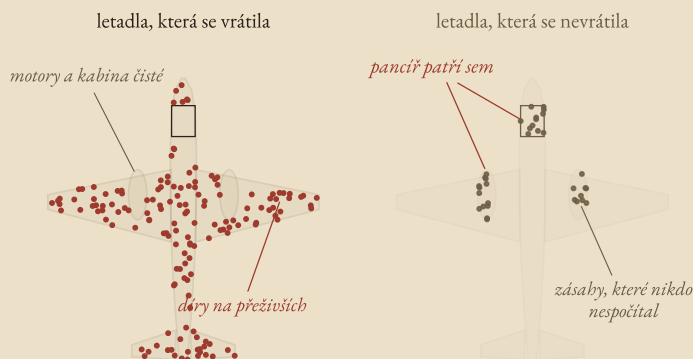
Memoranda nikdy nevyšla jako kniha s vtipnou pointou. Zůstala v šuplících vojenských analytiků a pracovala: tytéž metody se používaly ještě ve válkách v Koreji a ve Vietnamu. Veřejnost je poznala až z reprintu z roku 1980 a statistikům je o čtyři roky později připomněl rozbor v odborném časopise. Teprve pak se z tiché metody začal stávat hlasitý mem.

Proč tu demontáž podnikáme? Protože mem ti prodá vtip — a vtip si zapamatuješ místo lekce. Skutečný Wald je lekce: data nelhala čísla. Každá díra na každém vráceném letadle byla spočítána správně. Data lhala vzorkem — tím, kdo se do tabulky vůbec směl dostat. A síto pracovalo přesně proti směru otázky: přežívaly stroje zasažené tam, kde zásah nevádí, takže čím pečlivěji jsi díry počítal, tím jistěji tě tabulka vedla pancéřovat nesprávná místa. Čísla seděla. Právě proto byla nebezpečná.



*Prameny: W. Allen Wallis, „The Statistical Research Group, 1942–1945“, JASA 75/1980; Waldova memoranda souborně jako reprint CRC 432 (Center for Naval Analyses, 1980); M. Mangel & F. J. Samaniego, „Abraham Wald’s Work on Aircraft Survivability“, JASA 79/1984; provenience obrázku: Cameron Moll ≈ 2005, Wikipedie/McGeddon 2016.*

*Abraham Wald zemřel roku 1950 při letecké havárii v jižní Indii, cestou na přednáškové turné.*



*Vlevo to, co viděla armáda; vpravo to, co neviděl nikdo. I tahle kresba je moderní ilustrace — tečky jsou ilustrační, věrná je informační struktura: o pancíři rozhoduje právě letadlo, a právě to ve statistice chybí.*

Ta past má jméno: **klam přeživších**. Vzorek sis nevybral ty — vybral ho výsledek, na který se ptáš. Poznáš ji všude, jakmile ji jednou uvidíš: knihy rozhovorů s úspěšnými zakladateli („všichni vytrvali navzdory pochybám“ — jenže vytrvali i ti, o kterých knihy nevyšly), obdiv ke starým domům, „které se ještě uměly stavět“ (ty špatně postavené už nestojí, takže se do vzorku nedostaly), i tvůj vlastní dashboard: telemetrie měří jen požadavky, které doletěly. Ve všech případech je součet správně a vzorek křivý.

*klam přeživších (survivorship bias) — vzorek prošel sítí, které souvisí s měřeným výsledkem; čísla pak popisují síť, ne svět.*

→ kap. 21: monitoring měří jen to, co doletělo  
— tam se k síti vrátíme v produkci.



EINSTEIN · MATEMATIKA — přeskočitelné

**Waldova rovnice v kostce.** Označ  $p_i$  pravděpodobnost, že zásah dopadne do zóny  $i$  letadla (odhadnutelná z ploch — flak nemíří), a  $s_i$  pravděpodobnost, že stroj zásah do zóny  $i$  přežije. Mezi vrátivšími se letadly s jedním zásahem pak zónu  $i$  uvidíš s četností

$$P(\text{zóna } i \mid \text{vrátil se}) = \frac{s_i \cdot p_i}{\sum_j s_j \cdot p_j}$$

Pozorované rozdělení děr je skutečné rozdělení zásahů *převážené přežitím*. Armáda četla levou stranu, jako by to bylo  $p_i$ ; Wald rovnici obrátil a z děr na přeživších dopočítal  $s_i$  — čísla o letadlech, která nikdy neviděl. (Skutečná memoranda zvládají i vícenásobné zásahy a stroje bez šrámu; princip je tenhle zlomek.)

## DEMO

[lesk-a-bida-vibecodingu.gaussalgo.com/d/bombarder](http://lesk-a-bida-vibecodingu.gaussalgo.com/d/bombarder)

Rozmístí pancíř sám: klikěj podle děr na vrácených letadlech — a pak se ukážou ta, která se nevrátila.



Klam přeživších je ale jen nejslavnější člen početnější rodiny. Rodinné pravidlo zní: *výběr podmíněný následkem obývá čísla*. A nejtěžší případ té rodiny nezná internet zdaleka tak dobře jako bombardér — přestože je to možná nejpoučnější statistická detektivka dvacátého století.

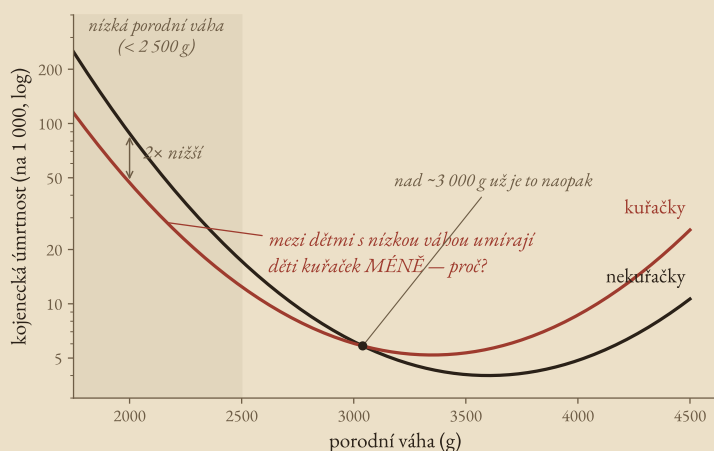
## Paradox porodní váhy

Jacob Yerushalmy byl biostatistik na Berkeley a skeptik z povolání. V roce 1971 publikoval rozbor kalifornské kohorty: zhruba třináct tisíc těhotenství sledovaných od prvních měsíců. Nejdřív potvrdil, co se vědělo. Kuřáčky rodí lehčí děti — děti pod 2 500 gramů měly výrazně víc než nekuřáčky. A nízká porodní váha je krutý prediktor: kojenecká úmrtnost je proti dětem s normální váhou řádově dvacetkrát vyšší. Kouření → nízká váha → smrt; řetěz vypadal uzavřený.

Pak Yerushalmy udělal, co dělá poctivý statistik: srovnal srovnatelné, děti stejné váhy s dětmi stejné váhy. A dostal výsledek, který se nelíbil vůbec nikomu. Mezi dětmi s nízkou porodní váhou umíraly děti kuřáček *méně* — zhruba o polovinu — než stejně lehké děti nekuřáček. V pásmu, kde jde o život nejvíc, vycházela cigareta jako ochrana.

Yerushalmy z toho vyvodil vlastní závěr: „ne kouření, ale kuřáčka“ — ženy, které kouří, se od nekuřáček liší v tolika věcech, že observační data o škodlivosti kouření nic nedokazují. V tomhle se mýlil; kouření v těhotenství škodí a dnes o tom není spor. Ale vyvrátit ho tehdy neuměl nikdo. Čísla seděla, paradox se replikoval v dalších kohortách i zemích, tabákovému průmyslu se náramně hodil — a epidemiologie s ním žila přes třicet let.

→ kap. 12: tam lhala správnost testu bez prevalence, tady lze úmrtnost podmíněná výběrem. Obě čísla jsou spočítaná správně — jen odpovídají na jinou otázku, než kterou sis položil.



tvar křivek schematický; poměry (22× a 2×) dle Yerushalmy 1971

Pramen: J. Yerushalmy, „The relationship of parents' cigarette smoking to outcome of pregnancy...“, American Journal of Epidemiology 93/1971, 443–456. Křivky schematicky, poměry dle původních dat; všimni si i obratu nad  $\approx 3\,000$  g.

Rozuzlení nepřinesla větší data ani přesnější přístroje, ale obrázek. V roce 2006 nakreslila trojice epidemiologů Hernández-Díaz, Schisterman a Hernán **kauzální graf (DAG)** celé situace: obrázek, v němž

uzly jsou veličiny a šipka neznamena „souvisí spolu“, nýbrž „působí na“. Z obrázku je vysvětlení vidět. Nízká porodní váha nemá jedinou příčinu: kouření je jedna — a vrozené vady, podvýživa a další poruchy jsou druhá, vzácnější a mnohem smrtelnější. Podmínka „jen děti pod 2 500 gramů“ se pak ptá: *tohle dítě je malé — čím to?* U dítěte kuřačky je po ruce poměrně milosrdné vysvětlení: cigarety. U dítěte nekuřačky cigarety k dispozici nejsou, zbývají příčiny zlověstnější. Vzorek lehkých dětí nekuřaček je proto prosycen těžkými diagnózami víc než vzorek lehkých dětí kuřaček — a jejich vyšší úmrtnost nevyrábí ochranný účinek kouření, nýbrž skladba vzorku, kterou jsi vlastní podmínkou stvořil.

Uzlu, do něhož míří dvě šipky kauzálního grafu, se říká **collider**: sráží se v něm dvě příčiny. Dokud ho necháš na pokoji, je to nevinná křížovatka. Jakmile na něm ale podmíníš — stratifikuješ podle něj, vybereš podle něj vzorek, „očistíš“ o něj regresi — otevřeš mezi jeho příčinami falešný kanál: kdo ví, že dítě je lehké, a zároveň že matka nekouří, ví tím pádem něco o třetí, neviditelné příčině. Přesně tenhle kanál vyráběl třicet let ochranný účinek cigaret. A Waldova letadla? Tentýž mechanismus: podmínka „vrátilo se“ je následek místa zásahu — kdo počítá jen vrácené stroje, podmínil celý vzorek na následku, jen si to nikde nenapsal.

## Tři ozvěny téhož

Jakmile mechanismus jednou uvidíš, nepřestaneš ho vidět. Rok 1946, klinika Mayo: statistik Joseph Berkson ukázal, že v nemocničních datech vychází cukrovka jako „ochrana“ před zánětem žlučníku — v obecné populaci nic takového neplatí. Nemocnice je collider: do vzorku ses dostal jen proto, že tě něco přivedlo dovnitř, a kdo nemá jednu nemoc, „potřebuje“ na vysvětlení své přítomnosti tu druhou. Dvě nezávislé nemoci se v datech nemocnice navzájem vytěsňují.

Že to nepotřebuje žádné spiknutí, si spočítáš na ubrousek. Dej cukrovce i žlučníku po jednom procentu populace, nezávisle na sobě, a pošli do nemocnice každého, kdo má aspoň jednu z nich. Mezi hospitalizovanými má pak cukrovku zhruba polovina — ale mezi hospitalizovanými *se žlučníkem* jen jedno procento: koho přivedl žlučník, ten cukrovku k vysvětlení „nepotřeboval“. Silná záporná korelace je na světě a v populaci přitom žádná není.

### ■ DEMO

lesk-a-bida-vibecodingu.gaussalgo.com/d/berkson

Berksonův paradox naživo: dvě nezávislé vlastnosti, jedno síto přijetí — a záporná korelace je na světě.

*Prameny rozuzlení: S. Hernández-Díaz, E. Schisterman & M. Hernán, „The Birth Weight Paradox: Uncovered?“, Am. J. Epidemiology 164/2006; T. Vander Weele, Int. J. Epidemiology 43/2014. Mechanismus se v literatuře jmenuje collider-stratification bias.*

*collider — uzel, do něhož míří dvě šipky kauzálního grafu. Podmínění na collideru (nebo na jeho následku) otevře mezi jeho příčinami falešnou asociaci. Skloňuje se počestně: bez collideru, na collideru.*

*Pramen: J. Berkson, „Limitations of the Application of Fourfold Table Analysis to Hospital Data“, Biometrics Bulletin 2/1946. Dnes se tomu říká Berksonův paradox.*

*Týž mechanismus mimochodem vysvětluje odvěké pozorování, že krásní bývají protivní. Nebývají. Jen jsi nikdy nešel na rande s někým, kdo nebyl ani krásný, ani milý — a tvůj vzorek známostí je proto proužek, ve kterém čím víc krásy, tím méně vlídnosti. Vesmír je nevinný; ten filtr sis napsal sám.*

Univerzity znají tutéž past pod jménem přijímací řízení. Mezi přijatými studenty výběrové školy vychází korelace mezi přijímacím testem a pozdějšími známkami slabá, někdy i záporná — a někdo z toho vyvodí, že „testy nic neměří“. Jenže přijetí je collider: vzali tě buď za skvělý test, nebo za skvělé něco jiného. Uvnitř přijatého proužku se ty dvě přednosti vytěsňují úplně stejně jako Berksonovy diagnózy.

#### ■ DEMO

lesk-a-bida-vibecodingu.gaussalgo.com/d/sat

Proč mezi přijatými přestane test korelovat se známkami: posouvej latku přijetí a dívej se, jak korelace mizí.

A medicína má i pokračování příběhu s váhou, tentokrát dospělou: mezi pacienty se srdečním selháním mívají oběžní nižší úmrtnost než štíhlí — říká se tomu paradox obezity. Zní ti to povědomě? Podmínka „být pacient“ je zase následek: štíhlý člověk s těžkou diagnózou má na ni častěji jiný, horší důvod — pokročilou nemoc, úbytek svalů, skryté onemocnění. Část paradoxu podle všeho vyrábí přesně tahle stratifikace na následku.

#### ■ DEMO

lesk-a-bida-vibecodingu.gaussalgo.com/d/bmi

Paradox obezity v datech: zapni a vypni podmínku „jen pacient“ a sleduj, jak se křivka úmrtnosti přetočí.

A čtvrtá ozvěna je celá dnešní. Velké jazykové modely, které v kapitole 15 rozebereme do šroubků, se učily na textu, který prošel sítí: na kódu, který někdo uznal za hodný zveřejnění, na odpovědích, které čtenáři odhlasovali nahoru, na projektech, jež se dožily vlastního repozitáře. Miliardy pokusů smazaných v hodině svého vzniku model nikdy neviděl — jeho korpus je flotila vrátivších se letadel. Vzpomeň si na to, až ti vygenerovaný kód bude připadat, jako by na světě neexistovaly slepé uličky: model se učil od přeživších a sebejistota, kterou z něj cítíš, je sebejistota vzorku, ne světa.

*O vysvětlení paradoxu obezity se vede spor (reverzní kauzalita, kouření, svalový úbytek); collider je jeden z hlavních podezřelých — Banack & Kaufman 2015.*

*→ kap. 5: bad požirající vlastní ocas — na čem se modely vyučily. → kap. 16: proč plynulá sebejistota není kalibrace — a co s tím dělá výcvik na potlesk.*

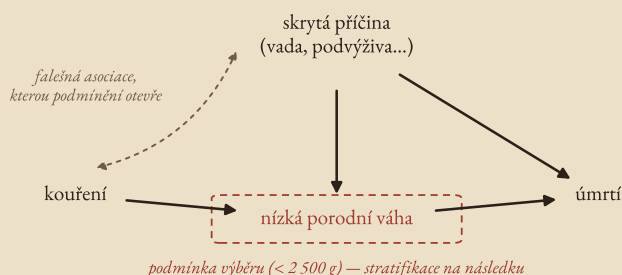
## Tužka před regresí

Co si z těch čtyř příběhů odnést do vlastní dílny? Jednu tužku a tři otázky.

Checklist datového detektiva — polož ho vedle každé tabulky, která ti dává za pravdu:

```
před_regresí(data):
1. kdo ve vzorku chybí?      # nevrácená letadla
2. co bylo podmínkou vstupu?
   → je následkem zkoumaných veličin?
3. kdo měřil a proč?        # rodný list dat
   nakresli kauzální graf    # tužkou, 5 uzlů
   stratifikuj podle příčin  # ne podle následků
```

Poslední dvě řádky checklistu jsou ta tužka. Takhle vypadá kauzální graf paradoxu porodní váhy, jak ho načmáráš na okraj zápisníku za dvacet vteřin:



Nakreslit ho *před* regresí je zákon řemesla z téže rodiny jako „testovací data jsou svatá“. Ne proto, že by tužka znala pravdu — ale protože tě donutí říct nahlas, čemu věříš: každá šipka je hypotéza o mechanismu, kterou jde napadnout, a každá podmínka výběru dostane na papíře tvar, aby bylo vidět, do čeho se strefuje. Regrese bez grafu je výpočet bez otázky.

A pravidlo, které zůstane, i kdybys všechno ostatní z checklistu zapomněl: **podmínka výběru je tichý collider**. Ty jsi možná žádnou stratifikaci nedělal — ale výběr ji udělal za tebe, mlčky, ještě než jsi data poprvé otevřel. Nemocnice, přijímačky, „vrátilo se“, „dokončil dotazník“, „zůstal zákazníkem“: každá z těch podmínek je následkem něčeho, co možná právě měříš.

Ta poslední řádka checklistu potřebuje dovysvětlit, protože **stratifikace** — rozdělení dat do vrstev a srovnávání uvnitř vrstev — je zároveň první pomoc, ne jen past. Když těžké případy tečou k lepší léčbě a lehké k horší, souhrnná čísla srovnávají přídělky pacientů místo léčeb; rozdělit data podle závažnosti je jediná záchrana. Tentýž nůž, dva směry řezu: stratifikace podle *příčiny* léčí, stratifikace podle *následku* zraňuje. A jak

poznat směr? Nijak — z tabulky to nevyčteš. Poznáš ho jedině z grafu. Proto tužka před regresí.

A třetí otázka checklistu — kdo měřil a proč — je nejméně matematická, a proto se na ni zapomíná nejsnáze. Data nepadají z nebe; vždycky je zplodila něčí otázka a něčí pohodlí. Dotazník spokojenosti vyplní jen ten, kdo je pořád zákazníkem; logy pokladny nezachytí nikoho, kdo odešel s prázdnýma rukama; „průměrná doba opravy“ se počítá z tiketů, které se někomu chtělo založit. Než začneš počítat, vyslechni rodný list dat: jaká otázka je zplodila a jaké síto je vychovalo. Bez toho výsledku čteš odpověď, aniž znáš otázku.

A teď věž. To kreslení tužkou má totiž překvapivě tvrdá pravidla: Judea Pearl a jeho škola z něj udělali kalkulu, ve kterém se otázka „smím na tomhle podmínit?“ rozhoduje mechanicky — stejným okem přečteš bombardéry, matky, nemocnice i přijímačky.

Judea Pearl — Causality (2000), Turingova  
cena 2011. V praxi ti graf zkontroluje nástroj  
DAGitty (dagitty.net): nakreslíš šipky,  
správné sady proměnných k podmínění vypíše  
sám. Kreslit ale musíš ty — nástroj hlídá  
logiku, ne znalost světa.



EINSTEIN · MATEMATIKA — přeskočitelné

**Tři atomy a jedno pravidlo.** Kauzální graf je plným jménem orientovaný acyklický graf: uzly jsou veličiny, šipka  $X \rightarrow Y$  je přímý vliv, cykly jsou zakázány. Asociace mezi veličinami teče *cestami* — posloupnostmi šipek bez ohledu na jejich směr — a každá cesta se skládá ze tří atomů:

- **řetěz**  $S \rightarrow W \rightarrow Y$ : přenáší vliv; podmíněním na prostřední uzel se ucpe.
- **vidlice**  $A \leftarrow Z \rightarrow B$ : společná příčina  $Z$  vyrábí asociaci bez kauzality; podmíněním na  $Z$  se ucpe.
- **collider**  $S \rightarrow W \leftarrow U$  (jedinkrát česky: kolizní proměnná): ucpaný sám od sebe; podmíněním na  $W$  — nebo kterémkoli jeho potomku — se otevře.

Pravidlo **d-separace**: cesta je blokována, leží-li na ní řetězový či vidlicový uzel, na němž podmiňuješ, anebo collider, na němž (ani na jeho potomcích) nepodmiňuješ. Jsou-li všechny cesty mezi  $X$  a  $Y$  blokovány, plyne z grafu, že  $X$  a  $Y$  jsou při daném podmínění nezávislé. Graf ti tak dovolí přechít, které nezávislosti čekat, dřív než sáhneš na data — a naopak: podmínění není nevinné čtení, je to zásah, který cesty otvírá a zavírá.



**Procvič na obrázku z dílny.** Vezmi graf porodní váhy:  $S \rightarrow W \leftarrow U$ , k tomu  $W \rightarrow Y$  a  $U \rightarrow Y$ . Otázka první: je kouření  $S$  nezávislé na poruše  $U$ ? Mezi nimi vedou dvě cesty,  $S \rightarrow W \leftarrow U$  a  $S \rightarrow W \rightarrow Y \leftarrow U$  — na obou sedí collider, na ničem nepodmiňuješ, obě jsou blokové. Ano, nezávislé, přesně jak jsme svět stavěli. Otázka druhá: a při podmínění na  $W$ ? Cesta  $S \rightarrow W \leftarrow U$  se otevře, nezávislost padá a mezi kouřením a poruchou vznikne záporná asociace; a protože  $U$  táhne úmrtnost, zdědí ji i vztah kouření–úmrtní. Celý třicetiletý paradox je jedno cvičení na d-separaci — jakmile máš graf. Bez grafu to bylo třicet let záhady.



**Paradox porodní váhy v číslech.** Stačí hračka o třech parametrech. Skrytá porucha  $U$  postihne 1 % dětí a lehké dítě ( $W = 1$ ) z ní vzejde s pravděpodobností 90 %; kouření  $S$  zvedá pravděpodobnost nízké váhy z jiných příčin z 5 % na 10 %;  $U$  na kouření nezávisí. Bayes pro vzorek lehkých dětí:

$$P(U = 1 \mid W = 1, S) = \frac{0,9 \cdot 0,01}{0,9 \cdot 0,01 + p_S \cdot 0,99}$$

kde  $p_S$  je pravděpodobnost nízké váhy bez poruchy: 0,05 pro nekuřáčky, 0,10 pro kuřáčky. Dosad: mezi lehkými dětmi nekuřáček má poruchu 15,4 %, mezi lehkými dětmi kuřáček 8,3 % — skoro polovina. Rozhoduje-li o úmrtí hlavně  $U$ , umírají v tomhle výběru děti kuřáček méně, přestože jim kouření nijak nepomohlo a nikde v modelu žádný ochranný účinek není. Celé kouzlo dělá jmenovatel: kouření rozšiřuje zástup „nevinne lehkých“ dětí, a tím ředí podíl poruch.



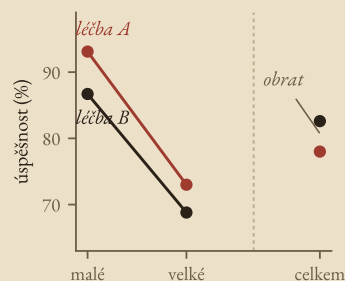
**Simpsonův obrat je aritmetika vah.** Skutečná data (Charig et al., *BMJ* 1986): 700 pacientů s ledvinovými kameny, léčba A = otevřená operace, léčba B = šetrnější odstranění vpichem.

|              | léčba A        | léčba B        |
|--------------|----------------|----------------|
| malé kameny  | 93 % (81/87)   | 87 % (234/270) |
| velké kameny | 73 % (192/263) | 69 % (55/80)   |
| celkem       | 78 % (273/350) | 83 % (289/350) |

A vede u malých kamenů i u velkých — a celkově prohrává. Tomu obratu se říká **Simpsonův paradox** a není v něm žádná magie, jen vážený průměr:

$$\frac{273}{350} = \frac{87}{350} \cdot \frac{81}{87} + \frac{263}{350} \cdot \frac{192}{263}$$

Souhrnná úspěšnost je průměr vrstev vážený tím, jaké případy léčba dostávala. Léčba A dostala tři čtvrtiny velkých, tedy těžkých kamenů; léčba B tři čtvrtiny malých. Souhrn nesrovnává léčby — srovnává léčby i s jejich přiděly pacientů, smíchané dohromady.



*Charig 1986: v obou vrstvách vede A (sanguine), v souhrnu B. Obrat nevězí v datech, ale ve vahách — kdo jaké případy dostal.*



**Který pohled tedy platí?** Tabulka ti to nikdy neřekne — rozhodne jedině graf. U kamenů je velikost *vidlice*: společná příčina volby léčby (lékaři posílali těžší případy na operaci) i výsledku. Vidlice se podmíněním ucpává, stratifikace je tu tedy správně a platí pohled po vrstvách: A je lepší. Ale představ si tutéž tabulku, kde prostřední veličinou není velikost kamene, nýbrž následek léčby — třeba „pacient dokončil celou kúru“. Stejná čísla, stejný obrat, jenže stratifikovat by teď znamenalo podmínit na collideru či přiškrtit vlastní mechanismus léčby — a platil by souhrn. Dvě identické tabulky, dva opačné správné závěry. Data sama kauzální otázku nerozhodují; rozhoduje protokol jejich vzniku.



**Proč „příhod“ do regrese všechno“ škodí.** Chceš-li z observačních dat účinek  $X$  na  $Y$ , hledáš množinu proměnných  $Z$ , na níž podmínit. Pearlovo kritérium zadních vrátek (backdoor):  $Z$  nesmí obsahovat potomky  $X$  a musí blokovat všechny cesty, které do  $X$  vstupují šipkou zvenčí. Pak platí korekční vzorec

$$P(Y \mid \text{do}(X = x)) = \sum_z P(Y \mid X = x, Z = z) \cdot P(Z = z)$$

kde  $\text{do}(X = x)$  značí zásah —  $X$  jsem *nastavil*, ne pozoroval; rozdíl mezi „vidět“ a „udělat“ je celá kauzalita v jedné závorce. Z toho kritéria plyne, proč reflex „kontrolovali jsme na všechno“ škodí: každá proměnná navíc je zásah do grafu, ne pojistka. Přidáš mediátor (uzel na cestě  $X \rightarrow Y$ ) a ucpeš přesně ten mechanismus, který jsi chtěl změřit. Přidáš collider a otevřeš falešnou cestu — porodní váha v regresi úmrtnosti je přesně tenhle hřích. Přidáš následek  $Y$  a podmínil jsi na výsledku. Víc sloupců v regresi není víc pravdy; je to víc podmínek, a každá z nich přepíná cesty v grafu.

## Tři dveře na obzoru

Čtyři příběhy, jeden zákon. Bombardéry, matky, nemocnice, přijímačky: ani jednou nebyla chyba v číslech — všechna seděla na desetinnou čárku, a kdyby sis je přepočítal, vyjdou ti znovu. Rozdíl mezi pravdou a klamem nebydlel v tabulce, ale v **protokolu vzniku dat**: kdo prošel sítí, co bylo podmínkou vstupu, kdo měřil a proč. Dvě identické tabulky s různou historií nesou různé pravdy — to je nejhlubší věta téhle kapitoly.

A je to i mostek do příští. Existuje totiž hra, v níž je celý protokol vidět najednou, v přímém přenosu: tři dveře, za jedněmi výhra, a moderátor, o němž potřebuješ vědět jedinou věc — jestli ví, co dělá, nebo jen náhodně otevírá. Tatáž otevřená vrata, dva různé protokoly, dvě různé pravděpodobnosti. Na téhle hře se v roce 1990 srazila tisícovka doktorátů s jednou novinářkou; a rozhodlo přesně to, co rozhodovalo u Walda — ne čísla, ale cesta, kterou vznikla.

A všimni si, co ti právě předvedla tahle kapitola sama o sobě: klamu přeživších podléhají i příběhy. Z dějin statistiky ti do obrazovky doplavalo to, co se nejlépe sdílí — obrázek s tečkami, ne osm memorandum; Twainův vtip, ne jeho nedoložený rodokmen. Kapitola o sítích k tobě prošla sítí. Ber to jako poslední, nepovinnou otázku checklistu: kdo vybral příběhy, kterými ses o výběru poučil?

Shrnuto: model z kapitoly 10 lhal mapou a bylo to vidět na chybě. Data téhle kapitoly lžou vzorkem — a na číslech to vidět není, protože

*Nití tří dveří: tady se křtí protokol vzniku dat  
→ kap. 14 ho předvede na třech dveřích →  
kap. 21 simulaci nasadí jako veřejnou službu  
→ část II ji postaví skutečnými nástroji.*

každé z nich je spočítané správně. Proti první lži pomáhá testovací sada; proti druhé žádný výpočet, jen tužka: kauzální graf, tři otázky checklistu a nedůvěra ke každé podmínce, která rozhodovala o vstupu do dat. Klam přeživších, Berksonova nemocnice, porodní váha i Simpsonův obrat jsou jeden mechanismus v různých kostýmech — podmínění na následku. Obrana stojí dvacet vteřin kreslení tužkou. Levnější pojistka v téhle knize není.

### „Jak bych poznal, že se pletu?“

Konkrétní test k této kapitole: až ti příště analýza potvrdí přesně to, co sis přál, vezmi tužku dřív než šampaňské. Nakresli kauzální graf; vypiš každou podmínku, kterou musel řádek splnit, aby se do dat vůbec dostal; u každé se zeptej, jestli je následkem veličin, mezi nimiž jsi našel vztah — pokud ano, máš tichý collider a závěr se může přetočit. A jestli chceš shodit celou kapitolu: vygeneruj si dva sloupce nezávislých náhodných čísel, ponech jen řádky s vysokým součtem a spočítej korelaci. Tvrdím, že vyjde záporná, přestože ji tam nikdo nevložil; vyjde-li nula, kapitola se plete. A nakonec otázka, kterou si od Walda odnes doslova: kde jsou letadla, která se nevrátila?