

Walenty Ostasiewicz

Akademia Ekonomiczna we Wrocławiu

UWAGI HISTORYCZNE O ROZWOJU STATYSTYKI

1. Początki statystyki księgowo-opisowej

Statystyka jest to „rodzona siostra” rachunku prawdopodobieństwa. Ponieważ „oboje” pojawili się mniej więcej w tym samym czasie, a mianowicie ok. roku 1660, to można ich nawet traktować jako bliźniaków.

Słowo „statystyka” jest pochodzenia włoskiego, wyraz *lo stato* oznacza państwo, *statista* zaś męża stanu, polityka. Słowem statystyka określono ogólną wiedzę o państwie.

Już w roku 1581 G. Ghilini określał w Mediolanie statystykę jako wiedzę o wszelkich przejawach życia w państwie, polityce i wojsku (w oryginale: *civile politica, statistica e militare scienza*). W takim samym duchu naukę tę rozwijano też na terenie niemieckim [Crombie 1994].

Pierwszy udokumentowany wykład o statystyce wygłosił w języku łacińskim profesor filozofii uniwersytetu w Helmstädt H. Conring (1606-1681) 20 listopada 1660 r. Wszystkie jego wykłady zostały zaś opublikowane przez jego uczniów dopiero w 1730 r.

W całej tej pracy, jak to stwierdził K. Pearson, liczby pojawiają się tylko jako numery stron. We współczesnych podręcznikach statystyki G. Achenwall (1719-1772) wymieniany jest zwykle jako ten, który po raz pierwszy użył słowa statystyka na określenie nauki o państwie. Prawda jest jednak taka, że on tylko jako pierwszy wygłosił wykład o statystyce w języku niemieckim. Wykład ten wydany został drukiem w 1749 pt. *Abris der neuesten Staatswissenschaft der vornehmsten europäischen Reiche und Republiken*. Achenwall określił statystykę jako państwowznawstwo, ale w dalszym ciągu traktowana ona była jako wiedza opisowa.

2. Pierwsze próby wnioskowania statystycznego

Nieco inaczej rozumiano i rozwijano statystykę w Anglii. Przede wszystkim zaczęto tam rejestrować fakty życia gospodarczego nie werbalnie, lecz za pomocą liczb. Liczby zaś, jak się okazało, mogą „mówić” coś więcej niż to, co się zarejestrowało. Trzeba je tylko do tego odpowiednio zachęcić.

Możliwość taką zademonstrowali tzw. arytmetycy polityczni. W Angli bowiem, opisując wiedzę o państwie, używano nazwy „arytmetyka polityczna”. Wybitnymi przedstawicielami takiej arytmetyki byli dwaj przyjaciele J. Graunt (1620-1674) oraz W. Petty (1623-1667). Wiedząc, że w 1666 r. spłonęło 13 200 domostw, oraz to, że stosunek *liczby zgonów* w tych gospodarstwach do ogólnej liczby zgonów wynosi 1 do 5, W. Petty oszacował, że liczba gospodarstw domowych wynosi 66 000. Ponieważ stosunek liczby pożarów w roku 1686 do liczby pożarów z roku 1666 wynosił 4 do 3, łatwo można obliczyć liczbę gospodarstw domowych w roku 1686 – wynosiła ona 88 000.

W 1662 roku J. Graunt wydał książkę pt. *Natural and Political Observations Mentioned in the Following Index and Made upon the Bills of Mortality*, która miała pięć wydań i znana była w wielu krajach Europy. Dzięki tej publikacji handlarz J. Graunt dostąpił zaszczytu członkostwa w elitarnym Królewskim Towarzystwie Naukowym w Londynie. Kwestionowano to, czy Graunt był autorem tej wspaniałej książki, autorstwo przypisywano jego przyjacielowi W. Pettiemu.

Inną wybitną pracę, w której dokonano wnioskowania na podstawie zebranych danych, był obszerny artykuł naukowy o długim tytule *An Estimate of the Degrees of Mortality of Mankind Drawn from Courions Tables of the Birth and Funerals of the City of Breslaw with an Attempt to Ascertain the Price of Annuities upon Lives*, opublikowany w 1693 r. przez E. Halleya. W artykule tym znany astronom opracował po raz pierwszy tablice trwania życia. Opracował je na podstawie danych, które K. Neumann (1648-1715) zebrał w dzisiejszym Wrocławiu.

3. Początki myślenia probabilistycznego

Metodami uzyskiwania nowej wiedzy na podstawie wiedzy już posiadanej zajmowano się w starożytnej Grecji. Grecy wyraźnie rozróżniali wiedzę pewną, tzn. taką, o której można orzekać w kategoriach prawdy i fałszu, oraz wiedzę niepewną. Wiedzę pewną Grecy określali mianem *episteme* i do uzyskiwania jej stosowali niezawodne wnioskowanie logiczne. Od słowa *episteme* pochodzi dzisiejsza nazwa epistemologia, odnosząca się do działu filozofii dotyczącego poznawania otaczającej nas rzeczywistości. Wiedzę niepewną Grecy określali mianem *doksa* – oznacza to mniej więcej to, co dziś nazywamy przypuszczeniem, sądem lub opinią. Platon pisał, że w sądach nikomu nie chodzi o prawdę, lecz tylko o przekonanie, a przekonują nas prawdopodobieństwa, bowiem ludzie karmią się prawdopodobieństwem, gdyż jest ono podobne do prawdy. Do oceny pewnej wiedzy naukowej służyło tylko jedno kryterium – kryterium prawdziwości. Do oceny wiedzy niepewnej Grecy używali wielu zbliżonych do siebie określeń (por. [Crombie 1994; von Furstenberg 1990]). Najbardziej typowe są następujące: *εἰκος* (tłumaczone na język łaciński jako *versimile*, czyli podobne do prawdy) oraz *παραγοαρια* (tłumaczone jako *persuasiv*, czyli przekonujące).

Rzymski mąż stanu z kolei – Marcus Tullius Cicero (106-43 p.n.e.), tłumacząc dzieła greckiej filozofii na język łaciński, uznał, że najbardziej odpowiednim słowem jest słowo *probabilis*. To właśnie słowo używane jest obecnie do odróżnienia myślenia logicznego od myślenia o wiedzy niepewnej, która nazywa się myśleniem probabilistycznym.

W starożytnej Grecji ten sposób myślenia rozwijany był i powszechnie stosowany w dwóch dziedzinach praktycznej działalności, a mianowicie w medycynie i w sądownictwie. Pojęcie prawdopodobieństwa w czasach starożytnej Grecji miało więc charakter wyłącznie retoryczny. Z punktu widzenia naukowego było ono traktowane na takim samym poziomie jak pojęcie prawdziwości. Jedno dotyczyło sądów, drugie zaś naukowej wiedzy pewnej.

4. Probabilizm i powstanie rachunku prawdopodobieństwa

Rozwój teorii prawdopodobieństwa i spokrewnionej z nią statystyki matematycznej jako uniwersyteckich dyscyplin naukowych dokonał się dopiero w XVII wieku. Do rozwoju rachunku prawdopodobieństwa przyczynił się B. Pascal (1623-1662), a do rozwoju statystyki inny geniusz tego okresu, a mianowicie I. Newton (1642-1727).

Zgodnie z duchem owych czasów, inspirację dla obu z nich stanowiły rozważania teologiczne. Historia przedstawiona w typowych podręcznikach dotyczy korespondencji listownej, jaką A. Gombaud (1607-1684), czyli kawaler de Mere, prowadził z Pascalem na temat podziału stawki w grze hazardowej.

Do rozwoju teorii prawdopodobieństwa przyczyniła się jednak o wiele bardziej korespondencja prawnika P. de Fermata (1601-1665) z Pascalem. Warto mocno podkreślić, że w korespondencji tej słowo „prawdopodobieństwo” w ogóle się nie pojawiało. Dotyczyła ona sprawiedliwej *ceny*, jaką trzeba zapłacić, aby uczestniczyć w grze hazardowej. Korespondencja ta trwała od lipca do października 1654 r. Dnia 23 listopada 1654 r. Pascal doznał olśnienia religijnego, przeniósł się do Port-Royal i poświęcił się całkowicie problematyce metafizycznej. W wyniku rozważań teologicznych sformułował podstawy wnioskowania probabilistycznego, które połączył z wnioskowaniem logicznym. Sformułował też podstawy logiczne racjonalnego podejmowania decyzji w warunkach niepewności i ryzyka. Rozwagał o ryzyku najgroźniejszym ze wszystkich, tzn. ryzyku wiecznego potępienia, którego można jednak uniknąć, podejmując właściwą decyzję o prowadzeniu życia zgodnego z Pismem Świętym.

Aby przekonać niewierzącego kawalera de Mere do wiary chrześcijańskiej, Pascal prowadził grę, w której stawką było zbawienie. Do osiągnięcia zbawienia potrzebne jest odnowienie utraconego ducha chrześcijańskiego. Zgodnie z doktryną Janseniusza (1585-1638) było to możliwe przez powrót do nauk św. Augustyna. Przeciwno takiemu stanowisku ostro zaprotestował Molina (1535-1601), uważany

za głównego twórcę kazuistyki jezuitów oraz głoszonej przez nich doktryny probabilizmu. Obalając tę doktrynę, Pascal usunął poważną przeszkodę w rozwoju myślenia probabilistycznego, była ona bowiem zaprzeczeniem takiego myślenia.

5. Rozwój teorii prawdopodobieństwa

Zrywając z dotychczasowym świeckim życiem, Pascal drastycznie ograniczył swe kontakty nawet z ludźmi nauki. Genialny Holender Ch. Huygens (1629-1695), twórca falowej teorii światła i autor wielu innych odkryć, zainteresował się wynikami Pascala i Fermata. Chcąc dokładnie poznać te wyniki, w roku 1655 udał się do Paryża, aby spotkać się z Pascalem. Do spotkania w tym roku jednak nie doszło. Obaj uczeni spotkali się kilka razy, ale dopiero pięć lat później.

Ch. Huygens napisał pracę po holendersku, którą później jego nauczyciel przetłumaczył na język łaciński. Pracę pt. *De ratiociniis in ludo aleae* wydano w 1657 r. (por. [Płocki 1992]). Zawarte w niej były problemy, które rozważali Pascal z Fermatem, sformułowanych też było wiele zadań nierozwiązanych, ale najważniejszą rzeczą godną odnotowania było to, że w opracowaniu tym po raz pierwszy zdefiniowano wartość oczekiwaną, określoną mianem *expectatio*, jako wartość gry hazardowej, czyli jako sprawiedliwą cenę gry. Pojęcie to Huygens określił w sposób werbalny: jeśli liczba przypadków, w których otrzymuje się sumę a , wynosi q , zaś liczba przypadków, w których otrzymuje się sumę b , wynosi q , a wszystkie przypadki mogą się zdarzyć jednakowo łatwo, to wartość oczekiwana wynosi $(pa + qb)/(p + q)$.

Książka Huygensa o rachunku w grze w kości była wznawiana kilkakrotnie i prawie przez 50 lat służyła jako podręcznik do rachunku prawdopodobieństwa. Słowo „prawdopodobieństwo” w podręczniku tym w ogóle nie występowało, bowiem w tym czasie wciąż było rozumiane jako argumentowanie, tzn. argumentowanie „za” pewnymi kategorycznymi stwierdzeniami lub „przeciw” nim. Dzieło Huygensa stanowiło inspirację dla Bernoullego do przygotowania przełomowego dzieła w dziejach rozwoju prawdopodobieństwa.

Jacques Bernoulli, którego imię zapisywane jest też jako Jacob lub James, urodził się dokładnie w tym samym roku, w którym rodził się rachunek prawdopodobieństwa, tzn. w roku 1654, zmarł zaś w 1705. Z pracą Huygensa zapoznał się on już w 1684 roku, a następnie przystąpił do opracowywania swego dzieła na temat rachunku prawdopodobieństwa. Dzieło to, pt. *Ars conjectandi*, zostało wydane pośmiertnie w 1713 r. Łacińskie określenie *ars conjectandi sive stochastice* oznacza po polsku sztukę przewidywania, czyli stochastykę. Dzieło Bernoullego nawiązuje do słynnego dzieła *Ars cogitandi*, czyli sztuki myślenia, wydanego bezimiennie w Paryżu w 1622 r. (wiadomo jednak, że jego autorami byli czołowi janseniści: A. Arnauld i P. Nicole, a także B. Pascal). W dziele tym po raz pierwszy pojawił się termin „prawdopodobieństwo” w sensie czegoś mierzalnego (por. [Hacking 1975]). Jednak pojęcie to po raz pierwszy zostało zdefiniowane i omówione w pracy Bernoullego. Zostało ono zdefiniowane jako *stopień pewności*, czyli jako uzasadnione

przekonanie czy też jako *stopień racjonalnej wiary* (por. [Shafer 1990]). Prawdopodobieństwo w tej pracy rozumiane jest metafizycznie, jako coś nieobserwowalnego bezpośrednio. J. Bernoulli był zwolennikiem determinizmu i odrzucał wszelką tzw. obiektywną przypadkowość. Uważał, że gdyby obserwację wszelkich zdarzeń prowadzono przez całą wieczność, prawdopodobieństwo przeszłoby w całkowitą pewność i doszlibyśmy do przeświadczenia o wszechmocy najwyższego Stwórcy.

Zarówno prawdopodobieństwo, jak i tzw. przypadkowość były traktowane jako kategorie subiektywne (por. [Januszkiewicz 1976]). Stosunek prawdopodobieństwa do pewności traktował Bernoulli tak samo jak stosunek części do całości. W swej pracy wyraził to następująco: *probabilitas enim est gradus certitudinis, et ab hac differt ut pars a toto* (por. [Shafer 1990]). Książka Bernoullego składa się z czterech części, oto oryginalne ich tytuły:

I. Tractatum Hugonii de Ratiociniis in Ludo Aleae.

II. Doctrinam de Permutationibus et Combinationibus.

III. Usum Praecedentis Doctrinae in Variis Sortitionibus et Ludis Aleae.

IV. Usum et Applicationem Praecedentis Doctrinae in Civilibus, Moralibus et Oeconomicis.

W pierwszej części omawiano dzieło Huygensa i rozwiązano wszystkie zadania przez niego postawione. W części drugiej podano wykład o kombinatoryce, część trzecią poświęcono różnym wariantom gier hazardowych. Część czwarta jest częścią, która „zrewolucjonizowała teorię prawdopodobieństwa” (por. [Hacking 1975]). W części tej udowodniono twierdzenie, które w czasie późniejszym S. Poisson nazwał prawem wielkich liczb. A ono z kolei umożliwiło stosowanie prawdopodobieństwa nie tylko w mało poważanej dziedzinie gier hazardowych, ale w szeroko rozumianych naukach społecznych.

Doniosłość dzieła Bernoullego polega na tym, że dokonał on w nim połączenia trzech koncepcji prawdopodobieństwa: ceny sprawiedliwej, subiektywnego przekonania oraz częstości.

6. Rozwój myślenia statystycznego

Niemal równolegle z rozwojem teorii prawdopodobieństwa formowały się podstawy myślenia statystycznego. Podobnie jak w przypadku prawdopodobieństwa źródło rozwoju tego rodzaju myślenia stanowiły także rozważania teologiczne. Nie kto inny, jak jeden z głównych twórców nowoczesnej statystyki, a mianowicie K. Pearson, twierdził, że idea wszechobecnej boskości utrzymującej stabilność średnich wartości statystycznych stanowiła podstawę rozwoju statystyki od Derhama, Süßmilcha, Niewentyta i Price’a do Queteleta i Nightingale.

Przyczyny, które doprowadziły De Moivre’a do odkrycia krzywej rozkładu normalnego lub Bayesa do jego twierdzenia, były bardziej teologiczne i socjologiczne aniżeli czysto matematyczne. Większy wpływ na te i inne wyniki miała teologia Newtona aniżeli jego matematyka (por. [Hacking 1965; 1975]).

Trzy lata przed opublikowaniem dzieła J. Bernoullego ukazała się drukiem praca J. Arbuhnota pt. *An Argument for Divine Providence Taken from the Constant Regularity Observed in the Births of Both Sexes*, w której po raz pierwszy do odrzucenia hipotezy zerowej, że chłopców i dziewczynek rodzi się tyle samo, zastosowany został test statystyczny. Argumentacja statystyczna i teologiczna zawarta w tej pracy uzasadniająca twierdzenie o istnieniu Opatrzności boskiej miała wielki wpływ na innych uczonych badających istotę determinizmu i losowości. W szczególności A. De Moivre w swym epokowym dziele o doktrynie losowości (*The Doctrine of Chances: or a Method of Caleulating the Probability of Events in Play*) przedstawił swoją opinię bardzo zbliżoną do tej, którą po raz pierwszy wyraził J. Arbuthnot. De Moivre udowodnił w dziele tym m.in. twierdzenie o aproksymacji rozkładu dwumianowego rozkładem normalnym, który odkrył już w 1733 r., a więc ponad 40 lat wcześniej, zanim się urodził F. Gauss.

Problem wnioskowania statystycznego o parametrze p w rozkładzie dwumianowym na podstawie obserwacji zmiennej o tym rozkładzie został rozwiązany dopiero przez T. Bayesa. Wynik ten znany dzisiaj jako twierdzenie lub wzór Bayesa został opublikowany pośmiertnie w 1764 r. W odróżnieniu od innych wyników wynik Bayesa przez długi czas nie wywarł prawie żadnego wpływu na rozwój statystyki. Wzór Bayesa na nowo odkrył P. Laplace, który technikę „odwrotnego prawdopodobieństwa” stosował oprócz rozkładu dwumianowego do wielu innych rozkładów. W wyniku tych badań powstało jego słynne centralne twierdzenie o aproksymacji rozkładem normalnym sumy zmiennych o dowolnych rozkładach.

Kolejnego przełomu w rozwoju statystyki dokonał genialny Belg A. Quetelet. Rozwijając idee, które wcześniej formułował J.P. Süßmilch, Quetelet udowodnił swoimi pracami, że statystyka może być wykorzystywana jako narzędzie doskonalenia społeczeństwa, tzn. jako narzędzie poprawy losu ludzkiego bytowania.

7. Regresja, korelacja i metoda najmniejszych kwadratów

Pracami Queteleta zafascynowany był sir F. Galton (1822-1911), autor tak powszechnie dzisiaj stosowanych pojęć, jak regresja i korelacja.

F. Galton wiele czasu poświęcił na badanie związków między dwiema cechami, np. między wzrostem rodziców i wzrostem dzieci. K. Pearson, uczeń Galtona, zebrał dane o wzroście ojców i ich dorosłych synów dla 1078 przypadków. Średni wzrost ojca wynosił 68 cali, a średni wzrost syna wynosił 69 cali, odchylenie w obu przypadkach było takie samo i wynosiło 2,7 cala. W celu udzielenia odpowiedzi na pytanie: „jak zależy wzrost syna, traktowany jako cecha Y , od wzrostu ojca, traktowanego jako niezależna cecha X ?”, Galton opracował specjalną metodę o nazwie *metoda regresyjna*. Ze względu na spotykane w literaturze błędne rozumienie terminu „regresja” rozpatrzmy szczegółowo zarówno samą metodę, jak i jej nazwę. Zauważmy przede wszystkim, że sformułowane wyżej pytanie nie może być przedstawione w postaci: „jaka jest wartość y , czyli jaki jest wzrost syna, jeśli

wiadomo, że wzrost ojca wynosi x cali?" Zależność między wartościami x a wartościami y nie jest bowiem zależnością funkcyjną typu $y = f(x)$. Jednej wartości x cechy X , np. wzrostowi ojca równemu 72 cale, odpowiada zwykle kilka lub wiele wartości cechy Y , które są mniej lub bardziej rozproszone. Stopień rozproszenia tych wartości może być różny i może zależeć od wartości cechy niezależnej X .

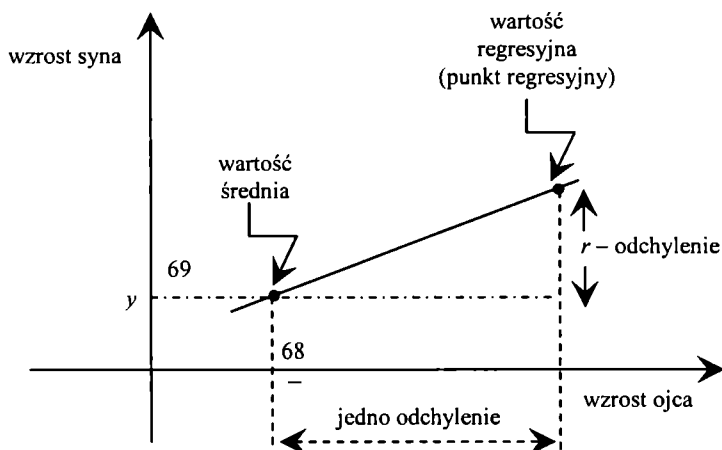
Istota myślenia „regresyjnego” polega na tym, że zamiast stosowania naturalnych jednostek do pomiaru rozproszenia, takich jak cale, kilogramy itp., stosowana jest jednostka o nazwie *odchylenie standardowe* (SD).

Szukając związku między cechami, zamieniamy pytanie: „jaka wartość y cechy Y odpowiada wartości x cechy X ?”, na pytanie: „jak zmieni się wzrost syna, jeśli wzrost ojca zwiększy się od średniej wartości wśród wszystkich ojców o jedno odchylenie standardowe?”. Zamiast zwrotu „o jedno odchylenie standardowe” mogą być użyte zwroty typu „o dwa odchylenia”, „o półtora odchylenia” itp. Odpowiedź na takie pytania podawana jest również w postaci: „o dwa odchylenia standardowe”, „o pół odchylenia” itp.

F. Galton podał następujące rozwiązanie tego problemu: jeśli wartość x cechy X wzrośnie o jedno odchylenie standardowe (SD_X) od wartości przeciętnej, to wartość y cechy Y wzrośnie też o jedno odchylenie standardowe cechy Y , ale pomnożone przez wielkość pewnego czynnika r , tzn. zwiększy się o wielkość $r \cdot SD_Y$.

F. Galton określił również wielkość tego czynnika, który nazwał *co-relation*. K. Pearson ok. roku 1895 podał wyrażenie algebraiczne określające ten czynnik i wprowadził do dziś stosowaną nazwę „współczynnik korelacji”.

Istotą metody regresyjnej, tzn. metody wykorzystywania współczynnika korelacji do szacowania *średniej wartości* cechy zależnej w zależności od *dowolnej wartości* cechy niezależnej, przedstawiono na rysunku na przykładzie zależności wzrostu syna w zależności od wzrostu ojca.



Rys. 1. Istota metody regresyjnej

Linia łącząca wszystkie możliwe punkty regresyjne nazywa się linią regresyjną lub krzywą regresji. Aby wyjaśnić pochodzenie samej nazwy „regresja”, rozpatrzmy jeszcze jedno ważne zagadnienie związane z regresją, a mianowicie zagadnienie tzw. *efektu regresyjnego*, odkrytego również przez F. Galtona (por. [Freedman i in. 1878]).

Zauważył on, że wyjątkowo wysokie osoby (lub osoby wyjątkowo uzdolnione matematycznie itp.) jako grupa osób mają często dzieci niższe od siebie (lub mniej uzdolnione). Natomiast osoby bardzo niskie (lub np. mało uzdolnione) mają dzieci wyższe od siebie (lub bardziej uzdolnione itp.). Zjawisko to Galton nazwał początkowo rewersją (*reversion*), później użył zwrotu *regression to mediocrity*, czyli regres do średniactwa. Efekt regresyjny szczególnie dobrze znany jest psychologom eksperymentatorom, którzy bardzo przestrzegają początkujących badaczy przed popełnieniem typowego błędu polegającego na zbyt pochopnym wyciąganiu wniosków przyczynowo-skutkowych w przypadku doboru do eksperymentu osób mających skrajne wartości badanej cechy niezależnej.

Stosując współczesną terminologię probabilistyczną, regresję definiuje się jako warunkową wartość oczekiwaną. Rozpatrzmy przypadek najprostszy, dotyczący dwuwymiarowej zmiennej losowej (X, Y) . Otóż w tym przypadku definiuje się dwie regresje: regresję zmiennej Y względem zmiennej X oraz regresję zmiennej X względem zmiennej Y .

Pierwsza z nich jest to *funkcja* $g(x)$ zdefiniowana następująco:

$$g(x) = E(Y | X = x),$$

druga zaś jest to następująca funkcja:

$$h(y) = E(X | Y = y).$$

Zatem regresja w sensie Galtona jest to warunkowa wartość oczekiwana pewnej zmiennej losowej względem innej zmiennej losowej. Do określenia tej wartości stosowane są różne metody estymacji (nie aproksymacji) parametrycznej lub nieparametrycznej. Jedną z metod jest metoda najmniejszych kwadratów.

Gdy rozkład prawdopodobieństwa jest znany, to nawet estymacja jest niepotrzebna, gdyż funkcję regresji można wówczas podać dokładnie.

Jeżeli np. dwuwymiarowa zmienna losowa (X, Y) ma gęstość określoną za pomocą funkcji:

$$f(x, y) = (x + y) / 16, \quad x, y \in [1, 3],$$

to regresja zmiennej Y względem zmiennej X jest następująca (por. [Ostasiewicz 2006]):

$$g(x) = (12x + 26) / (6x + 12).$$

Zauważmy, że jest to funkcja hiperboliczna, czyli regresja wcale nie musi być funkcją liniową.

Jeśli między zmiennymi X i Y zachodzi związek

$$E(Y | X) = a + b \cdot X,$$

to regresja jest funkcją liniową (por. [Linnik 1962]).

Pojęcie regresji jest też często stosowane w innym sensie, a mianowicie w sensie Gaussa. W niewielkim uproszczeniu ten rodzaj regresji oznacza, że znane jest prawo $y = f(x)$, określające związek funkcyjny badanych wielkości (por. [Spanos 1986]). Związek ten przedstawiony jest w postaci parametrycznej, np. w postaci funkcji liniowej.

$$y = \alpha_0 + x_1 + \alpha_1 x_2 + \dots + \alpha_n x_n.$$

To znaczy przyjmuje się, iż wiadomo, np. z teorii ekonomii, fizyki, biologii itp., że ogólna postać związku jest liniowa, nieznane są zaś parametry tego związku, które wyznacza się po przeprowadzeniu pomiarów. Pomiary obarczane są często błędami. W najprostszym przypadku przyjmuje się, że tylko pomiary zmiennej niezależnej obciążone są błędami, a wówczas uzyskuje się następującą zależność:

$$Y = \alpha_0 x_1 + \alpha_1 x_2 + \dots + \alpha_n x_n + u,$$

którą nazywa się gaussowskim modelem liniowym. Ostatnio zaś coraz częściej zależność taką nazywa się *liniową regresją gaussowską* lub krótko regresją liniową, a nawet modelem ekonometrycznym.

Problem w tym przypadku polega na *aproksymacji* zbioru obserwacji $(y_i, x_{0i}, x_{1i}, \dots, x_{ni})$, $i = 1, 2, \dots, N$ kombinacją liniową podanej wyżej postaci. Jednym z najczęściej stosowanych kryteriów aproksymacji jest suma kwadratów odchyleń hiperpłaszczyzny od zaobserwowanych punktów. Stąd też do znalezienia wartości współczynników $\alpha_0, \dots, \alpha_n$ tej kombinacji stosowana jest metoda najmniejszych kwadratów. Spór o autorstwo tej metody do dziś nie jest w pełni rozstrzygnięty. Prawie pewne jest, że zastosował ją po raz pierwszy C.F. Gauss w 1795 r., natomiast A.M. Legendre jako pierwszy w 1809 r. opublikował pracę na ten temat i to on traktowany jest jako „Ford statystyki”, gdyż metoda najmniejszych kwadratów nazywana jest *automobilem współczesnej statystyki* (por. [Stigler 1981]).

Z chwilą gdy podane zostało probabilistyczne uzasadnienie tej metody, tzn. wykazano, że współczynniki $\alpha_0, \alpha_1, \dots, \alpha_n$ wyznaczone za pomocą tej metody są najlepszymi *estymatorami* nieznanymi parametrów, jeśli tylko zakłócenia pomiarów są *losowe*, przy czym niezależne od siebie i mają rozkład normalny ze stałą wariancją, liniowy model gaussowski zaczęto nazywać *liniową regresją* lub *liniowym modelem regresyjnym*. Ostatnio zaś coraz częściej stosowane jest określenie „model regresji”. Określenie to z punktu widzenia językowego można traktować jako poprawne tylko w przypadku, gdy przez regresję rozumie się *zjawisko* powrotu do średniej. Po słowie „model” w języku polskim występuje bowiem rzeczownik w dopełniaczu, a więc model samolotu, model człowieka, model burzy itp.

W przypadku gdy błędem losowym obarczone są pomiary tylko jednej cechy, a mianowicie zmiennej zależnej, to metoda najmniejszych kwadratów jest w pewnym sensie najlepszą metodą estymacji nieznanymi parametrów. Sytuacja zmienia się drastycznie, gdy dwie zmienne obarczone są błędami. Niestety różnica między dwiema powyższymi sytuacjami jest często ignorowana w wielu publikacjach.

Zauważmy wyraźnie, że *prawo*, w myśl którego wartości cechy Y powiązane są z wartościami cechy X , w obu przypadkach jest takie samo. W najprostszym przypadku jest ono następujące:

$$y = \alpha_0 + \alpha_1 x.$$

Jeśli pomiary tylko cechy Y obarczone są błędem losowym, to po zastosowaniu metody najmniejszych kwadratów uzyskujemy optymalne estymatory parametrów α_0 i α_1 w klasie wszystkich estymatorów liniowych.

Jeśli pomiary obu cech obarczone są błędami, to metoda najmniejszych kwadratów w ogólnym przypadku w ogóle nie może być stosowana. Metoda ta nie pozwala bowiem wówczas uzyskać nawet estymatorów zgodnych. Metodę znajdowania takich estymatorów podał po raz pierwszy A. Wald w 1940 r. Jest to metoda dwóch punktów (zmodyfikowana później do trzech punktów).

Estymator parametru α_1 ma następującą postać [Wald 1940]:

$$\hat{\alpha}_1 = \frac{(y_1 + \dots + y_m) - (y_{m+1} + y_{m+2} + \dots + y_N)}{(x_1 + \dots + x_m) - (x_{m+1} + x_{m+2} + \dots + x_N)},$$

gdzie $m = N/2$, przy czym przyjęto, że liczba obserwacji N jest liczbą parzystą.

Estymator parametru α_0 ma zaś postać: $\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x}$.

Literatura

- Crombie A.C., *Style myśli naukowej w początkach nowożytnej Europy*, Polska Akademia Nauk, Warszawa 1994.
- Fienberg S.E., *A Brief History of Statistics in Three and One-half Chapters : a Review Essay*, „Statistical Science” 1992 nr 2, 208-225.
- Freedman D., Pisani R., Purves R., *Statistics*, W.W. Norton, 1878.
- Hacking I., *Logic of Statistical Inference*, Cambridge University Press, Cambridge 1965.
- Hacking I., *The Emergence of Probability*, Cambridge University Press, 1975.
- Juszkiewicz A.P. (red.), *Historia matematyki*, tom II, PWN, Warszawa 1976.
- Linnik J.W., *Metoda najmniejszych kwadratów i teoria opracowania obserwacji*, PWN, Warszawa 1962.
- Ostasiewicz W., *Propedeutyka probabilistyki*, AE, Wrocław 2006.
- Płocki A., *Propedeutyka rachunku prawdopodobieństwa i statystyki matematycznej dla nauczycieli*, PWN, Warszawa 1992.
- Shafer G., *The Unity and Diversity of Probability*, „Stat, Sci.” 1990 vol. 5, 435-462.

-
- Spanos A., *Statistical Foundations of Econometric Modeling*, Cambridge University Press, Cambridge 1986.
- Stigler S.M., *Gauss and the Invention of Least Squares*, „The Annals of Statistics” 1981 vol. 9, 465-474.
- von Furstenberg G.M., *Acting under Uncertainty*, Kluwer, London 1990.
- Wald A., *The Fitting of Straight Lines if Both Variables Are Subject to Error*, „The Ann. of Math. Stat.” 1940 vol. II, 284-300.

HISTORICAL REMARKS ON THE DEVELOPMENTS OF STATISTICS

Summary

This paper highlights some basic statistical concepts from the perspective of their historical developments. Special emphasis has been put on these notions which are often confused or even misunderstood. The notion of regression is one of typical example. The meaning of this concept is in great detail explained in this paper.