helion.pl
KSIĘGARNIA INTERNETOWA · WYDAWNICTWO

IDŹ DO
PRZYKŁADOWY ROZDZIAŁ
SPIS TREŚCI

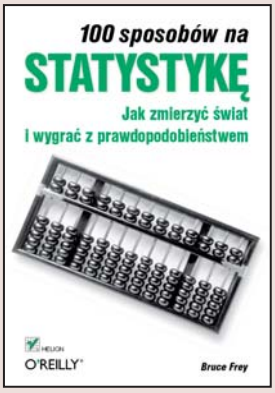
KATALOG KSIĄŻEK
KATALOG ONLINE
ZAMÓW DRUKOWANY KATALOG

TWÓJ KOSZYK
DODAJ DO KOSZYKA

CENNIK I INFORMACJE
ZAMÓW INFORMACJE O NOWOŚCIACH
ZAMÓW CENNIK

CZYTELNIA
FRAGMENTY KSIĄŻEK ONLINE

75 sposobów na statystykę. Jak zmierzyć świat i wygrać z prawdopodobieństwem



Autor: Bruce Frey
Tłumaczenie: Dariusz Biskup, Tomasz Misiorek
ISBN: 978-83-246-0708-2
Tytuł oryginału: Statistics Hacks: Tips & Tools for Measuring the World and Beating the Odds
Format: B5, stron: 336

**Zbiór metod i sztuczek statystycznych,
które pozwolą Ci zrozumieć naturę zdarzeń losowych**

- Jak trafnie prognozować przyszłe wydarzenia?
- W jaki sposób wykryć relacje między pozornie niepowiązanymi zjawiskami?
- Jak zarobić pieniądze dzięki rachunkowi prawdopodobieństwa?

Zazwyczaj szczęście sprzyja lepszym, choć czasem to „głupi ma szczęście". Jednak prawie zawsze szczęściu można pomóc, ponieważ wiele na pozór całkowicie przypadkowych zdarzeń rządzi się specyficznymi prawami wynikającymi z rachunku prawdopodobieństwa. Stosując odpowiednie metody statystyczne, można wykryć prawidłowości i relacje w wielu grach, odkryć sfałszowane dane, złamać szyfry czy odróżnić naprawdę losowe zjawiska od tych niezupełnie przypadkowych.

Książka „75 sposobów na statystykę. Jak zmierzyć świat i jak wygrać z prawdopodobieństwem" to zbiór technik i sztuczek statystycznych, które pozwolą Ci lepiej zrozumieć zjawiska zachodzące w świecie. Poznasz podstawowe metody statystyczne oraz nauczysz się stosować je do wykrywania niezauważalnych na pierwszy rzut oka relacji i trafnego szacowania prawdopodobieństwa różnych zdarzeń. Dowiesz się, jak podejmować najbardziej optymalne decyzje w grach hazardowych i jak zwiększyć prawdopodobieństwo wygranej w rozmaitych zabawach, takich jak Monopol czy rozgrywki sportowe. Przeczytasz także o gimnastyce umysłu, która ułatwia szybkie odkrywanie tajemnic losowych wydarzeń.

- Podstawowe metody statystyczne
- Wykrywanie i analizowanie relacji między zjawiskami
- Trafne szacowanie prawdopodobieństwa wydarzeń
- Optymalne podejmowanie decyzji w grach hazardowych
- Techniki pokazujące, jak grać, aby wygrać
- Ćwiczenia w zakresie poprawnego myślenia

**Chcesz, aby Twoim życiem przestał kierować przypadek?
Naucz się kontrolować swój los!**

Wydawnictwo Helion
ul. Kościuszki 1c
44-100 Gliwice
tel. 032 230 98 63
e-mail: helion@helion.pl

Spis treści

Mierzenie świata

Sposoby 23. – 34.

Zrozumienie zjawisk przez nadanie im wartości liczbowej jest bardzo cenne. Choć czasami podczas przekładu idei na liczbę tracimy coś ważnego, tworzenie wyników mających reprezentować interesujące nas zagadnienie pozwala je lepiej zrozumieć i dokonać niezbędne porównania. Wszystkie sposoby w tym rozdziale dotyczą pomiaru i interpretacji wyników.

Cała rodzina sposobów oparta jest na rozkładzie normalnym **[Sposób 23.]** i jego obecności wszędzie, gdzie byśmy nie spojrzeli. Dzięki krzywej rozkładu normalnego możemy stwierdzić, gdzie znajdujemy się w porównaniu z wszystkimi innymi **[Sposób 24.]**, możemy dowiedzieć się, jaki uzyskamy wynik testu, jeszcze zanim do niego zasiądziemy **[Sposób 25.]**, a także zinterpretować rezultaty naszych testów **[Sposoby 26. i 27.]**.

Skoro mowa o testach, nauczymy się układać dobre zbiory pytań **[Sposób 28.]** i przygotowywać wysokiej jakości testy **[Sposoby 31. i 32.]**. Jesteśmy w stanie rozpoznać złe elementy, bezwartościowe pytania i rozwiązać test z powodzeniem, nie znając odpowiedzi **[Sposób 29.]**. Możemy też poprawić wynik uzyskany w teście bez sięgania do książek **[Sposób 30.]**.

Wreszcie, poznając kilka solidnych podstaw pomiarów, możemy prognozować długość trwania epoki, osoby lub biznesu **[Sposób 33.]**, jak również dowiedzieć się, jak wykorzystywać informacje medyczne **[Sposób 34.]** do przedłużenia (być może) swojego życia.

Ziarnko do ziarnka, mamy oto rozdział pełen sposobów związanych z wszelkiej maści pomiarami.

SPOSÓB 23.

Spróbujmy zrozumieć świat

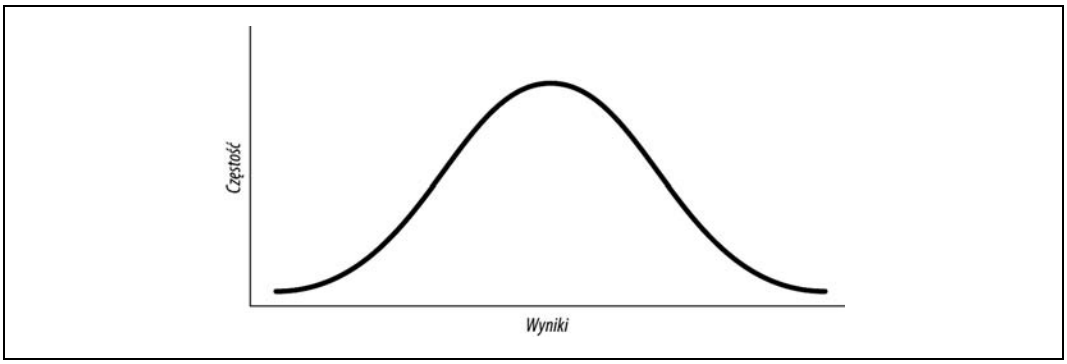
Niemal wszystko w otaczającym nas świecie rozkłada się w jakiś sposób. Wszystko, cokolwiek jesteśmy w stanie zmierzyć — a co osiąga różne wyniki — będzie miało dobrze nam znany „rozkład normalny". Jeśli dokładnie znamy kształt krzywej normalnej, możemy bardzo trafnie prognozować zachowania.

W świecie statystyki jest miejsce dla kilku cudów. Istnieją przynajmniej trzy narzędzia (trzy odkrycia), które są tak magiczne i wspaniałe, że gdy studenci statystyki dowiadują się o nich i zaczynają pojmować ich znaczenie, zdarza się nie raz i nie dwa, że eksplodują.

No dobrze, może nieco przesadzam, ale mamy trzy śliczne narzędzia, dzięki którym możemy lepiej zrozumieć świat. Oto one:

- współczynnik korelacji **[Sposób 11.]**,
- centralne twierdzenie graniczne **[Sposób 2.]**,
- krzywa rozkładu normalnego.

Ponieważ pierwsze dwa cuda omówiliśmy przy okazji prezentowania innych sposobów, poświęćmy trochę czasu na poznanie trzeciego: **krzywej dzwonowej**. Z największą przyjemnością przedstawiam krzywą dzwonową, rozkład normalny, krzywą normalną, taką jak na rysunku 3.1.



Rysunek 3.1. Krzywa normalna

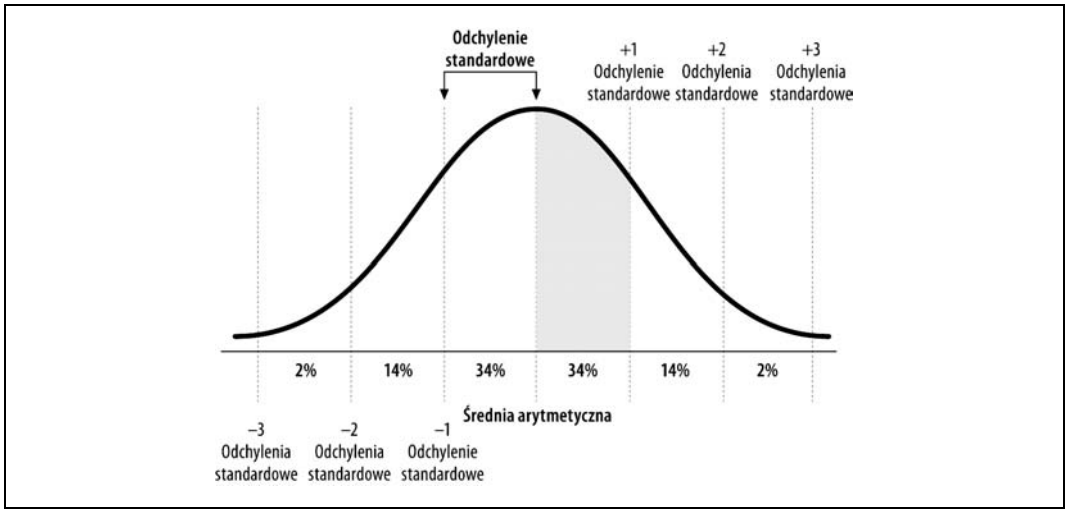
Stosowanie obszarów w krzywej normalnej

Statystycy bardzo starannie zdefiniowali krzywą normalną. Za pomocą obliczeń, jak również danych zbieranych przez setki lat, osiągnięto zgodne dla obu metod konkluzje co do dokładnego kształtu rozkładu normalnego. Na rysunku 3.2 widać istotne cechy krzywej dzwonowej. Średnia arytmetyczne jest w środku, a w miarę oddalania się od środka coraz mniej zostaje miejsca na wyniki.

Choć szerokość krzywej normalnej jest teoretycznie nieskończona, trzy odchylenia standardowe po każdej stronie średniej arytmetycznej zwykle mieszczą w sobie wszystkie wyniki.

Odchylenie standardowe rozkładu to przeciętna odległość każdego wyniku od średniej arytmetycznej **[Sposób 2.]**.

Prognozowanie wyników testów. Przywołajmy tu stwierdzenie, które padło wcześniej: wszystko, co mierzymy, rozkłada się w kształcie krzywej normalnej. Implikacja tego jest taka, że wszystko, co mierzymy, będzie miało większość wyników blisko średniej arytmetycznej i tylko kilka wyników będzie od niej oddalonych. Jeśli poddamy pomiarowi wystarczającą liczbę ludzi, otrzymamy też jakiś wynik skrajny, bardzo oddalony od średniej, ale wyniki oddalone od średniej będą rzadkie. Proporcja osób osiągających konkretny wynik zmniejsza się w miarę oddalania tego wyniku od średniej.



Rysunek 3.2. Obszary krzywej normalnej

Jak to będzie z następnym testem, jaki będziemy pisać? Nic nie wiem na temat moich czytelników, ale jestem gotów się założyć, że uzyskają wynik bliski średniej. Prognozuję wynik przeciętny. Może to być wynik powyżej lub poniżej przeciętnej, ale krzywa dzwonowa mówi mi, że będzie zbliżony do średniej arytmetycznej.

Żeby dokonywać takich prognoz, będąc całkiem przekonanym o ich trafności, możemy wykorzystać znane nam wymiary krzywej rozkładu normalnego do określenia odsetka wyników, które znajdą się pomiędzy dowolnymi dwoma punktami na osi X (poziomej linii na dole wykresu). Odsetek wyników pomiędzy parami punktów odchyleń standardowych na skali został pokazany na rysunku 3.2. Suma odsetek daje 100 procent (dzięki zaokrągleniom). Nie wolno zapominać, że niektóre wyniki, bardzo nieliczne, będą znajdowały się dalej od średniej niż trzy odchylenia standardowe.

Oto kilka kluczowych faktów na temat krzywej, które możemy wykorzystać do prognozowania rezultatów:

- Około 34 procent wyników mieści się pomiędzy średnią a jednym standardowym odchyleniem powyżej średniej. Spójrzmy na zakreślony na szaro fragment na rysunku 3.2. Jeśli wzięlibyśmy atrament i zaczernilibyśmy obszar mieszczący się pod krzywą normalną, 34 procent tego atramentu zużylibyśmy na ten fragment.
- Około 34 procent wyników mieści się pomiędzy średnią a jednym standardowym odchyleniem poniżej średniej.
- Około 14 procent wyników mieści się pomiędzy średnią a jednym do dwóch standardowych odchyleń powyżej średniej.
- Około 2 procent wyników mieści się pomiędzy średnią a dwoma do trzech standardowymi odchyleniami poniżej średniej.

Możemy też łączyć procenty, by stwierdzać inne fakty, takie jak:

- Około 68 procent wszystkich wyników znajdzie się w granicach jednego odchylenia standardowego od średniej.
- Około 50 procent wyników znajdzie się poniżej średniej.

Możesz wykorzystać te odsetki do prognozowania i stwierdzania prawdopodobieństwa. Możemy mówić o krzywej normalnej albo jako o **odsetku wyników**, które znajdują się w granicach danych obszarów krzywej, albo jako o **prawdopodobieństwie, że dana osoba zdająca test** znajdzie się w granicach danych obszarów:

- jest dwuprocentowa szansa, że w następnym teście dana osoba uzyska wynik wyższy od średniej arytmetycznej o więcej niż dwa odchylenia standardowe;
- jest tylko szesnastoprocentowa szansa, że kandydat uzyska wynik niższy od jednego odchylenia standardowego poniżej średniej arytmetycznej w naszym teście badającym umiejętności zawodowe.

Ustanawianie standardów. Autorzy polityki firmowej zakładają podczas definiowania oczekiwań wobec pracowników, że rozkład umiejętności jest normalny. Tak dobierają poziomy oczekiwań, aby zagwarantować sobie określony odsetek odpowiednich ludzi. Rozkład normalny jest nieocenionym narzędziem przy ustalaniu polityki naboru bądź oczekiwanej jakości usług, jeśli chcemy w magiczny sposób z góry wiedzieć, ile osób się zakwalifikuje.

Na przykład uczelnia pilnująca wysokich standardów kształcenia może wymagać od kandydatów, by ich średnia ocen, na podstawie której decyduje się o przyjęciu, była przynajmniej o jedno odchylenie standardowe wyższa od przeciętnej. W ten sposób zapewniają sobie, że przyjmowane będą wyłącznie osoby mieszczące się w 16 procentach najlepszych.

Podobnie, polityka edukacyjna w Stanach Zjednoczonych określa wyniki, jakie uczniowie muszą uzyskać w testach, by otrzymać specjalny status edukacyjny (i tym samym zakwalifikować się do stypendiów federalnych i stanowych). Wyniki kwalifikujące to konkretne wyniki, które dana osoba musi przekroczyć (lub znaleźć się poniżej). Jeśli autorzy polityki dysponują w budżecie pieniędzmi na dodatkowe świadczenia edukacyjne dla, powiedzmy, tylko dwóch procent wszystkich dzieci, ustawiają wyniki kwalifikujące na poziomie dwóch odchyleń standardowych od średniej. Zaufanie do krzywej rozkładu normalnego pozwala im obliczyć, ile dzieci w takim przypadku skorzysta ze stypendiów.

Doceńmy piękno krzywej normalnej

Aby docenić cud rozkładu normalnego, zawsze można stworzyć swój własny. Wyobraźmy sobie, że coś zmierzyliśmy (na przykład nastawienie, wiedzę, wzrost lub szybkość). Mamy jakiś system punktacji, w którym wyniki mogą być rozmaite (tak jak wyniki ankiety badającej nastawienie, wyniki egzaminów albo centymetry czy kilometry na godzinę). Mamy mnóstwo wyników, bo zmierzyliśmy mnóstwo ludzi, budynków lub wróbli. Następnie nanieśmy te wyniki na wykres, tak aby oś X odpowiadała wartości wyników, od najniższej

do najwyższej, od lewej do prawej (lub w innym kierunku, wedle uznania). Oś Y (pionowa linia po lewej stronie) powinna odpowiadać relatywnej częstości występowania każdej wartości w naszej grupie wyników.

Na takim wykresie wysokość słupka lub miejsce, w którym znajduje się punkt, odpowiada relatywnej proporcji wyników o określonej wartości. Zauważmy, że w przypadku krzywej normalnej najwyżej położone punkty znajdują się w jej środku, a te umieszczone najniżej — na krańcach. Wynik środkowy jest wynikiem przeciętnym, a także najczęściej występującym. Na krzywej normalnej mediana jest równa średniej arytmetycznej, która jest równa modalnej **[Sposób 21.]**.

Zauważmy też, że krzywa rozkładu normalnego jest symetryczna — możemy ją zgiąć na pół i jedna strona idealnie nałoży się na drugą. Inną cechą charakterystyczną krzywej normalnej, o której trzeba wiedzieć, jest to, że ciągnie się ona w nieskończoność. Jest to teoretyczna krzywa, więc dwa końce krzywej nigdy nie dotkną linii podstawowej.

Krzywa normalna dotyczy wszystkiego, łączy ze sobą całą naturę. Jest idealnie zrównoważona. Jest nieskończona. Jest wieczna. I wyglądem przypomina trochę dinozaura, co jest fajne.

SPOSÓB 24.

Tworzenie rang centylowych

Istnieje prosty, lecz potężny sposób interpretowania wyników testów, jednak wymaga on wykorzystania rang centylowych. Poniżej przedstawiony jest przepis na przekształcenie surowego, niewiele mówiącego wyniku w coś znacznie bardziej użytecznego i mającego większą wartość informacyjną.

W szkole nauczyciele (lub osoby odpowiedzialne za przekazywanie uczniom wyników testów) mogą przekazywać rezultaty bez podawania wyników. Zamiast tego przedstawiają liczbę wyglądającą jak wartość procentowa, a mającą informować o tym, jak dana osoba wypadła w porównaniu z innymi osobami piszącymi ten test. Ten rodzaj wyniku nazywany jest **rangą centylową**.

Jeśli zobaczymy rangę centylową odpowiadającą naszemu wynikowi w teście, nie będzie ona dla nas użyteczna, jeśli nie będziemy wiedzieli, co oznacza. Z drugiej strony, jeśli mielibyśmy wyjaśnić komuś, jak poradził sobie z testem, i podalibyśmy mu sumę uzyskanych przez niego punktów, też nie byłoby to szczególnie pomocne. Umiejętność tworzenia lub interpretowania rang centylowych jest użyteczna dla osób znajdujących się po obu stronach testu.

Pomiar różnicujący **[Sposób 26.]** to podejście zmierzające do uczynienia wyników bardziej czytelnymi przez porównanie ich ze sobą. Najczęściej spotykanym wynikiem różnicującym jest ranga centylowa. Definiujemy ją jako „odsetek wyników w rozkładzie mających wartość niższą od danego interesującego nas wyniku". Na przykład, jeśli udzieliliśmy prawidłowej odpowiedzi w 15 na 20 przypadków, a dokładnie połowa klasy uzyskała słabszy wynik, nasz centyl wynosi 50.

Tworzenie i podawanie rang centylowych

Dla każdego nauczyciela, osoby zarządzającej zasobami ludzkimi czy kogokolwiek, kto musi przedstawiać innym wyniki testów, możliwość przedstawienia rangi centylowej zamiast surowego wyniku pozwala pomóc osobom zdającym test zrozumieć, jak sobie poradziły, a osobom podejmującym decyzje — zrozumieć konsekwencje ustanawiania różnych standardów wydajności.

Organizacja danych. Tworzenie rang centylowych zaczyna się od zorganizowania wszystkich wyników testów. Dla niewielkiego zbioru danych stosunkowo łatwo można stworzyć **tablicę liczebności**, w której można znaleźć odpowiedzi na rozmaite pytania, no i oczywiście rangi centylowe. Oto przykładowy rozkład 30 wyników uzyskanych podczas szkolnego testu (ułożonych od najniższego do najwyższego), gdzie 100 punktów było wynikiem najwyższym z możliwych:

59, 65, 72, 75, 75, 75, 80, 83, 83, 85, 85, 85, 85, 85, 85, 86, 86, 86, 86, 88, 88, 88, 90, 90, 90, 90, 90, 92, 94, 97

Obliczenie częstości i odsetek. Ze względów praktycznych dane te mogą zostać przedstawione tak jak w tabeli 3.1, gdzie dla każdej wartości obliczona została też częstość występowania i odsetek w zbiorze.

Tabela 3.1. Łączna liczebność dla szkolnego testu

Wynik	Liczebność	Łączna liczebność	Odsetek	Łączny odsetek
59	1	1	3,33 procent	3,33 procent
65	1	2	3,33 procent	6,67 procent
72	1	3	3,33 procent	10,00 procent
75	3	6	10,00 procent	20,00 procent
80	1	7	3,33 procent	23,33 procent
83	2	9	6,67 procent	30,00 procent
85	6	15	20,00 procent	50,00 procent
86	4	19	13,33 procent	63,33 procent
88	3	22	10,00 procent	73,33 procent
90	5	27	16,67 procent	90,00 procent
92	1	28	3,33 procent	93,33 procent
94	1	29	3,33 procent	96,67 procent
97	1	30	3,33 procent	100,00 procent

Tabela 3.1 zawiera następujące informacje: wszystkie wyniki osiągnięte w teście, liczbę osób które osiągnęły poszczególne wyniki, łączną liczbę osób, które uzyskały dany bądź niższy wynik, odsetek osób, które osiągnęły poszczególne wyniki, i łączny odsetek osób, które osiągnęły dany bądź niższy wynik. W kolumnach „łączna liczebność" i „łączny odsetek" zawsze znajduje się suma osób (lub wyników) w rozkładzie (w naszych przypadku jest to 30) i łączny odsetek osób (zawsze 100%).

Określanie rangi centylowej. Aby określić rangę centylową dla dowolnego wyniku w rozkładzie, wykorzystujemy kolumnę „Łączny odsetek". Znajdujemy interesujący nas wynik i szukamy łącznego odsetka w wierszu **bezpośrednio nad** wierszem, w którym ów wynik się znajduje. Na przykład dla wyniku wynoszącego 94 ranga centylowa wynosi 93,33, czyli jest to mniej więcej 93. centyl. Dla wyniku wynoszącego 86 ranga centylowa wynosi 50.

Jeśli zapoznalibyśmy się z kilkunastoma podręcznikami poświęconymi statystyce lub pomiarowi, dowiedzielibyśmy się, że istnieją dwie konkurujące ze sobą definicje rangi centylowej. Ja wolę „odsetek wyników w rozkładzie mających wartość niższą od danego interesującego nas wyniku", ale niektóre książki podają: „odsetek wyników w rozkładzie mających wartość **równą** lub niższą od danego interesującego nas wyniku". Obie definicje są rozsądne i rangi centylowe mogą być za pomocą tablicy liczebności obliczane zgodnie z dowolną z nich. Według pierwszej definicji, setny centyl nie może istnieć. Według drugiej, nie ma centyla zerowego. Badacz winien wybrać tę definicję, która bardziej mu odpowiada, ale przy podawaniu rezultatów zawsze trzeba zaznaczać, której definicji się używa.

Interpretowanie rangi centylowej

Wyobraźmy sobie, że doradca zawodowy poinformował nas, że nasza ranga centylowa wynosi 93. Cóż to oznacza? Najprostsza interpretacja jest taka, że 93 procent wszystkich osób, które zdawały ten test, uzyskały niższy wynik. Prawdziwym będzie też stwierdzenie, że 7 procent osób uzyskało wynik wyższy lub równy. Możemy też odczytywać rangę centylową jako informację o tym, jak bardzo wynik odbiega od normy. Średnia ranga centylowa znajduje się zwykle koło 50. centyla, a dokładnie tam, jeśli rozkład wyników jest normalny, a zwykle tak właśnie jest. Dlatego możemy też powiedzieć, że 93. centyl to całkiem sporo powyżej przeciętnej.

Trzeba uważać, by nie powielać błędu popełnianego czasem przez wielu inteligentnych przecież praktyków statystyki. Wcześniej w tym podrozdziale użyliśmy jako przykładu wyniku testu, w którym udzieliliśmy prawidłowej odpowiedzi na 15 z 20 pytań, a połowa pozostałych uczniów uzyskała słabszy wynik. W tamtym przykładzie nasza ranga centylowa wynosiła 50. Zauważmy, że udzieliliśmy 75 procent poprawnych odpowiedzi, ale nasza ranga centylowa wynosi 50. Nie należy mylić ze sobą tych dwóch rzeczy! Znajomość rangi centylowej nie mówi nam, na ile pytań odpowiedzieliśmy poprawnie.

Gdzie to nie działa?

Nie wolno zapominać, że ranga centylowa jest przydatna tylko wtedy, gdy poszukujemy interpretacji różnicującej. Jeśli chcemy się dowiedzieć, czy opanowaliśmy jakiś zbiór umiejętności, wówczas to, że dowiemy się, jaki odsetek osób opanował te umiejętności w mniejszym lub większym stopniu niż my, nic nam nie powie. Żeby dowiedzieć się, jak wypadamy w odniesieniu do jakiegoś zbioru standardów, nie w odniesieniu do innych ludzi, potrzebujemy pomiaru sprawdzającego **[Sposób 26.]**. W takim przypadku większe znaczenie ma dla nas **odsetek udzielonych poprawnych odpowiedzi** niż **ranga centylowa**.

Zobacz również

- Jeśli założymy, że nasze wyniki rozkładają się normalnie, albo przynajmniej pochodzą z populacji o rozkładzie normalnym, możemy przekształcić dowolny standaryzowany wynik bezpośrednio w rangę centylową, wykorzystując informacje na temat obszarów w granicach krzywej rozkładu normalnego **[Sposób 25.]**.

SPOSÓB 25.

Przewidywanie przyszłości za pomocą krzywej normalnej

Ponieważ niemal wszystko, co mierzymy w świecie naturalnym, ma znany rozkład nazywany „krzywą normalną", możemy wykorzystać szczegóły tego rozkładu do przewidywania przyszłości i odpowiadania na wiele pytań o prawdopodobieństwo.

Wiele spośród zawartych w tej książce sposobów bazuje na zamiłowaniu statystyków do **krzywej normalnej**. „Spróbujmy zrozumieć świat" **[Sposób 23.]** pokazuje, jak wykorzystać krzywą normalną do ogólnego prognozowania osiągnięć w teście. Możemy jednak zrobić też coś więcej.

Dokładny kształt tej intrygującej krzywej znany jest tak doskonale, że możemy z wielką dokładnością prognozować prawdopodobieństwo tego, że uzyskany zostanie określony zakres wyników. Jest wiele typów pytań, które można zadać w związku z osiągnięciami testowymi, a statystyka może nam pomóc poznać odpowiedź na tego rodzaju pytania, zanim w ogóle napiszemy test!

Na przykład:

- Jakie są szanse na to, że osiągniemy wynik mieszczący się pomiędzy dwoma określonymi wynikami?
- Ile osób osiągnie wynik mieszczący się pomiędzy tymi wynikami?
- Jakie są szanse, że zdamy następny test?
- Czy zostaniemy przyjęci na prestiżową uczelnię?
- Jaki procent uczniów w kraju zakwalifikuje się do rządowych stypendiów?
- Jakie są szanse na to, że mój wujek Franek będzie w stanie zdać test kwalifikacyjny do Mensy?

Żeby poznać odpowiedź na tego rodzaju pytania, potrzebujemy konkretnego narzędzia. Niniejszy sposób daje to narzędzie — **tablicę obszarów w granicach krzywej normalnej**.

Tablica obszarów w granicach krzywej normalnej

Krzywa normalna definiowana jest przez średnią arytmetyczną oraz odchylenie standardowe rozkładu, a kształt krzywej jest zawsze taki sam, niezależnie od tego, co mierzymy (dopóty, dopóki system pomiaru pozwala na występowanie różnych wyników). Proporcje wyników mieszczących się w różnych obszarach krzywej, takich jak przestrzeń pomiędzy określonymi odchyleniami standardowymi i odległości od średniej, zostały przedstawione wcześniej.

Ten sposób jest oparty na tabeli, która wygląda na skomplikowaną, ale zawiera tyle użytecznych informacji, że szybko stanie się jednym z naszych ulubionych narzędzi statystycznych. Nie rozwodząc się dłużej nad złożonością tabeli, weźmy głęboki oddech i spójrzmy na nią (tabela 3.2).

Tabela 3.2. Obszary w granicach krzywej normalnej

Wynik typu z	Proporcja wyników pomiędzy średnią a z	Proporcja wyników w większym obszarze	Proporcja wyników w mniejszym obszarze
0,00	0,00	0,50	0,50
0,12	0,05	0,55	0,45
0,25	0,10	0,60	0,40
0,39	0,15	0,65	0,35
0,52	0,20	0,70	0,30
0,67	0,25	0,75	0,25
0,84	0,30	0,80	0,20
1,04	0,35	0,85	0,15
1,28	0,40	0,90	0,10
1,65	0,45	0,95	0,05
1,96	0,475	0,975	0,025
4,00	0,50	1,00	0,00

Odcyfrowywanie tabeli

Zanim zaczniemy używać tego sprytnego narzędzia, musimy wziąć głęboki oddech i rozeznać sytuację. Informacje przedstawione w tabeli uprościłem na kilka sposobów. Po pierwsze, umieściłem tam tylko niektóre z wartości, które można obliczyć. Tak naprawdę w podręcznikach statystyki znajdują się tablice, na których umieszczono wartości od 0,00 do 4,00, zwiększające się co 0,01. To mnóstwo informacji do zaprezentowania, dlatego postanowiłem pokazać tu tylko wycinek zawierający najczęściej wykorzystywane wartości, w tym wartości typu **z** niezbędne dla 90-procentowej ufności (1,65) oraz 95-procentowych przedziałów ufności (1,96). Więcej informacji na temat przedziałów ufności znaleźć można w rozdziale 1., w podrozdziale „Precyzyjne ocenianie” **[Sposób 6.]**.

Zaokrągliłem też proporcje do dwóch miejsc po przecinku. Wreszcie, użyłem w tabeli symbolu **z**, aby wskazać na odległość od średniej mierzoną w odchyleniach standardowych. Więcej informacji na temat wyników typu **z** znaleźć można w rozdziale 3., w podrozdziale „Opracowujemy surowe wyniki” **[Sposób 26.]**.

Po zrozumieniu tego, w jaki sposób tabela została uproszczona, pierwszym krokiem w stronę wykorzystania jej do prognozowania prawdopodobieństwa uzyskania danych wyników lub odpowiadania na pytania statystyczne jest zrozumienie tego, co znajduje się w każdej z czterech kolumn.

Kolumna z

Wyobraźmy sobie krzywą rozkładu normalnego **[Sposób 23.]**. Jeśli interesuje nas jakiś wynik, który może znaleźć się w jakimś miejscu dolnej poziomej linii, to będzie to jakaś odległość od średniej. Wynik ten może być większy lub mniejszy niż średnia. Odległość od średniej arytmetycznej wyrażona w odchyleniach standardowych to właśnie wynik typu **z**. Wynik typu **z** o wartości 1,04 opisuje wynik znajdujący się o odrobinę więcej niż jedno odchylenie standardowe od średniej. Ponieważ krzywa normalna jest symetryczna, nie zajmujemy się odnotowaniem, czy odległość jest dodatnia czy ujemna, i wszystkie wyniki typu **z** są przedstawiane jako wyniki dodatnie.

Proporcja wyników pomiędzy średnią a z

W tej przestrzeni pomiędzy danym wynikiem a średnią będzie się znajdowała określona proporcja wyników. Jest to prawdopodobieństwo tego, że losowy wynik znajdzie się w obszarze pomiędzy średnią a dowolnym wynikiem typu **z**.

Proporcja wyników w większym obszarze

Możemy też opisać obszar pomiędzy dowolnym **z** i **z** równym 4,00 albo końcem krzywej.

Teoretycznie krzywa nie ma prawdziwego końca, ale wynik typu **z** wynoszący 4,00 będzie zawierał prawie 100 procent wyników.

Jednak krzywa ma dwa końce. Jeśli tylko nasze **z** nie wynosi 0,0, odległość pomiędzy **z** a jednym końcem krzywej będzie większa niż odległość pomiędzy **z** a drugim jej końcem. Ta kolumna odnosi się do obszaru pomiędzy z a najbardziej oddalonym krańcem krzywej, a wartości w tej kolumnie to proporcja wyników, które znajdą się w tym obszarze. Innymi słowy, to szansa na to, że przypadkowa osoba wygeneruje wynik mieszczący się w tym obszarze.

Proporcja wyników w mniejszym obszarze

Ta kolumna odnosi się do obszaru pomiędzy **z** i najbliższym końcem krzywej. To proporcja wyników, które znajdą się w tym obszarze.

Oszacowanie szansy na uzyskanie wyniku wyższego lub niższego od innego wyniku

Jeśli chcemy dowiedzieć się, jakie mamy szanse, aby dostać się na wybraną uczelnię, musimy uzyskać informację, jaka liczba punktów na egzaminie wstępnym nam to umożliwi (czyli inaczej, jaki w tym przypadku będzie **próg dopuszczenia**). Gdy już znamy ten wynik, musimy znaleźć średnią arytmetyczną i odchylenie standardowe dla testu. (Wszystkie te informacje będą zapewne dostępne na stronie internetowej uczelni). Następnie przekształcamy nasz surowy wynik w wynik typu **z [Sposób 26.]**, po czym odnajdujemy ten lub zbliżony wynik typu **z** w tabeli 3.2.

Stwierdzamy, czy próg dopuszczenia znajduje się powyżej średniej arytmetycznej:

- Jeśli to prawda, patrzymy na kolumnę „Proporcja wyników w mniejszym obszarze". Określone są w niej nasze szanse na uzyskanie wyniku równego lub wyższego progowi dopuszczenia i tym samym na dostanie się na uczelnię.
- Jeśli próg dopuszczenia znajduje się poniżej średniej (co jest wysoce nieprawdopodobne, ale musimy założyć taką możliwość, aby dokładnie poznać zastosowanie tego sposobu), należy odwołać się do „Proporcji wyników w większym obszarze". Będzie to proporcja przyjmowanych studentów i tym samym nasze szanse na to, że zostaniemy przyjęci (o ile inne warunki będą równe).

Jeżeli chodzi o szanse na osiągnięcie wyniku **niższego** niż dany, proces jest dokładnie odwrotny w stosunku do opisanego powyżej. Szanse osiągnięcia wyniku niższego niż próg dopuszczenia znajdujący się poniżej średniej można odczytać z kolumny „w mniejszym obszarze". Szanse osiągnięcia wyniku niższego niż próg dopuszczenia znajdujący się powyżej średniej można odczytać z kolumny „w większym obszarze".

Oszacowanie szansy na osiągnięcie wyniku pomiędzy dwoma innymi wynikami

Szanse na osiągnięcie wyniku znajdującego się w dowolnym zakresie wyników można określić, badając proporcję wyników normalnie uzyskiwanych w tym zakresie.

Jeśli chcemy wiedzieć, jaka proporcja wyników wypada pomiędzy dowolnymi dwoma punktami krzywej, musimy określić te punkty jako wyniki typu **z** i obliczyć odpowiednią proporcję. W zależności od tego, czy oba wyniki znajdują się po tej samej stronie średniej, właściwą proporcję wyników pomiędzy dwoma innymi wynikami można uzyskać na jeden z dwóch sposobów:

- Jeśli wyniki typu **z** znajdują się po tej samej stronie krzywej, proporcje wyników odczytujemy zarówno z kolumny „w większym obszarze", jak i z kolumny „w mniejszym obszarze", a następnie wartość niższą odejmujemy od wyższej.
- Jeśli wyniki typu **z** znajdują się po obu stronach średniej, korzystamy z kolumny „proporcja wyników pomiędzy średnią a **z**". Odczytujemy wartości dla obu wyników i sumujemy je.

Tworzenie rang centylowych

Trzecim zastosowaniem tej tabeli jest tworzenie rang centylowych. Na temat takich wyników **różnicujących** więcej możemy dowiedzieć się z podrozdziału „Tworzenie rang centylowych" **[Sposób 24.]**. Dla wyników powyżej średniej arytmetycznej ranga centylowa wynosi tyle co „proporcja wyników pomiędzy średnią a z" plus 0,5. Dla wyników poniżej średniej ranga centylowa wynosi tyle co „proporcja wyników w mniejszym obszarze".

Określanie istotności statystycznej

Kolejnym zastosowaniem dla tego rodzaju tabel jest przypisywanie różnicom wyników istotności statystycznej **[Sposób 4.]**. Wiedząc, jaka proporcja wyników znajdzie się w określonej odległości od siebie bądź dalej, możemy takiemu rezultatowi przypisać poziom prawdopodobieństwa statystycznego.

Co bardziej użyteczne, inne wartości statystyczne, takie jak korelacje i proporcje, mogą być przekształcane w wyniki typu **z**, a powyższa tabela może być wykorzystywana do porównywania tych wartości z zerem lub ze sobą nawzajem.

Dlaczego to działa?

Sposób „Spróbujmy zrozumieć świat" **[Sposób 23.]** daje dobry obraz krzywej normalnej. Jednak dobre pojęcie na temat kształtu rozkładu normalnego można sobie wyrobić tylko przez przyjrzenie się temu, w jaki sposób zmieniają się wartości w tabeli 3.2. W pobliżu średniej, gdzie znajdują się wiersze z niewielkimi wynikami typu **z**, przypada spora proporcja wyników. W miarę jak coraz bardziej oddalamy się od średniej, potrzeba coraz to większych i większych obszarów krzywej do zawarcia takiej samej proporcji wyników.

Na przykład, aby objąć ostatnie 5 procent rozkładu, trzeba przeskoczyć od **z** równego 1,65 do 4. Natomiast w pobliżu średniej do objęcia 5 procent wyników wystarczy przeskoczyć od **z** = 0,12 do **z** = 0,15. Tabela ilustruje, jak pospolite jest to, co pospolite, i jak rzadkie jest to, co rzadko spotykane.

Zobacz również

- Własne dokładne obszary krzywej rozkładu normalnego możemy obliczyć, korzystając ze wskazówek zawartych na stronie internetowej *http://www.psychstat.missouristate.edu/introbook/sbk11m.htm*. Na części tej strony, którą zajmuje się David Stockburger, znajduje się dobre omówienie tematu i kilka interaktywnych kalkulatorów. Wybierając się tam z wizytą, nie należy dać się zmylić takim słowom jak **Mu** i **Sigma**. To w żargonie statystycznym nazwy średniej i odchylenia standardowego.

SPOSÓB 26.

Opracowujemy surowe wyniki

Surowy wynik testu znaczy niewiele lub zgoła nic. Wystarczy jednak przekształcić ten żałosny wynik w „wynik typu **z**", a trudno będzie uwierzyć, ile informacji zmieściło się w tej jednej małej superliczbie.

To zadziwiające, jak niewiele informacji jest przekazywanych przez jeden surowy wynik uzyskany na przykład w teście w szkole średniej. O co mi chodzi? Gdybym wrócił do domu ze szkoły i powiedział mamie, że w ważnym teście dostałem dziś 16 punktów, pewnie powiedziałaby między innymi: „Dlaczego w wieku 42 lat mieszkasz ciągle z nami?" oraz „To ładnie skarbie. A czy to dobrze?".

Gdy przekazujemy komuś wyłącznie surowy wynik, przekazujemy tak naprawdę bardzo niewiele informacji. Nie wiesz, czy 16 **to dobrze**. Nie wiemy, czy 16 to stosunkowo dużo, czy mało. Czy większość osób uzyskuje 16 i więcej, czy też większość uzyskuje mniej niż 16 punktów? Nawet jeśli znamy zakres wyników w teście, liczbę możliwych do uzyskania punktów itd., wciąż nie możemy porównać osiągnięć w tym teście z osiągnięciami w poprzednim teście, następnym teście lub w teście dotyczącym czego innego. Surowe wyniki są praktycznie bez znaczenia.

Nie martwmy się! Wciąż możemy zrozumieć nasze osiągnięcia i osiągnięcia innych. Wciąż możemy podejmować decyzje odnośnie selekcji, a także porównywać osiągnięcia różnych osób, w różnych testach. Wciąż jest dla nas nadzieja!

Surowe wyniki mogą zostać przeobrażone w nową liczbę, która robi to wszystko, do czego walczący w wadze koguciej surowy wynik nie jest zdolny. Surowe wyniki mogą zostać przeobrażone w superliczbę: **wynik typu z**. Inaczej niż surowy wynik, **z** mówi nam, czy osiągnięcia są powyżej czy poniżej przeciętnej, a także jak bardzo powyżej lub poniżej przeciętnej. **Z** pozwala nam również porównywać osiągnięcia w różnych testach i różnych przypadkach, a nawet pomiędzy różnymi osobami.

Wyliczanie wyników typu z

Wynik typu **z** to wynik surowy, który został przeobrażony w taki sposób, że nowo powstała liczba wskazuje, jak bardzo wynik surowy odbiega od średniej.

Oto równanie:

$$z = \frac{\text{surowy wynik} - \text{średnia arytmetyczna}}{\text{odchylenie standardowe}}$$

Aby zamienić surowy wynik w **z**, musimy odjąć od niego średnią arytmetyczną, a uzyskany wynik podzielić przez odchylenie standardowe. Odchylenie standardowe rozkładu to przeciętna odległość każdego wyniku od średniej **[Sposób 2.]**.

Zrozumienie osiągnięć

Wyniki typu **z** zwykle przybierają wartości pomiędzy –3 a +3. Przyjrzawszy się górnej części równania na wynik typu **z**, możemy zauważyć rzeczy następujące:

- jeśli surowy wynik jest większy niż średnia, **z** będzie pozytywny;
- jeśli surowy wynik jest poniżej średniej, **z** będzie negatywny;
- jeśli surowy wynik jest równy średniej, **z** będzie wynosił 0.

Wyniki typu **z** zazwyczaj wahają się od –3 do +3, ponieważ **rozkład normalny** wyników ma zwykle szerokość sześciu odchyleń standardowych **[Sposób 23.]**.

Bystrzy specjaliści od pomiaru wyników wykorzystują sztuczkę z wynikiem typu **z** przy podawaniu rezultatów. Zamiast dostarczać surowe wyniki, dają odbiorcom tylko wyniki oparte na wynikach typu **z**, generalnie znane jako wyniki standaryzowane **[Sposób 27.]**. Te wyniki standaryzowane mają znane, stabilne cechy. Dlatego też, jeśli znamy cechy tych wyników (średnią arytmetyczną i odchylenie standardowe), możemy zamienić je z powrotem na wyniki typu **z** i dzięki temu dowiedzieć się, jak wypadliśmy w porównaniu z innymi.

Aby zilustrować, w jaki sposób można wykorzystać ten wzór do odkrycia ukrytych informacji dotyczących naszych osiągnięć, przeanalizujemy testy ACT. Testy ACT (skrót od *American College Test*) są pisane przez uczniów drugich klas wielu szkół średnich w Stanach Zjednoczonych i wiele uczelni wyższych wymaga zaliczenia tego testu od kandydatów. Jest to test osiągnięć i zdolności, mający prognozować osiągnięcia w szkole wyższej.

Wyniki dla każdej części testu mieszczą się w zakresie od 1 do 36. Mimo że na przestrzeni ostatnich kilku dekad wyniki się poprawiały i statystyki się zmieniały, oficjalna średnia arytmetyczna dla testów ACT jest zwykle podawana jako 18, z odchyleniem standardowym równym 6. Wyobraźmy sobie, że 3 uczniów podeszło do ACT i osiągnęło trzy różne wyniki. Możemy wykorzystać średnią i odchylenie standardowe z rozkładu wyniku ACT do przekształcenia ich w wyniki typu **z**, tak jak zostało to pokazane w tabeli 3.3.

Tabela 3.3. Przekształcenie surowych wyników w wyniki typu z

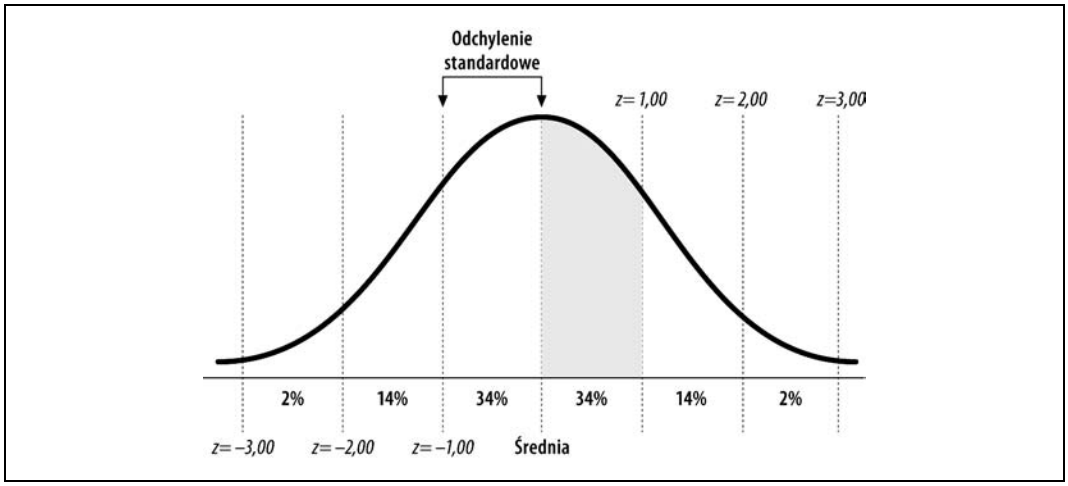
Uczeń	**Wynik w teście ACT**	$\frac{\text{surowy wynik} - \text{średnia arytmetyczna}}{\text{odchylenie standardowe}}$	**Wynik typu z**
Błażej	14	$\frac{14-18}{6} = -\frac{4}{6}$	–0,67
Eryk	18	$\frac{18-18}{6} = \frac{0}{6}$	0,00
Adrian	24	$\frac{24-18}{6} = \frac{6}{6}$	1,00

Z Błażeja jest ujemne, stąd wiemy, że osiągnął wynik poniżej średniej. Dokładnie rzecz biorąc, osiągnął wynik mieszczący się o dwie trzecie odchylenia standardowego poniżej średniej. Wynik typu **z** Eryka wynoszący 0,00 oznacza, że sprawił się przeciętnie w porównaniu z innymi, którzy na przestrzeni lat przystępowali do testu ACT. Adrian poradził sobie najlepiej, uzyskując wynik mieszczący się o pełne odchylenie standardowe powyżej średniej.

W prawdziwym ACT średnia i odchylenie standardowe wahają się z roku na rok. **Prawdziwa** średnia i odchylenie standardowe dla ostatnich kilku lat wynosiły około 21 w przypadku średniej i około 4,5 w przypadku odchylenia standardowego.

Rozpoznawanie wyjątkowości naszych osiągnięć

Choć wiedza na temat tego, jaki osiągnęliśmy wynik w porównaniu z innymi osobami piszącymi test, jest bardziej wartościowa niż znajomość samego surowego wyniku, prawdziwa potęga interpretacyjna wyników typu **z** bierze się z ich relacji z krzywą normalną. Rysunek 3.3 to wykres rozkładu normalnego, podobnego do tego znajdującego się w podrozdziale „Spróbujmy zrozumieć świat" **[Sposób 23.]**.



Rysunek 3.3. Wyniki typu z i krzywa normalna

Różnica pomiędzy rysunkiem z podrozdziału „Spróbujmy zrozumieć świat" **[Sposób 23.]** a tym, jest taka, że zamiast pokazywać odległość każdego odchylenia standardowego od średniej, rysunek 3.3 pokazuje te wartości jako wyniki typu **z**. Wykorzystując wiedzę na temat obszarów w granicach krzywej normalnej, możemy dowiedzieć się z wyniku typu **z** jeszcze więcej. Jeśli wyniki rozkładają się normalnie, można bardzo dużo powiedzieć na temat prawdopodobieństwa wystąpienia wyników w określonym zakresie.

Wyniki uczniów podane w tabeli 3.3 mogą też zostać zinterpretowane jako liczba uczniów, którzy poradzili sobie lepiej (lub gorzej) od innych. Wynik 0,00 Eryka oznacza, że poradził sobie lepiej niż 50 procent uczniów. Wyniki tych dzieciaków można też wyrazić jako prawdopodobieństwo. Jest 50-procentowa szansa na to, że Eryk osiągnie wynik typu **z** na poziomie 0,00 lub wyższy. Ponieważ szansa na osiągnięcie w tym bądź innym teście wyniku typu **z** równego lub wyższego 1,00 wynosi tylko 16 procent, Adrian — w porównaniu z innymi uczniami przystępującymi do testu — wypadł bardzo dobrze.

Dlaczego to działa?

Jeśli przekształcanie surowych wyników na wyniki typu **z** po to, by móc porównywać ze sobą osiągnięcia różnych osób, wydaje się nam sensowne, to nie jesteśmy w tej opinii odosobnieni. Przez pierwsze 100 lat w świecie pomiarów edukacyjnych socjologowie (i każdy, kto musiał mierzyć osiągnięcia innych ludzi) dawali się skusić przez prostotę **interpretacji różnicujących**. Jeśli nie jesteśmy pewni, co tak naprawdę oznacza uzyskany w teście wynik,

możemy porównać go z wynikami innych osób. Będziemy przynajmniej wiedzieli, czy mamy **więcej** lub **mniej** niż inni tego, co właśnie zmierzyliśmy.

Innym sposobem interpretowania wyników edukacyjnych i psychologicznych jest **interpretacja odniesiona do kryterium**. To podejście wymaga lepszej znajomości mierzonej cechy lub zawartości i określenia z góry, ile to jest „wystarczająco dużo". Pomiar w odniesieniu do kryterium pozwala wszystkim osiągnąć ten sam wynik dopóty, dopóki spełniają takie same kryteria. Poprzednie podejście było i wciąż jest najpopularniejszą metodą interpretacji, podczas gdy to drugie dopiero zaczęło się przyjmować.

SPOSÓB 27.

Standaryzowanie wyników

Co zaskakujące, wyniki żadnego z powszechnie znanych testów, od których wiele zależy, takich jak testy SAT, ACT lub testy na inteligencję, nie są podawane w formie surowej. Zamiast tego, ta bezużyteczna liczba jest przekształcana w liczbę mającą większą wartość informacyjną — taką, która może zostać wykorzystana do zrozumienia naszego osiągnięcia w porównaniu z osiągnięciami wszystkich innych, którzy podchodzili do testu. Gdy zrozumiemy wyniki „standaryzowane", będziemy potrafili sami je obliczać, a nawet tworzyć nowe.

Podrozdział „Opracowujemy surowe wyniki" **[Sposób 26.]**. omawia supermoce wyników typu **z**. Te standaryzowane wyniki dodają wszelkiego typu informacje do nic nie znaczących surowych wyników. Dzięki temu każdy Czytelnik tej książki może interpretować wyniki typu **z** i podejmować decyzje w oparciu o pozyskane informacje.

Jeśli jednak zechcemy interpretować wiele raportów (takich jak wyniki egzaminu SAT, do którego właśnie podchodziliśmy), nie zobaczymy tam wyników typu **z**, ale dziwaczny wynik standaryzowany, opracowany i używany wyłącznie przez daną organizację, coś na kształt wyniku typu **z**, ale różniący się od niego na tyle, by dla osoby niewtajemniczonej pozostał bezużyteczny.

Nie ma się czego bać. Oto narzędzia, których będziemy potrzebowali do interpretowania tych dziwacznych wyników standaryzowanych, a nawet, jeśli będziemy chcieli, do tworzenia własnych (na przykład gdy będziemy chcieli przedstawiać innym wyniki naszego dziwnego testu, który stanie się szalenie popularny i sprawi, że będziemy bogaci jak pan ACT, pan IQ, czy ktokolwiek inny, kto na tym zarabia w naszym testami stojącym społeczeństwie).

Problemy z wynikami typu z

Jest pewna, że tak się wyrażę, **szpetota** wyników typu **z**. To ona sprawia, że wyniki typu **z** nie są powszechnie używane do przekazywania informacji o osiągnięciach osobom, które przystąpiły do testu, rodzicom takich osób, uczelniom czy firmom, które podejmują decyzję o rekrutacji. Natomiast większość firm zajmujących się przeprowadzaniem testów używa wyniku typu **z** jako pierwszego kroku do stworzenia atrakcyjniejszego wyniku standaryzowanego.

Surowy wynik jest przekształcany w wynik typu **z** za pomocą następującego wzoru:

$$z = \frac{\text{surowy wynik} - \text{średnia arytmetyczna}}{\text{odchylenie standardowe}}$$

To równanie, opisane dokładniej w podrozdziale „Opracowujemy surowe wyniki" **[Sposób 26.]**, daje wynik typu **z**, zwykle wahający się od –3 do +3, przy czym jego przeciętna wartość to 0, a odchylenie standardowe wynosi 1. Choć jest on bardzo użyteczny jako narzędzie do interpretowania osiągnięć w testach, ludziom te liczby nie podobają się ze względu na kilka problemów:

- Mogą być ujemne. Tak naprawdę, połowa wszystkich wyników typu **z** będzie ujemna. Trudno przekonać osoby, które podeszły do testu, że wynik ujemny może oznaczać cokolwiek dobrego.
- Wynik 0 to wynik przeciętny! Jeśli nie możemy przekonać ludzi, że liczba ujemna niekoniecznie jest czymś złym, wyobraźmy sobie próbę przekonania rodziców, że spodziewamy się, iż ich mały Romuś otrzyma zero z ważnego egzaminu, i będziemy zadowoleni, gdy tak się stanie.
- Najwyższy możliwy wynik to 3, a osiągnie go tylko jedna z każdych stu osób, które przystąpią do testu. Cała ta ciężka praca przy przygotowaniach do testu tylko po to, żeby dostać marne 3!

Ludzie zajmujący się pomiarami szukali i znaleźli inne standaryzowane skale wyników testów, znacznie przyjemniejsze w odbiorze. Sztuczka polega na tym, aby wyjść od wyniku typu **z**, a następnie przekształcić go na jakąś inną skalę, której średnia i odchylenie standardowe wyglądają przyjaźniej.

Tworzenie i interpretowanie wyników typu z

Jednym z problemów z wynikami typu **z** jest to, że średnia wynosi w nich zero. Podawanie zera jako wyniku neutralnego źle działa na niektórych nauczycieli, rodziców i uczniów. Problem ten możemy rozwiązać, przechodząc w dół alfabetu, od **z** do **T**.

Wyniki typu **T** to przeobrażenie wyników typu **z** w nowy rozkład, w którym średnia wynosi 50, a odchylenie standardowe 10. Równanie dla wyniku typu **T** przekształca wynik wstecz. Oto wzór na wynik typu **T**:

$$\mathrm{T} = \mathrm{z}(10) + 50$$

Jeśli więc mały Romuś wypadł w ważnym teście przeciętnie i otrzymał wynik typu **z** wynoszący 0, zamiast przekazywać tę niepokojącą wartość jego rodzicom, możemy przekształcić ją w **T**:

$$\mathrm{T} = 0{,}00(10) + 50, \quad \mathrm{T} = 0{,}00 + 50, \quad \mathrm{T} = 50$$

Następnie informujemy rodziców, że wynik Romusia to 50. Gratulacje! Aby nadać temu wynikowi jakąś wartość informacyjną, dobry nauczyciel lub pedagog szkolny wyjaśni, że wyniki typu **T** wahają się zwykle od 20 do 80, przy czym 50 to przeciętna.

Wyniki typu **T** są wykorzystywane do przekazywania informacji o wynikach testów jako coś lepszego od wyników typu **z**. Wyniki te nie mogą być ujemne, a średnią jest w nich poważniej wyglądające 50.

Jednym z popularnych testów wykorzystujących rozkład typu **T** jest test Minnesota Multiphase Personality Inventory-II, pozwalający na ocenę stanu psychicznego człowieka. Średnia wyników na każdej podskali MMPI-II to 50, z odchyleniem standardowym wynoszącym 10. Umieszczając wynik każdego podtestu na jednej skali, możemy porównać poszczególne cechy i stworzyć profil wyników, by lepiej zrozumieć osobę poddaną testowi.

Tworzenie własnych wyników standaryzowanych

Twórcy testów odkryli też inne metody raportowania wyników standardowych. W tabeli 3.4 znajduje się lista wielu spośród najlepiej znanych ważnych testów, które większość ludzi przeszła albo pewnego dnia przejdzie.

Tabela 3.4. Pospolite standaryzowane rozkłady wyników

Test	Typowy zakres wyników	Średnia	Odchylenie standardowe
Wyniki typu **z**	–3,00 do 3,00	0	1
Wyniki typu **T**	20 do 80	50	10
American College Test (ACT)	1 do 36	18	6
SAT	200 do 800	500	100
Graduate Record Exam (GRE)	200 do 800	500	100
Graduate Management Admission Test (GMAT)	200 do 800	500	100
Law School Admission Test (LSAT)	120 do 180	150	10
Medical College Admission Test (MCAT)	1 do 15	8	2,5
Test na inteligencję Wechslera	55 do 145	100	15
Test na inteligencję Stanforda-Bineta	52 do 148	100	16

Ponieważ wyniki testów mają rozkład normalny, możemy interpretować każdy z tych wyników, umieszczając go na krzywej normalnej i sprawdzając, czy uzyskany wynik był przeciętny, wyjątkowo niski, czy też wyjątkowo wysoki **[Sposób 23.]**.

Tworzymy własny wynik standaryzowany

Dla zabawy możemy stworzyć swój własny standaryzowany rozkład wyników, z taką średnią i odchyleniem standardowym, jakie nam się podobają. Nie podoba nam się to, że uzyskaliśmy w teście SAT 350 punktów? Przekształćmy ten wynik w inny, o takim rozkładzie, jaki chcemy.

Wyobraźmy sobie na przykład, że wolelibyśmy, aby średnia arytmetyczna rozkładu wynosiła 752 365, a odchylenie standardowe 216 456 (no bo kto by nie wolał?). Nazwijmy ten rozkład **Rozkładem Wyniku Freya**. Uogólniając wzór na wynik typu **T**, możemy przekształcić nasz wynik SAT wynoszący 350 na wynik Freya. Musimy pamiętać, że wychodzimy od wyniku typu **z** dla wyniku SAT wynoszącego 350:

$$z = \frac{\text{surowy wynik} - \text{średnia}}{\text{odchylenie standardowe}} = \frac{350 - 500}{100} = \frac{-150}{100} = -1{,}50$$

Następnie przekształcamy to w wynik Freya:

$$\text{Frey} = -1{,}50(216\,456) + 752\,365 = -324\,684 + 752\,365 = 427\,681$$

No proszę, czy wynik 427 681 nie wygląda lepiej niż wynik 350? Ponieważ znamy średnią rozkładu Freya, oba wyniki będziemy interpretować w ten sam sposób. Oba są poniżej przeciętnej i cały czas znajdują się o półtora odchylenia standardowego poniżej średniej arytmetycznej. Nie zmieniliśmy więc rzeczywistości, a jedynie opisujące ją liczby.

Dlaczego to działa?

Rozkład wyników typu **z** ma średnią 0 i odchylenie standardowe 1. Jest tak ze względu na użyte równanie. Dzielenie grupy wartości przez ich odchylenie standardowe daje nam odchylenie standardowe nowego rozkładu wynoszące 1. Po odjęciu średniej od każdego wyniku w rozkładzie nowe wartości rozkładają się wokół średniej 0.

Jeśli chcemy, by wyniki miały konkretną średnią i wybrane odchylenie standardowe, możemy wziąć każdy wynik typu **z** i przekształcić go wspak, zastępując średnią, o wartości 0, dowolną liczbą i odchylenie standardowe, o wartości 1, również dowolną liczbą.

Zrozumienie oceniania różnicującego

Mówiliśmy o informacji zawartej w ocenianiu różnicującym i jej intuicyjnych zaletach z punktu widzenia statystyki, ale nie jest to jedyna droga do stworzenia użytecznych wyników, nie zawsze jest to też najlepsza metoda.

Jak omówiliśmy w podrozdziale „Opracowyujemy surowe wyniki" **[Sposób 26.]**, opracowując system oceniania i tworząc testy, możemy wybrać jeden z dwóch sposobów podejścia do problemu:

Ocenianie różnicujące
Zgodne z rozumowaniem, że aby najlepiej zrozumieć wyniki w jakiejś dziedzinie (takiej jak gra w filmie czy pisanie testu ACT), poziom osiągnięty przez jedną osobę powinien być porównany z osiągnięciami innych.

Ocenianie sprawdzające
Określa osiągnięcia na podstawie zbioru kryteriów takich jak zasób wiedzy, zbiór umiejętności, realizacja poleceń czy cechy diagnostyczne.

Jeśli podejście różnicujące wydaje nam się sensowne, to możemy używać przedstawionych tu narzędzi do interpretowania naszych osiągnięć w powszechnych dziś standaryzowanych testach.

SPOSÓB 28.

Zadawanie właściwych pytań

Nauczyciel, osoba prowadząca rozmowy kwalifikacyjne oraz każdy znajdujący się w sytuacji, w której chce ocenić czyjąś wiedzę, może zadawać pytania na rozmaite sposoby. Oto kilka narzędzi z dziedziny pomiaru edukacyjnego, które pozwalają na zadawanie właściwych pytań we właściwy sposób.

Przez ponad sto lat klasy w szkołach były środowiskiem pełnym pytań i odpowiedzi. Poza szkołą testy stają się coraz bardziej popularne w pracy i przy podejmowaniu decyzji o zatrudnieniu pracownika. Mało tego, wystarczy, że weźmiemy do ręki dowolną gazetę dla pań, a znajdziemy w niej test sprawdzający, czy w stosunku do ludzi spotkanych na imprezie jesteśmy „przyjaźni" czy „chłodni" (ja jestem „chłodny" — ktoś ma z tym problem?).

W wielu profesjach trzeba zadawać dobre pytania lub pisać dobre testy:

- Nauczyciele zadają uczniom pytania, czy to w czasie zajęć, czy podczas korepetycji, po to, aby ocenić poziom zrozumienia tematu przez ucznia.
- Szkoleniowcy piszą pytania, aby ocenić efektywność zajęć praktycznych.
- Kadrowcy przygotowują standardowe pytania mające na celu zmierzenie umiejętności kandydatów.

Każdy, kto musi oceniać poziom wiedzy innych osób, głowi się nad tym, jakie pytanie zadać, by trafić w samo sedno. Ten sposób stanowi rozwiązanie dwóch najczęściej spotykanych problemów przy pisaniu testów lub tworzeniu pytań, których zadaniem jest ocena wiedzy lub poziomu zrozumienia:

- Jak skonstruować dobre pytanie?
- O co należy zapytać?

Konstruowanie trafnego pytania

Jeśli celem działania jest szybkie i efektywne mierzenie wiedzy, formatem pytania, który trudno będzie pobić, jest **zadanie wielokrotnego wyboru**.

Pytania wielokrotnego wyboru to taki rodzaj zadań, które przedstawiają pytanie lub polecenie (znane jako **trzon zadania**), a następnie nakazują **wybór** właściwej odpowiedzi lub reakcji z listy dostępnych opcji.

Abyśmy mogli mówić o pisaniu dobrych zadań wielokrotnego wyboru, używając właściwej terminologii, niezbędne jest krótkie wprowadzenie.

Oto przykład zadania wielokrotnego wyboru:

Kto napisał powieść *Wielki Gatsby*?	**Trzon zadania**
A. Faulkner	Dystraktor
B. Fitzgerald	Odpowiedź prawidłowa (zgodnie z kluczem do zadania)
C. Hemingway	Dystraktor
D. Steinbeck	Dystraktor

Jak widać, każdy element zadania ma swoją nazwę. Odpowiedź prawidłowa jest nazywana **odpowiedzią prawidłową** (ach, ten naukowy żargon), a odpowiedzi nieprawidłowe nazywane są **dystraktorami**.

Przeprowadzono trochę (choć niewiele) badań nad cechami zadań wielokrotnego wyboru i tym, w jaki sposób pisać dobre zadania. Aby pisać dobre zadania wielokrotnego wyboru, możemy skorzystać z wyników tych badań w postaci zbioru najważniejszych wskazówek:

Należy zawrzeć od 3 do 5 opcji do wyboru.
Zadania powinny zawierać tyle opcji wyboru, aby zgadywanie było trudne, ale nie aż tyle, by dystraktory były niewiarygodne albo by udzielenie odpowiedzi trwało zbyt długo.

Należy unikać odpowiedzi „wszystkie z powyższych".
Niektóre osoby, postępując zgodnie ze strategią rozwiązywania testów, regularnie będą zgadywały, że to o tę opcję chodzi. Inne, zgodnie z tą samą strategią, będą jej unikały. Tak czy owak, opcja ta nie sprawdza się dobrze jako dystraktor. Co więcej, oszacowanie tego, czy opcja „wszystkie z powyższych" jest wiarygodna, wymaga umiejętności analitycznych, których poziom rozwoju jest różny u różnych osób. Ocenienie tej konkretnej umiejętności na ogół nie jest przedmiotem testu.

Należy unikać odpowiedzi „żadna z powyższych".
Ta wskazówka jest tu obecna z tego samego powodu, co poprzednia. Dodatkowo, z jakiegoś powodu, nauczyciele mają tendencję do tworzenia zadań, w których odpowiedź „żadne z powyższych" jest z największym prawdopodobieństwem odpowiedzią prawidłową i niektórzy uczniowie to wiedzą.

Wszystkie opcje muszą być wiarygodne.
Jeśli jakaś opcja jest ewidentnie nieprawidłowa, bo nie wydaje się być w ogóle związana z innymi opcjami albo pochodzi z dziedziny nie obejmowanej przez test, albo też nauczyciel umieścił ją dla żartu, nie spełnia w ogóle funkcji dystraktora.

Uczniowie nie będą brali jej pod uwagę, więc na przykład zadanie z czterema możliwościami odpowiedzi stanie się zadaniem z trzema odpowiedziami i znacznie łatwiej będzie odgadnąć właściwą.

Opcje należy szeregować logicznie lub losowo.

Niektórzy nauczyciele mają tendencję do takiego układania zadań, aby prawidłowa odpowiedź była zawarta w określonej opcji (na przykład B lub C). Uczniowie mogą się w tym zorientować. Dodatkowo, niektóre kursy uczące rozwiązywania testów wielokrotnego wyboru sugerują tę technikę jako element strategii rozwiązywania testów. Nauczyciele mogą panować nad takimi tendencjami, szeregując opcje wg jakiejś zasady (na przykład od najkrótszej do najdłuższej, alfabetycznie, chronologicznie).

Innym rozwiązaniem problemu szeregowania odpowiedzi jest przejrzenie szkicu testu w edytorze tekstów i ustawienie opcji w sposób losowy. Oczywiście komputerowa randomizacja jest też rozwiązaniem dla twórców komercyjnych testów standaryzowanych.

Trzon zadania powinien być dłuższy niż odpowiedzi.

Zadanie jest rozwiązywane szybciej, jeśli większość tego, co zdający musi przeczytać, znajduje się w trzonie zadania, zaś opcje odpowiedzi są zwięzłe.

Ponieważ dłuższy trzon zadania i krótsze opcje odpowiedzi ułatwiają rozwiązywanie testu osobom, które do niego przystępują, dobre zadanie wielokrotnego wyboru powinno wyglądać następująco:

```
==================================
===============
===============
===============
===============
```

Nie należy używać przeczeń.

Niektórzy uczniowie czytają dokładniej lub przyswajają treść dokładniej niż inni, a słowo „nie" łatwo przeoczyć. Nawet jeśli to słowo zostało podkreślone tak, że nie sposób go przeoczyć, wiedza zwykle nie jest przyswajana w formie zbioru rzeczy nie będących faktami albo fałszywych twierdzeń — zwykle jest zbiorem wiadomości potwierdzonych, a nie zaprzeczeń.

Należy zadbać o to, by opcje były zgodne pod względem gramatycznym z trzonem z dania.

Na przykład, jeśli forma gramatyczna trzonu zadania wskazuje na to, że właściwa odpowiedź jest rodzaju żeńskiego lub w liczbie mnogiej, należy zadbać o to, by wszystkie opcje odpowiedzi były rodzaju żeńskiego lub w liczbie mnogiej.

Należy używać pełnych zdań jako trzonów zadań.

Jeśli trzon zadania to pełne zdanie zakończone znakiem zapytania albo pełne polecenie zakończone kropką, uczniowie mogą zacząć zastanawiać się nad odpowiedzią, zanim jeszcze spojrzą na opcje. Jeśli trzon zadania kończy się pustym znakiem, przecinkiem lub po prostu jest niekompletny, uczniowie muszą bardziej się wysilać. Takie utrudnienie zwiększa możliwość popełnienia błędu.

Zadawanie pytania na właściwym poziomie

Określenie właściwego poziomu pytania, które mamy zadać, to drugi poważny problem, z którym trzeba się uporać, tworząc test. Niektóre pytania są łatwe, oceniają tylko umiejętności do przypominania sobie informacji i wskazują stosunkowo niski poziom wiedzy. Inne pytania są trudniejsze — udzielenie odpowiedzi na nie wymaga połączenia posiadanej wiedzy lub zastosowania jej do nowego problemu lub sytuacji. Ponieważ różne poziomy pytań mierzą różne poziomy zrozumienia, aby przedsięwzięcie odniosło jakikolwiek skutek, właściwe pytanie musi zostać zadane na właściwym poziomie.

Bystry badacz problematyki nauczania Benjamin Bloom, piszący w latach pięćdziesiątych XX wieku, zasugerował sposób postrzegania pytań i poziomu zrozumienia niezbędnego do udzielenia prawidłowej odpowiedzi. Jego system klasyfikacji stał się znany jako **taksonomia Blooma**, system klasyfikujący cele nauczania w oparciu o poziom zrozumienia niezbędny do opanowania celu. Bloom i jego współpracownicy zaproponowali rozróżnienie sześciu etapów poznawczych w procesie uczenia się. Oto one, od najniższego do najwyższego:

1. **Wiedza**
   Umiejętność przywoływania słów, faktów i pojęć.
2. **Zrozumienie**
   Umiejętność zrozumienia tematu i wymiany informacji.
3. **Zastosowanie**
   Umiejętność zastosowania ogólnej wiedzy do rozwiązania nowego problemu.
4. **Analiza**
   Umiejętność rozłożenia koncepcji na części i zrozumienia ich wzajemnych relacji.
5. **Synteza**
   Umiejętność zbudowania z istniejącej wiedzy nowego wzoru lub koncepcji.
6. **Ewaluacja**
   Umiejętność dokonywania rzetelnych osądów co do wartości nowych koncepcji.

Dobór właściwego poziomu poznawczego. Rozważmy to zagadnienie na przykładzie testów tworzonych przez nauczycieli. Nauczyciele wybierają odpowiedni poziom poznawczy jako cel nauczania, a pomiar jakości jest tak zaprojektowany, by mierzył, czy te cele zostały zrealizowane. Większość zadań stworzonych przez nauczycieli oraz tych zamieszczonych w zeszytach ćwiczeń i podręcznikach jest na poziomie **wiedzy**. Badacze generalnie uważają, że nie jest to najlepsze, bo cele nauczania powinny być (i zwykle są) na wyższych poziomach poznawczych niż proste zapamiętywanie informacji.

Jednak gdy wprowadzany jest nowy materiał (niezależnie od wieku ucznia — od przedszkola do zaawansowanych szkoleń specjalistycznych), badanie powinno zawierać element sprawdzający, czy przyswojone zostały podstawowe nowe fakty. Gdy nauczyciele postanawiają dokonać pomiaru wykraczającego poza poziom wiedzy, to, jaki poziom zadań będzie odpowiedni, zależy od poziomu rozwoju uczniów. Poziom poznawczy uczniów, a szczególnie ich zdolność do myślenia i rozumienia abstrakcyjnego oraz ich umiejętność

wielostopniowego rozwiązywania problemów powinny określać najwłaściwszy poziom celu nauczania i tym samym najwłaściwszy poziom zadań testowych. Badacze uważają, że nauczyciele powinni testować to, czego uczą, na tym samym poziomie, na którym uczą.

Dlatego za każdym razem, gdy będziemy chcieli zmierzyć wiedzę ukrytą w czyjejś głowie, zastanówmy się nad poziomem zrozumienia, który chcemy zmierzyć. Czy wystarczy nam ocena zapamiętanej wiedzy? Jeśli tak, odpowiednim poziomem dla pytania będzie poziom **wiedzy**. Chcemy dowiedzieć się, czy osoba ubiegająca się o pracę potrafi używać swojej wiedzy do rozwiązywania problemów, z którymi nigdy się nie spotkała? Zadajemy pytanie na poziomie **zastosowania**, a będzie musiała zademonstrować tę umiejętność.

Projektowanie pytań na różnych poziomach poznawczych. Posługując się wskazówkami z tabeli 3.5, możemy tworzyć zadania lub pytania na każdym poziomie taksonomii Blooma.

Tabela 3.5. Pytania z różnych poziomów poznawczych

Poziom Blooma	Cechy pytania	Przykładowe pytanie lub zadanie
Wiedza	Wymaga jedynie zapamiętywania informacji i takich umiejętności jak przywoływanie informacji z pamięci, rozpoznawanie i powtarzanie.	Kto napisał powieść *Wielki Gatsby*? A. Faulkner B. Fitzgerald C. Hemingway D. Steinbeck
Zrozumienie	Wymaga takich umiejętności jak parafrazowanie, podsumowywanie i wyjaśnianie.	Co to jest chwytny ogon?
Zastosowanie	Wymaga takich umiejętności jak wykonywanie działań i rozwiązywanie problemów, zawiera takie polecenia jak **użyj**, **oblicz** czy **przygotuj**.	Jeśli rolnik posiada 40 hektarów gruntu i kupi jeszcze 16 hektarów, ile hektarów gruntu będzie posiadał?
Analiza	Wymaga takich umiejętności jak szkicowanie, słuchanie, logiczne myślenie i obserwacja, wykorzystuje takie polecenia jak **oznacz** czy **przeanalizuj**.	Narysuj mapę swojej okolicy i oznacz każdy dom.
Synteza	Wymaga takich umiejętności jak organizowanie i projektowanie, wykorzystuje takie polecenia jak **porównaj** czy **zestaw ze sobą**.	Wykorzystując swoją wiedzę o postaciach, opisz dalsze losy bohaterów *Kwiatów dla Algernona*.
Ewaluacja	Wymaga umiejętności krytycznego osądu i formułowania opinii, wykorzystuje takie polecenia jak **wyjaśnij** i **uzasadnij**.	Który aktor filmów muzycznych był najbardziej wysportowany? Odpowiedź uzasadnij.

Kiedy stosować taksonomię Blooma? Kategorie Blooma są uszeregowane hierarchicznie, przy czym **wiedza** to najniższy poziom poznania, a **ewaluacja** — najwyższy i najbardziej złożony. Każdy, kto pisze pytania mające sprawdzać wiedzę, może tworzyć zadania na

dowolnym poziomie. Nauczyciele mogą określać poziom wybranych celów nauczania i tworzyć zadania sprawdzające na tym właśnie poziomie. Jest stosunkowo łatwo uchwycić niższe poziomy taksonomii Blooma, trudniej jest dokonywać pomiarów na wyższych poziomach, ale nie jest to niemożliwe.

Nie powinniśmy się zanadto przejmować dokładnym rozróżnieniem pomiędzy sześcioma poziomami, tak jak je definiuje Bloom. Na przykład **zrozumienie** i **zastosowanie** są często traktowane jako jedno i to samo, bo właśnie zdolność zastosowania tego, czego oceniany się nauczył, wskazuje na poziom zrozumienia. W dzisiejszych czasach większość teoretyków i nauczycieli największą uwagę przykłada do rozróżnienia pomiędzy poziomem **wiedzy** a innymi poziomami. Większość nauczycieli, o ile tylko nie naucza początkowych etapów zupełnie nowych dziedzin, woli uczyć i sprawdzać cele nauczania na poziomie wyższym niż poziom **wiedzy**.

Zobacz również

- Razem z kilkoma kolegami napisaliśmy też nieco bardziej naukowe opracowanie tego tematu — zobacz: B.B. Frey, S.E. Petersen, L.M. Edwards, J.T. Pedrotti i V. Peyton, *Item-writing rules: Collective wisdom*, „Teaching and Teacher Education" 2005 r., nr 21, s. 357 – 364.
- Dobre omówienie zasad konstruowania zadań znajduje się w następującym opracowaniu: T.M. Haladyna, S.M. Downing i M.C. Rodriguez, *A review of multiple-choice item-writing guidelines for classroom assessment*, „Applied Measurement in Education", 2002 r., 15(3), s. 309 – 334.
- Wpływowe idee taksonomii Blooma zostały zaprezentowane w książce: B.S. Bloom (red.) *Taxonomy of educational objectives: The classification of educational goals. Handbook 1. Cognitive domain,* McKay, Nowy Jork 1956.
- B.S. Bloom, J.T. Hastings i G.F. Madaus, *Handbook on formative and summative evaluation of student learning*, McGraw-Hill, Nowy Jork, 1971.
- G.D. Phye, *Handbook of classroom assessment: Learning, adjustment, and achievement*, Academic Press, San Diego 1997.

SPOSÓB 29.

Sprawiedliwe testowanie

Nauczyciele regularnie tworzą własne testy, aby móc sprawdzać postępy w nauce swoich uczniów. Zwykle martwią się, czy ich testy nie są za trudne lub za łatwe i czy mierzą to, co mają mierzyć. Rozwiązaniem tego problemu są narzędzia do analizowania zadań.

Sprawdzanie wiedzy uczniów to chyba najczęstsza działalność we współczesnym nauczaniu. Nauczyciele cały czas przygotowują i sprawdzają testy, uczniowie cały czas uczą się i zdają testy, a cały proces ma na celu to, aby nauka była efektywniejsza. Testy nie mogę być za trudne (ani za łatwe) i muszą mierzyć to, co nauczyciel chce zmierzyć. Wyniki testów i oceny są najważniejszymi informacjami przekazywanymi rodzicom, uczniom i administratorom szkoły, dlatego też ocena z każdego testu musi być sprawiedliwa. Musi właściwie informować o poziomie wiedzy uczniów i powinna być skutkiem rzetelnego sprawdzenia ich umiejętności.

Zatroskani nauczyciele bezustannie pracują nad udoskonaleniem swoich testów, ale zwykle działają na ślepo, nie mając porządnych danych, na których mogliby się oprzeć. Co może zrobić inteligentny, zaangażowany nauczyciel, aby udoskonalić swoje testy lub zwiększyć efektywność oceniania? Nauczycielom, którzy chcą wypracować sobie sprawiedliwe metody sprawdzania i oceniania wiedzy, w sukurs przychodzi grupa metod statystycznych znana jako **analiza zadań**.

Analiza zadań

Analiza zadań to proces badania praktycznej przydatności poszczególnych zadań testowych. Nauczyciel może chcieć ocenić wyniki poszczególnych części testu, aby stwierdzić, które obszary są opanowane przez studentów, a które potrzebują dalszej pracy. Komercyjny twórca testów tworzący egzaminy dla szkoły pielęgniarek zapewne będzie chciał wiedzieć, które zadania w jego teście są trafne, a które zdają się mierzyć coś zupełnie innego i z tego powodu powinny zniknąć z testu.

W obu przypadkach twórcę testu interesuje poziom trudności zadań i to, czy są trafne. Choć w jednym przykładzie jest mowa o nauczycielu ze szkoły ponadgimnazjalnej przygotowującym testy dla uczniów, a w drugim o dużej firmie nastawionej na zysk, obaj twórcy testów są zainteresowani tym samym rodzajem danych i obaj mogą zastosować do analizowania zadań te same narzędzia.

Trzy rodzaje problemów z mierzeniem wiedzy

Każdy nauczyciel zatroskany o skuteczność wykorzystywanych metod sprawdzania wiedzy musi odpowiedzieć sobie na trzy rodzaje pytań. Na szczęście istnieją trzy narzędzia do analizowania zadań, które dostarczą trzech wymaganych rodzajów informacji.

Czy pytania są za trudne? Trudność poszczególnych zadań w teście może być stosunkowo łatwo określona za pomocą wzoru na **indeks trudności**. Możemy otrzymać indeks trudności dla zadania, obliczając odsetek studentów, którzy dane zadanie wykonują prawidłowo. Im większy odsetek, tym więcej jest osób przystępujących do testu, które posiadały informacje mierzone przez zadanie.

Określenie **„indeks trudności”** jest nieintuicyjne, bo tak naprawdę mierzy **łatwość** zadania, a nie jego **trudność**. Zadanie o wysokim indeksie trudności to łatwe zadanie, a nie trudne.

Jak znaleźć właściwy poziom trudności? O tym każdy musi zdecydować sam. Niektórzy nauczyciele uważają, że zadania o indeksie 0,5 lub niższym są za trudne, bo to oznacza, że większość uczniów ich nie wykona. Oczywiście możemy mieć wyższe standardy. Jeśli uważamy, że większość uczniów powinna opanować dany materiał, a indeks trudności dla zadania informuje nas, że znaczna część klasy nie była w stanie go wykonać, może to oznaczać, że jest ono za trudne.

Czy każde pytanie mierzy to, co ma mierzyć? Specjaliści od pomiarów twierdzą, że jeśli zadanie mierzy to, co powinno mierzyć, to jest trafne **[Sposób 32.]**. Podstawową miarą trafności zadania jest **indeks różnicujący**, mierzący też jego rzetelność. Indeks różnicujący mierzy stopień, w jakim za pomocą zadania można rozróżnić tych, którzy z całości testu otrzymali wysoką notę, od tych, którzy otrzymali niską ocenę.

Choć jego obliczenie składa się z kilku kroków, raz wyliczony wskaźnik może być interpretowany jako miara tego, do jakiego stopnia ogólna wiedza w danej dziedzinie lub opanowanie zbioru umiejętności przekłada się na umiejętność rozwiązania zadania.

Indeks różnicujący nie nazywa się tak, jak się nazywa dlatego, że wskazuje **obciążenie** testu. **Jest** to umiejętność rozróżnienia, czy osoba, która wypełniła zadanie prawidłowo, znajduje się w grupie tych, które osiągnęły wysoką notę, czy w grupie tych, które osiągnęły notę niską.

Dlaczego uczniowie nie wykonują prawidłowo zadania? Poza zbadaniem jakości całego zadania testowego, nauczyciele są często zainteresowani zbadaniem jakości poszczególnych dystraktorów (nieprawidłowych opcji odpowiedzi) w zadaniach wielokrotnego wyboru przez **analizę opcji odpowiedzi**. Obliczając odsetek uczniów, którzy wybierają poszczególne odpowiedzi, nauczyciele mogą zobaczyć, jakiego rodzaju błędy popełniają uczniowie. Czy opacznie pojęli pewne koncepcje? Czy jakieś elementy materiału są często błędnie rozumiane?

Aby poprawić skuteczność zadania z punktu widzenia pomiaru wiedzy, nauczyciele sprawdzają również, które dystraktory „działają" i wydają się atrakcyjne dla uczniów nie znających prawidłowej odpowiedzi, a które dystraktory zabierają tylko miejsce i są wybierane przez niewielu uczniów.

Dla wyeliminowania zgadywanek, owocujących przypadkowym udzielaniem prawidłowych odpowiedzi, nauczyciele i twórcy testów wprowadzają tyle prawdopodobnych dystraktorów, ile tylko można. Analizy udzielonych odpowiedzi pozwalają nauczycielom na dostrojenie i ulepszenie zadań, które chcą wykorzystać ponownie dla innych grup uczniów.

Przeprowadzanie analizy zadania i interpretowanie rezultatów

Oto procedury dla obliczeń związanych z analizowaniem zadań, z wykorzystaniem przykładowych danych. Wyobraźmy sobie klasę złożoną z 25 uczniów rozwiązujących test, w którym znajdowało się zadanie z tabeli 3.6 (należy przy tym pamiętać, że nawet twórcy testów standaryzowanych, do których podchodzą setki tysięcy ludzi, używają tych samych procedur).

Gwiazdka przy jednej z opcji w tabeli 3.6 oznacza, że odpowiedź B jest prawidłowa.

Tabela 3.6. Przykładowe zadanie do przeanalizowania

Odpowiedz na pytanie: Kto napisał powieść *Wielki Gatsby*?	**Liczba studentów, którzy wybrali poszczególne odpowiedzi**
A. Faulkner	4
B. Fitzgerald*	16
C. Hemingway	5
D. Steinbeck	0

Aby obliczyć indeks trudności:

1. Należy policzyć osoby, które udzieliły prawidłowej odpowiedzi.
2. Uzyskany wynik podzielić przez ogólną liczbę osób, które pisały test.

W zadaniu z tabeli 3.6 prawidłowej odpowiedzi udzieliło 16 z 25 osób:

$$16 : 25 = 0{,}64$$

Indeksy trudności wahają się od 0,00 do 1,00. Zadanie z naszego przykładu ma indeks trudności równy 0,64. Oznacza to, że prawidłową odpowiedź znało 64 procent studentów.

Jeśli nauczyciel uważa, że 64 procent to za mało, może podjąć kilka działań. Może postanowić zmienić sposób nauczania, aby lepiej zrealizować cel nauczania mierzony przez to zadanie. Kolejna interpretacja może być taka, że zadanie było zbyt trudne, mylące lub nietrafne, a w takim przypadku nauczyciel może zastąpić lub zmodyfikować zadanie, na przykład wykorzystując informacje z indeksu różnicującego albo analizując opcje odpowiedzi.

Aby obliczyć indeks różnicujący:

1. Należy podzielić test na podstawie wyników i stworzyć dwie grupy: **wysokie noty**, złożone z górnej połowy wyników, i **niskie noty**, z dolnej połowy.
2. Dla każdej z grup należy obliczyć indeks trudności zadania.
3. Należy odjąć indeks trudności grupy, która otrzymała niskie noty, od indeksu trudności grupy, która otrzymała noty wysokie.

Wyobraźmy sobie, że w naszym przykładzie 10 z 13 uczniów w grupie not wysokich i 6 z 12 uczniów w grupie not niskich odpowiedziało na pytanie prawidłowo. Dla grupy not wysokich indeks trudności wynosi 0,77 (10/13) zaś dla grupy not niskich 0,5 (6/12), możemy więc wyliczyć następujący indeks różnicujący:

$$0{,}77 - 0{,}50 = 0{,}27$$

Indeks różnicujący dla tego zadania wynosi 0,27. Indeksy różnicujące wahają się od –1,0 do 1,0. Im wyższa jest wartość dodatnia (im bliżej wskaźnikowi do 1,00), tym silniejsza jest relacja pomiędzy ogólnym wynikiem testu a odpowiedzią na to zadanie.

Jeśli indeks różnicujący jest ujemny, to oznacza, że z jakiegoś powodu uczniowie, którzy uzyskali w teście niski wynik, częściej odpowiadali na to pytanie prawidłowo. To dziwna sytuacja i sugeruje ona słabą trafność zadania albo to, że klucz odpowiedzi był niewłaściwy. Nauczycielom zwykle zależy na tym, aby każde zadanie w teście odwoływało się do tej samej wiedzy lub umiejętności, co reszta testu.

Wzór na obliczanie indeksu różnicującego jest tak stworzony, że jeśli prawidłowy wynik wybierze więcej uczniów z grupy wysokich not niż uczniów z grupy niskich not, liczba będzie dodatnia. Zatem nauczyciel powinien mieć nadzieję przynajmniej na to, że wynik będzie dodatni, bo to by wskazywało na fakt, że prawidłowe odpowiedzi zostały udzielone dzięki posiadanej wiedzy.

Możemy wykorzystać informacje z tabeli 3.6 do przeanalizowania popularności różnych opcji odpowiedzi, tak jak w tabeli 3.7.

Tabela 3.7. Analiza zadania „Kto napisał powieść Wielki Gatsby?"

Odpowiedź	Popularność odpowiedzi	Indeks trudności
A. Faulkner	4/25	0,16
B. Fitzgerald*	16/25	0,64
C. Hemingway	5/25	0,20
D. Steinbeck	0/25	0,00

Analiza opcji odpowiedzi wykazuje, że uczniowie, którzy nie odpowiedzieli prawidłowo, niemal w równej proporcji wskazywali na odpowiedź A i odpowiedź C. Żaden z uczniów nie wybrał odpowiedzi D, więc odpowiedź D nie posłużyła jako dystraktor. Uczniowie nie wybierają w tym zadaniu pomiędzy czterema opcjami — tak naprawdę wybierają jedynie pomiędzy trzema, bo nie biorą odpowiedzi D w ogóle pod uwagę.

To zwiększa szanse na odgadnięcie prawidłowej odpowiedzi i tym samym szkodzi trafności zadania. Nauczyciel może zinterpretować te dane jako dowód na to, że większość uczniów potrafi ze sobą powiązać Fitzgeralda i *Wielkiego Gatsby'ego* oraz na to, że uczniowie, którzy tego związku nie widzą, mają problemy z rozróżnieniem między Faulknerem a Hemingwayem.

Podwyższanie jakości testów

Aby podwyższyć jakość testów, można za pomocą analizy zadań wyłapać te, które są za trudne (albo za łatwe, jeśli o to obawia się nauczyciel), nie rozróżniają tych, którzy się przygotowali, od tych, którzy tego nie zrobili, albo mają nieodpowiednie dystraktory.

Jeśli występując w roli nauczyciela, mamy obawy o to, czy test jest rzetelny, możemy zmienić sposób nauczania, zmienić sposób testowania, albo też zmienić sposób oceniania testów:

Zmiana sposobu nauczania

Jeśli niektóre zadania są zbyt trudne, możemy zmienić sposób nauczania, np. przyłożyć większą wagę do nieopanowanego materiału albo zastosować inną strategię przekazywania wiedzy. Możemy zmodyfikować konkretne instrukcje, eliminując nieporozumienia co do obszaru obejmowanego przez zadanie.

Zmiana sposobu testowania

Jeśli zadania mają niskie lub ujemne wartości indeksu różnicującego, mogą być usunięte z aktualnego testu i możemy je również usunąć z puli zadań do kolejnych testów. Możemy też przyjrzeć się zadaniu, rozpoznać, co było w nim złego, i je zmienić. Gdy dystraktory okazują się niefunkcjonalne (żaden z uczniów ich nie wybiera), nauczyciele mogą zmodyfikować zadanie, wprowadzając nowy dystraktor. Celem trafnego i rzetelnego testu jest zmniejszenie szansy na to, że prawidłowa odpowiedź zostanie wybrana wskutek zgadywania. Im większa jest liczba wiarygodnych dystraktorów, tym bardziej udany, trafny i rzetelny będzie test.

Zmiana sposobu oceniania

Możemy wykorzystać informacje uzyskane w procesie analizy zadań do stwierdzenia, że materiał nie został prawidłowo przekazany i przez wzgląd na uczciwość usunąć zadanie z testu i przeliczyć wyniki. Najprostszym sposobem stosowanym przez nauczycieli jest podliczenie liczby **złych** zadań i dodanie tego wyniku do wyniku każdego ucznia. Z technicznego punktu widzenia, nie jest to to samo, co obliczenie wyników tak, jakby zadania nigdy nie było, ale dzięki temu uczniowie, którzy jednak rozwiązali trudne lub podchwytliwe zadanie, zostaną za to nagrodzeni, co większości nauczycieli wydaje się sprawiedliwsze.

Obawy, jakie nauczyciele mają co do jakości ich testów, nie różnią się zbytnio od pytań badawczych zadawanych przez naukowców. Tak samo jak naukowcy, nauczyciele mogą gromadzić dane od swoich uczniów, analizować te dane i interpretować rezultaty.

SPOSÓB 30.

Poprawianie swoich wyników bez żadnego wysiłku

Jeśli nie podoba nam się wynik, który uzyskaliśmy w ważnym teście, może powinniśmy podejść do niego ponownie. Powinniśmy?

Omawialiśmy już kwestię rzetelności badania **[Sposób 6.]**. **Rzetelność** to konsekwencja, z jaką test zwraca określony wynik. Innymi słowy, rzetelny test daje stabilny wynik, a nierzetelny test go nie daje. Ponieważ testy, które nie są idealnie rzetelne, dają wyniki, na które wpływ przynajmniej w części miał przypadek, wyniki te mogą się wahać w możliwy do przewidzenia sposób. Ponieważ nasz wynik przy powtórnym podejściu do testu będzie miał tendencję do zbliżania się do wyniku przeciętnego dla tego testu, efekt ten jest nazywany **regresją w kierunku średniej**.

Gdy podchodzimy do ważnych testów, takich jak SAT, ACT, GRE, LSAT lub MCAT, zwykle mamy możliwość powtórnego podejścia, by poprawić uzyskany wynik. Decyzja o tym, czy warto poświęcać czas, ciężką pracę i pieniądze na próbę poprawienia wyniku, powinna zostać podjęta przy pełnym zrozumieniu rzetelności testu i tego, jak może się zmienić wynik z uwagi na zjawisko regresji w kierunku średniej.

Regresja w kierunku średniej

Najpierw sprawmy, by regresja do średniej nastąpiła, aby było jasne, że wyniki mogą zmienić się w przewidywalnym kierunku wyłącznie z powodu właściwości krzywej rozkładu normalnego **[Sposób 23.]**. Zobaczyć znaczy uwierzyć i mam nadzieję, że uda mi się ten niewidzialny magiczny fenomen wywołać tu i teraz.

Aby to zrobić, należy poprosić setkę znajomych o wypełnienie testu typu **prawda-fałsz** takiego, jaki znajduje się w tabeli 3.8. No dobrze, powiedzmy, niech to będzie dziesięć osób, wliczając w to nas samych. Tysiąc osób byłoby nawet lepsze, ale wystarczy mi tylu, aby przekonać niedowiarków, że regresja rzeczywiście ma miejsce. Idąc dalej, pamiętajmy o tym, że gdyby w tym szalenie trudnym (lub też nadzwyczaj łatwym) teście wzięło udział 100 lub 1000 osób, wyniki byłyby jeszcze bardziej przekonujące.

Tabela 3.8. Test z zaawansowanej fizyki kwantowej

Pytanie	Zakreśl odpowiedź
1.	Prawda *lub* Fałsz
2.	Prawda *lub* Fałsz
3.	Prawda *lub* Fałsz
4.	Prawda *lub* Fałsz
5.	Prawda *lub* Fałsz
6.	Prawda *lub* Fałsz
7.	Prawda *lub* Fałsz
8.	Prawda *lub* Fałsz
9.	Prawda *lub* Fałsz
10.	Prawda *lub* Fałsz

Och, a jeśli chodzi o sam test, nie trzeba nawet widzieć pytań. Wyniki testu będą się zmieniać niezależnie od poddawanej pomiarowi konstrukcji **[Sposób 32.]**. Dlatego w tym teście można jedynie zgadywać. Ponieważ odpowiedzi na pytania są typu prawda-fałsz, będziemy mieli 50 procentową szansę na udzielenie prawidłowej odpowiedzi, a przeciętny wynik dla naszej grupy 10 osób poddanych testowi (albo 100, jeśli traktujemy tę próbę naprawdę poważnie… może chociaż 30? jacyś chętni?) powinien wynieść 5 z 10.

Poproś o napisanie „Testu z zaawansowanej fizyki kwantowej" wszystkie osoby, które uda się namówić. Odpowiadając na pytania, nie wolno oszukiwać, choć klucz do testu znajduje się zaledwie parę centymetrów poniżej (w tabeli 3.9)!

Tabela 3.9. Klucz odpowiedzi do testu z zaawansowanej fizyki kwantowej

1. Prawda	2. Prawda	3. Fałsz	4. Fałsz	5. Prawda
6. Fałsz	7. Fałsz	8. Prawda	9. Prawda	10. Fałsz

Zbierz wypełnione arkusze testów (dopilnuj, by znalazły się na nich nazwiska osób podchodzących do testu!) i podlicz punkty, korzystając z klucza w tabeli 3.9.

Następnie wybieramy osobę, która uzyskała najwięcej prawidłowych odpowiedzi (to będzie ktoś taki jak my, ktoś, kto w testach standaryzowanych, takich jak SAT, uzyskuje ponadprzeciętne wyniki), i osobę, która uzyskała ich najmniej (to będzie ktoś zupełnie różny od nas, kto uzyskuje wyniki niższe od przeciętnych). Prosimy te dwie osoby, aby powtórnie wypełniły test (wciąż nie pokazując im prawidłowych odpowiedzi) i powtórnie podliczamy punkty.

I tu zaczyna działać **regresja w kierunku średniej**. Jestem prawie przekonany (nie znając osób, które wzięły udział w teście, ani nie widząc ich wyników) o tym, że:

- osoba, która uzyskała najniższy wynik przy pierwszym podejściu, przy drugim uzyska wynik wyższy niż poprzednio;
- osoba, która uzyskała najwyższy wynik przy pierwszym podejściu, przy drugim uzyska wynik niższy niż poprzednio.

Jeśli tak się stało, to wypada tylko zapytać: „A nie mówiłem?". Jeśli jest inaczej, to przecież uprzedzałem, że jestem „prawie na pewno przekonany" o tym, że to zadziała. O wiele większa szansa jest na to, że zadziała przy większej liczebności próbki.

Dlaczego to działa?

Oczekiwaliśmy, że przy powtórzeniu testu wszystkie wyniki poniżej 5 (lub też poniżej średniej, jakąkolwiek by ona nie była) zwiększą się, a wyniki powyżej 5 — zmniejszą. To mogło się przydarzyć lub nie w przypadku naszych dwóch wyników, ale jest to najbardziej prawdopodobne.

Należy pamiętać, że był to test, w którym wiedza nie miała żadnego wpływu na wyniki. Za oboma razami wynik zależał tylko od przypadku. Jednak ten efekt występuje też w przypadku prawdziwych testów, gdzie wiedza ma wpływ na wynik. Dzieje się tak dlatego, że żaden prawdziwy test nie jest doskonale rzetelny i w każdym z nich pewną rolę odgrywa przypadek. Powyższa demonstracja tylko nasiliła ten efekt przez wykorzystanie testu, w którym przypadek odpowiadał za wyniki w stu procentach.

Dlaczego zatem wyniki mają tendencję do zmienienia się i za drugim razem zbliżają się do średniej? Na dłuższą metę, ze zbiorem wyników liczącym 100 lub 1000, moglibyśmy oczekiwać, że wyniki będą miały coś w rodzaju normalnego rozkładu. Tak samo jak w przypadku rzutu monetą (gdzie może wypaść orzeł lub reszka i każdy wynik ma 50 procent szans). W tabeli 3.10 znajdują się możliwe wyniki i szansa na to, że osoba przystępująca do „Testu z zaawansowanej fizyki kwantowej" je osiągnie.

Dlaczego skrajne wyniki miałyby się stawać mniej skrajne przy powtarzaniu testu? Porównajmy prawdopodobieństwo uzyskania dwóch skrajnych wyników (czyli na przykład wyniku 2 i kolejnego wyniku 2) z prawdopodobieństwem uzyskania wyniku 2 (prawdopodobieństwo = 0,44), a następnie wyniku 4 (prawdopodobieństwo = 0,205). Jest niemal pięciokrotnie wyższa szansa na to, że osoba, która za pierwszym razem otrzymała wynik 2, za drugim razem otrzyma wynik 4, niż że powtórnie otrzyma 2. Tak naprawdę istnieje niemal 95-procentowa pewność, że uzyska wynik wyższy od 2 (1–0,044–0,010–0,001 = 0,945).

Tabela 3.10. Prawdopodobny rozkład wyników testu

Wynik	Prawdopodobieństwo
0	0,001
1	0,010
2	0,044
3	0,117
4	0,205
5	0,246
6	0,205
7	0,177
8	0,044
9	0,010
10	0,001

Określenie „regresja w kierunku średniej" zawdzięcza swoją nazwę sławnemu Francisowi Galtonowi (dalekiemu kuzynowi Karola Darwina), który badał wpływ wzrostu rodziców na wzrost dzieci. Odkrył, że przeciętny wzrost dzieci był bliższy średniej wszystkich dzieci niż przeciętnej średniej ich rodziców. Galton nazwał to zjawisko „regresją w kierunku mierności" (Galton nie był znany z dyplomatycznego języka), my jesteśmy łagodniejsi. Nie ma to nic wspólnego z genetyką, natomiast jest to zjawisko (jakże by inaczej) statystyczne.

W przypadku tego testu, którego wyniki są całkowicie losowe, istnieje 65,6-procentowa szansa uzyskania wyniku równego lub bliskiego średniej (to połączone prawdopodobieństwo uzyskania 4, 5 lub 6 punktów). W przypadku większości testów, w których zadań jest więcej i wyniki podlegają prawidłom rozkładu normalnego, szansa na uzyskanie średniej lub zbliżonej do średniej liczby punktów wynosi 68 procent **[Sposób 23.]**.

Przewidywanie prawdopodobieństwa otrzymania wyższego wyniku

To wszystko bardzo interesujące, ale jak ma nam pomóc w podjęciu decyzji, czy warto po raz drugi podchodzić do testu? Podchodzenie do ważnych testów (takich, w których wyższa nota ma realne znaczenie) po raz drugi kosztuje pieniądze, czas, stres i zapewne przygotowanie, więc przy podejmowaniu decyzji o powtórnym podejściu do testu należy myśleć strategicznie.

Oczywiście możemy uzyskać wyższą notę z testu, zwiększając swój poziom wiedzy mierzonej przez ten test. Większą szansę mamy, jeśli będziemy się uczyć, robić testy próbne, uczęszczać na korepetycje i tak dalej. Jeśli jednak uzyskamy bardzo niski wynik, istnieje duże prawdopodobieństwo, że za drugim razem pójdzie nam lepiej, nawet jeśli pomiędzy testami nie będziemy nic robić, a to ze względu na zjawisko regresji w kierunku średniej. Można leniuchować w oczekiwaniu na drugi termin, a wynik najprawdopodobniej i tak będzie wyższy. To się nazywa szczęście!

Prawdopodobieństwo tego, że wypadniemy w teście lepiej tylko dlatego, że podchodzimy do niego ponownie, zależy od dwóch rzeczy: naszego wyniku za pierwszym razem i rzetelności testu.

Nasz wynik

Ponieważ wyniki mają tendencję do zbliżania się do średniej (losowo), szansa na to, że za drugim podejściem uzyskamy lepszy wynik, zależy od tego, czy za pierwszym razem nasz wynik znalazł się poniżej, czy powyżej średniej. Możemy sobie wyobrazić średnią jako wielki wir, ściągający do siebie wszystkie wyniki w całym rozkładzie. Wyniki poniżej średniej mają większą szansę na poprawę niż wyniki powyżej średniej.

Rzetelność testu

Statystycy zajmujący się mierzeniem wiedzy stosują pojęcie rzetelności, która — wyrażona w formie liczby — odpowiada proporcji zmienności wyniku, przy czym proporcja ta **nie jest** uzależniona od przypadku. Im wyższa rzetelność, tym w mniejszym stopniu o kształcie wyniku będzie decydował los. Wyniki rzetelne to wyniki stabilne, a siła wiru, jakim jest średnia, nie jest w stanie im sprostać.

Statystycy opracowali wzór, który możemy zastosować, by zorientować się, jak duże mamy pole manewru przy naszym wyniku. Jeśli miejsca na jego zwiększenie jest dużo, możemy wziąć pod uwagę drugie podejście. Użytecznym narzędziem, z którego tu skorzystamy, jest **standardowy błąd pomiaru**. Oto wzór na standardowy błąd pomiaru **[Sposób 6.]**:

$$\text{Błąd standardowy} = \text{Odchylenie standardowe}\sqrt{1 - \text{Rzetelność}}$$

Większość testów standaryzowanych jest zaopatrzona w informację na temat ich rzetelności i spodziewanym odchyleniu standardowym dla setek tysięcy wyników dawanych przez test przy każdorazowym jego przeprowadzaniu. Podstawiając te wartości do równania na standardowy błąd pomiaru, można się zorientować, na ile wyniki pomiędzy pierwszym a drugim podejściem do testu mogą się zmienić bez żadnych starań ze strony osoby testowanej.

Jednak nawet błąd standardowy jest w przypadku wartości skrajnych mylący. Wyniki bardzo niskie i bardzo wysokie częściej — z powodu czystego przypadku — przesuwają się dalej, niż by to sugerował błąd standardowy. Im bardziej oddalasz się od normy, tym trudniej pokonać jej siłę przyciągania. Wyniki skrajne nie potrafią oprzeć się sile wiru, chyba że są idealnie rzetelne.

A zatem, przed podjęciem decyzji o drugim podejściu do testu, warto wziąć pod uwagę następujące rady:

- Jeśli, relatywnie rzecz biorąc, uzyskaliśmy wynik bardzo wysoki, choć dla nas niezadowalający, podejście do testu po raz drugi najprawdopodobniej nie będzie warte starania.
- Jeśli uzyskaliśmy wynik bardzo niski (daleko poniżej średniej), jest niemal pewne, że za drugim razem wynik będzie wyższy. Spróbujmy ponownie. A może tym razem warto przysiąść też trochę nad książkami?

— Neil Salkind

SPOSÓB 31.

Ustalanie rzetelności

Ludzie, którzy wykorzystują, tworzą i zdają ważne testy, mają żywotny interes w ustaleniu tego, jak precyzyjne są ich wyniki. Na szczęście dziedzina pomiaru edukacyjnego i psychologicznego oferuje kilka metod, za pomocą których można zarówno sprawdzić, czy wynik testu jest konsekwentny i dokładny, jak i określić, na ile jest wiarygodny.

Każdy, kto wykorzystuje testy do podejmowania ważkich w skutkach decyzji, musi być przekonany, że otrzymane wyniki są dokładne i że nie mają na nie wpływu czynniki losowe, np. takie jak niedyspozycja zdrowotna ucznia zdającego egzamin. Twórcy testów muszą sprawić, by były one rzetelne, aby mogli przekonać swoich klientów, że mogą polegać na ich rezultatach.

Co jednak chyba najważniejsze, gdy podchodzimy do testu, którego wynik zadecyduje o tym, czy zostaniemy przyjęci na uczelnię albo czy otrzymamy awans na stanowisko głównego kipera, musimy wiedzieć, że test odzwierciedli nasze możliwości. Ten sposób przedstawia kilka procedur mierzenia rzetelności testów.

Dlaczego rzetelność jest ważna

Dlaczego powinniśmy szukać informacji na temat rzetelności ważnych testów, do których chcemy podejść? Testy i inne narzędzia pomiarowe powinny działać z konsekwencją, zarówno **wewnętrzną** (mierząc ten sam konstrukt, zachowujący się w podobny sposób), jak i **zewnętrzną** (dając podobne rezultaty przy kolejnych powtórzeniach testu). To kwestie **rzetelności**.

Rzetelność jest mierzona statystycznie i można uzyskać liczbę odpowiadającą poziomowi spójności testu. Większość wskaźników rzetelności opartych jest na korelacjach **[Sposób 11.]** pomiędzy odpowiedziami na zadania testowe albo pomiędzy dwoma zbiorami wyników testu przeprowadzanego dwa razy.

Do ustalania, czy test daje wyniki nieobciążone zanadto losową zmiennością, wykorzystuje się cztery rodzaje rzetelności:

Rzetelność wewnętrzna
Czy wyniki uzyskiwane przez każdą testowaną osobę są konsekwentne na przestrzeni poszczególnych zadań w danym teście?

Rzetelność powtórnego testowania
Czy wyniki uzyskiwane przez każdą testowaną osobę są konsekwentne na przestrzeni dwóch podejść do danego testu?

Porównywalność
Czy jeśli dwie różne osoby oceniają test, to ich oceny każdego testowanego są zbliżone?

Rzetelność wersji równoległych
Czy wyniki uzyskiwane przez każdą osobę podchodzącą do testu są konsekwentne na przestrzeni różnych wersji tego samego testu?

Obliczanie rzetelności

Jeśli stworzyliśmy test, który chcemy stosować (niezależnie od tego, czy do badania poziomu wiedzy uczniów, poziomu kwalifikacji kandydatów do pracy czy też stanu pacjentów), musimy określić, czy przeprowadzane za jego pomocą pomiary będą rzetelne. Metody wykorzystywane do obliczania poziomu precyzji testu zależą od tego, jaki rodzaj rzetelności nas interesuje.

Rzetelność wewnętrzna. Najczęściej podawaną miarą rzetelności jest miara wewnętrznej konsekwencji nazywana współczynnikiem alfa (albo współczynnikiem Cronbacha). **Współczynnik alfa** to liczba, która niemal zawsze mieści się w przedziale od 0,00 do 1,00. Im wyższa jest ta liczba, tym większa wewnętrzna konsekwencja charakteryzuje zadania w teście.

Gdybyśmy podzielili test na połowę — na przykład zadania nieparzyste umieścilibyśmy po jednej stronie, a nieparzyste po drugiej — moglibyśmy obliczyć korelację pomiędzy tymi dwiema połowami. Wzór na korelację tych połówek to wzór na współczynnik korelacji **[Sposób 11.]**. Jest to tradycyjna metoda ustalania rzetelności, aczkolwiek w dzisiejszych czasach uznawana za nieco staroświecką.

Z matematycznego punktu widzenia, wzór na współczynnik alfa daje średnią korelacji pomiędzy wszystkimi możliwymi połówkami testu i zastępuje korelację pomiędzy dwiema połowami w roli najczęściej używanej metody na stwierdzanie rzetelności wewnętrznej. Ze względu na stopień skomplikowania równania, do obliczania tej wartości zwykle wykorzystywane są komputery:

$$alfa = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} OS_i}{OS}\right)$$

Gdzie n to liczba zadań w teście, OS oznacza odchylenie standardowe testu (całkowitą wariancję skali), a $\sum_{i=1}^{n} OS_i$ to suma odchyleń standardowych każdego z n zadań testu (wariancja i-tego elementu sumy).

Rzetelność powtórnego testowania. Rzetelność wewnętrzna jest zwykle uznawana za odpowiedni dowód na rzetelność testu, ale w niektórych przypadkach niezbędne też jest zademonstrowanie konsekwencji na przestrzeni czasu.

Jeśli mierzona wielkość nie powinna się zmieniać z upływem czasu, albo powinno się zmieniać bardzo powoli, wyniki tej samej grupy powinny być z grubsza takie same, jeśli test zostanie powtórzony w innym terminie. Korelacja pomiędzy tymi dwoma zbiorami wyników będzie odzwierciedlać konsekwencję testu na przestrzeni czasu.

Porównywalność. Możemy też obliczyć rzetelność w przypadku, gdy więcej niż jedna osoba ocenia test lub dokonuje obserwacji. Gdy za wyniki odpowiadają różni oceniający,

należy zademonstrować to, że ich noty są spójne. Nawet jeśli ocenia tylko jedna osoba (na przykład nauczyciel), jeśli sposób oceniania jest subiektywny, jak w przypadku większości wypracowań i zadań domowych, tego typu rzetelność ma wielkie znaczenie teoretyczne.

Aby zademonstrować fakt, że w takich przypadkach otrzymany wynik odpowiada możliwościom osoby ocenianej, należy wykazać, że nie ma znaczenia, kto dokonuje oceny. Poziom porównywalności jest zwykle ustalany jako korelacje pomiędzy notami otrzymanymi przez serię osób albo odsetek wskazujący na to, jak często ocena była zgodna.

Rzetelność wersji równoległych. Wreszcie, możemy przekonać o rzetelności testu przez dowiedzenie, że niezależnie od tego, jaką wersję testu otrzyma dana osoba, uzyskany przez nią wynik będzie mniej więcej taki sam. Zademonstrowanie rzetelności wersji równoległych jest niezbędne tylko wtedy, gdy zadania, z których złożony jest test, dobierane są z większej puli zadań.

Na przykład w przypadku większości standaryzowanych testów (takich jak amerykańskie SAT i ACT) różne osoby podchodzące do testu dostają różne zadania obejmujące ten sam temat. Firmy odpowiedzialne za tworzenie tych testów stworzyły kilkaset pytań i budują różne wersje tego samego testu, wykorzystując różne próby tych pytań. Z tego powodu osoba, która w sobotę rano w stanie Maine podchodziła do testu, nie może zadzwonić do swojego kuzyna w Kalifornii i podać mu zadań, na które powinien się przygotować, bo kuzyn najprawdopodobniej dostanie inny zestaw zadań.

Gdy firmy tworzą różne wersje tego samego testu, muszą dowieść, że testy mają taką samą trudność i podobne właściwości statystyczne. Co najważniejsze, muszą wykazać, że osoba zdająca wersję testu z Maine uzyskałaby taki sam wynik, gdyby podeszła do testu w wersji z Kalifornii.

Interpretowanie dowodów rzetelności

Jest wiele metod ustalania rzetelności testów, a testy — w zależności od swego przeznaczenia — powinny być zaopatrzone w odpowiednie dowody rzetelności. Przy podejmowaniu decyzji, czy stworzony przez nas test wymaga udoskonalenia, możemy oprzeć się na wielkości współczynników rzetelności. Jeśli tylko podchodzimy do testu albo chcemy wykorzystać dostarczane przez test informacje, możemy wykorzystać wartość rzetelności do tego, by zdecydować, czy rezultaty testu są godne zaufania.

Rzetelność wewnętrzna
: Test zaprojektowany po to, by na jego podstawie można było podjąć ważną decyzję, powinien mieć bardzo wysoką rzetelność wewnętrzną, tak by otrzymany wynik bardzo dokładnie odzwierciedlał możliwości testowanego. Aby można było uznać, że test jest rzetelny wewnętrznie, zwykle wymaga się współczynnika alfa na poziomie 0,7 lub wyższego, choć jest to tylko ogólna zasada. W przypadku testów, do których podchodzimy albo które tworzymy, sami decydujemy o wymaganej wysokości współczynnika alfa.

Rzetelność powtórnego testowania
Test wykorzystywany do badania postępujących z czasem zmian, znajdujący zastosowanie w rozmaitych naukowych projektach badawczych, powinien charakteryzować się wysoką rzetelnością powtórnego testowania, aby różnice pomiędzy kolejnymi testami nie wynikały z czynnika losowego. Odpowiednia wielkość korelacji stabilności zależy od tego, jak bardzo w teorii konstrukt powinien zachowywać stabilność. Następnie, w zależności od swojej charakterystyki, test powinien dawać wyniki o korelacji w zakresie od 0,60 do 1,00.

Porównywalność
Porównywalność będzie nas interesowała tylko wtedy, gdy test jest oceniany subiektywnie, czyli na przykład gdy zadanie polega na napisaniu wypracowania. Obiektywne, oceniane przez komputer testy wielokrotnego wyboru powinny zapewniać idealną porównywalność, więc zwykle dla takich testów dane na temat porównywalności są pomijane. Jeśli podaje się korelacje porównywalności w celu określenia porównywalności testu, za minimalną dopuszczalną wartość można przyjąć 0,80.

Czasami rzetelność not wystawianych przez różnych oceniających jest podawana w formie odsetka przypadków, w których noty były zbieżne. W takiej sytuacji **odsetek zgodności** na poziomie 85 procent jest zwykle uznawany za wystarczający.

Rzetelność wersji równoległych
Jedynie testy mające różne wersje mogą być opisane jako posiadające rzetelność wersji równoległych. Profesorowie na uczelni raczej nie potrzebują ustalać rzetelności wersji równoległych, bo wszyscy ich studenci podchodzą do testów złożonych z takiego samego zestawu zadań, ale wielkie firmy zajmujące się tworzeniem testów muszą o to zadbać.

Rzetelność wersji równoległych powinna być bardzo wysoka, tak żeby ludzie mogli poważnie traktować wyniki testu, niezależnie od jego wersji. Korelacja pomiędzy dwoma wersjami testu powinna być wyższa niż 0,90. Firmy tworzące testy przeprowadzają badania, podczas których grupa osób podchodzi do obu wersji testu — po to, aby ocenić jego współczynnik rzetelności.

Zanim podejdziemy do ważnego testu, który może zadecydować o naszej przyszłości, powinniśmy upewnić się, że test ma uznawane poziomy rzetelności. Rodzaj rzetelności, który powinien być udowodniony i podany, zależy od przeznaczenia testu.

Poprawienie rzetelności testu

Najprostszą drogą do zapewnienia wysokiego współczynnika alfa albo jakiegokolwiek innego współczynnika rzetelności jest wydłużenie testu. Im więcej będzie zadań dotyczących tego samego pojęcia i im więcej okazji osoby podchodzące do testu będą miały, by wyjaśnić swoje podejście lub wykazać się wiedzą, tym bardziej rzetelna będzie ich łączna nota z testu. Teoretycznie to ma sens, ale też zwiększa rzetelność matematycznie, ze względu na kształt wzoru wykorzystywanego do obliczania rzetelności.

Spójrzmy na równanie na współczynnik alfa. W miarę zwiększania długości testu, zmienność sumarycznego wyniku zwiększa się szybciej niż zmienność dla zadań. We wzorze oznacza to, że wartość w nawiasie będzie tym większa, im dłuższy będzie test. Część ***n*/*n*–1** również się zwiększa wraz ze zwiększaniem się liczby zadań. Na skutek tego dłuższe testy zwykle cechuje wyższy poziom rzetelności.

Dlaczego to działa?

Korelacje porównują dwa zbiory poszeregowanych w pary wyników, tak że każda para wyników opisuje jedną osobę. Jeśli większość ludzi osiąga wyniki konsekwentnie (kolejne ich wyniki są wysokie, niskie lub średnie w porównaniu z innymi osobami, albo wysoki wynik w jednym teście zgadza się z niskim wynikiem w kolejnym), korelacja będzie bliska 1,00 lub –1,00.

Niekonsekwentne relacje pomiędzy wynikami dają korelację bliską zeru. Powtarzalność wyników albo korelacja testu z nim samym ma zgodnie z kryteriami ustalonymi przez klasyczną teorię testów **[Sposób 6.]** wskazywać na to, że wynik jest rzetelny. Klasyczna teoria testów wskazuje między innymi na to, że błąd losowy to jedyny powód, dla którego wyniki uzyskiwane przez daną osobę będą się od siebie różnić, jeśli ten sam test zostanie wielokrotnie powtórzony.

SPOSÓB 32.

Ustalanie trafności

Najważniejszą właściwością testu jest to, czy jest on przydatny w celu, w jakim został stworzony. Ustalenie trafności jest ważne, jeśli ktokolwiek ma zaufać temu, że wynik testu oznacza to, co ma oznaczać. Możemy przekonać siebie i innych, że nasz test jest trafny, jeśli przedstawimy pewne rodzaje dowodów.

Dobry test mierzy to, co w założeniu miał mierzyć. Na przykład ankieta mająca ustalić, jak często uczniowie szkół średnich zapinają pasy w samochodzie, powinna zawierać pytania dotyczące wykorzystania pasów bezpieczeństwa. Ankieta, w której nie znalazłyby się takie pytania, mogłaby zupełnie słusznie zostać zakwestionowana jako **nietrafna**. Trafność to zakres, w jakim coś mierzy to, co ma mierzyć. Ankiety, testy i eksperymenty muszą być trafne, by można je było uznać za dopuszczalne. Jeśli tworzymy test mający na celu badanie wiedzy lub osobowości albo jeśli chcemy upewnić się, że nasz test może być stosowany, powinniśmy zatroszczyć się o ustalenie jego trafności.

Trafność to nie jest coś, co test ma albo czego nie ma. Trafność to argument przedstawiany przez autora testu, osoby korzystające z jego wyników lub kogokolwiek, komu zależy na akceptacji testu bądź jego wyników.

Weźmy na przykład test poprawnej pisowni, w którym zadania będą wymagały rozwiązywania problemów matematycznych. Oczywiście test z zadaniami z zakresu matematyki nie jest trafnym testem ortograficznym. Choć jednak nie jest to trafny test mierzący poziom umiejętności ortograficznych, może być z powodzeniem trafnym testem mierzącym wiedzę z zakresu matematyki. Trafność testu lub ankiety nie leży w samym instrumencie, ale w interpretacji rezultatów.

Test może być trafny dla jednego zastosowania, a dla drugiego nie. Interpretowanie wyniku dyktanda napisanego przez dziecko jako wskaźnika jego wiedzy matematycznej nie ma sensu. Taki wynik może powiedzieć nam coś na temat umiejętności posługiwania się słowami, ale nie powie nam nic na temat swobody posługiwania się liczbami. Sam wynik nie jest trafny ani nietrafny. To znaczenie wiązane z wynikiem jest albo trafne, albo nietrafne.

Dla zilustrowania rozwiązania problemu ustalenia trafności wyobraźmy sobie, że stworzyliśmy nowy sposób sprawdzenia umiejętności poprawnego pisania. Chcemy sprzedać nasze testy szkołom w cały kraju, ale wpierw musimy przedstawić ewidentne dowody, że nasze testy mierzą umiejętność poprawnego pisania, a nie coś innego, takiego jak zasób słów, umiejętność czytania, poziom stresu egzaminacyjnego albo (jako inne czynniki mogące mieć wpływ na wynik) płeć czy rasę.

Strategie zwyciężania w sporze o trafność

Spór o trafność może wydawać się sporem, którego nie sposób rozstrzygnąć, ponieważ ze względu na niewidoczny wskaźnik jakości, trafności nie można jednoznacznie ustalić. Jednak jako twórcy testów chcemy przekonać osoby podchodzące do testów i każdego, kto będzie potem korzystał z ich rezultatów, że nasz test mierzy w odpowiednim zakresie to, co ma mierzyć. Na szczęście istnieje kilka powszechnie przyjętych sposobów na udowodnienie trafności testu.

Najpowszechniej przyjmowanym rodzajem dowodu trafności jest, co zaskakujące, najsłabszy argument za trafnością, jaki można przedstawić. To argument **trafności fasadowej**, który przedstawia się następująco: ten test jest trafny, ponieważ wygląda (z fasady) jakby mierzył to, co ma mierzyć. Osoby przedstawiające albo przyjmujące argument trafności fasadowej wierzą, że test, o którym mowa, składa się z zadań, których oczekujemy w takim teście. Na przykład ankieta na temat stosowania pasów bezpieczeństwa, o której mówiliśmy wcześniej, zostałaby uznana za trafną, gdyby znalazły się w niej pytania o zapinanie pasów.

Argument trafności fasadowej jest słaby, bo polega tylko na ludzkiej ocenie, ale może być przekonujący. Zdrowy rozsądek to mocny argument, może nawet najmocniejszy w przekonywaniu kogokolwiek do zaakceptowania dowolnego aspektu pomiaru. Choć trafność fasadowa wydaje się mieć niższą wartość naukową niż inne rodzaje dowodów trafności (i biorąc to dosłownie, rzeczywiście **ma** mniejszą wartość naukową), niewiele instrumentów testowych byłoby do zaakceptowania dla tych, którzy je tworzą i którzy z nich korzystają, gdyby nie miały trafności fasadowej. Jeśli my, jako twórcy lub użytkownicy testów, nie możemy dostarczyć dowodów na trafność omówionych w pozostałej części tego sposobu, to powinniśmy przynajmniej zaprezentować test mający choćby fasadową trafność.

Na użytek testu umiejętności poprawnego pisania, jeśli osoby do niego podchodzące będą musiały podawać poprawną pisownię słów, możemy uznać, że trafność pozorna została dowiedziona.

Osoby polegające na pomiarach generalnie akceptują cztery bardziej naukowe rodzaje dowodów trafności. Wszystkie one są częścią zakresu argumentów mogących przemawiać za trafnością.

Argumenty na trafność wewnętrzną (treściową)
Czy zadania w teście są reprezentatywne dla zadań, które mogłyby się w nim znaleźć? Jeśli test ma obejmować jakąś określoną dziedzinę wiedzy, to czy pytania są dobrą próbką z tej dziedziny?

Argumenty na trafność zewnętrzną (kryterialną)
Czy wyniki testu będą pozwalały na oszacowanie oczekiwanych wyników jakiegoś innego testu?

Argumenty na trafność teoretyczną
Czy wynik testu jest reprezentatywny dla cechy lub właściwości, którą chcemy zmierzyć?

Argumenty na trafność konsekwencyjną
Czy osoby, które podejdą do testu, skorzystają na tym doświadczeniu? Czy test jest obciążony na rzecz jakichś grup? Czy podejście do testu powoduje tak silny stres, że niezależnie od wyniku nie warto do niego podchodzić?

Argumenty na trafność treściową

Jeśli postanawiamy dokonać pomiaru jakiegoś pojęcia, mamy wiele aspektów tego pojęcia i możemy zadać w związku z nim wiele różnych pytań. Argumentem treściowym przemawiającym za trafnością naszego testu byłaby jakaś demonstracja, że zadania, które wybraliśmy do testu, są reprezentatywne dla wszystkich możliwych zadań.

Te wymagania wydają się trudne do spełnienia. Tradycyjnie tego rodzaju dowody były uważane za ważniejsze w przypadku testów osiągnięć. W dziedzinach osiągnięć (medycynie, prawie, języku, matematyce) istnieją dobrze określone dziedziny i obszary, z których powinien czerpać trafny test. Nauczyciel też najprawdopodobniej określa zbiór celów nauczania lub obszarów wiedzy, które test powinien mierzyć. Jednak na użytek testów osobowości, wiedzy czy nastawienia takie ściśle określone aspekty tematów rzadko są dostępne. W rezultacie przedstawienie wiarygodnego argumentu, że wybraliśmy pytania reprezentatywne dla jakiejś wyimaginowanej puli wszystkich możliwych pytań, jest trudne.

Co zatem jest niezbędne dla przedstawienia dowodów trafności treściowej w konstrukcji testu? Wydaje się, że konstrukcja testu wymaga przynajmniej jakiejś zorganizowanej metody budowy lub doboru pytań. Na przykład przy przeprowadzaniu pomiaru poczucia własnej wartości pytania mogą dotyczyć tego, jak osoba podchodząca do testu czuje się w różnych okolicznościach (na przykład w pracy, w domu lub w szkole), w różnych dziedzinach aktywności (w sporcie, w nauce lub podczas wykonywania pracy) i co myśli o różnych aspektach swojej osoby (na przykład o swoim wyglądzie, inteligencji lub umiejętności nawiązywania kontaktów).

Dobrą metodą dla nauczyciela mierzącego przyrost wiedzy uczniów w ciągu ostatnich tygodni jest **tabela specyfikacji** (uporządkowana lista obejmowanych przez materiał tematów oznaczonych pod względem ważności).

Decyzja co do tego, jak uporządkować koncepcję lub jak rozłożyć ją na składniki, należy do twórcy testu. Twórca może czerpać inspirację z badań lub innych testów, albo też może postępować zgodnie z planem dyktowanym przez zdrowy rozsądek. Kluczem jest przekonanie samego siebie, tak żebyśmy mogli przekonać innych, że obejmujemy testem kluczowe aspekty dziedziny, którą mierzymy.

Dla naszego testu poprawnego pisania, jeśli będziemy w stanie stwierdzić, że słowa, których poprawną pisownię mają podawać nasi uczniowie, są reprezentatywne dla większej puli słów, które nasi uczniowie powinni umieć poprawnie napisać, damy w ten sposób dowód trafności treściowej testu.

Argumenty na trafność zewnętrzną

Dowody trafności zewnętrznej testu wykazują, że odpowiedzi na zadania testowe prognozują wyniki w jakiejś innej sytuacji. Słowo „wyniki" może tu oznaczać sukces w pracy, wynik testu, oceny u innych i tak dalej.

Jeśli odpowiedzi udzielone w teście są związane z wynikami przez kryteria, które można zmierzyć natychmiast, dowód trafności jest nazywany dowodem na **trafność diagnostyczną**. Jeśli odpowiedzi udzielone w teście są związane z wynikami przez kryteria, które będzie można zmierzyć dopiero w przyszłości (na przykład ewentualne ukończenie studiów, sukces terapii lub ewentualne popadnięcie w nałóg), dowód trafności jest nazywany dowodem na **trafność prognostyczną**.

Jest oczywiste, że środki, które wybierzemy do podtrzymania trafności zewnętrznej, powinny być logiczne — kryteria powinny w jakiś sposób na poziomie teoretycznym pozostawać w relacji. Ta forma dowodu trafności jest najbardziej przekonująca i naprawdę ważna, gdy przeznaczeniem testu jest sprawdzenie lub prognozowanie wyników w jakiejś innej kwestii.

Dowody zewnętrzne są mniej przekonujące i niezbyt istotne dla testów, które nie mają niczego prognozować ani nie mają służyć do szacowania wyników w innej dziedzinie. Na przykład taki dowód wcale nie musi być użyteczny dla naszego testu poprawnej pisowni. Z drugiej strony, jest możliwe, że uda nam się udowodnić, iż osoby, które osiągnęły dobre wyniki w naszym teście, poradzą sobie dobrze podczas olimpiady językowej.

Argumenty na trafność teoretyczną

Trzecią kategorią dowodów trafności są dowody na trafność teoretyczną. **Konstrukt** to teoretyczna koncepcja lub cecha, którą test ma mierzyć. Wiemy, że takich konstruktów jak inteligencja czy wiara w swoje możliwości nie da się zmierzyć bezpośrednio. W pomiarach psychologicznych stosujemy metody pośrednie. Zadajemy serię pytań, co do których mamy nadzieję, że zmuszą odpowiadającego do wykorzystania tej części umysłu, którą badamy, lub odwołania się do tej części pamięci, która zawiera informacje na temat dawnych zachowań lub wiedzy, albo też przynajmniej zdołają nakłonić odpowiadającego do zastanowienia się nad swoją postawą i odczuciami na dany temat.

Dalej, mamy nadzieję, że osoby przystępujące do testu trafnie i uczciwie odpowiedzą na pytania. W praktyce rezultaty testów są zwykle traktowane jak bezpośrednia miara konstruktu, ale nie wolno nam zapominać, że tak naprawdę to tylko założenia. Powodzenie całego procesu zależy od kolejnego zbioru przypuszczeń: że prawidłowo zdefiniowaliśmy konstrukt, który próbujemy zmierzyć, i że nasze testy odzwierciedlają tę definicję.

Dowody na trafność teoretyczną zwykle zawierają zarówno argumenty na obronę samego zdefiniowanego konstruktu, jak i stwierdzenie, że wykorzystane instrumenty odpowiadają tej definicji. Dowody na trafność konstruktu mogą zawierać demonstrację tego, że odpowiedzi są takie, jakie powinny być zgodnie z teorią. Dowody na trafność konstruktu są gromadzone za każdym razem, gdy test (czy ankieta) jest przeprowadzany i — podobnie jak wszystkie argumenty trafności — nigdy nie będą całkowicie przekonujące. W pewnym sensie argumenty na trafność teoretyczną zawierają w sobie zarówno argumenty na trafność treściową, jak i zewnętrzną, bo wszelkie dowody na trafność polegają na powiązaniu ze sobą koncepcji i działania, które ma je mierzyć.

Dla naszego testu poprawnego pisania mogą istnieć badania natury **umiejętności poprawnego pisania**, przedstawiające ją jako czynność kognitywną lub cechę osobowości, lub inny dobrze zdefiniowany byt. Jeśli możemy zdefiniować, co rozumiemy pod pojęciem umiejętności poprawnego pisania, i przedstawić, że wyniki naszego testu zachowują się tak, jak wynika z definicji, możemy stwierdzić, że oto mamy dowód na trafność teoretyczną testu. Czy teoria sugeruje, że osoby, które lepiej czytają, popełniają mniej błędów? Jeśli wykażemy tę relację, może nawet popartą współczynnikiem korelacji **[Sposób 11.]**, to mamy dowód na trafność, który może przekonać innych.

Argumenty na trafność konsekwencyjną

Jeszcze dziesięć czy dwadzieścia lat temu ludzie zajmujący się pomiarami, ustalając ich trafność, starali się udowodnić tylko to, że wynik testu jest reprezentatywny dla konstruktu. Ze względu na rosnącą troskę o to, że pewne testy mogą być niesprawiedliwe wobec pewnych grup osób, a także z uwagi na obawy o społeczne konsekwencje powszechnego wykorzystania testów, teoretycy pomiaru naukowego i osoby odpowiedzialne za decyzje strategiczne przyglądają się dziś konsekwencjom, na jakie naraża się osoba podchodząca do testu.

Chodzi o to, że tak przyzwyczailiśmy się już do testowania i podejmowania ważnych decyzji w oparciu o wyniki testów, że czasami powinniśmy spojrzeć z boku i zapytać, czy społeczeństwo rzeczywiście korzysta na tym, iż podejmujemy te decyzje na podstawie testów. To odpowiada rozszerzeniu definicji trafności **z wyniku reprezentatywnego dla konstruktu na test spełniający zamierzoną rolę**. Założenie jest takie, że testy są po to, by ulepszyć nasz świat, nie pogorszyć, i dowody na trafność konsekwencyjną pomagają w przedstawieniu wartości, jaką mają testy dla społeczeństwa.

Tak jak agenci rządowi ze starych kawałów, testy „są tutaj po to, aby nam pomóc".

W przypadku naszego testu poprawnej pisowni najważniejsza negatywna konsekwencja, którą chcemy wykluczyć, to obciążenie. Jeśli w naszej teorii umiejętność poprawnego pisania jest niezależna od płci, rasy czy statusu socjoekonomicznego, wyniki testów powinny być takie same, niezależnie do której grupy należy podchodząca do testu osoba. Jeśli uzyskamy dowody na równość pomiędzy grupami, na przykład przy użyciu testu **t [Sposób 17.]**, to będziemy mieli już mocne argumenty na to, że nasz test jest sprawiedliwy i trafny.

Wybór z listy opcji trafności

Rozmaite kategorie opisanych tutaj dowodów trafności odpowiadają strategicznej liście opcji. Jeśli chcemy dowieść trafności, możemy wybierać z różnych rodzajów argumentów na trafność.

Oczywiście nie wszystkie testy muszą być zaopatrzone we wszystkie rodzaje dowodów trafności. Mały test z historii przygotowany przez nauczyciela dla grupy 25 uczniów może wymagać tylko argumentów na trafność treściową, aby nauczyciel mógł spokojnie zaufać jego wynikom. Argumenty na trafność zewnętrzną będą niepotrzebne, bo zadaniem tego testu nie jest prognozowanie wyników uczniów w innym teście.

Z drugiej strony, testy, od których wiele zależy, np. testy rekrutacyjne na uczelnie wyższe (w Stanach Zjednoczonych to testy ACT, SAT i GRE) oraz testy na inteligencję, mające wyłonić uczniów spełniających warunki niezbędne do otrzymania stypendiów, powinny być poparte argumentami ze wszystkich czterech obszarów trafności. Dla naszego testu poprawnej pisowni możemy sami zadecydować, jaki rodzaj dowodów i jaki rodzaj argumentów będzie najbardziej przekonujący.

SPOSÓB 33.

Prognozowanie żywotności

Wielu z nas instynktownie wierzy w to, że rzeczy trwające już długo, najprawdopodobniej przetrwają jeszcze dłużej, zaś rzeczy, które znamy krótko, wręcz przeciwnie. Formalizacja tego przypuszczenia znana jest jako zasada Gotta, a związane z nią obliczenia nie są trudne.

Fizyk J. Richard Gott III trafnie przewidział upadek muru berlińskiego i wyliczył czas grania 44 spektakli na Broadwayu[1]. Przedstawił też kontrowersyjną prognozę, w myśl której rasa ludzka przetrwa jeszcze od 5100 do 7,8 miliona lat, ale nie dłużej. Twierdzi, że to dobry powód dla tworzenia samowystarczalnych kolonii w kosmosie — jeśli rasa ludzka przełoży część jajek do innego koszyka, będzie mogła zapewnić dłuższe przetrwanie swego gatunku na wypadek uderzenia asteroidy albo wojny nuklearnej na macierzystej planecie[2].

Gott wierzy, że jego proste obliczenia mogą zostać rozciągnięte na niemal wszystko, o ile są zachowane określone warunki. Aby za pomocą tych obliczeń przewidzieć, jak długo coś przetrwa, wystarczy wiedzieć, jak długo **dotąd istniało**.

[1] Timothy Ferris, *How to Predict Everything*, „The New Yorker", 12 lipca 1999 r.

[2] J. Richard Gott III, *Implications of the Copernican Principle for Our Future Prospects*, „Nature", nr 363, 27 maja 1993 r.

W praktyce

Gott swoje obliczenia oparł na czymś, co nazwał zasadą kopernikańską (a co niektórzy ludzie w tym konkretnym zastosowaniu nazywają zasadą Gotta). Zasada mówi, że gdy wybieramy moment na obliczenie żywotności zjawiska, ten moment jest najprawdopodobniej zwyczajny, nie wyjątkowy lub uprzywilejowany, tak samo jak Kopernik powiedział nam, że Ziemia nie zajmuje uprzywilejowanej roli we wszechświecie.

Ważnym jest, by dobierać obiekty w zwykłym momencie ich życia. Obciążanie testu przez dobieranie obiektów z dużym prawdopodobieństwem znajdujących się na początku lub pod koniec okresu życia (takich jak na przykład niemowlęta na oddziale noworodkowym szpitala albo pensjonariusze domu starców) da marne rezultaty. Co więcej, zasada Gotta jest mniej użyteczna tam, gdzie już istnieją dane aktuarialne. Ponieważ mamy mnóstwo danych aktuarialnych na temat długości ludzkiego życia, zasada Gotta gorzej się sprawdza przy jej obliczaniu.

Gdy już wybraliśmy moment, pora go zbadać. Jeśli nie występują inne czynniki, mamy 50-procentową szansę na to, że moment znajduje się gdzieś w środkowych 50 procentach okresu życia zjawiska, 60-procentową szansę, że w środkowych 60 procentach, 95-procentową szansę, że w środkowych 95 procentach i tak dalej. Dlatego jest tylko 25 procent szans na to, że wybraliśmy moment w pierwszej ćwiartce okresu życia zjawiska, 20 procent, że w pierwszej piątej części, 2,5 procent, że w ostatnich 2,5 procent okresu życia i tak dalej.

W tabeli 3.11 znajdują się równania dla poziomów ufności: 50 procent, 60 procent i 95 procent. Zmienna $t_{przeszłość}$ odpowiada temu, jak długo obiekt istniał, a $t_{przyszłość}$ odpowiada temu, jak długo ma jeszcze trwać.

Tabela 3.11. Poziomy ufności dla zasady Gotta

Poziom ufności	**Minimalna $t_{przyszłość}$**	**Maksymalna $t_{przyszłość}$**
50 procent	$t_{przeszłość}/3$	$3t_{przeszłość}$
60 procent	$t_{przeszłość}/4$	$4t_{przeszłość}$
95 procent	$t_{przeszłość}/39$	$39t_{przeszłość}$

Spójrzmy na prosty przykład. Odpowiadamy szybko: czyja muzyka ma większe szanse wciąż być słuchana za 50 lat — Jana Sebastiana Bacha czy Britney Spears? Pierwsze dzieło Bacha zostało wykonane około 1705 roku, czyli około 300 lat temu. Pierwszy album Britney Spears został wydany w styczniu 1999 r., około 6,5 roku albo 79 miesięcy przed napisaniem tej książki.

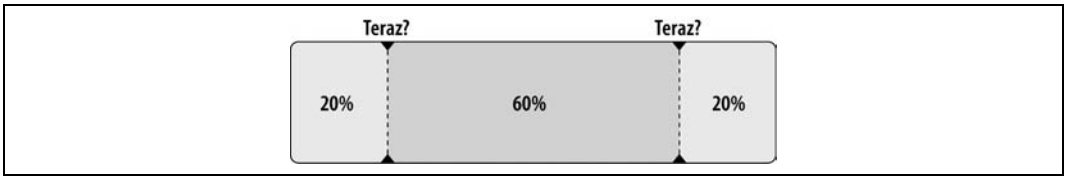
W tabeli 3.11, odczytując wartości dla poziomu ufności 60 procent, widzimy że minimalna $t_{przyszłość}$ to $t_{przeszłość}/4$, a maksymalna to $4t_{przeszłość}$. Ponieważ $t_{przeszłość}$ dla muzyki Britney wynosi 79 miesięcy, mamy 60-procentową szansę na to, że muzyka Britney będzie słuchana jeszcze od 79/4 miesięcy do 79×4 miesięcy. Innymi słowy, mamy 60-procentową pewność, że Britney będzie częścią naszej kultury jeszcze przynajmniej przez 19,75 miesięcy (1,6 lat), a najwyżej przez 316 miesięcy (26,3 lata).

Sześćdziesiąt procent to dobry współczynnik ufności dla szybkiej oceny — nie dość, że szansa jest większa niż pół na pół, to czynniki ¼ i 4 są łatwe w użyciu.

Na tej samej zasadzie możemy z 60-procentową pewnością prognozować, że ludzie będą słuchali muzyki Bacha jeszcze od 300/4 lat do 300×4 lat, czyli jeszcze przez 75 – 1200 lat. W ten sposób możemy przewidzieć, że muzyka Britney umrze wraz z jej fanami, a muzyka Bacha będzie słuchana jeszcze w 4-tym tysiącleciu.

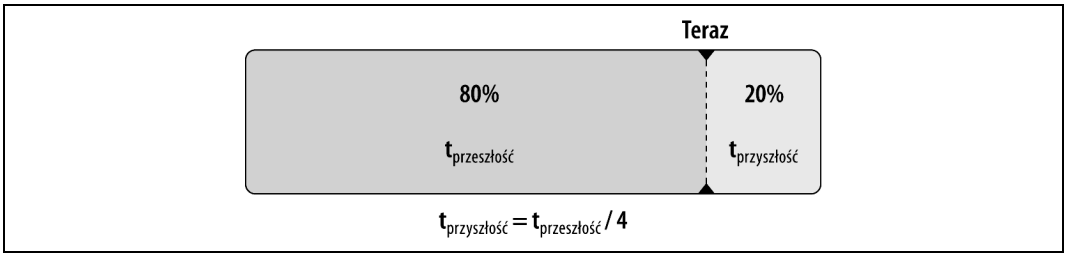
Jak to działa?

Przypuśćmy, że badamy żywotność pewnego obiektu, który nazwiemy sobie **celem**. Jak się już przekonaliśmy, mamy 60 procent szans na to, że znajdujemy się gdzieś w środkowych 60 procentach długości życia obiektu (rysunek 3.4).[3]



Rysunek 3.4. Środkowe 60 procent długości życia

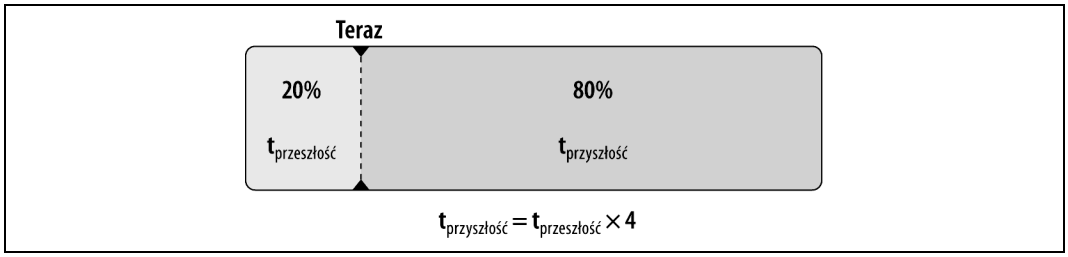
Jeśli znajdujemy się na samym końcu tych środkowych 60 procent, będziemy w drugim punkcie oznaczonym jako „teraz?" na rysunku 3.4. W tym momencie pozostało tylko 30 procent żywotności celu (rysunek 3.5), co oznacza, że $t_{przyszłość}$ jest równa jednej czwartej $t_{przeszłość}$ (80 procent). Jest to minimalna pozostała długość życia, jakiej z 60-procentową pewnością możemy oczekiwać.



Rysunek 3.5. Minimalny pozostały okres życia (poziom ufności 60 procent)

Podobnie, jeśli znajdujemy się na początku środkowych 60 procent (pierwszy punkt oznaczony jako „teraz?" na rysunku 3.4), 80 procent okresu istnienia celu należy do przyszłości, jak to przedstawiono na rysunku 3.6. A zatem $t_{przyszłość}$ (80 procent) jest równe $4 \times t_{przeszłość}$ (20 procent). Jest to maksymalna pozostała długość życia, jakiej z danym poziomem ufności możemy oczekiwać.

[3] J. Richard Gott III, „A Grim Reckoning" *http://pthbb.org/manual/services/grim* (strona w języku angielskim — *przyp. tłum.*).



Rysunek 3.6. Maksymalny pozostały okres życia (poziom ufności 60 procent)

Ponieważ szansa na to, że trafimy pomiędzy te dwa punkty wynosi 60 procent, możemy z 60-procentową ufnością obliczyć, że okres dalszego trwania celu ($t_{przyszłość}$) znajdzie się pomiędzy $t_{przeszłość}/4$ a $4 \times t_{przeszłość}$.

W codziennym życiu

Przypuśćmy, że zamierzamy zainwestować w pewną firmę i żeby określić, czy inwestycja będzie udana, chcielibyśmy oszacować, jak długo firma będzie jeszcze funkcjonować. Możemy w tym celu wykorzystać zasadę Gotta. Choć akcje tej firmy nie znajdują się na giełdzie, weźmy jako przykład wydawnictwo O'Reilly Media.

Oczywiście nie wybrałem wydawnictwa O'Reilly losowo, a na temat żywotności firm jest dostępnych mnóstwo informacji, ale mimo to spróbujmy wykorzystać zasadę Gotta do zgrubnego oszacowania żywotności wydawnictwa O'Reilly. W końcu istnieją też obszerne dane na temat żywotności spektakli na Broadwayu, ale to nie powstrzymało Gotta przed analizowaniem tej żywotności — a teraz, gdy O'Reilly opublikował moją książkę, waham się przed stwierdzeniem, że wydawnictwo przetrwa wieki.

W Wikipedii znajdziemy informację, że O'Reilly rozpoczęła działalność w roku 1978, jako firma consultingowa zajmująca się problematyką techniczną. W lipcu 2005 roku, gdy piszę te słowa, firma O'Reilly ma za sobą około 27 lat działania. Jak długo O'Reilly będzie jeszcze istniała według naszych oczekiwań?

Oto prawdopodobna żywotność firmy O'Reilly Media, obliczona z 50-procentowym poziomem ufności:

Minimalna:
27/3 = 9 lat (do lipca 2014)

Maksymalna:
27×3 = 81 lat (do lipca 2086)

Oto nasze prognozy dla poziomu ufności 60 procent:

Minimalna:
27/4 = 6 lat i 9 miesięcy (do kwietnia 2012)

Maksymalna:
27×4 = 108 lat (do lipca 2113)

Wreszcie, oto nasze prognozy dla poziomu ufności 95 procent:

Minimalna:
27/39 = 0,69 lat = około 8 miesięcy i 1 tygodnia (do połowy marca 2006)

Maksymalna:
27×39 = 1053 lat (do lipca 3058)

W gospodarce, która przeżyła krach dot-comów, te liczby wyglądają całkiem nieźle. Na przykład Apple Computers nie wypada wiele lepiej, a Microsoft powstał w 1975 roku, więc można o nim powiedzieć to samo. Prawdziwy inwestor wziąłby pod uwagę wiele innych czynników, takich jak roczny obrót i cena akcji, ale na pierwszy rzut oka wygląda na to, że O'Reilly Media ma taką samą szansę na przeżycie hipotetycznego inwestora jak na bankructwo w następnym dziesięcioleciu.

— Ron Hale-Evans

SPOSÓB 34.

Podejmujemy rozsądne decyzje dotyczące naszego zdrowia

Testy medyczne (badania) dostarczają informacji diagnostycznych, które zwykle rozumiane są opacznie przez pacjentów, a czasami nawet przez lekarzy. Zrozumienie cech probabilistycznych takich jak „czułość" i „specyficzność" może sprawić, że ujrzymy obraz wyraźniejszy i (czasami) bardziej pocieszający.

Jako konsumenci informacji medycznych musimy podejmować decyzję odnośnie dalszego postępowania, leczenia, zasięgania drugiej opinii i tak dalej. Najprawdopodobniej przy podejmowaniu tych decyzji będziemy polegali na informacjach medycznych takich jak artykuły z gazet, porady naszego lekarza i wyniki badań. Jednak większość informacji medycznych otrzymywanych od lekarza jest obarczonych błędem o znanej wielkości. Jest to prawdziwe zwłaszcza w odniesieniu do wyników badań, które wskazują na możliwość występowania u nas określonego schorzenia.

Sposób ten omawia wykorzystanie informacji na temat właściwości testów medycznych do uzyskania wyraźniejszego obrazu rzeczywistości i (miejmy nadzieję) podejmowania lepszych decyzji na temat leczenia.

Statystyki i badania

Aby mądrze korzystać z informacji znajdujących się w wynikach badań, musimy wpierw dowiedzieć się, co w przypadku tych testów znaczy pojęcie **dokładności**. W tabeli 3.12 są przedstawione cztery możliwe pod względem dokładności wyniki testu medycznego (badania).

Tabela 3.12. Możliwe wyniki badania

	Pacjent naprawdę ma daną przypadłość (A)	**Pacjent naprawdę nie ma danej przypadłości (B)**
Wynik testu wskazuje na to, że pacjent ma daną przypadłość	Prawdziwy pozytywny (wynik jest prawidłowy)	Fałszywy pozytywny (wynik jest nieprawidłowy)
Wynik testu wskazuje na to, że pacjent nie ma danej przypadłości	Fałszywy negatywny (wynik jest nieprawidłowy)	Prawdziwy negatywny (wynik jest prawidłowy)

Rzetelność **[Sposób 6.]** badań medycznych jest sumą dwóch proporcji znanych jako **czułość** i **specyficzność**. Generalnie, osoby posługujące się wynikami takich badań interesują trzy kwestie związane z dokładnością:

- Jeśli osoba jest chora, jakie są szanse na to, że wynik testu będzie pozytywny? To prawdopodobieństwo to **czułość**. Jaki odsetek osób w kolumnie A otrzyma pozytywny wynik testu?
- Jeśli osoba **nie** jest chora, jakie są szanse na to, że wynik testu będzie negatywny? To prawdopodobieństwo to **specyficzność**. Jaki odsetek osób w kolumnie B otrzyma negatywny wynik testu?
- Jeśli osoba otrzyma pozytywny wynik testu, jakie są szanse na to, że jest chora? Z perspektywy pacjenta jest to pytanie najważniejsze i można traktować je jako podstawową kwestię trafności dla takich testów. Pani doktor, czy mogę polegać na wynikach tego badania, czy też mogą one być mylące?

Należy zauważyć, że w tabeli 3.12, w kolumnach A i B znajdują się różne osoby. Ludzie mający daną chorobę są w kolumnie A, zaś ludzie nie mający danej choroby — w kolumnie B. Osoba z kolumny A nie może uzyskać w teście wyniku fałszywego pozytywnego, bo wynik pozytywny będzie prawdziwy. Natomiast osoba z kolumny B nie może uzyskać wyniku fałszywego negatywnego, bo negatywny wynik będzie prawdziwy.

To, w której kolumnie kto się znajdzie, zależy od naturalnego rozkładu choroby. Prawdopodobieństwo tego, że osoba znajdzie się w kolumnie A (czyli prawdopodobieństwo tego, że ma daną chorobę), zależy od **bazowego wskaźnika** rozpowszechnienia. Jeśli na daną chorobę cierpi 5 procent populacji, to te 5 procent znalazłoby się w kolumnie A.

Zrozumienie badań przesiewowych raka piersi

Rak piersi to przykład bardzo poważnej choroby wykrywanej za pomocą badań przesiewowych. Badanie na obecność raka piersi zaczyna się od badania mammograficznego. Pozytywny wynik tego badania prowadzi do dalszych badań — kolejnej mammografii, badania ultrasonograficznego lub biopsji.

Na pierwszym miejscy interesują nas odpowiedzi na pytania o czułość i specyficzność badań przesiewowych raka piersi. Uzbrojeni w te informacje i wiedzę na temat bazowego wskaźnika rozpowszechnienia raka, będziemy mogli odpowiedzieć na najważniejsze pytanie:

> Jeśli kobieta uzyska pozytywny wynik badania, to jakie jest prawdopodobieństwo, że ma raka piersi?

Zasięgając informacji u lekarza lub sięgając do źródeł, możemy dowiedzieć się, że czułość badania mammograficznego wynosi około 90 procent. Specyficzność około 92 procent.

Dokładna czułość i specyficzność badań przesiewowych raka piersi zmienia się z czasem, w zależności od badanych populacji. Młodsze kobiety częściej są badane mammografem niż dawniej i dla młodszych kobiet test ten ma niższą czułość i specyficzność. Oczywiście informacji na temat aktualnej dokładności badania należy zasięgnąć u lekarza lub innego specjalisty.

Odpowiednie liczby umieściliśmy w tabeli 3.13, skonstruowanej tak samo jak tabela 3.12. Ponieważ suma podmiotów w kolumnach A i B musi wynosić 100 procent w każdej, możemy też ocenić liczbę fałszywych wyników negatywnych i fałszywych pozytywnych.

Tabela 3.13. Teoretyczne wyniki badania mammograficznego 10 000 kobiet

	Pacjentka ma raka piersi (A) L = 120	**Pacjentka nie ma raka piersi (B) L = 9 880**
Mammogram wskazuje na obecność raka	Czułość	Fałszywe pozytywne
	90 procent	8 procent
	L = 108	L = 790
Mammogram nie wskazuje na obecność raka	Fałszywe negatywne	Specyficzność
	10 procent	92 procent
	L = 12	L = 9090

W tabeli 3.13 znajdują się wyniki dla hipotetycznych 10 000 kobiet, w oparciu o wskaźnik rozpowszechnienia raka piersi w populacji wynoszący około 1,2 procent.

Okazuje się, że prawidłowe określenie częstości występowania raka piersi jest trudne ze względu na różnice w definiowaniu populacji, której badania dotyczą, oraz ograniczenia badań wykrywających raka. Wykorzystuję tu często powtarzaną i powszechnie przyjmowaną liczbę kobiet pomiędzy 40 a 84 rokiem życia cierpiących na raka piersi.

Zanim zinterpretujemy wyniki badania, wróćmy do trzeciego pytania z listy ważnych pytań, które powinniśmy zadać. Jeśli osoba otrzyma pozytywny wynik testu na obecność choroby, to jakie jest prawdopodobieństwo, że rzeczywiście jest chora? Z 10 000 kobiet poddanych przesiewowemu badaniu na obecność raka piersi 898 otrzyma wynik pozytywny. Dla 790 spośród nich ten wynik nie jest prawdziwy — tak naprawdę nie mają raka. Dla 108 kobiet wynik jest prawdziwy — są chore na raka piersi. Innymi słowy, jeśli dana osoba otrzyma pozytywny wynik testu, istnieje tylko 12-procentowe prawdopodobieństwo

tego, że naprawdę jest chora. Najczęstszy rezultat dodatkowego badania pacjentki, która otrzymała pozytywny wynik badania mammograficznego, jest taki, że tak naprawdę nie ma raka.

Co z dokładnością wyniku negatywnego? Z 9102 kobiet, które otrzymają podczas badania wynik negatywny, 12 naprawdę jest chorych. To stosunkowo niewielka wartość, 1/10 z 1 procenta, ale badanie przepuszcza te osoby całkowicie i nie mają one szans na właściwe leczenie.

Dlaczego to działa?

Dokładność badań przesiewowych polega na konkretnym zastosowaniu ogólnego podejścia do prawdopodobieństwa warunkowego przypisywanego Thomasowi Bayesowi, osiemnastowiecznemu filozofowi i matematykowi. Pytanie o prawdopodobieństwo warunkowe brzmi „Jeśli coś się zdarzyło, jakie są szanse, że zdarzy się…".

Podejście Bayesa do prawdopodobieństwa warunkowego[4] polegało na przyjrzeniu się naturalnie występującej częstości zdarzeń. Podstawowy wzór na prawdopodobieństwo tego, że dana osoba jest chora, jeśli test dał wynik pozytywny, wygląda następująco:

$$\frac{\text{Prawdziwe pozytywne}}{\text{Prawdziwe pozytywne} + \text{Fałszywe pozytywne}}$$

Wyrażony w prawdopodobieństwach warunkowych wzór przedstawia się tak:

$$\frac{\text{Wskaźnik rozpowszechnienia} \times \text{Czułość}}{(\text{Wskaźnik rozpowszechnienia} \times \text{Czułość}) + (1 - \text{Wskaźnik rozpowszechnienia})(1 - \text{Specyficzność})}$$

Aby odpowiedzieć na najważniejsze pytanie w naszym przykładzie z rakiem piersi („Jeśli kobieta uzyska pozytywny wynik badania, jakie jest prawdopodobieństwo na to, że ma raka piersi?"), należy wykonać następujace działania w celu wyznaczenia wartości liczbowej:

$$\frac{0{,}012 \times 0{,}90}{(0{,}012 \times 0{,}90) + (1 - 0{,}012)(1 - 0{,}92)} = 0{,}1202$$

Podejmowanie świadomych decyzji

Testy medyczne są wykorzystywane do stwierdzenia, czy pacjent może być chory lub zagrożony chorobą. Stwierdzenie lub wykluczenie obecności takiej choroby jak rak zwykle wymaga przynajmniej dwustopniowego postępowania. Stopień pierwszy polega na zbadaniu

[4] Prawdopodobieństwem warunkowym zajścia zdarzenia A pod warunkiem zajścia zdarzenia B, gdzie $P(B) > 0$ nazywamy liczbę $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. Jest to iloraz prawdopodobieństwa części wspólnej zdarzeń A, B i prawdopodobieństwa zdarzenia B — *przyp. red*.

pacjenta metodą przesiewową, który to test najczęściej jest stosunkowo prosty i nieinwazyjny, szukający u pacjenta objawów określonego schorzenia. Jeśli rezultat jest pozytywny, drugi stopień polega na przeprowadzeniu drugiego badania (lub serii badań), które zwykle są bardziej złożone, inwazyjne i kosztowne, ale też o wiele dokładniejsze, mogące potwierdzić lub wykluczyć wcześniejszą diagnozę.

Badania medyczne nie są idealnie trafne i rzetelne. Wyniki testów mogą być nieprawidłowe. Każdy, kto poddaje się badaniu, może znaleźć się w jednej z czterech grup. Może być chory, co stwierdzi badanie, może też nie być chory i badanie nie stwierdzi obecności choroby. W takich przypadkach test działa prawidłowo, a wyniki są trafne.

Wyniki badania mogą jednak stwierdzać coś zupełnie przeciwnego w stosunku do stanu zdrowia pacjenta, dając pozytywny rezultat fałszywie wskazujący na obecność choroby, której nie ma, lub negatywny rezultat fałszywie wskazujący na to, że pacjent jest zdrowy. W takich przypadkach test nie zadziałał prawidłowo i wyniki nie są trafne. Tabela możliwości jest podobna do tych istniejących, gdy akceptuje się lub odrzuca hipotezę w podejmowaniu decyzji za pomocą statystyki **[Sposób 4.]**.

Badania przesiewowe na obecność raka piersi bardzo skutecznie wykrywają prawdziwe przypadki raka. Jednak wadą tak czułego testu dla rzadko występującej choroby jest to, że o wiele więcej zdrowych osób zostanie poinformowanych, że mogą mieć raka. W badaniach medycznych szuka się kompromisu pomiędzy czułością a specyficznością. Testy o większej czułości dają zwykle więcej fałszywych wyników pozytywnych, ale w sytuacjach tak poważnych, gdy ważą się sprawy życia i śmierci, to skutek uboczny, z którym powinniśmy się pogodzić.

Zobacz również

- G. Gigerenzer, *Calculated risks. How to know when numbers deceive you*, Simon and Schuster, Nowy Jork 2002.