PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 93–124
DOI: 10.31648/przegldpsychologiczny.9461

Wprowadzenie do teorii wnioskowania przyczynowego dla psychologów: testowalne i nietestowalne założenia przyczynowe i statystyczne

Borysław Paulewicz[1]
Uniwersytet Jagielloński, Instytut Psychologii
https://orcid.org/0000-0002-1270-2988

Streszczenie

Celem badań podstawowych jest udzielenie odpowiedzi na pytania przyczynowe. Na ogół tylko jeden etap tego procesu, tj. analiza statystyczna, przebiega częściowo formalnie i według jasno określonych reguł, natomiast analiza relacji przyczynowych bywa niejawna i podatna na trudne do wykrycia błędy. Wprowadzenie ma pokazać psychologom, że korzystając ze współczesnej formalnej teorii wnioskowania przyczynowego, można i warto robić to inaczej. W tej części omawiam nieoczywisty status i rolę założeń przyczynowych i statystycznych we wnioskowaniu przyczynowym. Po przeanalizowaniu ogólnego schematu wnioskowania o wpływie na podstawie założeń przyczynowych, statystycznych, i wyników badania, objaśniam w zarysie i ze współczesnej perspektywy granice użyteczności regresji liniowej, a następnie wprowadzam od podstaw część formalnej teorii wnioskowania przyczynowego opartą na grafach. Korzystając z tych narzędzi, analizuję wyniki eksperymentu dotyczącego przeszukiwania pamięci krótkoterminowej i omawiam poprawki tylnych drzwi i przednich drzwi. Żeby przedstawić we względnie przystępny sposób, nie upraszczając jej przy tym nadmiernie, matematyczną część teorii, ilustruję jej sens za pomocą symulacji napisanych w coraz częściej używanym przez psychologów języku R.

Słowa kluczowe: przyczynowość, wnioskowanie przyczynowe, rachunek przyczynowy, metodologia badań, metateoria, wnioskowanie statystyczne, wnioskowanie bayesowskie

Głównym celem tego wprowadzenia jest przekonanie prowadzących badania naukowe psychologów, że formalnej teorii wnioskowania przyczynowego nie

[1] Adres do korespondencji: e-mail: boryslaw.paulewicz@uj.edu.pl. Aktualizacje dotyczące artykułu: https://github.com/boryspaulewicz/przyczynowosc1.

wypada lekceważyć, ponieważ jej twierdzenia, wynikające dedukcyjnie z aksjomatów wyrażających elementarne intuicje na temat przyczynowości, mają doniosłe znaczenie dla całego obszaru metodologii badań psychologicznych.

Starałem się zminimalizować ewentualny opór Czytelnika przed matematyką. Nie mogę jej jednak unikać, bo teoria wnioskowania przyczynowego to teoria matematyczna, co z konieczności sprawia, że poziom trudności jest zróżnicowany. Moje doświadczenia w uczeniu podstaw tej teorii wskazują na dydaktyczną użyteczność symulacji komputerowych. Współcześnie wielu psychologów potrafi takie symulacje tworzyć i modyfikować, a ułatwiają one oswojenie się z podstawowymi zagadnieniami, zanim opanuje się definicje terminów technicznych, co nie dzieje się szybko. Czytelnikowi, który nie jest przyzwyczajony do operowania wyrażeniami oznaczającymi rozkłady prawdopodobieństwa, powinno być dzięki temu łatwiej zrozumieć część matematyczną, a dzięki temu zmierzyć się z literaturą przedmiotu, w której takich wyrażeń nie brakuje.

Posługuję się tutaj przyczynowym modelem strukturalnym Pearla (ang. *Structural Causal Model*, SCM, Pearl, 2000; Pearl i in., 2016; Pearl i Mackenzie, 2021) i opartym na nim rachunkiem przyczynowym albo interwencyjnym (ang. *do-calculus*). Nie omawiam najważniejszej alternatywnej teorii o podobnym statusie, tj. teorii wyników potencjalnych Neymana-Rubina (Rubin, 2005), ponieważ wiemy już, że aksjomaty jednej teorii wynikają z aksjomatów drugiej i vice versa (Galles i Pearl, 1998), przy czym teoria Pearla jest od pewnego czasu intensywniej rozwijana, a dzięki zastosowaniu grafów i temu, że nie wszystko musi być w niej zdefiniowane w kategoriach kontrfaktycznych, nadaje się bardziej do zastosowań w psychologii.

Sposób, w jaki objaśniam tę teorię, różni się od tego, jak jest ona przedstawiana w popularnonaukowych *Przyczynach i skutkach* (Pearl i Mackenzie, 2021), w *Causal inference in statistics: A primer* (Pearl i in., 2016), który jest już pełnowartościowym podręcznikiem dla początkujących lub średnio zaawansowanych czy w wymagającym *Causality* (Pearl, 2000). Staram się konsekwentnie stosować „możliwościową" interpretację jakościowych relacji przyczynowych, podkreślam wagę nietestowalności niektórych założeń przyczynowych, sporo miejsca poświęcam założeniom i metodom statystycznym, oraz omawiam przykłady badań psychologicznych, w tym jeden szczegółowo. Odnoszę się przy okazji do dwóch najpopularniejszych polskich podręczników do statystyki i metodologii badań skierowanych do psychologów. Żeby zmniejszyć ryzyko wystąpienia nieporozumień, liberalnie korzystam z akcentowania ważnych fragmentów pismem pochyłym. Mam nadzieję, że dzięki temu to wprowadzenie będzie dopasowane do potrzeb odbiorcy.

„Korelacja to nie związek przyczynowy"

Rozważmy na początek korelację między dochodami rocznymi (X) i zadowoleniem z życia (Y). W przeciwieństwie do *deklarowanego* zadowolenia z życia zadowolenie z życia nie jest co prawda obserwowalne, ale dla uproszczenia pominę na razie wynikające stąd komplikacje.

Od nietrywialnych sądów empirycznych, będących elementami rozumowań czy teorii naukowych, nie możemy wymagać, żeby były z pewnością prawdziwe, ponieważ sądy empiryczne to nie twierdzenia matematyczne. Możemy i powinniśmy za to wymagać, żeby takie sądy były w jakimś stopniu empirycznie i teoretycznie *uzasadnione*. Wnioskowanie o wpływie w określonym kierunku na podstawie *samej* korelacji jest całkowicie nieuzasadnione, nic więc dziwnego, że w powszechnej opinii badaczy ten błąd jest poważny i elementarny. Bliższa analiza pokazuje jednak, że nie jest wcale oczywiste, kiedy i w jakim dokładnie sensie korelacja nie oznacza wpływu.

W omawianym przykładzie mamy dwa rodzaje założeń, te dotyczące planu badawczego:

1.1 Zmienna *X* jest obserwowana.

1.2 Zmienna *Y* jest obserwowana.

i te dotyczące analizy statystycznej i jej wyników:

1.3 Zmienne *X* i *Y* są dodatnio skorelowane.

Mamy też ewentualny wniosek:

1.4 *X* wpływa na *Y*.

Na mocy samych przesłanek 1.1–1.3 wniosek 1.4, że *X* wpływa na *Y*, nie jest uzasadniony, co nie znaczy, że nie jest prawdziwy. Wykryta zależność statystyczna (tutaj korelacja) może wynikać częściowo lub całkowicie z wpływu jakieś zmiennej na *X* i *Y* lub z wpływu *Y* na *X*, a założenia 1.1–1.3 nie wykluczają tych możliwości. Gdyby dodać, niezależne od wyników badania, teoretyczne lub empiryczne argumenty uzasadniające wniosek końcowy, byłby on uzasadniony ze względu na te argumenty, ale nadal nie byłby *w ogóle* uzasadniony *jako wniosek z badania*. Z perspektywy krytycznego odbiorcy taki wniosek jako wniosek z badania byłby zatem wyssany z palca.

Wnioskowanie statystyczne to stosowany rachunek prawdopodobieństwa, a ten jest teorią rozkładów, czyli, w interpretacji częstościowej, teorią relatywnych częstości występowania różnych możliwych zdarzeń, gdy proces generujący dane jest *ustalony*. Kiedy mówimy o wpływie, z konieczności rozważamy jednak, w ramach tego samego modelu, *różne możliwe procesy generujące dane*, a nie tylko częstości różnych zdarzeń. Aby wniosek przyczynowy nie był wyssany z palca, musimy więc dodać założenia *pozastatystyczne*.

Ze względu na wnioskowanie o relacjach przyczynowych w tym wypadku ważne jest, że zmienna *X* jest obserwowana, ale nie przypisywana losowo, jednak w języku statystyki tej własności nie da się wyrazić, ponieważ sama ta własność nie ma nic wspólnego z tym, jak często występują różne możliwe wartości *X*. Założenia na temat własności planu badawczego są z kolei ważne, tylko jeżeli pozwalają wykluczyć alternatywne wyjaśnienia przyczynowe korelacji. Minimalna poprawna logicznie wersja rozważanego rozumowania musi więc wyglądać tak:

2.1 *Y* nie wpływa na *X*.

2.2 *X* i *Y* nie mają wspólnych przyczyn.

2.3 *X* i *Y* są skorelowane.

Wobec tego:

2.4 *X* wpływa na *Y*.

Założenie 2.3 jest tutaj jedynym założeniem (a także wnioskiem z danych i pominiętych założeń) statystycznym, a założenia 2.1–2.2 są założeniami

przyczynowymi. Założenia 2.1 i 2.2 będą uzasadnione m.in. wtedy, gdy zmienna *X* jest przypisywana losowo (inaczej „randomizowana"). Wtedy żadna zmienna różna od mechanizmu losowania nie może wpływać na *X*, co gwarantuje prawdziwość założeń 2.1 i 2.2.

W rozumowaniu 2.1–2.4 korzystamy z metazałożenia, mówiącego, że zależność statystyczna, o ile nie jest błędem próby, musi mieć *jakieś* źródło przyczynowe. W tym sensie z korelacji co prawda wynika wpływ, ale nie wiadomo, w jakim kierunku, ani czy w jednym konkretnym kierunku. Nawet gdy obserwowana korelacja jest artefaktem, ten artefakt też musi wynikać z przyczynowej struktury procesu generującego dane. Dlatego w minimalnej poprawnej wersji rozumowania prowadzącego do wniosku o wpływie nie pojawia się, nieodgrywające ważnej roli, założenie, że *X może wpływać* na *Y*. Takie niby-założenie mogłoby zwiększyć czytelność rozumowania, ale byłoby zbędne, ponieważ „może wpływać" oznacza tutaj, że może, ale nie musi, albo że nie wiadomo, czy wpływa, czy nie.

Metazałożenie o braku przypadkowych zależności statystycznych w populacji można zastąpić słabszą wersją typu „zależność statystyczna w populacji może wynikać albo z relacji przyczynowych między zmiennymi, albo może być przypadkowa"[2]. W praktyce nie ma to znaczenia, ponieważ o ewentualnych zależnościach przypadkowych nigdy nie możemy wiedzieć na pewno, że są (w tym znaczeniu) przypadkowe.

Wyobraźmy sobie teraz, że zmienna *X* była randomizowana. Wtedy założenia o braku wpływu *Y* na *X* i o braku wspólnych przyczyn byłyby uzasadnione na mocy założenia o randomizacji, czyli na podstawie znanej własności planu badawczego. Jednocześnie żadne z uzasadnionych w takim wypadku założeń przyczynowych o braku wpływu nie byłoby wtedy *testowalne*. Już na tym prostym przykładzie widzimy więc, że model przyczynowy może być zarazem uzasadniony, użyteczny i nietestowalny na podstawie danych pochodzących z badania, którego dotyczy.

Kilka uwag na temat regresji liniowej

Jednym z celów, który przyświecał mi podczas pisania tej części, było rozwianie pewnych wątpliwości, które Czytelnik – jeżeli uczył się ze *Statystycznego drogowskazu* (Bedyńska i in., 2012) czy *Metodologii badań psychologicznych* (Brzeziński, 2022) – może mieć na temat modeli statystycznych stosowanych w psychologii. Ponieważ w rozważanym rozumowaniu była mowa o korelacji, wniosek wymagał zastosowania *regresji liniowej*, czyli modelu statystycznego o postaci ogólnej:

[2] Definicja zależności statystycznej nie wymaga, aby ta wynikała z jakiejkolwiek relacji przyczynowej. Zależność statystyczna może więc być przypadkowa przynajmniej w dwóch znaczeniach: z powodu błędu próby może się tylko wydawać, że zależność istnieje albo zależność może faktycznie występować, ale z powodu przypadkowej synchronizacji, albo dlatego, że populacja jest skończona i tak się przypadkiem składa, że wartości pewnej zmiennej nie mają takiego samego rozkładu dla każdej wartości jakiejś innej zmiennej lub zmiennych.

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_n X_n + \epsilon$$

gdzie Y to *zmienna zależna*, X_i to *predyktory* albo *zmienne niezależne*, α_i to *współczynniki regresji*, a ϵ to *błędy regresji* o rozkładzie normalnym, który ma średnią 0 i odchylenie standardowe σ. Greckie litery oznaczają tutaj *wolne parametry*, czyli nieznane na ogół wielkości, które szacujemy na podstawie danych. Wyrażenie po prawej stronie, z wyłączeniem ϵ, to *część systematyczna*, która opisuje *wartość oczekiwaną* (inaczej średnią) rozkładu Y ze względu na predyktory.

Posługuję się tutaj współczesną terminologią, zgodnie z którą model regresji, niekoniecznie liniowej, to dowolny model statystyczny *rozkładu warunkowego*. W tym wypadku jest to model rozkładu zmiennej Y, zdefiniowanego jako funkcja zmiennych X_i, traktowanych jako znane wielkości stałe. Taki rozkład zapisujemy w skrócie jako $p(Y \mid X_1,\ldots,X_n)$. Modele analizy wariancji[3] są więc również modelami regresji, ponieważ opisują rozkład warunkowy zmiennej zależnej jako funkcję czynników.

Korelacja, o której była mowa w analizowanym wcześniej rozumowaniu, jest równa nachyleniu linii regresji, jeżeli (jedyny) predyktor i zmienna zależna mają odchylenie standardowe 1, co można łatwo zagwarantować, dzieląc te dwie zmienne przez ich własne odchylenia standardowe. Jeżeli odejmiemy od obserwowanych wartości Y liczby, które wynikają z samej części systematycznej dopasowanego do danych modelu, uzyskamy *reszty*, które będą przypominać błędy regresji, ale ponieważ model jest tylko dopasowany, nie będą z nimi tożsame.

Wbrew temu, co zdają się sugerować autorzy dwóch wymienionych podręczników, ta regresja jest liniowa nie dlatego, że może opisać tylko linie proste lub płaszczyzny, ale dlatego, że jest *liniowa w parametrach*. Na przykład, model $Y = X^{\alpha} + \epsilon$ jest *nieliniowy* (w parametrach), ale $Y = \alpha \sin^2(X) + \epsilon$, gdzie $\sin^2(X)$ to całkiem zwyczajny predyktor, tyle że przekształcony, to model *krzywoliniowy* (ze względu na X), który jest również modelem liniowym (w parametrach), ponieważ oznaczając $\sin^2(X)$ przez Z, możemy go zapisać jako $Y = \alpha_0 + \alpha_1 Z_1 + \epsilon$, spełniając w ten sposób wymagania definicji regresji liniowej.

W badaniach psychologicznych regresja liniowa nigdy nie jest modelem prawdziwym, bo nie może być. Zmienne obserwowalne, którymi posługujemy się w psychologii, są zwykle dyskretne i obustronnie ograniczone, a jeżeli są ciągłe, to są przynajmniej jednostronnie ograniczone, np. czas reakcji jest zawsze dodatni. Błędy regresji nie mogą więc mieć rozkładu normalnego, bo ten jest ciągły i nie ma granic. Żadne przekształcenie nie zagwarantuje przy tym, że rozkład błędów stanie się normalny. Żeby takie przekształcenie było znane, rozkład również musiałby być znany, a o rozkładach zmiennych ciągłych badanych w psychologii wiemy niewiele. Dlatego zalecane w wymienionych podręcznikach testowanie odstępstw od normalności rozkładu reszt, ponieważ w dodatku nie nadaje się nawet

[3] O modelach analizy wariancji Brzeziński (2022) twierdzi, że te nadają się do analizy wyników badań eksperymentalnych, natomiast modele regresji nadają się jego zdaniem do analizy wyników badań obserwacyjnych, które nazywa badaniami „odpowiadającymi modelowi korelacyjnemu".

dobrze do ustalania, czy lub w jakim stopniu te odstępstwa wymagają zastosowania metod alternatywnych, nie ma za wiele sensu. Można się o tym dowiedzieć m.in. z podręczników do współczesnych metod odpornych albo odpornościowych (np. Wilcox, 2011), czyli metod wnioskowania w warunkach, w których założenia modelu statystycznego mogą nie być spełnione. Metody odporne są współczesnymi alternatywami dla zalecanych przez autorów dwóch wymienionych podręczników klasycznych, nie opartych na teorii odporności, metod nieparametrycznych.

Oszacowania punktowe współczynników regresji uzyskujemy często metodą najmniejszych kwadratów (MNK). Co będzie ważne w dalszej części, *oszacowania MNK mają deskryptywny sens również wtedy, gdy założenia regresji liniowej n i e są spełnione*. Ta uniwersalna deskryptywna użyteczność wynika z tego, że uzyskane metodą MNK współczynniki regresji zawsze odpowiadają projekcji ortogonalnej wektora zmiennej zależnej na płaszczyznę liniową rozpinaną przez wektory predyktorów[4]. Dlatego oszacowanie MNK z próby np. punktu przecięcia (inaczej wyrazu wolnego) i nachylenia jest nieobciążonym oszacowaniem *takiej linii prostej, która minimalizuje* średni *kwadrat błędu*. Jest tak niezależnie od tego, czy rozkład błędów jest normalny i ma jednorodną wariancję (w psychologii nigdy nie jest i prawie nigdy nie może mieć), albo od tego, czy dane są niezależne ze względu na model (zwykle nie można być tego pewnym, zob. np. Greenland, 2022), albo od tego, czy prawdziwa zależność daje się najlepiej opisać równaniem liniowym (prawie na pewno nie), czy może raczej liniowe (w parametrach) efekty istnieją tylko „wirtualnie" jako „trendy". Zawsze można więc powiedzieć, że korelacja wyraża stopień *liniowej* współzmienności, chociaż nie zawsze można powiedzieć, że korelacja wyraża stopień współzmienności. Podobnie średnia z próby jest *zawsze* nieobciążonym oszacowaniem średniej rozkładu, z którego próba pochodzi, co wynika z liniowości operatora wartości oczekiwanej. Brak obciążenia to własność polegająca na tym, że w hipotetycznym nieskończonym ciągu replikacji tego samego doświadczenia oszacowanie jest średnio równe temu, co szacuje, czyli w pewnym sensie szacuje to, co powinno.

Z centralnego twierdzenia granicznego wynika, że w szerokim zakresie sytuacji, rozkład ważonej sumy zmiennych losowych o niekoniecznie takich samych rozkładach będzie się zbliżał do rozkładu normalnego wraz ze wzrostem liczby sumowanych zmiennych. Oszacowania MNK współczynników regresji liniowej są ważonymi sumami zmiennych losowych, a więc niezależnie od prawdziwości modelu oszacowania *przedziałowe* (przedziały ufności) i odpowiadające tym przedziałom poziomy istotności będą często w najgorszym razie bardziej konserwatywne (szerokie), niż byłoby to osiągalne za pomocą innych metod. Ponieważ w psychologii prawie zawsze z góry wiadomo, że model statystyczny nie jest prawdziwy, oszacowania przedziałowe i tak można interpretować tylko jako przybliżone miary niepewności dotyczące własności uproszczonego opisu rozkładu[5].

[4] O ile za predyktor uznamy również wektor złożony z samych jedynek, odpowiadający wyrazowi wolnemu, gdy taki występuje. Znakomite wprowadzenie do geometrycznie rozumianych modeli liniowych znajdzie Czytelnik u Saville'a i Wooda (2012).

[5] Drugie wydanie podręcznika *Statistical rethinking*, dotyczącego jednocześnie wnioskowania bayesowskiego i przyczynowego, jest znakomitym przykładem konsekwentnego stosowania tego sposobu myślenia o modelach statystycznych (McElreath, 2020).

W przeciwieństwie do, często katastrofalnych, konsekwencji błędnych założeń przyczynowych, tego rodzaju problemy można jednak minimalizować, zwiększając wielkość próby albo zmieniając metodę analizy.

Jednym z ważnych powodów, dla których należy się przejmować *stopniem i charakterem* odstępstw od założeń modelu statystycznego, a nie tym, *czy* te założenia są spełnione, bo na ogół nie mogą być, a nawet, gdyby były i tak nie można się upewnić, że są spełnione, jest *efektywność estymatora*. To jest stopień, w jakim oszacowania punktowe są przeciętnie rozrzucone (niekoniecznie wokół wartości prawdziwych!) w hipotetycznych replikacjach tego samego doświadczenia. Chcemy, żeby oszacowania punktowe były przeciętnie jak najmniej rozrzucone, w dodatku wokół wartości prawdziwych (czyli wartości odpowiadających *uproszczonemu, ale dobrze zdefiniowanemu opisowi rozkładu albo populacji*) i żeby związana z nimi niepewność była możliwie mała. Częste występowanie znacznych odstępstw od założeń sprawia, że metody odporne w psychologii działają pod obydwoma względami lepiej niż zwykła regresja liniowa[6], o której wiadomo, że jest w pewnym technicznym sensie najmniej odporna (Field i Wilcox, 2017; Wilcox, 2011).

Dobór metody statystycznej jest zagadnieniem złożonym i nie da się podać prostego algorytmu, pozwalającego w typowych sytuacjach wskazać najlepsze rozwiązanie. Dlatego Czytelnika, który najpewniej stosuje głównie regresję liniową lub jej uogólnienia, być może pocieszy fakt, że, zachowując odpowiednie środki ostrożności, o których z braku miejsca nie mogę tutaj więcej powiedzieć, zwykle można na tym modelu polegać. W dopasowanym *nieprawdziwym* liniowym modelu statystycznym może bowiem nie być nic *błędnego* w tym znaczeniu, że możemy wyprowadzić za jego pomocą uzasadnione i poprawne wnioski, o ile tylko poprawnie interpretujemy ten model jako *część metody szacowania uproszczonego opisu rozkładu*.

Dotyczy to również większości sytuacji, w których skala zmiennej zależnej zdawałaby się na to nie pozwalać. W interpretacji częstościowej o zdarzeniach pewnego rodzaju mówimy, że mają rozkład, zakładając, że określone jest, być może niejawnie, jakieś teoretycznie powtarzalne doświadczenie. Gdy doświadczenie jest ustalone, spełnione są pewne warunki ogólne[7], i możliwym wynikom przypisujemy w jakiś sposób liczby, powołujemy do istnienia rozkład o dobrze zdefiniowanej średniej. Na przykład, średnia z losowej próby zakodowanej jako 0 lub 1 *nominalnej* zmiennej płeć biologiczna jest nieobciążonym oszacowaniem średniej w populacji liczb 0 i 1, gdy te są przypisane do płci biologicznej wszystkich osób w populacji. Mimo tego, że zmienna jest nominalna, ta średnia ma oczywistą interpretację jako proporcja kobiet (lub mężczyzn) w populacji. Używanie modelu liniowego do opisu rozkładu zmiennej nominalnej o więcej niż dwóch wartościach jest co prawda niewygodne (można to zrobić, np. kodując każdy poziom za pomocą osobnej binarnej zmiennej wskaźnikowej), ale takie zmienne jako zmienne zależne występują w psychologii stosunkowo rzadko.

[6] Dzieje się tak jednak zwykle kosztem pewnego obciążenia.

[7] Dodałem to zastrzeżenie, ponieważ nie każdy rozkład wartości liczbowych ma dobrze określoną średnią, ten problem nie dotyczy jednak skali pomiarowej.

To, jak bardzo należy się przejmować skalą porządkową, jest do pewnego stopnia kwestią sporną (zob. np. Liddell i Kruschke, 2018; Paulewicz i Blaut, 2022). Zwracam jednak uwagę, że nie da się nawet wysłowić „problemu skali porządkowej", nie zakładając istnienia *dwóch* zmiennych – tej, o wartościach której mają mówić coś wartości liczbowe przypisane realizacjom zmiennej porządkowej i samej zmiennej porządkowej, zakładając przy tym zwykle, że ta pierwsza zmienna wpływa na tę drugą. Na przykład, takie same różnice w miejscach na podium interpretowanych jako liczby 1, 2 i 3 nie będą odpowiadały takim samym różnicom w czasach ukończenia wyścigu, które są przyczynami miejsc na podium. To są jednak dwie różne zmienne; dopóki interesuje nas *obserwowana zmienna porządkowa jako taka*, np. wyrażona liczbą całkowitą odpowiedź na skali Likerta, a nie jej hipotetyczne *nieobserwowane źródło*, np. faktyczny stopień przekonania, problem skali porządkowej w ogóle się nie pojawia, a współczynniki MNK regresji liniowej są nieobciążonymi oszacowaniami dobrze zdefiniowanych – chociaż psychologicznie mniej interesujących – ilościowych własności uproszczonego opisu rozkładu warunkowego takiej zmiennej.

Przyczynowy charakter symulacji komputerowych i pojęcie interwencji

Wprowadzenie do wnioskowania przyczynowego warto zaczynać od omówienia symulacji komputerowych, ponieważ komputer jest urządzeniem programowalnym, a programowanie polega na *zmienianiu sposobu, w jaki komputer działa*. W szczególności, za pomocą symulacji można stosunkowo łatwo ilustrować – ale nie dowodzić, bo symulacja to nie dowód – sens i prawdziwość twierdzeń rachunku przyczynowego.

Wykonanie poniższego kodu napisanego w języku R (R Core Team, 2022) uruchamia proces, w którym jedyną relacją przyczynową między X i Y jest wpływ X na Y. Dla uproszczenia wszystkie efekty są liniowe, a wszystkie zmienne „całkowicie losowe" (wkrótce nazwę je inaczej) mają standardowy rozkład normalny.

```
set.seed(1234)
n = 10000
U_X = rnorm(n)
U_Y = rnorm(n)
X = U_X
Y = 1 + 2 * X + U_Y
```

Pierwsze dwie instrukcje przygotowują warunki symulacji: instrukcja `set.seed(1234)` ustawia tzw. ziarno losowania na wybraną arbitralnie liczbę 1234, dzięki czemu przebieg symulacji jest powtarzalny, a `n` jest nazwą dla ustalonej liczby próbek. Instrukcja `rnorm(n)` generuje pseudolosowe próbki ze standardowego (czyli o średniej 0 i wariancji 1) rozkładu normalnego (`rnorm` to skrót od „random normal").

Znak równości nie oznacza tutaj równości matematycznej, tylko *operację przypisania wartości do zmiennej*. Wykonanie tej operacji polega na tym, że interpreter języka R ewaluuje wyrażenie po prawej stronie, np. tekst „10000" w instrukcji `n = 10000` jest interpretowany jako liczba 10000 i powstaje reprezentacja tej liczby w pamięci komputera, a uzyskana wartość jest zapisana w zmiennej występującej po stronie lewej. To, co pojawia się po prawej stronie operacji przypisania, jest więc *przyczyną* stanu czy wartości zmiennej, która pojawia się po stronie lewej. Instrukcja `U_X = rnorm(n)` sprawia, że powstaje n pseudolosowych próbek ze standardowego rozkładu normalnego, które stają się wartościami zmiennej U_X, a następująca po niej instrukcja `U_Y = rnorm(n)` sprawia, że powstaje n *nowych i niezależnych* próbek z rozkładu normalnego, które stają się wartościami U_Y. W symulowanym procesie Y ani bezpośrednio, ani pośrednio nie wpływa na X, ponieważ wartość wyrażenia określającego proces powstawania wartości X nie zależy ani bezpośrednio, ani pośrednio od Y. Widać też, że X i Y nie mają wspólnych przyczyn.

Zmienne powstające ze względu na dany kontekst (tutaj pseudo-) losowo, albo w bliżej nieokreślony sposób, we wnioskowaniu przyczynowym nazywamy *egzogennymi*, czyli pochodzącymi z zewnątrz (modelu lub modelowanego procesu). Zmienne generowane przez nie(pseudo)losową część kodu lub procesu nazywamy *endogennymi*, czyli generowanymi wewnątrz (modelu). Wybór zmiennych endogennych jest równoznaczny ze wskazaniem fragmentu rzeczywistości poddawanego analizie przyczynowej, jest więc arbitralny i wynika zwykle z tego, co akurat interesuje badacza, i z tego, jak badacz wyobraża sobie ewentualne nieobserwowalne konstrukty teoretyczne.

Zgodnie z konwencją stosowaną czasem w literaturze zmienne endogenne, które będę nazywał *modelowanymi*, oznaczam dużymi literami z wyłączeniem litery U, a odpowiadające im zmienne egzogenne, które będę nazywał *niemodelowanymi*, oznaczam literą U z odpowiednim indeksem dolnym. Nie będę odtąd nazywał zmiennych egzogennymi lub endogennymi, żeby podkreślić, że to rozróżnienie nie dotyczy samych zmiennych, tylko sposobu, w jaki są traktowane. Ponieważ dla każdej zmiennej modelowanej V, odpowiadająca jej zmienna niemodelowana U_V reprezentuje *wszystkie* niemodelowane przyczyny V, to jeżeli V nie ma żadnych przyczyn modelowanych, jak tutaj X, to U_V reprezentuje po prostu wszystkie przyczyny V, a więc zbiór wartości V można bez utraty ogólności utożsamić ze zbiorem wartości U_V (tutaj $X = U_X$).

Ponieważ tworzymy proces generujący dane, możemy wiernie *symulować skutki interwencji*, zamieniając wybrane instrukcje na wartości stałe. Na przykład, fizyczne ustalenie wartości X na 44 odpowiada następującej wersji procesu:

```
set.seed(1234)
n = 10000
U_X = rnorm(n)
U_Y = rnorm(n)
X = 44
Y = 1 + 2 * X + U_Y
```

Jeżeli W i V to zmienne modelowane lub zbiory takich zmiennych, to wyrażenie $p\big(W|do(V=v)\big)$ oznacza *rozkład interwencyjny*, generowany przez proces powstający z wyjściowego przez zastąpienie źródeł zmienności V stałą (lub wektorem, jeżeli V to zbiór) v. Powyższa symulacja generuje więc próbki z rozkładu $p\big(Y|do(X=44)\big)$. Tego rodzaju rozkład reprezentuje tzw. *całkowity* efekt przyczynowy interwencji. Z kodu wynika, że interwencja $do(Y=y)$ nie może zmienić rozkładu X, tj. $p\big(X|do(Y=y)\big)=p(X)$ dla każdego y, czyli (całkowity) efekt przyczynowy Y na X nie występuje.

Formalny język rachunku przyczynowego różni się od języka rachunku prawdopodobieństwa tylko obecnością operatora do[8]. Ten abstrakcyjny operator pozwala definiować również interwencje, które nie mają definicji operacyjnej, takie jak hipotetyczne interwencje wpływające *tylko* na ciśnienie krwi albo płeć. W ten sposób można formułować pytania przyczynowe, na które nie da się wprost odpowiedzieć eksperymentalnie i mimo to można, w sprzyjających warunkach, znaleźć uzasadnione odpowiedzi na takie pytania.

Struktura i interpretacja grafów przyczynowych

Mimo utrudniającej badanie wielowymiarowej intra- i interindywidualnej zmienności nieobserwowalnego psychologicznego procesu reagowania *zawsze* możemy w taki sposób wymienić wykluczone i niewykluczone, ale poza tym bliżej nieokreślone ilościowo, relacje przyczynowe między zmiennymi obserwowanymi, nieobserwowanymi i zmiennymi nieobserwowalnymi, które wyobraża sobie część społeczności badaczy, żeby mieć powody sądzić, że wszystkie te założenia są *prawdziwe*. Ponadto na podstawie m.in. dowodów selektywnego wpływu, dysocjacji, albo interferencji (Sternberg, 2001) możemy czasem zasadnie wnosić o istnieniu odrębnych, scharakteryzowanych *jakościowo* struktur latentnych, tj. podsystemów, modułów, komponentów, procesów, albo etapów procesu reagowania. Na podstawie tego rodzaju wzorców wyników mamy np. powody sądzić, że istnieją różne rodzaje pamięci albo że na etapie selekcji reakcji czy podejmowania decyzji może występować wąskie gardło (Levy i in., 2006).

Dlatego dla psychologów szczególnie przydatna jest część teorii dotycząca samych jakościowych relacji przyczynowych, które w teorii Pearla są reprezentowane za pomocą *grafów skierowanych*, tj. takich, na których każda *krawędź*, łącząca zawsze dwa *wierzchołki*, inaczej *węzły*, ma kierunek, który oznaczamy grotem strzałki. Dla wygody posługujemy się również łukami, tj. połączeniami dwukierunkowymi, które oznaczają możliwość istnienia bliżej nieokreślonej wspólnej przyczyny. Z powodu komplikacji, jakie się z tym wiążą, nie będę rozważał tutaj grafów *cyklicznych*, tj. takich, że dla przynajmniej jednego wierzchołka da się na nich wrócić do tego wierzchołka, idąc zgodnie z kierunkiem strzałek.

[8] Tłumacz *Przyczyn i skutków* (Pearl i Mackenzie, 2021) zaproponował polską wersję „wykonaj" zamiast oryginalnego „do", która jednak staje się moim zdaniem uciążliwa, gdy trzeba operować złożonymi wyrażeniami interwencyjnymi.

Stosowanie grafów przyczynowych w najgorszym razie pozwoli dostrzec, że z powodu braku dostatecznej znajomości badanych procesów niewykluczonych relacji wpływu jest zbyt wiele, żeby dało się oszacować poszukiwane wielkości. Taki rezultat może być zniechęcający, ale warto o nim wiedzieć, bo pozwala uniknąć formułowania lub nadmiernego przywiązania do sformułowanych przez innych autorów nieuzasadnionych wniosków. Jeżeli okaże się, że pewne efekty przyczynowe da się oszacować, pozostanie tylko ustalić ogólną postać estymatora (wyrażoną w kategoriach bliżej nieokreślonych rozkładów zmiennych obserwowanych) i znaleźć dla tej postaci dobre statystyczne przybliżenie (czyli zwykle funkcję uproszczonych opisów takich rozkładów). W najprostszych sytuacjach będzie to polegało na dopasowaniu jakiegoś modelu regresji.

Symulowany wcześniej proces można przedstawić za pomocą grafu przyczynowego $X \rightarrow Y$. Jeżeli nie stwierdzamy wyraźnie inaczej, strzałki na takich grafach interpretujemy jako *teoretyczną możliwość* wpływu, a więc każda strzałka lub łuk to *brak* założenia przyczynowego. Ponieważ zachowując sens grafu, można nad każdą strzałką i każdym łukiem napisać „nie wiadomo", oznaczonych krawędzi nie trzeba uzasadniać. To braki dających się narysować na grafie strzałek lub łuków odpowiadają wymagającym uzasadnienia założeniom przyczynowym, bo to właśnie braki strzałek lub łuków mają kategoryczne konsekwencje statystyczne i umożliwiają szacowanie wielkości przyczynowych. Strzałki oznaczają możliwość wpływu *bezpośredniego*, gdzie „bezpośredni" nie znaczy natychmiastowy, tylko nie zapośredniczony przez inne zmienne w modelu.

Zmiennych niemodelowanych zwykle nie oznaczamy, ponieważ można wywnioskować, w jakie relacje przyczynowe wchodzą. Musimy jednak oznaczyć zmienne niemodelowane, które mogą być zależne statystycznie. Za wyjątkiem zależności pozornej wynikającej z działania zderzacza, o której powiem później, na mocy metazałożenia „nie ma korelacji bez przyczynowości", zależność między zmiennymi niemodelowanymi jest możliwa tylko wtedy, gdy jedna z nich wpływa na drugą, lub gdy obie mają wspólną przyczynę. Taka zależność ma ważne konsekwencje, dlatego musimy ją oznaczyć również wtedy, gdy pomijamy zmienne niemodelowane. Robimy to za pomocą dwukierunkowego łuku ponieważ wynika z niej, że dwie zmienne modelowane mają co najmniej jedną wspólną przyczynę niemodelowaną. Graf $X \rightarrow Y$ to zatem uproszczona wersja grafu $U_X \dashrightarrow X \rightarrow Y \dashleftarrow U_Y$, gdzie zmienne U_X i U_Y są z założenia niezależne, ponieważ nie łączy ich (ani zmiennych X i Y na grafie uproszczonym) łuk.

Graf przyczynowy jest reprezentacją jakościowych relacji przyczynowych w tym znaczniu, że zakładając sam graf, nie zakładamy niczego na temat ilościowych własności procesu, dzięki czemu zawsze można skonstruować prawdziwy lub prawdopodobnie prawdziwy graf badania psychologicznego. W szczególności, inaczej niż w przypadku typowych modeli SEM[9] (Blalock, 2018; Bollen, 1989; Duncan, 2014; Wright, 1921), które są z założenia liniowe i nie dopuszczają efektów interakcyjnych, w jakościowych modelach przyczynowych dopuszczamy

[9] Tak będę nazywał liniowe modele równań strukturalnych, żeby odróżnić je od ogólniejszych strukturalnych modeli przyczynowych.

relacje liniowe (w tym krzywoliniowe) i nieliniowe, poza ewentualną niezależnością nie zakładamy niczego na temat rozkładów zmiennych niemodelowanych, a gdy jakaś zmienna jest pod bezpośrednim wpływem więcej niż jednej zmiennej, dopuszczamy wpływ interakcyjny.

Na mocy aksjomatów *teoria gwarantuje prawdziwość wniosków opartych na grafie, o ile nie brakuje strzałki lub łuku odpowiadającego faktycznie zachodzącej relacji wpływu*. W szczególności, gdy strzałek lub łuków jest więcej niż potrzeba, w znaczeniu, że nie wszystkie odpowiadają relacjom rzeczywistego wpływu, ale nie brakuje żadnych strzałek ani łuków, wnioski nadal będą prawdziwe, tyle tylko, że być może mniej będzie z grafu wynikało i mniej będzie się dało wywnioskować z danych. Możliwościowa interpretacja strzałek pozwala również uwzględniać na grafie konstrukty teoretyczne, tj. zmienne nieobserwowalne, co do których nie można mieć pewności, że istnieją, nie sprawiając, że graf staje się fałszywy, jeżeli taka zmienna nie istnieje[10].

Analiza grafów przyczynowych to przede wszystkim analiza ścieżek (ang. *path*), czyli niepustych i skończonych (w tym również jednoelementowych) ciągów stykających się strzałek bez powtórzeń[11], w których kierunek może się zmieniać. Analiza własności ścieżek sprowadza się często do korzystania z własności: łańcucha $X \to Y \to Z$ (ang. *chain*), rozwidlenia $X \leftarrow Y \to Z$ (ang. *fork*) i zderzacza $X \to Y \leftarrow Z$ (ang. *collider*). Zapamiętanie własności tych trzech struktur bardzo ułatwia korzystanie z części teorii opartej na grafach. Ponieważ ścieżki wyglądają jak grafy, ale są z założenia tylko częściami (być może bliżej nieokreślonych) grafów, od tego momentu Czytelnik będzie musiał zwracać uwagę na to, czy rozważane struktury przyczynowe są ścieżkami czy grafami.

Jeżeli X i Z łączy łańcuch $X \to Y \to Z$, to X i Z mogą, ale nie muszą, być zależne statystycznie. Jeżeli X i Z nie łączy inna ścieżka bez zderzacza, to X i Z muszą być niezależne w każdej *warstwie Y,* tj. w każdym podzbiorze populacji, w którym *naturalnie występuje* albo tylko jest *obserwowana* jakaś konkretna wartość Y: dla każdego y, gdy popatrzymy na podzbiór losowych próbek taki, że $Y = y$, z dokładnością do błędu próby nie będzie widać zależności między X i Z. Mówiąc jeszcze inaczej, dla każdego y, X i Z będą niezależne w rozkładzie warunkowym $p(X, Z|y)$, co zapisujemy jako $p(Z|X, Y) = p(Z|Y)$ lub krócej jako $X \perp\!\!\!\perp Z|Y$. Wreszcie, w każdej warstwie *potomka* zmiennej Y na grafie, którego częścią jest ten łańcuch, zależność statystyczna między X i Z może być osłabiona. *Wpływ* nie ulegnie wtedy zmianie, ponieważ warstwowanie to z definicji selektywna obserwacja, która nie zmienia *przebiegu procesu*, tylko sprawia, że *inaczej patrzymy na wynik procesu*. O tyle, o ile warstwowanie po potomku Y sprawi, że w danych nie ujawni się w pełni zmienność Y, ewentualny wpływ zapośredniczony przez Y być może również nie będzie mógł się w pełni ujawnić, chociaż będzie przebiegał bez zmian.

[10] Można np. przyjąć konwencję, że zmienne oznaczające nieistniejące konstrukty teoretyczne to de facto stałe.

[11] Ścieżka nie zawsze jest definiowana jako ciąg bez powtórzeń, a ogólnie w teorii (niekoniecznie przyczynowych) grafów definicja ścieżki może dopuszczać, aby ciąg był pusty.

Wszystkie wymienione własności łańcucha możemy zilustrować za pomocą symulacji. Dla uproszczenia możemy przyjąć, że efekty są liniowe, punkty przecięcia są równe 0, nachylenia są równe 1 i że każda zmienna niemodelowana ma standardowy rozkład normalny. Trzymając się tej konwencji, łańcuch $X \to Y \to Z$ interpretowany jako graf przyczynowy możemy „zrealizować" tak:

```
X = rnorm(n)
Y = rnorm(n) + X
Z = rnorm(n) + Y
```

Żeby łatwiej było zobaczyć w kodzie odpowiadający mu graf, pozbyłem się tutaj powtarzających się instrukcji (resetującej ziarno losowania i ustalającej liczbę próbek), nie nadałem nazw zmiennym niemodelowanym i zmieniłem kolejność sumowanych wyrażeń. Możemy się teraz przekonać, że, z dokładnością do błędu próby, kontrolowanie statystyczne Y sprawia, że zależność między X i Z nie jest istotna statystycznie:

```
confint(lm(Z ~ X + Y))
#              2.5%  97.5%
# (Intercept) -0.02  0.02
# X           -0.04  0.01
# Y            0.98  1.02
```

Znak # jest interpretowany jako początek komentarza i sprawia, że R ignoruje tekst, który pojawia się w linijce za tym znakiem. Odtąd będę w ten sposób dodawał do kodu wyniki ewaluacji wyrażeń (zaokrąglone do dwóch miejsc po przecinku). Funkcja `lm` dopasowuje model liniowy, a `confint` zwraca domyślnie 95-procentowe przedziały ufności dla wszystkich efektów. Ponieważ 95-procentowe przedziały ufności dla nachylenia X zawierają 0, to nachylenie nie jest istotne na poziomie $\alpha = ,05$. Poza tym widzimy, że efekt statystyczny Y na Z jest zgodny z efektem przyczynowym Y na Z, danym przez nachylenie równe 1. Możemy się też przekonać, co powinno zainteresować osoby stosujące analizę mediacji, że kontrolowanie statystyczne zmiennej będącej potomkiem Y, którą można interpretować jako idealnie trafny (chociaż nie „idealnie rzetelny") pomiar Y, działa inaczej niż kontrolowanie Y:

```
V = rnorm(n) + Y
confint(lm(Z ~ X + V))
#              2.5%  97.5%
# (Intercept) -0.03  0.02
# X            0.46  0.52
# V            0.48  0.52
```

Jak widać, przedziały ufności wokół nachylenia dla V nie mają nic wspólnego z efektem przyczynowym Y, a efekt statystyczny X pozostaje istotny, co – nie znając struktury procesu – można by błędnie zinterpretować jako powód odrzucenia założenia o mediacji całkowitej.

Pod względem statystycznym rozwidlenie $X \leftarrow Y \rightarrow Z$ zachowuje się tak samo jak łańcuch, tzn. jedyną testowalną konsekwencją obydwu ścieżek jako grafów jest $X \perp\!\!\!\perp Z|Y$ i warstwowanie po potomku Y może osłabić obserwowaną zależność między X i Z. Te dwie ścieżki są więc *obserwacyjnie nieodróżnialne jako grafy*.

Zderzacz zachowuje się niemal całkowicie odwrotnie i zarazem nieintuicyjnie, dlatego zbiór jego własności nazywamy paradoksem Berksona (Berkson, 1946). Jeżeli X i Z łączy ścieżka ze zderzaczem i nie łączy żadna ścieżka bez zderzacza, to X i Z są niezależne, ale *mogą być zależne w warstwach Y*. Na przykład, jeżeli popatrzymy tylko na te próby dwóch niezależnych rzutów kostką X i Z, dla których suma Y jest parzysta, gdzie suma jest skutkiem X i Z, to z tego, że X jest liczbą parzystą, od razu będziemy mogli wywnioskować, że Z też jest liczbą parzystą, czyli wtedy nieprawda, że $X \perp\!\!\!\perp Z|Y$. Patrząc przez warstwy zmiennej na jej niezależne przyczyny, możemy więc zobaczyć – i zwykle zobaczymy – zależność pozorną (systematycznie zniekształconą) między przyczynami. Jedyną kategoryczną testowalną konsekwencją grafu $X \rightarrow Y \leftarrow Z$ jest jednak $X \perp\!\!\!\perp Z$. Czytelnik powinien być w tym momencie w stanie napisać kod symulacji zderzacza ilustrującej paradoks Berksona, do czego osoby początkujące zachęcam, ponieważ nieintuicyjne własności tej często występującej struktury warto sobie dobrze przyswoić.

Żeby od razu skorzystać z wprowadzonej części teorii, zastanówmy się nad konsekwencjami występowania w badaniu zderzacza, nazywanymi stronniczością próby. Dobór próby w badaniach psychologicznych zwykle polega na korzystaniu z osób, do których badacz ma wygodny dostęp. Jeżeli oznaczymy zbiór zmiennych, które interesują badacza, jako Z, a zbiór wszystkich innych zmiennych takich, że z powodu sposobu pobierania próby, pod względem tych zmiennych tego rodzaju próby są specyficzne, przez X, to *dla każdej pary zmiennych V i W należących do Z, jeżeli V i W wpływają na jakąś zmienną w zbiorze X, zależność statystyczna między V i W będzie w tej próbie systematycznie zniekształcona*.

Rozważmy badanie dotyczące związku między płcią biologiczną urodzeniową (P) i inteligencją ogólną (I). Ponieważ, zgodnie z obecną wiedzą, płeć biologiczna jest wynikiem działania mechanizmu losowego, regresja niemal każdej zmiennej Y (w tym też I) na P szacuje wpływ P na Y, czyli $p(Y|P = p) = p\big(Y|do(P = p)\big)$, mimo tego, że badanie jest w zasadzie obserwacyjne. Jeżeli jednak *proces pobierania próbek* będzie taki, że do badania, z większym lub mniejszym prawdopodobieństwem niż wynosi proporcja studentów psychologii w populacji, będą trafiali studenci psychologii, to ponieważ zarówno inteligencja, jak i płeć z pewnością mają (silny) wpływ na to, czy ktoś jest studentem (częściej studentką) psychologii, już tylko z tego powodu zależność statystyczna między płcią i inteligencją będzie systematycznie zniekształcona.

Ponieważ mogą naturalnie generować zależność statystyczną, ścieżki bez zderzacza nazywamy *aktywnymi*, a ścieżki ze zderzaczem nazywamy nieaktywnymi. Zależności statystyczne wynikające z paradoksu Berksona nazywamy *pozornymi* (ang. *spurious*), a te wynikające ze ścieżek z rozwidleniem nazywamy czasem, czego akurat unikam, „nieprzyczynowymi" (ang. *noncausal*).

Graf przyczynowy mówi to samo, co lista tzw. *nieparametrycznych*, w znaczeniu abstrakcyjnych lub bliżej nieokreślonych, *równań strukturalnych*, nazywanych tak, ponieważ w przeciwieństwie do zwykłych symetrycznych równań

matematycznych opisują asymetryczne relacje wpływu. Na przykład, graf $X \rightarrow Y \leftarrow Z$ wyraża te same założenia, co następujący nieparametryczny model strukturalny:

$$X = f_X(U_X)$$
$$Z = f_Z(U_Z)$$
$$Y = f_Y(X,Z,U_Y)$$

gdzie, jak wcześniej wyjaśniłem, możemy przyjąć, że $f_X(U_X) = U_X$ i $f_Z(U_Z) = U_Z$. Zarówno graf $X \rightarrow Y \leftarrow Z$, jak i powyższy model strukturalny to zatem dwie reprezentacje kategorycznych założeń o braku wpływu bezpośredniego X na Z, Z na X, Y na X, Y na Z, i założenia, że zmienne niemodelowane są niezależne.

Ze względu na interpretację znaku równości model strukturalny można rozumieć jako abstrakcyjny opis programu komputerowego. Mówiąc jeszcze inaczej, w obydwu przypadkach znak równości oznacza operację przypisywania wartości do zmiennej, tyle że albo ogólnie przez Naturę, albo przez tę szczególną część Natury, którą jest komputer. I tak, równanie strukturalne $Z = f_Z(X,Y,U_Z)$ oznacza, że wartości zmiennej Z powstają w sposób, o którym zakładamy, że nie zależy od żadnej zmiennej spoza zbioru swoich argumentów $\{X,Y,U_Z\}$, a od zmiennych z tego zbioru może zależeć, ale nie musi[12], co odpowiada „możliwościowej" albo „niewykluczającej" interpretacji strzałek. O zmiennych niemodelowanych w nieparametrycznych modelach strukturalnych zakładamy jedynie, że mają jakiś łączny, bliżej nieokreślony – poza konsekwencjami braku łuków – rozkład.

Własności ścieżek i grafów powstających na skutek interwencji analizujemy tak jak przed interwencją, trzeba tylko wcześniej usunąć wszystkie strzałki wchodzące do zmiennych poddawanych interwencji. W ten sposób w ramach jednego modelu możemy rozważać różne procesy czy wersje tego samego procesu, czego nie da się zrobić w rachunku prawdopodobieństwa. Definicja interwencji, jako operacji na grafie polegającej na usunięciu strzałek wchodzących do zmiennych poddawanych interwencji, odpowiada definicji strukturalnej, która wymaga, żeby prawą stronę każdego równania strukturalnego, określającego sposób powstawania wartości poddawanej interwencji zmiennej modelowanej, zastąpić wartością stałą. Definicja strukturalna z kolei odpowiada temu, jak symuluję tutaj skutki interwencji.

Interwencja rozumiana jest więc jako odcięcie od naturalnych albo wcześniejszych źródeł zmienności. Gdyby jakieś własności procesu mogły ulec zmianie w sposób inny niż z góry określony przez (niekoniecznie znany) model, skutek interwencji byłby niezdefiniowany. Dlatego te abstrakcyjne interwencje są z definicji lokalne w tym znaczeniu, że z założenia pozostałe funkcje strukturalne nie ulegają zmianie. Szacowanie efektów przyczynowych polega więc na szacowaniu skutków pewnego, być może nieosiągalnego, teoretycznego ideału, jakim są hipotetyczne idealnie selektywne interwencje.

[12] Definicja funkcji matematycznej dopuszcza, aby funkcja ignorowała swoje argumenty.

Przykład zastosowania wnioskowania przyczynowego do analizy i interpretacji wyników psychologicznego badania eksperymentalnego

Dotychczasowe ustalenia znajdują zastosowanie w przypadku dowolnych planów badawczych, postaci zależności statystycznych i poszukiwanych efektów przyczynowych na poziomie populacji osób albo populacji rozumianej jako hipotetyczne replikacje doświadczenia na tej samej osobie.

Zanim omówię typową sekwencję kroków analizy przyczynowej, muszę wprowadzić definicję ważnej relacji, którą można rozumieć jako formalizację pojęcia blokowania przepływu informacji. Mówimy, że zbiór zmiennych S *d-separuje* ścieżkę p między zmiennymi X i Y, jeżeli p zawiera zderzacz taki, że ani jego środkowa zmienna, ani żaden jej potomek, nie znajduje się w S – wtedy p nie jest aktywna, a warstwowanie po S nie umożliwia wystąpienia zależności pozornej ze względu na p – lub p zawiera łańcuch lub rozwidlenie takie, że środkowa zmienna należy do S – wtedy warstwowanie po tym elemencie S uniemożliwia przepływ informacji przez ścieżkę p, niezależnie od tego, czy p jest aktywna. Kryterium d-separacji wynika więc z omówionych wcześniej własności łańcucha, rozwidlenia i zderzacza. Jeżeli X i Y to dwa niepuste i rozłączne zbiory zmiennych, to mówimy, że zbiór S *d-separuje zbiory* X i Y, jeżeli d-separuje każdą ścieżkę między zmiennymi należącymi odpowiednio do zbiorów X i Y.

Typową sekwencję kroków w analizie przyczynowej prześledzimy na przykładzie interpretacji wyników badania na przeszukiwanie pamięci krótkoterminowej (Sternberg, 1969), w którym osoby badane rozpoznawały nowe lub stare bodźce docelowe po oglądaniu losowych zestawów szeregowo prezentowanych bodźców do zapamiętania. Najważniejszymi zmiennymi randomizowanymi były wielkość zestawu (W) i to, czy bodziec był nowy czy stary, a miarami były czas reakcji (RT) i poprawność rozpoznawania (ACC). Dla uproszczenia omówię tylko warunek, w którym bodźce testowe były nowe.

Sternberg założył w jednym z modeli, że nieobserwowalny proces generowania reakcji polega na szeregowym przeszukiwaniu pamięci, w ramach którego reprezentacja bodźca testowego jest porównywana w losowej kolejności bez powtórzeń z reprezentacjami bodźców zapisanymi w pamięci i każde porównanie zajmuje średnio tyle samo czasu (μ_T). Dla warunku z nowymi bodźcami testowymi te założenia możemy wyrazić modelem strukturalnym zawierającym równanie $RT = \sum_{i=1}^{W} T_i + U_{RT} = W\mu_T + \epsilon_T + U_{RT}$, gdzie T_i to czas i-tego porównywania, ϵ_T to suma odchyleń czasów porównań od średniej rozkładu czasów porównywania μ_T, a U_{RT} to łączny czas trwania pozostałych etapów procesu, takich jak kodowanie bodźca testowego i generowanie reakcji motorycznej. Zgodnie z tym modelem efekt W na średnią RT powinien być prostoliniowy i nachylenie powinno być równe średniemu czasowi porównywania, bo nowy bodziec testowy wymaga sprawdzenia wszystkich zapisanych w pamięci elementów. Zwracam jednocześnie uwagę, że ten model zakłada optymistyczny – bo nie ma na nim wielu możliwych strzałek i łuków – graf $W \rightarrow RT \leftarrow T$.

Pierwszym krokiem analizy przyczynowej może być *podział zmiennych modelowanych* na *obserwowane* (W, RT i ACC) i *nieobserwowane* (T).

Drugim krokiem będzie, dla wszystkich par zmiennych modelowanych, *narysowanie strzałek i łuków, których nie można wykluczyć lub uznać za pomijalne*. Żeby upewnić się, że sprawdziliśmy wszystkie pary, możemy uporządkować zmienne zgodnie z następstwem czasowym, np. W, T, RT, ACC, i wybierać kolejno pary (i,j) dla $i = 1, ..., n\text{-}1$ i $j = i + 1, ..., n$, gdzie n to liczba zmiennych, tj. (W,T), (W,RT) (W,ACC), (T,RT), (T,ACC), (RT,ACC). Ponieważ ten etap polega na formułowaniu argumentów teoretycznych, Czytelnik może się ze mną nie zgadzać i usunąć niektóre oznaczane przeze mnie krawędzie. Każdy taki krok będzie jednak wymagał uzasadnienia teoretycznego *ze strony Czytelnika*, bo rysując strzałki i łuki, twierdzę tylko ostrożnie, że czegoś nie wiadomo.

Konstruowany graf będzie dotyczył procesów rozgrywających się w trakcie pojedynczej próby, dlatego jego zastosowanie będzie być może wymagało analizy statystycznej danych nieuśrednionych. Gdyby analizy statystyczne były przeprowadzane na danych uśrednionych po próbach, konieczne byłoby rozważenie *relacji przyczynowych między próbami nie będących rozwidleniami*. Na przykład, poprawność w próbie t może wpływać na przebieg procesu reagowania w próbie $t + 1$, jeżeli osoba badana może się czasem zorientować po fakcie, że popełniła błąd, albo jeżeli w zadaniu pojawia się informacja zwrotna. Średni czas reakcji i średnia poprawność są deterministycznymi funkcjami czasów reakcji i poprawności w uśrednianych próbach, a więc możliwość tego rodzaju wpływu sprawia, że należy wtedy oznaczyć na grafie strzałkę $ACC \rightarrow RT$, której, jak się przekonamy, nie musimy oznaczać, jeżeli analiza jest przeprowadzona na danych nieuśrednionych.

Dogodnie jest zacząć od łuków. Ponieważ zmienna W jest randomizowana, żadna strzałka nie może wchodzić do W, ale już T może łączyć łuk z RT i ACC, bo trudno wykluczyć, że efektywność czy łatwość porównywania zależy od innych czynników, które mogą wpływać na poziom wykonania, takich jak np. chwilowe dystrakcje. Wreszcie, z oczywistych powodów, takich jak motywacja, strategia wykonania, uczenie się, zmęczenie, ale też z powodu wspomnianego przed chwilą możliwego wpływu poprawności w poprzedniej próbie, RT i ACC musimy również połączyć łukiem. W może wpływać bezpośrednio na każdą inną zmienną modelowaną choćby dlatego, że osoba badana może czasem zmieniać sposób wykonania zadania w zależności od spostrzeganej trudności zadania. Zmienna T może oczywiście wpływać na RT i, jeżeli np. występuje presja czasowa, również na ACC. Zauważmy, że RT i ACC to dwie własności tej samej reakcji, przy czym ACC jest również własnością bodźca (warunku nowy/stary), czego nie oznaczyłem na grafie, bo rozważam tylko próby z nowymi bodźcami docelowymi. W szczególności, zmienna RT *nie* jest tożsama np. z czasem przeszukiwania, a zmienna ACC *nie* jest tożsama z poziomem trudności czy łatwością przetwarzania, to są tylko obserwowane konsekwencje stanu tych i innych zmiennych, przy czym RT jest zmienną, której stan jest ustalany dopiero przez program komputerowy, dlatego RT i ACC nie mogą wpływać na siebie nawzajem. Z powodu następstwa w czasie te dwie zmienne nie mogą też wpływać na T. W ten sposób uzyskujemy graf przedstawiony na rysunku 1.

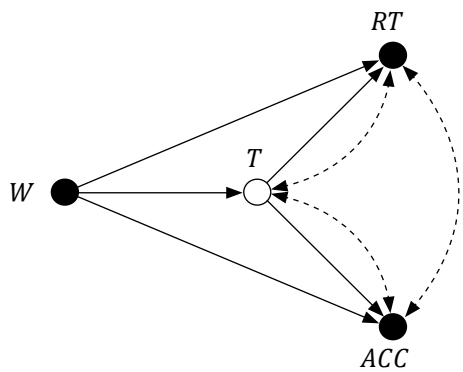
Być może Czytelnik się ze mną nie zgodzi, ale moim zdaniem status zmiennej T nie jest tutaj oczywisty. Na przykład, nie wiemy, czy przeszukiwanie pamięci polega na porównywaniu dyskretnych reprezentacji, czy może raczej na bardziej „rozmytym" procesie kumulacji świadectw albo na czymś jeszcze innym. Przyjmując

możliwościową interpretację również wobec wybranych zmiennych latentnych (tutaj tylko *T*), możemy jednak spokojnie założyć, że ten graf jest prawdziwy.

Trzecim krokiem może być *testowanie założeń przyczynowych*. Testowalne konsekwencje grafu przyczynowego to (być może warunkowe) niezależności między zmiennymi obserwowanymi wynikające z d-separacji. Należy pamiętać, że prawie każdy graf przyczynowy będzie miał „sobowtóry" w postaci grafów, które są od niego statystycznie nieodróżnialne. Mówimy o takich grafach, że są *statystycznie równoważne* albo że należą do tej samej *klasy równoważności*.

Rysunek 1

Graf reprezentujący teoretycznie możliwe relacje przyczynowe między wielkością zestawu W, latentnym czasem porównywania T, czasem reakcji RT i poprawnością ACC w warunku z nowymi bodźcami w zadaniu na przeszukiwanie pamięci krótkoterminowej



Nierozróżnialne statystycznie są m.in. wszystkie modele jakościowe, które mają te same zmienne modelowane, ten sam *szkielet* – zmienne połączone strzałką lub łukiem w jednym są też połączone strzałką lub łukiem w drugim, i vice versa – i w których występują te same struktury V-kształtne, czyli struktury z rodzicami nie połączonymi pojedynczymi strzałkami ani łukami (Verma i Pearl, 2022). Zmieniając kierunek dowolnej strzałki na grafie, tworzymy więc model nieodróżnialny statystycznie, czyli przez obserwację wszystkich zmiennych modelowanych (w tym również latentnych!), od modelu wyjściowego, o ile nie usuwamy istniejącej, ani nie tworzymy nowej struktury V-kształtnej.

Możemy teraz dokonać przekładu uwagi dotyczącej korelacji między zadowoleniem z życia i dochodami rocznymi na precyzyjny i ogólny język teorii wnioskowania przyczynowego. Zamiast *„z korelacji między X i Y nie wynika, że X wpływa na Y, ponieważ ta korelacja może występować z powodu istnienia wspólnych przyczyn lub z powodu wpływu Y na X"* możemy teraz powiedzieć ogólniej *„dla każdego rozkładu X i Y istnieje nieskończenie wiele procesów o grafie przyczynowym innym niż X → Y generujących ten rozkład"*. Aby wykazać, że wnioskowanie o wpływie w określonym kierunku na podstawie samej korelacji jest błędne, wystarczy oczywiście wykazać, że istnieje jeden taki proces.

Wracając do eksperymentu Sternberga, w tym wypadku żaden zbiór zmiennych obserwowanych nie d-separuje żadnej pary zmiennych obserwowanych, a więc graf nie jest w ogóle testowalny – każdy możliwy rozkład trzech zmiennych może być generowany przez pewien proces opisany przez ten graf. Ten graf nie musi być jednak testowalny, tylko ma wyrażać formalnie teoretycznie możliwe relacje przyczynowe, które trzeba brać pod uwagę, interpretując wyniki badania.

Oparty na spekulatywnych założeniach jakościowych i ilościowych szeregowy model Sternberga jest testowalny, bo istnieją możliwe wzorce wyników, które są z nim niezgodne. Każdy wzorzec wyników zgodny z tym modelem można jednak wyjaśnić, przyjmując zupełnie inne założenia na temat procesu przeszukiwania. Na przykład, jak za pomocą pracochłonnej analizy formalnej udowodnili Townsend i in. (1983), ze względu na wyniki tego rodzaju badania rozważany model szeregowy jest nieodróżnialny empirycznie m.in. od pewnych teoretycznie akceptowalnych modeli równoległych. I bez takich formalnych analiz można zresztą zauważyć, że – jak to w psychologii zwykle bywa – wzorzec wyników jest prosty i umiarkowanie zaskakujący, a pytanie ambitne.

Warto przy okazji omawiania testowania założeń przyczynowych zwrócić uwagę na mniej oczywiste konsekwencje randomizacji. W przypadku procesów stochastycznych nie da się zagwarantować braku błędu próby. Dlatego oczekiwanie, że faktycznie losowo wydzielone grupy będą pod każdym ważnym względem jednakowe, jest równie niemądre jak oczekiwanie, że losowa próba będzie miała wszystkie cechy populacji albo że dwie wielokrotnie niezależnie losowane wartości nigdy nie będą istotnie skorelowane. Na ogół *nietestowalne*[13] założenie o niezależności *mechanizmu decydującego o przynależności do warunków* nadal będzie wtedy spełnione. Ze względu na wnioski końcowe tylko to się liczy, ponieważ gwarancje związane z wnioskowaniem statystycznym dotyczą własności reguł decyzyjnych i mają charakter *asymptotyczny*. Statystyczne kontrolowanie wartości pomiarów dokonanych przed interwencją eksperymentalną może być uzasadnione z różnych powodów. Jeżeli jednak faktycznie zrandomizowane grupy będą się istotnie różniły pod względem zmiennych, których wartości były ustalone przez proces generujący dane jeszcze *przed* randomizacją, to będzie to tylko wynik błędu próby. Stosowanie *z tego powodu* poprawek statystycznych będzie przykładem stosowania *reguły decyzyjnej polegającej na wnioskowaniu z błędu próby*, czyli na wróżeniu z fusów. Kontrolowanie statystyczne zmiennych, których wartości były ustalone *po* interwencji, może systematycznie zniekształcić oszacowanie efektu całkowitego interwencji, dlatego taki zabieg będzie miał sens tylko w szczególnych sytuacjach, np. w kontekście analizy mediacji kiedy to celem nie jest szacowanie efektu całkowitego.

Czwartym i w tym wypadku ostatnim krokiem będzie *ustalenie warunków identyfikowalności poszukiwanych wielkości przyczynowych* i oszacowanie tego, co da się oszacować. Z grafu wynika natychmiast, że regresja każdej zmiennej

[13] Jeden z recenzentów zwrócił uwagę, że randomizacja może być nieskuteczna i może się pojawić konieczność sprawdzenia, czy była skuteczna. Wtedy jednak strukturę przyczynową zawodnego procesu randomizacji należałoby również oznaczyć na grafie.

obserwowanej różnej od W na W daje, z dokładnością do przybliżenia za pomocą modelu statystycznego, poprawne oszacowanie całkowitego wpływu W na tę zmienną. Jednocześnie widzimy, że „wyczyszczenie" danych przez usunięcie błędnych reakcji może generować zależność pozorną między W i RT. Dzieje się tak z dwóch powodów: ACC jest zderzaczem W i T, jak również zderzaczem W i każdej przyczyny niemodelowanej RT i ACC. Wynika stąd, że taki zabieg może na różne sposoby systematycznie zniekształcić wpływ W na RT. Chwila namysłu pozwala również zrozumieć, że nie można spokojnie założyć, że wszystkie reakcje poprawne to takie, które są generowane przez interesujący badacza proces, bo ludzie to nie roboty wykonujące proste zadania i ich reakcje mogą być, i na pewno w nieznanej części prób są, poprawne poniekąd przez pomyłkę. Co więcej, interesujący badacza proces może zachodzić w jakościowo taki sam sposób dla wszystkich lub dla niektórych reakcji poprawnych i dla wszystkich lub dla niektórych reakcji błędnych, ale może się systematycznie różnić pod względem własności ilościowych między reakcjami poprawnymi i błędnymi, przez co analizowanie tylko czasów reakcji poprawnych może prowadzić do systematycznie błędnych wniosków.

Jak widać, bez dobrej teorii niewiele można się dowiedzieć na temat ukrytego złożonego procesu reagowania. Dobrze uzasadnione wnioski z tego badania sprowadzają się w zasadzie do tego, że w pewnych granicach zwiększanie zestawu do zapamiętania wydłuża mniej więcej prostoliniowo średni czas reakcji i zwiększa prawdopodobieństwo błędu. Dowiemy się też na pewno, jeżeli zbadamy więcej niż raz więcej niż jedną osobę, że te efekty są inter i intraindywidualnie zróżnicowane. Na przykład, w stopniu różnym dla różnych osób będzie widać postępującą w miarę ćwiczenia poprawę poziomu wykonania.

Badacz, który chciałby twierdzić, że z tych danych można wywnioskować coś więcej, musiałby się zmierzyć z tym, że korzystając z analizy przyczynowej, można zdefiniować proste symulacje, ilustrujące alternatywne wyjaśnienia obserwowanych zależności statystycznych. Jedynym sposobem, żeby wyprowadzić mocniejsze niż to wynika z samego uzasadnionego grafu wnioski na temat badanych procesów, jest skorzystanie z teorii, z której wynikają dostatecznie silne ograniczenia *ilościowe*. Wtedy trzeba jednak mieć dobre powody, żeby twierdzić, że ta teoria jest *bliska prawdy jako opis procesu reagowania*, który jest przecież wielowymiarowy, złożony, nieobserwowalny, niestacjonarny i idiosynkratyczny.

Kategoryczność powyższych uwag wynika z dwóch ważnych twierdzeń, z których jedno zostało udowodnione dopiero niedawno. Po pierwsze, wiemy już od dłuższego czasu, że rachunek przyczynowy jest zupełny: dla dowolnego grafu przyczynowego i rozłączonych zbiorów zmiennych modelowanych X, Y i Z na tym grafie, gdzie zbiór Z może być pusty, rozkład interwencyjny $p(Y|do(X = x), Z)$ jest identyfikowalny wtedy i tylko wtedy, gdy jego identyfikowalność można ustalić, stosując trzy reguły rachunku przyczynowego (zob. Shpitser i Pearl, 2008, gdzie podano również ogólne warunki identyfikowalności dla wielkości kontrfaktycznych). Pewne efekty przyczynowe mogą być czasem co prawda identyfikowalne ze względu na dodatkowe założenia ilościowe, takie jak liniowość, brak interakcji, lub monotoniczność, ale w sytuacji braku wiedzy na temat ilościowych własności badanych procesów, a więc m.in. w przypadku typowych badań psychologicznych,

poleganie na tego rodzaju założeniach będzie często przejawem daleko idącego optymizmu; wnioski końcowe będą wtedy uzasadnione przede wszystkim ze względu na te optymistyczne założenia, a w mniejszym stopniu, o ile w ogóle, ze względu na wyniki badania. Po drugie, od niedawna wiemy też, że poziom kontrfaktyczny jest w pewnym technicznym sensie nieredukowalny do poziomu interwencyjnego (Bareinboim i in., 2022), z czego z kolei wynika, że na pewne pytania dotyczące ilościowych własności procesu generującego dane nie da się udzielić odpowiedzi za pomocą badań eksperymentalnych. To być może oznacza, że ilościowa część teorii musi mieć do pewnego stopnia uzasadnienie czysto teoretyczne, co zresztą zdarza się w psychologii (zob. np. teorie oparte na analizie racjonalnej, Chater i Oaksford, 1999).

Na badanie Sternberga można popatrzeć w jeszcze inny sposób, zakładając graf przedstawiony na rysunku 2. Ponieważ wybór zmiennych modelowanych nie ma znaczenia dla poprawności wniosków, dla uproszczenia pominąłem na tym grafie poprawność. Nie oznaczyłem też łuków, dlatego że każda ścieżka między zmiennymi obserwowanymi (tutaj tylko W i RT) przebiegająca przez ewentualne łuki zawierałaby zderzacz, a więc byłaby nieaktywna. Ze względu na dopuszczalne wyjaśnienia przyczynowe rozkładu $p(RT|W)$ moglibyśmy więc założyć, że ten graf jest prawdziwy, gdyby nie to, że dla uproszczenia optymistycznie wykluczyłem również strzałkę $TT \rightarrow P$, chociaż długi całkowity czas przeszukiwania pamięci mógłby być przyczyną utraty informacji w pamięci.

W eksperymentach psychologicznych celem często jest wykazanie, że jakaś jedna psychologiczna zmienna latentna (tutaj P) wpływa na jakąś inną zmienną latentną (tutaj TT), np. że nastrój wpływa na pamięć. Dlatego grafy przypominające ten na rysunku 2 będą się pojawiały w każdym takim eksperymencie.

Sternbeg założył eksplicite, że W może wpływać na latentną liczbę elementów w pamięci P, a P może wpływać na RT za pośrednictwem latentnego całkowitego czasu przeszukiwania TT. Pomijając konsekwencje znanych własności planu badawczego (randomizacja i następstwo czasowe), to są w istocie albo tylko dwa kategoryczne założenia, o istnieniu zmiennych latentych P i TT, albo, przyjmując możliwościową interpretację konstruktów teoretycznych, brak jakichkolwiek kategorycznych założeń przyczynowych, czyli jedynie wyraz *intencji* badacza. W szczególności, brakuje powodów, żeby przyjąć optymistyczny graf $W \rightarrow P \rightarrow TT \rightarrow RT$, a tym samym wykluczyć którąkolwiek ze ścieżek $W \rightarrow TT \rightarrow RT$, $W \rightarrow P \rightarrow RT$ i $W \rightarrow RT$.

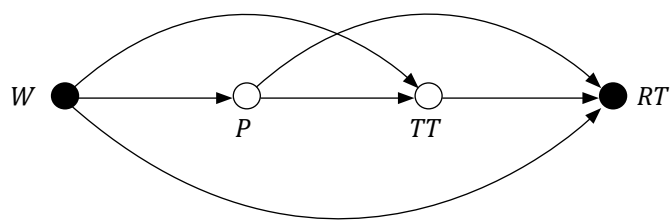
Randomizacja W gwarantuje jedynie, że efekt statystyczny W na dowolną inną zmienną wynika z działania *jakiejś* ścieżki kierunkowej, idącej od zmiennej W do tej zmiennej. Jak większość miar stosowanych w psychologii RT może być jednak pod *systematycznym* wpływem różnych czynników, o których zwykle niewiele wiadomo (Borsboom, 2005; Millsap, 2012; Paulewicz i Blaut, 2022; Van Bork i in., 2022). Ponadto o psychologicznych manipulacjach eksperymentalnych nie możemy zwykle spokojnie założyć, że nie mają niezamierzonych przez badacza skutków. Dlatego często nie da się wykluczyć niektórych lub wszystkich dodatkowych ścieżek oznaczonych na grafie przedstawionym na rysunku 2 nawet w relatywnie prostych eksperymentach, przeprowadzanych we względnie kontrolowanych warunkach.

W tym wypadku można np. zapytać, dlaczego osoby badane nie miałyby reagować wolniej częściowo dlatego, że widząc większy zestaw do zapamiętania (*W*), zniechęcają się czasem do wykonywania zadania, albo dlatego, że w czasie prezentacji kolejnych bodźców do zapamiętania rośnie prawdopodobieństwo lub stopień utraty koncentracji. Takie efekty mogłyby być mediowane przez czas przeszukiwania ($W \rightarrow TT \rightarrow RT$) albo przez czas kodowania, albo przez szybkość emitowania reakcji motorycznej będącej pod wpływem *P*, albo przez coś jeszcze innego ($W \rightarrow P \rightarrow RT$, $W \rightarrow RT$).

Czytelnikowi przyszło już zapewne do głowy, że może warto czasem próbować uzyskać dowody niezależności zmiennych, aby na tej podstawie usuwać problematyczne strzałki lub łuki. Wymaga to jednak przyjęcia założeń statystycznych, które nie będą łącznie prawdziwe i o których w dodatku nie będzie wiadomo, czy są łącznie tak bliskie prawdy, jak tego może wymagać w danym kontekście wykazywanie niezależności statystycznej na potrzeby wnioskowania przyczynowego. Założenie, że z niezależności statystycznej wynika brak aktywnych ścieżek jest akurat zwykle co najmniej prawdopodobne. W typowych sytuacjach wykazywanie niezależności zmiennych polega jednak na *wykazywaniu prawdziwości hipotezy punktowej*, np. że różnica między średnimi lub korelacja wynosi *dokładnie* 0*,* a nie na samym nieodrzuceniu takiej hipotezy. Użyteczną alternatywą jest zastosowanie metod wnioskowania przyczynowego, które pozwalają korzystać z założeń interwałowych, ale z braku miejsca nie będę pisał o tych zaawansowanych rozwiązaniach.

Rysunek 2

Uproszczony (zob. objaśnienie w tekście) graf reprezentujący teoretycznie możliwe relacje przyczynowe między wielkością zestawu W, latentnym obciążeniem pamięci P, latentnym całkowitym czasem przeszukiwania TT i czasem reakcji RT w zadaniu przeszukiwania pamięci krótkoterminowej



Ścieżki zakłócające i sposoby radzenia sobie z nimi

Możliwość wyprowadzenia uzasadnionych wniosków przyczynowych często znika, gdy zmienna, której skutki interesują badacza, nie jest poddana randomizacji, czyli wtedy, gdy ze względu na interesującą badacza przyczynę badanie, w którym poza tym mogą być jakieś inne zmienne randomizowane, ma charakter obserwacyjny. Chociaż następstwo czasowe będzie wtedy może pozwalało wykluczyć niektóre strzałki, to – zwłaszcza w psychologii – zwykle nie będzie się dało

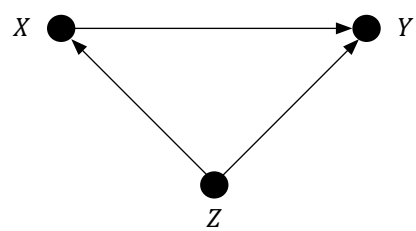
ani wykluczyć istnienia wspólnych przyczyn, ani dobrze uzasadnić założenia, że ich rola jest pomijalna.

Zmienne odgrywające rolę wspólnych przyczyn, będących jednocześnie alternatywnymi wyjaśnieniami albo współwystępującymi źródłami interpretowanej jako miara efektu przyczynowego zależności statystycznej, to tzw. zmienne *zakłócające* (ang. *confounding variables*). Tak naprawdę jednak to nie *zmienne* zakłócające są problematyczne, tylko występowanie w modelu aktywnych *ścieżek* z rozwidleniem, które dostarczają alternatywnych wyjaśnień obserwowanych zależności lub mogą zniekształcać sposób, w jaki w tych zależnościach ujawnia się poszukiwany efekt przyczynowy. W dodatku żaden z dwóch najważniejszych sposobów radzenia sobie z tego rodzaju ścieżkami, tj. poprawki tylnych drzwi i przednich drzwi, nie wymaga, żeby zmienna zakłócająca była obserwowana. Dlatego wygodniej jest posługiwać się pojęciem *ścieżki zakłócającej*.

Wyobraźmy sobie, że zakładamy poprawnie graf widoczny na rysunku 3, …

Rysunek 3

Graf z jedną ścieżką zakłócającą ($X \leftarrow Z \rightarrow Y$) i jedną zmienną zakłócającą (Z) ze względu na $p(Y|do(X))$



… a proces działa tak:

```
U_X = rnorm(n)
U_Y = rexp(n) - 1
U_Z = rbinom(n, size = 1, prob = 0.5)
Z = U_Z
X = U_X + 1 + 2 * Z
Y = U_Y + 3^Z + X^(Z + 1)
```

W języku R znak ^ oznacza operację podnoszenia do potęgi. Instrukcja `rbinom(n, size = 1, prob = 0.5)` generuje n pseudolosowych próbek o wartościach 0 lub 1 z rozkładu $p(0) = p(1) = 0{,}5$. Żeby zilustrować relatywne znaczenie założeń statystycznych i przyczynowych we wnioskowaniu przyczynowym zmienna U_Y ma tym razem przesunięty rozkład wykładniczy (`rexp(n) - 1`); odjęcie stałej 1 sprawia, że rozkład U_Y ma średnią 0.

Jeżeli interesuje nas efekt $p(Y|do(X))$, a nie *łączny* efekt $p(Y|do(X),do(Z))$, to musimy sobie poradzić ze ścieżką zakłócającą $X \leftarrow Z \rightarrow Y$. Gdybyśmy chcieli oszacować łączny wpływ X i Z na Y, wystarczyłoby dopasować model regresji Y na X i Z. Regresja

$p(Y|X,Z)$ szacuje tutaj $p\big(Y|do(X),do(Z)\big)$, ponieważ uwzględnienie jako predyktora Z blokuje jedyną problematyczną ścieżkę, nie generując przy tym zależności pozornej[14].

Przyjmijmy na początek, że chcemy oszacować średnią rozkładu Y w sytuacji $do(X = 0)$, ale *tylko w warstwie* $Z = 0$. W tym celu dopasujemy nieprawdziwą regresje liniową – rozkład $p(Y|X,Z)$ jest tutaj rozkładem (czasem przesuniętym) wykładniczym, a nie normalnym – do podzbioru próbek, dla których zachodzi $Z = 0$. W tej warstwie zakłócający wpływ zmiennej nie może się zamanifestować, bo Z przyjmuje tylko jedną wartość. Zależność statystyczna między X i Y w warstwie $Z = 0$ może więc wynikać tylko z wpływu X na Y i problem ścieżki zakłócającej znika. Inaczej mówiąc, z dokładnością do przybliżenia za pomocą modelu statystycznego, regresja Y na X w warstwie $Z = 0$ mówi wszystko o wpływie X na Y w tej warstwie, bo szacuje rozkład $p(Y|X,Z = 0)$, a $p(Y|X = x,Z = 0) = p\big(Y|do(X = x),Z = 0\big)$ dla każdego x. Dla uproszczenia skupię się jednak na interwencji $do(X = 0)$.

```
confint(lm(Y ~ X, subset = Z == 0))
#             2.5% 97.5%
# (Intercept) 0.96  1.04
# X           0.96  1.02
```

Z kodu symulacji wynika, że gdy obserwujemy $Z = 0$ i wymuszamy $do(X = 0)$, zmienna Y powstaje jako $3^Z + X^{Z+1} + U_Y = 3^0 + 0^{0+1} + U_Y = 1 + U_Y$, gdzie U_Y ma rozkład o średniej 0, czyli Y ma wtedy rozkład wykładniczy o średniej 1.

Ponieważ w każdej warstwie Z efekt statystyczny X na Y jest równy efektowi przyczynowemu X na Y, a wyraz wolny (ang. *intercept*) w dopasowanej regresji liniowej reprezentuje średnią rozkładu Y gdy $X = 0$, to oszacowanie wyrazu wolnego jest jednocześnie oszacowaniem poszukiwanej wielkości przyczynowej. Potwierdza to fakt, że prawdziwa średnia rozkładu interwencyjnego $p\big(Y|do(X = x),Z = 0\big)$, równa 1, mieści się w 95-procentowych przedziałach ufności dla wyrazu wolnego.

Tak samo możemy postąpić dla warstwy $Z = 1$. Ponieważ regresja liniowa pozwala, żeby predyktorami były jednocześnie różne funkcje zmiennych niezależnych, o ile zbiór wszystkich predyktorów nie jest współliniowy (w przypadku dwóch zmiennych współliniowość to tyle, co korelacja równa 1 lub -1), żeby dobrze opisać systematyczną część zależności statystycznej w warstwie $Z = 1$, wystarczy stworzyć dodatkową zmienną równą kwadratowi zmiennej X.

```
kwadratX = X^2
confint(lm(Y ~ kwadratX, subset = Z == 1))
#              2.5% 97.5%
# (Intercept)  2.91  3.02
# kwadratX     1.00  1.01
```

[14] Autorzy wymienionych wcześniej podręczników do metodologii badań i statystyki dla psychologów utrzymują, że korelacja między predyktorami jako taka, która tutaj oczywiście występuje, stanowi poważny problem, gdy stosuje się regresję liniową, albo wręcz że taka korelacja jest niezgodna z założeniami regresji liniowej (nie jest).

Jak łatwo sprawdzić, albo symulując skutki interwencji, albo obliczając dokładną wartość teoretyczną, uzyskane przedziałowe oszacowanie wyrazu wolnego zawiera prawdziwą średnią rozkładu interwencyjnego $p(Y|do(X=0),Z=1)$, równą 3. Oczekiwana wartość Y w sytuacji $do(X=0)$ jest więc czasem równa 1, a czasem 3, zależnie od tego, na którą warstwę Z akurat patrzymy. Jeżeli pomnożymy 1 i 3 przez prawdopodobieństwa, z jakimi te dwa możliwe skutki $do(X=0)$ występują, czyli odpowiednio przez $p(Z=0)$ i $p(Z=1)$, uzyskamy średnią rozkładu Y, gdy $do(X=0)$ zachodzi *w populacji*.

Uogólniając to rozumowanie na cały rozkład interwencyjny (a nie tylko średnią), arbitralne zmienne lub rozłączne zbiory zmiennych dyskretnych X i Y, i arbitralną interwencję na X, uzyskujemy *poprawkę tylnych drzwi*. Jeżeli ścieżek zakłócających będzie więcej niż jedna, konieczne będzie zablokowanie ich wszystkich. O każdym zbiorze zmiennych S takim, że żaden *potomek (elementu zbioru) X nie jest w S* i S d-separuje wszystkie ścieżki między X i Y takie, że zaczynają się od strzałki *wchodzącej do* (zmiennej w zbiorze) X, mówimy, że jest zbiorem *wystarczającym* (ze względu na zastosowanie poprawki tylnych drzwi do szacowania $p(Y|do(X))$). Istnienie takiego zbioru pozwala zastosować poprawkę tylnych drzwi analogicznie do sposobu, w jaki to właśnie zostało zrobione, tyle że rolę zmiennej Z będą wtedy spełniały wszystkie zmienne należące do S (sumowanie albo całkowanie i warstwowanie będzie przebiegało po nich wszystkich):

$$\begin{aligned} p(Y|do(X=x)) &= \sum_S p(Y|do(X=x),S=s)p(S=s) \\ &= \sum_S p(Y|X=x,S=s)p(S=s) \end{aligned}$$

W ostatnim wyrażeniu nie występuje operator do. Jak już wiemy, wynika to stąd, że ze względu na wpływ na Y w każdej warstwie S *obserwowanie* X jest statystycznie równoważne *interwencji* na X, a co z kolei wynika z d-separacji wszystkich ścieżek zakłócających przez elementy zbioru S. Ponieważ ostatnie wyrażenie zawiera same wielkości nieinterwencyjne, wartość tego wyrażenia można szacować na podstawie wyników badania *obserwacyjnego* – $p(Y|X,S)$ możemy szacować za pomocą regresji, a $p(S)$ możemy szacować, dopasowując rozkład parametryczny. Uzyskane wyrażenie jest *uniwersalnym estymatorem całkowitego efektu przyczynowego X na Y w każdej sytuacji, w której istnieje zbiór wystarczający, niezależnie od ilościowych własności relacji przyczynowych*. Trzeba tylko pamiętać, że gdyby któraś ze zmiennych w S była ciągła, zamiast sumy indeksowanej pojawiłaby się całka.

Ten ważny generyczny estymator jest nazywany poprawką tylnych drzwi, bo zmienne w S blokują „tylne wejścia" do zmiennej lub zmiennych, których całkowity wpływ chcemy oszacować. Kontrolowane w ten sposób statystycznie zmienne nie muszą być zmiennymi zakłócającymi, o czym warto wiedzieć, bo nie wszystkie zmienne zakłócające mogą być obserwowane albo obserwacja innej niż zakłócająca zmiennej blokującej może być mniej kosztowna. Co więcej, jeżeli jakaś blokująca zmienna obserwowana jest na grafie bliżej – w znaczeniu liczby strzałek na ścieżce – zmiennej Y niż zmienna zakłócająca, to użycie takiej bliższej Y zmiennej do blokowania zakłóceń zwiększy precyzję estymatora, jeżeli tylko procent wariancji „wyjaśnionej" w Y będzie dzięki temu większy.

Teoria wnioskowania częstościowego nie będzie szczególnie pomocna, gdy chcemy uzyskać oszacowanie *przedziałowe* tego rodzaju wielkości, bo w ogólnym przypadku teoretyczny rozkład z próby estymatora powstającego przez zastosowanie poprawki tylnych drzwi nie będzie znany. Czytelnik znający podstawy wnioskowania bayesowskiego będzie mógł jednak w wielu tego rodzaju sytuacjach skonstruować właściwy estymator, zastępując $p(S)$ i $p(Y|X = x,S)$ lub $p(Y|X,S)$ próbkami z odpowiednich rozkładów aposteriori. Trzeba wtedy uważać na jakość przybliżenia efektów statystycznych zmiennych w zbiorze wystarczającym S. W szczególności, blokowanie ścieżek zakłócających w badaniach psychologicznych będzie często utrudnione przez to, że węzły na tych ścieżkach będą reprezentowały zmienne latentne, a kontrolowanie statystyczne wyniku pomiaru zmiennej ma inne konsekwencje niż kontrolowanie statystyczne samej tej zmiennej.

Gdyby zastosowany model regresji nie uchwycił dobrze efektu S na Y, albo użyto *pomiaru* zmiennej blokującej zamiast samej tej zmiennej, resztowe zależności, w takich sytuacjach nazywane czasem resztowymi zakłóceniami (ang. *residual confounding*), mogłyby systematycznie zniekształcać oszacowanie efektu przyczynowego X na Y. Na przykład, jeżeli proces to rozwidlenie $X \leftarrow Z \rightarrow Y$, Z ma standardowy rozkład normalny, Z_{01} to zmienna Z zdychotomizowana według kryterium $Z > 0$, to $X \perp\!\!\!\perp Y|Z$, ale nieprawda, że w ogólnym przypadku $X \perp\!\!\!\perp Y|Z_{01}$. Jak łatwo się przekonać za pomocą symulacji, próba szacowania (tutaj zerowego) efektu $p\big(Y|do(X)\big)$ przez zastosowanie w poprawce tylnych drzwi regresji $p(Y|X,Z_{01})$ i oszacowania rozkładu $p(Z_{01})$ doprowadzi bo błędnego wniosku, że X wpływa na Y.

O poprawności uzyskanego wcześniej estymatora można się przekonać, symulując interwencję i obliczając średnią już nie z warstwy Z, tylko z całego zbioru wygenerowanych wartości Y, czyli z próbek z rozkładu $p\big(Y|do(X = 0)\big)$:

```
U_Y = rexp(n) - 1
U_Z = rbinom(n, size = 1, prob = 0.5)
Z = U_Z
X = 0
Y = U_Y + 3^Z + X^(Z + 1)
mean(Y)
# 2.00
```

Dla porównania naiwne, jeżeli celem jest oszacowanie wpływu X, dopasowanie regresji liniowej Y na X lub na X^2, lub na X i X^2, bez uwzględnienia roli Z, daje 95-procentowe przedziały ufności wokół wyrazu wolnego równe odpowiednio [-1,61; -1,39], [0,79; 0,90] i [0,42; 0,57]. W każdym przypadku prawdziwa wartość interwencyjna leży daleko poza przedziałami ufności (mierząc odległość szerokością interwałów).

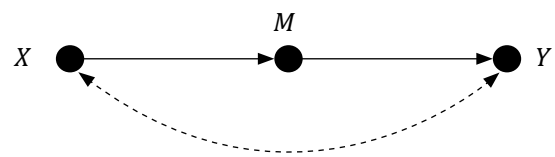
Poprawkę tylnych drzwi albo coś do niej podobnego można próbować stosować np. w badaniach dotyczących relatywnego wpływu genów (G) i cech rodziców lub innych własności środowiska rodzinnego (R) na cechy dziecka w wieku dorosłym (D). Z powodu następstwa czasowego możemy wykluczyć strzałkę $R \leftarrow D$, a więc można takie badanie przedstawić za pomocą z pewnością prawdziwego

grafu $R \rightarrow D + R \leftarrow X \rightarrow D$[15], gdzie X to wszystkie zmienne zakłócające, włączając w to geny. W ogólnym przypadku nie możemy bezpiecznie założyć, że $X = G$, ale możemy potencjalnie uzyskać wyniki wskazujące na to, że to założenie jest dobrym przybliżeniem. Jeżeli np. stwierdzimy na podstawie wyników badania na dużej próbie, że zależność statystyczna między R i D staje się znacznie słabsza i bliska zeru, kiedy poprawnie kontrolujemy efekt statystyczny G, to uzasadniony będzie wniosek, że związek statystyczny między R i D wynika co najmniej prawie całkowicie z wpływu G.

Jeżeli mamy powody wierzyć w prawdziwość grafu przedstawionego na rysunku 4, możemy skorzystać z faktu, że X blokuje wszystkie tylne wejścia ze względu na $p\big(Y|do(M)\big)$.

Rysunek 4

Warunki umożliwiające oszacowanie efektu przyczynowego X na Y za pomocą poprawki przednich drzwi



W każdej takiej sytuacji efekt $p\big(M|do(X)\big)$ można oszacować za pomocą regresji, a efekt $p\big(Y|do(M)\big)$ za pomocą poprawki tylnych drzwi ze względu na zbiór wystarczający $\{X\}$. Żeby oszacować na tej podstawie efekt przyczynowy X na Y stosujemy *poprawkę przednich drzwi*. Niestety, nie jest łatwo wskazać badania podstawowe w psychologii, w których można by tę poprawkę zastosować. Mediator będzie przecież zwykle zmienną latentną, mediacja będzie prawdopodobnie częściowa, model pomiarowy mediatora będzie spekulatywny i bardzo uproszczony i nie będzie można zwykle wykluczyć, że efekty $p\big(M|do(X)\big)$ lub $p\big(Y|do(M)\big)$ są zakłócone (Rohrer i in., 2022). Dlatego zainteresowanego tą poprawką Czytelnika odsyłam do *Przyczyn i skutków* (Pearl i Mackenzie, 2021) albo nawet do *Primera* (Pearl i in., 2016), bo do tej lektury powinien być już przygotowany.

Uwagi końcowe

Teoria wnioskowania statystycznego pozwala znacznie zminimalizować ryzyko, że oparte na danych wnioski *na temat rozkładów* będą błędne. Teoria wnioskowania przyczynowego odgrywa analogiczną rolę na etapie *analizy teoretycznej*, pozwalając część tej analizy sformalizować, a dzięki temu m.in. ułatwia identyfikację

[15] Operacja dodawania grafów, którą wprowadziłem dla wygody, polega na utożsamieniu wierzchołków o tych samych nazwach.

wszystkich możliwych typów alternatywnych wyjaśnień. Niestety, psychologowie nadal próbują odpowiadać na pytania przyczynowe, polegając w dużym stopniu, o ile nie całkowicie, na porównaniach dopasowania modeli statystycznych. Mnie też zdarzało się popełniać ten fundamentalny błąd, czasem więcej niż raz w tej samej publikacji (np. Paulewicz i in., 2007). To jest w istocie ten sam błąd, polegający na niewłaściwym uwzględnieniu statusu i roli założeń przyczynowych i statystycznych, od którego omówienia zacząłem to wprowadzenie, tylko że poza znajomym kontekstem korelacji dwóch zmiennych nie jest tak łatwy do wykrycia.

Im mniej jest strzałek w modelu przyczynowym, tym zwykle prostszy będzie odpowiadający mu model statystyczny, bo będzie miał mniej wolnych parametrów, potencjalnie tym bardziej będzie testowalny, i tym więcej będzie wynikało z modelu przyczynowego, a więc tym *ciekawsze* będą zwykle wynikające z zastosowania takiego modelu wnioski przyczynowe. Można to było zauważyć m.in. na przykładzie modelu szeregowego Sternberga. Wydawałoby się więc, że choćby z tych powodów, zgodnie z zasadą brzytwy Ockhama, w sytuacjach budzących wątpliwości czasem lepiej jest usuwać strzałki, łuki lub zmienne latentne, zamiast je pozostawiać. Prostota i testowalność empiryczna, wraz ze związaną z tą pierwszą uogólnialnością modelu *statystycznego*, są przecież czymś, czego oczekujemy zwykle od teorii czy hipotez empirycznych.

Takie podejście służy jednak poprawie *predykcyjnych*, a nie *eksplanacyjnych* właściwości modelu. Jak może się przekonać Czytelnik, który powinien być teraz w stanie porównać np. za pomocą testu ilorazu wiarygodności regresje $p(Y|X)$ i $p(Y|X,Z)$ dopasowane do wyników symulacji procesów o statystycznie nieodróżnialnych grafach $X \rightarrow Y + X \leftarrow Z \rightarrow Y$ i $X \rightarrow Y + X \rightarrow Z \leftarrow Y$, to czy jeden model statystyczny pasuje lepiej niż drugi, jest własnością *logicznie niezależną* od tego, który z nich dostarcza interpretowalne oszacowania. Między innymi dlatego np. dodawanie predyktorów, bo wydają się „jakoś" związane z interesującymi badacza zmiennymi, i wybieranie modelu regresji na podstawie testów statystycznych w nadziei, że w ten sposób uzyska się bardziej interpretowalne oszacowania, jest nieporozumieniem, które prędzej czy później, ale nieuchronnie, prowadzi do całkowicie błędnych wniosków (Cinelli i in., 2021).

Brak dobrej teorii można wyraźnie odczuć w psychometrii (Borsboom, 2005), która przecież zajmuje się tym, w jaki sposób możemy i powinniśmy „docierać" do psychologicznych zmiennych latentnych. O psychologicznych zmiennych latentnych z perspektywy wnioskowania przyczynowego będę mógł może coś więcej napisać w planowanej następnej części tego wprowadzenia, ale już teraz zwracam uwagę na pewną ważną okoliczność. Nie ma mianowicie sensu uzasadnianie wniosku, że pewien zestaw pozycji testowych jest pod systematycznym wpływem jednej tylko zmiennej latentnej, którą wyobraża sobie część społeczności badaczy, tym, że dobrze lub źle pasuje jakiś model czynnikowy. Sama jakościowa struktura przyczynowa modelu jednoczynnikowego *nie* jest testowalna, bo czynnik w tym uogólnionym rozwidleniu jest z definicji nieobserwowalny: *każdy* rozkład odpowiedzi na pozycje testowe może być generowany przez *pewien* proces o takiej strukturze, nawet jeżeli nie każdy taki rozkład wygląda tak, jakby generował go proces *liniowy* o takiej strukturze. Testowalne w tych modelach są wyłącznie, w psychologii przyjęte tylko dla uproszczenia, nawet nie spekulatywne, tylko

zwyczajnie odległe od prawdy, założenia o liniowości efektów, a także o niezależności i normalności błędów, a dokładniej *p r e d y k c y j n a użyteczność dalekich od prawdy, ułatwiających analizę założeń ilościowych, przy założeniu nietestowalnej i nierzadko w najlepszym razie wątpliwej struktury przyczynowej*. Ta pesymistyczna obserwacja wynika już z części teorii, którą tu przedstawiłem, i tego samego można się dowiedzieć również ze współczesnych opracowań na temat modeli SEM, w których przyczynowość jest traktowana poważnie (np. Hoyle, 2012; Kline, 2015). O tym, że nie bardzo wiadomo nawet, gdzie należy *szukać* zadowalającego rozwiązania problemu systematycznych źródeł błędu pomiaru psychologicznych zmiennych latentnych, można się dowiedzieć m.in. z niepozostawiającej złudzeń monografii Millsapa dotyczącej niezmienności pomiaru (2012), a także z mojego i Blaut skromnego wkładu w tą literaturę (Paulewicz i Blaut, 2022).

Wiemy, że rachunek przyczynowy jest zupełny i nie są znane kontrprzykłady pokazujące wadliwość jego aksjomatów. W dodatku teoria wnioskowania przyczynowego jest już na tyle rozwinięta, że o niektórych ważnych klasach problemów wiemy nie tylko, że ta teoria dostarcza *jakieś* rozwiązania tego rodzaju problemów, ale również że dostarcza *wszystkie możliwe* rozwiązania. Istnieją w wersji użytecznej i rozwiniętej, albo na naszych oczach powstają, części teorii dotyczące mediacji (Pearl, 2012), brakujących obserwacji (zastępując przestarzałe metody klasyczne, spełnienie wymagań których jest tak trudne do ustalenia w praktyce, że rzadko wiadomo, kiedy można je stosować, zob. Mohan i in., 2013), integracji wyników podobnych lub tylko powiązanych ze sobą badań obserwacyjnych lub eksperymentalnych (zastępując i poszerzając obszar zastosowań przyczynowo ślepych metod metaanalizy, zob. Bareinboim i Pearl, 2016), sposobów radzenia sobie ze stronniczością próby (Bareinboim i in., 2022) i innych zagadnień o kluczowym znaczeniu dla badań podstawowych lub aplikacyjnych. Podstawy metodologii badań psychologicznych, włączając w to teorię planowania i analizy wyników badań i teorię tworzenia narzędzi pomiarowych, oceny ich właściwości psychometrycznych i interpretacji wyników pomiaru, były do niedawna oparte tylko na jednej teorii matematycznej, tj. na rachunku prawdopodobieństwa, i na opartej na tym rachunku teorii wnioskowania statystycznego. Najważniejsze problemy metodologiczne są jednak w pierwszej kolejności problemami przyczynowymi, a dopiero w drugiej statystycznymi. Jak pokazują przykłady, które omówiłem, poleganie na intuicji w sytuacjach, do których stosują się twierdzenia rachunku przyczynowego, jest równie rozsądne jak lekceważenie twierdzeń rachunku prawdopodobieństwa czy zasad logiki. Wydaje się zatem, że psychologowie mają obecnie do wyboru albo przyswoić sobie tę teorię, albo wyprowadzać dalej z badań wnioski, o których teraz można często dedukcyjnie stwierdzić coś, z czego – jak sądzę – wielu i bez tego zdaje sobie nie najgorzej sprawę.

Bibliografia

Bareinboim, E., Correa, J. D., Ibeling, D., Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. W: R. Dechter, J. Halpern i H. Geffner (red.),

Probabilistic and causal inference: The works of Judea Pearl (s. 507–556). ACM Books.

Bareinboim, E., Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345–7352. https://doi.org/10.1073/pnas.1510507113

Bareinboim, E., Tian, J., Pearl, J. (2022). Recovering from selection bias in causal and statistical inference. W: R. Dechter, J. Halpern i H. Geffner (red.), *Probabilistic and causal inference: The works of Judea Pearl* (s. 433–450). ACM Books.

Bedyńska, S., Książek, M., Cypriańska, M. (2012). *Statystyczny drogowskaz*. Wydawnictwo Akademickie Sedno.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2*(3), 47–53. https://doi.org/10.2307/3002000

Blalock, H. M. (2018). *Causal inferences in nonexperimental research*. UNC Press Books.

Bollen, K. A. (1989). *Structural equations with latent variables* (T. 210). John Wiley & Sons.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Brzeziński, J. (2022). *Metodologia badań psychologicznych*. Wydawnictwo Naukowe PWN.

Chater, N., Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65. https://doi.org/10.1016/s1364-6613(98)01273-x

Cinelli, C., Forney, A., Pearl, J. (2021). A crash course in good and bad controls. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221099552

Duncan, O. D. (2014). *Introduction to structural equation models*. Elsevier.

Field, A. P., Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Galles, D., Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *3*(1), 151–182. https://doi.org/10.1023/A:1009602825894

Greenland, S. (2022). The causal foundations of applied probability and statistics. W: R. Dechter, J. Halpern i H. Geffner (red.), *Probabilistic and causal inference: The works of Judea Pearl* (s. 605–624). ACM Books.

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford Press.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Levy, J., Pashler, H., Boer, E. (2006). Central interference in driving: Is there any stopping the psychological refractory period? *Psychological Science*, *17*(3), 228–235. https://doi.org/10.1111/j.1467-9280.2006.01690.x

Liddell, T. M., Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Mohan, K., Pearl, J., Tian, J. (2013). Graphical models for inference with missing data. W: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani i K. Q. Weinberger (red.), *Advances in Neural Information Processing System*, 26 (NIPS-2013) (s. 1277-1285). Curran Associates, Inc.

Paulewicz, B., Blaut, A. (2022). The general causal cumulative model of ordinal response. *PsyArXiv*. https://doi.org/10.31234/osf.io/e7a3x

Paulewicz, B., Chuderski, A., Nęcka, E. (2007). Insight problem solving, fluid intelligence, and executive control: A structural equation modeling approach. W: S. Vosniadou, D. Kayser i A. Protopapas (red.), *Proceedings of the European Cognitive Science Conference 2007* (s. 586–591). Psychology Press.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Pearl, J. (2012). The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science*, *13*(4), 426–436. https://doi.org/10.1007/s11121-011-0270-1

Pearl, J., Glymour, M., Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., Mackenzie, D. (2021). *Przyczyny i skutki. Rewolucyjna nauka wnioskowania przyczynowego*. Copernicus Center Press.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna. https://www.R-project.org/

Rohrer, J. M., Hünermund, P., Arslan, R. C., Elson, M. (2022). That's a lot to PROCESS! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095827

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *10* (469), 322–331. https://doi.org/10.1198/016214504000001880

Saville, D. J., Wood, G. R. (2012). *Statistical methods: The geometric approach*. Springer Science & Business Media.

Shpitser, I., Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, *9*, 1941–1979. https://doi.org/10.5555/1390681.1442797

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457. https://www.jstor.org/stable/27828738

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica*, *106*(1), 147–246. https://doi.org/10.1016/s0001-6918(00)00045-7

Townsend, J. T., Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge University Press.

Van Bork, R., Rhemtulla, M., Sijtsma, K., Borsboom, D. (2022). A causal theory of error scores. *Psychological Methods*. https://doi.org/10.1037/met0000521

Verma, T. S., Pearl, J. (2022). Equivalence and synthesis of causal models. W: R. Dechter, J. Halpern i H. Geffner (red.), *Probabilistic and causal inference: The works of Judea Pearl* (s. 221–236). ACM Books.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.