

4.1 Tło historyczne (można śmiało pominąć, ale po co?)

Pojęcie hipotezy statystycznej ewoluowało przez setki lat. Pierwsze zachowane wzmianki o koncepcie hipotezy można znaleźć w pracy „Teoria Matematyki” greckiego filozofa Geminus’a (pierwsze dziesięciolecia naszej ery). Termin „hipoteza” był przez wieki używany w astrologii oraz fizyce. Przykładami są prace Gottfrieda Wilhelma Leibniz’a („Nowe hipotezy fizyczne”, 1671) oraz Isaaca Newtona („Hipotezy o świetle”, 1675).

Wzmianki o pierwszej hipotezie zweryfikowanej na gruncie analizy statystycznej dotyczą pracy medyka Johna Arbuthnota (1667 – 1735), który w roku 1710 opublikował w Royal Society pracę „An argument for Divine Providence, taken from the constant regularity observ’d in the births of both sexes”. W pracy tej przedstawił roczne liczby urodzeń chłopców oraz dziewcząt w Londynie z lat 1625-1710 oraz zauważył, że w każdym roku rodziło się więcej chłopców niż dziewcząt. Przyjmując, że częstość urodzin chłopców jest równa $1/2$, prawdopodobieństwo, że przez 86 lat co roku rodziło się więcej chłopców niż dziewcząt jest równe $1/2^{86} < 10^{-24}$ czyli jest niezmiernie małe. Było to dla niego dowodem na to, że częstość urodzin chłopców jest statystycznie istotnie większa niż częstość urodzin dziewcząt.

Wnioskowanie Johna Arbuthnota zostało skrytykowane między innymi przez Nicholasa Bernoulli’ego, co prawdopodobnie spowodowało, że pierwszeństwo w konstrukcji pierwszego statystycznego testu hipotezy statystycznej często przypisuje się Pierre-Simon Laplace’owi (1749 – 1827). Przedstawił on w roku 1796 fizykalno-matematyczne uzasadnienie „nebular hypothesis” (hipotezy mgławicowej, nazywanej też hipotezą Kanta-Laplace’a), opisującej genezę powstania Układu Słonecznego. Hipoteza ta opiera się na przypuszczeniu, że planety powstały w wyniku stopniowego odrywania się od wirującego Słońca pierścieni gazowej materii, przekształcających się z czasem w zwarte kule. Jako uzasadnienie tej hipotezy Laplace wykazał, że ekliptyki planet nie są losowe, lecz leżą blisko jednej pierwotnej ekliptyki. Laplace był zwolennikiem subiektywnej interpretacji prawdopodobieństwa, dlatego przeprowadzone przez niego wnioskowanie było bliskie wnioskowaniu Bayesowskiemu.

Przez kolejny wiek uczeni stawiając i weryfikując hipotezy statystyczne kierowali się intuicją. Dopiero w latach dwudziestych XX wieku aksjomatyczne podstawy dla zagadnienia testowania opracowali Jerzy Sława-Neyman (matematyk polskiego pochodzenia) i Egon Pearson (syn znakomitego statystyka Karla Pearsona). Matematyczne wykształcenie Jerzego Neymana, wspólnie z intuicjami jego współpracownika Egona Pearsona, pozwoliły na spójne przedstawienie teorii testowania hipotez. W latach od 1928 do 1933 napisali oni wspólnie wiele istotnych prac o procesie testowania hipotez, testach statystycznych, testach najefektywniejszych, rozmiarach testu, poziomach istotności itp. Jako pierwsi do procesu

testowania wprowadzili pojęcie hipotezy alternatywnej (dziś oczywiste i niekwestionowane), przez co byli długo krytykowani przez współczesne im autorytety, między innymi Ronalda Aylmera Fishera.

Jerzy Spława-Neyman podczas studiów wyższych w Charkowie doskonale poznał, wywodzącą się z Rosji, częstościową interpretację prawdopodobieństwa, której aksjomatyczne podwaliny zostały stworzone przez Andrieja Kołomogorowa. Na tej interpretacji prawdopodobieństwa oparł swoją aksjomatyczną teorię testowania, przez co jest ona nazywana również częstościowym, lub klasycznym ujęciem testowania hipotez. Na cześć jej twórców jest ona również nazywana Neymanowsko-Pearsonowską teorią testowania hipotez. Takie ujęcie procesu testowania hipotez dominowało od lat 30-40 do końca XX wieku.

Metody testowania były i są rozwijane nie tylko przez matematyków, ale również przez biologów, fizyków oraz chemików. Przykładowo, znaczący wkład do statystyki wniosły prace podpisywane pseudonimem „Student”, których autorem był pracownik browaru Guinness, chemik William Gosset, ukrywający swoje nazwisko.

W teorii testowania hipotez ważny jest wybór poziomu istotności, czyli dopuszczalnego prawdopodobieństwa odrzucenia prawdziwej hipotezy zerowej. Najczęściej przyjmowany poziom istotności to 0.05, co oznacza, że średnio błędnie odrzucimy hipotezę zerową nie częściej niż raz na 20 razy. Rozważmy jednak zbiór 100 hipotez zerowych, każda orzekająca, że pewien lek nie wpływa na stan pacjenta. Wykonajmy 100 testów, każdy na poziomie istotności $\alpha = 0.05$. Nawet, jeżeli żaden z rozważanych leków nie wpływa na zdrowie pacjenta, to w około 5 przypadkach test odrzuci błędnie hipotezę zerową, a przyjmie hipotezę alternatywną. Czy to oznacza, że te 5 leków istotnie wpływa na stan pacjenta? Oczywiście nie, jednak ze wzrostem liczby przeprowadzonych testów na ustalonym poziomie istotności, rośnie prawdopodobieństwo błędnego odrzucenia przynajmniej jednej hipotezy zerowej.

Zwrócił na to uwagę biostatystyk Graham Martin, który w grudniu 1984 roku w *The Lancet* opublikował list zatytułowany „Munchausen Statistical Grid, that makes all trials look significant”. W liście tym Graham Martin oskarżał badaczy stosujących metody statystyczne, szczególnie w medycynie, o powtarzanie eksperymentu wielokrotnie lub też wykorzystywanie różnych testów, tak długo, aż w któryś test odrzuci hipotezę zerową i „potwierdzi” słuszność ich przypuszczeń. Taka procedura nazywana została statystyczną siatką Munchausena. Nazwa pochodzi od nazwiska Barona von Munchausena, który znany był z opowiadania niesamowitych i nierzeczywistych historii (historia o tym, jak to Baron von Munchausen wyciągnął się z bagna za własne paski od butów pojawia się często w genezie nazwy metody bootstrap). Zarzuty tego typu sprawiły, że przy publikacji wyników zaczęto stawiać wymóg stosowania korekty poziomu istotności uwzględniającej liczbę weryfikowanych hipotez. Historycznie pierwszą i wciąż najpopularniejszą korektą jest korekta Bonferroniego, polegająca na podzieleniu poziomu istotności przez liczbę rozważanych hipotez. Korekta ta jest

z powodzeniem wykorzystywana do dziec w przypadku, gdy testowane jest kilka lub kilkadziesiąt hipotez.

4.2 Prawda a testowanie hipotez

Statystyczna teoria testowania hipotez dotyczy weryfikacji przypuszczenia dotyczącego pewnego parametru. Dlatego też istnieje pewna rozbieżność pomiędzy pytaniem interesującym praktyka (czy jestem chory na raka) a odpowiedzią, którą statystyk może udzielić (przy przyjętych założeniach na poziomie istotności 5% nie można odrzucić hipotezy o braku istotnych statystycznie objawów choroby).

Rozważmy następujący przykład. Pewien naukowiec wyprodukował chemicznie nawóz, i chce sprawdzić czy nawożenie tym produktem zwiększy średnią masę pomidorów. Oczywiście, każdy pomidor jest inny, będziemy więc porównywać średnią masę pomidora w grupie pomidorów nawożonych nowym specyfikiem i w grupie pomidorów hodowanych tradycyjnie. Przyjmujemy pewien model, w tym przypadku że waga pomidora w kilogramach może być opisana przez zmienną losową o rozkładzie $\mathcal{N}(\mu_x, \sigma^2)$, gdzie $\mu_x = \mu_{test}$ to średnia waga pomidora wyhodowanego gdy krzak nie był nawożony nowym specyfikiem, a $\mu_x = \mu_{nawoz}$ to średnia waga pomidora, którego krzak był nawożony. Ani nasz naukowiec ani my nie wiemy jakie są parametry μ_{nawoz} i μ_{test} , chcemy jednak sprawdzić, czy istnieją dowody że te wartości są istotnie różne.

Możliwe są dwie sytuacje, mianowicie $\mu_{nawoz} > \mu_{test}$, czyli nawożenie powoduje że średnia waga pomidora jest większa, lub $\mu_{nawoz} \leq \mu_{test}$. W trakcie przeprowadzania procesu testowania, możemy podjąć jedną z dwóch decyzji, możemy uznać, że na podstawie obserwacji średnia waga pomidora z nawożonego krzaka jest większa, lub możemy uznać, że nie ma ku temu wystarczających przesłanek. Możliwe sytuacje opisuje poniższa tabelka.

prawda	decyzja	decyzja: nawóz pomaga	decyzja: nawóz nie pomaga
$\mu_{nawoz} > \mu_{test}$		decyzja poprawna	błąd pierwszego rodzaju
$\mu_{nawoz} \leq \mu_{test}$		błąd drugiego rodzaju	decyzja poprawna

W procesie testowania nie można jednocześnie minimalizować błędów obu rodzajów, przyjmuje się więc błąd pierwszego rodzaju za ważniejszy i to ten błąd chcemy minimalizować.

4.3 Popularne testy statystyczne

TODO: opisać test na równość średnich (t-studenta), równość wariancji (F-test), zgodność (ks-test, shapiro-wilka) i niezależność (znakow i chi2)