2.4. Podstawowe metody statystyczne w epidemiologii

Daniel Rabczenko, Bogdan Wojtyniak

2.4.1. Wstęp

„*Statystyka* jest nauką i sztuką gromadzenia, zwięzłego przedstawiania i analizowania danych, które podlegają zmienności losowej"[33].

Biostatystyka jest to statystyka stosowana w zagadnieniach związanych ze zdrowiem populacji, w tym szczególnie w badaniach epidemiologicznych w zdrowiu publicznym i medycynie.

Efektywne wykorzystanie danych było zawsze istotnym składnikiem działań w dziedzinie zdrowia publicznego. Jego znaczenie wzrosło niesłychanie w ostatnich latach, gdy dąży się do tego, by wszystkie działania były oparte na faktach i miały solidne podstawy naukowe.

Trudno sobie obecnie wyobrazić badania dotyczące problemów zdrowotnych ludności, które byłyby prowadzone bez znajomości podstaw demografii oraz stosowania odpowiedniego aparatu statystycznego. Statystyka jest podstawowym narzędziem do oceny wyników badań epidemiologicznych. Pozwala na obiektywizację oceny i interpretacji uzyskanych danych. Przedmiotem niniejszego rozdziału są podstawowe, najczęściej stosowane w epidemiologii metody analizy statystycznej.

Większość przykładów w niniejszym rozdziale powstała na podstawie zbioru danych zawierającego pomiary ciśnienia skurczowego i rozkurczowego grupy 100 osób. Dostępne w nim są również płeć i wiek badanych, a także informacja o dodawaniu soli do potraw. Zbiór jest częścią pakietu *epicalc* wchodzącego w skład statystycznego programu R for Windows (http://r.meteo.uni.wroc.pl/web/packages/epitools/index.html).

Dane w formie tabelarycznej podane są poniżej (tab. 2.9).

Tabela 2.9. Zbiór danych dotyczących ciśnienia krwi i stosowania soli

Nr	Płeć	Skurczowe ciśnienie tętnicze	Rozkurczowe ciśnienie tętnicze	Solone (potrawy)	Data urodzin	Wiek
1	Mężczyzna	110	80	tak	1943-01-06	58
2	Kobieta	85	55	nie	1969-01-03	32
3	Mężczyzna	167	112	tak	1933-06-10	68
4	Kobieta	145	110	tak	1946-11-23	55
5	Kobieta	180	120	nie	1941-01-03	60

[33] Last J.M.: *A Dictionary of Epidemiology, fourth edition*. Oxford University Press, Oxford 2001.

Tabela 2.9. cd.

6	Mężczyzna	112	78	nie	1942-04-16	59
7	Kobieta	110	70	nie	1969-01-11	32
8	Mężczyzna	198	119	tak	1938-07-25	63
9	Mężczyzna	171	102	tak	1943-04-24	58
10	Kobieta	133	75	tak	1955-09-22	46
11	Kobieta	150	72	tak	1938-12-16	63
12	Kobieta	178	128	tak	1938-02-07	63
13	Mężczyzna	118	72	nie	1949-03-28	52
14	Mężczyzna	140	90	nie	1934-09-17	67
15	Mężczyzna	192	118	nie	1959-02-12	42
16	Kobieta	126	78	nie	1955-08-29	46
17	Kobieta	182	96	tak	1941-10-08	60
18	Kobieta	132	80	nie	1975-12-08	26
19	Kobieta	148	94	nie	1953-06-18	48
20	Kobieta	128	84	nie	1937-05-14	64
21	Kobieta	202	118	tak	1945-03-25	56
22	Kobieta	184	148	tak	1940-03-30	61
23	Mężczyzna	118	72	nie	1961-03-02	40
24	Mężczyzna	122	86	nie	1965-10-05	36
25	Mężczyzna	126	92	nie	1960-01-14	41
26	Mężczyzna	136	90	nie	1948-02-03	53
27	Kobieta	110	82	tak	1965-11-24	36
28	Mężczyzna	136	84	nie	1957-12-03	44
29	Mężczyzna	112	82	nie	1949-03-15	52
30	Kobieta	115	73	tak	1961-02-11	40
31	Mężczyzna	125	86	tak	1946-03-07	55
32	Mężczyzna	160	98	nie	1937-12-08	64
33	Mężczyzna	125	84	tak	1956-09-03	45
34	Mężczyzna	125	75	tak	1960-02-05	41
35	Mężczyzna	145	92	nie	1962-05-04	39
36	Mężczyzna	122	95	nie	1945-10-02	56
37	Mężczyzna	115	85	nie	1944-06-27	57
38	Mężczyzna	215	114	tak	1945-04-13	56
39	Mężczyzna	170	98	nie	1935-04-20	66
40	Mężczyzna	126	76	tak	1955-09-13	46
41	Kobieta	203	158	tak	1941-10-25	60
42	Kobieta	204	116		1943-06-02	58
43	Kobieta	202	108	tak	1949-04-04	52
44	Kobieta	206	140	tak	1959-12-07	42

Tabela 2.9. cd.

45	Kobieta	196	128		1948-08-19	53
46	Kobieta	238	136		1943-07-01	58
47	Kobieta	208	138		1944-08-23	57
48	Kobieta	142	118		1966-05-16	35
49	Kobieta	182	122	tak	1942-06-22	59
50	Kobieta	206	102		1945-04-15	56
51	Kobieta	182	96		1942-08-06	59
52	Kobieta	152	108		1950-10-13	51
53	Kobieta	146	80	nie	1942-12-18	59
54	Kobieta	106	78		1970-11-27	31
55	Kobieta	100	70	nie	1962-12-07	39
56	Kobieta	120	80	tak	1964-01-06	37
57	Kobieta	122	84		1964-08-04	37
58	Kobieta	108	72		1971-01-01	30
59	Kobieta	118	82	tak	1970-05-05	31
60	Kobieta	128	86		1957-11-08	44
61	Kobieta	110	70		1963-05-30	38
62	Mężczyzna	180	118		1941-12-20	60
63	Mężczyzna	184	98		1943-12-21	58
64	Mężczyzna	202	150	tak	1960-07-30	41
65	Mężczyzna	180	100		1941-04-07	60
66	Mężczyzna	224	130		1941-10-03	60
67	Mężczyzna	164	86	nie	1942-04-15	59
68	Mężczyzna	120	82	nie	1970-09-27	31
69	Mężczyzna	110	72		1970-06-25	31
70	Mężczyzna	123	85		1967-01-15	34
71	Kobieta	235	117		1966-03-16	35
72	Mężczyzna	182	105	nie	1943-09-07	58
73	Kobieta	182	105	tak	1953-09-30	48
74	Mężczyzna	172	105	tak	1948-07-01	53
75	Kobieta	160	110	nie	1953-07-30	48
76	Mężczyzna	165	122	nie	1960-08-28	41
77	Mężczyzna	116	75	nie	1972-10-12	29
78	Kobieta	224	144	tak	1952-12-20	49
79	Kobieta	193	132	tak	1957-01-17	44
80	Mężczyzna	125	84	nie	1971-09-22	30
81	Mężczyzna	115	82	nie	1967-03-07	34
82	Kobieta	143	97	nie	1956-04-12	45
83	Mężczyzna	136	90	nie	1950-06-23	51

Tabela 2.9. cd.

84	Kobieta	148	108	tak	1946-06-19	55
85	Kobieta	201	124	nie	1930-11-14	71
86	Mężczyzna	200	110	nie	1935-10-12	66
87	Mężczyzna	120	68	tak	1970-08-14	31
88	Kobieta	185	125	tak	1941-03-05	60
89	Kobieta	223	102	tak	1938-03-04	63
90	Kobieta	90	70	tak	1965-04-03	36
91	Kobieta	80	62	tak	1970-08-03	31
92	Kobieta	185	105	tak	1965-12-05	36
93	Kobieta	160	108	tak	1954-04-25	47
94	Kobieta	110	80	tak	1958-03-10	43
95	Kobieta	80	60	nie	1956-02-04	45
96	Kobieta	174	105	tak	1953-04-09	48
97	Mężczyzna	215	130	tak	1947-10-31	54
98	Kobieta	220	135	tak	1956-08-26	45
99	Mężczyzna	165	115	tak	1946-05-11	55
100	Mężczyzna	170	120	tak	1947-07-25	54

2.4.2. Dane

Pobieranie danych

W zależności od typu badania epidemiologicznego w jego trakcie zostają zebrane informacje dotyczące osób biorących udział w badaniu lub charakteryzujące badaną populację. Szczególnym typem badania jest badanie wyczerpujące. Polega ono na zebraniu danych dotyczących wszystkich jednostek w populacji. Z taką sytuacją mamy do czynienia na przykład podczas badania małej populacji – zamieszkującej skażony teren czy też spożywającej posiłek w stołówce, w której zanotowano zatrucia pokarmowe. Częściej jednak badanie statystyczne polega na pobraniu tak zwanej próby, czyli grupy osób z badanej populacji, i wnioskowaniu na jej podstawie o prawidłowościach w populacji, którą reprezentuje pobrana próba.

Podstawowym warunkiem tego, aby na podstawie zależności obserwowanych w próbie móc wyciągnąć prawdziwe wnioski dotyczące całej populacji, jest losowy dobór próby i pobranie jej w taki sposób, by jak najdokładniej przypominała ona populację pod względem kluczowych z punktu widzenia badania cech, np. rozkładu płci, wieku, struktury zamieszkania czy też zatrudnienia. W związku z tym pierwszym bardzo ważnym aspektem przeprowadzenia badania staje się sposób pobierania próby.

Podstawowym zagadnieniem, jakie trzeba rozważyć, jest reprezentatywność próby. Jeżeli okazałoby się, że próba nie jest reprezentatywna (na przykład dane ze względów logistycznych zostały zebrane jedynie od mieszkańców dużych miast), wnioski dotyczące całej populacji (na przykład rozpowszechnienia występowania chorób odzwierzęcych) mogłyby być fałszywe. Istnieje wiele sposobów losowego pobierania próby – wyboru ludzi do badania. Przedstawimy najważniejsze z nich.

Podstawową metodą jest metoda oparta na *losowaniu prostym*. Przy zastosowaniu tej metody każda osoba z badanej populacji ma takie samo prawdopodobieństwo znalezienia się w próbie. Badana grupa jest losowana bezpośrednio z populacji. Minusem tego rozwiązania jest konieczność posiadania listy wszystkich osób w populacji. W dużych badaniach jest to oczywiście niemożliwe ze względów logistycznych.

Drugą z podstawowych metod losowania jest *losowanie warstwowe*. Polega ono na podzieleniu badanej populacji na homogenne pod względem pewnych cech (wiek, płeć, region geograficzny) grupy – tak zwane warstwy, a następnie na przeprowadzeniu w obrębie każdej z nich losowania prostego. Końcową próbę otrzymuje się, łącząc próby uzyskane w każdej z warstw. Schemat ten zapewnia uzyskanie próby zawierającej reprezentantów wszystkich wybranych podgrup. Można również tak przygotować schemat losowania, by liczebności prób z poszczególnych warstw były proporcjonalne do liczebności populacji w tych warstwach. Schemat ten jest szczególnie użyteczny, gdy występuje związek między badaną cechą a cechami wyznaczającymi warstwy.

Trzecią z najważniejszych metod losowania jest *losowanie zespołowe*. Polega ono na losowym wybraniu nie indywidualnych badanych jednostek, ale ich zespołów i włączeniu ich do próby (często stosowana jest również modyfikacja polegająca na dwustopniowej realizacji losowania i przeprowadzeniu losowania prostego wewnątrz każdego z wylosowanych zespołów). Przykładem losowania zespołowego może być losowy wybór szkół z terenu województwa i przebadanie w nich wszystkich (bądź części) dzieci, np. pod względem występowania chorób kręgosłupa. Koszty tak zaprojektowanego badania są mniejsze niż na przykład przebadanie określonej liczby dzieci z każdej ze szkół.

W praktyce bardzo często stosuje się kombinacje omówionych powyżej metod. Na przykład w pierwszym etapie losowane są zespoły (np. szkoły), w których przeprowadzone jest losowanie warstwowe (np. warstwy wyznaczone są przez klasę).

Zagadnieniem związanym z losowaniem jest występująca w przypadku badań interwencyjnych potrzeba podziału próby na podgrupy – na przykład leczone według różnych programów terapeutycznych. Aby przypisanie osoby do grupy badawczej nie było obciążone wyborem badacza, przeprowadza się procedurę randomizacji – losowego przydziału osób biorących udział

w badaniu do grupy badawczej. Jest wiele sposobów przeprowadzenia tej procedury, zagadnienie to wykracza jednak poza zakres tego rozdziału.

Typy danych

Rozróżnienie typów danych, choć może wydawać się zagadnieniem czysto teoretycznym, ma bardzo duże znaczenie, ponieważ różnym typom danych (wyznaczonym przez omówione poniżej skale pomiarowe) odpowiadają różne techniki opisu statystycznego.

Skale pomiarowe

Wybór odpowiednich metod opisu statystycznego zależy od skali, w jakiej mierzona jest dana cecha. Rozróżniamy trzy podstawowe skale: nominalną, porządkową i interwałową (ilościową).

Skala nominalna

Dane mierzone w skali nominalnej opisane są za pomocą kategorii, których nie potrafimy uporządkować. Typowymi cechami mierzonymi w tej skali są: rasa, płeć, stan cywilny. Szczególnym przykładem danych mierzonych w skali nominalnej są wyniki doświadczenia, w którym możliwe są tylko dwa wyniki – zdarzenie występuje albo nie (choroba obecna/nieobecna, stężenie cholesterolu w normie/powyżej normy).

Skala porządkowa

Wyniki doświadczenia opisujemy za pomocą zestawu pewnych uporządkowanych kategorii. Przykładem danych mierzonych w skali porządkowej może być samopoczucie określane jako słabe/normalne/dobre. Możemy określić kierunek natężenie badanej cechy, jednakże nie potrafimy stwierdzić, że np. samopoczucie słabe jest gorsze o tyle samo od normalnego, co normalne od dobrego. Innymi przykładami danych mierzonych w skali porządkowej są poziom wykształcenia czy też stopień zaawansowania nowotworu.

Skala interwałowa

W skali interwałowej mierzymy dane, co do których mamy informację zarówno o kierunku natężenia badanej cechy, jak i odległości pomiędzy war-

tościami pomiarów. Za przykład zmiennych mierzonych w tej sali mogą posłużyć wszystkie dane wyrażane za pomocą liczb, jak: wzrost, waga, ciśnienie krwi, miano przeciwciał itp.

Pomiary powiązane (skorelowane, zależne)

Niezwykle istotnym elementem planowania analizy statystycznej jest uwzględnienie w niej ewentualnego powiązania pomiarów. Problem ten pojawia się w sytuacji, gdy u osób uczestniczących w badaniu wykonuje się więcej niż jeden pomiar tej samej wielkości. Typową sytuacją jest ocena stanu chorego przed zabiegiem i po nim lub dwukrotne wykonanie testów diagnostycznych u tej samej osoby.

2.4.3. Statystyka opisowa

Szereg rozdzielczy

Rozkład analizowanej cechy o charakterze ciągłym (mierzonej w skali interwałowej) można bardzo łatwo zbadać, tworząc tzw. szereg rozdzielczy. Polega to na zliczeniu liczby obserwacji badanej cechy, które mieszczą się w określonych wcześniej zakresach (przedziałach klasowych). Graficzną reprezentacją szeregu rozdzielczego jest histogram. Przy tworzeniu szeregu rozdzielczego bardzo ważny jest wybór szerokości przedziału. Zarówno zbyt wąskie, jak i zbyt szerokie przedziały utrudnić mogą wyciąganie wniosków o prawidłowościach występujących w danych.

Przykład 1

Prześledźmy kilka szeregów rozdzielczych utworzonych dla zmiennej SBP (systolic blood pressure – skurczowe ciśnienie tętnicze).

Tabela 2.10. Kategorie ciśnienia tętniczego co 20 mmHg

SBP	Liczebność	Procent
Poniżej 100	4	4,0
100–120	21	21,0
121–140	19	19,0
141–160	12	12,0
Powyżej 161	44	44,0

Kategorie zostały wybrane bez odniesienia do istniejących norm dotyczących ciśnienia skurczowego krwi. Szereg rozdzielczy opisuje rozkład badanej cechy, jednak takie przedstawienie danych stwarza trudności interpretacyjne.

Tabela 2.11. Dwie kategorie ciśnienia tętniczego: do 120 i powyżej 120 mmHg

SBP	Liczebność	Procent
0–119	22	22,0
120+	78	78,0

Szereg rozdzielczy zawierający tylko dwie kategorie bardzo ogranicza możliwości interpretacyjne zebranych danych. Wyższa kategoria nie różnicuje osób z bardzo wysokim ciśnieniem, a niższa – z bardzo niskim ciśnieniem.

Tabela 2.12. Kategorie wyznaczone przez stan pacjenta[34]

SBP	Liczebność	Procent
Niedociśnienie (< 90)	3	3,0
Ciśnienie w normie (90–119)	19	19,0
Stan poprzedzający nadciśnienie (120–139)	21	21,0
Nadciśnienie 1 stopnia (140–159)	10	10,0
Nadciśnienie 2 stopnia (co najmniej 160)	47	47,0

Kategorie zostały wybrane na podstawie literatury, dzięki czemu oprócz prostego opisu rozkładu badanej cechy możemy opisać badaną grupę również pod względem rozpowszechnienia zaburzeń ciśnienia różnego rodzaju i o różnym natężeniu.

Miar tendencji centralnej używamy do określenia przeciętnych wartości badanej cechy. Z wielu takich miar przedstawimy trzy: średnią arytmetyczną, medianę oraz wartość modalną.

Najczęściej używaną miarą położenia jest *średnia arytmetyczna* (arithmetic mean). Oblicza się ją według wzoru:

$$\bar{\mu} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i ,$$

gdzie:

$x_1,\ldots,x_n$ – obserwacje badanej cechy u każdej z n jednostek badania.

---

[34] Chobanian A.V. i in.: *Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure*. Hypertension, 2003, 42(6), 1206–52.

Średnia arytmetyczna może być obliczana tylko dla danych w skali interwałowej. Dobrze opisuje dane o rozkładzie symetrycznym, natomiast nie powinno się jej używać w sytuacji, gdy w zbiorze danych istnieją pomiary znacznie odbiegające od innych (rozkład jest skośny). Średnia arytmetyczna jest estymatorem punktowym średniej wartości analizowanej cechy w badanej populacji.

Drugą ważną miarą położenia jest *mediana* (median). Jest to wartość cechy dzieląca zbiór elementów próby na połowy. Może być obliczana jedynie dla danych w skali interwałowej. Obserwacje znacznie odbiegające od innych nie zmieniają jej wartości w tak dużym stopniu jak w przypadku średniej arytmetycznej.

Wartość modalna to najczęściej występująca w zbiorze danych wartość badanej cechy.

Miary rozproszenia używane są do opisu zmienności badanej cechy w populacji. Konkretne miary rozproszenia są połączone z konkretnymi miarami położenia. Najczęściej stosowanymi miarami rozproszenia są odchylenie standardowe (odpowiadające średniej arytmetycznej) i kwantyle (odpowiadające medianie).

Odchylenie standardowe (standard deviation) dane jest wzorem:

$$sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})}.$$

Pozwala ono zatem stwierdzić, o ile przeciętnie różnią się wartości analizowanej cechy występujące u badanych jednostek od średniej arytmetycznej wartości tej cechy. Odchylenie standardowe może być obliczane jedynie dla danych w skali interwałowej.

Kwantyle są tymi wartościami analizowanej cechy, które dzielą badaną zbiorowość na podgrupy o równych liczebnościach. Najczęściej są to: *tercyle* – trzy podgrupy, *kwartyle* – cztery podgrupy, *kwintyle* – pięć podgrup, *decyle* – dziesięć podgrup i *centyle* – sto podgrup. *Percentylem* (rzędu q, $0 < q < 100$) nazywamy wartość, poniżej której zawiera się q% obserwacji (uporządkowanych rosnąco).

Przedziały ufności

Opisane wyżej miary tendencji centralnej i rozproszenia noszą wspólną nazwę estymatorów punktowych. Oznacza to, że dana miara w populacji generalnej oszacowana jest przez jedną wartość liczbową. Prawie pewne jest jednak, że szacowany parametr w populacji generalnej nie jest identyczny z wartością estymatora punktowego z próby. Chcąc oszacować nie-

pewność dotyczącą estymatora punktowego, możemy obliczyć jego *błąd standardowy*.

W przypadku średniej arytmetycznej dany jest on wzorem:

$$se = \frac{sd}{\sqrt{n}},$$

gdzie:

sd (standard deviation) – odchylenie standardowe,

n – liczba pomiarów.

Zauważmy, że za pomocą odchylenia standardowego oceniamy rozproszenie wyników, a za pomocą błędu standardowego średniej niepewność dotyczącą jej wartości.

Ważną rolę odgrywa metoda estymacji przedziałowej. Polega ona na konstrukcji dla badanego parametru (na podstawie jego błędu standardowego) tak zwanego *przedziału ufności* (confidence interval), w którym prawdziwa wartość badanego parametru leży z zadanym prawdopodobieństwem. Szczególnie ważne są przedziały ufności dla wartości średniej oraz dla odsetka osób w populacji posiadających wybraną cechę.

Przedział ufności dla wartości średniej dany jest wzorem:

$$(\bar{x} - t_\alpha \cdot \frac{s}{\sqrt{n-1}};\ \bar{x} + t_\alpha \cdot \frac{s}{\sqrt{n-1}}),$$

gdzie:

$\bar{x}$ – średnia arytmetyczna z próby,

s – odchylenie standardowe,

t_α – odpowiednia wartość krytyczna rozkładu t-Studenta.

Przedział ufności dla wskaźnika struktury (procentu) dany jest natomiast wzorem:

$$(p - u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}};\ p + u_\alpha \cdot \sqrt{\frac{p(1-p)}{n}}),$$

gdzie:

p – odsetek osób z daną cechą,

n – liczebność populacji,

u_α – odpowiednia wartość krytyczna rozkładu normalnego.

Przedziały ufności oblicza się nie tylko dla statystyk opisowych, ale również dla miar ryzyka, o których będzie mowa poniżej.

Rozkłady prawdopodobieństwa

W początkowej fazie analizy bardzo istotny jest opis częstości występowania konkretnych wartości badanej cechy. Należy zbadać jej rozkład prawdopodobieństwa, który charakteryzuje badaną populację pod względem interesującej nas zmiennej, poprzez przypisanie wartościom tej zmiennej prawdopodobieństwa ich wystąpienia.

W przypadku zmiennej o charakterze ciągłym możemy podać tylko prawdopodobieństwo, że konkretna jej wartość należy do pewnego przedziału. Rozkład prawdopodobieństwa wyznaczony jest przez jego parametry. Oznacza to, że do oceny danego zjawiska o danym rozkładzie wystarcza nam ocena parametrów tego rozkładu. Funkcję, która wartościom badanej zmiennej przypisuje prawdopodobieństwo ich wystąpienia, nazywamy *funkcją gęstości prawdopodobieństwa*. Rozkładami mającymi największe zastosowanie w epidemiologii są rozkład normalny, dwumianowy i Poissona.

Rozkład dwumianowy opisuje prawdopodobieństwo wystąpienia dokładnie określonej liczby zdarzeń w pewnej określonej liczbie powtórzeń. Zależy od dwóch parametrów – liczby powtórzeń danego doświadczenia oraz prawdopodobieństwa sukcesu w pojedynczym doświadczeniu. Z sytuacją taką mamy do czynienia na przykład przy opisie wyników leczenia (pozytywny/negatywny) w grupie pacjentów o określonej liczebności. Badanym zdarzeniem jest fakt wyleczenia, powtórzeniami jest liczba pacjentów, a nieznane prawdopodobieństwo sukcesu (wyleczenia) możemy oszacować na podstawie uzyskanych danych.

Rozkład Poissona ma szczególne znaczenie w epidemiologii, gdyż za jego pomocą opisuje się liczbę interesujących nas zdarzeń w sytuacji, gdy mogą się one pojawiać w sposób w pełni losowy (swobodny) – na przykład rozkład dziennej liczby zgonów w danym mieście w ciągu roku. Rozkład ten wyznaczony jest przez jeden parametr – średnią liczbę zdarzeń.

Prawdopodobieństwo wystąpienia dokładnie k zdarzeń w rozkładzie Poissona (funkcja gęstości) dane jest wzorem:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

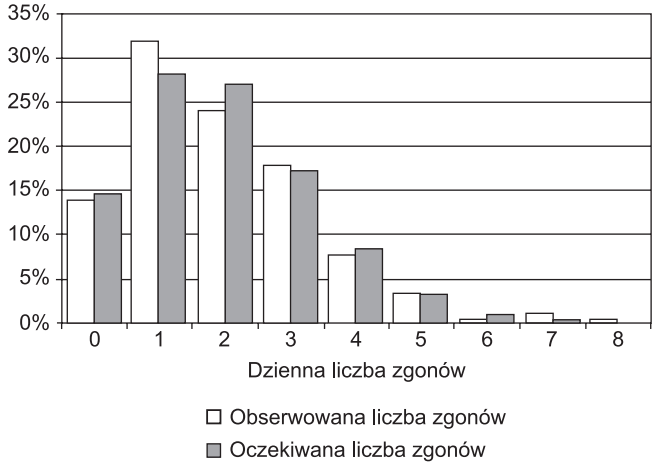
gdzie:

k – liczba zdarzeń,

λ – średnia liczba zdarzeń.

Przykład 2

W przykładzie przeanalizowano rozkład obserwowanej i oczekiwanej dziennej liczby zgonów z powodu chorób układu oddechowego osób w wieku co najmniej 70 lat w Warszawie w 2000 r. Oczekiwana liczba zgonów została obliczona ze wzoru na gęstość rozkładu Poissona ze średnią 1,918 oszacowaną na podstawie danych empirycznych. Dobra zgodność rozkładu empirycznego (obserwowanego) i teoretycznego świadczy o tym, że obserwowane liczby zgonów w kolejnych dniach są zdarzeniami niezależnymi (realizacją procesu Poissona).

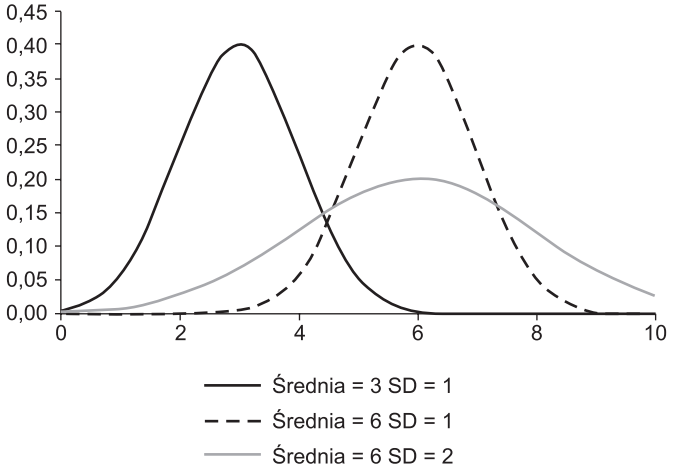


Ryc. 2.17. Rozkład dziennej obserwowanej i oczekiwanej liczby zgonów z powodu chorób układu oddechowego w Warszawie w 2000 roku.

Rozkład normalny (*Gaussa*) jest bardzo często wykorzystywany – można za jego pomocą opisać rozpowszechnienie w populacji wielu cech mierzonych w skali interwałowej, takich jak wzrost, waga, wyniki badań laboratoryjnych itp. Rozkład ten jest bardzo ważny również ze względu na jego właściwości matematyczne – na jego podstawie powstało wiele metod analizy danych. Rozkład normalny wyznaczony jest przez dwa parametry – średnią i odchylenie standardowe. Rycina 2.18 przedstawia funkcje gęstości rozkładu normalnego dla różnych wartości średnich i odchylenia standardowego. Dowolny rozkład normalny daje się przekształcić w tak zwany standardowy rozkład normalny (o średniej 0 i odchyleniu standardowym 1), którego wartości są ujęte w tablicach.

Znajomość rozkładu danej zmiennej pozwala na obliczenie (odczytanie) prawdopodobieństwa wartości mniejszych (większych) od danej war-

tości, a więc na określenie, czy jest prawdopodobne, że dana wartość pochodzi z rozkładu o założonych średniej i odchyleniu standardowym. Ma to kluczowe znaczenie we wnioskowaniu statystycznym, podczas którego obliczamy prawdopodobieństwo tego, że analizowane zależności są wynikiem jedynie zmienności losowej (sprawdzamy, czy jest prawdopodobne, że obliczona statystyka testowa pochodzi z określonego rozkładu prawdopodobieństwa).



Ryc. 2.18. Przykłady funkcji gęstości rozkładu normalnego przy różnych wartościach średniej i odchylenia standardowego.

Znajomość rozkładu prawdopodobieństwa badanej zmiennej jest istotna również z tego względu, że pozwala na ocenę, czy wystąpienie konkretnej wartości tej zmiennej jest prawdopodobne, czy nie. Będzie to miało kluczowe znaczenie we wnioskowaniu statystycznym. Polega ono na ocenie prawdopodobieństwa tego, czy posiadane dane potwierdzają postawioną hipotezę, czy jej przeczą. We wszystkich testach prezentowanych poniżej prawdopodobieństwo to oznacza się literą p i interpretowane jest jako prawdopodobieństwo tego, że obserwowane różnice wynikają jedynie ze zmienności losowej. Jeżeli prawdopodobieństwo okaże się małe (zwykle za graniczną wartość przyjmuje się wartość $p = 0,05$), to interpretujemy to tak, że przyczyną obserwowanych różnic jest inny czynnik niż tylko losowy. W przypadku porównywania dwóch prób (populacji) fakt taki będziemy interpretować jako istnienie istotnych statystycznie różnic pomiędzy nimi co do poziomu lub częstości występowania badanej cechy.

2.4.4. Analiza zmiennych skategoryzowanych

Współczynniki i ich standaryzacja

Podstawową miarą opisu stanu zdrowia populacji są współczynniki umieralności, chorobowości, zapadalności itp. Współczynniki te omówione zostały dokładnie w rozdziale „Epidemiologia – narzędzia badawcze i metody". Zauważmy, że obliczenie współczynnika lub procentu osób chorych czy zmarłych odpowiada obliczeniu statystyk opisowych dla zmiennej o charakterze kategorycznym, posiadającej dwie wartości: tak/nie (żyje/ /zmarł, zdrowy/chory itp.). W przypadku, gdy chcemy porównać dwie lub więcej populacje pod względem częstości rozpowszechnienia pewnej cechy, konieczne jest rozważenie, czy obserwowane różnice lub ich brak są wynikiem różnic w rozpowszechnieniu badanej cechy, czy może różnic pewnych charakterystyk populacji, np. rozkładu płci i wieku, które mają bezpośredni wpływ na występowanie tej cechy.

Dla zilustrowania tego problemu prześledzimy dane przedstawione w przykładzie 3.

Przykład 3

Rozważmy problem porównania współczynników umieralności w dwóch hipotetycznych populacjach (dla uproszczenia w analizie nie uwzględniamy płci). Widać, że we wszystkich grupach wieku współczynniki cząstkowe są wyższe w populacji B, jednak ogólny (rzeczywisty) współczynnik umieralności wyższy jest w populacji A. Różnice wynikają z innego rozkładu wieku w porównywanych populacjach. W populacji A przeważają starsze, a w populacji B młodsze grupy wieku. Natężenie umieralności jest w nich odmienne.

Tabela 2.13

Wiek	Populacja A			Populacja B		
	Liczba mieszkańców	Liczba zgonów	Cząstkowy współczynnik umieralności	Liczba mieszkańców	Liczba zgonów	Cząstkowy współczynnik umieralności
0–20	100 000	10	10,0	400 000	100	25,0
21–40	200 000	100	50,0	300 000	195	65,0
41–60	300 000	300	100,0	200 000	300	150,0
60 +	400 000	800	200,0	100 000	400	400,0
Ogółem	1 000 000	1210	121,0	1 000 000	995	100,0