JAK BŁĘDY I PARADOKSY MOGĄ ZAPEWNIĆ MIEJSCE W ANNAŁACH NAUKI? PIONIERSKIE DOKONANIA JOSEPHA BERKSONA

Katarzyna Ostasiewicz
Uniwersytet Ekonomiczny we Wrocławiu, Polska
e-mail: katarzyna.ostasiewicz@ue.wroc.pl
ORCID: 0000-0002-0115-3696

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 18(24)

ISSN 1644-6739
e-ISSN 2449-9765

Quote as: Ostasiewicz, K. (2020). Jak błędy i paradoksy mogą zapewnić miejsce w annałach nauki? Pionierskie dokonania Josepha Berksona. *Śląski Przegląd Statystyczny, 18*(24).

DOI: 10.15611/sps.2020.18.23

Współcześni lekarze częstokroć współpracują ze statystykami. Bez tej współpracy niemożliwa byłaby epidemiologia, szukająca zależności pomiędzy różnymi czynnikami zewnętrznymi a zapadalnością na choroby. Niemożliwe byłoby również testowanie nowych leków. Jesteśmy tak przyzwyczajeni do standardowych obecnie procedur – podwójnie ślepych prób i statystycznej analizy wyników – że łatwo zapominamy, iż aby coś stało się rutyną i oczywistością, ktoś wcześniej musiał włożyć duży wysiłek, by przekonać środowisko do tego, co, nim stało się rutyną, było nowością, a być może i rewolucją.

Zasadnicze zręby paradygmatu obowiązującego na początku XXI wieku w medycynie opartej na dowodach naukowych (EBM – *Evidence Based Medicine*) wykuwane były od ponad stu lat przez takich tytanów, jak Ignaz Semmelweis, Robert Koch czy Ronald Fisher. Mniej znane jest nazwisko wieloletniego szefa Sekcji Biometrii i Statystyki Medycznej kliniki Mayo[1], którego wkład w dyskusję nad metodologią badań klinicznych i epidemiologicznych zasługuje za przypomnienie.

[1] Słynna klinika, istniejąca od połowy XIX wieku, powszechnie uważana za jeden z najlepszych – bądź wręcz najlepszy – amerykański szpital.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

Indywidualista, który nie tolerował głupców

Joseph Berkson miał szczególnie solidne podstawy do pracy na styku medycyny i statystyki. Uzyskawszy tytuł naukowy z fizyki na Uniwersytecie Columbia w 1922 roku, popełnił kolejno dwa doktoraty – z medycyny, w 1927 roku, i ze statystyki, w 1928 roku, oba na Uniwersytecie Johnsa Hopkinsa. Większość życia zawodowego – ponad trzydzieści lat – spędził jednak w klinice Mayo. W obecnych czasach wąskich specjalizacji podziw budzi polihistor, który z równym zaangażowaniem brał udział zarówno w rozwoju czysto teoretycznej statystyki matematycznej, jak i w dyskusji nad metodyką badań klinicznych i konkretnych zagadnień epidemiologicznych. Wśród współpracowników znany był jako indywidualista z trudem tolerujący głupców, „maszerujący w takt własnego bębna" (Armitage i Colton, 1998). Bardzo żywiołowo bronił swoich racji i angażował się w zaciekłe polemiki na łamach prasy naukowej. Metoda największej wiarygodności czy najmniejszych kwadratów? U mało kogo ten dylemat budzi takie emocje, jakie rodził u Berksona. On zaciekle bronił najmniejszych kwadratów – wręcz „lobbował" za nimi. Tak samo z ogromną energią napędzał dyskusję nad kolejną alternatywą, mało istotną dla większości współczesnych użytkowników statystyki: model logitowy czy probitowy? Nieświadomi ogromnej batalii, jaka rozegrała się z udziałem naszego bohatera, obecnie zwykle wybieramy jeden czy drugi, albo oba równolegle, nie bacząc na to, że duch Berksona zżyma się nad taką beztroską. Ten temperament i – wydawałoby się – niespożyta energia odziedziczone zostały być może po wschodnich przodkach. Berkson był bowiem szóstym dzieckiem rosyjskich imigrantów, którzy na nowym kontynencie przyjęli bardziej angielsko brzmiącą wersję oryginalnego nazwiska – Berkman. Żywiołowość i intelektualne nienasycenie pozwoliły Berksonowi zaznaczyć swoją obecność w wielu dziedzinach badań naukowych, i nie tylko *stricte* naukowych. Przykładowo w czasie drugiej wojny światowej, służąc w stopniu pułkownika, odniósł takie sukcesy w obszarze prewencji epidemiologicznej wojsk amerykańskich, że zasłużył na odznaczenie Legią Zasługi. Odznaczenie niby wojskowe, ale – koniec końców – będące efektem pasji do nauki.

Nie sposób omówić wszystkich wątków prac badawczych Berskona. Jeszcze przed trzydziestką miał na swoim koncie publikacje we współautorstwie z Louisem Flexnerem (biochemikiem, badaczem syntezy białek w mózgu i ich roli w uczeniu się) czy Lowellem Reedem (twórcą statystycznej metody oceniania dawki skutecznej medialnej i modelu Reeda-Frosta). Inny wątek, który w tamtych latach zajmował Berksona, jest obecnie zupełnie anachroniczny. Dzięki szybkim komputerom, o których przed II wojną jeszcze nie śniło się nie tylko filozofom, probabilistyczne nomogramy, opracowywane skrupulatnie jako niezbędne podówczas

narzędzia obliczeniowe, nie tylko przestały być potrzebne, ale wydaje się, że mało kto potrafiłby się nimi współcześnie posługiwać. Całe szczęście, że prace te nie stanowiły *clou* dorobku Berksona, bo musiałby być ogromnie sfrustrowany, śledząc w ciągu swego długiego życia, jak odchodzą one do lamusa. W roku śmierci Berksona, 1982, IBM wprowadził pecety, Microsoft system operacyjny MS-DOS, Intel 16-bitowy procesor taktowany zegarem do 25 MHz, komputer został maszyną roku tygodnika „Time", a możliwości obliczeniowe pikowały w górę. Chociaż tak właściwie Berkson może i był zachwycony rozwojem mechanizacji obliczeń. W końcu w klinice Mayo to on właśnie był pionierem wprowadzenia do rejestracji i klasyfikacji chorób i pacjentów dziurkowanych kart Holleritha (których był również protoplastą IBM-a). Nie brakowało mu przy tym bynajmniej pewności siebie – odrzucił oficjalną klasyfikację przyczyn śmierci (ICD), akceptowaną przez Amerykańskie Towarzystwo Medyczne, i wprowadził własną (Kurland i Molgaard, 1981). W każdym razie dorobek Berksona był na tyle obfity i różnorodny, że fakt, iż nomogramy nie przetrwały próby czasu, niemal niezauważalnie uszczuplił to, co wciąż jest doniosłe.

Przyjrzyjmy się bliżej kilku zaledwie z frontów, na których prowadził potyczki Berkson. Część z bitew, jakie stoczył, jest już w dużej mierze zapomniana. Z wielką szkodą dla rozwoju naukowego nowych pokoleń statystyków, bo kontrowersje i problemy w nich poruszane były i są najwyższej wagi. Paradoksalnie (i pechowo) bitwa, która została Berksonowi najlepiej zapamiętana, została przez niego niechlubnie przegrana – i chyba niesłusznie pogrążyła jego nazwisko w lekkiej niesławie, rzucając niekiedy cień na uczciwość badawczą zarówno jego samego, jak i innych naukowców po tej samej stronie barykady.

Ale po kolei.

Funkcja logistyczna

Pod koniec XVIII wieku Thomas Malthus wywołał panikę, przepowiadając to, co dziś nazywamy „bombą populacyjną". Jeśli każda para ludzi (czy jakiegokolwiek innego organizmu) będzie miała liczbę potomstwa większą niż dwa (licząc potomstwo przeżywające do takiego wieku, by również móc się zreprodukować), proste rachunki wskazują, iż populacja będzie przyrastała w postępie geometrycznym. Formalnie równanie opisujące zmianę liczebności populacji w czasie (oznaczonej przez $N(t)$) przybiera postać:

$$\frac{dN}{dt} = \alpha N(t),$$

gdzie α jest (dodatnim) współczynnikiem reprodukcji.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

Rozwiązaniem tego równania różniczkowego jest funkcja:

$$N(t) = N_0 e^{\alpha t},$$

z N_0 jako wartością początkową – i bombę (populacyjną) mamy gotową.

Sławny belgijski statystyk Adolphe Quételet zauważył, że taki model jest nieco nierealistyczny. Nim jakakolwiek populacja osiągnie takie zagęszczenie, że osobniki będą musiały wchodzić sobie nawzajem na głowy, by pomieścić się w dostępnym habitacie, wzrośnie śmiertelność i zmaleje rozrodczość – choćby z przyczyny braku pożywienia. Quételet poprosił swego ucznia, Pierre'a Françoisa Verhulsta, o opracowanie bardziej adekwatnego modelu. Szczęśliwie się złożyło, że ten ostatni został właśnie wydalony z Państwa Kościelnego po nieudanych próbach namówienia papieża do wprowadzenia reform i konstytucji, mógł zatem zająć się mniej istotnymi dla ludzkości sprawami. Efektem jego namysłu była następująca prosta propozycja. Każdy habitat ma jakąś określoną ograniczoną pojemność dla danego gatunku – $N_{\max}$ (współcześnie nazywaną po angielsku *carrying capacity*), a im bardziej wielkość populacji zbliża się do tego pułapu, tym bardziej spowalnia jej wzrost. Zatem równanie różniczkowe opisujące dynamikę liczebności populacji zostałoby zmodyfikowane do postaci:

$$\frac{dN}{dt} = \beta N(t)[N_{\max} - N(t)],$$

która ma wbudowany hamulec bezpieczeństwa – gdy $N(t)$ zbliża się do $N_{\max}$, tempo wzrostu maleje do zera, zatem liczebność nigdy nie przekroczy górnej granicy.

Rozwiązaniem tego równania różniczkowego jest następująca funkcja:

$$\frac{N(t)}{N_{\max}} = \frac{\exp[\gamma + \beta t]}{1 + \exp[\gamma + \beta t]}$$

ze współczynnikiem γ zależnym od wartości początkowej wielkości populacji.

Funkcję tę Verhulst ochrzcił *funkcją logistyczną*.

Niemal wiek później ta sama funkcja została „odkryta" ponownie przez dwóch naukowców – Raymonda Pearla oraz Lowella J. Reeda – nieznających uprzednio prac Verhulsta. Ten ostatni, matematyk z zacięciem do stosowania matematyki w biologii, kilka lat później rozszerzył użycie funkcji logistycznej z modelowania wzrostu populacji na opis reakcji autokatalitycznych. Pracę na ten temat napisał z innym naukowcem, interesującym się zarówno matematyką, jak i naukami o życiu – z Josephem Berksonem.

Raz znalazłszy się w królestwie rządzonym przez funkcję logistyczną, stał się Berkson jej najwierniejszym akolitą i zbrojnym ramieniem, wyruszając na podbój nowych terenów. W szczególności zajadle wręcz lobbował za zastąpieniem modelu probitowego modelem logitowym.

Model probitowy narodził się i rozwijał w pierwszej połowie XX wieku na gruncie nauk o życiu. Rodziców miał wielu, ojcem chrzestnym zaś został Charles Ittner Bliss, który nazwę ukuł jako zbitkę słowną pochodzącą od określenia ***prob**ability un**it*** („jednostka prawdopodobieństwa"). Istoty biologiczne – rozumowali twórcy modelu – nie reagują jednakowo na określone bodźce, ale podlegają losowości. Jeśli do próbki określonego materiału przyłożymy określone napięcie, prąd popłynie, lub nie, za każdym razem tak samo – to charakterystyka materiału. Jeśli jednak istotę ludzką wystawimy na działanie określonych czynników chorobotwórczych, jednym razem zachoruje, innym nie. Istoty ludzkie, inaczej niż próbki materiału, charakteryzują się nieprzewidywalną zmiennością. Aby modelować „odpowiedź" jednostek biologicznych na różnego typu czynniki, uwzględniano zatem zarówno wielkość bodźca, jak i składnik losowy, opisywany rozkładem normalnym. Dodatkowo, co jest już bardziej techniczną kwestią, parametry modelu estymowane były metodą największej wiarygodności.

Joseph Berkson jako pierwszy zaproponował (Berkson, 1944) w latach czterdziestych XX wieku, i zaciekle bronił przez kilka kolejnych dekad, zastąpienie modelu probitowego modelem *logitowym* – odpowiednikiem tego pierwszego, z rozkładem normalnym podmienionym na rozkład logistyczny. Początkowo propozycja wywołała kpiny, włącznie z samą nazwą modelu, ale szybko okazało się, że przeciwnika nie da się tak łatwo zignorować. Rozkład logistyczny jest na oko praktycznie nierozróżnialny od rozkładu normalnego, za to jego użycie, ze względu na istnienie prostej postaci analitycznej dystrybuanty, było znacznie łatwiejsze. Rzecz nie do przecenienia w czasach, gdy komputery były rzadkością, a nawet jeśli były dostępne, ich moce obliczeniowe były z obecnego punktu widzenia śmiechu warte. Dochodziło zatem do sytuacji, gdy ci sami uczeni, którzy na seminariach przerzucali się argumentami na rzecz modelu probitowego, w zaciszach własnych gabinetów posługiwali się modelem logitowym.

A jakie to były argumenty?

Głównie z „naturalności". Historycznie nominalna „normalność" rozkładu normalnego wcale nie oznacza „zwykłości", ale pochodzi od bardziej skomplikowanej matematycznej własności ortogonalności płaszczyzny związanej z pewnym układem równań, zawierającym ten rozkład prawdopodobieństwa. Niemniej dość szybko rozkład normalny został – i wciąż jest – uznawany za rozkład wielkości „naturalnych". Wedle tego rozumowania na przykład rozkład płac, jako rezultat działań człowieka, z pewnością nie jest normalny, ale ilorazu inteligencji – ukształtowanego przez Naturę – już tak.

„Wybierać jedną funkcję zamiast innej, z powodu mniejszej ilości pracy a nie zważając na jej zgodność z rozważaniami teoretycznymi wydaje się bardzo chybioną praktyką" (Fisher, 1954a) – grzmiał Ronald

Fisher w 1954 roku, dodając, że w ciągu poprzednich 15 lat wielokrotnie używał rozkładu normalnego i nieprawdziwe jest stwierdzenie Berksona, iż użycie funkcji logistycznej skraca obliczenia ponadtrzydziestokrotnie. „Zgodziłem się skomentować uwagi sir Ronalda Fishera po długim wahaniu", ripostował Berkson w tym samym numerze czasopisma. „Akapity jego artykułu odnoszące się do mojej pracy są tak nieadekwatne, że czytając je po raz pierwszy uznałem, iż został błędnie poinformowany na temat moich stwierdzeń" (Fisher, 1954b). W dalszym ciągu tekstu Berkson zaproponował, by Fisher wycofał swoje stwierdzenia w odniesieniu do stwierdzeń tego pierwszego, ale spotkało się to ze zdecydowaną odmową. To oznaczało wojnę. Następiły szczegółowe wyliczenia przykładowego czasu potrzebnego na uzyskanie estymacji parametrów modelu probitowego w porównaniu z modelem logitowym. Berkson zajadle bronił też metody χ^2 jako lepszej od metody największej wiarygodności, wskazując, że ma ona solidne oparcie na gruncie teoretycznym. Bo Fisher przeoczył przecież, że χ^2 było inherentną składową modelu probitowego, lansowanego przez Berksona! Berkson był tak zdegustowany tym niedopatrzeniem, że aż tytuł jednego ze swoich artykułów zakończył wykrzyknikiem: „Minimum chi-kwadrat, a nie maksimum wiarygodności!" (Berkson, 1980).

Jak wiele innych zajadłych sporów z przeszłości, tak i ta – wraz z adwersarzami – odeszła do lamusa, a współcześnie modele probitowe i logitowe nie budzą już takich emocji, podobnie jak metoda chi-kwadrat i największej wiarygodności. Paradoksalnie, choć jeden z głównych argumentów – oszczędność czasu obliczeniowego – wobec rozwoju komputerów obecnie stracił właściwie rację bytu, to kwerenda prac z dziedziny statystyki pokazuje, że od kiedy w latach 70. ubiegłego wieku model logitowy zyskał przewagę nad probitowym, przewagę tę, ustaloną pod koniec wieku na około dwukrotną, wciąż utrzymuje.

Pamiętając o wykształceniu i miejscu pracy Josepha Berksona, nie można się dziwić, że jego zainteresowania – nawet jako statystyka – zwracały się ku zastosowaniom w naukach biologicznych. Model probitowy/logitowy używany był częstokroć do obliczania tak zwanych dawek śmiertelnych. Ponieważ nie każdy organizm reaguje identycznie na te same bodźce, szuka się na przykład wartości LD_{50} (*lethal dose* 50) – jest to wartość dawki, po której podaniu 50% zwierząt poddanych eksperymentowi umrze. Najczęstszym zwierzęciem, z jakim miał Berkson do czynienia w klinice Mayo, był człowiek. Na ludziach nie prowadzono badań mogących skutkować obliczeniami LD_{50}, pojawiały się za to różne ważkie kwestie epidemiologiczne. Czy czynnik X podnosi czy zmniejsza ryzyko zachorowania na chorobę Y, a może kompletnie nie ma znaczenia?

Sadząc po zainteresowaniu, jakie budzą sezonowe sensacje dotyczące odmładzających właściwości zielonej herbaty czy antyoksydantów

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

zapewniających niemalże nieśmiertelność, czy – z drugiej strony – panice budzonej przez parabeny lub GMO, należy stwierdzić, że metodologia badań epidemiologicznych jest kwestią bardzo żywotną, z punktu widzenia zarówno interesu społecznego, jak i poszczególnego zjadacza chleba (najlepiej z antyoksydantami i niemodyfikowanego genetycznie).

Szpital jako miejsce, przez które przewija się mnóstwo ludzi – przeważnie chorych i poddanych rygorystycznym badaniom, których obiektywizmowi nie można zarzucić arbitralności odpowiedzi, jak na przykład w badaniach ankietowych, wydaje się idealnym miejscem na prowadzenie badań nad zależnościami pomiędzy różnymi czynnikami i schorzeniami. A przynajmniej wydawał się, póki Berkson nie wskazał na fatalny błąd, mogący obciążać konkluzje wyciągane z takich badań.

Efekt ten nazywany jest „błędem Berksona" (*Berkson fallacy*) lub „paradoksem Berksona" (*Berkson paradox*). Sam uczony wolał tę drugą nazwę, co – z powodu dwuznaczności pierwszego określenia – raczej nie dziwi...

Paradoks Berksona

Zanim przejdziemy do bardziej skomplikowanych wywodów, prezentujących paradoks Berksona na podstawie jego przełomowej pracy odnoszącej się do badań w szpitalu, przedstawmy sedno tego zagadnienia na prostszym przykładzie.

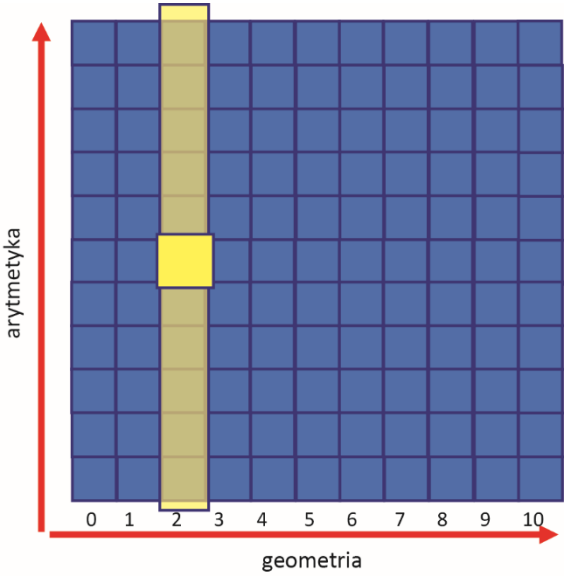
W konkursie matematycznym bierze udział 121 uczniów. Każdy z nich dostaje punkty w dwóch kategoriach: arytmetyki i geometrii, od zera do dziesięciu. Tak się złożyło, że wśród nich uzyskano wszystkie możliwe kombinacje punktów: dwa zera, zero z geometrii i jeden z arytmetyki, jeden z geometrii i zero z arytmetyki… i tak dalej, aż do dwóch dziesiątek. Na rysunku 1 każdy niebieski kwadracik oznacza wyniki pojedynczego ucznia.

Jaki jest średni wynik z arytmetyki uczniów, którzy uzyskali z geometrii, powiedzmy, dwa punkty? Oczywiście, średni ich wynik to 5 punktów. Tak samo zresztą jak dla uczniów, którzy zdobyli z geometrii zero, pięć czy dziesięć punktów. Średnia z arytmetyki nie zależy od średniej z geometrii i zawsze wynosi pięć punktów. W drugą stronę jest tak samo. Średnia z geometrii nie zależy od arytmetyki i też wynosi pięć punktów dla każdej zadanej liczby punktów z arytmetyki. Czyli – umiejętności geometryczne i arytmetyczne w ogóle od siebie nie zależą!

Zasada kwalifikacji uczniów do drugiego etapu konkursu jest następująca. Przechodzą uczniowie, którzy mają w sumie co najmniej dziesięć punktów. Są to zatem ci uczniowie, którzy na rys. 2 zaznaczeni są zielonymi kwadracikami.

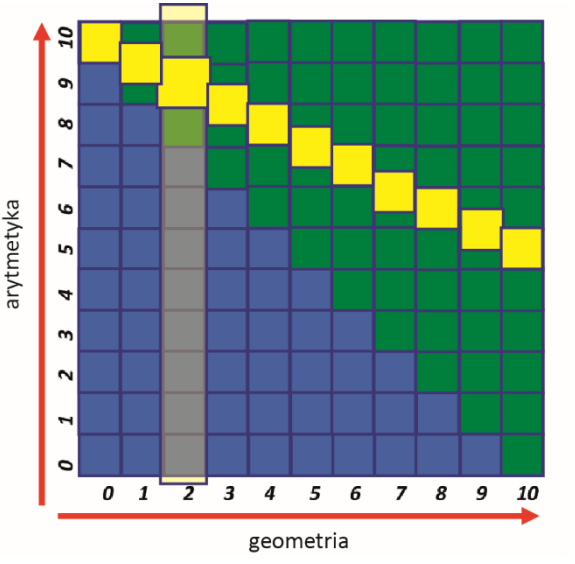
ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY

Nr 18(24)



Rys. 1. Wyniki z geometrii i arytmetyki 121 uczniów

Źródło: opracowanie własne.



Rys. 2. Pozorna zależność pomiędzy wynikami z arytmetyki i geometrii powstająca na skutek selekcji uczniów

Źródło: opracowanie własne.

Skupmy się na tych wyselekcjonowanych uczniach. Jaki będzie średni wynik z arytmetyki wśród tych, którzy uzyskali z geometrii trzy punkty? Tym razem nie będzie to pięć. Aby „nadrobić" słaby wynik z geometrii, uczniowie muszą mieć więcej punktów z arytmetyki – średnio aż dziewięć punktów! Im więcej punktów z geometrii, tym mniej punktów trzeba mieć z arytmetyki, by przejść do kolejnego etapu. Zatem im więcej punktów z geometrii, tym niższa średnia z arytmetyki – jak pokazują żółte kwadraty na rys. 2.

I w tym tkwi sedno: na skutek pewnej reguły selekcji (do kolejnego etapu konkursu) pojawiła się zależność pomiędzy wynikami z geometrii a wynikami z arytmetyki, która wyjściowo wśród wszystkich uczniów nie istniała!

Podobnym przykładem, bardziej humorystycznym, ilustruje paradoks Berksona matematyk i popularyzator nauki Jordan Ellenberg (2017). Oto wyjaśnienie tajemnicy, nad którą głowił się niejeden kawaler czy niejedna panna: dlaczego przystojni mężczyźni są głupi, a mądrzy – brzydcy? (i to samo w odniesieniu do kobiet). Otóż właśnie jest to złudzenie, wynikające z preselekcji. Głupiego brzydala odrzucimy w przedbiegach i nie damy mu w ogóle szansy na randkę. Jeśli umawiamy się na spotkania z mężczyznami, którzy wstępnie się kwalifikują – łącznie za urodę i błyskotliwość otrzymują jakąś minimalną liczbę punktów – nic dziwnego, że w grupie, z której z góry usunięto osoby o najniższych łącznych notach, zaistnieje ujemna zależność między urodą a intelektem…

Oryginalna praca Berksona (1946), dotycząca paradoksu nazwanego później jego imieniem, powstała w 1946 roku i wskazywała na pozorne korelacje, mogące zostać „zaobserwowane" w nieprawidłowo dobranej próbie. Przykład analizowany w tej pracy był nieco bardziej skomplikowany niż uproszczenia zakreślone powyżej.

W praktyce eksperymentalnej, z użyciem zwierząt laboratoryjnych, badacz ma pełną swobodę, by podzielić wszystkie na przykład myszki na grupę eksperymentalną i tak zwaną kontrolną, po czym tę eksperymentalną poddać działaniu czynnika, który ma być analizowany, a grupę kontrolną tylko obserwować. Badając następnie obie grupy i porównując częstość występowania danych efektów, można wyciągać wnioski, czy efekty te są powiązane z analizowanym czynnikiem. Jeśli częstości w obu grupach różnią się znacznie, można postulować, że efekt powiązany jest z tym czynnikiem; jeśli natomiast różnice częstości są znikome, przypisuje się to dziełu przypadku (oczywiście, można posłużyć się testem istotności, by obliczyć prawdopodobieństwo, że efekt jest czy też że nie jest dziełem przypadku). Jeśli chodzi o badania nad ludźmi, zazwyczaj nie ma niestety – czy raczej na szczęście – takiej swobody. Zwłaszcza w przypadku chorób współcześnie żaden chyba badacz nie odważyłby się nawet składać takiego podania do zaopiniowania komisji etyki („grupę młodych zdrowych ludzi

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

zarażamy wirusem dengi..."). Pozostaje zatem pracować na zastanym materiale. A ten – jeśli chodzi o ludzkie choroby – najlepiej dostępny jest nie gdzie indziej jak w szpitalu.

W klinice właśnie pracował Berkson i z badań na pacjentach kliniki on i jego koledzy próbowali wyciągać wnioski dotyczące epidemiologii i wzajemnych powiązań różnych chorób. W pewnym momencie uczonego uderzyło, że choć tworzone w praktyce klinicznej tabele kontyngencji wyglądają identycznie jak te tworzone w praktyce laboratoryjnej, różnią się one w zasadniczym aspekcie. Najprostszą tabelą kontyngencji jest tabela dwudzielcza i takim właśnie przykładem – w ślad za oryginalną pracą Berksona – tutaj się posłużymy.

Zacznijmy od sytuacji eksperymentalnej. Dzielimy myszki na dwie grupy, eksperymentalną oraz kontrolną. Grupie eksperymentalnej aplikujemy jakiś bodziec, po czym badamy myszki, czy wystąpił u nich efekt czy też nie wystąpił. Wyniki zbieramy w tablicy dwudzielczej:

Tabela 1. Tabela dwudzielcza dla badania eksperymentalnego

	efekt	brak efektu
eksperymentalna	*a*	*b*
kontrolna	*c*	*d*

Źródło: opracowanie własne.

„Efekt" obserwuje się u $\frac{a}{a+b}$ grupy eksperymentalnej oraz u $\frac{c}{c+d}$ osobników z grupy kontrolnej. Porównujemy te ułamki, czy różnią się znacząco, aby zawyrokować, czy „bodziec" powiązany jest z „efektem".

W praktyce klinicznej nie mamy możliwości *odgórnego* podzielenia na grupy eksperymentalną i kontrolną. Zwykle wygląda to tak, że aby zbadać potencjalną zależność pomiędzy chorobami A i B, jako grupę kontrolną wybieramy pacjentów obarczonych takim schorzeniem, które „ponad wszelką wątpliwość" nie ma wpływu ani na A, ani na B. Berkson podaje przykład badania nad zależnością pomiędzy zapaleniem pęcherzyka żółciowego a cukrzycą, gdzie grupę kontrolną stanowią pacjenci z wadami wzroku, zgłaszający się do kliniki w celu dobrania soczewek korekcyjnych.

Na pierwszy rzut oka tabela dwudzielcza wygląda identycznie i tak samo można porównywać odsetki osób z cukrzycą, które zachorowały na zapalenie woreczka żółciowego, z odsetkiem tych z grupy „kontrolnej", czyli pacjentów z wadami wzroku (i zapaleniem woreczka).

Czy zatem sytuacja jest w całości analogiczna? Pomimo powierzchownego podobieństwa tabel, wskazuje Berkson, istotna różnica polega na tym, iż w tym drugim przypadku nie było możliwości odgórnego podzielenia na grupy eksperymentalną i kontrolną.

Tabela 2. Tabela dwudzielcza dla badania w klinice

	zapalenie	brak zapalenia
cukrzyca	*a*	*b*
wada wzroku	*c*	*d*

Źródło: opracowanie własne.

Sęk w tym, że nie każdy cukrzyk i nie każdy chory na zapalenie woreczka żółciowego są od razu hospitalizowani. Nie każdy krótkowidz czy dalekowidz będzie też potrzebował nowych okularów w okresie przeprowadzania analizy. Gdyby tak było lub gdybyśmy w inny sposób mogli przebadać wszystkich cukrzyków, wszystkie osoby z wadami wzroku i wszystkich chorujących na zapalenie woreczka – wnioski wyciągane na temat zależności mogłyby być uprawnione[2]. Tymczasem jeśli mamy dostęp tylko do próby osób „wybranych" w sposób nielosowy, bo „wybrać" możemy tylko osoby hospitalizowane lub przynajmniej odwiedzające klinikę, ryzykujemy popełnienie poważnego błędu, a nawet wyciągnięcie wniosków całkowicie odwrotnych od prawdy.

Uproszczę nieco oryginalny przykład Berksona i od razu posłużę się konkretnymi (acz zupełnie nierealistycznymi[3]) liczbami zamiast ogólnej prezentacji na symbolach.

Niech w naszej ogólnej populacji znajduje się 20 tysięcy diabetyków oraz 100 tysięcy osób z wadami wzroku, przy czym spośród nich 10 tysięcy ma zarówno cukrzycę jak i wadę wzroku. Załóżmy dalej, że zachorowalność na zapalenie pęcherzyka żółciowego jest taka sama wśród wszystkich tych osób i wynosi 5%. Mielibyśmy zatem 200 cukrzyków z zapaleniem pęcherzyka i tysiąc osób z wadą wzroku i zapaleniem. Każde ze schorzeń w jakiś sposób predysponuje do zgłoszenia się do szpitala. Przyjmijmy dalej, że prawdopodobieństwo hospitalizacji z powodu cukrzycy wynosi 10%, z powodu zapalenia woreczka 30%, a prawdopodobieństwo zgłoszenia się po okulary osoby z wadą wzroku – 60%. Załóżmy, że prawdopodobieństwa te są niezależne, czyli że nosiciel dwóch lub trzech chorób zgłosi się do szpitala z odpowiednio *wyższym* prawdo-

[2] Zwróćmy jednakże uwagę, że nawet w takim przypadku możliwości wyciągania wniosków są wciąż ograniczone wobec sytuacji eksperymentalnej. W tej ostatniej badacz ma podstawy do postulowania zależności przyczynowo-skutkowych, na mocy starożytnej zasady *post hoc ergo propter hoc*, która tutaj niekoniecznie byłaby błędem logicznym. Jeśli zakażamy wirusem myszki z grupy eksperymentalnej, a następnie obserwujemy u nich objawy, których wcześniej nie było (i które nie wystąpiły w grupie kontrolnej), mamy solidne podstawy do wnioskowania o przyczynie i skutku. W przypadku omawianych badań na ludziach, nawet przy dostępie do całej populacji, jeśli nie obserwujemy rozwoju sytuacji w czasie, mamy podstawy tylko do wniosków o współzależności, a nie przyczynowości.

[3] Wybranymi tak, by wychodziły w miarę okrągłe i wyraziste wyniki.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

podobieństwem, liczonym zgodnie z regułami. Berkson tłumaczy je obrazowo: „O osobie z dwoma schorzeniami możemy myśleć jak o bliźniętach syjamskich, z których każde choruje na jedno schorzenie. Prawdopodobieństwo, że bliźnięta udadzą się do szpitala jest równe prawdopodobieństwu, że którykolwiek z nich pójdzie do szpitala, przy czym choroba jednego nie wpływa na chorobę drugiego" (Berkson, 1946, s. 513). Sumujemy zatem te prawdopodobieństwa i odejmujemy prawdopodobieństwo, że *obaj* chcą iść do szpitala, aby nie liczyć go podwójnie – prawdopodobieństwo takiej sytuacji ujęte jest zarówno w wielkości prawdopodobieństwa dla pierwszego z bliźniąt chcącego iść do szpitala (nigdzie nie jest powiedziane, że drugie jednocześnie też nie chce, z własnej przyczyny), jak i gdy bierzemy pod uwagę drugie z bliźniąt. Jeśli zatem dla jednego z bliźniąt prawdopodobieństwo wynosi p_1, a dla drugiego p_2, trafią oni do szpitala z prawdopodobieństwem $p_{1+2} = p_1 + p_2 - p_1 \cdot p_2$. Alternatywnie odejmowanie ostatniego wyrazu można wyjaśnić tak. Prawdopodobieństwo p_{1+2} otrzymamy, sumując prawdopodobieństwa, że tylko pierwsze z bliźniąt chce do szpitala, a drugie nie (czyli $p_1 \cdot (1 - p_2)$), że tylko drugie z bliźniąt chce do szpitala, a pierwsze nie (czyli $p_2 \cdot (1 - p_1)$), oraz że chcą oba naraz (czyli $p_1 \cdot p_2$). Po zsumowaniu i uproszeniu otrzymujemy: $p_{12} = p_1 \cdot (1 - p_2) + p_2 \cdot (1 - p_1) + p_1 \cdot p_2 = p_1 + p_2 - p_1 \cdot p_2$.

Zatem dla współwystępujących dwóch chorób (oznaczonych ogólnie 1 i 2) prawdopodobieństwo wynosi $p_{1+2} = p_1 + p_2 - p_1 \cdot p_2$, a dla współwystępujących trzech chorób (1, 2 oraz 3) można analogicznie obliczyć: $p_{1+2+3} = p_1 + p_2 + p_3 - p_1 \cdot p_2 - p_2 \cdot p_3 - p_1 \cdot p_3 - p_1 \cdot p_2 \cdot p_3$.

Stwórzmy teraz tabele dwudzielcze liczebności dla całej populacji oraz chorych przebadanych w klinice (tab. 3). Za grupę kontrolną przyjmiemy osoby mające wadę wzroku, które *nie* chorują jednocześnie na cukrzycę (zatem liczebność grupy kontrolnej równa jest liczebności osób z wadami wzroku minus liczebność osób z wadami wzroku oraz jednocześnie cukrzycą). Jedynką oznaczmy cukrzycę, dwójką – wadę wzroku, a trójką zapalenie woreczka żółciowego. Na górze przestawione zostały tabele dwudzielcze z wartościami otrzymanymi za pomocą obliczeń, które dla możliwości sprawdzenia widnieją w dolnej parze tabel.

Odsetek chorujących na zapalenie woreczka wśród diabetyków wynosi około 7,37%, natomiast w grupie kontrolnej – 5,94%. Różnica wynosi niemal półtora punktu procentowego.

Można by wyciągnąć pochopny wniosek: cukrzyca predysponuje do zapalenia woreczka żółciowego. I gdybyśmy nie byli w tej uprzywilejowanej sytuacji, że z góry wiedzieliśmy, iż odsetek zapadających na zapalenie woreczka jest taki sam w całej populacji, pewnie trudno byłoby nie ulec tej pokusie...

Tabela 3. Wyniki hipotetycznego badania zapadalności na zapalenie woreczka żółciowego

	cała populacja		
	zapalenie	brak	łącznie
cukrzyca	1 000	19 000	20 000
grupa kontrolna	5 000	95 000	100 000

	chorzy w klinice		
	zapalenie	brak	łącznie
cukrzyca	559	7 030	7 571
grupa kontrolna	3 600	57 000	60 600

	cała populacja	
	zapalenie	brak
cukrzyca	$n_1 \cdot f_3$	$n_1 \cdot (1 - f_3)$
grupa kontrolna	$(n_2 - n_{1+2}) \cdot f_3$	$(n_2 - n_{1+2}) \cdot (1 - f_3)$
	chorzy w klinice	
	zapalenie	brak
cukrzyca	$(n_1 - n_{1+2}) \cdot f_3 \cdot p_{1+3} + n_{1+2} \cdot f_3 \cdot p_{1+2+3}$	$(n_1 - n_{1+2}) \cdot (1 - f_3) \cdot p_1 + n_{1+2} \cdot (1 - f_3) \cdot p_{1+2}$
grupa kontrolna	$(n_2 - n_{1+2}) \cdot f_3 \cdot p_{2+3}$	$(n_2 - n_{1+2}) \cdot (1 - f_3) \cdot p_2$

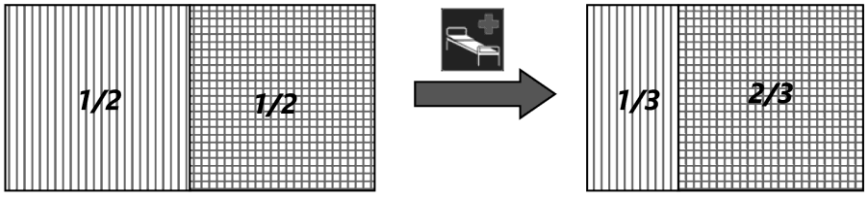
$n_{1,2,1+2}$ – liczby wszystkich osób z chorobami: cukrzyca (1), wada wzroku (2), cukrzyca + wada wzroku (1+2);
f_3 – zachorowalność na zapalenie woreczka żółciowego;
$p_{\{ch\}}$ – odsetek hospitalizowanych pacjentów z grupy chorób $\{ch\}$;
$n_1 = 20\,000, n_2 = 110\,000, n_{1+2} = 10\,000, f_3 = 0.05, p_1 = 0.10, p_2 = 0.60, p_3 = 0.30, p_{1+2} = p_1 + p_2 - p_1 \cdot p_2 = 0.64,$
$p_{1+3} = p_1 + p_3 - p_1 \cdot p_3 = 0.37,$
$p_{2+3} = p_2 + p_3 - p_2 \cdot p_3 = 0.72,$
$p_{1+2+3} = p_1 + p_2 + p_3 - p_1 \cdot p_2 - p_2 \cdot p_3 - p_1 \cdot p_3 - p_1 \cdot p_2 \cdot p_3 = 0.748.$

Źródło: opracowanie własne.

Spróbujmy krok po kroku prześledzić, w którym miejscu pojawia się błąd.

Wystarczy tylko założenie, że pacjenci obarczeni większą liczą chorób pojawią się w szpitalu z większym prawdopodobieństwem (całkiem realistyczne założenie), by pojawiła się rozbieżność pomiędzy odsetkiem osób z objawami H w całej populacji chorych na chorobę A, a odpowiednim odsetkiem wśród osób hospitalizowanych. Powiedzmy, że chorobie A w połowie przypadków towarzyszy przypadłość H. Sama choroba A nie jest bezwzględnym wskazaniem do hospitalizacji i pacjenci nieobarczeni towarzyszącą przypadłością H zgłoszą się do szpitala tylko w połowie przypadków. Za to pacjenci z obydwiema chorobami bezwzględnie trafiają do szpitala. Co się okazuje? O ile ogólnie H występuje w 50% przypadków A, o tyle w szpitalu objawy H ma 67% pacjentów z A!

Na razie otrzymaliśmy zaburzony odsetek w populacji szpitalnej (w porównaniu z odsetkiem w całej populacji). Można by jednak pomyśleć, że nie jest to tak groźny efekt, jeśli chcemy tylko porównywać dwa



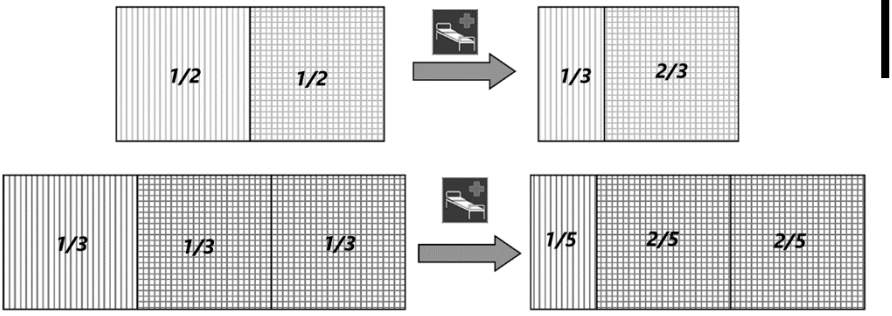
Rys. 3. Kreski pionowe – choroba A, kratka – choroby A oraz H

Źródło: opracowanie własne.

odsetki. Jasne, że w szpitalu rzadziej spotyka się ludzi zdrowych niż na ulicy. Ale jeśli naszym celem jest tylko porównanie, czy choroba A bardziej predysponuje do objawów H niż choroba B, może trzeba po prostu ograniczyć się do porównań, mając na uwadze, że odsetki obserwowane w szpitalu są ogólnie zawyżone wobec całości populacji? Gdyby w szpitalu chorzy na A mieli w 67% przypadków objawy H, natomiast wśród chorych na B ten odsetek wynosiłby 33%, wnioskowalibyśmy, że i w całej populacji objawy H dwukrotnie częściej towarzyszą chorobie A niż chorobie B, nie wyrokując niczego w odniesieniu do konkretnych ich wartości wśród ogółu chorych, również tych niehospitalizowanych. Czy byłby to wniosek uprawniony? Niestety, niekoniecznie. Tylko pod warunkiem, że wyjściowe proporcje są takie same (wskaźniki objawów H wśród chorych na A i B) oraz że do szpitala kładą się identyczne odsetki chorych tylko z A i tylko z B oraz takie same odsetki chorych z A+H i B+H. Porównajmy dwie przykładowe sytuacje.

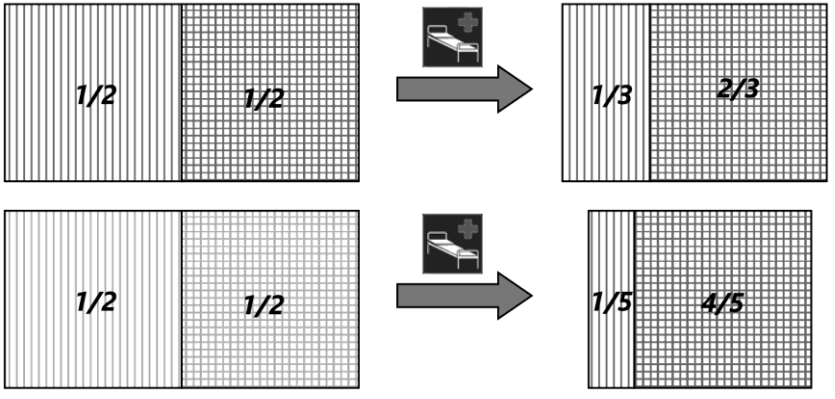
Najpierw niech wyjściowe wskaźniki będą różne. Chorobie A objawy H towarzyszą w połowie przypadków, natomiast chorobie B – w dwóch trzecich przypadków. Zatem stosunek wskaźników wynosi: $\frac{2/3}{1/2} = \frac{4}{3}$. Jak wcześniej, niech połowa chorych z tylko jedną chorobą (tylko A lub tylko B) idzie do szpitala oraz wszyscy chorzy z objawami H (czyli wszyscy A+H i B+H). W szpitalu okazuje się zatem, że 2/3 chorych z A ma objawy H oraz 4/5 chorych z B. To oznacza, że wśród hospitalizowanych objawy H występują w grupie B $\frac{4/5}{2/3} = \frac{6}{5}$ razy częściej niż w grupie A. Stosunek wskaźników różni się od ich relacji w ogóle populacji.

Teraz niech wyjściowe wskaźniki będą takie same, ale różnią się odsetki hospitalizowanych. Tak jak wcześniej połowa chorych z A i wszyscy z A + H idą do szpitala, natomiast w przypadku drugiej choroby – co czwarty chory tylko z B oraz wszyscy chorzy z B + H. Jaki odsetek objawów H otrzymamy w hospitalizowanej grupie chorych na B? Będzie to oczywiście 4/5, odmiennie niż w przypadku objawów H wśród hospitalizowanych chorych na A (jak wcześniej ten odsetek wynosi 2/3).



Rys. 4. Kreski pionowe – choroba A lub B, kratka – choroby A lub B oraz H

Źródło: opracowanie własne.



Rys. 5. Kreski pionowe – choroba A lub B, kratka – choroby A lub B oraz H

Źródło: opracowanie własne.

Przykład Berksona (tak jak praktyka kliniczna) jest dużo bardziej skomplikowany. Wchodzą tam w grę trzy choroby, przy czym współwystępować może każda ich kombinacja (czyli choroba pojedyncza, dwie różne choroby, wszystkie trzy naraz), a zarówno wyjściowe liczebności zachorowań, odsetki współwystępowania, jak i współczynniki hospitalizacji różnią się od siebie. Skoro w takich uproszczonych przypadkach jak przedstawione w poprzednich akapitach występują niezgodności pomiędzy wskaźnikami w całej populacji i w grupie hospitalizowanej, nic dziwnego, że tym bardziej skrzywione muszą być wyniki analiz, gdy kilka różnych czynników może dawać wkład do błędnej konkluzji.

Początkujących w nauce statystyki zawsze przestrzega się przed zwodniczym charakterem korelacji. Podkreśla się, że korelacja nie oznacza związku przyczynowo-skutkowego. Występowanie korelacji pomiędzy

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

C i D nie musi oznaczać, że C wywołuje D bądź na odwrót. Najprostszym wyjaśnieniem alternatywnym może być to, że zarówno C, jak i D powodowane są przez tę samą przyczynę, P, i dlatego współwystępują. Jeśli waga jest skorelowana ze stopniem choroby próchnicowej u dzieci, niekoniecznie oznacza to, że lecząc próchnicę, odchudzimy dziecko lub że dając mu tabletki na odchudzanie, poprawimy stan zębów. Bardziej wiarygodnym wyjaśnieniem jest to, że korelacja jest wynikiem wspólnej przyczyny i nadwagi, i próchnicy – nadmiernej ilości zjadanych słodyczy. To jest podręcznikowa wiedza. Paradoks Berksona rozszerzył stwierdzenie dotyczące złudnych związków o efektowny drugi człon:

Korelacje mogą być wynikiem zarówno wspólnej przyczyny, jak i wspólnych skutków.

Jak każdy dobry *bon mot*, tak i ten przyjmowany być musi z pewną dozą tolerancji na nieścisłości. Powodowane wspólną przyczyną korelacje są *rzeczywiste*, tylko nie można interpretować ich w kategoriach przyczynowo-skutkowych. Korelacje powodowane wspólnym skutkiem (trafienie do szpitala, uwzględnienie w planach randkowych jak w przykładzie Ellenberga itd.) nie tylko nie mogą być ujmowane w ramach par przyczyna/skutek, ale są *rzekome* w tym sensie, że obserwuje się je *tylko* w próbie, której dotyczy dany skutek, a nie w całej populacji. Wspólna przyczyna zaś odnosi się zazwyczaj do wszystkich przypadków.

Swoją analizą Berkson dał do ręki potężne narzędzie przyszłym badaczom. Jako ciekawy współczesny przykład warto przytoczyć kwestię kasków motocyklowych. Czy faktycznie są skuteczne? Niektórzy twierdzili, że w razie wypadku zwiększony ciężar noszony na głowie zwiększa ryzyko poważnych obrażeń kręgosłupa, zwłaszcza odcinka szyjnego. Inni podnosili, że nawet jeśli kask chroni, to jego noszenie wystawia na ryzyko tak zwanego hazardu moralnego – motocyklista w kasku czuje się bardziej bezpiecznie, zatem pozwala sobie na bardziej brawurowe zachowania. Ryzyko to nie jest kompensowane przez zwiększone bezpieczeństwo w trakcie ewentualnego wypadku, co koniec końców wychodzi na niekorzyść. Ze względu na liczbę poszkodowanych motocyklistów (w kaskach) regularnie trafiających na oddziały ratunkowe szpitali sądzić można, że jest coś na rzeczy. Może nawet zaobserwowano, że obrażenia tych, którzy nosili kaski, są poważniejsze od obrażeń kierowców mniej zabezpieczonych? A te dodatkowe problemy z udzieleniem pierwszej pomocy, gdy należy ściągnąć kask bez dodatkowego zaszkodzenia ofierze wypadku?

Nim zasugerujemy inne rozwiązanie kwestii kasków, czas na dygresję. Trwa druga wojna światowa, US Navy głowi się nad rozwiązaniem problemu zestrzeleń samolotów. Może opancerzyć dodatkowo któryś element kadłuba lub skrzydeł? Do akcji wkracza statystyk – Abraham Wald. Analizuje dane dotyczące dziur w opancerzeniu. Okazuje się, że uszkodzenia koncentrują się w obrębie końcówek skrzydeł i powierzchniach sterowych,

natomiast relatywnie mało jest ich w centralnej części kadłuba. Jaka była rekomendacja Walda? Wzmocnić... Nie, nie skrzydła, ale właśnie środek kadłuba! Czy to nonsens? Przecież tam prawie wcale nie było uszkodzeń.

Otóż to. Wziąć pod uwagę trzeba fakt, że nie każdy samolot wrócił, by przejść badanie. Te z uszkodzonymi skrzydłami – wróciły. Czyli postrzał tego typu nie był zabójczy ani dla załogi, ani dla samego samolotu. Czy brak samolotów z dziurami w środkowej części kadłuba wynikał z faktu, że ostrzał nieprzyjaciela skrupulatnie omijał te części? Czy raczej z tego, że tak postrzelone samoloty nie były w stanie wrócić do bazy, a piloci ginęli bądź dostawali się do niewoli?

Z powrotem na oddziale ratunkowym szpitala. Zastanowić się możemy, czy nadreprezentacja kierowców w kaskach nie może wynikać z tego, że wielu z tych bez kasków nie znalazło się w szpitalu dlatego, że pieczę nad nimi musiała przejąć inna, bardziej ostateczna, instytucja? Czyli, że ponieśli śmierć na miejscu? Kilka lat temu dwóch kanadyjskich naukowców przeprowadziło taką analizę, popierając ją liczbami (Woodfine i Redelmeier, 2015).

Rok 2020, pandemia koronawirusa na świecie. Niemal tego samego dnia znaleźć można w sieci sprzeczne informacje. Palacze mają większe ryzyko zachorowania. Palacze rzadziej mają ciężkie symptomy (Guan et al., 2020). Co jest prawdą? Faktycznie, w jednym badaniu we Francji zaobserwowano, że wśród zdiagnozowanych i hospitalizowanych chorych odsetek palaczy jest mniejszy niż odsetek osób, które otrzymały pozytywny wynik testu, a których stan pozwalał na pozostanie w domu (Myiara et al., 2020). Ludzie masowo zaczęli wykupywać tytoń, rząd wprowadził limity, naukowcy poczęli wymyślać domniemane mechanizmy ochronnej funkcji nikotyny. Zadowoleni nałogowcy szybko zostali jednak oblani kubłem zimnej wody przez badaczy wskazujących na możliwe źródła skrzywienia wyników. Berskson zapewne też by się do tych głosów przyłączył: być może te liczby są tylko efektem nielosowości próby? Nie zostały przecież przebadane całe populacje Chin i Francji, a właściwie przebadano niewielkie ich ułamki. A co jeśli pokasłujący palacze nie są podejrzewani o zakażenie, bo ich kaszel składa się na karb uzależnienia, natomiast podobne objawy u osoby o dotychczas czystych płucach budzą od razu niepokój? Powierzchowne wnioski, wyciągane na podstawie wyników badań na nielosowych próbach, powinny z miejsca wywoływać zdrowy sceptycyzm. Mam nadzieję, iż nie wrócimy do zgoła średniowiecznych metod zabezpieczenia się przed chorobą za pomocą trucizny (papierosami przed Covidem), jak niegdyś, gdy rtęcią leczono syfilis[4].

[4] Choć wciąż nowotwory leczone są za pomocą chemoterapii – może to świadczyć, że w tej dziedzinie nauka pozostaje jednak w tyle w stosunku do choćby bakteriologii czy wirusologii.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

Palenie tytoniu stało się zresztą już za czasów Berksona polem ogromnej bitwy, w której jako broni użyty został jego eponimiczny paradoks – niefortunnie w roli obrońcy nałogu. Zaangażowanie naukowca w tę dyskusję przez wielu uznane zostało za jego największą życiową i naukową porażkę. Nim jednak przejdziemy do tego błędu, błędu w sensie potocznym, zatrzymajmy się na innym, który był kolejnym doniosłym wkładem do statystyki teoretycznej: na modelu błędu Berksona.

Błąd, który nie był błędem

Sformułujmy bardzo prosty (pozornie) problem: ile wynosi gęstość wody? Każdy wie, że „gęstość to masa przez objętość". Wystarczy zatem wziąć trochę wody, zmierzyć masę, zmierzyć objętość – podzielić pierwsze przez drugie i mamy wynik. Proste? Proste, ale nie do końca poprawne. Każdy pomiar obarczony może być błędem. Robimy jeden pomiar pary wielkości: masa i objętość. Otrzymujemy wynik ρ_1. Powtarzamy pomiary i obliczamy wynik, otrzymując ρ_2, które różni się nieco od pierwszego. Które jest lepsze, bardziej wiarygodne? Przecież istnieje jakaś (prawdziwa) gęstość wody. Przy stałych warunkach (ciśnieniu i temperaturze) ta prawdziwa gęstość będzie występowała jako współczynnik kierunkowy w równaniu:

$$\textbf{\textit{masa = gęstość x objętość.}}$$

Gdybyśmy potrafili mierzyć doskonale dokładnie, wystarczyłby nam jeden pomiar masy i odpowiadający jej pomiar objętości, by otrzymać prawidłową wartość gęstości. Niestety, nasze przyrządy są ułomne i za każdym pomiarem otrzymamy nieco inny wynik, spośród których nie mamy jak wybrać tego „najpoprawniejszego".

Na szczęście nie jesteśmy w obliczu tego faktu tak kompletnie bezradni, a to dzięki bohaterskim statystykom, którzy opracowali przydatne narzędzia.

Aby nie przywiązywać się za bardzo do wody i jej gęstości, rozważmy dwie dowolne wielkości, *x* i *y*, powiązane ze sobą prawem liniowym:

$$y_i = \alpha x_i + \beta,$$

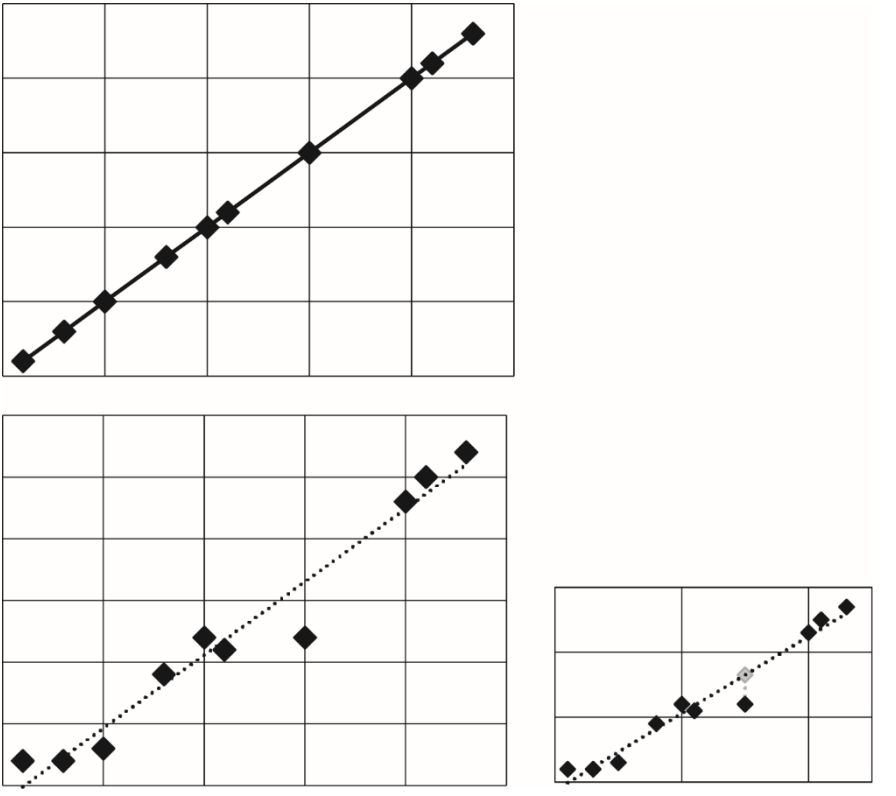
przy czym naszym zadaniem jest znalezienie parametrów tego prawa, czyli wartości α i β[5]. Indeks i w powyższym równaniu oznacza, że zależność ta obowiązuje dla wielu różnych par pomiarów, czyli na przykład dla kieliszka wody o masie y_1 i objętości x_1, dla kubka wody o masie y_2 i objętości x_2 i tak dalej. Indeks można też traktować jako numerację pomiarów

[5] Z podobnym zapisem i podobnymi obliczeniami mamy do czynienia w sytuacji konceptualnie odmiennej: gdy wielkości *x* i *y* są niezależnymi zmiennymi losowymi, a naszym zadaniem jest wyznaczenie wartości oczekiwanej jednej z wielkości przy ustalonej drugiej.

tego samego obiektu: wejdźcie raz po razie na tę samą wagę elektroniczną o dużej dokładności – możliwe, że pokaże dwa nieco odmienne wyniki (ze względu na ułomność przyrządów pomiarowych).

Wielkościami x i y mogą być wymieniane już masa i objętość danej substancji (współczynnikiem kierunkowym jest wówczas gęstość), napięcie i natężenie prądu w danym przewodniku (współczynnikiem kierunkowym jest wtedy odwrotność oporu przewodnika) czy wysokość dochodu i wysokość odprowadzonego podatku w systemie liniowym (współczynnikiem kierunkowym jest w tej sytuacji stopa opodatkowania).

Gdybyśmy mierzyli x i y perfekcyjnie, idealnie dokładnie, po czym nanieśli na wykres punkty odpowiadające pomiarom (każdy punkt ma współrzędne (x_i, y_i)), sytuacja wyglądałaby jak na rysunku 6a, a wartości parametrów moglibyśmy odczytać z wykresu bądź podstawiając do równania prostej współrzędne dwóch dowolnych punktów.



Rys. 6. Zbiór punktów i dopasowana do nich prosta: (a) możliwe jest idealne dopasowanie (b) nie istnieje prosta przechodząca przez wszystkie punkty

Źródło: opracowanie własne.

Jeśli jednakże pomiar obarczony jest błędem, przy czym błąd może dotyczyć albo tylko wielkości oznaczonych jako *x*, albo tylko wielkości *y*, albo obydwu – sytuacja wygląda mniej więcej tak, jak na rysunku 6b. W takim przypadku staramy się na podstawie tych ułomnych danych „odgadnąć" (estymować) wartości parametrów prostej. W dalszym ciągu zakładać będziemy – i jest to całkiem rozsądne założenie – że błąd, jaki popełniamy przy pomiarze (którejkolwiek z wielkości), jest błędem losowym z wartością oczekiwaną zero (i jest niezależny od wielkości mierzonej). Czyli równie prawdopodobne jest zarówno zawyżenie, jak i zaniżenie wyniku, nie ma błędu systematycznego.

Ponieważ wielkości, jakie mierzymy, nie są prawdziwymi wartościami *x*, oznaczymy je przez u. Podobnie zamiast *y* użyjemy *v*, przy czym:

$$u_{ij} = x_i + \epsilon_{ij},$$

$$v_{ij} = y_i + \eta_{ij},$$

a ϵ_{ij} oraz η_{ij} są błędami pomiaru. Podwójne indeksy oznaczają, że dla każdej konkretnej prawdziwej wartości x_i kolejne pomiary mogą dać różne błędy; tak samo dla y_i[6]. W dalszej części pominiemy podwójne indeksy – załóżmy, że każde x_i i każde y_i mierzymy dokładnie raz.

Estymacja metodą najmniejszych kwadratów opiera się na następującym kryterium: wśród wszystkich możliwych prostych (czyli wszystkich możliwych par (a, b)) szukamy takiej, by suma kwadratowych odchyleń punktów (u_i, v_i) od prostej $\hat{v}_i = au_i + b$ była jak najmniejsza.

Zauważmy dwa niuanse w zapisie tej prostej. Po pierwsze, v_i zyskało „daszek". Zyskało dlatego, że – jak wynika z poglądowego rys. 6a – nie istnieje prosta, która przechodziłaby przez zbiór punktów (u_i, v_i). Musimy się zatem liczyć z tym, że dla zmierzonej wartości u_i wartość funkcji różni się od zmierzonego v_i (i na odwrót). Zostało to zilustrowane dla jednej z wartości na rysunku 6a (mniejsza wstawka, różnica pomiędzy czarnym punktem – prawdziwa obserwacja, a szarym – wielkość wynikająca z „dopasowanej" linii). Po drugie, zamiast α i β pojawiają się a i b. Dlaczego? Dlatego, że ponieważ nie dysponujemy prawdziwymi parami (x_i, y_i), wcale nie mamy gwarancji, że nasza metoda estymacji da nam prawdziwe parametry prawa, α i β. Otrzymamy jakieś wartości a i b. A jak one się mają do α i β?

[6] Gdybyśmy indeksami oznaczali kolejne pomiary tej samej wielkości (np. masy i objętości tej samej szklanki wody), wówczas oznaczenia musiałyby być następujące: $u_i = x + \epsilon_i$, $v_i = y + \eta_i$. x i y utraciłyby w tej sytuacji indeks i, ponieważ, choć błąd za każdym razem może być inny (zatem i inne wartości zmierzone: u_i, v_i), to wartości prawdziwe, x i y, są za każdym razem takie same.

Stosując metodę najmniejszych kwadratów, łatwo wyprowadzić wzór na parametr a. Jeśli założymy, że średnie wartości u i v wynoszą zero (jeśli od każdej zmierzonej wartości odejmiemy wartość średnią, otrzymamy przeskalowane wielkości o średniej równej zero, a takie przeskalowanie nie wpływa na wartości szacowanych parametrów), wówczas otrzymamy:

$$a = \frac{\sum_i u_i \cdot v_i}{\sum_i u_i^2}$$

i jest to wynik doskonale znany z każdego podręcznika statystyki.

Zazwyczaj jednakże w podręcznikach nie dyskutuje się wspomnianej wyżej kwestii – jak ten „odgadnięty" współczynnik kierunkowy ma się do prawdziwego współczynnika – α, a to on przecież jest przedmiotem zainteresowania.

Gdybyśmy byli w stanie dokonywać pomiarów bez błędów, czyli $\epsilon_i \equiv 0$, $\eta_i \equiv 0$, wówczas, oczywiście, wyznaczona wartość a byłaby dokładnie równa prawdziwej wartości α:

$$a = \frac{\sum_i x_i \cdot y_i}{\sum_i x_i^2} = \alpha.$$

Jeśli jednakże popełniamy błąd przy pomiarze choćby jednej z wielkości, x lub y, sytuacja może być odmienna. O dziwo, różni się ona w zależności od tego, czy z błędem mierzymy zmienną zależną (y) czy niezależną (x). Łatwo się o tym przekonać.

Jeśli to y mierzone jest z błędem i zamiast igrekami dysponujemy zmierzonymi wartościami: $v_{ij} = y_i + \eta_{ij}$, wówczas:

$$a = \frac{\sum_i x_i \cdot v_i}{\sum_i x_i^2} = \frac{\sum_i x_i \cdot (y_i + \eta_i)}{\sum_i x_i^2} = \frac{\sum_i x_i \cdot y_i}{\sum_i x_i^2} + \frac{\sum_i x_i \cdot \eta_i}{\sum_i x_i^2}.$$

Ponieważ przyjęliśmy, iż wartość oczekiwana błędu wynosi zero, gdy mamy dostatecznie dużą liczbę pomiarów[7], drugi wyraz w wyrażeniu na a znika i otrzymujemy:

$$a = \frac{\sum_i x_i \cdot y_i}{\sum_i x_i^2} = \alpha.$$

W przedziwny sposób z ułomnych danych otrzymujemy prawdziwą wartość parametru funkcji liniowej.

Inaczej jednak sytuacja wygląda, gdy ułomność pomiarów dotyczy (tylko) zmiennej niezależnej – x. W tym przypadku bowiem zerowa wartość oczekiwana błędu nas nie ratuje:

[7] Dopiero w granicy, gdy liczba pomiarów dąży do nieskończoności (w praktyce – jest duża), można uznać, że błąd „na plus" będzie kasował się z błędem „na minus" – to wynika z definicji wartości oczekiwanej (oraz intuicji).

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

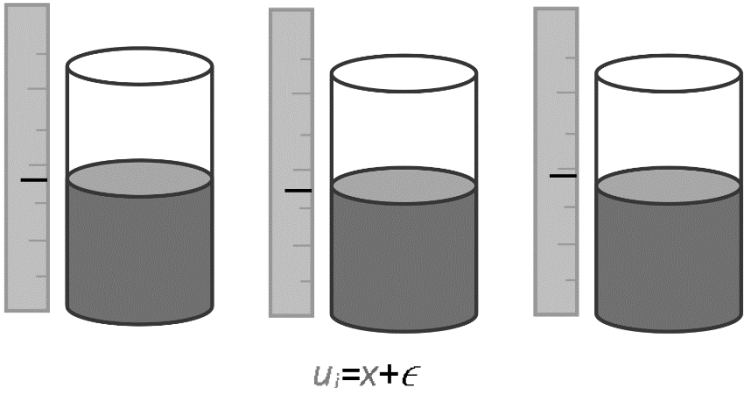
$$a = \frac{\sum_i (x_i+\epsilon_i)\cdot y_i}{\sum_i (x_i+\epsilon_i)^2} = \frac{\sum_i x_i \cdot y_i + \sum_i \epsilon_i \cdot y_i}{\sum_i x_i^2 + 2\sum_i \epsilon_i \cdot x_i + \sum_i \epsilon_i^2},$$

i choć ponownie na mocy zerowej wartości oczekiwanej błędu drugi wyraz w liczniku i środkowy wyraz w mianowniku znikają, to nieubłaganie pozostaje jeszcze ten ostatni:

$$a = \frac{\sum_i (x_i+\epsilon_i)\cdot y_i}{\sum_i (x_i+\epsilon_i)^2} = \frac{\sum_i x_i \cdot y_i}{\sum_i x_i^2 + \sum_i \epsilon_i^2} = \alpha \cdot \frac{\sum_i x_i^2}{\sum_i x_i^2 + \sum_i \epsilon_i^2} = \alpha \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\epsilon^2},$$

gdzie σ_x^2 oznacza wariancję wielkości x, a σ_ϵ^2 – wariancję ϵ[8].

Opisany tu został pokrótce tak zwany klasyczny model błędu. Zrekapitulujmy. Mamy do czynienia z sytuacją, gdy pewna ustalona i prawdziwa wielkość mierzona jest wielokrotnie, za każdym razem z pewnym błędem, wynikającym z ułomności przyrządów pomiarowych. Pomiar jednej prawdziwej wielkości daje nam w wyniku pewien rozrzut obserwacji, a z nich mamy wyciągać wnioski na temat tej prawdziwej wartości. Na laboratorium z fizyki studenci ze znudzeniem powtarzają takie pomiary: położyć odważnik na wagę, zanotować wynik, wyzerować przyrząd pomiarowy, położyć ten sam odważnik na wagę…



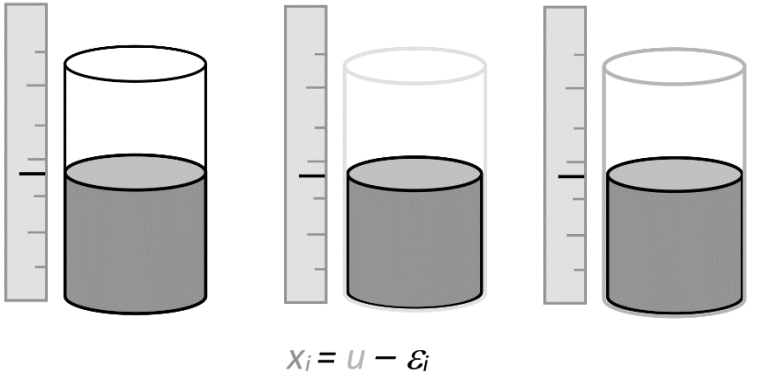
Rys. 7. Błąd klasyczny (powtarzając kilkakrotnie pomiar poziomu płynu w tej samej szklance, za każdym razem otrzymujemy odrobinę inny wynik – różny poziom czarnej kreski na linijce)

Źródło: opracowanie własne.

W niektórych eksperymentach laboratoryjnych praktyka jest jednakże odmienna, a logika całkowicie odwrócona. Jeśli chcemy określić odsetek

[8] Wariancja jest miarą rozrzutu wokół wielkości średniej. Jeśli wartość średnia jakiejś zmiennej wynosi zero, to suma (dużej liczby) kwadratów obserwacji, podzielona przez liczbę obserwacji, równa jest właśnie wariancji.

myszek, które przeżyją podanie określonej dawki pewnego środka, każdej z nich powinna zostać zaaplikowana ta sama dawka. Napełniamy menzurkę do poziomu określonej podziałki i aplikujemy substancję jednej myszce. Napełniamy inne naczynko do tej samej wysokości, aplikujemy kolejnej myszce… Tym razem każdy odczyt jest identyczny, ale nie możemy być pewni, że naprawdę każda z nalanych objętości była idealnie taka sama. Jesteśmy przekonani, że nalaliśmy dokładnie do tej samej podziałki, ale nasze oko i precyzja wykonywania pomiaru są ułomne – zatem faktyczna ilość cieczy może się odrobinę różnić w poszczególnych menzurkach. Czyli mamy do czynienia z sytuacją odwrotną. Nie mamy jednej prawdziwej wielkości i wielu różnych odczytów, ale wciąż ten sam odczyt i potencjalnie wiele różnych prawdziwych wartości. Taki błąd nazywany jest błędem Berksona, gdyż to on jako pierwszy sformułował to zagadnienie i szczegółowo omówił (Berkson, 1950).



Rys. 8. Błąd Berksona (nalewając do trzech różnych szklanek wodę do poziomu oznaczonego czarną kreską, za każdym razem otrzymujemy faktycznie odrobinę inną ilość cieczy)

Źródło: opracowanie własne.

Przedstawmy w dużym skrócie doniosły wynik Berksona.

Ponownie naszym celem jest jak najlepsze zbliżenie się do prawa: $y_i = \alpha x_i + \beta$. Tym razem jednak nie dysponujemy wieloma różnymi odczytami u_i, dotyczącymi tej samej prawdziwej wartości x, ale potencjalnie różnymi prawdziwymi wartościami x_j, odpowiadającym temu samemu odczytowi u. To samo dotyczyć może drugiej zmiennej lub obu naraz, czyli:

$$x_j = u - \epsilon_j,$$

$$y_j = v - \eta_j,$$

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

gdzie ϵ_i oraz η_i ponownie są błędami.

Ustalmy – powiedzmy – różne dawki specyfiku podawanego myszkom na poziomach u_j i zmierzmy ich reakcje – v_j. Przyjmijmy na początek, że dawkę, czyli x, mierzymy bez błędu: $x_i = u_i$. Jeśli estymujemy metodą najmniejszych kwadratów współczynnik kierunkowy w równaniu regresji, $\hat{v}_i = au_i + b$, otrzymamy:

$$a = \frac{\sum_i x_i \cdot (y_i + \eta_i)}{\sum_i x_i^2} = \frac{\sum_i x_i \cdot y_i + \sum_i x_i \cdot \eta_i}{\sum_i x_i^2},$$

a ponieważ wartość oczekiwana błędu wynosi zero, to dla dużej liczby pomiarów wartość estymowana parametru będzie równa wartości prawdziwej:

$$a = \frac{\sum_i x_i \cdot y}{\sum_i x_i^2} = \alpha.$$

Na razie wynik jest taki sam jak dla eksperymentu niekontrolowanego: jeśli zmienna niezależną (tę z „daszkiem") mierzymy bez błędu, parametr α otrzymujemy nieobciążony, czyli mamy gwarancję, że jeśli wykonamy dostatecznie wiele pomiarów i zrobimy na ich podstawie obliczenia, będziemy bardzo blisko prawdziwej wartości.

Przypomnijmy, że dla eksperymentu niekontrolowanego sytuacja, gdy to zmienna niezależna była mierzona z błędem, różniła się całkowicie. Estymator był bowiem obciążony wariancją zarówno wielkości x, jak i błędu. Z tego powodu studenci statystyki uczeni są, że istnieją dwie regresje: linia regresji x od y nie pokrywa się z regresją y od x. Dlatego też swój przełomowy artykuł z 1950 roku Berkson zatytułował: *Czy istnieją dwie regresje?* Pokazuje w nim bowiem – i to kulminacja całego wywodu – że w przypadku eksperymentu kontrolowanego nawet w sytuacji, gdy to zmienna zależna mierzona jest z błędem, estymator pozostaje nieobciążony! Sprawdźmy:

$$a = \frac{\sum_i u_i \cdot y_i}{\sum_i u_i^2} = \frac{\sum_i u_i \cdot (\alpha x_i + \beta)}{\sum_i u_i^2} = \frac{\sum_i u_i \cdot (\alpha \cdot (u_i - \epsilon_i) + \beta)}{\sum_i u_i^2} = \frac{\alpha \sum_i u_i^2 + \alpha \sum_i u_i \cdot \epsilon_i + \beta \sum_i u_i}{\sum_i u_i^2}.$$

Ponownie korzystając z tego, że zarówno wartość oczekiwana błędu jest równa zero jak i średnia z wartości u jest równa zero (pamiętamy, wielkości zostały przeskalowane tak, że o każdej z nich odjęto wartość średnią wszystkich pomiarów danej wielkości), należy stwierdzić, że:

$$a = \frac{\alpha \sum_i u_i^2}{\sum_i u_i^2} = \alpha.$$

I proszę! W przypadku eksperymentu kontrolowanego istnieje tylko jedna regresja!

W rzeczywistości eksperymentalnej występuje zazwyczaj jakiś miks obu rodzajów błędów. Na etapie projektowania można zastanowić się,

którą wielkość warto kontrolować, by zminimalizować całkowity błąd. Bo sprawa, jak widać, jest znacznie bardziej skomplikowana niż rzucone mimochodem „plus minus margines" czy „w granicach błędu"…

Opisany powyżej model błędu Berksona nie był bynajmniej jego błędem, a raczej sporym osiągnieciem na niwie statystyki. Tym, co współcześnie uważa się za prawdziwy błąd Berksona i położyło się cieniem – niekoniecznie słusznie – na reputacji wspaniałego naukowca, jest jego stanowisko w tak zwanej wielkiej debacie dotyczącej palenia tytoniu.

Wielka Debata

Zgodnie z regulacjami Unii Europejskiej, od 2016 roku 65% powierzchni paczki papierosów pokrywają drastyczne zdjęcia z ostrzeżeniami przed skutkami palenia. „Palenie jest przyczyną zawałów serca". „Palenie powoduje 90% przypadków raka płuc". „Palenie powoduje raka jamy ustnej i gardła". Sigmund Freud zmarł na raka jamy ustnej? Przed oczami od razu staje obraz twórcy psychoanalizy z nieodłącznym cygarem. Wiele osób wie też, że nieszczęsny Kowboj Marlboro zmarł na raka płuc[9].

Związek palenia z chorobami nowotworowymi (i nie tylko) wydaje się dziś całkowicie oczywisty. Jest uważany za jedno z kilku zaledwie żelaznych i niekwestionowanych ustaleń epidemiologicznych. Taki na przykład cholesterol wciąż miewa swoje wzloty i upadki, raz obwiniany o wszelkie zło, innym razem ułaskawiany. Palenie za to tylko i wyłącznie zbiera kolejne cięgi, będąc sukcesywnie relegowanym z życia publicznego[10].

Nie tak dawno jeszcze było jednak zupełnie inaczej. Początkowo tytoń uważany był przez wielu lekarzy i uczonych za panaceum na wszelkie dolegliwości. Wśród nich prym wiódł Jean Nicot, od którego nazwiska nazwę zaczerpnęła główna substancja czynna tytoniu i za którego radą Katarzyna Medycejska zażywała tabakę w celu uśmierzenia nawracających migren. Jeszcze w latach trzydziestych XX wieku niczym niezwykłym nie były reklamy zachwalające prozdrowotne właściwości papierosów. Przede wszystkim zalecano je w celu zrzucenia wagi i jako metodę uspokojenia nerwów. Pierwsze sprzeciwy dotyczące papierosów nie miały bynajmniej przyczyn zdrowotnych, ale raczej społeczne. W momencie bowiem, kiedy to kobiety sięgnęły po używkę zarezerwowaną dotychczas dla mężczyzn, rozległy się głosy oburzenia. Spowodowało to oczywistą reakcję – papieros stał się symbolem emancypacji. Wkrótce, nie bez zaangażowania

[9] Cała prawda jest jeszcze gorsza: trzech modeli reklamujących Marlboro zmarło na raka płuc, a czwarty – na chorobę obturacyjną.

[10] W czasie pandemii Covid-19 w 2020 roku dorzucono mu kolejny zarzut: palacze mają ciężej przechodzić zakażenie wirusem i mieć większe prawdopodobieństwo zgonu.

przemysłu tytoniowego, dla którego stuprocentowy wzrost rynku był nie do pogardzenia, żony paliły na równi z mężami.

I tak oto nadeszła znana jeszcze wielu z nas rzeczywistość. Zadymione samoloty, korytarze uczelni i pociągi. Papieros był czymś tak naturalnym, że twórcom „Nostromo" nie przyszłoby do głowy, iż ledwie kilka dekad później widok Sigourney Weaver i jej kolegów palących papierosy na pokładzie statku kosmicznego może być dla widza bardziej surrealistyczny niż ósmy pasażer.

Co bystrzejsi obserwatorzy szybko zaczęli jednakże obserwować pewne koincydencje.

Jeszcze na przełomie XIX i XX wielu rak płuc był zjawiskiem tak rzadkim, że wielu lekarzy w trakcie całej swojej praktyki nie miało okazji zetknąć się z choćby jednym przypadkiem. A potem nastąpił lawinowy wzrost.

Pierwszym, który zasugerował związek między rakiem a paleniem, był Isaac Adler (1912). Było to w drugiej dekadzie XX wieku i nic dziwnego, że jego spostrzeżeń nie przyjęto bezkrytycznie. Świat ulegał wówczas szybkim przemianom. Za gwałtowny wzrost zachorowań na nowotwory płuc mogły odpowiadać równie dobrze inne czynniki: zanieczyszczenie powietrza, będące skutkiem szybkiej industrializacji; asfaltowe wyziewy z nowo kładzionych dróg; gazy bojowe używane podczas Wielkiej Wojny czy choćby powikłania grypy hiszpanki. Bardziej rygorystyczne badania przeprowadzone zostały dopiero w nazistowskich Niemczech. W Stanach Zjednoczonych palenie było w bardzo dobrym tonie, a producenci zapewniali o jego nieszkodliwości, ale Adolf Hitler, zagorzały przeciwnik nikotynizmu, stworzył w Rzeszy sprzyjające warunki do eksploracji tematu[11].

Już pierwsze obserwacje, opublikowane w 1939 roku przez Franza Hermanna Müllera (1940), prowadzone w szpitalu w Kolonii nie pozostawiały wątpliwości: wśród chorych na raka płuc palacze stanowili znacznie większy odsetek niż w porównywalnej grupie zdrowych!

Współcześnie każdy student potrafi wypunktować słabości badania z punktu widzenia metodologii statystycznej: na jakim poziomie istotności te różnice? Czy grupa badana i kontrolne były losowe? W sposób oczywisty nie były.

Łatwo wykpiwać niedostatki warsztatu badaczy z pierwszej polowy XX wieku, ale pamiętać trzeba o dwóch sprawach. Po pierwsze, rygory metody naukowej dopiero były w trakcie tworzenie. Po drugie, w badaniach medycznych częstokroć nawet współcześnie nie jest możliwe (z przyczyn choćby etycznych) sprostanie im. Dodatkowo w kwestiach życia i śmierci szybkie działanie może być ważniejsze niż metodyka. Gdy

[11] W III Rzeszy palenie było zabronione w większości miejsc publicznych, zakazane były też reklamy tytoniu.

w 1854 roku, podczas epidemii cholery w Londynie, John Snow zauważył, że przypadki zachorowań koncentrują się wokół jednej z pomp wodnych – zamiast dywagować na temat nielosowości próby i prawdopodobieństwa popełnienia błędu któregoś rodzaju[12] – po prostu doprowadził do zakręcenia pompy, kładąc epidemii kres.

Niestety, doniesienia Müllera i podobne większość establishmentu zignorowała. A ludzie umierali. Szacuje się, iż milion wypalonych papierosów przekłada się na jedną śmierć. Jeśli porównać to z zyskiem osiąganym przez koncerny tytoniowe, można przeliczyć, że jedno życie równa się 10 tysiącom dolarów wpadającym do kabzy handlarzy chorobą (Proctor, 2012). Było więc o co walczyć.

Klinicyści byli jednakże uparci. Posypały się kolejne wyniki.

W kontynentalnej części Europy Niemcy kontynuowali badania nad wpływem palenia, wykazując jego szkodliwość, a już po II wojnie dołączyli do nich z impetem naukowcy brytyjscy i amerykańscy. Najsławniejsze, uważane za przełomowe i kładące kropkę nad „i", było badanie kohortowe, którego wyniki zostały opublikowane w 1954 roku przez Richarda Dolla i A. Bradforda Hilla z Anglii (Doll i Hill, 1954). Obiektami w tym badaniu było około 40 tysięcy brytyjskich lekarzy, którzy tworzyli „kohortę". Retrospektywnie odnotowywano dotyczące ich dane, w tym jako kluczowe – liczbę wypalanych dziennie papierosów oraz długość okresu nałogu. Lekarze byli o tyle wdzięcznym przedmiotem badania, że – z racji profesji – ich relacje w odniesieniu do stanu zdrowia, nawyków i obciążeń dziedzicznych uważane były za dokładne i wiarygodne. Następnie porównywano zachorowalność na różnego typu choroby w czterech grupach, wyodrębnionych na podstawie intensywności palenia.

Wyniki były wymowne. Sam Richard Doll, do tej pory nikotynista, z miejsca rzucił palenie, a większość establishmentu medycznego uznała zmianę konsensusu za dokonaną. Ale nie statystycy. W ich gronie zawrzało.

Oto na ich teren wkroczyli dyletanci, którzy źle traktowali ich ukochane dzieci. Którzy posługiwali się pięknie zdefiniowanymi i precyzyjnie określonymi pojęciami w sposób nieścisły, nie dorastając do wyśrubowanych rygorów, ustanowionych przez ojców – założycieli statystyki matematycznej. Trzech z nich najsilniej zaznaczyło swój głos po stronie anty-antynikotynowej: Berkson (1955, 1958, 1959), Neyman (1955) i Fisher (1958), a ich argumenty opierały się na trzech filarach: obciążeniu próby; braku przesłanek do wyciągania wniosków o przyczynowości; braku mechanizmu klinicznego.

[12] Oczywiście, te pojęcia powstały znacznie później, więc szczęśliwie dla Londyńczyków nikt nawet nie mógłby na tej podstawie podważać jego intuicji, która okazała się zbawienna dla wielu osób.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

Jeśli chodzi o obciążenie próby, prym wiódł tutaj oczywiście autor opisu paradoksu opartego na nielosowości prób szpitalnych, czyli Berkson. W tamtych czasach metodologia eksperymentalnych badań klinicznych dopiero się wykuwała. Współczesnym ideałem jest losowy przydział jednostek do dwóch lub więcej grup i poddawanie tych grup odmiennym działaniom. Próba ma być ślepa, a najlepiej podwójnie ślepa – ani badana jednostka, ani nawet eksperymentator nie mają wiedzieć, komu podawany jest na przykład lek, a komu placebo. Wszystko jest kodowane, a identyfikacja następuje dopiero po zakończeniu eksperymentu.

Ideał ideałem, ale w praktyce, przynajmniej z ludźmi, taka procedura jest niemalże niemożliwa do zastosowania. Nawet jeśli chcieć prowadzić badania, na które komisje etyki wydałyby zgodę, to uczestnictwo w eksperymencie nie może zostać nikomu nakazane, z góry zatem trzeba się liczyć z obciążeniem próby. W przypadku podejrzenia, że papierosy powodują nowotwory, nie byłoby etyczne (ani prawdopodobnie wykonalne) podzielenie niepalących na dwie grupy i jedną z nich zmuszenie do nałogu. Na szczęście możliwa była interwencja odwrotna. Richard Doll, osoba najbardziej zasłużona w dziele udowodnienia szkodliwości palenia, podzielił palaczy na dwie grupy i jedną z nich namawiał do rzucenia palenia, obserwując wpływ tej zmiany na dynamikę choroby wrzodowej żołądka. Taki eksperyment, niestety, również był daleki od ścisłych rygorów naukowych. Po pierwsze, bardzo wysoki odsetek odmów powodował, że próbie daleko było do losowości. Po drugie, z całą pewnością nie była ona ślepa.

Wszystkim zastrzeżeniom Berksona wtórował Jerzy Neyman, który wskazywał, że w badaniach szpitalnych mamy do czynienia tylko z pacjentami – no cóż – żywymi. Ci, którzy nie trafili do szpitala, nim pokonał ich nowotwór, mogliby znacząco zmienić wyniki. Być może – hipotetyzował, choć przyznawał, że z medycznego punktu widzenia jest to raczej nieprawdopodobne – palenie zwiększa przeżywalność raka płuc i to z tego względu w szpitalach wśród chorych na ten nowotwór jest więcej palaczy…

Drugim wielkim argumentem był brak wynikania przyczynowości z korelacji, i tym głównie argumentem szermował Ronald Fisher.

Współwystępowanie dwóch cech jest bardzo kuszące i naturalne jest, że próbujemy implikować pomiędzy nimi przyczynowość. Osoby, które więcej ćwiczą, mają silniejsze mięśnie – *ergo*, ćwiczenia powodują przyrost masy mięśniowej. Osoby, które więcej jedzą, mają większy obwód w pasie – *ergo*, jedzenie powoduje tycie. Osoby, które częściej grają w wojenne gry, są bardziej agresywne w życiu codziennym – zatem, agresywne gry powodują wzrost agresji… Czy faktycznie powodują? A może jest tak, że ludzie, którzy z natury są bardziej agresywni, zarówno chętniej grają w wojenne gry, jak i są gwałtowni na co dzień?

Być może pacjenci skłonni do uzależnienia się od nikotyny mają też genetyczną skłonność do zapadalności na nowotwór płuc? A może palenie

jest bardziej rozpowszechnione w pewnych kręgach społecznych (na przykład wśród klasy robotniczej), w których opieka zdrowotna jest na niższym poziomie i to z tej przyczyny zapadalność na choroby płuc wyższa? Wielki psycholog, Hans Eysenck, który też włączył się do debaty (Eysenck, Tarrant, Woolf, i England, 1960) (a sam nie stronił w swojej dziedzinie od narzędzi statystycznych), hipotetyzował, że palaczami są częściej neurotycy[13], a silne osobowości wolne są od nałogu – i że ten osobowościowy czynnik również wpływa na zapadalność na choroby płuc. Sama korelacja nie mówi nic o przyczynowości.

Aby uprawdopodobnić przyczynowość, należałoby przedstawić jakiś wiarygodny mechanizm biologiczny, co stanowiło trzeci filar krytyki stanowiska antynikotynowego.

Jak obecnie wiadomo, mechanizm powstawania każdego nowotworu jest wieloetapowy i współgrają w nim czynniki zarówno genetyczne, jak i środowiskowe – a w dużej mierze wszystko opiera się na ślepym trafie. Przy zakażeniach drobnoustrojami dużo łatwiej wykazać spełnienie postulatów Kocha – między innymi to, by czynnik chorobotwórczy obecny był w każdym przypadku i zawsze prowadził do choroby. Tymczasem nie każdy palacz (na szczęście) zapada na nowotwór i niepalenie nie zawsze chroni (niestety) przed zapadnięciem na niego. Rzeczywistość jest bardziej złożona, co nie znaczy, że żadnych wniosków wyciągnąć nie można.

Właściwie dziwić może, że Joseph Berkson postanowił zaatakować wnioski klinicystów właśnie od tej strony. „Nie będziemy naprawdę wiedzieć, czy palenie powoduje raka, dopóki nie dowiemy się czegoś precyzyjnego o sposobie, w jaki to powoduje” (Berkson, 1960), pisał w 1960 roku. Bardzo wyśrubowane wymaganie, nawet współcześnie, a co dopiero ponad pół wieku wcześniej. W 2020 roku zamknięto granice, zamrożono usługi i sporą część handlu, miliony osób objęto kwarantanną, a jeszcze większą ich liczbę – znacznymi ograniczeniami kontaktów – i to pomimo tego, że nie poznano jeszcze precyzyjnie mechanizmu, w jaki Covid-19 atakuje płuca i prowadzi do śmierci w dużo większym odsetku przypadków niż zwykła grypa.

Każde z zastrzeżeń powinno być – i było – brane na poważnie. W pewnym jednakże momencie gros naukowców postanowił wziąć pod uwagę nie tyle pojedyncze badania, ale cały korpus dowodów. Czyli zarówno tych z badań klinicznych, w tym wspominanych badań kohortowych Dolla i Hilla, jak i z badań laboratoryjnych. Wykazywano w nich związek między ekspozycją na dym papierosowy a nowotworami u zwierząt laboratoryjnych. W hodowlach komórkowych uzyskano ponadto nieprawidłowo

[13] Eysenck – jako zagorzały krytyk psychoanalizy – wrzucał tym samym kolejny kamyk do ogródka Zygmunta Freuda, zamiłowanego palacza. Ale co na to Fisher ze swoją nieodłączną fajką?

uformowane molekuły, powstające pod wpływem niektórych substancji, zawartych w dymie papierosowym.

Takie „całościowe" stanowisko przyjmował Abraham M. Lilienfeld, jeden z głównych naukowców kojarzonych z badaniami skutków palenia. Richard Doll, w komentarzu do niemal dwudziestostronicowego artykułu Berksona wykazującego niedostatki projektowania badań i wnioskowania statystycznego w pracach naukowców próbujących wykazywać związki pomiędzy paleniem a rakiem, zwięźle zauważył: „Poglądy autora są interesujące, ale nie zmieniają faktu, że względna umieralność wśród palaczy na skutek raka płuc (w porównaniu do niepalaczy) jest wielokrotnie większa niż w przypadku innych chorób" (Doll, 1959).

Trudno nie podziwiać zwięzłości i zawartości informacyjnej tego komentarza. Autor jednym zdaniem wskazuje na dużo większą zapadalność na raka płuc osób palących i odrzuca sugestię, iż mogłaby ona być powodowana ogólną gorszą kondycją zdrowotną tej grupy, mogącą wynikać z innych przyczyn, na przykład ekonomiczno-społecznych.

Koniec końców, jak wiadomo, to Richardowi Dollowi i innym z jego „narożnika" oficjalnie przyznano słuszność. W ciągu dekad nagromadziło się wiele dowodów. Poznano mechanizmy biologiczne kancerogenności tytoniu i wykazano je w badaniach laboratoryjnych. Wypracowano nowe metody analizy statystycznej. Choćby takie, które biorą pod uwagę całe zestawy różnych badań – tak zwane metaanalizy, zbierające w jedną całość wyniki różnych badaczy z różnych miejsc, dzięki czemu możliwe jest osiągnięcie wielkości prób nie do pomyślenia w pojedynczym eksperymencie.

Naukowcy z drugiej strony boiska oskarżani bywali (i wciąż są, nawet pośmiertnie) o zaprzedanie się kompaniom tytoniowym. Faktem jest, że wielu z nich było hojnie opłacanymi konsultantami, w tym sam Joseph Berkson i wielki Ronald Fisher. Nie musi to jednakże jednoznacznie wskazywać na to, by mieli głosić opinie niezgodne z ich prawdziwymi przekonaniami. W końcu sir Fisher udowodnił szczerość własnym życiem i własną śmiercią. Zagorzały palacz zmarł w wieku 72 lat z powodu powikłań pooperacyjnych. Chorował na raka okrężnicy. Jak wiadomo – silnie zależnego od tytoniu. Choć są i tacy, którzy kwestionują jego uczciwość. Podobno Fisher wyznał Davidowi Daubowi (wedle słów jego syna), iż swoje stanowisko podtrzymywał jedynie dla pieniędzy (Proctor, 2011). Czy wierzyć temu? A jeśli chodzi o Berksona? Czy on sam był palaczem? Nie udało mi się tego ustalić. Na żadnym znanym mi zdjęciu nie został uwieczniony z papierosem czy fajką. Ale ani brak zdjęcia, ani nawet ewentualny fakt niepalenia nie muszą jednakże jednoznacznie wskazywać na sprzedajność.

Wszak Berkson sam uczył, by nie wyciągać pochopnych wniosków.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 18(24)

Pokłosie

Joseph Berkson, zmarł w 1982 roku, przeżywszy 83 lata. Jego naukowy dorobek jest imponujący. Pozostawił również dobre wspomnienia. Choć biologicznych dzieci nie miał, to nawiązał dobrą więź z synem z pierwszego małżeństwa swojej żony, Susanny Cacioli. Była ona zatrudniona jako tłumaczka z włoskiego w tej samej co Berkson instytucji, czyli klinice Mayo. Para przeżyła wspólnie wiele szczęśliwych lat, utrzymując bliskie relacje z Frankiem, synem Susanny, i jego rodziną. Również koledzy po fachu – w tym i adwersarze – cenili i szanowali Berksona.

Wśród kolegów zapamiętany został jako wielki skrupulant, wzdragający się przed używaniem narzędzi statystycznych, niemających – w jego mniemaniu – dostatecznie solidnych matematycznych podstaw. A swoich racji bronił niezwykle żarliwie. Żona wspominała, jak korespondencja od innych naukowców potrafiła rozpalić jego emocje. „Nie mogą mi tego zrobić", krzyczał, biegnąc do swego biura, by spisać ciętą zazwyczaj ripostę. Ciętą i klarowną, bo słynął z wysokiej jakości swoich publikacji. „Był najlepszym pisarzem w tym interesie", twierdził inny wielki statystyk, William Gemmell Cochran. „Nigdy nie miałem wątpliwości, co Joe miał na myśli w danym zdaniu" (Armitage i Colton, 1998).

Jeszcze za swego życia, w 1978 roku, został uhonorowany w sposób zapewne szczególnie mu miły: jego imię nadano bowiem głównej sali konferencyjnej Kliniki Mayo. W trakcie uroczystości, wśród innych luminarzy, Jerzy Neyman oddał Berksonowi zasługi wprowadzenia go w meandry wielu zagadnień statystycznych, w tym analizy przeżycia i *competing risks*.

Nieco ponad pół roku po śmierci Berksona, w sierpniu 1983 r., jedna z sesji spotkania Amerykańskiego Towarzystwa Statystycznego, organizowana przez Williama S. Taylora, poświęcona została pamięci wielkiego statystyka.

Zarówno sam Taylor, jak i inna wychowanka Berksona, Lila R. Elveback, długo jeszcze po przejściu mentora na emeryturę kontynuowali tradycję statystycznej kontroli badań prowadzonych w Klinice Mayo – tradycję podtrzymywaną po dziś dzień i zapewniającą wysoką jakość publikacji firmowanych przez tę instytucję.

We wspomnieniu pośmiertnym o swoim nauczycielu Taylor napisał: „Moja pierwsza publikacja została zapoczątkowana postawionym przez Berksona pytaniem. Bardzo mi przy tym pomógł. Pomógł mi, gdy rozpoczynałem pracę, i 15 lat później, gdy przejmowałem jego funkcje w Klinice Mayo. Zawsze pomocny, nigdy natrętny, zawsze inspirujący" (Taylor, 1983).

Literatura

Adler, I. (1912). *Primary malignant growths of the lungs and bronchi*. Longmans, Green, and Company.

Armitage, P., i Colton, T. (red.). (1998). *Encyclopedia of Biostatistics*, 6 Volumes, Chichester: Wiley.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, *39*(227), 357-365.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2*(3), 47-53.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, *45*(250), 164-180.

Berkson, J. (1955). The statistical study of association between smoking and lung cancer. *In Proceedings of Staff Meetings of the Mayo Clinic* (vol. 30, no. 15, pp. 319-48).

Berkson, J. (1958). Smoking and lung cancer: some observations on two recent reports. *Journal of the American Statistical Association*, *53*(281), 28-38.

Berkson, J. (1959). The statistical investigation of smoking and cancer of the lung. *In Proceedings of Staff Meetings of the Mayo Clinic* (vol. 34, no. 8, s. 206-25).

Berkson, J. (1960, June). Smoking and cancer of the lung. W: Proceedings of the staff meetings. Mayo Clinic (vol. 35, p. 367).

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *The Annals of Statistics*, *8*(3), 457-487.

Doll, R. (1959). Komentarz do: J. Berkson, The statistical investigation of smoking and cancer of the lung. W: proceedings of staff meetings of the mayo clinic (vol. 34, no. 8, s. 206-25).

Doll, R., i Hill, A. B. (1954). The mortality of doctors in relation to their smoking habits. *British Medical Journal*, *1*(4877), 1451.

Ellenberg, J. (2017). Jak się nie pomylić, czyli potęga matematycznego myślenia, Helion.

Eysenck, H. J., Tarrant, M., Woolf, M., i England, L. (1960). Smoking and personality. *British Medical Journal*, *1*(5184), 1456.

Fisher, R. (1954a). The analysis of variance with various binomial transformations. *Biometrics*, *10*(1), 130-139.

Fisher, R. (1954b). Discussion of the analysis of variance with various binomial transformations. *Biometrics*, *10*(1), 140-151.

Fisher, R. A. (1958). Cancer and smoking. *Nature*, *182*(4635), 596-596.

Guan, Wei-jie, et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*.

Kurland, L. T., i Molgaard C. A., (1981). The patient record in epidemiology. *Scientific American*, *245*(4), 54-63.

Müller, F. H. (1940). Tabakmissbrauch und lungencarcinom. *Zeitschrift für Krebs-forschung*, *49*(1), 57-85.

Myiara, M. et al. (2020). *Low incidence of daily active tobacco smoking in patients with symptomatic COVID-19*. https://doi.org/10.32388/WPP19W.3

Neyman, J. (1955). Statistics--servant of all sciences. *Science*, *122*(3166), 401-406.

Proctor, R. N. (2012). The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, *21*(2), 87-91.

Proctor, R. N., (2011). *Golden holocaust: origins of the cigarette catastrophe and the case for abolition*. Univ of California Press.

Taylor, W. F. (1983). Obituary: Joseph Berkson, 1899-1982. *Journal of the Royal Statistical Society: Series A (General)*, *146*(4), 438-439.

Woodfine, J. D., i Redelmeier, D. A. (2015). Berkson's paradox in medical care. *Journal of Internal Medicine*, *278*(4), 424-426.