Wstęp

W dłuższej perspektywie wkład statystyki w rozwój świata nie zależy zbytnio od wykształcenia wielu wysoko wykwalifikowanych statystyków pracujących w przemyśle. Jest to raczej potrzeba stworzenia statystycznie myślącego pokolenia fizyków, chemików, inżynierów i innych, którzy na różne sposoby będą rozwijać gospodarkę jutra.

W.A. Shewhart & W.E. Deming

STATYSTYKA ma długą prehistorię, a krótką historię. Jej pochodzenie można wywodzić od początków ludzkości, ale dopiero w ostatnich czasach okazała się dziedziną o wielkim znaczeniu praktycznym. Czy statystyka jest oddzielną dziedziną wiedzy jak fizyka, chemia, ekonomia czy inne uznane od lat nauki? Nie ma przecież w Polsce tytułów naukowych ze statystyki, można być profesorem matematyki, ale nie statystyki. Matematyk oddaje się dedukowaniu twierdzeń na podstawie danych aksjomatów i reguł wnioskowania. Ekonomista wyjaśnia, co, kiedy i za ile produkować. Fizyk opisuje otaczający nas świat za pomocą praw i reguł nim rządzących, a chemik wyjaśnia zachodzące w świecie reakcje. Każda z tych dziedzin ma swoje potrzeby i własne metody ich rozwiązywania, które nadają im status oddzielnych nauk. Czy istnieją zatem czysto statystyczne problemy, które statystyka mogłaby rozwiązywać? Jeśli nie, to czy jest to rodzaj sztuki, logiki lub techniki stosowanej do rozwiązywania problemów w innych naukach? Przez wiele lat słowo „statystyka" nie było używane ani często, ani poprawnie. Zapatrywano się na nią sceptycznie, niewielu rządowych specjalistów czy pracowników naukowych stosowało jej narzędzia. Inaczej niż dzisiaj, gdy istnieje ogromny rynek pracy dla statystyków – w rządzie, przemyśle, nauce – albowiem niemal zawsze musimy wyciągać wystarczające wnioski z niewystarczających przesłanek. W zasadzie nie ukazują się żadne poważne prace naukowe w wielu dziedzinach, w których nie byłoby metod statystycznych.

Jak przewidywać społeczno-ekonomiczne charakterystyki ludności na podstawie bieżących tendencji? Jak podejmować decyzje sprzyjające wzrostowi dobrobytu społeczeństwa? Czy jutro będzie padać deszcz? Czy ubezpieczyć się na wypadek klęski nieurodzaju, śmierci, katastrofy? Jak łatwo się domyślić, główną przeszkodą w udzieleniu odpowiedzi na te pytania jest niepewność – brak jednoznacznej relacji między przyczyną i skutkiem. Dopiero z początkiem poprzedniego wieku nauczono się wyznaczać niepewność[1]. Dane trzeba przetwarzać, aby dowiedzieć się, do jakiego stopnia można z nich usunąć element niepewności. Wiedza o zasobie niepewności zawartej w danych jest kluczem do podjęcia odpowiedniej decyzji. Statystyka jest więc logiką, za pomocą któ-

[1] Datuje się to od roku 1933, kiedy KOŁMOGOROW wprowadził aksjomatyczną definicję prawdopodobieństwa.

rej można wspiąć się po drabinie od danych do informacji o jeden szczebel wyżej. Statystyka to pewna metodologia podejmowania decyzji, czyli wnioskowania w warunkach niepewności. Wiedza osoby zajmującej się statystyką (czyli statystyka) pozwala zaufać głoszonym przez niego sądom lub podejmowanym decyzjom z uwzględnieniem nieuniknionego ryzyka. Musimy pamiętać, że zawsze istnieje ryzyko porażki, popełnienia błędu, ale zdecydowanie lepiej wiedzieć coś jedynie z pewnym prawdopodobieństwem, niż nie wiedzieć nic z całą pewnością.

Osoby uprawiające statystykę często popełniają różnorakie błędy, spowodowane kilkoma przyczynami:

- Zdecydowana większość ludzi korzystających z metod statystycznych to specjaliści w zupełnie innych dziedzinach, dla których statystyka odgrywa rolę pomocniczą – ekonomiści, biolodzy, chemicy itp.
- Klasyczna teoria statystyki powstawała ponad pół wieku temu i z braku wówczas odpowiednio wydajnych komputerów opiera się na zaawansowanych metodach analitycznych (czytaj: długich i skomplikowanych wzorach) oraz koniecznych do ich wyprowadzenia założeniach, nie zawsze spełnianych w praktyce (raczej rzadko) i często nierozumianych (lub rozumianych błędnie) przez niestatystyków.
- Próba wyjaśnienia tej złożonej teorii na kursie lub w podręczniku dla niestatystyków kończy się zwykle katalogiem przepisów „kiedy stosować który test". Niestety, żaden katalog nie uwzględni wszystkich przypadków, z którymi możemy mieć do czynienia, i nie zastąpi zrozumienia podstaw.
- Podstawową konsekwencją rozpowszechnienia komputerów jest ułatwienie dostępu do tych skomplikowanych metod: z wczytaniem danych do pakietu statystycznego jakoś sobie poradzimy, potem tylko trzeba „doklikać się" do testu i komputer zawsze „wyrzuci" wynik. Komputer jednak nie przyjmie odpowiedzialności za dobór metody do problemu, poprawne sformułowanie hipotezy oraz wyciągnięcie właściwych wniosków.

Te wszystkie przyczyny popełniania błędów powodują, że w społeczeństwie modne jest powiedzenie, że „statystyka kłamie". Nie jest to jednak prawda, dużo bliższe prawdy jest stwierdzenie: „Liczby nie kłamią, ale kłamcy liczą"[2]. Książka ta ma służyć jako podręcznik do nauki statystyki dla początkujących, aby nie stali się takimi „kłamcami". Znajduje się w niej wiele uwag, które w kluczowych momentach przestrzegają przed popełnianiem powszechnych błędów. Do zrozumienia materiału wymagana jest jedynie minimalna wiedza matematyczna, raczej niewykraczająca zakresem poza materiał szkoły średniej, a w wielu miejscach nawet gimnazjum. Część trudniejszych problemów została jedynie zarysowana (pominięto niemal wszystkie wyprowadzenia wzorów) w celu lepszego zrozumienia całości materiału (materiał dodatkowy oznaczono gwiazdką). Wszystkie omawiane techniki są bogato ilustrowane

[2]Słowa wypowiedziane przez C.H. Grosvenora (1833–1917), amerykańskiego pułkownika podczas wojny domowej w USA w latach 1860–1865.

przykładami. Szczególną uwagę zwrócono na wizualizację metod statystycznych za pomocą wykresów i tabel.

Ponieważ obecnie większość metod statystycznych jest dość żmudna obliczeniowo, wszystkie przykłady zostały przeprowadzone również z użyciem pakietu statystycznego R[3], który zyskał obecnie ogromną popularność na świecie. Wszystkie przytoczone w książce funkcje, pakiety oraz zbiory danych zebrano w odpowiednich indeksach na końcu książki. Wszystkie zbiory opisane są przy pierwszym wystąpieniu, przy kolejnych wzmiankowany jest jedynie pakiet, z którego pochodzą. W dobie Internetu (wyszukiwania informacji) niezbędna jest znajomość języka angielskiego również w statystyce, z tego względu wszystkie istotne pojęcia podano również w tym języku. Wyszukanie pojęć ułatwiają indeksy pojęć w języku polskim i angielskim.

Każdy rozdział zakończony jest zadaniami o zróżnicowanym poziomie trudności (trudniejsze zadania oznaczone są gwiazdką), które pozwalają lepiej zrozumieć oraz utrwalić materiał. Część zadań wymaga jedynie kartki oraz czegoś do pisania, natomiast znaczna część została przewidziana do rozwiązania za pomocą R. W przypadku odpowiedzi używane zbiory nie są dołączane (**attach**), natomiast używane są, jakby były dołączone.

Podręcznik ten powstał na podstawie prowadzonych przeze mnie od wielu lat zajęć na Uniwersytecie im. Adama Mickiewicza w Poznaniu.

Na koniec chciałbym podziękować Panu Doktorowi Maciejowi Łuczakowi, bez którego kształt tej książki byłby zupełnie inny.

[3] *http://www.r-project.org/*