Marta Zalewska
Wojciech Zieliński
Stanisław Jaworski
Konrad Furmańczyk

Zbiór zadań z podstaw statystyki

Spis treści

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Przedmowa

Niniejsza książeczka powstała na bazie wieloletnich doświadczeń dydaktycznych prowadzenia przedmiotów statystycznych na Warszawskim Uniwersytecie Medycznym. Jej celem jest zebranie podstawowych informacji pojawiających się w trakcie wykładów oraz udostępnienie Studentom zadań i najbardziej niezbędnych tablic statystycznych.

Książeczka składa się z rozdziałów będących jednocześnie tematami kolejnych spotkań wykładowych i ćwiczeniowych. Na początku każdego rozdziału podane są najważniejsze fakty i wzory, następnie rozwiązane są przykładowe zadania i podanych jest kilkanaście zadań do samodzielnego rozwiązania. Nie należy traktować książki jako wykładu ze statystyki czy ekonometrii, lecz jako przewodnik do lepszego zrozumienia wykładów i ćwiczeń. Na końcu zebrane są najważniejsze tablice statystyczne niezbędne przy rozwiązywaniu zadań.

W dobie komputeryzacji wiele zadań można rozwiązać posługując się odpowiednim oprogramowaniem. Pokazano wykorzystanie arkusza kalkulacyjnego Excel, niemniej jednak zadania oczywiście można również rozwiązać posługując się innym oprogramowaniem statystycznym.

Do zadań nie podano odpowiedzi. Wynika to stąd, że celem tych zadań jest nie tyle uzyskanie konkretnego wyniku liczbowego ile „zmuszenie" Czytelnika do maksymalnie samodzielnej analizy zagadnienia. Poza tym, każde z prezentowanych zadań można rozwiązać korzystając z różnych technik statystycznych.

Na zakończenie podano garść informacji bibliograficznych. Literatura przedmiotu jest bardzo bogata i wybór tych a nie innych książek po pierwsze nie wyczerpuje spektrum bibliograficznego, a po drugie podyktowany jest ich dostępnością w księgarniach i bibliotekach. Ze względu na podstawowy charakter wykładu w spisie literatury można znaleźć zarówno pozycje najnowsze jak i sprzed kilkudziesięciu lat. Należy zwrócić uwagę, że ich „zabytkowość" w niczym nie umniejsza ich wartości merytorycznej i poznawczej.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

1. Analiza danych

Analiza danych jest działem statystyki zajmującym się syntetycznym opisem zbiorów danych. Techniki analizy danych stosowane są zazwyczaj tam, gdzie nie są znane mechanizmy rządzące obserwowanymi zjawiskami i na podstawie uzyskiwanych informacji budowany jest pewien model zjawiska oraz formułowane są różnorakie przypuszczenia, które mogą być weryfikowane technikami zaliczanymi do prezentowanej wcześniej grupy metod wnioskowania statystycznego. Spośród wielu mierników stosowanych w analizie danych prezentowanych jest tylko kilka częściej spotykanych w zastosowaniach praktycznych.

Zbierane dane mogą być przedstawione w jednej z dwóch postaci. Jedną z nich jest **próba prosta** lub **dane indywidualne**, tzn. do dyspozycji są kolejno zbierane informacje $X_1, X_2, \ldots, X_n$. Druga postać danych, to **szereg rozdzielczy** lub **dane skumulowane**

Przedział klasowy	Liczebność
$x_0 - x_1$	n_1
$x_1 - x_2$	n_2
$\vdots$	$\vdots$
$x_{k-1} - x_k$	n_k

Pojęcie przedziału klasowego może rzeczywiście opisywać pewien przedział na prostej, ale też może być to pojedyncza wartość (np. liczba oczek na kostce) lub wielkość opisowa (np. barwa). W dalszym ciągu zajmować się będziemy tylko obserwacjami cech ilościowych.

W przypadku próby prostej niech $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ będzie uporządkowanym ciągiem danych. Dla szeregu rozdzielczego konstruowany jest szereg skumulowany

Przedział klasowy	Liczebność skumulowana
$x_0 - x_1$	$n_{(1)} = n_1$
$x_1 - x_2$	$n_{(2)} = n_1 + n_2$
$\vdots$	$\vdots$
$x_{k-1} - x_k$	$n_{(k)} = n_1 + n_2 + \cdots + n_k (= n)$

Dla liczby p takiej, że $0 \leq p \leq 1$, niech x_p, n_p, h_p oznaczają początek, liczebność i długość przedziału zawierającego obserwację o numerze $[p \cdot n]$ oraz niech $n_{(p)}$ oznacza liczebność skumulowaną przedziału poprzedzającego przedział o początku x_p. Symbol $[z]$ oznacza największą liczbę całkowitą nie większą niż z.

Mierniki położenia są grupą charakterystyk opisujących poziom obserwowanej cechy, tzn. w sposób syntetyczny charakteryzujących wartości przyjmowane przez badaną cechę.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

1. Średnia określona jest wzorem

$$\bar{x} = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i, & \text{dla próby prostej,} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{k} \dot{x}_i n_i, & \text{dla szeregu rozdzielczego.} \end{cases}$$

Jest to liczba charakteryzująca „środek ciężkości” danych. Liczba $\dot{x}_i$ oznacza środek przedziału (x_{i-1}, x_i).

2. Mediana określona jest wzorem

$$Me = \begin{cases} X_{[n/2]:n}, & \text{dla próby prostej,} \\ x_{0.5} + \dfrac{h_{0.5}}{n_{0.5}}\left(\dfrac{n}{2} - n_{(0.5)}\right), & \text{dla szeregu rozdzielczego.} \end{cases}$$

Mediana charakteryzuje „geometryczny środek” danych. Połowa danych znajduje się poniżej mediany, zaś druga połowa powyżej.

3. Dolny kwartyl określony jest wzorem

$$Q_1 = \begin{cases} X_{[n/4]:n}, & \text{dla próby prostej,} \\ x_{0.25} + \dfrac{h_{0.25}}{n_{0.25}}\left(\dfrac{n}{4} - n_{(0.25)}\right), & \text{dla szeregu rozdzielczego.} \end{cases}$$

Dolny kwartyl rozdziela dane w proporcji 1 : 3, tzn. poniżej dolnego kwartyla znajduje się czwarta część danych, zaś powyżej reszta.

4. Górny kwartyl określony jest wzorem

$$Q_3 = \begin{cases} X_{[3n/4]:n}, & \text{dla próby prostej,} \\ x_{0.75} + \dfrac{h_{0.75}}{n_{0.75}}\left(\dfrac{3n}{4} - n_{(0.75)}\right), & \text{dla szeregu rozdzielczego.} \end{cases}$$

Górny kwartyl rozdziela dane w proporcji 3 : 1.

5. Dominanta (**moda**) jest najczęściej występującą wartością. W przypadku danych skumulowanych wyznaczana jest ona za pomocą wzoru

$$D = x_D + h_D \frac{n_D - n_{D-1}}{2n_D - n_{D+1} - n_{D-1}}.$$

Tutaj x_D, h_D oraz n_D są odpowiednio początkiem, szerokością oraz liczebnością przedziału o największej ilości danych (tzn. $n_D = \max\{n_1, \ldots, n_k\}$), natomiast n_{D-1} oraz n_{D+1} są liczebnościami przedziałów sąsiadujących z przedziałem o liczebności n_D. Należy zauważyć, że wyznaczanie dominanty ma sens dla szeregów o jednym maksimum.

Mierniki rozproszenia są grupą charakterystyk opisujących zróżnicowanie cechy, tzn. w sposób syntetyczny opisujących zróżnicowanie wartości przyjmowanych przez badaną cechę.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

1. Wariancja określona jest wzorem

$$\mathcal{S}^2 = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2, & \text{dla próby prostej,} \\ \frac{1}{n}\sum_{i=1}^{k}n_i(X_i - \bar{x})^2, & \text{dla szeregu rozdzielczego.} \end{cases}$$

Jest to liczba charakteryzująca rozrzut danych wokół ich „środka ciężkości". Stosowany jest również nieco inny sposób wyznaczania wariancji:

$$S^2 = \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2, & \text{dla próby prostej,} \\ \frac{1}{n-1}\sum_{i=1}^{k}n_i(X_i - \bar{x})^2, & \text{dla szeregu rozdzielczego.} \end{cases}$$

2. Odchylenie standardowe S jest pierwiastkiem z wariancji S^2. Zaletą odchylenia standardowego jest to, że wyrażone jest ono w tych samych jednostkach co oryginalne dane (jednostką wariancji jest kwadrat jednostek pomiarowych).

3. Współczynnik zmienności określony wzorem

$$V = \frac{\mathcal{S}}{\bar{x}}100\%$$

opisuje względne zróżnicowanie danych, tzn. udział odchylenia standardowego w wartości średniej.

4. Odchylenie przeciętne określone jest wzorem

$$d = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{x}|, & \text{dla próby prostej.} \\ \frac{1}{n}\sum_{i=1}^{k}n_i|X_i - \bar{x}|, & \text{dla szeregu rozdzielczego,} \end{cases}$$

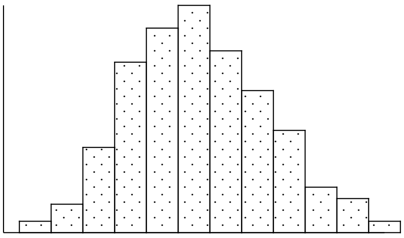
Odchylenie przeciętne, podobnie jak odchylenie standardowe mierzy rozrzut danych wokół średniej. W wartości odchylenia przeciętnego każda z danych ma taki sam udział, natomiast w wartości odchylenia standardowego większy udział mają dane „dalekie" od średniej, tzn. im obserwacja jest bardziej odległa od średniej, tym jej udział w odchyleniu standardowym jest większy.

5. Rozstęp R jest różnicą między największą a najmniejszą daną i pokazuje zakres zmienności zjawiska.

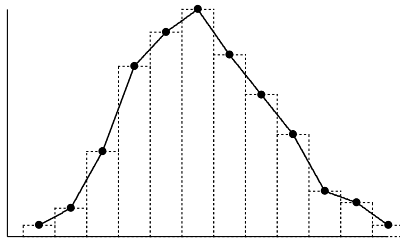
6. Odchylenie ćwiartkowe określone jest wzorem

$$Q = \frac{Q_3 - Q_1}{2}$$

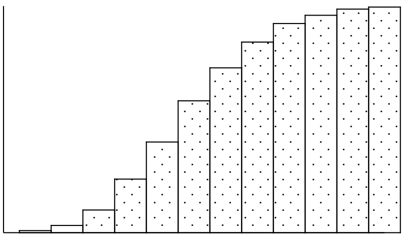
Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk



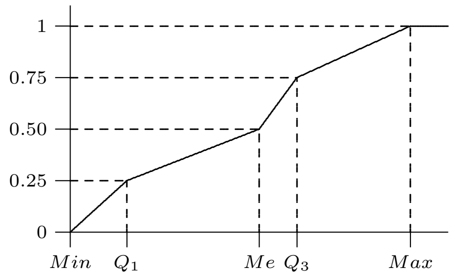
Histogram



Wielobok częstości



Histogram skumulowany

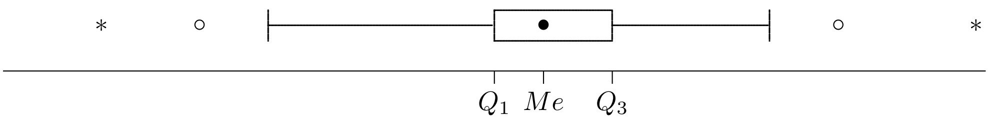


Wykres kwartylowy

i opisuje zakres zmienności środkowych 50% danych.

W analizie danych pomocne są **prezentacje graficzne**. Na rysunku pokazane są cztery najczęściej stosowane sposoby prezentacji: histogram (wykres słupkowy), wielobok częstości, skumulowany histogram oraz wykres kwartylowy. Pierwsze dwa z tych wykresów mogą być kojarzone z funkcją gęstości rozkładu prawdopodobieństwa, zaś pozostałe z dystrybuantą.

Inną formą prezentacji danych są wykresy pudełkowe. Przykładowy wykres:



Wykres pudełkowy

Na wykresie zaznaczone są także **obserwacje odstające** (oznaczone na wykresie symbolem $\circ$) oraz **obserwacje ekstremalne** (oznaczone na wykresie symbolem $*$). Obserwację X nazywamy **odstającą**, jeżeli $X > Q_3 + 1.5(Q_3 - Q_1)$ lub $X < Q_1 - 1.5(Q_3 - Q_1)$, natomiast nazywamy ją **ekstremalną**, jeżeli $X > Q_3 + 2 \cdot 1.5(Q_3 - Q_1)$ lub $X < Q_1 - 2 \cdot 1.5(Q_3 - Q_1)$.

Tablice kontyngencji. W wielu badaniach obserwowane są dwie cechy X oraz Y, przy czym zbiór wartości cechy X jest podzielony na k klas, natomiast zbiór wartości

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

cechy Y – na m klas. Wyniki tych badań zapisywane są następującej postaci tablic kontyngencji:

Klasy cechy X	Klasy cechy Y: 1	2	$\dots$	m
1	n_{11}	n_{12}	$\dots$	n_{1m}
2	n_{21}	n_{22}	$\dots$	n_{2m}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	n_{k1}	n_{k2}	$\dots$	n_{km}

Wielkość n_{ij} oznacza, że w wśród n danych było n_{ij} obserwacji i-tej klasy cechy X i jednocześnie j-tej klasy cechy Y. Oczywiście $\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij} = n$.

Na podstawie tak zebranych informacji odpowiada się na pytanie, czy obserwowane cechy są czy nie są zależne. W tym celu wyznaczane są warunkowe rozkłady jednej cechy względem drugiej:

Rozkłady warunkowe cechy X względem Y

Klasy cechy X	Klasy cechy Y: (1)	(2)	$\dots$	(m)
1	$n_{11}/n_{\cdot 1}$	$n_{12}/n_{\cdot 2}$	$\dots$	$n_{1m}/n_{\cdot m}$
2	$n_{21}/n_{\cdot 1}$	$n_{22}/n_{\cdot 2}$	$\dots$	$n_{2m}/n_{\cdot m}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	$n_{k1}/n_{\cdot 1}$	$n_{k2}/n_{\cdot 2}$	$\dots$	$n_{km}/n_{\cdot m}$

lub

Rozkład warunkowe cechy Y względem X

Klasy cechy X	Klasy cechy Y: 1	2	$\dots$	m
(1)	$n_{11}/n_{1\cdot}$	$n_{12}/n_{1\cdot}$	$\dots$	$n_{1m}/n_{1\cdot}$
(2)	$n_{21}/n_{2\cdot}$	$n_{22}/n_{2\cdot}$	$\dots$	$n_{2m}/n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
(k)	$n_{k1}/n_{k\cdot}$	$n_{k2}/n_{k\cdot}$	$\dots$	$n_{km}/n_{m\cdot}$

W powyższych rozkładach $n_{i\cdot} = \sum_{j=1}^{m} n_{ij}$ oznacza ogólną liczbę obserwacji i-tej klasy cechy X, natomiast $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$ – ogólną liczbę obserwacji j-tej klasy cechy Y.

Wnioskowanie jest następujące: jeżeli rozkłady warunkowe cechy X względem Y (lub rozkłady warunkowe cechy Y względem X) są „mniej więcej” takie same, to możemy podejrzewać, że obserwowane cechy są niezależne. W przeciwnym przypadku cechy należy uznać za zależne.

Przykład. *W celu zbadania istnienia związku między wykształceniem (W) a zarobkami (Z) wylosowano 950 osób. Wartości badanych cech podzielono na następujące klasy:*

Zarobki	≤ 500	$500-1000$	$1000-1500$	$1500-2000$	≥ 2000
Klasa	Z_1	Z_2	Z_3	Z_4	Z_5

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Wykształcenie	*podstawowe*	*średnie*	*wyższe*	*ponad wyższe*
Klasa	W_1	W_2	W_3	W_4

Uzyskano następujące dane

	W_1	W_2	W_3	W_4
Z_1	21	41	93	47
Z_2	33	37	35	53
Z_3	45	75	27	43
Z_4	30	48	50	55
Z_5	71	47	49	50

Czy powyższe świadczą o istnieniu zależności między wykształceniem i zarobkami?

Rozwiązanie. *Zbadano łącznie* $N = 950$ *osób.*

Zaobserwowano następujące liczby osób w poszczególnych klasach każdej z cech:

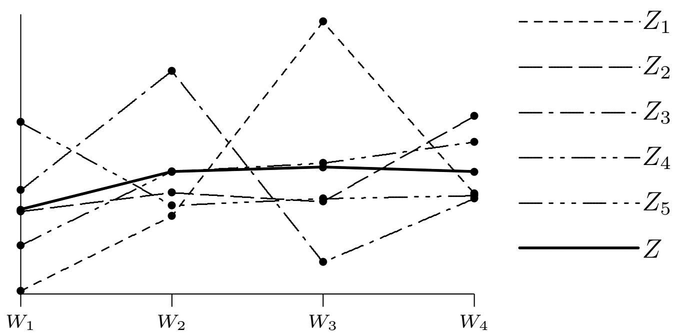
Zarobki	Z_1	Z_2	Z_3	Z_4	Z_5
Liczebność	$n_{1\cdot} = 202$	$n_{2\cdot} = 158$	$n_{3\cdot} = 190$	$n_{4\cdot} = 183$	$n_{5\cdot} = 217$

Wykształcenie	W_1	W_2	W_3	W_4
Liczebność	$n_{\cdot 1} = 200$	$n_{\cdot 2} = 248$	$n_{\cdot 3} = 254$	$n_{\cdot 4} = 248$

Rozkład warunkowy wykształcenia w grupach zarobkowych:

	W_1	W_2	W_3	W_4
(Z_1)	0.104	0.203	0.460	0.233
(Z_2)	0.209	0.234	0.222	0.335
(Z_3)	0.237	0.395	0.142	0.226
(Z_4)	0.164	0.262	0.273	0.301
(Z_5)	0.327	0.217	0.226	0.230
(Z)	0.211	0.261	0.267	0.261

Rozkład wykształcenia w poszczególnych grupach zarobkowych można przedstawić graficznie w następujący sposób:



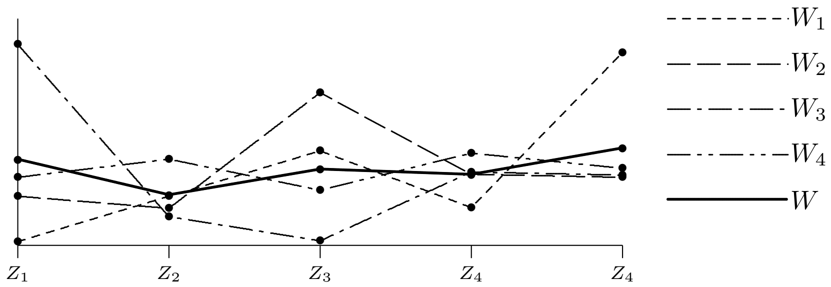
Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Warunkowe rozkłady w poszczególnych grupach zarobkowych nie są takie same (odpowiednie linie na wykresie nie są równoległe). Wnioskujemy więc, że wykształcenie i zarobki nie są niezależnymi cechami.

W podobny sposób można analizować rozkłady zarobków w poszczególnych klasach wykształcenia. Rozkład warunkowy zarobków względem wykształcenia:

	Z_1	Z_2	Z_3	Z_4	Z_5
(W_1)	0.105	0.165	0.225	0.150	0.355
(W_2)	0.165	0.149	0.302	0.194	0.190
(W_3)	0.366	0.138	0.106	0.197	0.193
(W_4)	0.190	0.214	0.173	0.222	0.202
(W)	0.213	0.166	0.200	0.193	0.228

Rozkład zarobków w poszczególnych klasach wykształcenia można przedstawić graficznie w następujący sposób:



Podobnie jak wcześniej wnioskujemy, że wykształcenie i zarobki nie są niezależnymi cechami. □

Zadania do samodzielnego rozwiązania

W poniższych zadaniach wyznaczyć wskaźniki położenia oraz rozproszenia analizowanych zmiennych. Podać interpretację wyznaczonych wskaźników.

1.1. Zużycie papieru (w kg) w Polsce w latach 1960 – 1971 na jednego mieszkańca: 2.2, 2.2, 2.1, 2.3, 2.3, 2.3, 2.3, 2.4, 2.4, 2.5, 2.4, 2.5. Przyjąć upraszczające założenie, że liczba mieszkańców w Polsce w tym okresie była stała.

1.2. Liczba koni (w mln. szt.) w Polsce w latach 1947 – 1974 wynosiła: 2.0, 2.3, 2.7, 2.8, 2.9, 2.7, 2.7, 2.6, 2.6, 2.5, 2.6, 2.7, 2.8, 2.8, 2.7, 2.7, 2.6, 2.6, 2.6, 2.6, 2.6, 2.7, 2.6, 2.6, 2.5, 2.4, 2.4, 2.3.

1.3. Procentowa zawartość tłuszczu w mleku 50 krów: 3.35, 4.16, 3.24, 4.23, 3.42, 3.73, 3.56, 3.98, 3.70, 4.47, 3.94, 3.92, 3.62, 3.53, 3.93, 4.16, 3.22, 4.10, 3.72, 4.26, 3.92, 3.66, 3.78, 3.96, 3.81, 4.28, 3.50, 3.39, 3.83, 4.27, 4.26, 3.71, 3.93, 4.27, 4.06, 3.78, 3.96, 3.89, 3.93, 4.06, 3.99, 3.77, 4.22, 3.78, 3.66, 3.41, 3.53, 3.54, 4.08, 3.44.

Ponadto, skonstruować szereg rozdzielczy (od 3.2 co 0.2) i na podstawie tego szeregu również wyznaczyć wskaźniki położenia i rozproszenia. Porównać uzyskane wyniki.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

1.4. Procentowa zawartość skrobi w każdym z 80 ziemniaków wylosowanych z partii ziemniaków:

Zawartość skrobii	9–11	11–13	13–15	15–17	17–19	19–21	21–23	23–25
Liczba ziemniaków	1	2	7	20	30	16	3	1

1.5. Czas dojazdu pracowników z miejsca zamieszkania do pracy:

Czas dojazdu	5–15	15–25	25–35	35-45	45–55	55–65
Liczba pracowników	3	5	25	15	5	2

W poniższych zadaniach wyznaczyć wskaźniki położenia oraz rozproszenia analizowanych zmiennych i na tej podstawie przeprowadzić analizę porównawczą.

1.6. Powierzchnia użytkowa mieszkań na wsiach w latach 1978 i 1988.

Powierzchnia	1978	1988
20 – 40	50	60
40 – 60	300	300
60 – 80	400	350
80 – 100	150	150
100 – 120	80	100
120 – 140	20	40

1.7. Struktura bezrobocia wśród mężczyzn i kobiet

Miesiące bez pracy	mężczyźni	kobiety
0 – 3	214	153
3 – 6	161	139
6 – 9	121	116
9 – 12	108	96
12 – 15	396	496

1.8. Powierzchnia użytkowa mieszkań na wsiach i w miastach.

Powierzchnia	miasto	wieś
20 – 40	80	50
40 – 60	350	300
60 – 80	400	400
80 – 100	100	150
100 – 120	50	80
120 – 140	20	20

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

1.9. Struktura wynagrodzeń miesięcznych w przemyśle i budownictwie.

Płaca	przemysł	budownictwo
0 – 200	250	100
200 – 400	450	350
400 – 600	150	400
600 – 800	100	100
800 – 1000	50	50

1.10. Opinie konsumentów dotyczące dwóch gatunków kawy na podstawie badań sondażowych. Każdy z sześćdziesięciu konsumentów oceniał każdą z dwóch kaw w skali punktowej.

Ocena	kawa Szatanex	kawa Lureksja
3	4	6
4	14	16
5	17	21
6	19	13
7	6	4

W poniższych zadaniach przeprowadzić graficzną analizę zależności pomiędzy badanymi cechami.

1.11. Właściciel palarni kawy twierdzi, że stopień palenia kawy nie ma wpływu na jej smak, a dokładnie na gorzkość. W celu udowodnienia tej tezy wybrano pewną mieszankę kawy i poddano ją procesowi palenia w różnym stopniu. Uzyskano następujące wyniki:

	smak kawy		
	normalna	gorzka	bardzo gorzka
słabo palona	5	9	4
mocno palona	2	12	8
bardzo mocno palona	1	7	14

W oparciu o powyższe dane odpowiedzieć na pytanie, czy właściciel palarni ma rację?

1.12. Poniższa tabela przedstawia liczbę psów zdrowych i chorych na nosówkę w zależności od tego, czy pies ma rodowód, czy go nie ma. Zbadać, czy istnieje zależność między zdrowotnością psa a posiadaniem przez niego rodowodu.

	Psy z rodowodem	Psy bez rodowodu
Psy zdrowe	300	200
Psy chore	40	20

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

1.13. Przypuszczano, że sposób zapewniania sobie posiłków w pracy przez pracowników, którym firma nie zapewnia regularnego wyżywienia, zależy od płci. W tym celu wylosowano pewną grupę pracowników i uzyskano następujące wyniki:

Płeć	śniadanie z domu	obiad na mieście	zamówienie do pracy
Mężczyźni	68	36	23
Kobiety	36	50	18

W oparciu o powyższe dane odpowiedzieć na pytanie, czy przypuszczenie można uznać za uzasadnione.

1.14. Pracownicy fabryk pewnego zjednoczenia charakteryzują się różną absencją. Wysunięto przypuszczenie, że absencja zależy do płci. Zweryfikować to przypuszczenie na podstawie poniższych danych.

Liczba dni	Płeć	
nieobecności	Kobiety	Mężczyźni
0–5	300	500
5–20	80	70
20 i więcej	20	30

1.15. W badaniach budżetów rodzinnych wylosowano 2000 gospodarstw domowych i zanotowano średni miesięczny dochód na głowę oraz fakt posiadania magnetowidu. Czy można na tej podstawie powiedzieć, że fakt posiadania magnetowidu jest wskaźnikiem zamożności rodziny?

Dochód	Magnetowid	
na głowę	jest	nie ma
poniżej 200	404	231
200– 400	486	300
400– 600	242	137
600– 800	57	44
800–1000	29	28
1000 i więcej	24	18

1.16. Zbadać, czy istnieje zależność między stopniem związania kiełbasy a jej smakowitością.

	słabo związana	związana	dobrze związana
dostateczna	9	5	3
dobra	4	12	6
b. dobra	1	6	14

1.17. W pewnym doświadczeniu chemicznym bada się grubość powłoki niklowej uzyskiwanej dla trzech różnych rodzajów kąpieli galwanicznych. Uzyskano następujące

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

wyniki. Czy na tej podstawie można powiedzieć, że grubość powłoki zależy od rodzaju kąpieli?

Grubość powłoki	Liczba pomiarów w kąpieli A	B	C
4– 8	32	51	68
8–12	123	108	80
12–16	10	26	26
16–20	41	34	28
20–24	18	20	24

1.18. W ankiecie rozesłanej wśród pracowników pewnego konsorcjum pytano, czy chcieliby zmienić obecne miejsce pracy. Uzyskano następujące wyniki. Czy chęć zmiany pracy zależy od aktualnych zarobków?

Zarobek aktualny	Odpowiedź Tak	Nie
500– 700	46	62
700– 900	94	146
900–1100	249	501
1100–1300	126	326
1300–1500	43	135
1500–1700	26	70

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

2. Elementy rachunku prawdopodobieństwa

Rachunek prawdopodobieństwa zajmuje się analizą praw rządzących zdarzeniami losowymi. Pojęciami pierwotnymi są: *zdarzenie elementarne* ω oraz *zbiór zdarzeń elementarnych* Ω.

Doświadczenie losowe to realizacja określonego zespołu warunków wraz z góry określonym zbiorem wyników.

Zdarzenie losowe A jest podzbiorem zbioru zdarzeń elementarnych Ω.

Prawdopodobieństwo (definicja aksjomatyczna) jest taką funkcją określoną na zbiorze zdarzeń losowych, że

1. $P(A) \in \langle 0, 1 \rangle$

2. $P(\Omega) = 1$

3. $P(A \cup B) = P(A) + P(B)$, o ile $A \cap B = \emptyset$

Prawdopodobieństwo (definicja klasyczna). Jeżeli Ω składa się z n jednakowo prawdopodobnych zdarzeń elementarnych, to prawdopodobieństwo zdarzenia A składającego się z k zdarzeń elementarnych wyraża się wzorem

$$P(A) = \frac{k}{n}.$$

Prawdopodobieństwo warunkowe zajścia zdarzenia A pod warunkiem realizacji zdarzenia B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (P(B) > 0).$$

Prawdopodobieństwo całkowite. Jeżeli zdarzenia $B_1, \ldots, B_n$ są takie, że $B_i \cap B_j = \emptyset$ dla wszystkich $i \neq j$, $B_1 \cup \ldots \cup B_n = \Omega$ oraz $P(B_i) > 0$ dla wszystkich i, to dla dowolnego zdarzenia A zachodzi

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n).$$

Twierdzenie Bayesa. Jeżeli zdarzenia $B_1, \ldots, B_n$ są takie, że $B_i \cap B_j = \emptyset$ dla wszystkich $i \neq j$, $B_1 \cup \ldots \cup B_n = \Omega$ oraz $P(B_i) > 0$ dla wszystkich i, to dla dowolnego takiego zdarzenia A, że $P(A) > 0$ zachodzi

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)}.$$

Niezależność zdarzeń. Zdarzenia A oraz B są niezależne, jeżeli $P(A|B) = P(A)$ oraz $P(B|A) = P(B)$. Równoważnie: $P(A \cap B) = P(A)P(B)$.

Zmienna losowa (cecha) jest funkcją określoną na zbiorze zdarzeń elementarnych o wartościach rzeczywistych. **Rozkładem zmiennej losowej** nazywamy zbiór wartości zmiennej losowej oraz prawdopodobieństwa z jakimi są te wartości przyjmowane.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Dystrybuanta F jest funkcją określoną na zbiorze liczb rzeczywistych $\mathbf{R}$ wzorem

$$F(x) = P\{X \leq x\}, \qquad x \in \mathbf{R}.$$

Najważniejsze własności dystrybuanty:

1. $0 \leq F(x) \leq 1$

2. $F(-\infty) = 0,\ F(\infty) = 1$

3. dystrybuanta jest funkcją niemalejącą

4. $P\{a < X \leq b\} = F(b) - \lim_{x \to a+} F(x)$

Funkcja gęstości rozkładu prawdopodobieństwa f jest funkcją określoną na zbiorze liczb rzeczywistych $\mathbf{R}$ wzorem

$$f(x) = \begin{cases} F'(x), & \text{jeżeli } F'(x) \text{ istnieje,} \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Najważniejsze własności funkcji gęstości:

1. $f(x) \geq 0$

2. $P\{a < X \leq b\} = \int_a^b f(x)dx$

Zmienna losowa skokowa (dyskretna) jest to zmienna, której zbiór wartości jest skończony lub przeliczalny. Jeżeli x_1 oraz x_2 są kolejnymi wartościami zmiennej losowej skokowej, to nie przyjmuje ona żadnych wartości między x_1 a x_2.

Zmienna losowa ciągła jest to zmienna przyjmująca wszystkie wartości z pewnego przedziału (najczęściej zbioru liczb rzeczywistych). Jeżeli x_1 oraz x_2 są wartościami zmiennej losowej ciągłej, to może ona przyjąć dowolną wartość między x_1 a x_2.

Wartość oczekiwana (średnia) EX zmiennej losowej X jest liczbą charakteryzującą położenie zbioru jej wartości

$$EX = \begin{cases} \sum x_i P\{X = x_i\}, & \text{dla zmiennej losowej skokowej,} \\ \int x f(x)dx, & \text{dla zmiennej losowej ciągłej.} \end{cases}$$

Wariancja D^2X zmiennej losowej jest liczbą charakteryzującą rozrzut zbioru jej wartości wokół wartości średniej EX

$$D^2X = \begin{cases} \sum (x_i - EX)^2 P\{X = x_i\}, & \text{dla zmiennej losowej skokowej,} \\ \int (x - EX)^2 f(x)dx, & \text{dla zmiennej losowej ciągłej.} \end{cases}$$

Odchylenie standardowe DX zmiennej losowej X jest liczbą charakteryzującą rozrzut zbioru jej wartości wokół wartości średniej EX

$$DX = \sqrt{D^2X}.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Kwantyl rzędu p zmiennej losowej X jest to taka liczba x_p, że

$$F(x_p) = p.$$

Frakcja. Jeżeli A jest danym podzbiorem zbioru wartości zmiennej losowej X, to frakcją nazywamy liczbę

$$p = P\{X \in A\}.$$

Zadania do samodzielnego rozwiązania

2.1. Tarcza strzelecka składa się z trzech koncentrycznych kół o promieniach odpowiednio 1, 2 i 3. Za trafienie w środkowe koło zdobywa się trzy punkty, za trafienie w kolejne pierścienie (licząc od środka koła) odpowiednio dwa i jeden punkt. Jakie jest prawdopodobieństwo uzyskania co najmniej trzech punktów w dwóch strzałach? (Zakładamy, że każdy strzał trafia w tarczę.)

2.2. W grupie studenckiej jest 20 osób. Na ćwiczeniach Student do odpowiedzi losowany jest na podstawie wyniku rzutu kostką dwudziestościenną. Jakie jest prawdopodobieństwo, że ten sam Student zostanie wyrwany do odpowiedzi trzykrotnie z rzędu?

2.3. Autobus przyjeżdża na przystanek co piętnaście minut. Jakie jest prawdopodobieństwo tego, że przychodząc na przystanek w losowym momencie będziemy czekać na autobus nie dłużej niż pięć minut?

2.4. Pan Roztargniony zapomniał ostatniej cyfry telefonu do znajomego. W związku z tym wykręcając numer telefonu ostatnią cyfrę wybiera losowo. Jakie jest prawdopodobieństwo tego, że dodzwoni się, jeżeli ma do dyspozycji cztery żetony telefoniczne?

2.5. Na egzamin przygotowanych jest 100 pytań. Student zna odpowiedź na 80 z nich. Egzaminator przerywa egzamin w chwili, gdy Student nie umie odpowiedzieć na pytanie, lecz nie później niż po piątym pytaniu. Ocena końcowa jest równa liczbie pytań, na które odpowiedział Student. Jakie jest prawdopodobieństwo tego, że Student otrzyma ocenę co najmniej dobrą?

2.6. Rzucono trzy kostki. Jakie jest prawdopodobieństwo, że przynajmniej na jednej kostce wypadnie jedynka, jeżeli na każdej kostce wypadnie inna liczba oczek?

2.7. Z talii 52 kart wyciągnięto losowo jedną kartę. Jakie jest prawdopodobieństwo, że jest to siódemka, jeżeli wiadomo, że wyciągnięta karta nie jest ani figurą ani asem?

2.8. Z talii 52 kart wyciągamy losowo jedną kartę. Rozpatrzmy zdarzenia:
A – wyciągnęliśmy asa,
B – wyciągnęliśmy kartę koloru czerwonego,
C – wyciągnęliśmy asa karo,
D – wyciągnęliśmy dziewiątkę.
Które z par zdarzeń są wzajemnie niezależne?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

2.9. Rzucamy czterokrotnie symetryczną monetą. Obliczyć prawdopodobieństwo uzyskania $0, 1, 2, 3$ oraz 4 orłów. Dane są następujące zdarzenia:
A — wypadły cztery orły,
B — wypadła parzysta liczba orłów,
C — wypadło więcej orłów niż reszek.
Obliczyć prawdopodobieństwa następujących zdarzeń:

$$P(A), P(B), P(C), P(A|B), P(B|A), P(A|C), P(C|A), P(B|C), P(C|B).$$

2.10. Troje dzieci: Ania, Basia i Czesio zmywają szklanki. Najstarsza Ania zmywa dwa razy częściej niż młodsza Basia, zaś Basia trzy razy częściej niż najmłodszy Czesio. Wiadomo, że prawdopodobieństwo zbicia szklanki w czasie zmywania wynosi dla Ani 0.01, dla Basi wynosi 0.04 natomiast dla Czesia 0.5. Jakie jest prawdopodobieństwo, że w czasie zmywania zostanie zbita jedna szklanka? Pewnego dnia po powrocie z pracy mama zauważyła, że jedna ze szklanek jest zbita, a żadne z dzieci nie chce się przyznać do zniszczenia szklanki. Które z dzieci najprawdopodobniej zmywało tego dnia?

2.11. Przedsiębiorstwo zawarło umowy z zakładami Z_1, Z_2 oraz Z_3 na dostawę podzespołów. Zakład Z_1 dostarcza 50%, zakład Z_2 dostarcza 35% natomiast zakład Z_3 dostarcza 15% potrzebnych podzespołów. Wiadomo, że 95% dostaw zakładu Z_1, 80% dostaw zakładu Z_2 oraz 85% dostaw zakładu Z_3 odpowiada wymaganiom technicznym. Jakie jest prawdopodobieństwo, że jeden wylosowany podzespół odpowiada wymaganiom technicznym? Do punktu serwisowego zgłasza się klient z urządzeniem, w którym uszkodzony jest podzespół. Jakie jest prawdopodobieństwo, że producentem zepsutego podzespołu był zakład Z_1?

2.12. Na wspólnej klasówce z matematyki spotkali się Studenci I roku z dwóch grup. W pierwszej grupie jest 15 pań oraz 10 panów, zaś w drugiej jest 12 panów i 13 pań. Prawdopodobieństwo, że pani z grupy pierwszej rozwiąże zadanie na klasówce wynosi 0.8, natomiast prawdopodobieństwo to dla pana wynosi 0.7. W drugiej grupie prawdopodobieństwa te kształtują się odpowiednio 0.9 oraz 0.85. Jak duży odsetek wszystkich Studentów rozwiąże zadanie na klasówce? Przy sprawdzaniu prac okazało się, że ktoś przygotował ściągawkę. Określić, kim najprawdopodobniej był autor ściągawki (tzn. określić płeć i grupę autora).

2.13. Wśród 300 zdających egzamin wstępny z matematyki jest 200 absolwentów klas matematyczno–fizycznych, 75 absolwentów klas ogólnokształcących oraz 25 absolwentów klas humanistycznych. Prawdopodobieństwo zdania egzaminu przez absolwenta klasy matematyczno–fizycznej wynosi 0.9, klasy ogólnokształcącej wynosi 0.25, zaś klasy humanistycznej 0.1. Jakie jest prawdopodobieństwo, że losowo wybrany przystępujący do egzaminu zda go pomyślnie? Jaki jest odsetek absolwentów klas matematyczno–fizycznych, klas humanistycznych oraz klas ogólnokształcących wśród osób, które zdały egzamin?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

3. Rozkład dwumianowy

Zmienna losowa X ma **rozkład dwumianowy** $B(n,p)$ z parametrami n oraz p, jeżeli

$$P\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Schemat Bernoulliego. Wykonujemy dwuwynikowe doświadczenie. Wyniki nazywane są umownie *sukcesem* oraz *porażką*. Prawdopodobieństwo sukcesu wynosi p. Doświadczenie wykonujemy w sposób niezależny n krotnie. Niech zmienną losową X będzie ilość sukcesów. Zmienna X ma rozkład $B(n,p)$. Zmienną losową związaną z wynikiem pojedynczego doświadczenia nazywamy **dwupunktową** i oznaczamy $D(p)$.

Rozkład dwumianowy ma następującą własność:

$$P_{n,p}\{X=k\} = P_{n,1-p}\{X=n-k\}.$$

Wynika to z dowolności nazwania jednego z dwóch możliwych wyników pojedynczego doświadczenia sukcesem, a drugiego porażką.

Oczekiwana liczba sukcesów w rozkładzie dwumianowym $B(n,p)$ wynosi np, natomiast wariancja tych wyników jest równa $np(1-p)$:

$$EX = np, \; D^2X = np(1-p).$$

Rozkład dwumianowy jest stablicowany dla typowych wartości n oraz p. W tablicy 1 podano wartości

$$Q(k; n, p) = P_{n,p}\{X \geq k\} = \sum_{i=k}^{n} P\{X=i\} \text{ dla } p \leq 0.5.$$

Dla $p > 0.5$ mamy $Q(k;n,p) = 1 - Q(n-k+1; n, 1-p)$. Dla dużych n oraz wartości p takich, że $np(1-p) > 9$ oraz $\frac{1}{n+1} < p < \frac{n}{n+1}$ rozkład dwumianowy z parametrami (n,p) przybliża się rozkładem normalnym $N(np, np(1-p))$.

Do wyznaczania prawdopodobieństw różnych zdarzeń losowych związanych ze zmienną losową o rozkładzie dwumianowym można skorzystać z dostępnej w arkuszu Excel funkcji

`ROZKŁ.DWUM(liczba_s; próby; prawdopodobieństwo_s; skumulowany)`

Argument `skumulowany` jest argumentem logicznym. Jeżeli chcemy wyznaczyć wartość funkcji rozkładu prawdopodobieństwa, to przyjmujemy `skumulowany=0`. Jeżeli chcemy wyznaczyć wartość dystrybuanty, to przyjmujemy `skumulowany=1`.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Przykład. *Wezwania pogotowia mogą być uzasadnione lub nie. Prawdopodobieństwo tego, że kolejne wezwanie będzie nieuzasadnione wynosi 5%. Obliczyć prawdopodobieństwo, że wśród kolejnych dziesięciu wezwań*

a. *co najmniej dwa będą nieuzasadnione,*

b. *dokładnie trzy będą nieuzasadnione,*

c. *co najwyżej jedno będzie nieuzasadnione.*

Rozwiązanie.

1. *Doświadczenie. Analiza zasadności kolejnego wezwania pogotowia.*

2. *Badana cecha. Obserwowaną cechą jest zasadność wezwania. Jest to cecha dwupunktowa o rozkładzie*

$$(\text{nieuzasdnione}, p = 0.05;\ \text{uzasadnione}, 1 - p = 0.95).$$

Nas interesuje zmienna losowa X opisująca liczbę nieuzasadnionych wezwań wśród dziesięciu. Ta zmienna losowa ma rozkład dwumianowy z parametrami $n = 10$ oraz $p = 0.05$.

3. *Obliczenia. W punkcie* **a.** *mamy obliczyć prawdopodobieństwo $P\{X \geq 2\}$.*

$$\begin{aligned} P\{X \geq 2\} &= P\{X = 2\} + P\{X = 3\} + \cdots + P\{X = 10\} \\ &= \sum_{i=2}^{10} P\{X = i\} = Q(2; 10, 0.05) = 0.08614. \end{aligned}$$

Ostatnia wartość została odczytana w tablicy 1.

W Excelu:

$$P\{X \geq 2\} = 1 - P\{X \leq 1\} = 1 - \texttt{ROZKŁ.DWUM}(1; 10; 0.05; 1)$$

W punkcie **b.** *mamy obliczyć prawdopodobieństwo $P\{X = 3\}$.*

$$\begin{aligned} P\{X = 3\} &= P\{X \geq 3\} - P\{X \geq 4\} = Q(3; 10, 0.05) - Q(4; 10, 0.05) \\ &= 0.01150 - 0.00103 = 0.01047. \end{aligned}$$

W Excelu:

$$P\{X = 3\} = \texttt{ROZKŁ.DWUM}(3; 10; 0.05; 0)$$

W punkcie **c.** *mamy obliczyć prawdopodobieństwo $P\{X \leq 1\}$.*

$$\begin{aligned} P\{X \leq 1\} &= 1 - P\{X \geq 2\} = 1 - Q(2; 10, 0.05) \\ &= 1 - 0.08614 = 0.91386. \end{aligned}$$

W Excelu:

$$P\{X \leq 1\} = \texttt{ROZKŁ.DWUM}(1; 10; 0.05; 1)$$

4. *Odpowiedź. Prawdopodobieństwo tego, że co najmniej dwa wezwania będą nieuzasadnione wśród kolejnych dziesięciu wynosi* 0.08614*, odnotowania trzech nieuzasadnionych wezwań wśród dziesięciu wynosi* 0.01047*, zaś odnotowania co najwyżej jednego nieuzasadnionego wezwania jest równe* 0.91386.

Zadania do samodzielnego rozwiązania

3.1. Co jest bardziej prawdopodobne: wygrać z równorzędnym przeciwnikiem trzy partie z pięciu czy dwie z trzech?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

3.2. Wyjeżdżamy na czternastodniowy urlop. Jakie jest prawdopodobieństwo tego, że będziemy mieli dziesięć dni pięknej pogody, jeżeli prawdopodobieństwo niepogodnego dnia wynosi $1/6$?

3.3. Załóżmy, że prawdziwa jest hipoteza Mendla, iż dla krzyżówki grochu w drugim pokoleniu stosunek nasion żółtych do zielonych jest jak $3 : 1$. Wylosowano dziesięć nasion. Obliczyć prawdopodobieństwo, że będą co najwyżej cztery nasiona żółte.

3.4. Środek owadobójczy zabija przeciętnie 90% owadów. Środek ten zastosowano na dziesięciu owadach. Obliczyć prawdopodobieństwo, że co najwyżej dwa osobniki przeżyją.

3.5. Wadliwość procesu produkcyjnego wynosi 10%. Obliczyć prawdopodobieństwo, że na osiem wylosowanych produktów będą co najwyżej dwa złe.

3.6. W pewnym gatunku zwierząt prawdopodobieństwo urodzenia osobnika płci męskiej wynosi 0.6. Obliczyć prawdopodobieństwo, że w miocie, w którym urodziło się pięcioro młodych będą co najmniej cztery osobniki męskie.

3.7. W stawie hodowlanym są dwa gatunki ryb w proporcji $8 : 2$. Obliczyć prawdopodobieństwo, że wśród dziesięciu złowionych ryb będzie co najmniej siedem ryb liczniejszego gatunku.

3.8. W jeziorze jest tysiąc ryb, w tym sto ryb zaobrączkowanych. Obliczyć prawdopodobieństwo, że wśród dziesięciu złowionych ryb będzie co najmniej siedem ryb zaobrączkowanych.

3.9. Właściciel kurzej fermy stwierdził, że kogutków wykluwa się trzy razy więcej niż kurek. Obliczyć prawdopodobieństwo, że z pięciu losowo wybranych jajek wykluje się co najmniej jeden kogutek, ale nie mniej niż dwie kurki.

3.10. Producent podaje, że w co czwartym jajku niespodziance znajduje się zajączek Ribbon. Jakie jest prawdopodobieństwo, że wśród dwudziestu kupionych jajek jest a) przynajmniej pięć jajek z zajączkiem Ribbon; b) nie więcej niż piętnaście jajek bez zajączka. Jaka jest najbardziej prawdopodobna ilość jajek z zajączkami?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

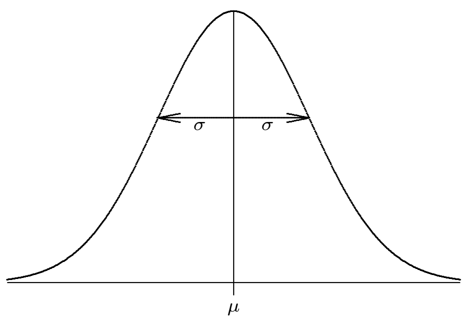
4. Rozkład normalny

Zmienna losowa X o wartości oczekiwanej μ i wariancji σ^2 ma **rozkład normalny** $N(\mu, \sigma^2)$, jeżeli funkcja gęstości jej rozkładu prawdopodobieństwa wyraża się wzorem

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}.$$

Jeżeli X ma rozkład $N(\mu, \sigma^2)$, to zmienna losowa $Z = (X-\mu)/\sigma$ ma rozkład $N(0, 1)$ zwany *standardowym rozkładem normalnym*. Gęstość i dystrybuanta tego rozkładu wyrażają się wzorami

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad F(x) = \int_{-\infty}^{x} f(t)dt.$$



Dystrybuanta $F(x)$ standardowego rozkładu normalnego dla $x \geq 0$ jest stablicowana. Dla $x < 0$ zachodzi $F(x) = 1 - F(-x)$. W tablicy 2 podano pięć pierwszych miejsc po przecinku wartości dystrybuanty $F(x)$ dla $x \geq 0$. W tablicy 3 podano wartości kwantyli rzędu $\alpha \geq 0.5$ rozkładu $N(0, 1)$. Dla $\alpha < 0.5$ mamy $u_\alpha = -u_{1-\alpha}$.

Prawdziwy jest następujący wzór

$$P\{X \in (a, b)\} = P\left\{Z \in \left(\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}\right)\right\} = F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right).$$

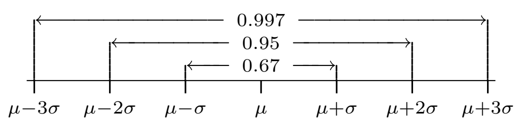
W celu wyznaczenia tego prawdopodobieństwa można posłużyć się arkuszem kalkulacyjnym Excel, w którym dostępna jest funkcja

`ROZKŁ.NORMALNY(x; średnia; odchylenie_std; skumulowany)`

Argument `skumulowany` jest argumentem logicznym. Jeżeli chcemy wyznaczyć wartość funkcji gęstości, to przyjmujemy `skumulowany=0`. Jeżeli chcemy wyznaczyć wartość dystrybuanty, to przyjmujemy `skumulowany=1`.

Dla zmiennej losowej o rozkładzie $N(\mu, \sigma^2)$ zachodzi prawo trzech sigm

$$P\{|X-\mu| < \sigma\} \approx 0.67$$
$$P\{|X-\mu| < 2\sigma\} \approx 0.95$$
$$P\{|X-\mu| < 3\sigma\} \approx 0.997$$



Przykład. *W pewnym doświadczeniu medycznym bada się czas snu (w minutach) pacjentów chorych na pewną chorobę. Można przyjąć, że czas snu ma rozkład normalny ze średnią* 500 *minut i odchyleniem standardowym* 20 *minut. Długość snu uważa się za prawidłową, jeżeli mieści się w przedziale* $(460, 520)$*. Jaki jest odsetek pacjentów o prawidłowej długości snu?*

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Rozwiązanie.

1. *Doświadczenie. Badanie czasu snu pacjentów chorych na pewną chorobę.*

2. *Badana cecha. Obserwowaną cechą X jest czas snu mierzony w minutach. Można przyjąć, że jest to cecha ciągła o rozkładzie normalnym $N(500, 20^2)$.*

3. *Obliczenia. Interesuje nas obliczenie prawdopodobieństwa tego, że czas snu losowo wybranego pacjenta będzie prawidłowy, czyli $P\{X \in (460, 520)\}$. Stosując podany wyżej wzór otrzymujemy:*

$$\begin{aligned} P\{X \in (460, 520)\} &= F\left(\frac{520-500}{20}\right) - F\left(\frac{460-500}{20}\right) = F(1) - F(-2) \\ &= F(1) - (1 - F(2)) = F(2) + F(1) - 1 \\ &= 0.97725 + 0.84134 - 1 = 0.81859. \end{aligned}$$

W Excelu:

$$P\{X \in (460, 520)\} = \\ \texttt{ROZKŁ.NORMALNY}(520; 500; 20; 1) - \texttt{ROZKŁ.NORMALNY}(460; 500; 20; 1)$$

4. *Odpowiedź. Odsetek pacjentów o prawidłowej długości snu wynosi* $0.81859 \times 100\% \approx 82\%$.

Zadania do samodzielnego rozwiązania

4.1. Niech X będzie zmienną losową o rozkładzie $N(-5, 100)$. Obliczyć:

$$P\{X \leq -9\},\ P\{X \in (-7, 1)\},\ P\{X \geq -7\},\ P\{|X + 5| \leq 10\}.$$

4.2. Niech X będzie zmienną losową o rozkładzie $N(10, 25)$. Obliczyć:

$$P\{X \leq 8\},\ P\{X \in (9, 13)\},\ P\{X \geq 9\},\ P\{|X - 10| \leq 5\}.$$

4.3. Wzrost kobiety jest zmienną losową o rozkładzie $N(158, 100)$. Obliczyć jaki jest procent kobiet o wzroście pomiędzy 148 a 168.

4.4. Dzienna mleczność krowy jest zmienną losową o rozkładzie $N(10, 4)$. Jaki jest procent krów o mleczności mniejszej niż 7, a jaki jest procent krów o mleczności większej niż 12.

4.5. Plon ziemniaka z poletka jest zmienną losową o rozkładzie $N(20, 25)$. Jaki procent poletek da plon między 18 a 23.

4.6. Ciężar jajek dostarczanych do skupu ma rozkład normalny ze średnią 2 dag i wariancją 0.01. Jajko kwalifikuje się do pierwszego gatunku, jeżeli jego waga wynosi co najmniej 2.096 g. Jaki procent jajek dostarczanych do skupu można uznać za jajka pierwszego gatunku?

4.7. Średni czas żarzenia się liści tytoniu wynosi 17 sekund. Liście tlące się krócej niż 12 sekund są dyskwalifikowane. Jaki jest procent liści przydatnych do produkcji, jeżeli wariancja wspomnianej cechy wynosi 6.25?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

4.8. Przyjmując, że przeciętna waga (w kilogramach) noworodka jest zmienną losową o rozkładzie $N(3, 0.25)$ określić procent noworodków o wadze z przedziału $(3, 3.5)$.

4.9. Stwierdzono, że 80% ludzi o IQ powyżej 90 jest w stanie nauczyć się posługiwać pewnym urządzeniem. Jaki jest to procent całej populacji jeśli wiadomo, że iloraz inteligencji jest cechą o rozkładzie $N(100, 100)$?

4.10. Aby zdać egzamin ze statystyki należy prawidłowo rozwiązać co najmniej 70% zadań z testu egzaminacyjnego. Przyjmując, że wyniki testu dla Studentów zdających w pierwszym terminie mają rozkład normalny ze średnią 76% i odchyleniem standardowym 8.0%, obliczyć jaki procent Studentów zda egzamin w pierwszym terminie.

4.11. Automat tokarski produkuje nity, których średnica ma rozkład normalny z odchyleniem standardowym 0.04 mm. Wartość średnia tej zmiennej losowej może być dowolnie regulowana przez odpowiednie ustawienie automatu. Nit uważa się za dobry, jeżeli jego średnica mieści się w przedziale $(2.9, 3.1)$. Jakie jest prawdopodobieństwo wyprodukowania braku, gdy automat tokarski ustawiony jest tak, że średnia średnica jest równa 3.05 mm? Jak powinien być ustawiony automat, by wadliwość procesu produkcyjnego była najmniejsza?

4.12. Automat produkuje nity, których średnica jest zmienną losową o rozkładzie $N(2, 0.01)$. Dla jakiego ε średnica nitu z prawdopodobieństwem 0.95 znajduje się w przedziale $(2-\varepsilon, 2+\varepsilon)$?

4.13. Obliczyć odchylenie standardowe przyrządu pomiarowego o którym wiadomo, że z prawdopodobieństwem 0.95 daje błąd nie przekraczający trzech jednostek. Zakładamy, że rozkład błędu jest normalny z wartością średnią zero.

4.14. Cecha X o rozkładzie normalnym przyjmuje wartości z przedziału $\langle 28, 46\rangle$ (przedział ten ustalony został na mocy prawa trzech sigm). Wartości $40-46$ zaliczamy do pierwszej klasy, wartości $28-34$ do trzeciej klasy. Jakie jest prawdopodobieństwo tego, że losowo wybrana wartość cechy X należy do pierwszej lub trzeciej klasy?

4.15. Udowodnić prawo trzech sigm.

Rozkład normalny i dwumianowy

Przykład. *Błąd pomiaru pewnym urządzeniem ma rozkład normalny o wartości średniej zero (pomiar nieobciążony) i odchyleniu standardowym 20 mm. Obliczyć prawdopodobieństwo, że błąd przynajmniej jednego z trzech niezależnych pomiarów nie przekroczy 4 mm.*

Rozwiązanie. *W przykładzie mamy do czynienia z dwoma mechanizmami losowymi. Jeden z nich związany jest z błędem pomiaru, zaś drugi z liczbą popełnionych „dużych" błędów w trakcie wykonywania trzech pomiarów. Niech X oznacza liczbę pomiarów o błędzie nie przekraczającym 4 mm (wśród trzech wykonywanych pomiarów). Jest to oczywiście zmienna losowa o rozkładzie dwumianowym z $n = 3$ oraz prawdopodobieństwem sukcesu p. Jako sukces potraktujemy otrzymanie pomiaru o błędzie nie przekraczającym 4 mm. Jeżeli przez Y oznaczymy zmienną losową reprezentującą wielkość błędu, to $p = P\{|Y| < 4\}$, przy czym wiemy, że zmienna losowa Y ma rozkład normalny o wartości oczekiwanej $\mu = 0$ oraz odchyleniu standardowym $\sigma = 20$. Możemy zatem obliczyć prawdopodobieństwo sukcesu:*

$$p = P\{|Y| < 4\} = F\left(\frac{4-0}{20}\right) - F\left(\frac{-4-0}{20}\right) = F(0.2) - F(-0.2) = 0.15852.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Możemy teraz przystąpić do wyznaczenia prawdopodobieństwa tego, że błąd przynajmniej jednego z trzech niezależnych pomiarów nie przekroczy 4 *mm. W terminach zmiennej losowej* X *oznaczającej liczbę pomiarów o błędzie mniejszym niż* 4 *mm, poszukiwane prawdopodobieństwo możemy zapisać jako* $P\{X \geq 1\}$. *Przypomnijmy, że zmienna losowa* X *ma rozkład dwumianowy* $B(3, 0.15852)$. *A zatem*

$$P\{X \geq 1\} = 1 - P\{X = 0\} = 1 - (1-p)^3 = 0.40146.$$

Odpowiedź. Prawdopodobieństwo, że błąd przynajmniej jednego z trzech niezależnych pomiarów nie przekroczy 4 *mm wynosi* 0.40146.

Zadania do samodzielnego rozwiązania

4.16. Błąd pomiaru pewnym urządzeniem ma rozkład normalny o wartości średniej zero (pomiar nieobciążony) i odchyleniu standardowym 20 mm. Ile co najmniej należy wykonać pomiarów, by z prawdopodobieństwem co najmniej 0.99 błąd przynajmniej jednego z nich nie przekroczył 4 mm.

4.17. Zawartość tłuszczu w mleku pewnej rasy krów ma rozkład normalny o wartości średniej 5% i wariancji 4. Mleko uważa się za bardzo tłuste, jeżeli zawartość tłuszczu przekracza 7%. Obliczyć prawdopodobieństwo, że przynajmniej jedna z trzech niezależnych próbek mleka będzie uznana za bardzo tłustą.

4.18. Wzrost dzieci jest zmienną losową o rozkładzie normalnym o wartości średniej 110 cm i wariancji 400. Obliczyć prawdopodobieństwo, że przynajmniej jedno z trójki losowo wybranych dzieci będzie miało wzrost większy od przeciętnej.

4.19. Średnica nitu ma rozkład normalny o wartości oczekiwanej 2 mm i wariancji 0.01. Obliczyć prawdopodobieństwo, że wśród trzech losowo pobranych nitów wszystkie okażą się brakami, jeżeli za brak uważany jest nit o średnicy mniejszej niż 1.8 mm lub większej niż 2.2 mm.

4.20. Ciężar jajka kurzego jest zmienną losową o rozkładzie normalnym o wartości średniej 10 g i wariancji 4. Obliczyć prawdopodobieństwo, że z trzech losowo wybranych jajek wszystkie będą miały ciężar poniżej przeciętnej.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

5. Analiza jednej populacji

Obserwujemy cechę X w pewnej populacji. Naszym celem będzie oszacowanie charakterystyk badanej cechy w obserwowanej zbiorowości oraz porównanie tych charakterystyk z pewnymi ustalonymi wartościami. Analiza ta będzie oparta na próbie $X_1, \ldots, X_n$.

Wnioskowanie o wartości średniej. Zakładamy, że obserwowana cecha X ma w badanej zbiorowości rozkład normalny $N(\mu, \sigma^2)$ o nieznanej wariancji σ^2. Zadanie polega na ocenie wartości średniej μ. W tym celu na podstawie próby konstruowany jest przedział ufności na poziomie ufności $1 - \alpha$:

$$\left(\bar{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}}\right),$$

gdzie $\bar{X}$ oznaczające średnią arytmetyczną obserwacji:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

jest punktową oceną wartości średniej μ,

$$S = \sqrt{S^2}, \quad S^2 = \frac{\mathrm{var}X}{n-1}, \quad \mathrm{var}X = \sum_{i=1}^{n}(X_i - \bar{X})^2,$$

jest punktową oceną odchylenia standardowego σ oraz $t(\alpha; n-1)$ jest wartością krytyczną rozkładu t z $n-1$ stopniami swobody (tablica 4). Estymator S^2 jest punktową oceną wariancji σ^2. Wielkość $\mathrm{var}X = \sum_{i=1}^{n}(X_i - \bar{X})^2$ oznacza sumę kwadratów odchyleń od średniej.

Wszystkie obliczenia można wykonać w arkuszu kalkulacyjnym Excel wykorzystując następujące jego funkcje:

`ŚREDNIA(obszar)`	średnia arytmetyczna $\bar{X}$
`ODCH.STANDARD.PRÓBKI(obszar)`	odchylenie standardowe S
`ROZKŁ.T.ODWR.DS`(α; `stopnie_swobody`)	wartość krytyczna rozkładu t

W pewnych zagadnieniach praktycznych interesuje nas zgodność wartości średniej μ z pewną z góry daną liczbą μ_0 (daną np. w normach). W tym celu weryfikowana jest hipoteza $H_0 : \mu = \mu_0$ za pomocą statystyki testowej

$$t_{\mathrm{emp}} = \frac{\bar{X} - \mu_0}{S}\sqrt{n}.$$

Jeżeli $|t_{\mathrm{emp}}| > t(\alpha; n-1)$, to testowana hipoteza jest odrzucana na poziomie istotności α. Jeżeli $|t_{\mathrm{emp}}| < t(\alpha; n-1)$, to nie ma podstaw do odrzucenia hipotezy.

Korzystając z arkusza kalkulacyjnego Excel można zastosować inny sposób wnioskowania. Za pomocą funkcji

`ROZKŁ.T.DS`(t_{emp}; `stopnie_swobody`)

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

można obliczyć poziom krytyczny testu (p-wartość). Jeżeli uzyskany wynik jest mniejszy od zadanego poziomu istotności, to weryfikowaną hipotezę należy odrzucić. W przeciwnym przypadku nie ma podstaw do jej odrzucenia.

Warto zauważyć, że hipoteza $H_0 : \mu = \mu_0$ nie jest odrzucana na poziomie istotności α tylko wtedy, gdy hipotetyczna wartość μ_0 leży w przedziale ufności (na poziomie ufności $1-\alpha$) dla wartości średniej μ.

Jeżeli wariancja σ^2 rozkładu cechy w populacji jest znana, to w powyższych postępowaniach zamiast oceny S wykorzystuje się odchylenie standardowe σ oraz przyjmuje się nieskończoną liczbę stopni swobody. Z właściwości rozkładu t wynika, że $t(\alpha;\infty) = u_{1-\alpha/2}$, gdzie u_q oznacza q–ty kwantyl rozkładu normalnego $N(0,1)$ (tablica 3).

Wnioskowanie o wariancji. Zakładamy, że obserwowana cecha X ma w badanej zbiorowości rozkład normalny $N(\mu,\sigma^2)$ o nieznanej średniej μ. Zadanie polega na ocenie wariancji σ^2. W tym celu na podstawie próby konstruowany jest przedział ufności na poziomie ufności $1-\alpha$:

$$\left(\frac{\mathrm{var}X}{\chi^2\left(\frac{\alpha}{2};n-1\right)}, \frac{\mathrm{var}X}{\chi^2\left(1-\frac{\alpha}{2};n-1\right)}\right),$$

gdzie $\chi^2(\alpha;n-1)$ oznacza wartość krytyczną rozkładu chi–kwadrat z $n-1$ stopniami swobody (tablica 5).

Obliczenia można wykonać w arkuszu kalkulacyjnym Excel wykorzystując następujące jego funkcje:

`WARIANCJA.PRÓBKI(obszar)`	wariancja S^2
`ROZKŁAD.CHI.ODW(prawd; stopnie_swobody)`	wartość krytyczna rozkładu chi-kwadrat

W pewnych zagadnieniach praktycznych interesuje nas zgodność wariancji σ^2 z pewną z góry daną liczbą σ_0^2 (daną np. w normach). W tym celu weryfikowana jest hipoteza $H_0 : \sigma^2 = \sigma_0^2$ za pomocą statystyki testowej

$$\chi^2_{\mathrm{emp}} = \frac{\mathrm{var}X}{\sigma_0^2}.$$

Jeżeli $\chi^2_{\mathrm{emp}} > \chi^2\left(\frac{\alpha}{2};n-1\right)$ lub $\chi^2_{\mathrm{emp}} < \chi^2\left(1-\frac{\alpha}{2};n-1\right)$, to testowana hipoteza jest odrzucana na poziomie istotności α. W przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy.

Korzystając z arkusza kalkulacyjnego Excel można zastosować inny sposób wnioskowania. Za pomocą funkcji

$$1 - \mathtt{ROZKŁ.CHI}(\chi^2_{\mathrm{emp}};\mathtt{stopnie\_swobody})$$

można obliczyć poziom krytyczny testu (p-wartość). Jeżeli uzyskany wynik jest mniejszy od zadanego poziomu istotności, to weryfikowaną hipotezę należy odrzucić. W przeciwnym przypadku nie ma podstaw do jej odrzucenia.

Zauważmy, że hipoteza $H_0 : \sigma^2 = \sigma_0^2$ nie jest odrzucana na poziomie istotności α tylko wtedy, gdy hipotetyczna wartość σ_0^2 leży w przedziale ufności (na poziomie ufności $1-\alpha$) dla wariancji σ.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Jeżeli średnia μ rozkładu cechy w populacji jest znana, to w powyższych postępowaniach zamiast oceny $\bar{X}$ wartości średniej wykorzystuje się znaną wartość μ oraz przyjmuje się liczbę stopni swobody równą n.

Wnioskowanie o frakcji. Obserwowana cecha X ma rozkład dwupunktowy o prawdopodobieństwie sukcesu p. Zadanie polega na ocenie prawdopodobieństwa p na podstawie n elementowej próby, w której zaobserwowano k sukcesów. Punktową oceną jest $\hat{p} = \frac{k}{n}$. Przedział ufności na poziomie ufności $1-\alpha$ ma postać

$$\left(\beta^{-1}\left(k+1, n-k; \frac{\alpha}{2}\right),\ \beta^{-1}\left(k, n-k+1; 1-\frac{\alpha}{2}\right)\right),$$

gdzie $\beta^{-1}(\cdot, \cdot; \cdot)$ oznacza kwantyl rozkładu Beta. Dla $k=0$ lewy koniec przyjmujemy 0, a dla $k=n$ prawy koniec przyjmujemy 1.

Końce przedziału ufności można także wyznaczyć ze wzoru:

$$\left(\frac{k}{k+(n-k+1)F\left(\frac{\alpha}{2}; 2(n-k+1), 2k\right)}, \frac{(k+1)F\left(\frac{\alpha}{2}; 2(k+1), 2(n-k)\right)}{n-k+(k+1)F\left(\frac{\alpha}{2}; 2(k+1), 2(n-k)\right)}\right)$$

gdzie $F(\alpha; r_1, r_2)$ jest wartością krytyczną rozkładu F o (r_1, r_2) stopniach swobody (tablica 6). Dla małych wartości n końce przedziału ufności są stablicowane (Zieliński 2000, Zieliński i Zieliński 1990)

Obliczenia można wykonać w arkuszu kalkulacyjnym Excel wykorzystując funkcję wyznaczającą kwantyle rozkładu beta:

$$\texttt{ROZKŁAD.BETA.ODW}(\texttt{prawd}; \texttt{a}, \texttt{b}).$$

Dla dużych wartości n przybliżony przedział ufności można wyznaczać w następujący sposób

$$\left(\hat{p} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

gdzie u_q jest kwantylem rzędu q rozkładu normalnego $N(0,1)$ (tablica 3).

W celu stwierdzenia, czy nieznana wartość p prawdopodobieństwa sukcesu jest daną liczbą p_0, weryfikowana jest hipoteza $H_0: p = p_0$. W przypadku dużej ilości n obserwacji stosuje się statystykę testową

$$u_{\mathrm{emp}} = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} = \frac{k-np_0}{\sqrt{n\hat{p}(1-\hat{p})}}.$$

Hipoteza jest odrzucana na poziomie istotności α, jeżeli $|u_{\mathrm{emp}}| > u_{1-\alpha/2}$.

Poziom krytyczny (p-wartość) w tym teście można obliczyć w Excelu w następujący sposób:

$$\texttt{ROZKŁAD.NORMALNY.S}(\texttt{MODUŁ.LICZBY}(u_{\mathrm{emp}})).$$

Zauważmy, że hipoteza $H_0: p = p_0$ nie jest odrzucana na poziomie istotności α tylko wtedy, gdy hipotetyczna wartość p_0 leży w przedziale ufności (na poziomie ufności

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

$1-\alpha$) dla prawdopodobieństwa p. Ten sposób wnioskowania można stosować zarówno dla małych, jak i dużych prób.

Przykład. *Podejrzewa się, że jedną z przyczyn nieprawidłowości rozwoju dzieci jest spędzanie zbyt dużej ilości czasu przed telewizorem. W celu oszacowania przeciętnego czasu spędzanego przed telewizorem wylosowano grupę dwudziestu dzieci w wieku 8−9 lat i zarejestrowano następujące czasy oglądania telewizji:* 107.8, 149.9, 132.1, 114.7, 51.1, 97.2, 95.7, 117.3, 119.5, 122.2, 65.7, 109.8, 84.0, 84.9, 134.1, 133.8, 49.5, 99.2, 116.5, 135.9. *Oszacować na tej podstawie średni czas spędzany przed telewizorem.*

Rozwiązanie.

1. *Badana zbiorowość. Dzieci w wieku* $8-9$ *lat.*

2. *Badana cecha. Cecha* X *— czas spędzany przez dzieci na oglądaniu telewizji. Jest to cecha ciągła.*

3. *Założenia. Cecha* X *ma rozkład normalny* $N(\mu,\sigma^2)$. *Liczba* μ *reprezentuje przeciętny czas oglądania telewizji, zaś* σ^2 *zróżnicowanie.*

4. *Cel. Zgodnie z wcześniejszymi ustaleniami liczba* μ *reprezentuje przeciętny czas oglądania telewizji. Zadanie sprowadza się do oszacowania liczby* μ.

5. *Procedura statystyczna. Liczbę* μ *oszacujemy wyznaczając przedział ufności:*

$$\left(\bar{X} - t(\alpha; n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t(\alpha; n-1)\frac{S}{\sqrt{n}}\right).$$

6. *Poziom ufności. Przyjmujemy* $1-\alpha = 0.95$.

7. *Wartość krytyczna. Ponieważ przyjęliśmy poziom ufności* $1-\alpha = 0.95$, *więc w tablicach wartości krytycznych rozkładu* t *odszukujemy* $t(0.05; 19) = 2.093$.

8. *Obliczenia.*

$$n = 20, \quad \sum_{i=1}^{n} X_i = 2120.9, \quad \sum_{i=1}^{n} X_i^2 = 239741.$$

Zatem

$$\bar{X} = \frac{2120.9}{20} = 106.045,$$

$$\mathrm{var}X = \sum_{i=1}^{n} X_i^2 - n(\bar{X})^2 = 239741 - 20 \cdot 106.045^2 = 14830.17,$$

$$S = \sqrt{S^2} = \sqrt{\frac{\mathrm{var}X}{n-1}} = \sqrt{\frac{14830.17}{20-1}} = 27.938.$$

Możemy teraz obliczyć granice przedziału ufności:

$$\left(106.045 - 2.093 \cdot \frac{27.938}{\sqrt{20}},\ 106.045 + 2.093 \cdot \frac{27.938}{\sqrt{20}}\right) = (92.97,\ 119.12).$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Te same obliczenia w Excelu.

	A	*B*	*C*
1	*107.8*	*poziom ufności*	`0.95`
2	*149.9*	*liczba obserwacji*	`20`
3	*132.1*	*średnia z próby*	`=ŚREDNIA(A1:A20)`
4	*114.7*	*odchylenie standardowe*	`=ODCH.STANDARD.PRÓBKI(A1:A20)`
5	*51.1*	*wartość krytyczna*	`=ROZKŁ.T.ODWR.DS(1-C1;C2-1)`
6	*97.2*	*lewy koniec*	`=C3-C5*C4/PIERWIASTEK(C2)`
7	*95.7*	*prawy koniec*	`=C3+C5*C4/PIERWIASTEK(C2)`
⋮	⋮		
20	*135.9*		

9. *Wniosek. Na poziomie ufności* $1-\alpha = 0.95$ *liczba* μ *zawiera się w przedziale* $(92.97,\ 119.12)$

10. *Odpowiedź. Na podstawie wyników możemy stwierdzić, że przeciętny czas spędzany na oglądaniu telewizji przez dzieci w wieku* $8-9$ *lat wyraża się liczbą z przedziału* $(92.97,\ 119.12)$ *z* 95% *zaufaniem do wniosku.*

Przykład. *Postawiono zadanie oszacowania odsetka zgłaszających się do szpitala pacjentów mających grupę krwi „0". W tym celu spośród pacjentów szpitala wylosowano* 1000 *wśród których* 240 *miało grupę krwi „0". Oszacować na tej podstawie odsetek pacjentów z tą grupą krwi.*

Rozwiązanie.

1. *Badana zbiorowość. Pacjenci szpitala.*

2. *Badana cecha. Cecha* X *– grupa krwi. Jest to cecha dwuwartościowa: grupa* $,,0''(1)$, *inna grupa*(0).

3. *Założenia. Cecha* X *ma rozkład dwupunktowy* $D(p)$. *Liczba* p *reprezentuje prawdopodobieństwo przyjęcia przez cechę* X *wartości* 1.

4. *Cel. Zgodnie z wcześniejszymi ustaleniami liczba* p *reprezentuje odsetek pacjentów z grupą krwi „0". Zadanie sprowadza się do oszacowania liczby* p.

5. *Procedura statystyczna. Liczbę* p *oszacujemy wyznaczając zarówno przybliżony, jak i dokładny przedział ufności.*

6. *Poziom ufności. Przyjmujemy* $1-\alpha = 0.95$.

7. *Wartość krytyczna. Ponieważ przyjęliśmy poziom ufności* $1-\alpha = 0.95$, *więc w tablicach kwantyli rozkładu normalnego odszukujemy* $u_{1-0.05/2} = 1.96$.

8. *Obliczenia. Wyznaczenie przybliżonego przedziału ufności.*

$$n = 1000, \quad k = 240, \quad \widehat{p} = \frac{240}{1000} = 0.24,$$

zatem

$$\left(0.24 - 1.96\sqrt{\frac{0.24\cdot(1-0.24)}{1000}},\ 0.24 + 1.96\sqrt{\frac{0.24\cdot(1-0.24)}{1000}}\right) = (0.2135, 0.2665)\,.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Obliczenia w Excelu (dokładny przedział ufności).

	A	*B*
1	*poziom ufności*	`0.95`
2	*liczba wszystkich pacjentów*	`1000`
3	*liczba pacjentów z grupą „0”*	`240`
4	*lewy koniec*	`=ROZKŁAD.BETA.ODW((1-B1)/2;B3+1;B2-B3)`
5	*prawy koniec*	`=ROZKŁAD.BETA.ODW((1+B1)/2;B3;B2-B3+1)`

9. *Wniosek. Na poziomie ufności* $1-\alpha = 0.95$ *liczba* p *zawiera się w przedziale* $(0.2135,\ 0.2665)$.

10. *Odpowiedź. Na podstawie wyników badań możemy stwierdzić, że odsetek pacjentów z grupą krwi „0” wyraża się liczbą z przedziału* $(21.35\%,\ 26.65\%)$ *z* 95% *zaufaniem do wniosku.*

Zadania do samodzielnego rozwiązania

5.1. Pokazać, że

$$\mathrm{var}X = \sum_{i=1}^{n} X_i^2 - n(\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2.$$

5.2. Pokazać, że hipoteza $H_0 : \mu = \mu_0$ nie jest odrzucana na poziomie istotności α tylko wtedy, gdy hipotetyczna wartość μ_0 leży w przedziale ufności (na poziomie ufności $1-\alpha$) dla wartości średniej μ.

5.3. Pokazać, że hipoteza $H_0 : \sigma^2 = \sigma_0^2$ nie jest odrzucana na poziomie istotności α tylko wtedy, gdy hipotetyczna wartość σ_0^2 leży w przedziale ufności (na poziomie ufności $1-\alpha$) dla wariancji σ^2.

5.4. Pokazać, że hipoteza $H_0 : p = p_0$ nie jest odrzucana na poziomie istotności α tylko wtedy, gdy hipotetyczna wartość p_0 leży w przedziale ufności (na poziomie ufności $1-\alpha$) dla prawdopodobieństwa p.

5.5. Wśród losowo wybranych osób przeprowadzono ankietę na temat czasu dojazdu do pracy. Wyniki ankiety przedstawiały się następująco: 28.63, 16.45, 22.34, 26.02, 22.72, 26.90, 25.82, 28.49, 28.46, 27.54, 31.28, 26.24, 20.69, 30.39, 20.55, 22.56, 27.37, 22.77, 23.99, 22.33. Zakładając normalność rozkładu badanej cechy, na podstawie wyników ankiety oszacować średni czas dojazdu do pracy.

5.6. Przy wycenie użytkowności mlecznej buhaja wzięto pod uwagę mleczność jego córek w okresie pierwszych stu dni laktacji. Uzyskano następujące wyniki oznaczania zawartości białka (w procentach): 3.14, 2.68, 2.90, 3.04, 2.91, 3.07, 3.03, 3.13, 3.13, 3.10, 3.24, 3.05, 2.84, 3.20, 2.83. Oszacować na tej podstawie średnią zawartość białka w mleku. Przyjąć, że badana cecha ma w populacji rozkład normalny.

5.7. Przez dwieście dni obserwowano liczbę znoszonych jaj przez kury i uzyskano następujące wyniki: 249, 128, 187, 223, 190, 232, 221, 248, 247, 238, 275, 225, 170, 267, 169, 189, 237, 191, 203, 186, 224, 244, 233, 224, 143, 184, 215, 253, 227, 168. Na podstawie zebranych danych oszacować przeciętną dzienną liczbę jaj znoszonych przez kury.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

5.8. Na podstawie zaobserwowanych mleczności krów oszacować średnią mleczność krowy: 10.181, 9.573, 9.867, 10.051, 9.886, 10.095, 10.041, 10.174, 10.173, 10.127, 10.314, 10.062, 9.784, 10.269, 9.778, 9.878, 10.118, 9.889, 9.950, 9.867, 10.055, 10.157, 10.099, 10.055, 9.650, 9.855, 10.011, 10.200, 10.069, 9.773, 10.003, 9.973, 9.838, 10.323, 9.979, 9.709, 9.837, 9.785, 9.804, 9.901.

5.9. Notowano wielkość suchej masy roślin pewnej odmiany jęczmienia i uzyskano wyniki: 1.181, 0.573, 0.867, 1.051, 0.886, 1.095, 1.041, 1.174, 1.173, 1.127, 1.314, 1.062, 0.784, 1.269, 0.778, 0.878, 1.118, 0.889, 0.950, 0.867, 1.055, 1.157, 1.099, 1.055, 0.650, 0.855, 1.011, 1.200, 1.069, 0.773, 1.003, 0.973, 0.838, 1.323, 0.979, 0.709, 0.837, 0.785, 0.804, 0.901. Oszacować przeciętną suchą masę rośliny.

5.10. Z poletek zebrano plon pewnego zboża i uzyskano wyniki: 21.8, 15.7, 18.7, 20.5, 18.9, 21.0, 20.4, 21.7, 21.7, 21.3, 23.1, 20.6, 17.8, 22.7, 17.8, 18.8, 21.2, 18.9, 19.5, 18.7, 20.6, 21.6, 21.0, 20.5, 16.5, 18.6, 20.1, 22.0, 20.7, 17.7. Oszacować przeciętny plon z poletka.

5.11. Należy oszacować średnią żywotność (w godzinach świecenia) wyprodukowanej partii świetlówek. Wiadomo, że czas świecenia świetlówek ma rozkład normalny. Wylosowana z tej partii próba świetlówek dała następujące rezultaty: 2909, 2544, 2720, 2831, 2732, 2857, 2825, 2905, 2904, 2876, 2988, 2837, 2671, 2962, 2667, 2727, 2871, 2733, 2770, 2720, 2833, 2894, 2859, 2833, 2590, 2713, 2807, 2920, 2842, 2664, 2802, 2784, 2703, 2994, 2788, 2626, 2702, 2671, 2683, 2741, 2894, 2873, 2880, 2645, 2794.

5.12. W pewnym doświadczeniu medycznym bada się czas snu (w minutach) pacjentów chorych na pewną chorobę. Można przyjąć, że czas snu ma rozkład normalny. Należy oszacować średni czas snu pacjentów, jeżeli pomiary w grupie pacjentów dały następujące rezultaty: 512, 299, 402, 466, 408, 482, 463, 509, 509, 493, 558, 470, 373, 543, 370, 406, 490, 409, 431, 402.

5.13. Ciężar jajka kurzego jest zmienną losową o rozkładzie normalnym. Na podstawie następujących pomiarów oszacować średni ciężar jaja kurzego: 6.87, 5.63, 6.23, 6.60, 6.27, 6.69, 6.58, 6.86, 6.85, 6.76, 7.14, 6.63, 6.06, 7.05, 6.05, 6.25, 6.74, 6.27, 6.40, 6.23, 6.61, 6.82, 6.70, 6.61, 5.79, 6.20, 6.52, 6.91, 6.64, 6.04, 6.51, 6.44, 6.17, 7.16, 6.46, 5.91, 6.17, 6.06, 6.10, 6.30, 6.82, 6.75, 6.77, 5.97, 6.48, 6.51, 6.70, 6.78, 6.55, 7.12.

5.14. Zawartość tłuszczu w mleku pewnej rasy krów ma rozkład normalny. Na podstawie następujących pomiarów oszacować średnią zawartość tłuszczu w mleku: 3.87, 2.63, 3.23, 3.60, 3.27, 3.69, 3.58, 3.86, 3.85, 3.76, 4.14, 3.63, 3.06, 4.05, 3.05, 3.25, 3.74, 3.27, 3.40, 3.23, 3.61, 3.82, 3.70, 3.61, 2.79, 3.20, 3.52, 3.91, 3.64, 3.04.

5.15. Zbadano zawartość alkoholu w piwie pewnej marki. Producent piwa twierdzi, że przeciętna zawartość alkoholu w piwie tej marki wynosi 5%. Czy wyniki badania dowodzą, że twierdzenie producenta nie jest pozbawione podstaw?

Wyniki badań: 4.20, 4.74, 5.70, 2.32, 5.60, 3.81, 2.65, 3.73, 3.93, 4.63, 4.17, 5.84, 4.31, 6.47, 6.88, 6.16, 5.04, 6.42, 3.98, 4.12, 3.06, 2.98, 4.47, 4.18, 2.76, 1.92, 3.22, 5.26, 5.37, 3.13, 5.19, 3.56, 4.63, 4.64, 6.38, 4.57, 5.36, 3.58, 3.30, 5.21, 6.71, 5.80, 4.50, 4.47, 2.73, 4.94, 4.71, 5.74, 3.66, 4.07.

5.16. Przypuśćmy, że producent gwarantuje uzyskanie średnio 220 jaj od kury i przypuśćmy, że hodowca zakupił partię piskląt, z których uzyskał 25 kur z ukończoną nieśnością. Zbadać, czy w warunkach fermy hodowcy możliwa jest do uzyskania gwarantowana przez producenta nieśność.

Wyniki badań: 209, 218, 233, 179, 231, 203, 184, 201, 205, 216, 209, 235, 211, 245, 252, 240, 222, 245, 205, 208, 191, 189, 213, 209, 186.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

5.17. Według norm, przebieg opon samochodowych powinien wynosić 35 tys. km. W celu sprawdzenia, czy nowy rodzaj opon spełnia wymagania normy, zbadano 200 opon i zanotowano ich przebiegi. Czy można uznać, że nowy rodzaj opon spełnia wymagania normy?

Wyniki badań: 42.356, 25.223, 33.502, 38.682, 34.043, 39.925, 38.407, 42.161, 42.117, 40.828, 46.088, 38.993, 31.181, 44.836, 30.991, 33.820, 40.581, 34.115, 35.833, 33.492, 38.807, 41.672, 40.027, 38.795, 27.398, 33.180, 37.563, 42.866, 39.203, 30.865, 37.338, 36.483, 32.684, 46.339, 36.672, 29.068, 32.667, 31.188, 31.744, 34.477, 41.660, 40.654, 41.026, 29.976, 36.947, 37.414, 40.028, 41.163, 37.912, 45.810, 36.376, 39.464, 44.216, 37.964, 39.890, 37.028, 39.204, 30.443, 32.703, 32.245, 34.431, 34.476, 36.315, 36.865, 39.652, 30.272, 37.313, 49.992, 30.870, 40.470, 28.877, 30.045, 43.174, 32.470, 39.408, 34.407, 36.113, 39.933, 39.411, 33.813, 31.368, 32.909, 48.832, 36.987, 27.440, 33.415, 34.535, 37.793, 35.068, 40.079, 47.691, 28.642, 41.318, 40.053, 35.823, 33.143, 33.201, 45.922, 34.990, 46.070, 35.940, 39.589, 38.132, 38.446, 40.060, 32.786, 32.965, 36.799, 46.570, 34.340, 44.681, 43.024, 42.527, 42.666, 33.112, 38.434, 39.669, 35.501, 30.340, 36.240, 30.168, 34.076, 39.085, 45.995, 33.310, 35.740, 37.907, 24.016, 36.826, 35.504, 36.297, 26.527, 44.716, 40.652, 36.012, 39.737, 36.821, 34.924, 37.971, 34.342, 38.074, 40.994, 43.161, 25.643, 35.506, 47.114, 44.355, 34.369, 43.870, 46.149, 29.454, 34.517, 34.691, 40.366, 37.512, 41.746, 36.908, 27.913, 37.479, 38.466, 36.057, 43.511, 40.631, 42.966, 50.054, 31.798, 44.099, 41.378, 43.410, 39.265, 32.756, 26.785, 33.046, 44.251, 35.555, 42.414, 30.190, 34.203, 32.121, 37.114, 34.192, 36.466, 41.085, 43.112, 45.751, 33.928, 36.834, 29.243, 40.245, 41.613, 36.567, 33.353, 39.013, 34.986, 39.859, 37.064, 37.422, 40.715, 41.142, 41.695.

5.18. W poszukiwaniu tzw. markerów genetycznych bada się zależność między takimi czynnikami jak grupa krwi czy typ białka krwi (*transferyny*) a poziomem cech produkcyjnych. W pewnej populacji krów o średniej wydajności mleka $\mu_0 = 3500$ kg zidentyfikowano 100 krów z transferyną EE i zbadano ich wydajność. Czy można uznać, że krowy z tą transferyną różnią się wydajnością od innych krów?

Wyniki badań: 4122, 2235, 3147, 3718, 3207, 3855, 3687, 4101, 4096, 3954, 4533, 3752, 2892, 4395, 2871, 3182, 3927, 3215, 3404, 3146, 3731, 4047, 3866, 3730, 2475, 3112, 3594, 4178, 3775, 2857, 3570, 3476, 3057, 4561, 3496, 2659, 3055, 2892, 2954, 3255, 4046, 3935, 3976, 2759, 3527, 3578, 3866, 3991, 3633, 4503, 3464, 3804, 4327, 3639, 3851, 3536, 3775, 2810, 3059, 3009, 3250, 3254, 3457, 3518, 3825, 2792, 3567, 4963, 2857, 3915, 2638, 2767, 4212, 3034, 3798, 3247, 3435, 3856, 3798, 3182, 2912, 3082, 4835, 3531, 2480, 3138, 3261, 3620, 3320, 3872, 4710, 2612, 4008, 3869, 3403, 3108, 3114, 4515, 3311, 4531.

5.19. Dzienne zużycie wody w pewnej fabryce podlega wahaniom losowym. Notowano dzienne zużycie wody w metrach sześciennych Czy można przyjąć, że średnie dzienne zużycie wody wynosi 1000 m^3?

Wyniki badań: 1041.5, 999.5, 1019.8, 1032.5, 1021.1, 1035.6, 1031.8, 1041.1, 1040.9, 1037.8, 1050.7, 1033.3, 1014.1, 1047.6, 1013.6, 1020.6, 1037.2, 1021.3, 1025.5, 1019.8, 1032.8, 1039.9, 1035.8, 1032.8, 1004.8, 1019.0, 1029.8, 1042.8, 1033.8, 1013.3, 1029.2, 1027.1, 1017.8, 1051.3, 1027.6, 1008.9, 1017.7, 1014.1, 1015.5, 1022.2, 1039.8, 1037.4, 1038.3, 1011.1, 1028.3, 1029.4, 1035.8, 1038.6, 1030.6, 1050.0, 1026.9, 1034.4, 1046.1, 1030.8, 1035.5, 1028.5, 1033.8, 1012.3, 1017.8, 1016.7, 1022.1, 1022.2, 1026.7, 1028.1, 1034.9, 1011.9, 1029.2, 1060.3, 1013.3, 1036.9, 1008.4, 1011.3, 1043.5, 1017.3, 1034.3, 1022.0, 1026.2, 1035.6, 1034.3, 1020.6, 1014.6, 1018.3, 1057.4, 1028.4, 1004.9, 1019.6, 1022.3, 1030.3, 1023.6, 1035.9, 1054.6, 1007.9, 1039.0, 1035.9, 1025.5, 1018.9, 1019.1, 1050.3, 1023.5, 1050.7, 1025.8.

5.20. Czy można uznać, że przeciętny czas pracy bateryjki radiowej jest zgodny z normą wynoszącą 35 godzin, jeżeli dokonano pomiaru czasu pracy (w godzinach) i

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

otrzymano wyniki: 37.8, 19.3, 28.2, 33.8, 28.8, 35.1, 33.5, 37.6, 37.5, 36.1, 41.8, 34.1, 25.7, 40.4, 25.5, 28.6, 35.9, 28.9, 30.7, 28.2, 33.9, 37.0, 35.3, 33.9, 21.7, 27.9, 32.6, 38.3, 34.4, 25.4, 32.4, 31.4, 27.4, 42.0, 31.6.

5.21. Maszyna jest ustawiona w taki sposób, by produkowała kulki łożyskowe o średnicy 1 cm. Na podstawie poniższych pomiarów średnicy kulek wylosowanych z produkcji kulek stwierdzić, czy można uznać, że maszyna nie rozregulowała się w trakcie pracy?

Wyniki badań: 1.0067, 0.9976, 1.0020, 1.0048, 1.0023, 1.0054, 1.0046, 1.0066, 1.0066, 1.0059, 1.0087, 1.0049, 1.0008, 1.0080, 1.0007, 1.0022, 1.0058, 1.0023, 1.0032, 1.0020, 1.0048, 1.0064, 1.0055, 1.0048, 0.9988, 1.0018, 1.0042, 1.0070, 1.0050, 1.0006, 1.0040, 1.0036, 1.0016, 1.0088, 1.0037, 0.9996, 1.0016, 1.0008, 1.0011, 1.0025, 1.0064, 1.0058, 1.0060, 1.0001, 1.0038, 1.0041, 1.0055, 1.0061, 1.0044, 1.0086, 1.0035, 1.0052, 1.0077, 1.0044, 1.0054, 1.0039, 1.0050, 1.0004, 1.0016, 1.0013, 1.0025, 1.0025, 1.0035, 1.0038, 1.0053, 1.0003, 1.0040, 1.0108, 1.0006, 1.0057.

5.22. W pewnym rejonie Polski uzyskiwano przeciętny plon pszenicy równy 22.6 q/ha. Chcemy sprawdzić, czy po zmianie sposobu uprawy przeciętny plon pszenicy ulegnie zmianie. W tym celu, dla wylosowanych z tego rejonu gospodarstw zbadano, jaki uzyskały one plon przy nowym sposobie uprawy. Czy uzyskane wyniki świadczą o zmianie wielkości plonu przy nowym sposobie uprawy?

Wyniki badań: 22.5, 21.6, 22.1, 22.3, 22.1, 22.4, 22.3, 22.5, 22.5, 22.4, 22.7, 22.4, 21.9, 22.7, 21.9, 22.1, 22.4, 22.1, 22.2, 22.1, 22.3, 22.5, 22.4, 22.3, 21.7.

5.23. Automat produkuje blaszki o nominalnej grubości 0.9 mm. Dla wylosowanych z produkcji automatu blaszek zmierzono ich grubość. Czy można na podstawie tych wyników twierdzić, że grubość blaszek produkowanych przez automat jest równa grubości nominalnej.

wyniki badań: 0.961, 0.840, 0.898, 0.935, 0.902, 0.944, 0.933, 0.960, 0.960, 0.950, 0.988, 0.937, 0.882, 0.979, 0.881, 0.901, 0.949, 0.903, 0.915, 0.898, 0.936, 0.956, 0.945, 0.936, 0.855, 0.896, 0.927, 0.965, 0.939, 0.880.

5.24. Paczka powinna ważyć 1 kg. Kontroli poddano 64 paczki. Sprawdzić, czy badana partia spełnia wymagania odnośnie do ciężaru paczki.

Wyniki badań: 1.10, 0.64, 0.86, 1.00, 0.87, 1.03, 0.99, 1.09, 1.09, 1.06, 1.20, 1.01, 0.80, 1.16, 0.79, 0.87, 1.05, 0.88, 0.92, 0.86, 1.00, 1.08, 1.03, 1.00, 0.70, 0.85, 0.97, 1.11, 1.01, 0.79, 0.96, 0.94, 0.84, 1.20, 0.94, 0.74, 0.84, 0.80, 0.81, 0.89, 1.08, 1.05, 1.06, 0.76, 0.95, 0.96, 1.03, 1.07, 0.98, 1.19, 0.94, 1.02, 1.15, 0.98, 1.03, 0.95, 1.01, 0.78, 0.84, 0.83, 0.88, 0.89, 0.93, 0.95.

5.25. Wyprodukowano pewien nowy środek owadobójczy. Środek ten zastosowano na tysiącu owadach, z których 852 padły. Oszacować skuteczność tego środka owadobójczego.

5.26. Ocenić wadliwość procesu produkcyjnego wiedząc, że na pięćdziesiąt przebadanych wyrobów stwierdzono dwa braki.

5.27. Z pewnej partii owoców pobrano do badania 200 sztuk. Stwierdzono, że 60 jest zepsutych. Ocenić na tej podstawie procent owoców zepsutych w całej partii.

5.28. Na 150 wylosowanych studentów pewnej Akademii Medycznej 114 stwierdziło, że systematycznie pali papierosy. Ocenić ogólny odsetek palaczy wśród studentów Akademii Medycznej.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

5.29. Przeprowadzono ankietę dotyczącą poparcia dla pewnego ruchu społecznego. Wśród tysiąca ankietowanych 850 wyraziło poparcie. Ocenić odsetek ludności popierającej wspomniany ruch społeczny.

5.30. W pewnym zakładzie wśród losowo wybranych dwudziestu osób okazało się, że cztery z nich nigdy nie były na zwolnieniu chorobowym. Oszacować jaki odsetek pracowników tego zakładu nie korzystał ze zwolnienia lekarskiego?

5.31. Na sześćset przypadków wezwań pogotowia czterysta było uzasadnionych. Ocenić jaki procent wezwań pogotowia jest nieuzasadniony.

5.32. Skontrolowano 1000 pojazdów i okazało się, że 80 z nich nie posiadało aktualnego ubezpieczenia „OC". Oszacować na tej podstawie odsetek pojazdów jeżdżących po polskich drogach nie mających wykupionego obowiązkowego ubezpieczenia „OC".

5.33. W celu oceny popularności pewnej partii politycznej przeprowadzono wśród tysiąca osób ankietę. Okazało się, że tylko 240 osób popiera wspomnianą partię. Oszacować na tej podstawie odsetek osób popierających tę partię?

5.34. Lider pewnej partii politycznej powiedział w wywiadzie, że jego partia ma poparcie 25% społeczeństwa. W odpowiedzi przytoczono wyniki ankiety przeprowadzonej wśród tysiąca osób. Spośród ankietowanych tylko 240 osób popierało wspomnianą partię. Czy wyniki ankiety dowodzą, że lider nie jest zorientowany w rzeczywistym poparciu dla swojej partii?

5.35. Wyprodukowano pewien nowy środek owadobójczy. Producent gwarantuje 90% skuteczności. Środek ten zastosowano na tysiącu owadach, z których 852 padły. Czy środek ma taką skuteczność jaką gwarantuje producent?

5.36. Czy można twierdzić, że wadliwość procesu produkcyjnego wynosi 2%, jeżeli na 50 przebadanych wyrobów stwierdzono dwa braki.

5.37. Czy można stwierdzić, że w transporcie psuje się 25% owoców, jeżeli na 200 przebadanych owoców było 60 zepsutych.

5.38. Na 800 zbadanych pacjentów pewnego szpitala 320 miało grupę krwi „O". Zweryfikować hipotezę, że procent pacjentów z tą grupą wynosi 35.

5.39. Na 150 wylosowanych studentów pewnej Akademii Medycznej 114 stwierdziło, że systematycznie pali papierosy. Zweryfikować hipotezę, że palących studentów jest 60%.

5.40. Panuje opinia, że 5% samochodów jeżdżących po polskich drogach nie ma wykupionego obowiązkowego ubezpieczenia „OC". Wśród 1000 skontrolowanych pojazdów 80 nie posiadało „OC". Czy to dowodzi, że wspomniana opinia jest słuszna?

5.41. Wśród losowo wybranych 500 Studentów pewnej uczelni zaobserwowano 475 kobiet. Czy można na tej podstawie sądzić, że zainteresowanie ta uczelnią u obu płci jest jednakowe?

5.42. Obliczono, że emisja programu telewizyjnego jest opłacalna, jeżeli jego oglądalność wynosi 25%. Wśród 1200 ankietowanych osób 200 stwierdziło, że systematycznie ogląda ten program telewizyjny. Czy można na tej podstawie uznać, że emisja programu jest opłacalna?

5.43. W badaniach popularności pewnej partii politycznej stwierdzono, że na 1000 osób, poparcie zadeklarowało 45. Czy można na tej podstawie rokować, że partia ta w wyborach przekroczy pięcioprocentowy próg wyborczy?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

6. Porównanie dwóch populacji

Obserwujemy cechę X w dwóch populacjach. Oznaczmy tę cechę w populacji pierwszej przez X_1, zaś w drugiej przez X_2. Naszym celem będzie porównanie charakterystyk badanej cechy w obserwowanych zbiorowościach. Porównania będą robione na podstawie próby $X_{11}, \ldots, X_{1,n_1}$ z populacji pierwszej oraz $X_{21}, \ldots, X_{2,n_2}$ z populacji drugiej.

Porównanie dwóch wartości średnich. Zakładamy, że obserwowane cechy X_1 oraz X_2 mają w badanych zbiorowościach rozkłady normalne $N(\mu_1, \sigma_1^2)$ oraz $N(\mu_2, \sigma_2^2)$ odpowiednio. Zadanie polega na weryfikacji hipotezy $H_0 : \mu_1 = \mu_2$. Sposób postępowania zależy od tego, czy wariancje σ_1^2 oraz σ_2^2 są równe czy nie. Jeżeli $\sigma_1^2 = \sigma_2^2$, to stosowany jest test t, nazywany również testem Studenta, o statystyce testowej

$$t_{\text{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r},$$

gdzie $S_e^2 = \dfrac{\text{var}X_1 + \text{var}X_2}{n_1 + n_2 - 2}$ oraz $S_r = \sqrt{S_e^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$. Wielkość S_e^2 jest estymatorem wariancji cechy. Jeżeli $|t_{\text{emp}}| > t(\alpha; n_1 + n_2 - 2)$, to hipotezę $H_0 : \mu_1 = \mu_2$ odrzucamy. Liczba $t(\alpha; n_1 + n_2 - 2)$ jest wartością krytyczną rozkładu t z $n_1 + n_2 - 2$ stopniami swobody na poziomie istotności α (tablica 4).

W przypadku, gdy nie wiadomo, czy wariancje σ_1^2 oraz σ_2^2 są sobie równe, to stosowany jest test V Behrensa–Fishera o statystyce testowej

$$V_{\text{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r},$$

przy czym $S_r = \sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}$. Wartości krytyczne $V(\alpha; n_1 - 1, n_2 - 1, c)$ (dla $n_1 \le n_2$) są uzależnione od wielkości $c = \dfrac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}$ (tablica 10). Hipotezę $H_0 : \mu_1 = \mu_2$ odrzucamy, jeżeli $|V_{\text{emp}}| > V(\alpha; n_1 - 1, n_2 - 1, c)$.

W pewnych zagadnieniach praktycznych zainteresowani jesteśmy oceną wielkości różnicy między średnimi, czyli oceną wielkości $\mu_1 - \mu_2$. W tym celu można skonstruować przedział ufności (na poziomie ufności $1 - \alpha$). Przedział taki ma postać

$$\left(\bar{X}_1 - \bar{X}_2 - S_r \cdot (\textit{liczba});\ \bar{X}_1 - \bar{X}_2 + S_r \cdot (\textit{liczba})\right),$$

przy czym

$$\begin{cases} S_r = \sqrt{S_e^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)} \text{ oraz } \textit{liczba} = t(\alpha; n_1 + n_2 - 2), & \text{gdy } \sigma_1^2 = \sigma_2^2, \\ S_r = \sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}} \text{ oraz } \textit{liczba} = V(\alpha; n_1 - 1, n_2 - 1, c), & \text{gdy } \sigma_1^2 \ne \sigma_2^2. \end{cases}$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Jak łatwo można zauważyć, hipoteza o równości średnich nie jest odrzucana tylko wtedy, gdy przedział ufności dla różnicy średnich obejmuje zero.

Porównanie dwóch wariancji. Zakładamy, że obserwowane cechy X_1 oraz X_2 mają w badanych zbiorowościach rozkłady normalne $N(\mu_1, \sigma_1^2)$ oraz $N(\mu_2, \sigma_2^2)$ odpowiednio. Weryfikujemy hipotezę $H_0 : \sigma_1^2 = \sigma_2^2$. Do weryfikacji tej hipotezy stosowany jest test F o statystyce testowej

$$F_{\text{emp}} = \frac{S_1^2}{S_2^2}\ .$$

Hipotezę odrzucamy na poziomie istotności α, jeżeli $F_{\text{emp}} > F\left(\frac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$ lub $F_{\text{emp}} < F\left(1 - \frac{\alpha}{2}; n_1 - 1, n_2 - 1\right)$. Liczby $F(\alpha; u, v)$ są stablicowanymi wartościami krytycznymi rozkładu F (tablica 6). Między wartościami krytycznymi rozkładu F zachodzi następująca zależność:

$$F(\alpha; u, v) = \frac{1}{F(1-\alpha; v, u)}.$$

W celu oceny wielkości ilorazu σ_1^2/σ_2^2 wariancji wyznaczany jest przedział ufności (na poziomie ufności $1 - \alpha$)

$$\left(\frac{S_1^2}{S_2^2} \cdot F\left(1 - \frac{\alpha}{2}; n_1 - 1, n_2 - 1\right);\ \frac{S_1^2}{S_2^2} \cdot F\left(\frac{\alpha}{2}; n_1 - 1, n_2 - 1\right)\right).$$

Można zauważyć, że hipoteza o równości wariancji nie jest odrzucana tylko wtedy, gdy przedział ufności dla ilorazu wariancji obejmuje jedynkę.

Porównanie dwóch frakcji. Zakładamy, że obserwowane cechy X_1 oraz X_2 mają w badanych zbiorowościach rozkłady dwupunktowe $D(p_1)$ oraz $D(p_2)$ odpowiednio. Zadanie polega na weryfikacji hipotezy $H_0 : p_1 = p_2$. W przypadku dużych liczności n_1 oraz n_2 prób hipoteza weryfikowana jest testem przybliżonym o statystyce testowej

$$u_{\text{emp}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

gdzie $\hat{p}_1 = k_1/n_1$, $\hat{p}_2 = k_2/n_2$ oraz $\hat{p} = (k_1+k_2)/(n_1+n_2)$. Wielkości k_1 oraz k_2 są zaobserwowanymi ilościami sukcesów w pierwszej i drugiej próbie odpowiednio. Hipoteza jest odrzucana na poziomie istotności α, jeżeli $|u_{\text{emp}}| > u_{1-\alpha/2}$, gdzie u_α jest kwantylem rozkładu normalnego $N(0, 1)$ (tablica 3). W przypadku, gdy liczności prób są mniejsze niż 25, to stosowany jest tzw. test dokładny (R. Zieliński, W. Zieliński 1990).

Chcąc ocenić różnicę $p_1 - p_2$ między frakcjami można posłużyć się przedziałem ufności (poziom ufności $1 - \alpha$):

$$\left(\hat{p}_1 - \hat{p}_2 - u_{1-\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \hat{p}_1 - \hat{p}_2 + u_{1-\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\ \right).$$

Hipoteza o równości frakcji nie jest odrzucana tylko wtedy, gdy powyższy przedział ufności obejmuje zero.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Przykład. *Badano zawartość tłuszczu w serach żółtych produkowanych zimą i latem. W każdym z dwóch okresów zbadano zawartość tłuszczu w dziesięciu serach i otrzymano wyniki.*
Lato: 30.5; 31.2; 29.9; 31.2; 32.3; 29.1; 28.2; 28.1; 26.4; 26.8.
Zima: 26.8; 26.2; 26.5; 26.1; 25.2; 28.8; 29.1; 29.2; 26.1; 27.2.
Na podstawie uzyskanych wyników stwierdzić, czy zawartość tłuszczu w serze żółtym zależy od pory roku.

Rozwiązanie.

1. *Zbiorowości. W zadaniu mamy do czynienia z dwiema zbiorowościami. Jedną z nich są sery produkowane latem, zaś drugą sery produkowane zimą.*

2. *Obserwowana cecha. Obserwowaną cechą X jest zawartość tłuszczu. Przez X_1 oznaczymy zawartość tłuszczu w serach produkowanych latem, natomiast przez X_2 zawartość tłuszczu w serach produkowanych zimą.*

3. *Założenia. Zakładać będziemy, że cecha X_1 jest cechą ciągłą o rozkładzie normalnym $N(\mu_1, \sigma_1^2)$ oraz, że cecha X_2 jest cechą ciągłą o rozkładzie normalnym $N(\mu_2, \sigma_2^2)$. Liczby μ_1 oraz μ_2 oznaczają przeciętne zawartości tłuszczu w serach produkowanych latem i zimą odpowiednio. Wielkości σ_1^2 oraz σ_2^2 oznaczają odpowiednie wariancje obserwowanej cechy. Dodatkowo zakładać będziemy, że $\sigma_1^2 = \sigma_2^2$.*

4. *Problem. Pytanie, czy średnia zawartość tłuszczu w serach zależy od pory roku sprowadza się do weryfikacji hipotezy $H_0 : \mu_1 = \mu_2$.*

5. *Technika statystyczna. Zgodnie z wcześniejszymi założeniami do weryfikacji sformułowanej hipotezy zastosujemy test t porównania średnich.*

6. *Obliczenia. W naszym zadaniu mamy $n_1 = n_2 = 10$. Proste rachunki dają*

$$\bar{X}_1 = 29.37, \quad \mathrm{var} X_1 = 35.32, \quad \bar{X}_2 = 27.12, \quad \mathrm{var} X_2 = 18.18,$$

$$S_e^2 = \frac{\mathrm{var} X_1 + \mathrm{var} X_2}{n_1 + n_2 - 2} = \frac{35.321 + 18.176}{18} = \frac{53.497}{18} = 2.972,$$

$$S_r = \sqrt{S_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{2.792 \left(\frac{1}{10} + \frac{1}{10} \right)} = 0.771,$$

$$t_{\mathrm{emp}} = \frac{\bar{X}_1 - \bar{X}_2}{S_r} = \frac{29.37 - 27.12}{0.771} = 2.918.$$

Te same obliczenia w Excelu (po wprowadzeniu danych do kolumn A i B).

	C	D	E
1		*Lato:*	*Zima:*
2	*liczność*	`=ILE.LICZB(A:A)`	`=ILE.LICZB(B:B)`
3	*średnia*	`=ŚREDNIA(A:A)`	`=ŚREDNIA(B:B)`
4	*var*	`=(D2-1)*WARIANCJA.PRÓBKI(A:A)`	`=(E2-1)*WARIANCJA.PRÓBKI(B:B)`
5	*df*	`=D2+E2-2`	
6	s_e^2	`=(D4+E4)/D5`	
7	s_r	`=PIERWIASTEK(D6*(1/D2+1/E2))`	
8	t_{emp}	`=(D3-E3)/D7`	
9	$1-\alpha$	`0.05`	
10	t_{kryt}	`=ROZKŁ.T.ODWR.DS(D9;D5)`	

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

7. *Wniosek. Przyjmujemy poziom istotności* $\alpha = 0.05$ *i w tablicach wartości krytycznych rozkładu t znajdujemy odpowiednią wartość dla* $n_1 + n_2 - 2 = 18$ *stopni swobody:* $t(0.05; 18) = 2.1009$. *Ponieważ* $|t_{\text{emp}}| > t(0.05; 18)$, *więc hipotezę o równości średnich odrzucamy.*

8. *Odpowiedź. Średnie zawartości tłuszczu w serach produkowanych latem i zimą nie są takie same.*

9. *Dalsze wnioski. Ponieważ weryfikowana hipoteza została odrzucona, więc można pokusić się o ocenę różnic między średnimi zawartościami tłuszczu. W tym celu można skonstruować przedział ufności dla różnicy* $\mu_1 - \mu_2$ *średnich. Przyjmując poziom ufności* $1-\alpha = 0.95$ *oraz wykorzystując przeprowadzone wcześniej obliczenia otrzymujemy następujący przedział ufności*

$$(29.37 - 27.12 - 2.1009 \cdot 0.771,\ 29.37 - 27.12 + 2.1009 \cdot 0.771) = (0.631,\ 3.869).$$

Ponieważ oba końce tego przedziału są dodatnie, więc możemy stwierdzić, że średnia zawartość tłuszczu w serach produkowanych latem jest wyższa niż średnia zawartość tłuszczu w serach produkowanych zimą. Co więcej, przeciętnie w letnich serach jest tego tłuszczu więcej o co najmniej 0.631, *ale nie więcej niż* 3.869 *jednostek.*

10. *Dodatki. W punkcie trzecim założyliśmy, że wariancje* σ_1^2 *oraz* σ_2^2 *są sobie równe. Można teraz zweryfikować to założenie testując hipotezę* $H_0 : \sigma_1^2 = \sigma_2^2$. *Wartość statystyki testowej jest równa*

$$F_{emp} = \frac{S_1^2}{S_2^2} = \frac{35.32/9}{18.18/9} = \frac{3.92}{2.02} = 1.94.$$

Ze względu na zakres tablic wybieramy poziom istotności α *równy* 0.10. *Wartości krytyczne to*

$$F\left(\tfrac{\alpha}{2}; n_1 - 1, n_2 - 1\right) = F\,(0.05; 9, 9) = 3.179$$

oraz

$$F\left(1 - \tfrac{\alpha}{2}; n_1 - 1, n_2 - 1\right) = F\,(0.95; 9, 9) = \frac{1}{F\,(0.05; 9, 9)} = \frac{1}{3.179} = 0.3146.$$

Ponieważ zachodzi $F\,(0.95; 9, 9) < F_{\text{emp}} < F\,(0.05; 9, 9)$, *więc hipotezy o równości wariancji nie odrzucamy. Zatem zastosowanie testu t do porównania średnich było uzasadnione.*

Przykład. *Pewien importer owoców cytrusowych twierdzi, że owoce zawijane w papierki mniej się psują w transporcie od owoców, które importuje się starą metodą bez zawijania. Jednak wprowadzenie nowej metody wiąże się ze zwiększeniem kosztów. Dlatego importer przeprowadził eksperyment, który miał udowodnić, że owoce zawijane w papierki mniej się psują od nie zawijanych. Pobrał próbę losową* 200 *owoców zawijanych w papierki, z których uległo zepsuciu* 85, *oraz* 150 *owoców nie zawijanych w papierki, w których znaleziono* 60 *owoców zepsutych. Czy badania importera potwierdzają jego twierdzenie?*

Rozwiązanie.

1. *Zbiorowości. W zadaniu mamy do czynienia z dwiema zbiorowościami. Jedną z nich są owoce zawijane w papierki, zaś drugą owoce niezawijane.*

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

2. *Obserwowana cecha. Obserwowaną cechą X jest jakość owocu. Przez X_1 oznaczymy jakość owocu w partii zawijanej w papierki, natomiast przez X_2 jakość owocu w partii nie zawijanej. Pod względem jakości owoce klasyfikowane są jako dobre lub zepsute.*

3. *Założenia. Zakładać będziemy, że cecha X_1 jest cechą skokową o rozkładzie dwupunktowym $D(p_1)$ oraz, że cecha X_2 jest cechą skokową o rozkładzie dwupunktowym $D(p_2)$. Liczby p_1 oraz p_2 oznaczają odpowiednie frakcje zepsutych owoców.*

4. *Problem. Pytanie o porównanie dwóch metod transportowania owoców cytrusowych sprowadza się do weryfikacji hipotezy $H_0 : p_1 = p_2$.*

5. *Technika statystyczna. Zgodnie z wcześniejszymi założeniami oraz ze względu na to, że liczebności prób są duże, do weryfikacji sformułowanej hipotezy zastosujemy test przybliżony porównania frakcji.*

6. *Obliczenia. W naszym zadaniu mamy $n_1 = 200$, $n_2 = 150$, oraz $k_1 = 85$, $k_2 = 60$. Proste rachunki dają*

$$\hat{p}_1 = \frac{k_1}{n_1} = \frac{85}{200} = 0.425, \quad \hat{p}_2 = \frac{k_2}{n_2} = \frac{60}{150} = 0.4,$$

$$\hat{p} = \frac{(k_1 + k_2)}{(n_1 + n_2)} = \frac{(85 + 60)}{(200 + 150)} = 0.412,$$

$$u_{\text{emp}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.425 - 0.4}{\sqrt{0.412(1-0.412)(\frac{1}{200} + \frac{1}{150})}} = 0.47.$$

Obliczenia w Excelu.

	A	B
1	*poziom istotności*	`0.05`
2	*zawinięte w papierki*	`200`
3	*zepsute w papierkach*	`85`
4	*nie zawinięte w papierki*	`150`
5	*zepsute bez papierków*	`60`
6	*odsetek zepsutych w papierkach*	`=B3/B2`
7	*odsetek zepsutych bez papierków*	`=B5/B4`
8	*odsetek zepsutych*	`=(B3+B5)/(B2+B4)`
9	*licznik u_{emp}*	`=B6-B7`
10	*mianownik u_{emp}*	`=B8*(1-B8)*(1/B2+1/B4)`
11	u_{emp}	`=B9/PIERWIASTEK(B10)`
12	*wartośćkrytyczna*	`=ROZKŁ.NORMALNY.S.ODWR(1-B1/2)`

7. *Wniosek. Przyjmujemy poziom istotności $\alpha = 0.05$ i w tablicach wartości krytycznych kwantyli rozkładu normalnego znajdujemy odpowiednią wartość dla $u_{1-0.05/2} = 1.96$. Ponieważ $|u_{\text{emp}}| < u_{0.975}$, więc nie mamy podstaw do odrzucenia hipotezy o równości frakcji.*

8. *Odpowiedź. Można przyjąć, że frakcje zepsutych owoców w obu metodach są takie same. Czyli zawijanie owoców w papierki nie zmienia ich podatności na psucie.*

Zadania do samodzielnego rozwiązania

6.1. Pokazać, że hipoteza o równości średnich nie jest odrzucana tylko wtedy, gdy przedział ufności dla różnicy średnich zawiera zero.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

6.2. Pokazać, że hipoteza o równości wariancji nie jest odrzucana tylko wtedy, gdy przedział ufności dla ilorazu wariancji zawiera jedynkę.

6.3. Pokazać, że hipoteza o równości frakcji nie jest odrzucana tylko wtedy, gdy przedział ufności dla różnicy frakcji zawiera zero.

6.4. Badano ciężar owoców jabłoni dwóch odmian. Dla każdej z odmian zważono po pięćdziesiąt owoców i otrzymano wyniki:
odmiana I: 13.70, 13.12, 14.15, 18.68, 18.03, 11.43, 13.92, 11.94, 11.84, 15.55, 15.22, 9.74, 13.48, 8.48, 11.98, 11.18, 12.41, 16.59, 10.50, 12.37, 13.38, 13.53, 12.93, 9.22, 14.28, 14.97, 10.77, 12.06, 12.37, 11.50, 12.61, 13.72, 12.35, 16.16, 9.89, 15.20, 13.09, 13.57, 13.14, 15.37, 11.38, 13.14, 8.75, 11.88, 11.45, 14.30, 10.14, 10.89, 13.79, 12.67,
odmiana II: 9.42, 11.97, 12.95, 13.20, 10.07, 12.62, 12.93, 15.16, 17.34, 16.37, 11.48, 11.88, 19.19, 14.58, 13.52, 15.18, 11.87, 13.51, 11.47, 14.87, 12.87, 17.62, 13.71, 12.10, 13.57, 16.56, 16.72, 18.33, 16.21, 10.23, 13.86, 13.70, 16.35, 13.55, 15.96, 15.90, 12.09, 9.14, 17.60, 11.28, 15.89, 14.69, 11.98, 13.16, 15.61, 13.01, 14.45, 14.41, 12.42, 14.00.
Na podstawie uzyskanych wyników stwierdzić, czy odmiany różnią się pod względem średniego ciężaru owocu.

6.5. Spośród uczniów pewnego liceum wylosowano piętnastu z klas pierwszych oraz dwunastu z klas drugich i obliczono średnią ocen uzyskanych w semestrze dla każdego z uczniów. Otrzymano rezultaty:
klasy I: 3.94, 3.77, 4.08, 5.00, 5.00, 3.26, 4.00, 3.41, 3.38, 4.50, 4.40, 2.75, 3.87, 2.37, 3.42,
klasy II: 2.54, 3.30, 3.60, 3.67, 2.73, 3.50, 3.59, 4.26, 4.91, 4.62, 3.15, 3.27.
Zbadać, czy osiągnięcia klas pierwszych i klas drugich można uznać za takie same.

6.6. W środę oraz w niedzielę zmierzono przeciętną prędkość tramwajów. Wyniki pomiarów podane są poniżej.
środa: 16.1, 15.4, 16.7, 22.6, 21.8, 13.2, 16.4, 13.8, 13.7, 18.6, 18.1, 11, 15.9, 9.3, 13.9, 12.9, 14.5, 19.9, 12, 14.4, 15.7, 15.9, 15.1, 10.3, 16.9, 17.8, 12.3, 14.0, 14.4, 13.3, 14.7, 16.2, 14.4, 19.3, 11.2, 18.1, 15.3, 16, 15.4, 18.3, 13.1, 15.4, 9.7, 13.8, 13.2, 16.9, 11.5, 12.5, 16.3, 14.8, 17.3, 17.7, 9.1, 14.6, 17.2, 12.1, 13.5, 15.8, 15, 13.4, 14, 18, 15, 13.8, 13.9, 16.7, 10.9, 15.1, 13.2, 14.7, 16.7, 16.7, 15.1, 19, 10.5, 13.6, 18.7, 12.5, 14.5, 15.2, 14.9, 13.9, 11.9, 17.3, 14.7, 14.4, 14.4, 15.2, 15.4, 16.3, 16.7, 13.3, 13.4, 15, 12.5, 19.8, 13.4, 10.1, 17.1, 15.2,
niedziela: 12.1, 14.7, 15.7, 16, 12.7, 15.4, 15.7, 18, 20.3, 19.3, 14.2, 14.6, 22.2, 17.4, 16.3, 18, 14.6, 16.3, 14.2, 17.7, 15.6, 20.6, 16.5, 14.8, 16.4, 19.5, 19.6, 21.3, 19.1, 12.9, 16.7, 16.5, 19.3, 16.3, 18.8, 18.8, 14.8, 11.8, 20.6, 14, 18.8, 17.5, 14.7, 15.9, 18.5, 15.8, 17.3, 17.2, 15.2, 16.8, 14.6, 18.6, 15.5, 14.5, 13.4, 12.6, 17.2, 17.3, 15.1, 18.2, 16.2, 12.8, 15.7, 16.8, 15.2, 15, 13.9, 15.6, 15.3, 15.8, 16.7, 14.4, 15.3, 15.9, 18.4, 14.5, 15.4, 16.5, 16.9, 15.6, 18.4, 14.4, 13, 17.6, 16.9, 17.6, 16.8, 16, 14.2, 17.2.
Na podstawie uzyskanych wyników zbadać, czy tramwaje jeżdżą tak samo szybko w środy jak i w niedziele.

6.7. Dwóm grupom robotników zlecono wykonanie tej samej pracy z tym jednak, że robotnicy grupy pierwszej przeszli wcześniej przeszkolenie. Zaobserwowana wydajność pracy w pierwszej grupie kształtowała się następująco (w szt/h): 18.2, 18.1, 18.3, 19.4, 19.3, 17.6, 18.3, 17.8, 17.7, 18.7, 18.6, 17.2, 18.1, 16.9, 17.8, 17.6, 17.9, 18.9, 17.4, 17.9, 18.1, 18.2, 18, 17.1, 18.3, 18.5, 17.5, 17.8, 17.9, 17.6, podczas gdy w grupie drugiej zaobserwowano następujące wydajności: 16, 16.6, 16.8, 16.9, 16.1, 16.8, 16.8, 17.4, 17.9, 17.7, 16.5, 16.6, 18.4, 17.2, 17.0, 17.4, 16.6, 17.0, 16.5, 17.3, 16.8, 18, 17, 16.6, 17.0. Na poziomie istotności 0.05 sprawdzić, czy przeszkolenie zmieniło wydajność pracy robotnika.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

6.8. W dwóch przedsiębiorstwach A i B pobrano pobrano próby losowe pracowników w celu zbadania ich czasu dojazdu do pracy. Przedsiębiorstwo A było położone w centrum miasta, a przedsiębiorstwo B na jego peryferiach:
Przedsiębiorstwo A: 38.34, 39.90, 40.50, 40.66, 38.74, 40.30, 40.49, 41.86, 43.19, 42.59, 39.60, 39.84, 44.32, 41.50, 40.85, 41.87, 39.84, 40.84, 39.60, 41.68, 40.45, 43.36, 40.97, 39.98, 40.88, 42.71, 42.81, 43.80, 42.50, 38.84, 41.06, 40.96, 42.59, 40.87, 42.34, 42.31, 39.98, 38.17, 43.35, 39.48, 42.30, 41.57, 39.91, 40.63, 42.13, 40.54, 41.42, 41.39, 40.18, 41.15.
Przedsiębiorstwo B: 45.56, 45.23, 45.81, 48.40, 48.03, 44.26, 45.68, 44.55, 44.50, 46.61, 46.43, 43.30, 45.43, 42.58, 44.57, 44.12, 44.82, 47.21, 43.73, 44.80, 45.37, 45.46, 45.12, 43.00, 45.89, 46.28, 43.89, 44.62, 44.80, 44.30, 44.93, 45.57, 44.78, 46.96, 43.38, 46.41, 45.21, 45.48, 45.23, 46.51, 44.24, 45.24, 42.73, 44.52, 44.27, 45.90, 43.53, 43.96, 45.60, 44.97,
Czy można stwierdzić weryfikując odpowiednią hipotezę, że średnie czasów dojazdu do obu przedsiębiorstw są takie same?

6.9. Wysunięto przypuszczenie, że jakość produkcji pewnego wyrobu po wprowadzeniu nowej, tańszej technologii nie uległa zmianie. Wylosowano próbę 120 sztuk tego wyrobu spośród wyprodukowanych starą technologią i otrzymano 12 sztuk złych. Wśród 160 wylosowanych sztuk wyprodukowanych nową technologią było 20 sztuk wadliwych. Czy wysunięte przypuszczenie można w świetle uzyskanych wyników uznać za uzasadnione?

6.10. Na 200 przebadanych szczurów u 60 stwierdzono objawy obniżonego refleksu. Wśród chorych szczurów tylko 20 dostawało pewien preparat P, a wszystkich szczurów karmionych tym preparatem było 80. Czy można uznać, że karmienie preparatem P wpływa na obniżenie refleksu u szczurów?

6.11. Badano, czy młodzież męska nosząca modną fryzurę ma inne wyniki niż pozostali młodzieńcy. W tym celu zbadano 492 uczniów i okazało się, że wśród 94 modnych młodzieńców aż 51 miało złe wyniki w nauce. Wszystkich źle uczących się było 245. Czy na tej podstawie można sądzić, że fryzura ma wpływ na wyniki nauczania?

6.12. Na 180 przebadanych studentów i studentek stwierdzono, że 100 zdało egzamin ze statystyki. Wśród 130 studentek egzaminu nie zaliczyło 55. Czy można na tej podstawie uznać, że wynik egzaminu zależy od płci zdającego?

6.13. Przeprowadzono sondaż wśród 2600 losowo wybranych pasażerów warszawskiego metra. Siedemdziesiąt procent oceniło, że jest zadowolona z komfortu jazdy metrem. W podobnym badaniu przeprowadzonym wśród 3000 pasażerów stołecznych tramwajów 1650 osób dokonało pozytywnej oceny komfortu jazdy. Czy na podstawie przeprowadzonych badań można stwierdzić, że frakcje osób niezadowolonych z komfortu jazdy są takie same?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

7. Jednoczynnikowa analiza wariancji

Obserwujemy cechę X w k populacjach ($k > 2$). Przez X_i oznaczamy cechę X w i–tej populacji. Zakładamy, że X_i ma rozkład normalny $N(\mu_i, \sigma_i^2)$ oraz $\sigma_1^2 = \cdots = \sigma_k^2$. Weryfikujemy hipotezę

$$H_0 : \mu_1 = \cdots = \mu_k$$

na podstawie prób $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, k$. Do weryfikacji stosowany jest **test** F **analizy wariancji**. Proces testowania ujmowany jest w postaci następującej tabeli analizy wariancji.

Źródło zmienności	Stopnie swobody	Sumy kwadratów	Średnie kwadraty	F_{emp}
Czynnik	$k-1$	$\text{var}A$	$S_a^2 = \text{var}A/(k-1)$	S_a^2/S_e^2
Błąd losowy	$N-k$	$\text{var}E$	$S_e^2 = \text{var}E/(N-k)$	
Ogółem	$N-1$	$\text{var}T$		

$$\text{var}A = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \ \text{var}E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2, \ \text{var}T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \ \bar{X} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}, \ N = \sum_{i=1}^{k} n_i \ .$$

Jeżeli $F_{\text{emp}} > F(\alpha; k-1, N-k)$, to hipotezę $H_0 : \mu_1 = \cdots = \mu_k$ odrzucamy na poziomie istotności α (liczba $F(\alpha; k-1, N-k)$ jest wartością krytyczną w rozkładzie F, tablica 6). Odrzucenie powyższej hipotezy zerowej prowadzi do wniosku, że przynajmniej jedna ze średnich $\mu_1, \ldots, \mu_k$ jest inna od pozostałych.

W przypadku odrzucenia hipotezy o równości wszystkich k średnich dokonuje się podziału zbioru $\{\mu_1, \ldots, \mu_k\}$ na grupy jednorodne, czyli największe podzbiory o równych sobie średnich. Postępowania statystyczne zmierzające do podzielenia zbioru średnich na grupy jednorodne nazywamy procedurami porównań wielokrotnych. Poniżej przedstawiona jest jedna z częściej stosowanych w praktyce procedur — jednoczesne przedziały ufności Tukeya.

Załóżmy, że $n_1 = \cdots = n_k = n$. Na podstawie danych wyznaczana jest liczba

$$NIR = t(\alpha; k, N-k)S_e\sqrt{\frac{1}{n}}$$

zwana *najmniejszą istotną różnicą*. Liczba $t(\alpha; k, N-k)$ jest wartością krytyczną studentyzowanego rozstępu (tablica 7). Podziału na grupy jednorodne dokonujemy w ten sposób, że jeżeli $|\bar{X}_i - \bar{X}_j| < NIR$, to uznajemy, że $\mu_i = \mu_j$, czyli populacje o średnich μ_i oraz μ_j zaliczamy do jednej grupy jednorodnej.

W przypadku nierównolicznych prób jedną z modyfikacji procedury Tukeya jest

$$NIR_{ij} = t(\alpha; k, N-k)S_e\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Przykład. *W celu podniesienia wydajności pracy stosowane są różne sposoby motywacji pracowników. Zadanie polega na porównaniu wybranych sposobów motywacji. Wylosowano czterdziestu pracowników i podzielono ich losowo na pięć grup, w których zastosowano różne sposoby motywacji (grupa kontrolna działała na dotychczasowych warunkach). Otrzymano następujące rezultaty (przeciętna wydajność dzienna):*

Sposób				Kontrola
1	2	3	4	
4.87	3.87	2.80	4.40	2.53
4.20	3.67	2.93	3.89	2.40
4.19	3.92	3.40	4.60	2.40
4.52	3.21	2.82	4.25	2.51
4.82	3.91	3.45	4.82	2.82
4.37	3.41	2.98	4.21	2.00
3.98	3.52	3.61	4.55	2.09
4.58	3.00	3.25	4.10	2.20

Stwierdzić, czy sposób motywacji wpływa na wydajność pracy. Jeżeli tak, to dokonać szczegółowej analizy porównawczej badanych sposobów.

Rozwiązanie. *Badaną cechą w pięciu zbiorowościach ($k = 5$) jest wydajność pracy. Można przyjąć, że wydajność pracy, liczona jako przeciętna dzienna wydajność X pracownika, jest zmienną losową o rozkładzie normalnym. To znaczy, zakładamy, że wydajność X_i przy i–tym sposobie motywacji ma rozkład normalny $N(\mu_i, \sigma_i^2)$ ($i = 1, \ldots, 5$). Wielkość μ_i oznacza średnią przeciętną dzienną wydajność przy i–tym sposobie motywacji, zaś σ_i^2 może być interpretowane jako zróżnicowanie pracowników pod względem wydajności. Przyjmiemy, że zróżnicowania te są takie same we wszystkich grupach, tzn. $\sigma_1^2 = \cdots = \sigma_5^2$. Porównanie skuteczności sposobów motywacji sformułujemy formalnie w postaci hipotezy*

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

mówiącej, że sposób motywacji nie wpływa na średnią wydajność. Nieodrzucenie tej hipotezy prowadzić będzie do wniosku, iż proponowane sposoby motywacji nie powodują zmiany średniej wydajności. Do weryfikacji hipotezy zastosujemy test F analizy wariancji.

Poniżej przedstawione są kolejne kroki obliczeniowe.

	Sposób motywacji				Kontrola
	1	2	3	4	
n_i	8	8	8	8	8
$\bar{x}_{i\cdot}$	4.44125	3.56375	3.15500	4.35250	2.36875
$\bar{x}$			3.57625		
$n_i(\bar{x}_{i\cdot} - \bar{x})^2$	5.98584	0.00128	1.41960	4.82048	11.66448
$\mathrm{var}A$			23.89168		
$\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	0.69189	0.82039	0.68260	0.62955	0.49369
$\mathrm{var}E$			3.31812		

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tabela analizy wariancji

Źródło zmienności	Stopnie swobody	Sumy kwadratów	Średnie kwadraty	F_{emp}	F_{kryt}
Sposób	4	23.892	5.973	63.00	$F(0.05;4,35)$=2.64
Błąd losowy	35	3.318	0.095		
Ogółem	39	27.217			

Ponieważ otrzymana wartość funkcji testowej F_{emp} jest większa od wartości krytycznej, więc hipotezę o równości średnich badanej cechy w pięciu populacjach odrzucamy. Oznacza to, że badane sposoby motywacji pracowników (w tym także obiekt kontrolny) różnicują średnią wydajność pracy.

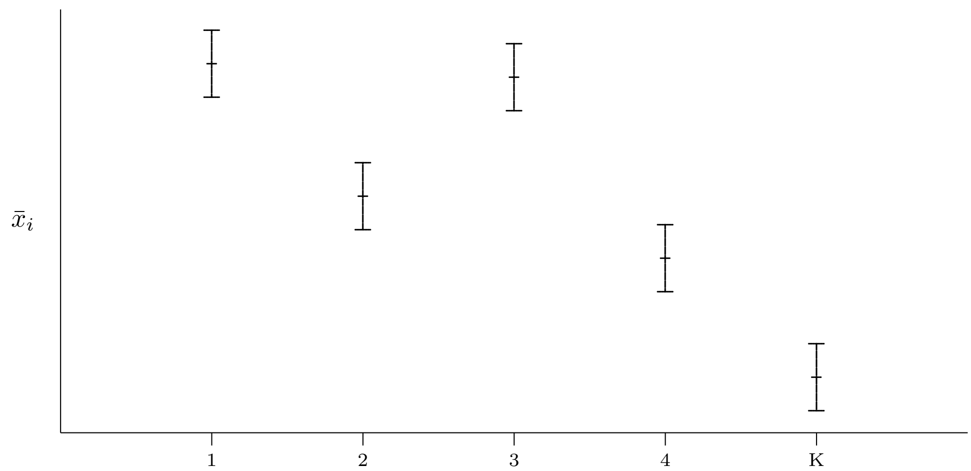
Ponieważ hipoteza o równości średnich została odrzucona stosujemy procedurę Tukeya do wyznaczenia grup jednorodnych. Przyjmijmy $\alpha = 0.05$. W tablicach studentyzowanego rozstępu znajdujemy wartość krytyczną $t(0.05; 5, 35) = 4.066$, a następnie wyznaczamy NIR:

$$NIR = 4.066 \cdot \sqrt{0.095} \cdot \sqrt{\frac{1}{8}} = 0.44308 \ .$$

Porządkujemy średnie próbkowe w ciąg nierosnący i badamy kolejne różnice pod względem ich wielkości w stosunku do NIR. Podkreślone zostały te średnie, których różnica jest mniejsza od NIR.

	Sposób 1	Sposób 3	Sposób 2	Sposób 4	Kontrola
$\bar{x}_{i\cdot}$	4.44125	4.35250	3.56375	3.15500	2.36875

Procedura Tukeya wydzieliła trzy grupy jednorodne. Możemy więc wyprowadzić następujące wnioski. Sposoby 1 oraz 3 dają taką samą średnią dzienną wydajność. Podobnie sposoby 2 oraz 4, chociaż inną niż sposoby 1 oraz 3. Tradycyjny sposób motywacji (kontrola) daje istotnie różne wyniki niż zastosowanie jakiegokolwiek innego sposobu (spośród czterech badanych). Wyniki procedury Tukeya można zilustrować graficznie (por. rysunek). Na osi pionowej zaznaczone są kolejne średnie próbkowe otoczone przedziałami o szerokości równej NIR. Jeżeli przedziały mają część wspólną, to odpowiadające im średnie populacyjne traktowane są jako równe sobie.



Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Zadania do samodzielnego rozwiązania

7.1. Chcąc stwierdzić, czy czas świecenia pewnego typu żarówek zależy od wielkości bańki szklanej przeprowadzono pewne pomiary:

Średnica bańki	czas świecenia (w minutach)										
Mniej niż 5 cm	10	15	18	18	12	15	13	15	12	18	17
Między 5 a 10 cm	18	19	20	19	22	25	19	22	25	25	24
Więcej niż 10 cm	38	40	45	44	42	41	43	40	35	30	29

Czy powyższe dane pozwalają stwierdzić, że wielkość bańki szklanej ma wpływ na czas świecenia żarówki?

7.2. W badaniu średniej zawartości żelaza w roślinach łąkowych uzyskano wyniki

Roślina	zawartość żelaza (w ppm na kg s.m.)					
Trawy	174.0	172.0	175.0	173.0	177.0	175.5
Turzycowate	134.0	135.0	137.0	138.0	135.0	135.2
Motylkowe	117.0	118.0	116.0	119.0	116.3	116.5
Zioła (chwasty)	111.0	112.0	113.0	111.5	112.7	114.2

Porównać średnią zawartość żelaza w roślinach łąkowych.

7.3. Przeciętna liczba błędów popełnionych przez tresowane szczury w toku przejścia przez labirynt ma rozkład normalny. Do pewnych doświadczeń wylosowano po pięć szczurów do czterech grup, które powinny być jednorodne pod względem stopnia wytresowania. Otrzymano dla tych szczurów następujące liczby popełnianych błędów:

Grupa	przeciętna liczba błędów				
I	10	8	7	6	11
II	7	10	6	14	5
III	8	13	15	6	3
IV	16	10	8	10	4

Czy można na podstawie uzyskanych wyników sądzić, że badane grupy są jednakowo wytresowane?

7.4. Trzech nauczycieli A, B, C statystyki oceniało w skali $\langle 0, 1\rangle$ prace jedenastu wylosowanych uczniów. Wyniki były następujące:

	ocena punktowa										
A	0.744	0.667	0.538	0.641	0.718	0.590	0.590	0.718	0.718	0.641	0.615
B	0.744	0.769	0.641	0.846	0.692	0.821	0.744	0.718	0.692	0.821	0.821
C	0.846	0.846	0.923	0.923	0.949	0.795	0.744	0.821	0.923	0.872	0.795

Czy można uznać, że wszyscy trzej nauczyciele są jednakowi w swoich ocenach?

7.5. Właściciel palarni kawy twierdzi, że wszystkie gatunki kawy które produkuje mają podobną zawartość kofeiny. W celu udowodnienia tej hipotezy wybrano trzy mieszanki kawy i po poddaniu ich procesowi palenia, uzyskano następujące zawartości kofeiny (w mg w jednej filiżance kawy) :

gatunek	zawartość kofeiny					
gatunek 1	122.24	122.49	123.22	123.09	120.43	121.26
gatunek 2	117.47	123.56	112.53	114.54	118.71	124.17
gatunek 3	129.96	114.32	118.30	115.65	127.74	128.02

W oparciu o powyższe dane odpowiedzieć na pytanie, czy właściciel palarni ma rację?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

7.6. Wysunięto przypuszczenie, że czas przeznaczony na kolokwium ma wpływ na jego wynik. Chcąc stwierdzić czy tak jest pobrano następującą próbę:

czas w minutach	wyniki kolokwium w punktach				
Od 15 do 20	11	12	13	12	10
Od 20 do 30	15	18	21	23	24
Od 30 do 60	30	25	32	28	20

Czy powyższe dane pozwalają odpowiedzieć na pytanie, czy czas przeznaczony na kolokwium ma wpływ na jego wyniki?

7.7. W doświadczeniu fizycznym przeprowadzonym trzema metodami A, B, C badano czas pewnego efektu świetlnego występującego w tym doświadczeniu. Wyniki eksperymentu:

metoda	średni czas (z pięciu powtórzeń)
A	3.22
B	3.24
C	3.14

Wiedząc również, że $s_e^2 = 0.004$ stwierdzić, czy można uznać, że średni czas występowania tego efektu jest dla wszystkich metod taki sam?

7.8. Mierzono czas świecenia (w godzinach) trzech typów żarówek.

typ żarówki	średni czas świecenia (z pięciu żarówek)
I	1864.8
II	1776.0
III	1827.8

Na podstawie poniższych danych stwierdzić, czy można uznać, że średni czas świecenia tych trzech typów żarówek jest taki sam ($s_e^2 = 8405.5$)?

7.9. Porównywano działanie trzech leków podawanych świniom chorym na różycę (skuteczność działania danego leku mierzono czasem trwania kuracji). Każde lekarstwo zostało zaaplikowane pięciu chorym zwierzętom. Uzyskano następujące wyniki:

lekarstwo	średni czas trwania kuracji
L_1	16
L_2	20
L_3	21

Obliczono również $s_e^2 = 5$. Czy na podstawie powyższych danych można przyjąć, że te trzy lekarstwa dają jednakowy efekt?

7.10. W pewnej miejscowości położonej blisko trasy szybkiego ruchu kierownik mleczarni stwierdził, że rolnicy pasą krowy w przydrożnych rowach. Jak wiadomo, zawartość metali ciężkich jest większa w roślinach rosnących przy drodze. Zbadano po dziesięć próbek mleka od dostawców A, B i C. Otrzymano następujące średnie zawartości metali ciężkich: $A = 4.41$, $B = 3.56$, $C = 4.35$. Ponadto obliczono $\text{var}E = 2.13$. Czy można na tej podstawie stwierdzić, który z dostawców pasie krowy przy szosie?

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

8. Testy zgodności

Obserwujemy cechę X. Interesuje nas pytanie o rozkład tej cechy w zbiorowości. W tym celu weryfikowana jest hipoteza

$$H_0: \text{ Cecha } X \text{ ma rozkład } G,$$

gdzie G jest pewnym teoretycznym rozkładem. Uniwersalnym testem do weryfikacji powyższej hipotezy jest test chi–kwadrat zgodności. Test ten można stosować bez względu na charakter obserwowanej cechy X. Jeżeli X jest cechą ciągłą, to można stosować test Kołmogorowa, zaś w przypadku badania zgodności z rozkładem normalnym (tzn. G jest rozkładem normalnym) najczęściej stosowany jest test Shapiro–Wilka. Poniżej omówiony jest tylko test chi–kwadrat zgodności. Inne testy zgodności można znaleźć w bogatej literaturze tematu.

Test chi–kwadrat zgodności. Hipotetyczny rozkład G może być dowolnym rozkładem prawdopodobieństwa. Omawiany test wymaga dużej ilości danych zapisanych w postaci szeregu rozdzielczego

Klasa	1	2	$\cdots$	k
Liczebność	n_1	n_2	$\cdots$	n_k

Statystyką testową jest

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k} \frac{(n_i - n_i^t)^2}{n_i^t},$$

gdzie

$$n_i^t = Np_i^t,\ N = \sum_{i=1}^{k} n_i,\ p_i^t = P_F\{X \text{ przyjęła wartość z klasy } i\}.$$

Liczby p_i^t noszą nazwę prawdopodobieństw teoretycznych, zaś n_i^t — liczebności teoretycznych. Są to wielkości opisujące czego można spodziewać się w poszczególnych przedziałach klasowych, jeżeli weryfikowana hipoteza jest prawdziwa.

Wartość statystyki testowej porównywana jest z wartością krytyczną $\chi^2(\alpha; k-u-1)$, gdzie u jest liczbą nieznanych parametrów hipotetycznego rozkładu G. Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; k-u-1)$, to hipotezę H_0 odrzucamy na poziomie istotności α.

Przykład. *Pracodawca podejrzewa, że liczba pracowników nieobecnych w różne dni tygodnia nie jest taka sama. W celu sprawdzenia swojego podejrzenia zanotował w pewnym okresie czasu liczby pracowników nieobecnych w kolejnych dniach tygodnia. Wyniki obserwacji podane są w poniższej tabeli*

Dzień tygodnia	Liczba nieobecnych
Poniedziałek	200
Wtorek	160
Środa	140
Czwartek	140
Piątek	100

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Rozwiązanie.

1. *Badana zbiorowość. Pracownicy firmy.*

2. *Badana cecha. Cechą obserwowaną X jest dzień, w którym pracownik nie był w pracy. Jest to cecha jakościowa o wartościach (poniedziałek, wtorek, środa, czwartek, piątek).*

3. *Cel. Przypuszczenie pracodawcy można sformułować w następujący sposób: prawdopodobieństwo nieobecności pracownika jest uzależnione od dnia tygodnia. My zweryfikujemy przeciwną hipotezę, a mianowicie o braku preferencji w opuszczaniu dni. Ta hipoteza może być przedstawiona w postaci jednakowych szans opuszczenia każdego z dni tygodnia przez pracownika.*

4. *Hipoteza statystyczna. Przypuszczenie pracodawcy możemy zapisać w postaci hipotezy*

$$H_0 : X \text{ ma rozkład } \frac{\begin{array}{ccccc}\text{Pon} & \text{Wtk} & \text{Śro} & \text{Czw} & \text{Ptk}\end{array}}{\begin{array}{ccccc}1/5 & 1/5 & 1/5 & 1/5 & 1/5\end{array}}$$

5. *Procedura statystyczna. Ze względu na jakościowy charakter obserwowanej cechy zastosujemy test chi–kwadrat zgodności.*

6. *Poziom istotności. Przyjmujemy $\alpha = 0.05$.*

7. *Obliczenia. Wszystkie niezbędne obliczenia prowadzące do wyznaczenia wartości statystyki testowej zostały przedstawione w poniższej tabeli*

	n_i	p_i^t	n_i^t	$n_i - n_i^t$	$\frac{(n_i - n_i^t)^2}{n_i^t}$
pon	200	0.2	148	52	18.27027
wtk	160	0.2	148	12	.97297
śro	140	0.2	148	−8	.43243
czw	140	0.2	148	−8	.43243
ptk	100	0.2	148	−48	15.56757
Suma	740	1.0	740	0	35.67567

Wartość statystyki testu chi–kwadrat zgodności wynosi 35.67567.

8. *Wartość krytyczna. Chcąc obliczyć stopnie swobody zauważmy, że mamy $k = 5$ przedziałów klasowych oraz, że teoretyczny rozkład jest całkowicie określony w hipotezie zerowej, czyli liczba u nieznanych parametrów rozkładu hipotetycznego wynosi zero. Zatem liczba stopni swobody jest równa $k - u - 1 = 5 - 0 - 1 = 4$. Wartość krytyczna na poziomie istotności $\alpha = 0.05$ wynosi $\chi^2\,(0.05; 4) = 9.4877$.*

9. *Wniosek. Ponieważ $\chi^2_{\text{emp}} > \chi^2\,(0.05; 4)$, więc hipotezę o zgodności z określonym rozkładem odrzucamy.*

10. *Odpowiedź. Przypuszczenie pracodawcy o nierównomiernym rozkładzie nieobecności w zakładzie pracy można uznać za uzasadnione.*

Zadania do samodzielnego rozwiązania

8.1. W celu sprawdzenia, czy kostka sześcienna do gry jest rzetelna wykonano 120 rzutów i otrzymano następujące wyniki:

Liczba oczek	1	2	3	4	5	6
Liczba rzutów	11	30	14	10	33	22

Czy badaną kostkę można uznać za rzetelną.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

8.2. Ścianki kostki do gry pomalowane są na dwa kolory: czarny i biały. Czy prawdziwe jest stwierdzenie, że białym kolorem pomalowana jest dokładnie jedna ścianka, jeżeli na 30 rzutów kolor biały pojawił się dokładnie 4 razy.

8.3. Badano stopień nieopanowania materiału ze statystyki przez studentów pewnego wydziału zliczając ilość ocen niedostatecznych na zaliczeniu przedmiotu w każdej z siedmiu grup.

Grupa	1	2	3	4	5	6	7
Ilość dwój	7	9	14	6	4	2	7

Czy na podstawie powyższych danych można sądzić, że stopień nieopanowania wiedzy jest we wszystkich grupach jednakowy?

8.4. W klasycznych doświadczeniach dotyczących selekcji grochu Mendel obserwował liczności występowania różnych rodzajów nasion otrzymywanych przy krzyżowaniu roślin z okrągłymi i żółtymi nasionami oraz roślin z pomarszczonymi i zielonymi nasionami. Uzyskał on następujące wyniki obserwacji:

pomarszczone i zielone	32
okrągłe i zielone	108
pomarszczone i żółte	101
okrągłe i żółte	315

Według teoretycznych rozważań prawdopodobieństwa otrzymania wymienionych rodzajów grochu winny być w stosunku 1 : 3 : 3 : 9. Na podstawie powyższych danych zweryfikować teorię.

8.5. Wyznaczono liczbę błędów na stronie przy korekcie 500–stronnicowej książki i otrzymano:

Liczba błędów	0	1	2	3	4	5	6	7	8
Liczba stron	67	139	134	90	44	15	6	4	1

Zbadać, czy można uznać, że liczba błędów na stronie ma rozkład Poissona[*]

8.6. Czy można twierdzić, że posiadane przez nas monety są symetryczne, jeżeli na 100 rzutów obiema monetami naraz 23 razy uzyskaliśmy dwa orły, 27 razy dwie reszki, a w pozostałych przypadkach uzyskaliśmy orła i reszkę.

8.7. Wykonano 200 serii po cztery rzuty monetą i uzyskano następujące liczby orłów:

Liczba orłów	0	1	2	3
Liczba rzutów	20	80	70	30

Czy na podstawie uzyskanych wyników można monetę traktować jako symetryczną?

[*] Rozkład Poissona określony jest wzorem $P\{X = k\} = (\lambda^k/k\ \ )e^{-\lambda}$ dla $k = 0, 1, 2 \ldots$. Liczba λ jest wartością średnią w tym rozkładzie.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

8.8. Właściciel sklepu podejrzewa, że czarne szale kupowane są dwa razy częściej niż brązowe, a te z kolei dwa razy częściej niż białe. Czy można uznać przypuszczenie sprzedawcy za uzasadnione, jeżeli na 350 sprzedanych szali 220 było czarnych, a 90 brązowych.

8.9. Przypuszcza się, że u ludzi zamieszkujących środkową Europę włosy naturalnie ciemne występują sześć razy częściej niż blond, a włosy blond dwa razy częściej niż rude. Czy można to przypuszczenie uznać za uzasadnione, jeżeli wśród 150 losowo wybranych osób stwierdzono 120 osób z włosami ciemnymi, a 15 z włosami blond.

8.10. W poniższej tablicy podano procentową zawartość skrobii w każdym z 80 ziemniaków wylosowanych z partii ziemniaków:

Zawartość skrobii	Liczba ziemniaków
9 – 11	1
11 – 13	2
13 – 15	7
15 – 17	20
17 – 19	30
19 – 21	16
21 – 23	3
23 – 25	1

Zbadać normalność rozkładu zawartości skrobi.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

9. Testy niezależności

Obserwujemy parę cech (X, Y), czyli każdemu obiektowi w zbiorowości przypisujemy dwie wielkości. Interesuje nas współzależność występowania wartości badanych cech. W tym celu weryfikowana jest hipoteza

$$H_0 : \text{cechy } X \text{ oraz } Y \text{ są niezależne.}$$

Do jej weryfikacji skonstruowano wiele różnych testów. Wśród nich najbardziej uniwersalnym testem jest test chi–kwadrat niezależności. Tym testem można badać niezależność cech dowolnych typów (jakościowe, skokowe czy ilościowe). W pewnych szczególnych sytuacjach można stosować inne, lepsze postępowania statystyczne takie jak np. test współczynnika korelacji rangowej Spearmana czy współczynnika korelacji rangowej Kendalla. Jeżeli obserwowane cechy mają rozkłady normalne, to stosowany jest test współczynnika korelacji Pearsona. Ze względu na jego szerokie zastosowania praktyczne współczynnikowi korelacji Pearsona poświęcony jest osobny rozdział *Analiza korelacji*. W bieżącym rozdziale omówiony zostanie tylko uniwersalny test chi–kwadrat niezależności.

Test chi–kwadrat niezależności. W celu zastosowania tego testu próba musi być zapisana w postaci dwywumiarowego szeregu rozdzielczego (tablicy dwudzielczej):

Klasy cechy X	Klasy cechy Y: 1	2	$\dots$	m
1	n_{11}	n_{12}	$\dots$	n_{1m}
2	n_{21}	n_{22}	$\dots$	n_{2m}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	n_{k1}	n_{k2}	$\dots$	n_{km}

Statystyka testowa ma postać

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n^t_{ij})^2}{n^t_{ij}} ,$$

gdzie

$$n^t_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}, \; n_{i\cdot} = \sum_{j=1}^{m} n_{ij}, \; n_{\cdot j} = \sum_{i=1}^{k} n_{ij}, \; N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} .$$

Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; (k-1)(m-1))$, to hipotezę H_0 : *Cechy X oraz Y są niezależne* odrzucamy. Liczby $n_{i\cdot}$, $i = 1, \dots, k$ oraz $n_{\cdot j}$, $j = 1, \dots, m$ noszą nazwę próbkowych liczebności brzegowych cech odpowiednio X oraz Y, zaś wartości $n_{i\cdot}/N$, $i = 1, \dots, k$ oraz $n_{\cdot j}/N$, $j = 1, \dots, m$ próbkowych rozkładów brzegowych.

Przykład. *Przypuszcza się, że wykształcenie i zarobki nie są cechami wzajemnie niezależnymi. W celu weryfikacji tego przypuszczenia zebrano informacje od* 950 *osób*

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

na temat ich wykształcenia i zarobków. Uzyskane dane przedstawione są w poniższej tabeli

	podstawowe	średnie	wyższe	ponad wyższe
≤500	21	41	93	47
500−1000	33	37	35	53
1000−1500	45	75	27	43
1500−2000	30	48	50	55
≥2000	71	47	49	50

Czy sformułowane przypuszczenie można uznać za uzasadnione?

Rozwiązanie.

1. *Badana zbiorowość. Osoby pracujące.*

2. *Badana cecha. Obserwowana jest para cech (X, Y), w której X reprezentuje zarobki, zaś Y wykształcenie. Cecha ciągła X została skategoryzowana, zaś cecha Y jest cechą jakościową.*

3. *Cel. Chcemy sprawdzić, czy przypuszczenie o wzajemnej zależności między badanymi cechami jest słuszne. Wygodniej będzie sformułować tezę przeciwną, czyli że obserwowane cechy są niezależne od siebie.*

4. *Hipoteza statystyczna. Będziemy weryfikować hipotezę*

$$H_0: \text{ cechy } X \text{ oraz } Y \text{ są niezależne.}$$

5. *Procedura statystyczna. Ze względu na jakościowy charakter obserwowanych cech zastosujemy test chi–kwadrat niezależności.*

6. *Poziom istotności. Przyjmujemy $\alpha = 0.05$.*

7. *Obliczenia. Wszystkie niezbędne obliczenia prowadzące do wyznaczenia wartości statystyki testowej zostały przedstawione poniżej. Zbadano łącznie $N = 950$ osób.*

Liczebności brzegowe:

$$n_{1\cdot} = 21 + 41 + 93 + 47 = 202,\ n_{2\cdot} = 158,\ n_{3\cdot} = 190,\ n_{4\cdot} = 183,\ n_{5\cdot} = 217,$$

$$n_{\cdot 1} = 21 + 33 + 45 + 30 + 71 = 200,\ n_{\cdot 2} = 248,\ n_{\cdot 3} = 254,\ n_{\cdot 4} = 248.$$

Liczebności teoretyczne n^t_{ij}

	podstawowe	średnie	wyższe	ponad wyższe
≤500	42.5263	52.7326	54.0084	52.7326
500−1000	33.2632	41.2463	42.2442	41.2463
1000−1500	40.0000	49.6000	50.8000	49.6000
1500−2000	38.5263	47.7726	48.9284	47.7726
≥2000	45.6842	56.6484	58.0189	56.6484

W następnym kroku wyznaczamy składniki statystyki testowej uzyskując następującą tabelę wartości $\frac{(n_{ij}-n^t_{ij})^2}{n^t_{ij}}$

	podstawowe	średnie	wyższe	ponad wyższe
≤500	10.8964	2.6104	28.1501	0.6232
500−1000	0.0021	0.4372	1.2423	3.3494
1000−1500	0.6250	13.0073	11.1504	0.8782
1500−2000	1.8870	0.0011	0.0235	1.0934
≥2000	14.0287	1.6433	1.4020	0.7803

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Wartość statystyki testowej jest sumą wartości podanych w powyższej tabeli:

$$\chi^2_{\text{emp}} = 93.8311.$$

8. *Wartość krytyczna. W tablicach znajdujemy wartość krytyczną zmiennej losowej o rozkładzie chi–kwadrat na poziomie istotności* $\alpha = 0.05$ *oraz dla* $(4-1)(5-1) = 12$ *stopni swobody:* $\chi^2(0.05; 12) = 21.0261$

9. *Wniosek. Ponieważ* $\chi^2_{\text{emp}} > \chi^2(0.05; 12)$, *więc hipotezę o niezależności badanych odrzucamy.*

10. *Odpowiedź. Istnieje zależność między wykształceniem i zarobkami.*

Zadania do samodzielnego rozwiązania

9.1. Właściciel palarni kawy twierdzi, że stopień palenia kawy nie ma wpływu na jej smak, a dokładnie na gorzkość. W celu udowodnienia tej tezy wybrano pewną mieszankę kawy i poddano ją procesowi palenia w różnym stopniu. Uzyskano następujące wyniki:

	smak kawy		
	normalna	gorzka	bardzo gorzka
słabo palona	5	9	4
mocno palona	2	12	8
bardzo mocno palona	1	7	14

W oparciu o powyższe dane odpowiedzieć na pytanie, czy właściciel palarni ma rację?

9.2. Poniższa tabela przedstawia liczbę psów zdrowych i chorych na nosówkę w zależności od tego, czy pies ma rodowód, czy go nie ma. Zbadać, czy istnieje zależność między zdrowotnością psa a posiadaniem przez niego rodowodu.

	Psy z rodowodem	Psy bez rodowodu
Psy zdrowe	300	200
Psy chore	40	20

9.3. Przypuszczano, że sposób zapewniania sobie posiłków w pracy przez pracowników, którym firma nie zapewnia regularnego wyżywienia, zależy od płci. W tym celu wylosowano pewną grupę pracowników i uzyskano następujące wyniki:

Płeć	śniadanie z domu	obiad na mieście	zamówienie do pracy
Mężczyźni	68	36	23
Kobiety	36	50	18

W oparciu o powyższe dane odpowiedzieć na pytanie, czy przypuszczenie można uznać za uzasadnione.

9.4. Pracownicy fabryk pewnego zjednoczenia charakteryzują się różną absencją. Wysunięto przypuszczenie, że absencja zależy do płci. Zweryfikować to przypuszczenie na podstawie poniższych danych.

Liczba dni	Płeć	
nieobecności	Kobiety	Mężczyźni
0–5	300	500
5–20	80	70
20 i więcej	20	30

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

9.5. W badaniach budżetów rodzinnych wylosowano 2000 gospodarstw domowych i zanotowano średni miesięczny dochód na głowę oraz fakt posiadania magnetowidu. Czy można na tej podstawie powiedzieć, że fakt posiadania magnetowidu jest wskaźnikiem zamożności rodziny?

Dochód na głowę	Magnetowid jest	Magnetowid nie ma
poniżej 200	404	231
200– 400	486	300
400– 600	242	137
600– 800	57	44
800–1000	29	28
1000 i więcej	24	18

9.6. Zbadać, czy istnieje zależność między stopniem związania kiełbasy a jej smakowitością.

	słabo związana	związana	dobrze związana
dostateczna	9	5	3
dobra	4	12	6
b. dobra	1	6	14

9.7. W pewnym doświadczeniu chemicznym bada się grubość powłoki niklowej uzyskiwanej dla trzech różnych rodzajów kąpieli galwanicznych. Uzyskano następujące wyniki. Czy na tej podstawie można powiedzieć, że grubość powłoki zależy od rodzaju kąpieli?

Grubość powłoki	Liczba pomiarów w kąpieli A	B	C
4– 8	32	51	68
8–12	123	108	80
12–16	10	26	26
16–20	41	34	28
20–24	18	20	24

9.8. W ankiecie rozesłanej wśród pracowników pewnego konsorcjum pytano, czy chcieliby zmienić obecne miejsce pracy. Uzyskano następujące wyniki. Czy chęć zmiany pracy zależy od aktualnych zarobków?

Zarobek aktualny	Odpowiedź Tak	Odpowiedź Nie
500– 700	46	62
700– 900	94	146
900–1100	249	501
1100–1300	126	326
1300–1500	43	135
1500–1700	26	70

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

10. Regresja liniowa

Obserwujemy zmienną Y oraz zmienne $X_1, \ldots, X_k$. Zakładamy, że między średnią wartością zmiennej Y a wartościami zmiennych $X_1, \ldots, X_k$ istnieje zależność liniowa:

$$E(Y|X_1 = x_1, \ldots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

Zmienną Y nazywamy **zmienną zależną** (objaśnianą), zmienne $X_1, \ldots, X_k$ **zmiennymi niezależnymi** (objaśniającymi), a funkcję

$$f(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

liniową **funkcją regresji**.

Klasyczny model regresji liniowej zapisujemy przy pomocy równań:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + +\beta_k x_{ki} + \varepsilon_i,\ \ i = 1, \ldots, n,$$

gdzie $\{(Y_i, x_{1i}, \ldots, x_{ki}) :\ i = 1, \ldots, n\}$ jest n – elementową próbą oraz ε_i, $i = 1, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym $N(0, \sigma^2)$.

Współczynniki $\beta_0, \beta_1, \ldots, \beta_k$ oraz wariancja σ^2 są nieznanymi parametrami modelu. Estymatory $\hat{\beta}_i$, $i = 1, \ldots, n$, współczynników funkcji regresji wyznaczamy przy pomocy metody najmniejszych kwadratów. Otrzymane tą drogą estymatory minimalizują sumę kwadratów

$$\sum_{i=1}^{n} (Y_i - \hat{f}(x_{1i}, \ldots, x_{ki}))^2,$$

w której $\hat{f}(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$ oznacza oszacowaną funkcję regresji. Za ocenę wariancji σ^2 przyjmujemy

$$S^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (Y_i - \hat{f}(x_{1i}, \ldots, x_{ki}))^2.$$

Podstawową hipotezą, którą weryfikujemy w analizie regresji jest

$$H_0 : \beta_1 = \ldots = \beta_k.$$

Jeżeli hipoteza ta jest prawdziwa, to w równaniu funkcji f nie występują zmienne niezależne. W przeciwnym przypadku występuje co najmniej jedna taka zmienna. Wtedy mówimy, że **regresja jest istotna.**

Po odrzuceniu powyższej hipotezy mogą nas interesować niektóre hipotezy cząstkowe oraz pewne obszary ufności. Wybrane aspekty zasygnalizowanej problematyki zostaną przedstawione na przykładzie regresji jednokrotnej (prostej) oraz dwukrotnej.

Regresja prosta. W modelu regresji prostej średnia wartość zmiennej Y zależy od jednej zmiennej X:

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Współczynniki funkcji regresji $f(x) = \beta_0 + \beta_1 x$ estymujemy na podstawie n–elementowej próby $\{(Y_i, x_i) : \; i = 1, \ldots, n\}$:

Estymacja punktowa	**Estymacja przedziałowa**
$\hat{\beta}_1 = \frac{\mathrm{cov}(Y,X)}{\mathrm{var}X}$	$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2)S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2)S_{\beta_1})$
$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$	$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2)S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2)S_{\beta_0})$

gdzie

$$\mathrm{cov}(Y, X) = \sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x}),$$

$$S^2_{\beta_1} = \frac{S^2}{\mathrm{var}X}, \; S^2_{\beta_0} = \frac{S^2}{\mathrm{var}X}\left(\frac{\mathrm{var}X}{n} + \bar{X}^2\right), \; S^2 = \frac{\mathrm{var}Y - \hat{\beta}_1 \mathrm{cov}(X,Y)}{n-2}.$$

Wielkość $S^2_{\beta_1}$ jest wariancją estymatora $\hat{\beta}_1$, $S^2_{\beta_0}$ jest wariancją estymatora $\hat{\beta}_0$, natomiast S^2 jest estymatorem wariancji σ^2. Liczba $t(\alpha; n-2)$ jest wartością krytyczną rozkładu t o $n-2$ stopniach swobody (tablica 4).

Istotność regresji sprawdzamy poprzez weryfikację hipotezy

$$H_0 : \beta_1 = 0.$$

Do jej weryfikacji stosujemy test F o statystyce testowej

$$F_{\mathrm{emp}} = \frac{\hat{\beta}_1^2}{S^2_{\beta_1}}.$$

Hipotezę odrzucamy, jeżeli $F_{\mathrm{emp}} > F(\alpha; 1, n-2)$, gdzie $F(\alpha; 1, n-2)$ jest wartością krytyczną rozkładu F (tablica 6).

Obszar ufności dla prostej regresji umożliwia nam wnioskowanie o wartościach zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X. Dokładniej mówiąc jest to obszar, dla którego zachodzi następująca równość:

$$P\{E(Y|X = x) \in (\hat{f}(x) - t(\alpha; n-2)S_Y; \hat{f}(x) + t(\alpha; n-2)S_Y), \; \forall x\} = 1 - \alpha,$$

gdzie $S^2_Y = S^2\left(\frac{1}{n} + \frac{(x-\bar{X})^2}{\mathrm{var}X}\right)$.

Obszar predykcji umożliwia nam wnioskowanie o wartościach zmiennej Y jednocześnie dla wielu wybranych wartości zmiennej X. Dokładniej mówiąc jest to obszar, dla którego zachodzi następująca równość:

$$P\{Y(x) \in (\hat{f}(x) - t(\alpha; n-2)S_{Y(x)}; \hat{f}(x) + t(\alpha; n-2)S_{Y(x)}), \; \forall x\} = 1 - \alpha,$$

gdzie $Y(x)$ oznacza wartość zmiennej Y dla wybranej wartości x zmiennej X oraz $S^2_{Y(x)} = S^2\left(1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\mathrm{var}X}\right)$.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Przykład. *W pewnym dużym przedsiębiorstwie postanowiono zbadać, czy wydajności pracy (Y) jest uzależniona od stażu pracy (X). W tym celu dla każdej ustalonej wielkości stażu pracy wylosowano po czterech pracowników. W sumie wylosowano 20 osób. Uzyskano następujące wyniki:*

$$\bar{x} = 10.2, \quad \bar{Y} = 22.1, \quad \mathrm{var}X = 2919.2, \quad \mathrm{var}Y = 15231.8, \quad \mathrm{cov}(X,Y) = 5491.6.$$

Rozwiązanie. *Dla każdego z pracowników notowane są dwie wielkości: wydajność pracy (cecha o charakterze losowym) oraz staż pracy (zmienna deterministyczna). Przyjmujemy, że związek ten można opisać za pomocą liniowej funkcji regresji*

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

Naszym zadaniem jest sprawdzenie, czy zaproponowana funkcja może opisywać wspomnianą zależność, a jeżeli tak to podać jej dokładną postać oraz wyprowadzić stosowne wnioski. Zakładamy, że są spełnione założenia modelu regresji liniowej.

Interesuje nas, czy przeciętna wydajność pracy zależy od wielkości stażu. W tym celu formułujemy hipotezę o braku tej zależności:

$$H_0 : \beta_1 = 0.$$

Przyjmujemy poziom istotności $\alpha = 0.05$. Wykonujemy niezbędne obliczenia:

$$\hat{\beta}_1 = \frac{549.16}{2919.2} = 1.88,$$

$$S^2 = \frac{15231.8 - 1.88 \cdot 5491.6}{20 - 2} = 272.644, \ S^2_{\beta_1} = \frac{272.644}{2919.2} = 0.093636,$$

$$F_{\mathrm{emp}} = \frac{3.5344}{0.093636} = 6.14.$$

Ponieważ $|F_{\mathrm{emp}}| > F\,(0.05; 1, 20 - 2)$, więc hipotezę odrzucamy i możemy uznać istnienie zależności przeciętnej wydajności pracy od wielkości stażu pracy. Chcemy znaleźć postać tej zależności. W tym celu liczymy

$$\hat{\beta}_0 = 22.1 - 1.88 \cdot 10.2 = 2.924.$$

Interesująca zależność przyjmuje postać:

$$\hat{f}(x) = 2.924 + 1.88x.$$

Skoro mamy postać tej zależności, możemy być zainteresowani przeciętną wydajnością pracy pracowników o stażu pracy $X = 11.2$. W tym celu wyznaczymy obszar ufności dla prostej regresji. Przyjmujemy poziom ufności $1 - \alpha = 0.95$. Potrzebne rachunki przedstawiają się następująco:

$$\hat{f}(11.2) = 2.924 + 1.88 \cdot 11.2 = 23.98, \quad S^2_Y = 272.644 \left(\frac{1}{20} + \frac{(11.2 - 10.2)^2}{2919.2} \right) = 13.73,$$

$$S_Y \cdot t(0.05, 18) = \sqrt{13.73} \cdot 2.1009 = 7.78.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Zatem możemy wyciągnąć wniosek, że przeciętna wydajność pracowników o stażu 11.2 *wynosi co najmniej* $23.98 - 7.78 = 16.2$*, ale nie więcej niż* $23.98 + 7.78 = 31.76$*. Poziom zaufania do wniosku wynosi* 95%*.*

W przypadku gdybyśmy byli zainteresowani tym jaka jest rozpiętość wydajności takich pracowników, należałoby wyznaczyć obszar predykcji.

Regresja dwukrotna. W modelu regresji dwukrotnej średnia wartość zmiennej Y zależy od dwóch zmiennych:

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Współczynniki funkcji regresji $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ estymujemy metodą najmniejszych kwadratów na podstawie n – elementowej próby $\{(Y_i, x_{1i}, x_{2i}) : \ i = 1, \ldots, n\}$.

Estymatory punktowe współczynników funkcji regresji wyznaczamy z układu

$$\begin{cases} \mathrm{var}(X_1)\hat{\beta}_1 + \mathrm{cov}(X_1, X_2)\hat{\beta}_2 = \mathrm{cov}(X_1, Y), \\ \mathrm{cov}(X_1, X_2)\hat{\beta}_1 + \mathrm{var}(X_2)\hat{\beta}_2 = \mathrm{cov}(X_2, Y), \\ \qquad \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 = \hat{\beta}_0. \end{cases}$$

Przedziały ufności dla współczynników funkcji regresji mają następującą postać:

$$\begin{aligned} \beta_0 &\in (\hat{\beta}_0 - t(\alpha; n-3)S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-3)S_{\beta_0}), \\ \beta_1 &\in (\hat{\beta}_1 - t(\alpha; n-3)S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-3)S_{\beta_1}), \\ \beta_2 &\in (\hat{\beta}_2 - t(\alpha; n-3)S_{\beta_2}, \hat{\beta}_2 + t(\alpha; n-3)S_{\beta_2}). \end{aligned}$$

gdzie

$$S^2 = \frac{\mathrm{var}Y - \hat{\beta}_1 \mathrm{cov}(X_1, Y) - \hat{\beta}_1 \mathrm{cov}(X_1, Y)}{n-3}, \ R^2 = \frac{\mathrm{cov}(X_1, X_2)^2}{\mathrm{var}X_1 \mathrm{var}X_2},$$

$$S^2_{\beta_0} = \frac{1}{n} + \frac{1}{1-R^2}\left(\frac{\bar{X}_1^2}{\mathrm{var}X_1} + \frac{\bar{X}_2^2}{\mathrm{var}X_2} - \frac{R^2}{\mathrm{cov}(X_1, X_2)}\right),$$

$$S^2_{\beta_1} = \frac{S^2}{(1-R^2)\mathrm{var}X_1}, \ S^2_{\beta_2} = \frac{S^2}{(1-R^2)\mathrm{var}X_2}$$

oraz $t(\alpha; n-3)$ jest wartością krytyczną rozkładu t o $n-3$ stopniach swobody. Wielkość $S^2_{\beta_0}$ jest wariancją estymatora $\hat{\beta}_0$, $S^2_{\beta_1}$ jest wariancją estymatora $\hat{\beta}_1$, $S^2_{\beta_2}$ jest wariancją estymatora $\hat{\beta}_2$, S^2 jest estymatorem punktowym wariancji σ^2.

Istotność regresji sprawdzamy poprzez weryfikację hipotezy

$$H_0 : \beta_1 = \beta_2 = 0.$$

Do jej weryfikacji stosujemy test F o statystyce testowej

$$F_{\mathrm{emp}} = \frac{S_R^2}{S^2},$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

gdzie $S_R^2 = (\hat{\beta}_1 \mathrm{cov}(X_1, Y) + \hat{\beta}_2 \mathrm{cov}(X_2, Y))/2$. Jeżeli $F_{\mathrm{emp}} > F\,(\alpha; 2, n-3)$, to hipotezę odrzucamy.

W przypadku odrzucenia powyższej hipotezy testowane są **hipotezy cząstkowe**:

$$H_0 : \beta_1 = 0, \qquad H_0 : \beta_2 = 0.$$

Do weryfikacji powyższych hipotez stosujemy test t o statystyce testowej $t_{\mathrm{emp}} = \hat{\beta}_i / S_{\beta_i}$. Hipotezę odrzucamy, jeżeli $|t_{\mathrm{emp}}| > t(\alpha; n-3)$.

Przykład. *Celem badania jest ilościowy opis zależności kosztu Y wyprodukowania jednego detalu na pewnych automatach od „wieku” automatu X_1 oraz wielkości dziennej produkcji X_2 tego automatu. W tym celu ustalono kilka poziomów wieku automatu oraz wielkości produkcji. Stosownie do tych ustaleń wybrano automaty oraz przeprowadzono odpowiedni eksperyment. Rozmiar zebranej próby wyniósł 50 obserwacji. Na podstawie uzyskanych wyników otrzymano:*

$\bar{x}_1 = 6$	$\bar{x}_2 = 300$	$\bar{y} = 1531$
$\mathrm{var} x_1 = 30$	$\mathrm{var} x_2 = 30$	$\mathrm{var} y = 400$
$\mathrm{cov}(x_1, x_2) = 20$	$\mathrm{cov}(x_1, y) = 40$	$\mathrm{cov}(x_2, y) = 350$

Rozwiązanie. *W zadaniu pojawia się jedna cecha o charakterze losowym (koszt) i dwie zmienne deterministyczne (wiek oraz dzienna produkcja). Przyjmujemy, że zależność ta opisana jest funkcją regresji $f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Zakładamy, że są spełnione wszystkie założenia modelu regresji liniowej. W pierwszym kroku chcemy sprawdzić, czy zależność ta jest istotna. W dalszych etapach wyznaczamy explicite postać funkcji oraz wyprowadzamy stosowne wnioski. Oczywiście te dalsze etapy wykonujemy wtedy, gdy stwierdzimy istotność zależności kosztu produkcji od wieku automatu i wielkości jego dziennej produkcji.*

W celu zbadania istnienia zależności średniej wartości cechy Y od zmiennych X_1 oraz X_2 formułujemy hipotezę

$$H_0 : \beta_1 = \beta_2 = 0.$$

Przyjmujemy poziom istotności $\alpha = 0.05$. Potrzebne rachunki przedstawiają się następująco. Tworzymy układ równań:

$$\begin{cases} 30\hat{\beta}_1 + 20\hat{\beta}_2 = 40, \\ 20\hat{\beta}_1 + 660\hat{\beta}_2 = 350 \end{cases}$$

i rozwiązujemy go ze względu na $\hat{\beta}_1, \hat{\beta}_2$:

$$\hat{\beta}_1 = 1, \quad \hat{\beta}_2 = 0.5.$$

Wyznaczamy pozostałe wartości:

$$S_R^2 = \frac{1 \cdot 40 + 0.5 \cdot 350}{2} = 107.5, \quad S^2 = \frac{400 - 1 \cdot 40 - 0.5 \cdot 350}{50-3} = 8.51,$$

$$F_{\mathrm{emp}} = \frac{S_R^2}{S^2} = 12.63.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Ponieważ $F_{\text{emp}} > F\,(0.05; 2, 50-3)$*, uznajemy istotność badanej zależności. Nie wiemy jednak, czy przeciętny koszt produkcji detalu zależy od obu pozostałych cech, czy tylko od jednej. W tym celu sprawdzamy hipotezy cząstkowe:*

$$H_0 : \beta_1 = 0, \qquad H_0 : \beta_2 = 0.$$

Przyjmujemy poziom istotności $\alpha = 0.05$ *i wykonujemy niezbędne rachunki:*

$$R^2 = \frac{20^2}{30 \cdot 660} = 0.02, \quad S^2_{\beta_1} = \frac{8.51}{(1-0.02)30}, \quad S^2_{\beta_2} = \frac{8.51}{(1-0.02)660},$$

$$t^{(1)}_{\text{emp}} = \frac{\hat{\beta}_1}{S_{\beta_1}} = 1.76, \quad t^{(2)}_{\text{emp}} = \frac{\hat{\beta}_2}{S_{\beta_2}} = 4.13.$$

Ponieważ $|t^{(1)}_{\text{emp}}| < t(0.05; 47)$ *oraz* $|t^{(2)}_{\text{emp}}| > t(0.05; 47)$*, stwierdzamy że przeciętny koszt produkcji detalu zależy jedynie od wielkości dziennej produkcji. W praktyce, w tego typu sytuacji, jeżeli chcemy znaleźć postać tej zależności, przyjmuje się* $\beta_1 = 0$*, a następnie, w zależności od potrzeb, stosuje metody analizy regresji jednokrotnej.*

Zadania do samodzielnego rozwiązania

10.1. Zbadano w dziesięciu wylosowanych zakładach przemysłowych wielkość zaplanowanego i wykonanego funduszu na akcję socjalną.

planowany	3.60	4.65	5.20	1.86	3.06	1.36	2.46	3.93	5.80	6.35
wykonany	3.56	4.59	5.13	1.84	3.02	1.35	2.43	3.89	5.72	6.26

Opisać ilościowo tę zależność przy pomocy liniowej funkcji regresji i zbadać jej istotność. Oszacować wykonanie funduszu przy funduszu zaplanowanym na poziomie 5.

10.2. Zbiorowość pracowników pewnego dużego przedsiębiorstwa postanowiono przebadać ze względu na zależność wydajności pracy (Y) od stażu pracy (X). W tym celu wylosowano 20 pracowników i uzyskano następujące wyniki:

$$\sum x_i = 204, \ \sum y_i = 442, \ \sum x_i^2 = 5000, \ \sum y_i^2 = 25000, \ \sum x_i y_i = 10000.$$

Opisać ilościowo tę zależność przy pomocy liniowej funkcji regresji i zbadać jej istotność. Oszacować przeciętną wydajność pracowników o stażu pracy $x = 10.2$. Zinterpretować uzyskany wynik.

10.3. Poniższe dane z dziesięciu poletek dotyczą efektywności nawożenia łąk azotem (w kg siana na 1 kg azotu) w zależności od poziomu nawożenia azotem:

$$\bar{x} = 40, \ \bar{y} = 16, \ \sum x_i^2 = 20200, \ \sum y_i^2 = 2564.625, \ \sum x_i y_i = 6295.$$

Wyznaczyć liniową funkcję zależności przeciętnej efektywności nawożenia łąk azotem od poziomu nawożenia. Czy zależność jest istotna? Jaki jest oczekiwany plon przy nawożeniu 45.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

10.4. Badano zależność między kosztami materiałowymi (X) a kosztami robocizny bezpośredniej (Y) przy produkcji pewnego artykułu. Na podstawie 20 losowych obserwacji i otrzymano wyniki

$$\bar{x} = 13,\ \mathrm{var}x = 50,\ \bar{y} = 15,\ \mathrm{var}y = 12,\ \mathrm{cov}(x, y) = 25.$$

Wyznaczyć liniową funkcję zależności przeciętnych kosztów robocizny bezpośredniej od kosztów materiałowych. Czy zależność jest istotna? Jaki jest oczekiwany koszt robocizny przy kosztach materiałowych 9.

10.5. W badaniu kosztów jednostkowych Y w zależności od wielkości X produkcji zanotowano

x_i	10	20	30	40	50	60	70	80	90	100
y_i	0.20	0.16	0.40	0.63	1.22	2.19	3.29	5.30	8.94	14.98

Skonstruować na tej podstawie wykładniczą funkcję regresji $f(x) = a \cdot e^{bx}$. Przeprowadzić pełną analizę regresji. Wskazówka: skorzystać z faktu, że jeżeli $y = a \cdot e^{bx}$, to $\ln y = \ln a + bx$.

10.6. Wyniki ogólnopolskiego konkursu maszynopisania dostarczyły informacji o średniej liczbie błędów na stronie popełnionych przez maszynistki zależnie od czasu pisania tego samego tekstu i ich wieku. Podane zestawienie uwzględnia odpowiednie informacje dla stu wylosowanych maszynistek:

wiek	26	30	22	28	32	25	27	28	24	21	24	33	25	20	37
czas	8.3	7.3	9.4	7.8	6.8	8.7	8.0	7.8	8.9	9.5	8.9	6.7	8.7	10.0	5.6
błędy	6.48	3.56	4.94	3.69	5.67	9.18	5.29	2.74	4.72	8.49	5.87	5.66	5.86	6.00	5.53

Zbadać istnienie zależności między liczbą popełnianych błędów a czasem pisania i wiekiem maszynistki. Opisać tę zależność za pomocą liniowej funkcji regresji.

10.7. Przypuszcza się, że miesięczne wydatki Y na zakup artykułów nieżywnościowych uzależnione są od wielkości dochodów X_1 oraz wysokości X_2 wydatków na zakup artykułów żywnościowych. Przeprowadzono odpowiednie badania 50 gospodarstw domowych i otrzymano wyniki:

$$\begin{array}{lll} \bar{x}_1 = 5 & \bar{x}_2 = 2.5 & \bar{y} = 1.5 \\ \sum x_{1i}^2 = 203.00 & \sum x_{2i}^2 = 50.75 & \sum y_i^2 = 18.59 \\ \sum x_{1i}x_{2i} = 101.50 & \sum x_{1i}y_i = 61.30 & \sum x_{2i}y_i = 30.65 \end{array}$$

Przeprowadzić na tej podstawie analizę regresji wydatków na artykuły nieżywnościowe od pozostałych zmiennych.

10.8. Do produkcji pewnego artykułu stosowane są dwa komponenty X_1 oraz X_2. Mierzona za pomocą pewnej cechy ilościowej Y jakość wyrobu uzależniona jest od ilości użytych komponentów. W wyniku zbadania czterdziestu próbek wyrobu produkowanego z różnymi ilościami obu składników otrzymano:

$$\begin{array}{lll} \bar{x}_1 = 27.00 & \bar{x}_2 = 8.00 & \bar{y} = 5.54 \\ \sum x_{1i}^2 = 5222.00 & \sum x_{2i}^2 = 457.00 & \sum y_i^2 = 237.09 \\ \sum x_{1i}x_{2i} = 1511.00 & \sum x_{1i}y_i = 1041.05 & \sum x_{2i}y_i = 315.73 \end{array}$$

Przeprowadzić analizę regresji jakości wyrobu od ilości użytych składników.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

10.9. W produkcji pewnego artykułu analizowano zysk Y ze sprzedaży w zależności od wielkości produkcji.

x_i	10	12	13	16	17	19	21	22	23	26	29	36
y_i	280.69	308.27	321.92	349.63	358.22	357.10	356.95	356.99	348.90	321.60	278.10	113.60

Przyjmując, że funkcja regresji opisująca zysk przedsiębiorstwa w zależności od wielkości produkcji ma postać $y = \beta_0 + \beta_1 x + \beta_2 x^2$ przeprowadzić analizę regresji. Czy istnieje produkcja maksymalizująca zysk?

10.10. Analizowano, jak wynik Studenta z egzaminu (Y) zależy od liczby kartkówek przeprowadzonych na ćwiczeniach (X_1) oraz od procentu opuszczonych przez Studenta ćwiczeń (X_2). Na podstawie 60 – elementowej próby otrzymano:

$$
\begin{array}{lll}
\bar{y} = 54 & \bar{x}_1 = 5 & \bar{x}_2 = 11 \\
\mathrm{var} y = 46602.57 & \mathrm{var} x_1 = 114.18 & \mathrm{var} x_2 = 57773.33 \\
\mathrm{cov}(y, x_1) = 1527.43 & \mathrm{cov}(y, x_2) = -12282.84 & \mathrm{cov}(x_1, x_2) = -55.33
\end{array}
$$

Przeprowadzić analizę regresji wyniku egzaminu od pozostałych zmiennych.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

11. Analiza korelacji

Korelacja jednokrotna. Zakładamy, że obserwowane zmienne X_1 oraz X_2 mają łączny rozkład normalny o macierzy korelacji

$$\mathcal{C} = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}.$$

Konsekwencją tego założenia jest to, że jeśli istnieje zależność między obserwowanymi zmiennymi, to ma ona charakter liniowy. Miarą tej zależności jest **współczynnik korelacji** ϱ. Wartość tego współczynnika należy do przedziału $\langle -1, 1\rangle$. Im $|\varrho|$ bliższe jest wartości jeden, tym wyraźniejsza jest zależność liniowa między zmiennymi.

Celem analizy korelacji jednokrotnej jest weryfikacja hipotezy

$$H_0 : \varrho = 0.$$

Do jej weryfikacji stosuje się **test współczynnika korelacji Pearsona** o statystyce testowej

$$R = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{var}X_1 \cdot \mathrm{var}X_2}}.$$

Hipotezę odrzucamy, jeżeli $|R| > r(\alpha; n)$, gdzie $r(\alpha; n)$ jest wartością krytyczną współczynnika korelacji (tablica 8).

Przykład. *Zadanie polega na zbadaniu istnienia zależności między miesięcznymi wydatkami na produkty tytoniowe (X_1) oraz na produkty alkoholowe (X_2). Na podstawie badania pięćdziesięciu rodzin otrzymano następujące wyniki:*

$$\mathrm{cov}(X_1, X_2) = 380, \ \mathrm{var}X_1 = 500, \ \mathrm{var}X_2 = 650.$$

Rozwiązanie. *Obserwujemy parę cech: (miesięczne wydatki na wyroby tytoniowe, miesięczne wydatki na produkty alkoholowe). Zakładać będziemy, że łączny rozkład tych cech jest normalny. W celu zbadania istnienia zależności między tymi zmiennymi posłużymy się współczynnikiem korelacji Pearsona. Formułujemy hipotezę o braku zależności, która dla cech o łącznym rozkładzie normalnym przyjmuje postać*

$$H_0 : \varrho = 0.$$

Hipotezę tę zweryfikujemy testem opartym na próbkowym współczynniku korelacji. Proste obliczenia pokazują, że wartość tego współczynnika w próbie wynosi

$$R = \frac{380}{\sqrt{500 \cdot 650}} \approx 0.67.$$

Na poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę, ponieważ $|R| > r(0.05; 50)$. Zatem przy założeniu normalności łącznego rozkładu zmiennych, wykazaliśmy, że istnieje między nimi zależność o charakterze liniowym. Możemy więc stwierdzić, że przeciętne miesięczne wydatki na produkty tytoniowe zależą liniowo od miesięcznych wydatków

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

na produkty alkoholowe. W dalszym ciągu posługując się technikami analizy regresji można wyznaczyć liniową funkcję zależności między obserwowanymi zmiennymi.

Korelacja wielokrotna. Zakładamy, że obserwowane zmienne X_1, X_2, X_3 mają łączny rozkład normalny. Wówczas para $(X_1, \lambda_1 X_2 + \lambda_2 X_3)$, gdzie λ_1 oraz λ_2 są dowolnymi liczbami rzeczywistymi, również ma łączny rozkład normalny. Niech $\varrho(X_1, \lambda_1 X_2 + \lambda_2 X_3)$ oznacza współczynnik korelacji między X_1 a $\lambda_1 X_2 + \lambda_2 X_3$. **Współczynnik korelacji wielokrotnej** $\varrho_{1|2,3}$ między X_1 a parą (X_2, X_3) definiujemy jako:

$$\varrho_{1|2,3} = \max_{\lambda_1,\lambda_2} \varrho(X_1, \lambda_1 X_2 + \lambda_2 X_3).$$

Wartość współczynnika korelacji wielokrotnej należy do przedziału $\langle 0, 1\rangle$.

Celem analizy korelacji wielokrotnej (dla trzech zmiennych) jest weryfikacja hipotezy

$$H_0 : \varrho_{1|2,3} = 0.$$

Do jej weryfikacji stosuje się test współczynnika korelacji wielokrotnej o statystyce testowej

$$R_{1|2,3} = \sqrt{1 - \frac{|C|}{|C_{11}|}},$$

gdzie $|C|$ oznacza wyznacznik macierzy korelacji

$$C = \begin{pmatrix} 1 & R_{12} & R_{13} \\ R_{21} & 1 & R_{23} \\ R_{31} & R_{32} & 1 \end{pmatrix}$$

oraz $|C_{11}|$ oznacza wyznacznik macierzy

$$C_{11} = \begin{pmatrix} 1 & R_{23} \\ R_{32} & 1 \end{pmatrix}.$$

Macierz C_{11} jest dopełnieniem algebraicznym elementu macierzy C stojącego w pierwszym wierszu jej pierwszej kolumny. Element R_{ij}, $i, j = 1, 2, 3$, oznacza współczynnik korelacji Pearsona między zmienną X_i oraz zmienną X_j.

Hipotezę powyższą odrzucamy, jeżeli $R_{1|2,3} > r(\alpha; 3, n)$, gdzie $r(\alpha; 3, n)$ jest wartością krytyczną współczynnika korelacji wielokrotnej (tablica 9). Po odrzuceniu hipotezy stwierdzamy, że istnieje zależność liniowa między średnią wartością zmiennej X_1 a wartościami zmiennych X_2 oraz X_3.

W analizie korelacji wielokrotnej możemy analizować także współczynniki $\varrho_{2|1,3}$ oraz $\varrho_{3|1,2}$. Definiuje się je w podobny sposób jak współczynnik $\varrho_{1|2,3}$.

Współczynnik korelacji cząstkowej. Zakładamy, że zmienne X_1, X_2, X_3 mają łączny rozkład normalny o macierzy korelacji

$$\mathcal{C} = \begin{pmatrix} 1 & \varrho_{12} & \varrho_{13} \\ \varrho_{21} & 1 & \varrho_{23} \\ \varrho_{31} & \varrho_{32} & 1 \end{pmatrix}.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Element ϱ_{ij}, $i, j = 1, 2, 3$, macierzy korelacji oznacza współczynnik korelacji między zmienną X_i oraz zmienną X_j.

Współczynnik korelacji cząstkowej $\varrho_{12|3}$ wyrażamy wzorem

$$\varrho_{12|3} = \frac{\varrho_{12} - \varrho_{13}\varrho_{23}}{\sqrt{(1-\varrho_{13}^2)(1-\varrho_{23}^2)}}.$$

Jest on miernikiem indywidualnego wpływu zmiennej X_2 na zmienną X_1 (z pominięciem wpływu zmiennej X_3). Istotność tego wpływu badamy poprzez weryfikację hipotezy

$$H_0 : \varrho_{12|3} = 0.$$

Statystyką testową jest współczynnik

$$R_{12|3} = \frac{R_{12} - R_{13}R_{23}}{\sqrt{(1-R_{13}^2)(1-R_{23}^2)}}.$$

Hipotezę odrzucamy, jeżeli $|R_{12|3}| > r(\alpha, 3, n-1)$, gdzie $r(\alpha, 3, n-1)$ jest wartością krytyczną współczynnika korelacji wielokrotnej.

Pojęcie korelacji wielokrotnej oraz współczynnika korelacji cząstkowej jest uogólnione na dowolną liczbę zmiennych.

Przykład. *Badano trzy charakterystyki mieszkań: cenę rynkową (X_1), powierzchnię całkowitą (X_2) oraz liczbę pokoi (X_3). Celem było ustalenie, czy istnieje zależność między ceną a pozostałymi wielkościami. Z ofert sprzedaży wybrano losowo $n = 100$ ofert. Dla każdej z nich zanotowano wartości interesujących nas cech i otrzymano następujące współczynniki korelacji Pearsona:*

$$R_{12} = 0.9214, \ R_{13} = 0.8346, \ R_{23} = 0.9216.$$

Czy na podstawie powyższych danych można mówić o zależności ceny od powierzchni oraz liczby pokoi?

Rozwiązanie. *Dla każdego wylosowanego mieszkania obserwowany jest wektor cech (X_1, X_2, X_3). W celu uzyskania odpowiedzi na postawione pytanie zastosujemy współczynnik korelacji wielokrotnej, a chcąc przeprowadzić bardziej szczegółowe badanie (oczywiście po stwierdzeniu niezerowości tego współczynnika) posłużymy się współczynikami korelacji cząstkowych.*

Wyznaczone współczynniki korelacji można zapisać w formie macierzy korelacji:

$$C = \begin{pmatrix} 1 & 0.9214 & 0.8346 \\ 0.9214 & 1 & 0.9216 \\ 0.8346 & 0.9216 & 1 \end{pmatrix}.$$

Współczynnik korelacji wielokrotnej zmiennej X_1 względem X_2, X_3 przyjmuje wartość

$$R_{1|2,3} = 0.9222.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Ponieważ jest to wartość stosunkowo bliska 1, więc wnioskujemy, że cena mieszkania prawdopodobnie zależy przynajmniej od jednej z dwóch obserwowanych cech: powierzchni całkowitej oraz liczby pokoi.

Jeżeli założymy, że zmienne X_1, X_2, X_3 mają łączny rozkład normalny, to istnienie zależności pomiędzy zmienną X_1 a zmiennymi (X_2, X_3) możemy badać bardziej formalnie. Sformułujemy hipotezę o braku takiej zależności:

$$H_0: \ \varrho_{1|2,3} = 0.$$

Przyjmując poziom istotności $\alpha = 0.05$ odrzucamy hipotezę, ponieważ

$$R_{1|2,3} > r(0.05, 3, 100).$$

Zatem interesująca nas zależność istnieje. Wartości w macierzy korelacji C sugerują, że cena mieszkania X_1 skorelowana jest zarówno z jego powierzchnią X_2 jak i liczbą pokoi X_3, chociaż wydaje się, że wpływ powierzchni ma nieco większe znaczenie w kształtowaniu ceny mieszkań. Zwróćmy uwagę na dość silną zależność pomiędzy powierzchnią a liczbą pokoi. Może to sugerować, iż jedna z tych cech wpływa na cenę mieszkania poprzez drugą, tzn. jedna z nich nie ma istotnego udziału w cenie mieszkania. Może być na przykład w ten sposób, iż liczba pokoi nie oddziałuje na cenę, czyli mieszkania o podobnej powierzchni mogą mieć taką samą cenę bez względu na liczbę pokoi. Chcąc dokładniej zbadać ten problem wyznaczamy współczynniki korelacji cząstkowych ceny mieszkania X_1 z każdą z pozostałych zmiennych:

$$R_{12|3} = 0.7122, \ R_{13|2} = -0.0972.$$

Po wyeliminowaniu wpływu liczby pokoi, współczynnik korelacji między ceną a powierzchnią mieszkania wynosi 0.7122. Wartość ta sugeruje dość silną zależność między ceną a powierzchnią. Współczynnik korelacji cząstkowej między ceną a liczbą pokoi jest niemalże zerowy. Pozwala to wyprowadzić wniosek, iż podstawową charakterystyką mieszkania (spośród dwóch rozważanych) wpływającą na jego cenę jest powierzchnia. Liczba pokoi nie ma specjalnego udziału w kształtowaniu ceny.

Wyprowadzony wniosek należy sprawdzić formalnie. Formułujemy dwie hipotezy:

$$H_0: \ \varrho_{12|3} = 0, \qquad H_0: \ \varrho_{13|2} = 0.$$

Przyjmując poziom istotności $\alpha = 0.05$ odrzucamy pierwszą hipotezę oraz nie odrzucamy drugiej, ponieważ

$$|R_{12|3}| > r(0.05, 99) \quad \text{oraz} \quad |R_{13|2}| < r(0.05, 99).$$

Zatem z formalnego punktu widzenia wyprowadzony wniosek możemy zaakceptować przy założeniu, że zmienne X_1, X_2, X_3 mają łączny rozkład normalny.

Zadania do samodzielnego rozwiązania

11.1. W dziesięciu gospodarstwach wiejskich badano przeciętne dzienne spożycie ziemniaków w kg (X) i wielkość spożycia artykułów zbożowych w kg (Y) przypadającą na jednego członka rodziny.

x_i	0.70	0.60	0.80	0.85	0.55	0.65	0.90	1.00	0.75	0.50
y_i	0.50	0.70	0.50	0.40	0.75	0.60	0.30	0.20	0.55	0.70

Zbadać, czy istnieje zależność między cechami X oraz Y. Jeżeli zależność istnieje, to opisać ją za pomocą liniowej funkcji regresji.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

11.2. Spośród studentów pewnego wydziału wylosowano niezależnie dziesięciu studentów IV roku i otrzymano dla nich następujące średnie oceny uzyskane na I roku oraz na IV roku.

I rok	3.5	4.0	3.8	4.6	3.9	3.0	3.5	3.9	4.5	4.1
IV rok	4.2	3.9	3.8	4.5	4.2	3.4	3.8	3.9	4.6	4.0

Zbadać, czy istnieje zależność między wynikami studiów na I i na IV roku. Jeżeli taka zależność istnieje, to opisać ją za pomocą liniowej funkcji regresji.

11.3. Badano istnienie zależności między zawartością tłuszczu (X) i białka (Y) w mleku krów. Dokonano dziesięciu pomiarów i otrzymano wyniki:

$$\sum x_i = 38.6,\ \sum x_i^2 = 150.16,\ \sum y_i = 36.2,\ \sum y_i^2 = 131.74,\ \sum x_i y_i = 140.34$$

Zbadać istnienie zależności, a jeżeli taka zależność jest, to opisać ilościowo tę zależność.

11.4. Na podstawie poniższych danych uzyskanych z dwudziestu pomiarów zbadać, czy istnieje zależność między grubością włókna lnu (X) i grubością łodygi (Y).

$$\sum x_i = 83.5,\ \sum x_i^2 = 1269.25,\ \sum y_i = 8.74,\ \sum y_i^2 = 18.5454,\ \sum x_i y_i = 145.780$$

Jeżeli tak, to opisać ilościowo tę zależność.

11.5. W pewnej miejscowości dokonano piętnastu pomiarów temperatury (Y) dla różnych głębokości pod powierzchnią ziemi (X). Otrzymano następujące wyniki:

$$\sum x_i = 4200,\ \sum x_i^2 = 3640000,\ \sum y_i = 144,\ \sum y_i^2 = 3988,\ \sum x_i y_i = 120000$$

Zbadać istotność tej zależności. Wyznaczyć liniową funkcję regresji opisującą zależność temperatury od głębokości pod powierzchnią ziemi.

11.6. Zanotowano średnie oceny dziesięciu Studentów pewnego kierunku studiów jakie uzyskali na maturze (X_1), egzaminie wstępnym (X_2) oraz czwartym roku studiów (X_3). Otrzymano następujące wyniki:

$$\begin{array}{lll} \bar{x}_1 = 3.5 & \bar{x}_2 = 3.6 & \bar{x}_3 = 3.8 \\ \sum x_{1i}^2 = 211.07 & \sum x_{2i}^2 = 220.19 & \sum x_{3i}^2 = 245.76 \\ \sum x_{1i}x_{2i} = 215.20 & \sum x_{1i}x_{3i} = 227.38 & \sum x_{2i}x_{3i} = 232.56 \end{array}$$

Przeprowadzić analizę korelacji badanych zmiennych.

11.7. Wylosowano sto średniej klasy mieszkań. Każde mieszkanie scharakteryzowano ze względu na cenę jednego metra kwadratowego (X_1), odległość od pasów szybkiego ruchu (X_2) oraz odległość od najbliższego centrum handlowego (X_3). Otrzymano następujące wyniki:

$$\begin{array}{lll} \bar{x}_1 = 826.06 & \bar{x}_2 = 3.06 & \bar{x}_3 = 9.94 \\ \mathrm{var} x_1 = 815803.20 & \mathrm{var} x_2 = 208.80 & \mathrm{var} x_3 = 854.82 \\ \mathrm{cov}(x_1, x_2) = -8093.18 & \mathrm{cov}(x_1, x_3) = -14189.81 & \mathrm{cov}(x_2, x_3) = -7.82 \end{array}$$

Przeprowadzić analizę korelacji badanych zmiennych.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

11.8. O czterdziestu czterech krajach, określanych jako rozwijające się, zebrano dane dotyczące średniego poziomu zmian wielkości produktu krajowego brutto w skali roku (X_1), średniego rocznego udziału inwestycji zagranicznych w produkcie krajowym brutto (X_2) oraz średniego rocznego udziału inwestycji krajowych w produkcie krajowym brutto (X_3). Na podstawie zebranych danych otrzymano:

$$\begin{array}{lll}
\bar{x}_1 = 4.047 & \bar{x}_2 = 0.306 & \bar{x}_3 = 21.181 \\
\mathrm{var}(x_1) = 251.782 & \mathrm{var}(x_2) = 4.576 & \mathrm{var}(x_3) = 2130.87 \\
\mathrm{cov}(x_1, x_2) = 12.220 & \mathrm{cov}(x_1, x_3) = -123.195 & \mathrm{cov}(x_2, x_3) = 25.726
\end{array}$$

Przeprowadzić analizę korelacji badanych zmiennych.

11.9. Badano sprawność systemu dostaw piwa. Obserwowano czas (X_1), wielkość (X_2) oraz moment (X_3) rozpoczęcia dostawy. Zebrano 50–elementową próbę. Otrzymano wyniki:

$$\begin{array}{lll}
\bar{x}_1 = 116.50 & \bar{x}_2 = 208.16 & \bar{x}_3 = 9.50 \\
\mathrm{var}(x_1) = 88540.55 & \mathrm{var}(x_2) = 222240.48 & \mathrm{var}(x_3) = 608.50 \\
\mathrm{cov}(x_1, x_2) = 109580.66 & \mathrm{cov}(x_1, x_3) = 2244.00 & \mathrm{cov}(x_2, x_3) = 3070.00
\end{array}$$

Przeprowadzić analizę korelacji badanych zmiennych.

11.10. Wylosowano 60 prywatnych szkół średnich. O każdej szkole zebrano informacje z pięciu lat. Każdą szkołę scharakteryzowano pod względem udziału uczniów, którzy dostali się na państwowe studia wyższe (X_1), średniego udziału pieniędzy, które szkoła przeznaczyła bezpośrednio na edukację (X_2) oraz procentu uczniów ostatnich klas, których średnia z przedmiotów ścisłych przewyższała o 30% średnią krajową (X_3). Na podstawie zebranych informacji otrzymano:

$$\begin{array}{lll}
\bar{x}_1 = 45.01 & \bar{x}_2 = 48.63 & \bar{x}_3 = 20.01 \\
\mathrm{var}(x_1) = 12060.98 & \mathrm{var}(x_2) = 8161.56 & \mathrm{var}(x_3) = 4749.86 \\
\mathrm{cov}(x_1, x_2) = 593.82 & \mathrm{cov}(x_1, x_3) = 7156.35 & \mathrm{cov}(x_2, x_3) = -10.98
\end{array}$$

Przeprowadzić analizę korelacji badanych zmiennych.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

12. Iloraz szans, ryzyko względne, czułość, swoistość

Szansa. Szansą zdarzenia A o prawdopodobieństwie zajścia $P(A)$ nazywamy wielkość

$$S(A) = \frac{P(A)}{1 - P(A)}.$$

Pojecie szansy (stawki) wywodzi się z zakładów bukmacherskich jest innym wyrażeniem prawdopodobieństwa. Jest to stosunek prawdopodobieństwa zajścia zdarzenia do prawdopodobieństwa zajścia zdarzenia przeciwnego

Przykład. *Wyznaczyć szansę wylosowania pika z talii* 52 *kart.*

Rozwiązanie. *Niech A oznacza zdarzenie polegające na wylosowaniu pika z talii kart. Ponieważ pików w talii jest* 13*, więc $P(A) = \frac{13}{52} = \frac{1}{4}$. A zatem szansa na wylosowanie pika wynosi:*

$$S(A) = \frac{1/4}{1 - 1/4} = \frac{1}{3}.$$

Oznacza to, że mamy trzykrotnie mniejsze szanse na wyciągnięcie pika niż karty innego koloru. Można to także zinterpretować w ten sposób, że średnio oczekujemy jednego pika na każde wyciągnięte cztery karty. □

Iloraz szans. Ilorazem szans (ang. *odds ratio*) zdarzenia A względem zdarzenia B nazywamy wielkość

$$OR = \frac{S(A)}{S(B)},$$

gdzie $S(A)$ oraz $S(B)$ oznaczają odpowiednio szanse zdarzeń A i B, odpowiednio.

Przykład. *Wyznaczyć iloraz szans wylosowania pika z talii* 52 *kart względem uzyskania parzystej liczby oczek w jednokrotnym rzucie sześcienną kostką.*

Rozwiązanie. *Niech A oznacza zdarzenie polegające na wylosowaniu pika z talii kart. Szansa wylosowania pika wynosi $S(A) = \frac{1}{3}$. Niech B oznacza zdarzenie polegające na uzyskaniu parzystej liczby oczek w jednokrotnym rzucie sześcienną kostką Ponieważ $P(B) = \frac{1}{2}$, więc:*

$$S(B) = \frac{1/2}{1 - 1/2} = 1.$$

A zatem szukany iloraz szans wynosi

$$OR = \frac{1/3}{1} = \frac{1}{3}.$$

Oznacza to, że mamy trzykrotnie mniejsze szanse na wyciągnięcie pika niż uzyskanie parzystej liczby oczek w jednokrotnym rzucie kostką. □

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Ryzyko względne. Ryzykiem względnym (ang. *relative risk*) zdarzenia A względem zdarzenia B nazywamy wielkość

$$RR = \frac{P(A)}{P(B)},$$

gdzie $P(A)$ oraz $P(B)$ oznaczają odpowiednio prawdopodobieństwa zajść zdarzeń A i B, odpowiednio.

Przykład. *Wyznaczyć ryzyko względne wylosowania pika z talii 52 kart względem uzyskania parzystej liczby oczek w jednokrotnym rzucie sześcienną kostką.*

Rozwiązanie. *Niech A oznacza zdarzenie polegające na wylosowaniu pika z talii kart. Prawdopodobieństwo wylosowania pika wynosi $P(A) = \frac{1}{4}$. Niech B oznacza zdarzenie polegające na uzyskaniu parzystej liczby oczek w jednokrotnym rzucie sześcienną kostką Prawdopodobieństwo tego zdarzenia wynosi $P(B) = \frac{1}{2}$. A zatem ryzyko względne jest równe:*

$$RR = \frac{1/4}{1/2} = \frac{1}{2}.$$

Oznacza to, że dwukrotnie łatwiej jest uzyskać parzystą liczbę oczek niż wylosować pika. □

Estymacja punktowa. W praktyce badań medycznych bardzo często dysponujemy danymi dotyczącymi zachorowań (na określoną chorobę) w pewnych grupach pacjentów. Dane zazwyczaj zapisane są w następującej postaci:

	chory	zdrowy
Grupa I	a	b
Grupa II	c	d

Niech A oznacza zdarzenie polegające na wylosowaniu osoby chorej z grupy I, natomiast B - zdarzenie polegające na wylosowaniu osoby chorej z grupy II.

Oszacowaniami prawdopodobieństw zachorowań w każdej z grup są:

$$\widehat{P(A)} = \frac{a}{a+b}, \qquad \widehat{P(B)} = \frac{c}{c+d},$$

zaś oszacowaniami szans zachorowania w poszczególnych grupach są:

$$\widehat{S(A)} = \frac{a}{b}, \qquad \widehat{S(B)} = \frac{c}{d},$$

Oszacowaniem ilorazu szans zachorowań w grupie I względem zachorowań w grupie II jest:

$$\widehat{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Oszacowaniem ryzyka względnego zachorowania w grupie I względem zachorowania w grupie II jest:

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}.$$

Przedziały ufności. Dla ilorazu szans oraz dla ryzyka względnego konstruowane są przybliżone przedziały ufności. Niech $1-\alpha$ będzie poziomem ufności.

Przybliżony przedział ufności dla logarytmu naturalnego ilorazu szans ($\ln(OR)$) ma postać

$$\left(\ln(\widehat{OR}) - u_{1-\alpha/2}S_{OR}, \ln(\widehat{OR}) + u_{1-\alpha/2}S_{OR}\right),$$

gdzie

$$S_{OR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}},$$

natomiast $u_{1-\alpha/2}$ jest kwantylem rozkładu normalnego $N(0,1)$. Stąd łatwo otrzymujemy przybliżony przedział dla ilorazu szans OR:

$$\left(\widehat{OR} \cdot \exp\left(-u_{1-\alpha/2}S_{OR}\right), \widehat{OR} \cdot \exp\left(u_{1-\alpha/2}S_{OR}\right)\right),$$

Przybliżony przedział ufności dla logarytmu naturalnego ryzyka względnego ($\ln(RR)$) ma postać

$$\left(\ln(\widehat{RR}) - u_{1-\alpha/2}S_{RR}, \ln(\widehat{RR}) + u_{1-\alpha/2}S_{RR}\right),$$

gdzie

$$S_{RR} = \sqrt{\frac{b/a}{a+b} + \frac{d/c}{c+d}},$$

natomiast $u_{1-\alpha/2}$ jest kwantylem rozkładu normalnego $N(0,1)$. Stąd łatwo otrzymujemy przybliżony przedział dla ilorazu szans OR:

$$\left(\widehat{RR} \cdot \exp\left(-u_{1-\alpha/2}S_{RR}\right), \widehat{RR} \cdot \exp\left(u_{1-\alpha/2}S_{RR}\right)\right),$$

Przykład. *Chcemy zbadać, czy jest zależność między paleniem a płcią. Dla każdej osoby z grupy badanej zanotowano dwie dane: płeć i informację, o tym czy pali papierosy. Dane przedstawione są w tabeli.*

	palący	niepalący	łącznie
mężczyźni	70	120	190
kobiety	65	140	205
łącznie	135	260	*395*

Rozwiązanie. *Niech A oznacza zdarzenie polegające na wylosowaniu palącego mężczyzny, natomiast B – polegające na wylosowaniu palącej kobiety.*

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Poniżej podane są oszacowania poszczególnych miar ryzyka.

$$\widehat{P(A)} = \frac{70}{190} = 0.368$$
$$\widehat{S(A)} = \frac{70}{120} = 0.583$$
$$\widehat{P(B)} = \frac{65}{205} = 0.317$$
$$\widehat{S(B)} = \frac{65}{140} = 0.464$$
$$\widehat{OR} = \frac{70 \cdot 140}{65 \cdot 120} = 1.256$$
$$\widehat{RR} = \frac{70 \cdot 205}{65 \cdot 190} = 1.162$$

Iloraz szans oraz ryzyko względne wskazuje podwyższenie ryzyka palenia mężczyzn w stosunku do kobiet. Analizując ryzyko względne wnioskujemy, że mężczyźni mają o 16% większe ryzyko bycia palącym niż kobiety.

W celu uzyskania odpowiedzi na postawione w zadaniu pytanie skonstruujemy przedziały ufności (na poziomie ufności 0.95) dla ilorazu szans oraz ryzyka względnego. Odpowiedni kwantyl rozkładu normalnego $N(0,1)$ wynosi 1.96. Wykonując podstawienia otrzymujemy:

$$S_{OR} = \sqrt{\frac{1}{70} + \frac{1}{120} + \frac{1}{65} + \frac{1}{140}} = 0.002, \quad S_{RR} = \sqrt{\frac{120/70}{190} + \frac{140/65}{205}} = 0.0004.$$

Przedziały ufności mają postać:

$$OR \in \left(1.256 \cdot e^{-1.96 \cdot 0.002}, 1.256 \cdot e^{1.96 \cdot 0.002}\right) = (0.82, 1.90),$$

oraz

$$RR \in \left(1.162 \cdot e^{-1.96 \cdot 0.0004}, 1.162 \cdot e^{1.96 \cdot 0.0004}\right) = (0.88, 1.53).$$

Zauważmy, że w obu przypadkach przedziały zawierają jedynkę. Stwierdzamy, na poziomie istotności 0.05, że wpływ płci na palenie jest nieistotny. □

Czułość i swoistość. Jednym z ważnych zagadnień jest ocena jakości testu diagnostycznego pewnej choroby. Celem takich testów diagnostycznych jest wykrywanie diagnozowanej choroby, przy czym zawsze może się zdarzyć iż osoba w rzeczywistości chora zostanie zdiagnozowana jako zdrowa i przeciwnie: zdrowa osoba zostanie zdiagnozowana jako chora. Konsekwencje działania takiego testu można podsumować w następującej postaci:

	Wynik diagnozy:	
	zdrowa	chora
osoba zdrowa	diagnoza prawidłowa	diagnoza nieprawidłowa
osoba chora	diagnoza nieprawidłowa	diagnoza prawidłowa

Powiedzmy, że zdiagnozowano pewną liczbę osób. Niech:

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

TP (ang. *true positive*) oznacza liczbę osób chorych, poprawnie zdiagnozowanych jako chore;

TN (ang. *true negative*) oznacza liczbę osób zdrowych, poprawnie zdiagnozowanych jako zdrowe;

FP (ang. *false positive*) oznacza liczbę osób zdrowych, błędnie zdiagnozowanych jako chore;

FN (ang. *false negative*) oznacza liczbę osób chorych, błędnie zdiagnozowanych jako zdrowe.

Uzyskane wyniki próbkowe można zestawić w następującą tabelę:

	Wynik diagnozy:	
	zdrowa	chora
osoba zdrowa	TN	FP
osoba chora	FN	TP

Do najważniejszych miar oceny jakości testu diagnostycznego zaliczamy jego czułość i specyficzność

Czułość testu TPF (ang. *true positive fraction* lub *sensitivity*) dotyczy populacji chorych. Jest prawdopodobieństwo zdiagnozowania choroby wśród osób rzeczywiście chorych. Oznacza prawidłowe rozpoznanie choroby przez test diagnostyczny. Czułość określa się wzorem:

$$TPF = \frac{TP}{TP + FN}$$

Swoistość testu TNF (ang. *true negative fraction* lub *specificity*) dotyczy populacji zdrowych. Jest prawdopodobieństwo rozpoznania zdrowej osoby wśród rzeczywiście zdrowych. Swoistość określa się wzorem:

$$TNF = \frac{TN}{TN + FP}$$

Zauważmy, że test diagnostyczny jest tym lepszy im jego czułość i swoistość są większe.

Przykład. *Pewien test pozwalający wykryć chorobę nowotworową został przeprowadzony na grupie* 4000 *osób, wśród których było* 200 *osób z chorobą nowotworową oraz* 3800 *zdrowych. Test prawidłowo wykrył chorobę nowotworową u* 190 *chorych, zaś u* 10 *chorych wynik testu był negatywny. W grupie zdrowych test prawidłowo zakwalifikował* 3610 *jako zdrowe oraz u* 190 *osób zdrowych wynik testu był pozytywny. Wyznaczyć czułość i swoistość tego testu.*

Rozwiązanie. *Zapis symboliczny danych:* $TP = 190$, $FN = 10$, $TN = 3610$, $FP = 190$. *Sporządzamy odpowiednią tabelę:*

	Wynik diagnozy:	
	zdrowa	*chora*
osoba zdrowa	3610	190
osoba chora	10	190

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Obliczmy czułość i swoistość testu diagnostycznego

$$TPF = \frac{190}{190+10} = 0.95$$

$$TNF = \frac{3610}{3610+190} = 0.95.$$

Czułość wynosząca 0.95 *oznacza, że test jest prawidłowo wykrywa chorobę wśród chorych w* 95% *przypadków, zaś swoistość wynosząca* 0.95 *oznacza, że test prawidłowo kwalifikuje osoby zdrowe w* 95%. *Prawidłowy test powinien mieć możliwie wysoką czułość jak i swoistość, co nie zawsze jest możliwe do osiągnięcia.* □

Zadania do samodzielnego rozwiązania

12.1. Wstępne badania dotyczące występowania astmy oskrzelowej polegały na zadaniu respondentom pytania, czy kiedykolwiek chorował lub może nadal choruje na astmę. W Katowicach, na dwa i pół tysiąca wylosowanych osób w wieku od dwudziestu do czterdziestu lat, twierdząco na to pytanie odpowiedziało sto pięćdziesiąt osób. Wśród odpowiedzi twierdzących było 40% kobiet i 60% mężczyzn, zaś wśród odpowiedzi negatywnych było po 50% kobiet i mężczyzn. Wyznaczyć iloraz szans wraz z przedziałem ufności, na poziomie ufności 0.95, dla występowania astmy dla kobiet względem mężczyzn. Czy jest istotna zależność występowania astmy względem płci?

12.2. Obserwowano skłonność do podrażnienia skóry wśród dzieci z atopowym zapaleniem skóry (AZS). Pobrana próba dała wyniki:

	Dziecko	
	z podrażnieniem	bez podrażnienia
dziecko z AZS	3018	10890
dziecko bez AZS	2122	12445

Wyznaczyć przedział ufności, na poziomie ufności 0.95, dla ilorazu szans osób z podrażnieniem skóry w grupie dzieci z AZS względem dzieci z podrażnieniem skóry bez AZS. Czy na poziomie istotności 0.05 można stwierdzić zależność badanych cech?

12.3. Dla danych przedstawionych poniżej wyznaczyć podstawowe iloraz szans oraz ryzyko względne dla chorych na Alzhaimera mężczyzn względem kobiety. Wyznaczyć również odpowiednie przedziały ufności (poziom ufności 0.95).

	chory	zdrowy
mężczyźni	170	220
kobiety	265	240

12.4. Celem badania jest zbadanie, czy kobiety poddane hormonalnej terapii zastępczej HRT (ang. *Hormonal Replacement Therapy*) mają mniejsze szanse na złamanie biodra niż kobiety nie poddane HRT. Wyniki obserwacji podane są w poniższej tabeli. Stosując iloraz szans wraz z odpowiednimi przedziałami ufności (na poziomie ufności 0.95), udzielić odpowiedzi na postawione pytanie.

	poddane HRT	niepoddane HRT
ze złamaniem	40	1287
bez złamania	239	3023

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

12.5. Przez pięć lat prowadzono obserwacje pacjentów z chorobą układu krążenia. Pacjenci byli losowo podzieleni na dwie grupy. Jedna grupa przyjmowała regularnie aspirynę, a druga placebo. Celem było sprawdzenie, czy aspiryna zmniejsza ryzyko zawału serca. Wyznaczyć iloraz szans oraz ryzyko względne wraz z odpowiednimi przedziałami ufności (poziom ufności 0.95).

	po zawale	bez zawału
placebo	189	10845
aspiryna	104	10933

12.6. Chcemy sprawdzić, czy ryzyko zachorowania na syfilis zależy od liczby partnerów. W tym celu wybrano 51 przypadków zachorowań oraz 107 przypadków wolnych od tej choroby. W każdej grupie zliczono liczbę partnerów, która miała dana osoba. Wyznaczyć iloraz szans wraz z przedziałem ufności (poziom ufności 0.95).

	chory	zdrowy
więcej niż dwóch partnerów	41	58
żadnego partnera	10	49

12.7. Na podstawie poniższych wyników wyznaczyć czułość i swoistość testu diagnostycznego.

	diagnoza dodatnia	diagnoza ujemna
zakażenie	90	10
brak zakażenia	5	40

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Bibliografia

ABT S. 1972: Matematyczno–statystyczne podstawy analizy rynku, PWE, Warszawa.

BOX G. E. P., HUNTER W. G., HUNTER J. S. 1978: Statistics for experimenters, Wiley.

BOX G. E. P., JENKINS G. M. 1983: Analiza szeregów czasowych, PWN, Warszawa.

BOŻYK Z., RUDZKI W. 1977: Metody statystyczne w badaniu jakości produktów żywnościowych i chemicznych, WNT, Warszawa.

BRANDT S. 1976: Metody statystyczne i obliczeniowe analizy danych, PWN, Warszawa.

BROSS J. B. 1965: Jak podejmować decyzje, PWN, Warszawa.

CRAMÉR H. 1957: Metody matematyczne w statystyce, PWN, Warszawa.

DĄBKOWSKI J. 1992:, Statgraphics, Komputerowa Oficyna Wydawnicza „HELP”, Warszawa.

DĄBROWSKI A., GNOT S., MICHALSKI A., SRZEDNICKA J. 1994: Statystyka, 15 godzin z pakietem Statgraphics, Wydawnictwo Akademii Rolniczej, Wrocław.

DITTMANN P. 1998: Metody prognozowania sprzedaży w przedsiębiorstwie, wydanie 3, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego, Wrocław.

DOMAŃSKI C. 1979: Statystyczne testy nieparametryczne, PWE, Warszawa.

DRAPER N. R., SMITH H. 1966: Applied Regression Analysis, Wiley.

ELANDT R. 1964: Statystyka matematyczna w zastosowaniu do doświadczalnictwa rolniczego, PWN, Warszawa.

FISZ M. 1958: Rachunek prawdopodobieństwa i statystyka matematyczna, PWN, Warszawa.

GAWĘCKI J., WAGNER W. 1984: Podstawy metodologii badań doświadczalnych w nauce o żywieniu i żywności, PWN, Warszawa.

GNANADESIKAN R. 1982: Statistical data analysis, Proceedings of Symposia in Applied Mathematics, Vol 28, American Mathematical Society, Providence, Rhode Island.

GOLDBERGER A. S. 1975: Teoria ekonometrii, PWE, Warszawa.

GÓRCZYŃSKI J. 1993: Podstawy statystyki z przykładami w arkuszach kalkulacyjnych, Fundacja „Rozwój SGGW”, Warszawa.

GREŃ J. 1984: Statystyka matematyczna. Modele i zadania, PWN, Warszawa

GUPTA R. P. 1975: Applied Statistics, North–Holland.

HALD A. 1957: Statistical Theory with Engineering Applications, Wiley.

HELLWIG Z. 1975: Elementy rachunku prawdopodobieństwa i statystyki matematycznej, PWN, Warszawa.

JAJUGA K. 1993: Statystyczna analiza wielowymiarowa, Biblioteka ekonometryczna, PWN, Warszawa.

JÓŹWIAK J., PODGÓRSKI J. 1993: Statystyka od podstaw, PWE, Warszawa.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

KALA R. 1996: Elementy wnioskowania parametrycznego dla przyrodników, Wydawnictwo Akademii Rolniczej w Poznaniu, Poznań.

KASSYK–ROKICKA H. 1997: Statystyka. Zbiór zadań, PWE, Warszawa.

KASSYK–ROKICKA H. 1998: Mierniki statystyczne, PWE, Warszawa.

KLEIN L. R. 1965: Wstęp do ekonometrii, PWE, Warszawa.

KOBUS P., PIETRZYKOWSKI R., ZIELIŃSKI W. 2000: Statystyka z pakietem STATISTICA, wydanie II, Fundacja „Rozwój SGGW", Warszawa.

KRISHNAIAH P. R. 1980: Handbook of Statistics, vol. 1 — Analysis of Variance, North–Holland.

KRYSICKI W., BARTOS J., DYCZKA W., KRÓLIKOWSKA K., WASILEWSKI M. 1994: Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, Część II. Statystyka matematyczna, wyd. 2, PWN, Warszawa.

MICHALSKI T. 2000: Statystyka, WSiP, Warszawa.

NOWAK E. 1994: Zarys metod ekonometrii, Zbiór zadań, PWN, Warszawa.

LANGE O. 1958: Wstęp do ekonometrii, PWN, Warszawa.

LEVIN R. I. 1987: Statistics for managament, Prentice–Hall.

LUSZNIEWICZ A., SŁABY T. 1997: Statystyka stosowana, PWE, Warszawa.

OKTABA W. 1982: Elementy statystyki matematycznej i metodyka doświadczalnictwa, PWN, Warszawa.

OKTABA W. 1982: Metody statystyki matematycznej w doświadczalnictwie, PWN, Warszawa.

PAWŁOWSKI Z. 1978: Ekonometria, PWN, Warszawa.

PAWŁOWSKI Z. 1981: Statystyka matematyczna, PWN, Warszawa.

PLATT C. 1981: Problemy rachunku prawdopodobieństwa i statystyki matematycznej, PWN, Warszawa.

PODGÓRSKI J. 1993: Statystyka z komputerem, Pakiet STATGRAPHICS, „DIAGRAM", Warszawa.

RAO C. R. 1982: Modele liniowe statystyki matematycznej, PWN, Warszawa.

RAO C. R. 1994: Statystyka i prawda, PWN, Warszawa.

REICHMANN W. J. 1968: Drogi i bezdroża statystyki, PWN, Warszawa.

RÓSZKIEWICZ M. 1993: Statystyka — kurs podstawowy, SGH, Warszawa.

SCHEFFÉ H. 1959: The Analysis of Variance, Wiley.

SEARLE E. S. R. 1971: Linear models, Wiley.

SEARLE E. S. R. 1987: Linear Models for Unbalanced Data, Wiley.

SEBER G. A. F. 1977: Linear Regression Analysis, Wiley.

SRIVASTAVA J. N. 1975: A Survey of Statistical Design and Linear Models, North--Holland.

SILVEY S. D. 1978: Wnioskowanie statystyczne, PWN, Warszawa.

SOBCZAK M. 1997: Statystyka, PWE, Warszawa.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

STANISZ T. 1993: Funkcje jednej zmiennej w badaniach ekonomicznych, Biblioteka ekonometryczna, PWN, Warszawa.

STANISZ A. 2001: Przystępny kurs statystyki w oparciu o program STATISTICA PL na przykładach z medycyny, Tom I, wydanie II, StatSoft Polska, Kraków.

STANISZ A. 2000: Przystępny kurs statystyki w oparciu o program STATISTICA PL na przykładach z medycyny, Tom II, StatSoft Polska, Kraków.

TUKEY J. W. 1977: Exploratory data analysis, Addison–Wesley Publishing Company.

WELFE W. 1977: Ekonometryczne modele rynku, tom 1, metody ekonometryczne, PWE, Warszawa.

WELFE W. 1990: Gospodarki Polski w okresie transformacji. Zasady modelowania ekonometrycznego, PWE, Warszawa

WÓJCIK A. R. 1987: Statystyka matematyczna z elementami rachunku prawdopodobieństwa i statystyki opisowej, SGGW, Warszawa.

WÓJCIK A. R., LAUDAŃSKI Z. 1989: Planowanie i wnioskowanie statystyczne w doświadczalnictwie, PWN, Warszawa.

ZIELIŃSKI R. 1976: Rachunek prawdopodobieństwa z elementami statystyki matematycznej, wyd. II, WSiP, Warszawa.

ZIELIŃSKI R., ZIELIŃSKI W. 1987: Podręczne tablice statystyczne, WNT, Warszawa.

ZIELIŃSKI R., ZIELIŃSKI W. 1990: Tablice statystyczne, PWN, Warszawa.

ZIELIŃSKI T. 1999: Jak pokochać statystykę, czyli STATISTICA do poduszki, StatSoft Polska, Kraków

ZIELIŃSKI W. 1998: Analiza regresji, Fundacja „Rozwój SGGW", Warszawa

ZIELIŃSKI W. 1999: Wybrane testy statystyczne, wydanie II poprawione, Fundacja „Rozwój SGGW", Warszawa

ZIELIŃSKI W. 2000: Tablice statystyczne, wydanie IV poprawione, Fundacja „Rozwój SGGW", Warszawa

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablice

Tablica 1. Wartości $Q(k;n,p)$ w rozkładzie dwumianowym.
Podane są wartości prawdopodobieństw $Q(k;n,p) = P_{n,p}\{X \geq k\}$ dla $p = 0.01, 0.05(0.05)0.30, 0.40, 0.50$ oraz $n = 3(1)10, 15, 20$.

Tablica 2. Dystrubuanta rozkładu normalnego $N(0,1)$.
Podanych jest pięć pierwszych miejsc po przecinku wartości dystrybuanty $F(x)$ dla $x = 0(0.01)3.99$.

Tablica 3. Kwantyle rozkładu normalnego $N(0,1)$.
Podane są wartości kwantyli u_α dla $\alpha = 0.5(0.01)0.9(0.001)0.999$.

Tablica 4. Wartości krytyczne rozkładu t–Studenta.
Podane są dwustronne wartości krytyczne $t(\alpha;r)$ dla $\alpha = 0.01, 0.025,$ 0.05, 0.1 oraz $r = 1(1)40(2)100(50)550, \infty$.

Tablica 5. Wartości krytyczne rozkładu chi–kwadrat.
Podane są wartości krytyczne $\chi^2(\alpha;r)$ dla $\alpha = 0.99, 0.975, 0.95, 0.9,$ $0.1, 0.05, 0.025, 0.01$ oraz $r = 1(1)30(5)100$.

Tablica 6. Wartości krytyczne rozkładu F–Snedecora.
Podane są wartości krytyczne $F(\alpha;r_1,r_2)$ dla $\alpha = 0.01, 0.05$ oraz $r_1 = 1(1)10(5)40(20)100, \infty$, $r_2 = 1(1)10(2)20(10)40(20)100, \infty$.

Tablica 7. Wartości krytyczne studentyzowanego rozstępu.
Podane są wartości krytyczne $t(\alpha;k,r)$ dla $\alpha = 0.01, 0.05$ oraz $k = 2(1)10$, $r = 3(1)20, 30, 40, 60, 120, \infty$.

Tablica 8. Wartości krytyczne współczynnika korelacji.
Podane są wartości krytyczne $r(\alpha;n)$ dla $\alpha = 0.01, 0.025, 0.05, 0.1$ oraz $n = 3(1)80$.

Tablica 9. Wartości krytyczne współczynnika korelacji wielokrotnej.
Podane są wartości krytyczne $r(\alpha;n,k)$ dla $\alpha = 0.01, 0.05$ oraz $n = 4(1)50(5)95$, $k = 3, 4, 5$.

Tablica 10. Wartości krytyczne w teście Behrensa–Fishera.
Podane są wartości krytyczne $V(\alpha;r_1,r_2,c)$ dla $\alpha = 0.01, 0.05$ oraz $r_1 = 5(1)10, 15, 20$, $r_2 = 2(1)r_1$, $c = 0.0(0.1)1.0$.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 1. Wartości $Q(k;n,p)$ w rozkładzie dwumianowym

n	k	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50
3	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.02970	.14263	.27100	.38588	.48800	.57813	.65700	.78400	.87500
	2	.00030	.00725	.02800	.06075	.10400	.15625	.21600	.35200	.50000
	3	.00000	.00013	.00100	.00338	.00800	.01563	.02700	.06400	.12500
4	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.03940	.18549	.34390	.47799	.59040	.68359	.75990	.87040	.93750
	2	.00059	.01402	.05230	.10952	.18080	.26172	.34830	.52480	.68750
	3	.00000	.00048	.00370	.01198	.02720	.05078	.08370	.17920	.31250
	4		.00001	.00010	.00051	.00160	.00391	.00810	.02560	.06250
5	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.04901	.22622	.40951	.55629	.67232	.76270	.83193	.92224	.96875
	2	.00098	.02259	.08146	.16479	.26272	.36719	.47178	.66304	.81250
	3	.00001	.00116	.00856	.02661	.05792	.10352	.16308	.31744	.50000
	4	.00000	.00003	.00046	.00223	.00672	.01562	.03078	.08704	.18750
	5		.00000	.00001	.00008	.00032	.00098	.00243	.01024	.03125
6	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.05852	.26491	.46856	.62285	.73786	.82202	.88235	.95334	.98438
	2	.00146	.03277	.11427	.22352	.34464	.46606	.57983	.76672	.89063
	3	.00002	.00223	.01585	.04734	.09888	.16943	.25569	.45568	.65625
	4	.00000	.00009	.00127	.00589	.01696	.03760	.07047	.17920	.34375
	5		.00000	.00005	.00040	.00160	.00464	.01094	.04096	.10938
	6			.00000	.00001	.00006	.00024	.00073	.00410	.01562
7	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.06793	.30166	.52170	.67942	.79028	.86652	.91765	.97201	.99219
	2	.00203	.04438	.14969	.28342	.42328	.55505	.67058	.84137	.93750
	3	.00003	.00376	.02569	.07377	.14803	.24359	.35293	.58010	.77344
	4	.00000	.00019	.00273	.01210	.03334	.07056	.12604	.28979	.50000
	5		.00001	.00018	.00122	.00467	.01288	.02880	.09626	.22656
	6			.00001	.00007	.00037	.00134	.00379	.01884	.06250
	7			.00000	.00000	.00001	.00006	.00022	.00164	.00781
8	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.07726	.33658	.56953	.72751	.83223	.89989	.94235	.98320	.99609
	2	.00269	.05724	.18690	.34282	.49668	.63292	.74470	.89362	.96484
	3	.00005	.00579	.03809	.10521	.20308	.32146	.44823	.68461	.85547
	4	.00000	.00037	.00502	.02135	.05628	.11382	.19410	.40591	.63672
	5		.00002	.00043	.00285	.01041	.02730	.05797	.17367	.36328
	6		.00000	.00002	.00024	.00123	.00423	.01129	.04981	.14453
	7			.00000	.00001	.00008	.00038	.00129	.00852	.03516
	8				.00000	.00000	.00002	.00007	.00066	.00391
9	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.08648	.36975	.61258	.76838	.86578	.92492	.95965	.98992	.99805
	2	.00344	.07121	.22516	.40052	.56379	.69966	.80400	.92946	.98047
	3	.00008	.00836	.05297	.14085	.26180	.39932	.53717	.76821	.91016
	4	.00000	.00064	.00833	.03393	.08564	.16573	.27034	.51739	.74609
	5		.00003	.00089	.00563	.01958	.04893	.09881	.26657	.50000
	6		.00000	.00006	.00063	.00307	.00999	.02529	.09935	.25391
	7			.00000	.00005	.00031	.00134	.00429	.02503	.08984
	8				.00000	.00002	.00011	.00043	.00380	.01953
	9					.00000	.00000	.00002	.00026	.00195

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

n	k					p				
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50
10	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.09562	.40126	.65132	.80313	.89263	.94369	.97175	.99395	.99902
	2	.00427	.08614	.26390	.45570	.62419	.75597	.85069	.95364	.98926
	3	.00011	.01150	.07019	.17980	.32220	.47441	.61722	.83271	.94531
	4	.00000	.00103	.01280	.04997	.12087	.22412	.35039	.61772	.82813
	5		.00006	.00163	.00987	.03279	.07813	.15027	.36690	.62305
	6		.00000	.00015	.00138	.00637	.01973	.04735	.16624	.37695
	7			.00001	.00013	.00086	.00351	.01059	.05476	.17188
	8			.00000	.00001	.00008	.00042	.00159	.01229	.05469
	9				.00000	.00000	.00003	.00014	.00168	.01074
	10					.00000	.00000	.00001	.00010	.00098
15	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
	1	.13994	.53671	.79411	.91265	.96482	.98664	.99525	.99953	.99997
	2	.00963	.17095	.45096	.68141	.83287	.91982	.96473	.99483	.99951
	3	.00042	.03620	.18406	.39577	.60198	.76391	.87317	.97289	.99631
	4	.00001	.00547	.05556	.17734	.35184	.53871	.70313	.90950	.98242
	5	.00000	.00061	.01272	.06171	.16423	.31351	.48451	.78272	.94077
	6		.00005	.00225	.01681	.06105	.14837	.27838	.59678	.84912
	7		.00000	.00031	.00361	.01806	.05662	.13114	.39019	.69638
	8			.00003	.00061	.00424	.01730	.05001	.21310	.50000
	9			.00000	.00008	.00078	.00419	.01524	.09505	.30362
	10				.00001	.00011	.00079	.00365	.03383	.15088
	11				.00000	.00001	.00012	.00067	.00935	.05923
	12					.00000	.00001	.00009	.00193	.01758
	13						.00000	.00001	.00028	.00369
	14							.00000	.00003	.00049
	15								.00000	.00003
20	0	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	
	1	.18209	.64151	.87842	.96124	.98847	.99683	.99920	.99996	1.00000
	2	.01686	.26416	.60825	.82444	.93082	.97569	.99236	.99948	.99998
	3	.00100	.07548	.32307	.59510	.79392	.90874	.96452	.99639	.99980
	4	.00004	.01590	.13295	.35227	.58855	.77484	.89291	.98404	.99871
	5	.00000	.00257	.04317	.17015	.37035	.58516	.76249	.94905	.99409
	6		.00033	.01125	.06731	.19579	.38283	.58363	.87440	.97931
	7		.00003	.00239	.02194	.08669	.21422	.39199	.74999	.94234
	8		.00000	.00042	.00592	.03214	.10181	.22773	.58411	.86841
	9			.00006	.00133	.00998	.04093	.11333	.40440	.74828
	10			.00001	.00025	.00259	.01386	.04796	.24466	.58810
	11			.00000	.00004	.00056	.00394	.01714	.12752	.41190
	12				.00000	.00010	.00094	.00514	.05653	.25172
	13					.00002	.00018	.00128	.02103	.13159
	14					.00000	.00003	.00026	.00647	.05766
	15						.00000	.00004	.00161	.02069
	16							.00001	.00032	.00591
	17							.00000	.00005	.00129
	18								.00001	.00020
	19									.00002
	20									.00000

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 2. Dystrubuanta $F(x)$ rozkładu normalnego $N(0,1)$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0.1	53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0.2	57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0.3	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0.4	65542	65910	66276	66640	67003	67364	67724	68082	68439	68793
0.5	69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0.6	72575	72907	73237	73565	73891	74215	74537	74857	75175	75490
0.7	75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0.8	78814	79103	79389	79673	79955	80234	80511	80785	81057	81327
0.9	81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1.0	84134	84375	84614	84849	85083	85314	85543	85769	85993	86214
1.1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1.2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1.3	90320	90490	90658	90824	90988	91149	91308	91466	91621	91774
1.4	91924	92073	92220	92364	92507	92647	92785	92922	93056	93189
1.5	93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1.6	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1.7	95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1.8	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1.9	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2.0	97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2.1	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2.2	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2.3	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2.4	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2.5	99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2.6	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2.7	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2.8	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2.9	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3.0	99865	99869	99874	99878	99882	99886	99889	99893	99896	99900
3.1	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3.2	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3.3	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3.4	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3.5	99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3.6	99984	99985	99985	99986	99986	99987	99987	99988	99988	99989
3.7	99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3.8	99993	99993	99993	99994	99994	99994	99994	99995	99995	99995
3.9	99995	99995	99996	99996	99996	99996	99996	99996	99997	99997

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 3. Kwantyle u_α rozkładu normalnego $N(0,1)$

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.5	0.0000	0.0251	0.0502	0.0753	0.1004	0.1257	0.1510	0.1764	0.2019	0.2275
0.6	0.2533	0.2793	0.3055	0.3319	0.3585	0.3853	0.4125	0.4399	0.4677	0.4958
0.7	0.5244	0.5534	0.5828	0.6128	0.6433	0.6745	0.7063	0.7388	0.7722	0.8064
0.8	0.8416	0.8779	0.9154	0.9542	0.9945	1.0364	1.0803	1.1264	1.1750	1.2265

α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.90	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
0.91	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
0.92	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
0.93	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
0.94	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
0.95	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
0.96	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
0.97	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
0.98	2.0537	2.0748	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
0.99	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

α	u_α
0.9995	3.290
0.9999	3.719
0.99995	3.891
0.99999	4.265
0.999995	4.417
0.999999	4.753
0.9999995	4.892
0.9999999	5.199

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 4. Wartości krytyczne $t(\alpha; r)$ rozkładu t–Studenta

$r \setminus \alpha$	0.1	0.05	0.025	0.01	$r \setminus \alpha$	0.1	0.05	0.025	0.01
1	6.3137	12.7062	25.4519	63.6559	42	1.6820	2.0181	2.3246	2.6981
2	2.9200	4.3027	6.2054	9.9250	44	1.6802	2.0154	2.3207	2.6923
3	2.3534	3.1824	4.1765	5.8408	46	1.6787	2.0129	2.3172	2.6870
4	2.1318	2.7765	3.4954	4.6041	48	1.6772	2.0106	2.3139	2.6822
5	2.0150	2.5706	3.1634	4.0321	50	1.6759	2.0086	2.3109	2.6778
6	1.9432	2.4469	2.9687	3.7074	52	1.6747	2.0066	2.3082	2.6737
7	1.8946	2.3646	2.8412	3.4995	54	1.6736	2.0049	2.3056	2.6700
8	1.8595	2.3060	2.7515	3.3554	56	1.6725	2.0032	2.3033	2.6665
9	1.8331	2.2622	2.6850	3.2498	58	1.6716	2.0017	2.3011	2.6633
10	1.8125	2.2281	2.6338	3.1693	60	1.6706	2.0003	2.2990	2.6603
11	1.7959	2.2010	2.5931	3.1058	62	1.6698	1.9990	2.2971	2.6575
12	1.7823	2.1788	2.5600	3.0545	64	1.6690	1.9977	2.2954	2.6549
13	1.7709	2.1604	2.5326	3.0123	66	1.6683	1.9966	2.2937	2.6524
14	1.7613	2.1448	2.5096	2.9768	68	1.6676	1.9955	2.2921	2.6501
15	1.7531	2.1315	2.4899	2.9467	70	1.6669	1.9944	2.2906	2.6479
16	1.7459	2.1199	2.4729	2.9208	72	1.6663	1.9935	2.2892	2.6458
17	1.7396	2.1098	2.4581	2.8982	74	1.6657	1.9925	2.2879	2.6439
18	1.7341	2.1009	2.4450	2.8784	76	1.6652	1.9917	2.2867	2.6421
19	1.7291	2.0930	2.4334	2.8609	78	1.6646	1.9908	2.2855	2.6403
20	1.7247	2.0860	2.4231	2.8453	80	1.6641	1.9901	2.2844	2.6387
21	1.7207	2.0796	2.4138	2.8314	82	1.6636	1.9893	2.2833	2.6371
22	1.7171	2.0739	2.4055	2.8188	84	1.6632	1.9886	2.2823	2.6356
23	1.7139	2.0687	2.3979	2.8073	86	1.6628	1.9879	2.2813	2.6342
24	1.7109	2.0639	2.3910	2.7970	88	1.6624	1.9873	2.2804	2.6329
25	1.7081	2.0595	2.3846	2.7874	90	1.6620	1.9867	2.2795	2.6316
26	1.7056	2.0555	2.3788	2.7787	92	1.6616	1.9861	2.2787	2.6303
27	1.7033	2.0518	2.3734	2.7707	94	1.6612	1.9855	2.2779	2.6291
28	1.7011	2.0484	2.3685	2.7633	96	1.6609	1.9850	2.2771	2.6280
29	1.6991	2.0452	2.3638	2.7564	98	1.6606	1.9845	2.2764	2.6269
30	1.6973	2.0423	2.3596	2.7500	100	1.6602	1.9840	2.2757	2.6259
31	1.6955	2.0395	2.3556	2.7440	150	1.6551	1.9759	2.2641	2.6090
32	1.6939	2.0369	2.3518	2.7385	200	1.6525	1.9719	2.2584	2.6006
33	1.6924	2.0345	2.3483	2.7333	250	1.6510	1.9695	2.2550	2.5956
34	1.6909	2.0322	2.3451	2.7284	300	1.6499	1.9679	2.2527	2.5923
35	1.6896	2.0301	2.3420	2.7238	350	1.6492	1.9668	2.2511	2.5899
36	1.6883	2.0281	2.3391	2.7195	400	1.6487	1.9659	2.2499	2.5882
37	1.6871	2.0262	2.3363	2.7154	450	1.6482	1.9652	2.2489	2.5868
38	1.6860	2.0244	2.3337	2.7116	500	1.6479	1.9647	2.2482	2.5857
39	1.6849	2.0227	2.3313	2.7079	550	1.6476	1.9643	2.2476	2.5848
40	1.6839	2.0211	2.3289	2.7045	∞	1.6449	1.9600	2.2414	2.5758

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 5. Wartości krytyczne $\chi^2(\alpha; r)$ rozkładu chi–kwadrat

$r \setminus \alpha$	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0.0^3157	0.0^3982	0.0^2393	0.0158	2.7055	3.8415	5.0239	6.6349
2	0.0201	0.0506	0.1026	0.2107	4.6052	5.9915	7.3778	9.2104
3	0.1148	0.2158	0.3518	0.5844	6.2514	7.8147	9.3484	11.3449
4	0.2971	0.4844	0.7107	1.0636	7.7794	9.4877	11.1433	13.2767
5	0.5543	0.8312	1.1455	1.6103	9.2363	11.0705	12.8325	15.0863
6	0.8721	1.2373	1.6354	2.2041	10.6446	12.5916	14.4494	16.8119
7	1.2390	1.6899	2.1673	2.8331	12.0170	14.0671	16.0128	18.4753
8	1.6465	2.1797	2.7326	3.4895	13.3616	15.5073	17.5345	20.0902
9	2.0879	2.7004	3.3251	4.1682	14.6837	16.9190	19.0228	21.6660
10	2.5582	3.2470	3.9403	4.8652	15.9872	18.3070	20.4832	23.2093
11	3.0535	3.8157	4.5748	5.5778	17.2750	19.6752	21.9200	24.7250
12	3.5706	4.4038	5.2260	6.3038	18.5493	21.0261	23.3367	26.2170
13	4.1069	5.0087	5.8919	7.0415	19.8119	22.3620	24.7356	27.6882
14	4.6604	5.6287	6.5706	7.7895	21.0641	23.6848	26.1189	29.1412
15	5.2294	6.2621	7.2609	8.5468	22.3071	24.9958	27.4884	30.5780
16	5.8122	6.9077	7.9616	9.3122	23.5418	26.2962	28.8453	31.9999
17	6.4077	7.5642	8.6718	10.0852	24.7690	27.5871	30.1910	33.4087
18	7.0149	8.2307	9.3904	10.8649	25.9894	28.8693	31.5264	34.8052
19	7.6327	8.9065	10.1170	11.6509	27.2036	30.1435	32.8523	36.1908
20	8.2604	9.5908	10.8508	12.4426	28.4120	31.4104	34.1696	37.5663
21	8.8972	10.2829	11.5913	13.2396	29.6151	32.6706	35.4789	38.9322
22	9.5425	10.9823	12.3380	14.0415	30.8133	33.9245	36.7807	40.2894
23	10.1957	11.6885	13.0905	14.8480	32.0069	35.1725	38.0756	41.6383
24	10.8563	12.4011	13.8484	15.6587	33.1962	36.4150	39.3641	42.9798
25	11.5240	13.1197	14.6114	16.4734	34.3816	37.6525	40.6465	44.3140
26	12.1982	13.8439	15.3792	17.2919	35.5632	38.8851	41.9231	45.6416
27	12.8785	14.5734	16.1514	18.1139	36.7412	40.1133	43.1945	46.9628
28	13.5647	15.3079	16.9279	18.9392	37.9159	41.3372	44.4608	48.2782
29	14.2564	16.0471	17.7084	19.7677	39.0875	42.5569	45.7223	49.5878
30	14.9535	16.7908	18.4927	20.5992	40.2560	43.7730	46.9792	50.8922
35	18.5089	20.5694	22.4650	24.7966	46.0588	49.8018	53.2033	57.3420
40	22.1642	24.4331	26.5093	29.0505	51.8050	55.7585	59.3417	63.6908
45	25.9012	28.3662	30.6123	33.3504	57.5053	61.6562	65.4101	69.9569
50	29.7067	32.3574	34.7642	37.6886	63.1671	67.5048	71.4202	76.1538
55	33.5705	36.3981	38.9581	42.0596	68.7962	73.3115	77.3804	82.2920
60	37.4848	40.4817	43.1880	46.4589	74.3970	79.0820	83.2977	88.3794
65	41.4436	44.6030	47.4496	50.8829	79.9730	84.8206	89.1772	94.4220
70	45.4417	48.7575	51.7393	55.3289	85.5270	90.5313	95.0231	100.4251
75	49.4751	52.9419	56.0541	59.7946	91.0615	96.2167	100.8393	106.3929
80	53.5400	57.1532	60.3915	64.2778	96.5782	101.8795	106.6285	112.3288
85	57.6339	61.3888	64.7494	68.7771	102.0789	107.5217	112.3933	118.2356
90	61.7540	65.6466	69.1260	73.2911	107.5650	113.1452	118.1359	124.1162
95	65.8983	69.9249	73.5198	77.8184	113.0377	118.7516	123.8580	129.9725
100	70.0650	74.2219	77.9294	82.3581	118.4980	124.3421	129.5613	135.8069

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 6. Wartości krytyczne $\begin{matrix} F(0.05;r_1,r_2) \\ F(0.01;r_1,r_2) \end{matrix}$ rozkładu F–Snedecora

$r_2 \diagdown r_1$	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
	4052.2	4999.3	5403.5	5624.3	5764.0	5859.0	5928.3	5981.0	6022.4	6055.9
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396
	98.502	99.000	99.164	99.251	99.302	99.331	99.357	99.375	99.390	99.397
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785
	34.116	30.816	29.457	28.710	28.237	27.911	27.671	27.489	27.345	27.228
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347
	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
	10.562	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
	7.314	5.178	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
	6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637	2.551
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831
	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

$r_2 \backslash r_1$	15	20	25	30	35	40	60	80	100	∞
1	245.95	248.02	249.26	250.10	250.69	251.14	252.20	252.72	253.04	254.32
	6157.0	6208.7	6239.9	6260.4	6275.3	6286.4	6313.0	6326.5	6333.9	6365.6
2	19.429	19.446	19.456	19.463	19.467	19.471	19.479	19.483	19.486	19.496
	99.433	99.448	99.459	99.466	99.470	99.477	99.484	99.484	99.491	99.499
3	8.703	8.660	8.634	8.617	8.604	8.594	8.572	8.561	8.554	8.526
	26.872	26.690	26.579	26.504	26.451	26.411	26.316	26.269	26.241	26.125
4	5.858	5.803	5.769	5.746	5.729	5.717	5.688	5.673	5.664	5.628
	14.198	14.019	13.911	13.838	13.785	13.745	13.652	13.605	13.577	13.463
5	4.619	4.558	4.521	4.496	4.478	4.464	4.431	4.415	4.405	4.365
	9.722	9.553	9.449	9.379	9.329	9.291	9.202	9.157	9.130	9.020
6	3.938	3.874	3.835	3.808	3.789	3.774	3.740	3.722	3.712	3.669
	7.559	7.396	7.296	7.229	7.180	7.143	7.057	7.013	6.987	6.880
7	3.511	3.445	3.404	3.376	3.356	3.340	3.304	3.286	3.275	3.230
	6.314	6.155	6.058	5.992	5.944	5.908	5.824	5.781	5.755	5.650
8	3.218	3.150	3.108	3.079	3.059	3.043	3.005	2.986	2.975	2.928
	5.515	5.359	5.263	5.198	5.151	5.116	5.032	4.989	4.963	4.859
9	3.006	2.936	2.893	2.864	2.842	2.826	2.787	2.768	2.756	2.707
	4.962	4.808	4.713	4.649	4.602	4.567	4.483	4.441	4.415	4.311
10	2.845	2.774	2.730	2.700	2.678	2.661	2.621	2.601	2.588	2.538
	4.558	4.405	4.311	4.247	4.201	4.165	4.082	4.039	4.014	3.909
12	2.617	2.544	2.498	2.466	2.443	2.426	2.384	2.363	2.350	2.296
	4.010	3.858	3.765	3.701	3.654	3.619	3.535	3.493	3.467	3.361
14	2.463	2.388	2.341	2.308	2.284	2.266	2.223	2.201	2.187	2.131
	3.656	3.505	3.412	3.348	3.301	3.266	3.181	3.138	3.112	3.004
16	2.352	2.276	2.227	2.194	2.169	2.151	2.106	2.083	2.068	2.010
	3.409	3.259	3.165	3.101	3.054	3.018	2.933	2.889	2.863	2.753
18	2.269	2.191	2.141	2.107	2.082	2.063	2.017	1.993	1.978	1.917
	3.227	3.077	2.983	2.919	2.871	2.835	2.749	2.705	2.678	2.566
20	2.203	2.124	2.074	2.039	2.013	1.994	1.946	1.922	1.907	1.843
	3.088	2.938	2.843	2.778	2.731	2.695	2.608	2.563	2.535	2.421
30	2.015	1.932	1.878	1.841	1.813	1.792	1.740	1.712	1.695	1.622
	2.700	2.549	2.453	2.386	2.337	2.299	2.208	2.160	2.131	2.006
40	1.924	1.839	1.783	1.744	1.715	1.693	1.637	1.608	1.589	1.509
	2.522	2.369	2.271	2.203	2.153	2.114	2.019	1.969	1.938	1.805
60	1.836	1.748	1.690	1.649	1.618	1.594	1.534	1.502	1.481	1.389
	2.352	2.198	2.098	2.028	1.976	1.936	1.836	1.783	1.749	1.601
80	1.793	1.703	1.644	1.602	1.570	1.545	1.482	1.448	1.426	1.325
	2.271	2.115	2.015	1.944	1.890	1.849	1.746	1.690	1.655	1.494
100	1.768	1.676	1.616	1.573	1.541	1.515	1.450	1.415	1.392	1.283
	2.223	2.067	1.965	1.893	1.839	1.797	1.692	1.634	1.598	1.427
∞	1.666	1.571	1.506	1.459	1.423	1.394	1.318	1.274	1.243	1.000
	2.039	1.878	1.773	1.696	1.638	1.592	1.473	1.404	1.358	1.000

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 7. Wartości krytyczne $\substack{t(0.05;k,r) \\ t(0.01;k,r)}$ studentyzowanego rozstępu

$r \diagdown k$	2	3	4	5	6	7	8	9	10
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462
	8.261	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826
	6.512	8.120	9.173	9.958	10.58	11.10	11.55	11.93	12.27
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995
	5.702	6.976	7.804	8.421	8.913	9.321	9.669	9.972	10.24
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493
	5.243	6.331	7.033	7.556	7.973	8.318	8.613	8.869	9.097
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158
	4.949	5.919	6.543	7.005	7.373	7.679	7.939	8.166	8.368
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918
	4.746	5.635	6.204	6.625	6.960	7.237	7.474	7.681	7.863
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739
	4.596	5.428	5.957	6.348	6.658	6.915	7.143	7.325	7.495
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599
	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.055	7.213
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
	4.392	5.146	5.621	5.970	6.247	6.476	6.672	6.842	6.992
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
	4.320	5.046	5.502	5.836	6.101	6.321	6.507	6.670	6.814
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
	4.260	4.964	5.404	5.727	5.981	6.192	6.372	6.528	6.667
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
	4.168	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.439
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
	4.131	4.786	5.192	5.489	5.722	5.915	6.079	6.222	6.349
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
	4.071	4.703	5.094	5.379	5.603	5.788	5.944	6.081	6.201
19	2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038
	4.046	4.670	5.054	5.334	5.554	5.735	5.889	6.022	6.141
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
	4.024	4.639	5.018	5.294	5.510	5.688	5.839	5.970	6.087
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
	3.956	4.546	4.907	5.168	5.374	5.542	5.685	5.809	5.919
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
	3.889	4.455	4.799	5.048	5.242	5.401	5.537	5.653	5.756
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
	3.825	4.367	4.696	4.931	5.114	5.265	5.392	5.502	5.599
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
	3.762	4.282	4.595	4.818	4.991	5.133	5.253	5.356	5.447
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
	3.702	4.200	4.497	4.709	4.872	5.005	5.118	5.214	5.299
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474
	3.643	4.120	4.403	4.603	4.757	4.882	4.987	5.078	5.157

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 8. Wartości krytyczne $r(\alpha; n)$ współczynnika korelacji

$n \diagdown \alpha$	0.1	0.05	0.025	0.01	$n \diagdown \alpha$	0.1	0.05	0.025	0.01
3	0.9877	0.9969	0.9992	0.9999	42	0.2573	0.3044	0.3456	0.3932
4	0.9000	0.9500	0.9750	0.9900	43	0.2542	0.3008	0.3415	0.3887
5	0.8054	0.8783	0.9237	0.9587	44	0.2512	0.2973	0.3376	0.3843
6	0.7293	0.8114	0.8680	0.9172	45	0.2483	0.2940	0.3339	0.3801
7	0.6694	0.7545	0.8166	0.8745	46	0.2455	0.2907	0.3302	0.3761
8	0.6215	0.7067	0.7713	0.8343	47	0.2429	0.2876	0.3267	0.3721
9	0.5822	0.6664	0.7318	0.7977	48	0.2403	0.2845	0.3233	0.3683
10	0.5494	0.6319	0.6973	0.7646	49	0.2377	0.2816	0.3200	0.3646
11	0.5214	0.6021	0.6669	0.7348	50	0.2353	0.2787	0.3168	0.3610
12	0.4973	0.5760	0.6400	0.7079	51	0.2329	0.2759	0.3137	0.3575
13	0.4762	0.5529	0.6159	0.6835	52	0.2306	0.2732	0.3106	0.3542
14	0.4575	0.5324	0.5943	0.6614	53	0.2284	0.2706	0.3077	0.3509
15	0.4409	0.5140	0.5748	0.6411	54	0.2262	0.2681	0.3048	0.3477
16	0.4259	0.4973	0.5570	0.6226	55	0.2241	0.2656	0.3021	0.3445
17	0.4124	0.4821	0.5408	0.6055	56	0.2221	0.2632	0.2994	0.3415
18	0.4000	0.4683	0.5258	0.5897	57	0.2201	0.2609	0.2967	0.3385
19	0.3887	0.4555	0.5121	0.5751	58	0.2181	0.2586	0.2942	0.3357
20	0.3783	0.4438	0.4993	0.5614	59	0.2162	0.2564	0.2917	0.3328
21	0.3687	0.4329	0.4875	0.5487	60	0.2144	0.2542	0.2892	0.3301
22	0.3598	0.4227	0.4764	0.5368	61	0.2126	0.2521	0.2869	0.3274
23	0.3515	0.4132	0.4660	0.5256	62	0.2108	0.2500	0.2845	0.3248
24	0.3438	0.4044	0.4563	0.5151	63	0.2091	0.2480	0.2823	0.3223
25	0.3365	0.3961	0.4472	0.5052	64	0.2075	0.2461	0.2801	0.3198
26	0.3297	0.3882	0.4386	0.4958	65	0.2058	0.2441	0.2779	0.3173
27	0.3233	0.3809	0.4305	0.4869	66	0.2042	0.2423	0.2758	0.3150
28	0.3172	0.3739	0.4228	0.4785	67	0.2027	0.2404	0.2737	0.3126
29	0.3115	0.3673	0.4155	0.4705	68	0.2012	0.2387	0.2717	0.3104
30	0.3061	0.3610	0.4085	0.4629	69	0.1997	0.2369	0.2697	0.3081
31	0.3009	0.3550	0.4019	0.4556	70	0.1982	0.2352	0.2678	0.3060
32	0.2960	0.3494	0.3956	0.4487	71	0.1968	0.2335	0.2659	0.3038
33	0.2913	0.3440	0.3896	0.4421	72	0.1954	0.2319	0.2641	0.3017
34	0.2869	0.3388	0.3839	0.4357	73	0.1940	0.2303	0.2623	0.2997
35	0.2826	0.3338	0.3784	0.4296	74	0.1927	0.2287	0.2605	0.2977
36	0.2785	0.3291	0.3731	0.4238	75	0.1914	0.2272	0.2587	0.2957
37	0.2746	0.3246	0.3681	0.4182	76	0.1901	0.2257	0.2570	0.2938
38	0.2709	0.3202	0.3632	0.4128	77	0.1888	0.2242	0.2554	0.2919
39	0.2673	0.3160	0.3586	0.4076	78	0.1876	0.2227	0.2537	0.2900
40	0.2638	0.3120	0.3541	0.4026	79	0.1864	0.2213	0.2521	0.2882
41	0.2605	0.3081	0.3497	0.3978	80	0.1852	0.2199	0.2505	0.2864

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 9. Wartości krytyczne $\genfrac{}{}{0pt}{}{r(0.05;n,k)}{r(0.01;n,k)}$ współczynnika korelacji wielokrotnej

$n \setminus k$	3	4	5
4	0.9987		
	0.9999		
5	0.9747	0.9992	
	0.9950	1.0000	
6	0.9297	0.9830	0.9994
	0.9765	0.9967	1.0000
7	0.8811	0.9501	0.9873
	0.9487	0.9834	0.9975
8	0.8356	0.9120	0.9612
	0.9173	0.9623	0.9872
9	0.7947	0.8743	0.9299
	0.8858	0.9373	0.9701
10	0.7584	0.8391	0.8978
	0.8554	0.9112	0.9493
11	0.7260	0.8067	0.8668
	0.8269	0.8852	0.9269
12	0.6972	0.7771	0.8378
	0.8004	0.8603	0.9042
13	0.6714	0.7502	0.8108
	0.7758	0.8365	0.8820
14	0.6481	0.7257	0.7858
	0.7531	0.8141	0.8605
15	0.6269	0.7032	0.7628
	0.7320	0.7931	0.8401
16	0.6077	0.6826	0.7414
	0.7125	0.7734	0.8206
17	0.5901	0.6636	0.7216
	0.6943	0.7548	0.8021
18	0.5738	0.6461	0.7032
	0.6774	0.7375	0.7846
19	0.5589	0.6298	0.6861
	0.6616	0.7211	0.7681
20	0.5450	0.6147	0.6701
	0.6468	0.7057	0.7524
21	0.5321	0.6006	0.6551
	0.6329	0.6912	0.7376
22	0.5201	0.5873	0.6410
	0.6198	0.6775	0.7235
23	0.5088	0.5749	0.6278
	0.6075	0.6646	0.7102
24	0.4982	0.5633	0.6154
	0.5959	0.6523	0.6975
25	0.4883	0.5523	0.6036
	0.5849	0.6407	0.6854
26	0.4789	0.5419	0.5925
	0.5744	0.6296	0.6739
27	0.4700	0.5321	0.5820
	0.5645	0.6191	0.6630
28	0.4616	0.5228	0.5720
	0.5551	0.6091	0.6525
29	0.4537	0.5139	0.5624
	0.5462	0.5995	0.6425
30	0.4461	0.5055	0.5534
	0.5376	0.5904	0.6329
31	0.4389	0.4975	0.5447
	0.5294	0.5816	0.6238
32	0.4320	0.4899	0.5365
	0.5216	0.5732	0.6150
33	0.4255	0.4825	0.5286
	0.5142	0.5652	0.6065
34	0.4192	0.4755	0.5210
	0.5070	0.5575	0.5984
35	0.4132	0.4688	0.5138
	0.5001	0.5501	0.5906
36	0.4075	0.4624	0.5068
	0.4935	0.5430	0.5831
37	0.4020	0.4562	0.5001
	0.4871	0.5361	0.5759
38	0.3967	0.4503	0.4937
	0.4810	0.5295	0.5689
39	0.3916	0.4446	0.4875
	0.4751	0.5231	0.5622
40	0.3867	0.4391	0.4815
	0.4694	0.5170	0.5557
41	0.3819	0.4338	0.4758
	0.4639	0.5110	0.5494
42	0.3774	0.4287	0.4702
	0.4586	0.5053	0.5433
43	0.3730	0.4237	0.4649
	0.4535	0.4997	0.5374
44	0.3687	0.4190	0.4597
	0.4486	0.4944	0.5317
45	0.3646	0.4144	0.4547
	0.4437	0.4892	0.5262
46	0.3606	0.4099	0.4498
	0.4391	0.4841	0.5208
47	0.3568	0.4056	0.4451
	0.4346	0.4792	0.5156
48	0.3531	0.4014	0.4405
	0.4302	0.4745	0.5105
49	0.3495	0.3973	0.4361
	0.4260	0.4698	0.5056
50	0.3460	0.3934	0.4318
	0.4219	0.4654	0.5009
55	0.3299	0.3753	0.4121
	0.4029	0.4447	0.4789
60	0.3159	0.3594	0.3949
	0.3863	0.4266	0.4595
65	0.3035	0.3455	0.3796
	0.3715	0.4105	0.4423
70	0.2925	0.3330	0.3660
	0.3584	0.3961	0.4269
75	0.2826	0.3218	0.3537
	0.3465	0.3831	0.4130
80	0.2736	0.3116	0.3426
	0.3358	0.3713	0.4004
90	0.2580	0.2939	0.3232
	0.3169	0.3506	0.3783

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Tablica 10. Wartości krytyczne $\frac{V(0.05;r_1,r_2,c)}{V(0.01;r_1,r_2,c)}$ w teście Behrensa–Fishera

r_2	r_1						c					
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
5	2	2.5706	2.4792	2.3949	2.3656	2.4365	2.6371	2.9681	3.3910	3.8269	4.1682	4.3027
		4.0321	3.8051	3.5834	3.5023	3.7136	4.3376	5.4048	6.8080	8.2859	9.4592	9.9250
	3	2.5706	2.4793	2.3876	2.3227	2.3080	2.3580	2.4753	2.6480	2.8500	3.0429	3.1824
		4.0321	3.8029	3.5684	3.3996	3.3607	3.4968	3.8211	4.3065	4.8815	5.4365	5.8408
	4	2.5706	2.4797	2.3856	2.3091	2.2668	2.2696	2.3204	2.4136	2.5353	2.6647	2.7765
		4.0321	3.8032	3.5650	3.3701	3.2619	3.2692	3.4012	3.6450	3.9653	4.3074	4.6041
	5	2.5706	2.4800	2.3850	2.3029	2.2476	2.2281	2.2476	2.3029	2.3850	2.4800	2.5706
		4.0321	3.8039	3.5645	3.3576	3.2183	3.1693	3.2183	3.3576	3.5645	3.8039	4.0321
6	2	2.4469	2.3726	2.3152	2.3169	2.4171	2.6404	2.9840	3.4089	3.8391	4.1721	4.3027
		3.7074	3.5284	3.3803	3.3862	3.6824	4.3731	5.4766	6.8812	8.3340	9.4747	9.9250
	3	2.4469	2.3718	2.3037	2.2651	2.2747	2.3439	2.4736	2.6518	2.8540	3.0445	3.1824
		3.7074	3.5250	3.3557	3.2574	3.2829	3.4688	3.8244	4.3222	4.8953	5.4420	5.8408
	4	2.4469	2.3716	2.2998	2.2476	2.2281	2.2493	2.3126	2.4126	2.5365	2.6654	2.7765
		3.7074	3.5241	3.3476	3.2181	3.1693	3.2231	3.3859	3.6454	3.9696	4.3095	4.6041
	5	2.4469	2.3716	2.2980	2.2393	2.2060	2.2045	2.2368	2.2996	2.3848	2.4803	2.5706
		3.7074	3.5240	3.3441	3.2002	3.1181	3.1145	3.1947	3.3514	3.5649	3.8046	4.0321
	6	2.4469	2.3716	2.2970	2.2346	2.1932	2.1788	2.1932	2.2346	2.2970	2.3716	2.4469
		3.7074	3.5240	3.3424	3.1903	3.0897	3.0545	3.0897	3.1903	3.3424	3.5240	3.7074
7	2	2.3646	2.3028	2.2638	2.2865	2.4064	2.6448	2.9966	3.4223	3.8480	4.1750	4.3027
		3.4995	3.3534	3.2543	3.3179	3.6706	4.4057	5.5323	6.9360	8.3694	9.4861	9.9250
	3	2.3646	2.3013	2.2493	2.2281	2.2538	2.3356	2.4733	2.6550	2.8570	3.0457	3.1824
		3.4995	3.3492	3.2224	3.1693	3.2360	3.4541	3.8296	4.3351	4.9061	5.4462	5.8408
	4	2.3646	2.3008	2.2441	2.2081	2.2034	2.2365	2.3080	2.4123	2.5375	2.6659	2.7765
		3.4995	3.3477	3.2112	3.1233	3.1119	3.1952	3.3775	3.6468	3.9733	4.3113	4.6041
	5	2.3646	2.3005	2.2415	2.1983	2.1792	2.1894	2.2299	2.2976	2.3848	2.4805	2.5706
		3.4995	3.3471	3.2059	3.1018	3.0555	3.0805	3.1803	3.3479	3.5655	3.8053	4.0321
	6	2.3646	2.3004	2.2400	2.1926	2.1652	2.1623	2.1851	2.2316	2.2964	2.3716	2.4469
		3.4995	3.3468	3.2028	3.0895	3.0239	3.0169	3.0719	3.1842	3.3415	3.5242	3.7074
	7	2.3646	2.3003	2.2391	2.1889	2.1562	2.1448	2.1562	2.1889	2.2391	2.3003	2.3646
		3.4995	3.3466	3.2010	3.0817	3.0039	2.9768	3.0039	3.0817	3.2010	3.3466	3.4995
8	2	2.3060	2.2537	2.2281	2.2659	2.4000	2.6491	3.0066	3.4327	3.8548	4.1772	4.3027
		3.3554	3.2333	3.1693	3.2741	3.6674	4.4340	5.5766	6.9786	8.3966	9.4948	9.9250
	3	2.3060	2.2517	2.2114	2.2026	2.2396	2.3303	2.4736	2.6576	2.8594	3.0467	3.1824
		3.3554	3.2286	3.1315	3.1100	3.2054	3.4459	3.8352	4.3456	4.9146	5.4496	5.8408
	4	2.3060	2.2509	2.2052	2.1806	2.1864	2.2278	2.3050	2.4123	2.5383	2.6663	2.7765
		3.3554	3.2266	3.1181	3.0589	3.0733	3.1768	3.3726	3.6486	3.9763	4.3127	4.6041
	5	2.3060	2.2505	2.2020	2.1697	2.1607	2.1790	2.2253	2.2963	2.3849	2.4807	2.5706
		3.3554	3.2256	3.1114	3.0347	3.0131	3.0575	3.1707	3.3459	3.5663	3.8060	4.0321
	6	2.3060	2.2502	2.2001	2.1634	2.1457	2.1508	2.1794	2.2295	2.2961	2.3717	2.4469
		3.3554	3.2251	3.1075	3.0208	2.9791	2.9913	3.0598	3.1802	3.3410	3.5244	3.7074
	7	2.3060	2.2501	2.1989	2.1592	2.1360	2.1326	2.1499	2.1864	2.2384	2.3003	2.3646
		3.3554	3.2247	3.1050	3.0119	2.9575	2.9495	2.9901	3.0764	3.1997	3.3466	3.4995
	8	2.3060	2.2500	2.1980	2.1563	2.1292	2.1199	2.1292	2.1563	2.1980	2.2500	2.3060
		3.3554	3.2245	3.1033	3.0057	2.9426	2.9208	2.9426	3.0057	3.1033	3.2245	3.3554

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

r_2	r_1	c: 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
9	2	2.2622	2.2172	2.2020	2.2512	2.3960	2.6532	3.0148	3.4410	3.8602	4.1790	4.3027
		3.2498	3.1461	3.1084	3.2444	3.6681	4.4584	5.6126	7.0125	8.4181	9.5016	9.9250
	3	2.2622	2.2149	2.1834	2.1840	2.2294	2.3267	2.4741	2.6598	2.8613	3.0474	3.1824
		3.2498	3.1408	3.0659	3.0675	3.1842	3.4411	3.8406	4.3544	4.9216	5.4524	5.8408
	4	2.2622	2.2139	2.1765	2.1605	2.1740	2.2216	2.3029	2.4124	2.5390	2.6667	2.7765
		3.2498	3.1385	3.0507	3.0125	3.0457	3.1640	3.3696	3.6504	3.9790	4.3140	4.6041
	5	2.2622	2.2133	2.1729	2.1488	2.1471	2.1714	2.2219	2.2954	2.3850	2.4809	2.5706
		3.2498	3.1373	3.0430	2.9863	2.9824	3.0411	3.1641	3.3447	3.5670	3.8065	4.0321
	6	2.2622	2.2130	2.1707	2.1419	2.1315	2.1424	2.1753	2.2280	2.2958	2.3717	2.4469
		3.2498	3.1365	3.0385	2.9711	2.9467	2.9728	3.0511	3.1774	3.3408	3.5246	3.7074
	7	2.2622	2.2128	2.1693	2.1373	2.1212	2.1237	2.1452	2.1845	2.2379	2.3003	2.3646
		3.2498	3.1361	3.0355	2.9613	2.9239	2.9296	2.9802	3.0726	3.1989	3.3466	3.4995
	8	2.2622	2.2126	2.1683	2.1341	2.1141	2.1106	2.1243	2.1541	2.1974	2.2499	2.3060
		3.2498	3.1358	3.0334	2.9545	2.9081	2.9000	2.9317	3.0012	3.1020	3.2244	3.3554
	9	2.2622	2.2125	2.1675	2.1318	2.1088	2.1009	2.1088	2.1318	2.1675	2.2125	2.2622
		3.2498	3.1355	3.0318	2.9495	2.8966	2.8784	2.8966	2.9495	3.0318	3.1355	3.2498
10	2	2.2281	2.1892	2.1820	2.2402	2.3933	2.6568	3.0216	3.4477	3.8645	4.1803	4.3027
		3.1693	3.0800	3.0629	3.2232	3.6708	4.4794	5.6424	7.0402	8.4355	9.5072	9.9250
	3	2.2281	2.1865	2.1620	2.1698	2.2218	2.3242	2.4748	2.6617	2.8629	3.0481	3.1824
		3.1693	3.0741	3.0163	3.0359	3.1688	3.4384	3.8455	4.3618	4.9273	5.4546	5.8408
	4	2.2281	2.1853	2.1544	2.1451	2.1646	2.2169	2.3015	2.4126	2.5396	2.6670	2.7765
		3.1693	3.0716	2.9997	2.9775	3.0251	3.1546	3.3677	3.6521	3.9812	4.3150	4.6041
	5	2.2281	2.1847	2.1505	2.1327	2.1368	2.1657	2.2194	2.2948	2.3851	2.4810	2.5706
		3.1693	3.0702	2.9913	2.9497	2.9594	3.0288	3.1592	3.3440	3.5677	3.8070	4.0321
	6	2.2281	2.1843	2.1481	2.1254	2.1205	2.1361	2.1722	2.2269	2.2957	2.3718	2.4469
		3.1693	3.0693	2.9862	2.9336	2.9223	2.9588	3.0446	3.1753	3.3407	3.5248	3.7074
	7	2.2281	2.1841	2.1465	2.1206	2.1099	2.1168	2.1417	2.1831	2.2376	2.3002	2.3646
		3.1693	3.0688	2.9828	2.9231	2.8985	2.9146	2.9727	3.0697	3.1983	3.3466	3.4995
	8	2.2281	2.1839	2.1454	2.1172	2.1025	2.1034	2.1204	2.1525	2.1969	2.2498	2.3060
		3.1693	3.0684	2.9805	2.9158	2.8821	2.8842	2.9235	2.9978	3.1011	3.2243	3.3554
	9	2.2281	2.1837	2.1446	2.1146	2.0970	2.0935	2.1048	2.1299	2.1669	2.2124	2.2622
		3.1693	3.0681	2.9787	2.9104	2.8701	2.8621	2.8879	2.9457	3.0307	3.1354	3.2498
	10	2.2281	2.1836	2.1439	2.1127	2.0928	2.0860	2.0928	2.1127	2.1439	2.1836	2.2281
		3.1693	3.0678	2.9774	2.9062	2.8609	2.8453	2.8609	2.9062	2.9774	3.0678	3.1693
15	2	2.1315	2.1105	2.1272	2.2114	2.3885	2.6698	3.0433	3.4686	3.8778	4.1846	4.3027
		2.9467	2.8998	2.9426	3.1734	3.6898	4.5506	5.7371	7.1262	8.4891	9.5240	9.9250
	3	2.1315	2.1068	2.1026	2.1312	2.2017	2.3183	2.4777	2.6678	2.8679	3.0500	3.1824
		2.9467	2.8919	2.8824	2.9522	3.1308	3.4357	3.8638	4.3862	4.9459	5.4620	5.8408
	4	2.1315	2.1052	2.0931	2.1027	2.1390	2.2046	2.2980	2.4137	2.5416	2.6680	2.7765
		2.9467	2.8885	2.8612	2.8836	2.9708	3.1314	3.3649	3.6589	3.9888	4.3186	4.6041
	5	2.1315	2.1043	2.0882	2.0883	2.1084	2.1501	2.2128	2.2933	2.3857	2.4815	2.5706
		2.9467	2.8865	2.8504	2.8508	2.8976	2.9964	3.1472	3.3432	3.5705	3.8088	4.0321
	6	2.1315	2.1037	2.0851	2.0798	2.0904	2.1185	2.1638	2.2240	2.2953	2.3720	2.4469
		2.9467	2.8853	2.8439	2.8317	2.8562	2.9214	3.0275	3.1704	3.3409	3.5256	3.7074
	7	2.1315	2.1033	2.0831	2.0741	2.0787	2.0980	2.1320	2.1792	2.2367	2.3002	2.3646
		2.9467	2.8845	2.8395	2.8193	2.8297	2.8740	2.9525	3.0622	3.1969	3.3468	3.4995
	8	2.1315	2.1030	2.0816	2.0701	2.0704	2.0836	2.1099	2.1479	2.1956	2.2497	2.3060
		2.9467	2.8839	2.8363	2.8105	2.8113	2.8413	2.9011	2.9885	3.0987	3.2242	3.3554
	9	2.1315	2.1028	2.0805	2.0671	2.0643	2.0731	2.0935	2.1249	2.1653	2.2122	2.2622
		2.9467	2.8834	2.8340	2.8041	2.7978	2.8176	2.8639	2.9352	3.0275	3.1350	3.2498

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

r_2	r_1						c					
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	10	2.1315	2.1026	2.0796	2.0647	2.0595	2.0649	2.0810	2.1073	2.1421	2.1833	2.2281
		2.9467	2.8831	2.8322	2.7991	2.7874	2.7995	2.8358	2.8949	2.9737	3.0673	3.1693
	11	2.1315	2.1025	2.0790	2.0629	2.0558	2.0585	2.0712	2.0933	2.1237	2.1604	2.2010
		2.9467	2.8828	2.8307	2.7951	2.7793	2.7853	2.8137	2.8633	2.9315	3.0142	3.1058
	12	2.1315	2.1024	2.0784	2.0614	2.0528	2.0533	2.0632	2.0820	2.1087	2.1417	2.1788
		2.9467	2.8825	2.8295	2.7919	2.7727	2.7739	2.7960	2.8380	2.8977	2.9714	3.0545
	13	2.1315	2.1023	2.0779	2.0601	2.0502	2.0490	2.0566	2.0727	2.0964	2.1263	2.1604
		2.9467	2.8823	2.8285	2.7892	2.7673	2.7645	2.7814	2.8172	2.8699	2.9363	3.0123
	14	2.1315	2.1022	2.0775	2.0591	2.0481	2.0454	2.0510	2.0649	2.0860	2.1132	2.1448
		2.9467	2.8822	2.8277	2.7870	2.7628	2.7567	2.7692	2.7998	2.8467	2.9069	2.9768
	15	2.1315	2.1021	2.0772	2.0582	2.0463	2.0423	2.0463	2.0582	2.0772	2.1021	2.1315
		2.9467	2.8820	2.8270	2.7851	2.7589	2.7500	2.7589	2.7851	2.8270	2.8820	2.9467
20	2	2.0860	2.0740	2.1024	2.1993	2.3879	2.6774	3.0548	3.4795	3.8846	4.1867	4.3027
		2.8453	2.8191	2.8911	3.1560	3.7053	4.5906	5.7872	7.1707	8.5165	9.5326	9.9250
	3	2.0860	2.0698	2.0754	2.1139	2.1932	2.3164	2.4798	2.6712	2.8706	3.0511	3.1824
		2.8453	2.8099	2.8232	2.9165	3.1163	3.4374	3.8750	4.3996	4.9558	5.4660	5.8408
	4	2.0860	2.0679	2.0650	2.0835	2.1277	2.1993	2.2968	2.4145	2.5427	2.6685	2.7765
		2.8453	2.8060	2.7995	2.8424	2.9476	3.1224	3.3651	3.6632	3.9931	4.3206	4.6041
	5	2.0860	2.0669	2.0595	2.0681	2.0956	2.1432	2.2101	2.2929	2.3860	2.4818	2.5706
		2.8453	2.8038	2.7874	2.8070	2.8706	2.9828	3.1426	3.3435	3.5723	3.8099	4.0321
	6	2.0860	2.0662	2.0562	2.0589	2.0768	2.1107	2.1600	2.2228	2.2952	2.3721	2.4469
		2.8453	2.8024	2.7801	2.7864	2.8270	2.9051	3.0204	3.1686	3.3413	3.5262	3.7074
	7	2.0860	2.0658	2.0539	2.0529	2.0644	2.0895	2.1277	2.1775	2.2363	2.3003	2.3646
		2.8453	2.8014	2.7753	2.7730	2.7991	2.8560	2.9437	3.0591	3.1965	3.3471	3.4995
	8	2.0860	2.0654	2.0523	2.0485	2.0558	2.0747	2.1051	2.1459	2.1951	2.2497	2.3060
		2.8453	2.8007	2.7717	2.7635	2.7797	2.8222	2.8912	2.9845	3.0977	3.2242	3.3554
	9	2.0860	2.0652	2.0510	2.0453	2.0493	2.0637	2.0884	2.1226	2.1646	2.2121	2.2622
		2.8453	2.8002	2.7691	2.7565	2.7654	2.7976	2.8532	2.9306	3.0262	3.1349	3.2498
	10	2.0860	2.0650	2.0500	2.0428	2.0443	2.0553	2.0757	2.1048	2.1412	2.1832	2.2281
		2.8453	2.7997	2.7671	2.7511	2.7545	2.7789	2.8244	2.8897	2.9720	3.0670	3.1693
	11	2.0860	2.0648	2.0493	2.0408	2.0404	2.0487	2.0656	2.0907	2.1227	2.1602	2.2010
		2.8453	2.7994	2.7654	2.7468	2.7460	2.7643	2.8019	2.8578	2.9296	3.0139	3.1058
	12	2.0860	2.0647	2.0486	2.0391	2.0372	2.0433	2.0574	2.0792	2.1077	2.1415	2.1788
		2.8453	2.7991	2.7641	2.7433	2.7390	2.7524	2.7838	2.8322	2.8956	2.9711	3.0545
	13	2.0860	2.0646	2.0481	2.0378	2.0345	2.0388	2.0507	2.0698	2.0953	2.1261	2.1604
		2.8453	2.7989	2.7630	2.7405	2.7333	2.7427	2.7689	2.8111	2.8677	2.9359	3.0123
	14	2.0860	2.0645	2.0477	2.0366	2.0323	2.0351	2.0450	2.0619	2.0849	2.1130	2.1448
		2.8453	2.7987	2.7620	2.7380	2.7285	2.7346	2.7565	2.7935	2.8443	2.9065	2.9768
	15	2.0860	2.0644	2.0473	2.0357	2.0304	2.0319	2.0402	2.0551	2.0760	2.1019	2.1315
		2.8453	2.7985	2.7612	2.7360	2.7244	2.7277	2.7459	2.7787	2.8246	2.8815	2.9467
	16	2.0860	2.0643	2.0469	2.0348	2.0287	2.0291	2.0360	2.0493	2.0683	2.0923	2.1199
		2.8453	2.7983	2.7605	2.7342	2.7209	2.7217	2.7369	2.7659	2.8076	2.8601	2.9208
	17	2.0860	2.0643	2.0466	2.0341	2.0273	2.0267	2.0324	2.0442	2.0617	2.0839	2.1098
		2.8453	2.7982	2.7599	2.7326	2.7178	2.7165	2.7290	2.7548	2.7929	2.8415	2.8982
	18	2.0860	2.0642	2.0464	2.0334	2.0260	2.0246	2.0292	2.0398	2.0558	2.0765	2.1009
		2.8453	2.7981	2.7594	2.7312	2.7152	2.7120	2.7221	2.7451	2.7800	2.8251	2.8784
	19	2.0860	2.0642	2.0462	2.0329	2.0249	2.0227	2.0264	2.0359	2.0506	2.0700	2.0930
		2.8453	2.7980	2.7589	2.7300	2.7128	2.7080	2.7160	2.7365	2.7686	2.8107	2.8609
	20	2.0860	2.0641	2.0460	2.0323	2.0239	2.0211	2.0239	2.0323	2.0460	2.0641	2.0860
		2.8453	2.7979	2.7585	2.7289	2.7106	2.7045	2.7106	2.7289	2.7585	2.7979	2.8453

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Często używane wzorki

Średnia arytmetyczna

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

Suma kwadratów odchyleń od średniej

$$\begin{aligned}\mathrm{var}X &= \sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \sum_{i=1}^{n} X_i^2 - n(\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\end{aligned}$$

Wariancja próbkowa

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{\mathrm{var}X}{n-1}$$

Suma iloczynów odchyleń

$$\begin{aligned}\mathrm{cov}(X,Y) &= \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\end{aligned}$$

Kowariancja próbkowa

$$\mathrm{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{\mathrm{cov}(X,Y)}{n-1}$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Wybrane pojęcia

Statystyka: nauka poświęcona metodom badania (analizowania) zjawisk masowych; polega na systematyzowaniu obserwowanych cech ilościowych i jakościowych oraz przedstawianiu wyników w postaci zestawień tabelarycznych, wykresów, itp.; posługuje się rachunkiem prawdopodobieństwa. *Encyklopedia Popularna PWN, Warszawa 1982*

Statystyka matematyczna: dział matematyki stosowanej oparty na rachunku prawdopodobieństwa; zajmuje się badaniem zbiorów na podstawie znajomości własności ich części. *Encyklopedia Popularna PWN, Warszawa 1982*

Populacja: zbiór obiektów z wyróżnioną cechą (cechami).

Próba: wybrana część populacji podlegająca badaniu.

Cecha: wielkość losowa charakteryzująca obiekty danej populacji.

Cecha jakościowa: cecha przyjmująca wartości nie będące liczbami (np. *kolor*, *płeć*, *smakowitość*)

Cecha (ilościowa) skokowa: cecha przyjmująca pewne wartości liczbowe i nie przyjmująca wartości pośrednich (np. *ilość bakterii*, *ilość pracowników*, *ilość pasażerów*). Cechy te nazywane są również **dyskretnymi**.

Cecha (ilościowa) ciągła: Cecha przyjmująca wartości z pewnego przedziału liczbowego (np. *wzrost*, *waga*, *plon*)

Estymacja parametrów rozkładu cechy Estymujemy parametr θ rozkładu cechy X na podstawie próby $X_1, X_2, \ldots, X_n$

Estymator (punktowy) jest funkcją próby:

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

w „rozsądny" sposób przybliżający wartość parametru θ.

Nieobciążoność. Jeżeli średnia wartość oceny $\hat{\theta}$ jest równa wartości parametru θ, to ocenę $\hat{\theta}$ nazywamy nieobciążoną:

$$(\forall\theta)\ E_\theta\hat{\theta} = \theta.$$

Minimalna wariancja. Nieobciążony estymator $\hat{\theta}$ parametru θ jest estymatorem o minimlanej wariancji, jeżeli dla wszystkich innych estymatorów nieobciążonych $\tilde{\theta}$ parametru θ zachodzi

$$(\forall\theta)\ D^2_\theta\hat{\theta} \le D^2_\theta\tilde{\theta}$$

z ostrą nierównością dla przynajmniej jednej wartości θ.

Minimalny błąd średniokwadratowy. Estymator $\hat{\theta}$ parametru θ jest estymatorem o minimalnym błędzie średniokwadratowym, jeżeli dla wszystkich innych estymatorów $\tilde{\theta}$ parametru θ zachodzi

$$(\forall\theta)\ E_\theta(\hat{\theta} - \theta)^2 \le E_\theta(\tilde{\theta} - \theta)^2$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

z ostrą nierównością dla przynajmniej jednej wartości θ.

Przedział ufności (estymator przedziałowy) jest przedziałem o końcach zależnych od próby, który ze z góry zadanym prawdopodobieństwem $1-\alpha$ pokrywa nieznaną wartość parametru θ

$$P\{\theta \in (\underline{\theta}(X_1,\ldots,X_n), \overline{\theta}(X_1,\ldots,X_n))\} = 1-\alpha.$$

Poziomem ufności nazywamy ustalone prawdopodobieństwo $1-\alpha$ pokrycia przez przedział ufności prawdziwej wartości parametru θ.

Weryfikacja hipotez statystycznych

Hipotezą statystyczną nazywamy dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa cechy (oznaczenie H_0).

Testem hipotezy statystycznej nazywamy postępowanie mające na celu odrzucenie lub nie odrzucenie hipotezy statystycznej.

Statystyką testową nazywamy funkcję próby na podstawie której wnioskuje się o odrzuceniu lub nie hipotezy statystycznej.

Błędem I rodzaju nazywamy błąd wnioskowania polegający na odrzuceniu hipotezy, gdy w rzeczywistości jest ona prawdziwa.

Błędem II rodzaju nazywamy błąd wnioskowania polegający na nieodrzuceniu hipotezy, gdy w rzeczywistości jest ona fałszywa.

Poziomem istotności nazywamy dowolną liczbę z przedziału $(0,1)$ określającą prawdopodobieństwo popełnienia błędu I rodzaju.

Mocą testu nazywamy prawdopodobieństwo odrzucenia hipotezy zerowej, gdy w rzeczywistości nie jest ona prawdziwa.

Współczynnik korelacji jest miernikiem liniowej zależności między cechami X Y.

Własności współczynnika korelacji

1. Współczynnik korelacji jest liczbą niemianowaną.

2. Współczynnik korelacji przyjmuje wartości z przedziału $\langle -1, 1\rangle$.

3. Jeżeli współczynnik korelacji jest dodatni, to zależność ma charakter rosnący.

4. Jeżeli współczynnik korelacji jest ujemny, to zależność ma charakter malejący.

5. Jeżeli współczynnik korelacji jest zerowy to bez względu na wartości przyjmowane przez jedną z cech, średnie wartości drugiej cechy są takie same.

6. Jeżeli współczynnik korealcji jest równy ± 1, to zależność ma charakter ściśle liniowy.

7. Jeżeli para (X,Y) ma dwuwymiarowy rozkład normalny, to zerowoć współczynnika korelacji jest równoważna niezależności cech X, Y.

gfgfgf

Analiza regresji. Obserwujemy cechę Y oraz zmienne deterministyczne $X_1,\ldots,X_k$.

Funkcją regresji nazywamy taką funkcję f, że

$$E(Y|X_1 = x_1,\ldots,X_k = x_k) = f(x_1,\ldots,x_k)$$

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Model liniowy Jeżeli funkcja regresji ma postać

$$f(x_1,\ldots,x_k)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k,$$

to mówimy o liniowym modelu regresji.

Metoda najmniejszych kwadratów. Estymatory $\hat{\beta}_i$ parametrów funkcji regresji są uzyskane metodą najmniejszych kwadratów, jeżeli

$$\sum_{i=1}^{n}(Y_i-(\hat{\beta}_0+\hat{\beta}_1x_{1i}+\cdots+\hat{\beta}_kx_{ki}))^2=\min_{\beta_0,\beta_1,\ldots,\beta_k}\sum_{i=1}^{n}(Y_i-(\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}))^2$$

Resztowa suma kwadratów reprezentuje rozrzut obserwacji cechy Y wokół dopasowanej funkcji regresji

$$\mathrm{var}R=\sum_{i=1}^{n}(Y_i-(\hat{\beta}_0+\hat{\beta}_1x_{1i}+\cdots+\hat{\beta}_kx_{ki}))^2.$$

Wariancja resztowa Jest to uśredniona resztowa suma kwadratów.

$$S^2_{y\cdot x}=\frac{\mathrm{var}R}{n-2}$$

Współczynnik determinacji jest miarą dopasowania funkcji regresji

$$D=R^2\cdot 100\%,\ \text{gdzie } R^2=\frac{\sum((\hat{\beta}_0+\hat{\beta}_1x_{1i}+\cdots+\hat{\beta}_kx_{ki})-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}$$

Wielkość R nazywana jest współczynnikiem korelacji wielokrotnej. Jeżeli $k=2$, to R jest zwykłym współczynnikiem korelacji.

Obszar ufności. jest takim obszarem o brzegach zależnych od próby, który ze z góry zadanym prawdopodobieństwem $1-\alpha$ pokrywa nieznaną funkcję regresji f. Na podstawie obszaru ufności wnioskujemy o **wartościach średnich** cechy Y jednocześnie dla wielu wybranych wartości zmiennych $X_1,\ldots,X_k$.

Obszar predykcji. Na podstawie obszaru ufności wnioskujemy o **wartościach** cechy Y jednocześnie dla wielu wybranych wartości cechy X.

Analiza szeregów czasowych

Szeregiem czasowym nazywamy ciąg y_t, $t=0,1,\ldots,n-1$ obserwacji zbieranych w czasie.

Trend (tendencja rozwojowa): funkcja opisująca generalny przebieg zjawiska, zmiany średniego poziomu zjawiska w czasie.

Wahania okresowe (sezonowe): powtarzające się regularnie zmiany poziomu zjawiska.

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk

Przydatne funkcje w arkuszu Excel

`ROZKŁ.DWUM(liczba_s;próby;prawd_s;skumulowany)`	rozkład dwumianowy
`ROZKŁ.NORMALNY(x;średnia;odch_std;skumulowany)`	rozkład normalny
`ROZKŁ.NORMALNY.S(x)`	dystrybuanta rozkładu $N(0,1)$
`ROZKŁAD.CHI(x;stopnie_swobody)`	dystrybuanta rozkładu chi-kwadrat
`ROZKŁAD.CHI.ODW(prawd;stopnie_swobody)`	wartość krytyczna rozkładu chi-kwadrat
`ROZKŁ.T.DS(x;stopnie_swobody)`	dwustronny ogon rozkładu t
`ROZKŁ.T.ODWR.DS(prawd;stopnie_swobody)`	wartość krytyczna rozkładu t
`ROZKŁAD.BETA.ODW(prawd;a,b)`	kwantyl rozkładu Beta
`ŚREDNIA(obszar)`	średnia arytmetyczna $\bar{X}$
`WARIANCJA.PRÓBKI(obszar)`	wariancja S^2
`ODCH.STANDARD.PRÓBKI(obszar)`	odchylenie standardowe S

Copyright © Marta Zalewska & Wojciech Zieliński & Stanisław Jaworski & Konrad Furmańczyk