Ryszard Zieliński

Statystyka matematyczna stosowana

Elementy

Mirosław Karpierz
Instytut Matematyczny
Polska Akademia Nauk
Śniadeckich 8
00-956 Warszawa
E-mail: R.Zielinski@impan.pl

Korekta: **Henryka Walas**

Redaktor merytoryczny: **Stanisław Janeczko**

Skład redakcji: **Małgorzata Zielińska, Anna Żubrowska**

Projekt graficzny i skład okładki: **Emilia Bojańczyk / Podpunkt**

Informacje o innych wydawnictwach tej serii dostępne pod adresem www.csz.pw.edu.pl

ISBN: 978-83-61993-03-2
Wydrukowano w Polsce

Spis treści

Not everything that can be counted counts,
and
not everything that counts can be counted.

Albert Einstein

1. Wprowadzenie

Odkąd komputery zbłądziły pod strzechy, także w statystyce można więcej uwagi przeznaczyć na myślenie niż na liczenie. Prezentowany wykład nie jest jeszcze jednym podręcznikiem statystyki matematycznej; jest już ich dostatecznie dużo — w Literaturze są wymienione te najnowsze i najłatwiej dostępne w języku polskim. W naszym wykładzie będziemy mówili o sprawach, które w tych podręcznikach rzadko są poruszane.

Gdy będziemy mówili o konstrukcji testu do weryfikacji hipotezy statystycznej, zasadniczą rolę będziemy przypisywali hipotezie alternatywnej. Już na początku wprowadzimy pojęcie *mocy testu statystycznego* i pokażemy, jak rzut oka na funkcję mocy pozwala wybrać, na przykład, odpowiedni test Studenta (test prawostronny, lewostronny lub dwustronny). Rozważając estymatory punktowe natychmiast będziemy podnosili kwestię oceny ich dokładności; za odpowiednie do tego narzędzie wybierzemy *przedziały ufności*. Żeby nie zaciemniać wykładu bardziej złożonymi modelami statystycznymi, dla demonstracji najważniejszych idei statystyki matematycznej zajmiemy się prostym i intuicyjnie dla każdego oczywistym *statystycznym modelem pomiaru*. Rozumiemy przez to następującą sytuację. Mamy zmierzyć pewną nieznaną wielkość liczbową μ. Wykonując pomiar otrzymujemy wynik X, który różni się od μ o błąd losowy ε; mamy więc wynik $X = \mu + \varepsilon$ i na tej podstawie chcemy odpowiedzieć na różne pytania dotyczące μ. Sytuacja jest bardzo ogólna. W tym schemacie mieści się zarówno pomiar długości lub ciężaru jakiegoś fizycznego obiektu, jak też przeciętny czas życia lub przeciętna cena danego papieru na giełdzie; wówczas „błąd losowy" ε jest po prostu odchyłką wielkości indywidualnego obiektu od wielkości przeciętnej (na przykład średniej lub mediany) dla całej populacji.

Rutynowe postępowanie polega na wielokrotnym powtórzeniu pomiaru X i wnioskowaniu o nieznanym parametrze μ na podstawie jakiegoś uśrednienia otrzymanych wyników pomiarów. Liczbę powtórzeń pomiarów w całym wykładzie będziemy oznaczali przez n, a poszczególne pomiary przez $X_1, \ldots, X_n$. Mamy więc obserwacje

$$X_j = \mu + \varepsilon_j, \quad j = 1, \ldots, n,$$

gdzie $\varepsilon_1,\ldots,\varepsilon_n$ jest ciągiem niezależnych błędów losowych o takim samym rozkładzie, powiedzmy o dystrybuancie F. Jeżeli o rozkładzie F nic nie wiemy, to oczywiście na podstawie obserwacji $X_1,\ldots,X_n$ nic nie potrafimy powiedzieć o mierzonej wielkości μ. W odpowiednim miejscu powiemy o sytuacji, w której zwielokrotnianie liczby n pomiarów nie ma sensu lub nawet może psuć wyniki wnioskowania.

W wykładzie omawiamy cztery, coraz bardziej ogólne modele statystyczne pomiaru. W pierwszym z nich rozważamy wnioskowanie o parametrze μ, gdy wiadomo, że błąd losowy ε ma standardowy rozkład normalny o średniej równej zeru i znanej wariancji σ^2 (znanym odchyleniu standardowym σ). Oznaczamy ten rozkład przez $N(0,\sigma)$. Odchylenie standardowe σ reprezentuje tu dokładność pomiaru (rozrzut, rozproszenie wokół wartości μ, precyzję przyrządu pomiarowego). Jest to elementarna część wykładu, raczej wszystkim znana, ale bardzo istotna, bo właśnie tu, na tym elementarnym technicznie poziomie, wprowadzamy oznaczenia i pojęcia, którymi będziemy się później intensywnie posługiwać. Drugi model różni się tylko tym od pierwszego, że odchylenie standardowe σ nie jest znane i dla wnioskowania o wielkości μ będziemy najpierw musieli oszacować σ na podstawie obserwacji $X_1,\ldots,X_n$. Trzeci z rozważanych dalej modeli to model ze znanym rozkładem F, już niekoniecznie rozkładem normalnym. Motywacja wynika ze współczesnych zastosowań statystyki matematycznej w finansach, ubezpieczeniach i ekologii; występują tu rozkłady, w których prawdopodobieństwo pojawienia się bardzo dużych obserwacji jest znacznie większe niż jest to realnie możliwe w modelach gaussowskich. Rozkłady te mogą nie mieć wartości oczekiwanej, a średnia arytmetyczna obserwacji może być bardziej rozproszona wokół estymowanej wielkości μ niż pojedyncza obserwacja. W takich modelach posługiwanie się standardowymi miarami, takimi jak średnia i wariancja, nie znajduje żadnego uzasadnienia. Czwarty model to model z nieznanym rozkładem F. W tym przypadku o rozkładzie F będziemy zakładali tylko tyle, że jest to rozkład z ciągłą i ściśle rosnącą dystrybuantą.

To, co do tej pory mówiliśmy o rozkładzie prawdopodobieństwa, dotyczy tylko rozkładu prawdopodobieństwa błędu losowego ε. Faktycznie interesuje nas rozkład prawdopodobieństwa obserwacji X, a ten różni się od rozkładu F błędu ε „przesunięciem" μ. Rozkład obserwacji X będziemy oznaczali przez F_μ. W rozważanym modelu mamy $F_\mu(x) = F(x-\mu)$. Wielkości μ nie znamy — właśnie mamy ją oszacować. Wobec tego rozkład prawdopodobieństwa obserwacji X także nie jest znany. Jeżeli interesuje nas jakieś zdarzenie losowe związane z obserwacją X, na przykład zdarzenie $\{X > 0\}$ polegające na tym, że w wyniku pomiaru otrzymamy wartość dodatnią, to na pytanie o prawdopodobieństwo tego zdarzenia nie ma jednoznacznej odpowiedzi, dopóki nie sprecyzujemy wielkości μ. Formalnie ujmujemy to zapisem $P_\mu\{X > 0\}$, pamiętając o tym, że w statystyce matematycznej nie ma jednego prawdopodobieństwa; w danym

przypadku jest ich tyle, ile różnych wartości μ w naszym modelu bierzemy pod uwagę. Mówiąc ogólnie, w statystyce matematycznej mamy całą rodziną prawdopodobieństw i trzeba zawsze wyjaśnić, z którym (z którymi) z nich mamy właśnie do czynienia. Służy temu dodatkowy indeks przy symbolu P lub, gdy mówimy o wartościach oczekiwanych, przy symbolu E. Indeks staje się niepotrzebny, gdy wiadomo, z jakim rozkładem mamy do czynienia. Na przykład, gdy mówimy o prawdopodobieństwie zdarzenia polegającego na tym, że zmienna losowa chi-kwadrat o ν stopniach swobody (oznaczamy ją przez χ^2_ν) nie przekroczy x, napiszemy $P\{\chi^2_\nu \leq x\}$, bo ta zmienna ma jednoznacznie określony rozkład (patrz rozdz. 2.2.2).

Najważniejszy w całym wykładzie jest rozdział 2.1. Wprowadzamy w nim i komentujemy wszystkie potrzebne później pojęcia (estymator, przedział ufności, test statystyczny, moc testu, poziom krytyczny testu). W istocie rzeczy w dalszej części wykładu powtarzamy rozumowania z tego rozdziału dla coraz bardziej złożonych (i bliższych życiu) modeli statystycznych.

W wykładzie nie ma dużo rachunków. Rozkład normalny, rozkład beta, niecentralny rozkład t Studenta i in. traktujemy jak pojęcia, dla których w licznych i łatwo dostępnych pakietach komputerowych istnieją oprogramowania za pomocą funkcji standardowych danego pakietu. Jeżeli mówimy, na przykład, że $B(x; p, q)$ jest wartością dystrybuanty rozkładu beta o parametrach p, q w punkcie x, to zakładamy, że Czytelnik w razie potrzeby potrafi dowiedzieć się od swojego komputera, ile to wynosi dla danych x, p, q.

Zakładamy, że Czytelnik zna podstawowe fakty z teorii prawdopodobieństwa i statystyki na poziomie przynajmniej jednej z książek wymienionych w Literaturze. Na przykład w rozdz. 2.1.2 piszemy, bez żadnego dodatkowego komentarza: „ponieważ obserwacja X ma rozkład $N(\mu, \sigma)$, to średnia $\bar{X}_n$ jest zmienną losową o rozkładzie normalnym $N(\mu, \sigma/\sqrt{n})$, czyli $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ jest zmienną losową o rozkładzie normalnym $N(0, 1)$ i dla dowolnie wybranego $\gamma \in (0, 1)$ mamy

$$P_\mu\{|\sqrt{n}(\bar{X}_n - \mu)/\sigma| \leq z_{(1+\gamma)/2}\} = \gamma,$$

gdzie z_α jest kwantylem rzędu α rozkładu normalnego $N(0, 1)$". Zakładamy, że żadne dodatkowe wyjaśnianie nie jest potrzebne.