

ADAM ŁOMNICKI

WYDANIE PIĄTE

WPROWADZENIE DO STATYSTYKI DLA PRZYRODNIKÓW



 PWN

WPROWADZENIE DO STATYSTYKI DLA PRZYRODNIKÓW

WYDANIE PIĄTE

ADAM ŁOMNICKI

WPROWADZENIE DO STATYSTYKI DLA PRZYRODNIKÓW



 PWN

[Kup książkę](#)

Projekt okładki i stron tytułowych **Michał Rosiński**

Redaktor **Krystyna Mostowik**

Czytelnicy poszukujący dodatkowych ćwiczeń, ewentualnych uzupełnień tej książki oraz kontaktu z jej autorem znajdą je w internecie na stronie **<http://www.eko.uj.edu.pl/lomnicki/statystyka>**

Książka, którą nabyłeś, jest dziełem twórcy i wydawcy. Prosimy, abyś przestrzegał praw, jakie im przysługują. Jej zawartość możesz udostępnić nieodpłatnie osobom bliskim lub osobiście znanym. Ale nie publikuj jej w internecie. Jeśli cytujesz jej fragmenty, nie zmieniaj ich treści i koniecznie zaznacz, czyje to dzieło. A kopiując jej część, rób to jedynie na użytek osobisty.

Szanujmy cudzą własność i prawo

Więcej na www.legalnakultura.pl

Polska Izba Książki

Copyright © by Wydawnictwo Naukowe PWN Sp. z o.o.
Warszawa 1995

Copyright © by Wydawnictwo Naukowe PWN SA
Warszawa 1999, 2003, 2010, 2014

ISBN 978-83-01-17948-9

Wydanie piąte poprawione
Warszawa 2014

Wydawnictwo Naukowe PWN SA
02-460 Warszawa, ul. Gottlieba Daimlera 2
tel. 22 69 54 321; faks 22 69 54 228
infolinia 801 33 33 88
e-mail: pwn@pwn.com.pl; www.pwn.pl
Druk i oprawa: OSDW Azymut Sp z o.o

Kup książkę

Spis treści

1. Wstęp	9
1.1. Do czego jest potrzebna statystyka?	9
1.2. Próba i populacja generalna, statystyka elementarna i statystyka matematyczna	14
1.3. Po co jeszcze jeden podręcznik statystyki?	16
2. Statystyka elementarna: pomiary w biologii	20
2.1. Skala nominalna, porządkowa i interwałowa	20
2.2. Procenty, proporcje, stosunki	22
2.3. Dokładność pomiarów i zaokrąglanie liczb	23
2.4. Graficzne przedstawienie danych i rozkłady frekwencji	26
2.5. Kodowanie i transformacja danych	30
Ćwiczenia	31
3. Statystyka elementarna: miary tendencji centralnej i rozproszenia	33
3.1. Średnia arytmetyczna, ważona, geometryczna i harmoniczna	33
3.2. Mediana i wartość modalna	38
3.3. Wariancja i odchylenie standardowe	40
3.4. Estymatory obciążone i nieobciążone	43
3.5. Współczynnik zmienności	46
Ćwiczenia	47
4. Wprowadzenie do statystyki matematycznej: rachunek prawdopodobieństwa i testowanie hipotez	49
4.1. Rachunek prawdopodobieństwa, dodawanie prawdopodobieństw zdarzeń wykluczających się	49
4.2. Mnożenie prawdopodobieństw zdarzeń niezależnych	51
4.3. Rozkład dwumianowy	57
4.4. Testowanie hipotez	60

4.5. Test znaków	68
Ćwiczenia	70
5. Rozkład normalny i test dla proporcji	72
5.1. Średnia i wariancja dla proporcji	72
5.2. Rozkład normalny	76
5.3. Standaryzacja pomiarów	79
5.4. Praktyczne posługiwanie się rozkładem normalnym	80
5.5. Test dla proporcji	84
5.6. Analiza mocy testu statystycznego	88
Ćwiczenia	89
6. Rozkład średnich z prób i przedziały ufności	92
6.1. Rozkład średnich z prób i błąd standardowy	92
6.2. Przedziały ufności dla średniej z dużej próby	96
6.3. Rozkład t	97
6.4. Przedziały ufności dla średniej	99
Ćwiczenia	102
7. Różnice między średnimi	104
7.1. Planowanie doświadczeń: zabieg i jego kontrola	104
7.2. Test różnicy między średnimi dla par wiązanych	108
7.3. Test różnic między średnimi	110
7.4. Różnice w mocy testów statystycznych	116
Ćwiczenia	117
8. Zastosowanie liczb losowych i wybór próby na powierzchni	118
8.1. Liczby losowe	118
8.2. Losowanie prób na powierzchni	120
8.3. Próby systematyczne	124
Ćwiczenia	126
9. Testy nieparametryczne dla różnic między dwoma próbami	128
9.1. Wstęp	128
9.2. Test Wilcoxona dla par wiązanych	129
9.3. Test serii Walda–Wolfowitza	132
9.4. Test U (Manna–Whitneya)	134
9.5. Test Kołmogorowa–Smirnowa	137
Ćwiczenia	139
10. Analiza wariancji: wprowadzenie	141
10.1. Wariancja wewnątrz grup i między grupami	142
10.2. Rozkład F	146
10.3. Sumy kwadratów odchyleń	149
10.4. Dwa modele	151
10.5. Założenia analizy wariancji	153
Ćwiczenia	155

11. Analiza wariancji: klasyfikacja prosta i dwukierunkowa	156
11.1. Klasyfikacja prosta	156
11.2. Porównania zaplanowane: kontrasty	161
11.3. Test <i>a posteriori</i>	163
11.4. Nieparametryczna alternatywa klasyfikacji prostej	166
11.5. Klasyfikacja dwukierunkowa	169
11.6. Interakcja i zagadnienia z nią związane	177
11.7. Ortogonalność i jej brak: niejednakowe liczebności w grupach i brak niektórych grup	179
11.8. Określenie składnika błędu w analizie wariancji przy klasyfikacji dwukierunkowej z powtórzeniami	180
11.9. Klasyfikacja dwukierunkowa bez powtórzeń	183
11.10. Hierarchiczna analiza wariancji	185
Ćwiczenia	189
12. Doświadczalnictwo rolnicze	192
12.1. Proste doświadczenie systematyczne	192
12.2. Proste doświadczenie losowe	194
12.3. Bloki losowe	195
12.4. Inne plany eksperymentów	197
Ćwiczenia	197
13. Analiza frekwencji	198
13.1. Test zgodności	199
13.2. Założenia i ograniczenia testu chi kwadrat	201
13.3. Porównanie szeregu rozdzielczego z rozkładem normalnym	202
13.4. Zastosowanie testu chi kwadrat do badania związku między dwoma skalami nominalnymi	205
Ćwiczenia	210
14. Korelacja i regresja	212
14.1. Szereg dwucechowy	212
14.2. Dwa modele szeregu dwucechowego	214
14.3. Dopasowanie linii prostej do danych empirycznych: regresja i oś główna zredukowana	217
14.4. Obliczenia współczynników regresji, osi głównej zredukowanej i korelacji	220
14.5. Statystyczna istotność regresji i korelacji	223
14.6. Opracowanie statystyczne szeregu dwucechowego w dwóch różnych modelach	227
14.7. Nieliniowość związku i transformacje szeregu dwucechowego	229
14.8. Nieparametryczna alternatywa współczynnika korelacji	233
14.9. Inne techniki statystyczne związane z analizą szeregów wielocechowych	235
Ćwiczenia	239

15. Analiza kowariancji	240
15.1. Wstępne uwagi i obliczenia	243
15.2. Analiza zmienności między grupami	247
15.3. Test jednorodności nachyleń prostych regresji między grupami ...	249
16. Wielokrotne porównania i ich testy	252
Piśmiennictwo zalecane i wspomniane w tekście	256
Rozwiązanie ćwiczeń	257
Tablice statystyczne	265
Skorowidz	279

1.1. Do czego jest potrzebna statystyka?

Jedną z najbardziej zauważalnych cech otaczającego nas świata jest zmienność. Cokolwiek usiłujemy pomierzyć: długość liści grabu, ilość siana zebranego z jednego ara łąki, masę dorosłych samców nornika burego, rozstaw rogów kozicy, długość życia samic muszek owocowych w laboratorium, zawartość tłuszczu w orzechach laskowych lub liczbę nasion w kłosie zboża, trudno znaleźć dwa pomiary dokładnie identyczne. Fakt, że koniczyna jest zwykle trójlistna, a czterolistna jest wyjątkiem i że ssaki mają dwoje oczu, a nie troje lub jedno, nie oznacza, iż w naukach przyrodniczych możemy się ograniczyć jedynie do badania cech o niezmiennych wymiarach. Przeciwnie, nawet w takich dziedzinach, jak systematyka roślin i zwierząt, która poszukuje głównie cech mało zmiennych, cechy ilościowe, z wartościami różnymi dla różnych osobników, są także ważnymi cechami systematycznymi.

Biolog spotykający się ze zmiennością stara się określić jej zakres. Na przykład w kluczu do oznaczania ptaków określa się zwykle zakres liczby składanych jaj podając, że sikorka bogatka składa ich od 6 do 14, a sikora modra od 7 do 12. Określając zakres, podawać też można masę główki kapusty lub wzrost mężczyzn w Polsce w wieku 18 lat.

Czy zakres jest najlepszym sposobem opisanie serii zmiennych pomiarów? Odpowiedź nie jest prosta. Po pierwsze, zakres nie mówi

nam, które pomiary są najczęstsze. Po drugie, zakres wzrasta wraz z liczbą pomiarów. Jeśli pomierzmy wzrost wszystkich studentów z jednego roku biologii, to istnieje niewielka szansa, że znajdziemy wśród nich osoby niższe niż 150 cm i wyższe niż 200 cm. Takich bardzo niskich i bardzo wysokich studentów możemy napotkać, jeśli nie na całym Uniwersytecie, to na pewno we wszystkich szkołach wyższych w Krakowie, natomiast na nadzwyczajnego karła i olbrzyma natrafimy po sprawdzeniu wzrostu mężczyzn w całej Polsce. Zakresy wzrostu i masy ciała ludzi oraz wielu innych pomiarów dla całego świata znajdziemy w księdze Guinnessa i chcąc być konsekwentym, zakresy te należałoby brać z tej książki. Zwykle nie postępujemy w ten sposób, ale raczej zawężamy zakresy do tych częściej spotykanych. Podobnie, opisując koniczynę w kluczu do oznaczania roślin, pomijamy interesujące zjawisko, że bardzo rzadko, ale można spotkać koniczynę czterolistną.

Oprócz podawania zakresu, innym sposobem postępowania ze zmiennością jest obliczenie średniej wartości dla rozważanych pomiarów. Tutaj powstaje problem: z ilu pomiarów trzeba obliczyć średnią. Jeśli istnieje zmienność, to najczęściej średnia z dwóch pomiarów różni się od każdego z nich, a średnia z trzech pomiarów różni się od średniej z dwóch. Gdy zwiększamy liczbę pomiarów, na których oparta jest średnia, otrzymujemy za każdym razem nieco inną jej wartość (tab. 1.1). Wraz ze wzrostem liczby pomiarów wahania średniej są coraz mniejsze, ale nigdy nie mamy gwarancji, czy dodanie dalszych pomiarów nie zmieni jej wartości.

Po tym wprowadzeniu można zadać pytanie: ile pomiarów powinien wykonać przyrodnik, aby uzyskać średnią, do której może mieć zaufanie, a następnie jak oszacować stopień tego zaufania. Można też zadać następne pytanie: jaka część pomiarów mieści się w założonym z góry zakresie wokół średniej lub jak częste są pomiary znacznie od niej odbiegające. Na takie właśnie pytania może odpowiedzieć statystyka.

Jakie jeszcze inne problemy można rozwiązać z pomocą statystyki? Wyobraźmy sobie, że chcemy ustalić, czy mysz domowa i nornica ruda różnią się zdolnościami uczenia się drogi w labiryncie. Każdy przyrodnik intuicyjnie czuje, że zbadanie jednej myszy i jednej nornicy niewiele daje. Wyobraźmy sobie, że mysz, poruszając się po raz drugi po tym samym labiryncie, popełniła dwa błędy, zaś nornica dziesięć

Tabela 1.1. Masy ciała poczwerek mącznika *Tenebrio molitor* (Coleoptera, Tenebrionidae) w miligramach, z hodowli prowadzonej na razowej mące pszennej w temperaturze 26 °C. Średnie podane w tabeli są obliczone ze wszystkich pomiarów masy w wierszach powyżej oraz w tym wierszu, w którym przedstawiono średnią. To samo dotyczy zakresów. Tak na przykład średnia i zakres w wierszu 7 jest średnią i zakresem pierwszych siedmiu pomiarów. Średnia w ostatnim wierszu jest konsekwentnie średnią ze wszystkich pomiarów

Numer pomiaru	Masa ciała	Średnia	Zakres
1	148	–	–
2	148	148,0	148 – 148
3	136	144,0	136 – ~
4	152	145,0	~ – 152
5	142	145,2	~ – ~
6	130	142,7	130 – ~
7	176	147,4	~ – 176
8	152	148,0	~ – ~
9	150	148,2	~ – ~
10	140	147,0	~ – ~
11	123	145,2	123 – ~
12	113	142,5	113 – ~
13	133	141,8	~ – ~
14	117	140,0	~ – ~
15	126	139,1	~ – ~
16	129	138,4	~ – ~
17	219	143,2	~ – 219
18	156	143,9	~ – ~
19	160	144,7	~ – ~
20	123	143,6	~ – ~

błędów. Nawet przy tak dużej różnicy można się obawiać, że wynika to ze zmienności osobniczej, a nie międzygatunkowej. Dlatego założymy, że badamy po pięć dorosłych dwumiesięcznych samców każdego gatunku i liczymy popełniane przez nie błędy. Te zastrzeżenia o wieku i płci naszych gryzoni czynimy po to, aby nie badać różnic między starymi i młodymi, samcami i samicami, lecz tylko różnice międzygatunkowe. Trzy różne hipotetyczne wyniki opisanych tu badań przedstawiono w tabeli 1.2. Popatrzmy na pierwszy wynik (I). Każdy znający się na rzeczy przyrodnik powie nam bez żadnej statystyki, że myszy dają sobie radę w labiryncie znacznie lepiej niż nornice.

Taki wynik jest oczywisty także bez opracowania statystycznego, ponieważ nornice popełniają pięciokrotnie więcej błędów niż myszy i najmądrzejsza nornica jest w labiryncie znacznie głupsza od najgłupszej myszy. Spójrzmy teraz na drugi wynik (II). Tu średnie są identyczne, podobnie jak i zakresy, dlatego nic nie wskazuje, że myszy różnią się od nornic liczbą błędów popełnianych w labiryncie. Gdyby wszystkie nasze wyniki badań były tak jednoznaczne, jak te dwa przedstawione powyżej, świat byłby prosty, a bez statystyki można by się było obejść. Tak jednak nie jest, a nasze wyniki podobne są najczęściej do wyniku trzeciego (III) z tabeli 1.2. Tutaj myszy popełniły mniej błędów niż nornice, ale różnica jest niewielka i zakresy zachodzą na siebie. Czy na podstawie tego wyniku można powiedzieć, że między tymi dwoma gatunkami istnieje różnica, czy też jest to czysty przypadek, że średnie dla tych dwóch gatunków różnią się od siebie? Na takie właśnie pytanie odpowiedź daje statystyka.

Tabela 1.2. Trzy hipotetyczne wyniki (I, II, III) przedstawiające liczbę błędów popełnionych w labiryncie przez 5 myszy domowych i 5 nornic rudych. Dla każdego z gatunków i każdego z wyników podano średnią liczbę błędów

	Mysz domowa		Nornica ruda	
	liczba błędów	średnia	liczba błędów	średnia
I	2, 3, 1, 2, 1	1,8	10, 8, 9, 11, 8	9,2
II	2, 1, 1, 2, 3	1,8	3, 3, 1, 1, 1	1,8
III	1, 1, 2, 3, 2	1,8	4, 2, 4, 1, 1	2,4

Opiszmy trzeci przypadek zastosowania statystyki. Wyobraźmy sobie, na przykład, genetyka, który, badając proste dziedziczenie mendlowskie jednej cechy u muszki owocowej, krzyżował homozygotę z heterozygotą, a następnie badał liczbę homozygot i heterozygot u 100 potomków z tej krzyżówki. Gdyby otrzymał on 50 homozygot i 50 heterozygot, oznaczałoby to, że jego wynik doskonale zgadza się z teorią i nie ujawniają się tam żadne komplikacje zakłócające proste dziedziczenie mendlowskie. Można by nawet podejrzewać, że wynik ten jest zbyt dobry i że nasz genetyk trochę go „poprawił”, ponieważ rzadko się zdarza, aby rzucając 100 razy monetą otrzymać za pierwszymi stu rzutami dokładnie 50 razy orła i 50 razy reszkę. Mamy tu do czynienia ze ścisłą analogią, ponieważ przewidywane proporcje homozygot

i heterozygot wynikają z losowego dobierania się gamet. Gdyby nasz genetyk otrzymał inny wynik, na przykład 10 homozygot i 90 heterozygot, to też sprawa byłaby jasna. Wynik taki wskazuje, że nie jest to normalne dziedziczenie mendlowskie i że zachodzą tam jakieś dodatkowe komplikacje genetyczne. W obu tych przypadkach uzyskane wyniki są oczywiście bez analizy statystycznej. Co jednak należy robić, gdy wynik powyższego eksperymentu nie jest tak oczywisty? Wydaje się, że wynik 49 homozygot i 51 heterozygot jest jeszcze wystarczający do potwierdzenia teorii, ale czy 45 homozygot i 55 heterozygot lub 40 homozygot i 60 heterozygot też ją potwierdza? To jest następne pytanie, na które odpowiada statystyka. Jest to zresztą problem, z którym zetknął się sam Grzegorz Mendel, gdy badane przez niego proporcje lekko odbiegały od teoretycznie przewidywanych.

Na zakończenie ostatni przykład. Wyobraźmy sobie, że w sześciu oddziałach pewnego kompleksu lasów ustalono, ile w każdym z nich jest gatunków drzew i gatunków ptaków. Załóżmy, że otrzymano następujące wyniki:

oddział	A	B	C	D	E	F
liczba gatunków drzew	2	5	5	8	9	10
liczba gatunków ptaków	3	2	8	6	9	8

Na pierwszy rzut oka wydaje się, że im więcej gatunków drzew, tym więcej gatunków ptaków, czyli że między tymi dwoma zestawami danych jest jakiś związek. Można by powiedzieć, że im więcej gatunków drzew, tym więcej różnych nisz ekologicznych i tym więcej mieści się tam gatunków ptaków. Być może w układzie tym działa jeszcze trzeci czynnik (na przykład zanieczyszczenia, o różnym nasileniu w różnych oddziałach), który wpływa ujemnie lub dodatnio zarówno na liczbę gatunków drzew, jak i ptaków. Krytyk takiej interpretacji danych mógłby jednak powiedzieć, że posuwamy się za daleko doszukując się tu jakiegokolwiek związku. Po prostu tak się przez przypadek złożyło, ale w rzeczywistości zwiększenie liczby gatunków drzew nie prowadzi do zwiększenia liczby gatunków ptaków. W oddziale B jest 5 gatunków drzew, a 2 gatunki ptaków, czyli mniej ptaków niż w oddziale A. Podobna niezgodność występuje w oddziałach E i F. Kto ma rację: badacz doszukujący się tu związku, czy jego krytyk? To także pytanie dla statystyki.

1.2. Próba i populacja generalna, statystyka elementarna i statystyka matematyczna

Przypomnijmy sobie podany w tabeli 1.1 przykład z masą ciała dwudziestu poczwerek mącznika. Tak naprawdę, to nikogo nie interesuje tych konkretnych dwadzieścia poczwerek, które kiedyś wyhodowano i zważono, a potem wyrzucono. Nikt nie zajmowałby się opracowaniem statystycznym tych danych i publikowaniem wyników, gdyby chodziło tylko o te poczwarki. W rzeczywistości, obliczając średnią masę ciała tychże, pragniemy ustalić masę wszystkich poczwerek mącznika *Tenebrio molitor*, gdyby hodowano je w opisanych tu warunkach. W języku statystyki mówimy, że badamy **próbę**, aby dowiedzieć się czegoś o **populacji generalnej**. W naszym przypadku próba to dwadzieścia pomiarów masy ciała poczwerek, natomiast populacja generalna to masy ciała wszystkich poczwerek tego gatunku, gdyby hodowano je w opisanych tu warunkach.

Populacja w sensie statystycznym jest czymś innym niż populacja w sensie ekologicznym lub populacja mendlowska w genetyce populacyjnej. Jeśli interesuje nas długość ogona myszy wielkookiej leśnej w Palearktyce, to naszą populacją generalną nie jest populacja tego gatunku myszy, ale populacja długości ogonów wszystkich myszy w Palearktyce.

Z logiki formalnej wiemy, że rozumowanie przez analogię jest zawodne. Jeśli pierwszy, drugi, a następnie dalsze, dziesiąty lub setny wróbel jest szary, to nie znaczy, że wszystkie wróble na świecie są szare. Może się trafić albinos. Podobnie, jeśli 20 poczwerek mącznika ma średnią masę 143,6 mg, to nie znaczy, że wszystkie poczwarki utrzymywane w warunkach opisanych w tabeli 1.1 mają taką samą średnią masę. Aby dowiedzieć się, jaką masę mają naprawdę, trzeba by hodować i ważyć wszystkie na świecie poczwarki tego gatunku, czyli populację generalną. Statystyka pozwala nam zdobyć pewną informację o całej populacji generalnej na podstawie jednej, odpowiednio dobranej próby. Statystyka jest zatem jednym wielkim fortelom (podobnym fortelom pana Zagłoby), wymyślonym dla ludzi leniwych, aby na podstawie niewielkiej próby dowiedzieć się czegoś o całej populacji.

Możemy obliczać nie tylko średnią pomiarów z próby, ale także zmienność w obrębie tej próby lub jakieś inne miary próby tę charakteryzujące. Jeśli dotyczą one próby, mówimy wówczas, że obliczamy