O nieuniknionej niedoskonałości testów.
Błędy I i II rodzaju

Testy, badające czy dana próbka (lub pojedyncza osoba) pochodzi z jakiejś dobrze znanej populacji (np. chorych na grypę), muszą być wykonywane. Czy mówimy o psychometrii, czy o badaniu lekarskim, albo o eksperymencie wysokich energii w fizyce, identyfikującym nową cząstkę elementarną – potrzebujemy narzędzi statystycznych do określania, czy dana próbka spełnia pewne warunki.

Jeśli produkujemy długopisy, nie sposób sprawdzać każdej sztuki z osobna w stosach gotowych kartonów wypełnionych długopisami przed wysłaniem ich w świat, bo trwałoby to nieskończenie długo. Jeśli dokonuje się rewizji na lotnisku, nie można jej poddać wszystkich jak leci, bo odprawy pasażerów stanęłyby w miejscu i słusznie wzbudziłoby to protesty ograniczania wolności podróżujących. Trzeba więc siłą rzeczy badać tylko losowe, lub wytypowane wstępnie, charakterystyczne porcje całości i na podstawie owej porcji wnioskować o całości. Tym właśnie zajmuje się statystyka.

W ogólności, chodzi o to, żeby móc ocenić jak najpewniej, czy dwie grupy punktów doświadczalnych należą do dwóch osobnych rozkładów (zdrowi/chorzy, bozon Higgsa/coś innego niż Higgs, sprawne długopisy/niesprawne długopisy), czy też do jednej, wspólnej. Odróżnić przypadki „trafione” od „nietrafionych”.

Czy można na podstawie testów psychologicznych (eksperymentów, kwestionariuszy) ocenić bez wahania, że dana osoba cierpi na jakieś zaburzenie? A może to tylko układ jej cech osobowości albo charakterystyczny styl radzenia sobie? Albo wada charakteru, dająca się naprostować? Okazuje się, że – jakkolwiek nie możemy zrezygnować z testów, bo nawet niedoskonałe są pożyteczne i pomocne – nie jest możliwe stworzenie pewnego i zawsze skutecznego testu.

Na lotnisku szczególną uwagę przykładamy do tego, żeby na pokład nie wniesiono bomby. Pobieżne sprawdzenie dotyczy każdego, ale nie daje ono całkowitej gwarancji. Pewność mielibyśmy tylko wówczas, gdybyśmy przy każdym pasażerze na wszelki wypadek uruchamiali alarm bombowy, podrywając na nogi ochronę i poddając go precyzyjnej osobistej rewizji, czy nie ma przy sobie komponentów, z których można złożyć bombę. Ale taka procedura całkowicie zablokowałaby lotnisko i zakorkowałaby je na amen – mielibyśmy co prawda pewność, że nikt nie wniesie bomby, ale samoloty stałyby uziemione, a pasażerowie w kilometrowych kolejkach do

odprawy i w fatalnych humorach, które mogłyby spowodować nieobliczalne skutki podczas lotu. Innymi słowy, to nie jest rozwiązanie!

To, czego obawiają się kontrolerzy na lotnisku, to popełnienie **błędu I rodzaju**, to znaczy przepuszczenia bagażu, którego absolutnie nie powinno się pozwolić przemycić. Podobnie naukowcy medyczni, opracowujący lekarstwo, chcą za pomocą rygorystycznych testów zagwarantować, że ich lek nie będzie szkodliwy dla wszystkich ludzi jacy są, a farmaceuci, że każda pigułka będzie działać identycznie i nie spowoduje powikłań. Również oni chcą jak najbardziej zminimalizować pojawienie się błędu I rodzaju. Niestety, okazuje się, że nie mogą oni wyeliminować błędu zupełnie, jest to fizycznie niemożliwe.

A to dlatego, że wraz z zaostrzaniem rygorów dotyczących wychwycenia wszelkich szkodliwych pierwiastków, osłabiają moc i skuteczność leku. Jeśli lek działa silnie, to zawsze jest ryzyko, że komuś może zaszkodzić. Jeśli osłabimy lek, to owszem, uczynimy go bezpiecznym, ale nie będzie on spełniał swojej podstawowej roli. W granicznym przypadku, pakując do pigułki wyłącznie placebo, zapewniamy, że będzie ona bezpieczna dla wszystkich – nie otrujemy nią nawet jednej osoby na milion – ale skuteczność takiego leku jest dokładanie zerowa.

Aby wyjaśnić, czemu się tak dzieje, wyobraźmy sobie jakiś profil (rozkład), który opisuje skuteczność wykrywania niebezpiecznych substancji – albo który opisuje, które przypadki osób, wypełniających dany test psychologiczny, uznamy za cierpiące na zaburzenie, które sprawdza test. Naszym celem jest uniknięcie przepuszczenia przez nasz test przypadku niebezpiecznego, przypadku zaburzenia. W tym celu, żeby jak najwięcej osób nawet z nieswoistymi, nietypowymi i słabymi objawami nasz test „oznaczył” jako „trafione”, musimy poszerzać i rozciągać jego rozkład. Ale wówczas coraz więcej przypadków, które wcale nie były zaburzeniem, tylko podobnie się przejawiały, nasz test również w swojej zapobiegliwości i skrupulatności uzna za zagrożone. Taki test przestanie nam odróżniać przypadki zdrowia od zaburzenia, automatycznie wszystkich „na wszelki wypadek” uznając za zaburzonych. To tak samo jak poddawanie rewizji każdego bagażu i każdego ubrania pasażera i wzbudzanie alarmu bombowego za każdym razem, dla wszelkiej pewności. Mówimy, że test stracił swoją *moc predykcyjną*, to jest *moc dyskryminującą*. Uznanie za trafiony przypadek, który jest normalny i wcale nie trafiony, jest tożsamy z popełnieniem **błędu II rodzaju**. Nasz test nie potrafi (lub boi się) uznać kogoś za zdrowego i nie potrafi odróżnić zdrowego od chorego. Taka diagnoza, czy to psychologiczna, czy lekarska, jest również niedopuszczalna!

Często mamy z błędem II rodzaju do czynienia potocznie, gdzie na podstawie jednej, o niczym nie przesądzającej cesze, za sprawą tylko naszej intuicji, albo na podstawie pojedynczego, jednostkowego zachowania, już wnioskujemy, że dana osoba na coś cierpi i że jest to jej stała

cecha. Człowiek kichnął, więc jest przeziębiony. Oceniamy stałą cechę osoby na podstawie szczególnego, pojedynczego przypadku, albo danego stylu zachowania się. Ktoś nosi przy sobie stale parasol, bo uważa, że tak jest wytwornie, a my oceniamy, że jest hipochondrykiem, który boi się deszczu. Albo jest introwertykiem i ogranicza swoje kontakty towarzyskie (cecha osobowości, ani zła, ani dobra), a my uznajemy go za autystycznego (choroba).

Powyższe rozważania powinny zaowocować już pewną intuicją Czytelnika – taką, że im bardziej staramy się zminimalizować wystąpienie błędu I rodzaju (przepuszczenie bomby), tym bardziej powiększamy automatycznie i nieuchronnie ryzyko wystąpienia błędu II rodzaju (uznanie za bombę zwyczajnego bagażu). Musimy sami ustalić, na czym nam bardziej zależy (czy na wykrywaniu wszystkich bomb, czy raczej na sprawnym funkcjonowaniu lotniska), i pójść na jakiś kompromis pomiędzy oboma błędami. Nie jest to kwestia niedokładności naszych testów, ale immanentna cecha przyrody, rzeczywistości, która jest nie do przeskoczenia.

Dla opisania precyzji testu, używa się pojęcia *poziomu ufności* (PU). Oznacza on liczbową miarę tego, ile „trafionych" przypadków na sto (czyli procentowo) jesteśmy gotowi przepuścić, przegapić przy użyciu testu, czyli – jak duże ma być ryzyko popełnienia błędu I rodzaju. Np. test o poziomie ufności 95% statystycznie przepuszcza pięć sztuk na sto, które powinniśmy byli zaliczyć za „trafione". Przy rozkładzie normalnym (Gaussa) odpowiada to m/w części rozkładu na dwa odchylenia standardowe w obie strony wokół średniej. Wszystkie przypadki w dalszych skrzydłach rozkładu zostaną przepuszczone.

W medycynie brak prawidłowej diagnozy choroby (lub identyfikacji trucizny w farmaceutyce) jest niedopuszczalny, dlatego przyjmuje się najczęściej PU = 99% albo 99,9% (tylko jeden szkodliwy przypadek na tysiąc zostanie zignorowany przez test). Jako się rzekło, nie można powiększać PU bezkarnie, bo lawinowo rośnie nam błąd II rodzaju i bardzo rygorystyczny test w ogóle nie odróżnia już przypadków zdrowych od niebezpiecznych – wszystkie uznaje za niebezpieczne. Albo już całkowicie wyeliminowaliśmy z badanego leku potencjalnie groźne substancje czynne i lek w rezultacie w ogóle przestał szkodzić, ale przy okazji także zupełnie przestał leczyć!

Przedział ufności (wielkość stowarzyszona blisko z poziomem ufności PU) określa nam, jak blisko leżące dwa punkty doświadczalne uznamy za niemożliwe do odróżnienia za pomocą tego testu, czyli jaka jest „zdolność rozdzielcza" testu – minimalna różnica, którą uznamy za znaczącą (istotną) statystycznie. To właśnie dlatego porównywanie surowych wyników testów psychologicznych jest błędem metodologicznym. Nie da się powiedzieć, że jeden wynik był większy od drugiego, jeśli

znajdują się one bliżej, niż określa przedział ufności. I właśnie dlatego zbiera się wyniki we wspólne kategorie: poziomu **niskiego, średniego i wysokiego.** Możemy zatem powiedzieć na podstawie np. testu osobowości (Wielkiej Piątki OCEAN – NEO-FFI), że cechujemy się wysoką sumiennością oraz przeciętnym poziomem ekstrawersji, ale nie, że nasza ekstrawersja wynosi np. 25 punktów. Ani nawet nie to, że znajduje się ona w szóstym lub ósmym decylu.

Jeśli druga osoba (z tej samej kategorii wiekowej, płci, etc., bo testy specjalnie się *standaryzuje* w wyniku badań dla danych podpopulacji i grup, aby ich wyniki wolno było ze sobą porównywać w ramach jednej grupy) osiągnęła wynik *wysoki* na nerwowość, a ja uzyskałem średni, to mogę uczciwie powiedzieć, że mam niższą nerwowość od tej osoby. Natomiast, jeśli oba nasze wyniki wpadają w ten sam przedział (np. wysoki), to nie jestem w stanie powiedzieć, że któryś z nas ma wyższą nerwowość od drugiego – ponieważ test o tym nie przesądził. Rozstęp i liczba kategorii wynikowych w teście pokazuje, jak precyzyjny jest test, jak dokładnie potrafi rozróżniać pomiędzy natężeniami danej cechy.

Nie ma testu doskonałego, idealnie dokładnego – ale musimy się posługiwać testami, aby w ogóle móc diagnozować chorych oraz nie przepuszczać bomb na lotnisku i wadliwych długopisów w sprzedaży. Nie mówiąc o przegapieniu bozonu Higgsa! Niedoskonałość testów jest dokładnie badana i na poziomie sprawdzonej laboratoryjnie jego tzw. *rzetelności*, jest on dopuszczany lub nie do: użytku domowego, do badań wspomagających diagnozę, do opierania na nim badań diagnostycznych, do badań laboratoryjnych wysokiej jakości. Na danym poziomie odpowiedzialności za wyniki można używać tylko tych testów, które zostały do niego certyfikowane. Nikt nie może stworzyć sobie testu i na jego podstawie stawiać diagnozy, zwłaszcza w psychologii i medycynie.

Między innymi, to właśnie dlatego wyłącznie profesjonalni psychologowie mają prawo przeprowadzać i interpretować testy psychologiczne, w oparciu o dane dostarczone im przez zespół badaczy nad skutecznością i standaryzacją danego testu, a zawarte w instrukcji i kluczu do niego. Jeśli klucz do testu przeniknie (wycieknie) do opinii publicznej (np. do internetu), na ogół zaleca się jego całkowitą eliminację z diagnostyki – bo ludzie mogą nauczyć się symulować, odpowiadać zgodnie z kluczem, co przekreśla sens dokonywania testu. Świadomie lub nieświadomie odpowiadać tak, jak chcieliby wypaść, a nie jak jest naprawdę. Przy czym owo „naprawdę” w testach jest traktowane tylko z danym PU.

Nb. Specjalna gałąź psychometrii i metodologii testów zajmuje się tworzeniem kryteriów diagnostycznych oceniających, czy i na ile badany zgadywał odpowiedzi, albo były one udzielane według z góry ustalonego klucza.