Rozdział 2

Kłamstwo dla początkujących

Przyjrzyjmy się teraz podstawom posługiwania się danymi w mylący sposób. Pewnie niektóre opisane tu sposoby są dość oczywiste, ale nie zaszkodzi przypomnieć, co może pójść nie tak, gdy ludzie posługują się danymi na poparcie swoich argumentów.

ZACIERANIE GRANIC MIĘDZY KORELACJĄ A ZWIĄZKIEM PRZYCZYNOWYM

Jeżeli dwa zdarzenia (nazwijmy je *A* i *B*) są ze sobą skorelowane, to znaczy, że z czasem zmiany ich wartości zachodzą w tym samym kierunku. Jeżeli rośnie *A*, rośnie też *B*, a jeżeli *A* maleje, maleje też *B*. Istnieją cztery możliwe wyjaśnienia tej prawidłowości:

1. *A* powoduje *B*.
2. *B* powoduje *A*.
3. Trzeci czynnik, *C*, powoduje *A* i *B*.
4. Między *A* i *B* nie występuje żadna zależność przyczynowa.

Tak naprawdę to nie jest wyczerpująca lista, istnieją również bardziej skomplikowane możliwości, takie jak „*A* powoduje *C*, a *C* powoduje *B*". Na potrzeby naszego przykładu wystarczą jednak cztery najprostsze kategorie.

Czwarte wyjaśnienie występuje na przykład w sytuacji, w której obserwacja nastąpiła w krótkim okresie i korelacja *A* i *B* wynika ze zwykłego przypadku. Pierwszy wniosek jest zatem taki, że jeśli analizujemy dane z krótkiego okresu, w ogóle nie ma sensu zaczynać szukać związków przyczynowych.

Zawsze warto natomiast zastanowić się nad trzecim wyjaśnieniem. Weźmy na przykład stwierdzenie, że łysienie powoduje choroby naczyniowe. Te dwa zjawiska mogą być ze sobą skorelowane, ale istnieje znacznie większe prawdopodobieństwo, że oba są powodowane przez trzeci czynnik, czyli starość, która znacząco zwiększa ryzyko występowania obu.

Ogólnie wygląda to tak, że gdy stwierdzimy występowanie korelacji między zdarzeniami *A* i *B*, jesteśmy skłonni zakładać, że mamy do czynienia z wyjaśnieniem pierwszym lub drugim. *Spurious Correlations*, znakomita książka i strona internetowa Tylera Vigena, opisuje zabawne przypadki par zjawisk, które nie mogą być powiązane przyczynowo, a mimo to są skorelowane. Na przykład w latach 1999–2008 wiek pań wybieranych jako Miss America wykazywał dość bliską korelację z morderstwami za pomocą pary wodnej, oparów i gorących przedmiotów. W podobnym okresie amerykański import ropy naftowej z Norwegii wykazywał korelację z liczbą kierowców, którzy zginęli w wypadkach z udziałem pociągów.

W tego rodzaju przypadkach na pierwszy rzut oka widać, że korelacja ma charakter przypadkowy, ponieważ trudno wyobrazić sobie jakikolwiek związek między tymi zdarzeniami. Mózg człowieka jest naturalnie zaprogramowany na poszukiwanie prawidłowości,

więc gdy mamy do czynienia z dwoma zjawiskami, w przypadku których można wyobrazić sobie zależność przyczynową, bardzo chętnie zakładamy jej występowanie.

W jaki sposób kłamcy mogą wykorzystywać tę naturalną dla człowieka skłonność? Weźmy na przykład słynną reklamę papierosów Camel z 1946 roku, w której zapewniano, że „Camele to najpopularniejsze papierosy wśród lekarzy". Były to lata, kiedy zaczynaliśmy rozumieć, że palenie może być dla nas niezbyt korzystne, a zatem autorzy tej reklamy postanowili wykorzystać powszechne przekonanie, że lekarze wiedzą, co jest zdrowe, więc palą najzdrowsze papierosy. Tym samym reklama sugerowała występowanie zależności przyczynowej między „wiedzą o tym, co jest zdrowe" a „paleniem papierosów Camel".

Zacznijmy od tego, że nie ma żadnych powodów, by sądzić, że lekarze jako eksperci od zdrowia sami dokonują zdrowych wyborów. Być może prawdą jest, że lekarze częściej sięgają po amfetaminę, niż wynosi średnia dla społeczeństwa, ale jeśli rzeczywiście tak jest, najprawdopodobniej wynika to z faktu, że ich praca jest obciążająca i stresująca, a nie z przekonania, że amfetamina jest zdrowa.

Tak się składa, że w tym przypadku podstęp miał charakter dwuwymiarowy. Badania, na których opierało się to zapewnienie, zostały przeprowadzone przez William Esty & Co., agencję reklamową zatrudnioną przez koncern tytoniowy R.J. Reynolds. Zastosowana metodologia nie należała do najrzetelniejszych: pracownicy agencji rozmawiali z lekarzami na konferencjach medycznych albo w ich gabinetach, a każdy wywiad zaczynali od wręczenia im kartonu Cameli.

Gdyby nawet jednak stwierdzenie opierało się na rzetelnych badaniach, nie byłoby żadnych powodów, żeby sądzić, że w tym przypadku korelacja oznacza związek przyczynowy.

Zapamiętaj: Gdy ktoś przedstawia ci korelację, szukaj czynników zakłócających i nie zakładaj od razu istnienia zależności przyczynowej.

UDAWAĆ, ŻE ZWIĄZEK PRZYCZYNOWY WYSTĘPUJE STALE

Wyobraź sobie, że pracujesz w firmie produkującej roboty kuchenne. Jednego roku dyrektor handlowy proponuje, żeby firma wyprodukowała ekspresy do kawy z górnej półki. Produkt spotyka się ze znakomitym przyjęciem rynku, w pierwszym roku obroty firmy rosną o 100 procent, a w roku następnym o kolejne 100 procent. W związku z tym dyrektor handlowy przygotowuje prezentację na zebranie ogólnoorganizacyjne. Stwierdza, że zwiększenie produkcji ekspresów do kawy przełożyło się na wzrost zysków i pokazuje wykres, na którym prognozowany wzrost obrotów w dwóch kolejnych latach nadal wynosi 100 procent.

Podejrzewam, że podchodziłbyś do tego dość sceptycznie – i bardzo słusznie. Zmiana produkowanych wyrobów mogła poskutkować wzrostem obrotów, ale prędzej czy później początkowy rynek docelowy zostanie nasycony, a wtedy firmie będzie już trudniej. W końcu zareagują też konkurenci i wprowadzą na rynek własne podobne produkty. Przykładem niech będzie sytuacja firmy Nintendo w latach 90. XX wieku. W tamtym okresie dzięki grze *Super Mario* osiągnęła dominującą pozycję na rynku, w Stanach Zjednoczonych Nintendo miała ponad 90 procent rynku gier wideo. Wystarczyło jednak kilka lat, aby pozycję rynkowego lidera odebrała jej Sega, producent między innymi gry *Sonic the Hedgehog*. Żadna firma nie jest w stanie zagwarantować, że jej sukcesy będą trwać, zwłaszcza jeśli ma tendencję do spoczywania na laurach.

Podobne zjawisko da się zaobserwować w środowisku naturalnym. Urodzaj jagód w lesie może przełożyć się na wzrost liczebności owadów, a to może spowodować wzrost liczebności tych gatunków ptaków, które się nimi żywią. Rozrastająca się populacja napotyka jednak na różne trudności, gdyby na przykład pojawiła się jakaś choroba, może się w dużej populacji bardziej rozwinąć, duża liczebność ptaków może też ściągnąć do lasu drapieżniki, które się nimi żywią. Pośrednia zależność przyczynowa między urodzajem jagód a liczebnością populacji ptaków naprawdę występuje, ale nie ma podstaw, żeby sądzić, że będzie trwać wiecznie.

Na koniec rozpatrzmy zależność między poziomem dochodów a szczęściem. Gdy rosną dochody osoby ubogiej, zwykle ma to istotny wpływ na jej styl życia i ogólnie pojęty poziom szczęścia. Przychodzi jednak taki moment, w którym osiągane dochody pozwalają na wygodne życie. Od tego momentu każdy wzrost dochodów podlega **prawu malejących korzyści krańcowych**. Znów zatem zależność przyczynowa występuje w pewnym zakresie, a dalej słabnie albo zanika.

Zapamiętaj: Korelacja może wynikać z przyczynowości, ale nawet gdy tak jest, nie można liczyć na to, że przyczynowość jest zależnością stałą.

BRAK DEFINICJI ZASTOSOWANEJ ŚREDNIEJ

Jeśli dysponujesz zbiorem danym, masz przynajmniej trzy sposoby na obliczenie średniej. Prawdopodobnie uczyłeś się tego w szkole, ale i tak przypomnę tutaj podstawowe wiadomości na ten temat.

Chcąc obliczyć średnią arytmetyczną, dodajesz wartości i dzielisz je przez liczbę dodanych wartości. Chcąc obliczyć medianę,

porządkujesz wartości od najmniejszej do największej i wybierasz wartość środkową. Chcąc obliczyć dominantę, znajdujesz tę wartość, która występuje w zbiorze danych najczęściej.

Poniżej przedstawiam wysokość dochodów 15 osób zamieszkujących przy tej samej, wyjątkowo zróżnicowanej zarobkowo ulicy:

1 000 000 funtów
150 000 funtów
50 000 funtów
50 000 funtów
40 000 funtów
40 000 funtów
40 000 funtów
20 000 funtów
20 000 funtów
19 000 funtów
18 000 funtów
17 000 funtów
16 000 funtów
15 000 funtów
5000 funtów

Kwoty te sumują się do 1 500 000 funtów, więc średnia arytmetyczna dochodów mieszkańców tej ulicy wynosi 100 000 funtów. Mediana wynosi 20 000 funtów, a dominanta – 40 000 funtów. Potencjalny kłamca ma tutaj zatem szerokie pole do popisu, ponieważ może wykorzystać w swojej argumentacji różne średnie i sam decyduje o tym, którą z nich się posłuży. Jeśli ktoś chce wyolbrzymić zamożność tej okolicy, może opowiadać głośno o tym, ile wynosi średnia arytmetyczna dochodów. Jeżeli chce tę zamożność zaniżyć, posłuży się medianą. Gdyby jednak żadna z tych wartości nie odpowiadała potrzebom naszego hipotetycznego kłamcy, zawsze

może sięgnąć po dominantę, która w tym przypadku daje mu wartość pośrednią.

Tak się składa, że często najbardziej reprezentatywną formą średniej jest mediana, a to głównie dlatego że dochody rozkładają się zwykle bardzo nierówno. Osób zarabiających mało jest stosunkowo dużo, a osób o najwyższych zarobkach jest garstka. Średnia arytmetyczna i mediana będą takie same lub choćby do siebie zbliżone, gdy zostały obliczone ze zbioru danych podlegającego rozkładowi normalnemu (czyli takiemu, który na wykresie tworzy krzywą dzwonową). Im bardziej wykres jest skośny po jednej lub drugiej stronie, tym większa będzie różnica między medianą a średnią.

Przydaje się również dominanta. Wyobraź sobie, że prezentujesz grupie ludzi prototyp noszonego namiotu (czyli po prostu płaszcza, z którego można zrobić namiot). Pytasz ich, ile powinna wynosić rozsądna cena tego produktu, a oni podają następujące odpowiedzi:

50 funtów
40 funtów
35 funtów
30 funtów
25 funtów
25 funtów
20 funtów
20 funtów
20 funtów
20 funtów
20 funtów

Mediana wynosi w tym przypadku 25 funtów, ale niewykluczone, że maksymalizacji sprzedaży gotowego produktu lepiej służyłaby cena w wysokości 20 funtów (albo nawet 19,99, gdybyś zdecydował się na zastosowanie starego tricku marketingowego,

mającego zwodzić konsumentów). Taka cena pozwala przypuszczać, że dla wszystkich uczestników badania cena gotowego produktu będzie rozsądna. Gdyby zastosować medianę, uważałaby tak tylko połowa ankietowanych.

Średnia arytmetyczna, mediana i dominanta mają swoje zastosowania. Tak naprawdę jednak trzeba pamiętać, że potocznie rozumiana „średnia" jest pojęciem ogólnym i można za jej pomocą ukryć wiele nieciekawych rzeczy.

Zapamiętaj: Gdy ktoś mówi o średniej, sprawdź, którą konkretnie miarę ma na myśli. Najlepiej, gdybyś po prostu zapoznał się z pełnym zbiorem danych.

STATYSTYCZNY CZŁOWIEK MA MNIEJ NIŻ DWOJE OCZU

Gdyby argumenty dotyczące średniej do kogoś nie dotarły, niech zapamięta ten prosty fakt. Żaden człowiek nie ma więcej niż dwojga oczu, a niektórzy mają ich mniej, więc statystyczny człowiek ma mniej niż dwoje oczu (średnia arytmetyczna prawdopodobnie wynosi w tym przypadku coś koło 1,9999). Oznacza to również, że niemal 100 procent ludzi ma więcej oczu, niż wynosi średnia.

To kolejny ciekawy przykład, który pokazuje, dlaczego na średnią należy uważać i dlaczego różne postacie średniej są przydatne w różnych sytuacjach. Mediana i dominanta liczby oczu niemal na pewno wynoszą w tym przykładzie dwa, więc byłyby zdecydowanie sensowniejszymi wersjami „średniej" niż średnia arytmetyczna.

Każdy przypadek, w którym dane podlegają rozkładowi innemu niż normalny, powoduje zamieszanie ze średnimi. W 2004 roku administracja prezydenta Busha zachwalała swoją obniżkę podat-

ków stwierdzeniem, że przeciętna amerykańska rodzina zaoszczędzi dzięki niej 1586 dolarów rocznie. Teoretycznie była to prawda, jeśli posługiwać się średnią arytmetyczną. Zarobki w społeczeństwie rozkładają się jednak bardzo nierówno. Niewielki odsetek gospodarstw domowych o bardzo wysokich dochodach oznaczał, że 98 procent[4] amerykańskich rodzin zaoszczędziła dzięki reformie podatkowej mniej niż 650 dolarów.

Wyobraźmy sobie szkołę, która specjalizuje się w koszykówce i w związku z tym ma stosunkowo dużo wysokich uczniów (oraz niewielką liczbę uczniów dość niskich). W takiej szkole ponad 50 procent uczniów może być wyższych niż średnia arytmetyczna wzrostu uczniów szkoły. Jeśli interesuje nas podział tych uczniów na połowy, powinniśmy posłużyć się medianą.

Zapamiętaj: Statystyczny człowiek nie istnieje. Większość członków danej populacji może znajdować się powyżej lub poniżej średniej.

KIEDY 50 PROCENT NIE OZNACZA 50 PROCENT?

Pamiętaj, że wartości procentowe potrafią być złudne, zwłaszcza gdy przedstawia się je bez kontekstu. W Wielkiej Brytanii w latach 2010–2019 siły policyjne stopniały ze 140 tysięcy funkcjonariuszy do 120 tysięcy funkcjonariuszy (dla uproszczenia podaję wartości

4 Te 98 procent zmyśliłem – źródło, z którego korzystałem, pisząc ten fragment, mówiło „o niemal wszystkich rodzinach", więc dobrałem sobie taki odsetek, żeby informacja wydawała się bardziej dramatyczna. Widzisz, jakie to proste? Tego rodzaju liczba z kapelusza bywa nazywana liczbą potiomkinowską, o której pisałem na stronie 3.

w zaokrągleniu). Oznacza to spadek o 14,3 procent (20 tysięcy podzielone przez 140 tysięcy). Gdybyśmy chcieli wrócić na poziom z 2010 roku, musielibyśmy zwiększyć zatrudnienie w policji o 16,6 procent (20 tysięcy podzielone przez 120 tysięcy).

Procentowy spadek nie jest zatem tak samo duży jak procentowy wzrost, ponieważ w tym drugim przypadku zmiana jest liczona od mniejszej wartości podstawowej. Mogą z tego wynikać naprawdę duże różnice. Na przykład 50-procentowa obniżka ceny, po której nastąpi wzrost ceny o 100 procent, będzie oznaczała, że cena końcowa jest dokładnie taka sama jak cena wyjściowa.

To samo zjawisko może wprowadzać w błąd na nieco różne sposoby. Wyobraźmy sobie trzech pracowników tej samej firmy. Prezes zarabia 500 tysięcy funtów, kierownik zarabia 50 tysięcy funtów, a sprzątacz zarabia 10 tysięcy funtów. Pod koniec roku ogłasza hojnie, że przyznaje 10-procentową podwyżkę sprzątaczowi, 2-procentową podwyżkę kierownikowi, a sam ograniczy się do podwyżki w wysokości 0,2 procent.

Indukujący dodo

Bardziej przemyślny, szesnastowieczny dodo mógłby się rozejrzeć po swoim rodzinnym Mauritiusie i wyciągnąć pewne wnioski co do tego, jak będzie wyglądać przyszłość. Na wyspie występowały wówczas tylko ptaki, owady i zwierzęta morskie. Jedyne ssaki, jakie docierały na tę odległą od lądu wyspę, to nietoperze i ssaki morskie.

Dodo miał pełne prawo dojść do wniosku, że jego gatunek nie ma naturalnych wrogów i że jego przyszłość jest bezpieczna. Gdyby był to dodo z zamiłowaniem do arkuszy kalkulacyjnych i prowadził bazę danych populacji przedstawicieli

swojego gatunku, mógłby prognozować stały wzrost populacji w kolejnych dekadach. Prawdziwy koniec tej historii wszyscy znamy: pod koniec XVI wieku wyspę skolonizowali Holendrzy. Wraz z ludźmi na wyspę przybyły szczury i koty, a to spowodowało błyskawiczne wyginięcie ptaka nielota.

To przykład zagrożeń związanych z indukcyjnym podejściem do myślenia. Najprościej rzecz ujmując, różnica między dedukcją a indukcją polega na tym, że w dedukcji wychodzi się od ogólnej zasady i przechodzi do przypadku szczególnego, natomiast w przypadku myślenia indukcyjnego zaczyna się od przypadku szczególnego i formułuje na tej podstawie zasadę ogólną.

Klasycznym przykładem, spopularyzowanym przez Nassima Nicholasa Taleba, jest czarny łabędź. Anglik żyjący w tych samych czasach co nasz dodo z powyższej opowieści mógłby spokojnie sformułować założenie, że wszystkie łabędzie są białe. Dopiero w XVII wieku, wraz z kolonizacją Australii, okazało się, że na świecie żyją również czarne łabędzie. Taleb porównuje to z chaosem, jaki nastąpił po wybuchu globalnego kryzysu finansowego. Wielu analityków i pracowników sektora finansowego posługiwało się modelami, które nie przewidywały możliwości trwałej zapaści cen na rynku nieruchomości. W rezultacie wystąpiło zjawisko „bezpodstawnego optymizmu", po którym na rynku nieruchomości mieszkalnych nastąpił krach o olbrzymich konsekwencjach.

Co oczywiste, indukcja prowadzi często do zupełnie rozsądnych wniosków. Wierzę, że jutro rano wzejdzie słońce, ponieważ dotychczas wschodziło codziennie. Nie mam całkowitej pewności, że dalej będzie wschodzić – że planeta nagle nie przestanie się obracać, a mój dom utknie po jej ciemnej

stronie – ale można przyjąć, że to założenie jest rozsądne. Należy po prostu zapamiętać, że indukowanie pozwala nam postrzegać świat w kategoriach probabilistycznych, przez co nigdy nie mamy całkowitej pewności co do przyszłości.

Po pracy wszyscy wiwatują na cześć hojnego prezesa, piją za jego zdrowie… aż w końcu sprzątacz dokonuje szybkich obliczeń na podstawce pod kufel i wychodzi mu, że wszyscy trzej dostali po 1000 funtów podwyżki.

Gdyby prezes ogłosił, że podnosi pensje o 10 procent, on dostałby 50 000 funtów więcej, kierownik dostałby dodatkowe 5000 funtów, a sprzątacz – 1000 funtów. Prezes mógł wybrać odpowiedni sposób prezentacji informacji w zależności od tego, czy chciał być postrzegany jako przyzwoity człowiek, czy też chciał maksymalizować swoją podwyżkę.

Zapamiętaj: Wartości procentowe same w sobie niewiele znaczą.