

Test chi-kwadrat

0. Ogólnie o teście. Test, zwany χ^2 („chi-kwadrat”) należy do kategorii testów nieparametrycznych, to znaczy, że może być stosowany do każdej zmiennej losowej, niezależnie od jej rozkładu. W szczególności nie musi to być rozkład normalny. Test ten ma dwie, dość podobne, odmiany: test niezależności i test zgodności. W każdej z nich przyjmujemy hipotezę zerową (właśnie niezależności albo zgodności z ustalonym rozkładem) i na ustalonym poziomie istotności sprawdzamy, czy są podstawy do jej odrzucenia, czy też podstaw takich nie ma. Odbywa się to – wzorem wielu innych testów – przez porównanie pewnej obliczonej wielkości z tzw. parametrem krytycznym dla danego rozkładu. Parametr ten wyznaczamy z tablic; można posłużyć się też niektórymi programami lub arkuszami kalkulacyjnymi, np. Excel.

1. **Test niezależności** można najlepiej zrozumieć na nieco sztucznym przykładzie. Przed wyborami prezydenckimi badamy, czy poparcie dla obu głównych kandydatów jest zależne, czy niezależne od płci. Innymi słowy, czy (statystycznie) mężczyźni mają inne preferencje wyborcze niż kobiety. Wybieramy stosowną próbkę n osób, w tym m mężczyzn i k kobiet. Układamy tabelę (zwaną tablicą dwudzielną):

Wartości obserwowane	Mężczyźni	Kobiety	Razem
Poparcie dla A	a	b	$a + b$
Poparcie dla B	c	d	$c + d$
Ogółem	$m = a + c$	$k = b + d$	$n = m + k = a + b + c + d$

Rozumujemy teraz tak. Poparcie dla kandydata A wynosi $\frac{a+b}{n}$, poparcie dla B wynosi $\frac{c+d}{n}$. Gdyby w obydwu grupach (tj. dla mężczyzn i dla kobiet) poparcie było takie same, to tablica wyglądałaby tak:

Wartości oczekiwane (teoretyczne)	Mężczyźni	Kobiety	Razem
Poparcie teoretyczne dla A	$\frac{a+b}{n} m$	$\frac{a+b}{n} k$	$a + b$
Poparcie teoretyczne dla B	$\frac{c+d}{n} m$	$\frac{c+d}{n} k$	$c + d$
Ogółem	m	k	$n = m + k = a + b + c + d$

Wartości z tej tabeli nazywamy wielkościami oczekiwanymi (bądź teoretycznymi). Mamy porównać te tabelki i odpowiedzieć na pytanie, jak bardzo się różnią. Tworzymy wielkość, zwaną właśnie statystyką chi-kwadrat:

$$\chi^2 = \sum_{i=1}^n \frac{(\text{obserw} - \text{oczekiw})^2}{\text{oczekiw}}$$

W opisywanej sytuacji będziemy mieli

$$\chi^2 = \frac{(a - \frac{a+b}{n}m)^2}{\frac{a+b}{n}m} + \frac{(b - \frac{a+b}{n}k)^2}{\frac{a+b}{n}k} + \frac{(c - \frac{c+d}{n}m)^2}{\frac{c+d}{n}m} + \frac{(d - \frac{c+d}{n}k)^2}{\frac{c+d}{n}k}$$

Zauważmy, że im mniejsze różnice między wielkościami teoretycznymi a obserwowanymi, tym mniejsza jest wartość wyliczonej statystyki. W szczególności, gdy zgodność jest idealna, to wszystkie różnice są równe zero, a zatem $\chi^2 = 0$. Rozpatrzmy to najpierw na bardziej konkretnych przykładach. Możemy powiedzieć, że im większa jest wartość χ^2 , tym bardziej prawdopodobne jest, że dane empiryczne (obserwowane) nie pasują do teoretycznych (oczekiwanych). Gałąź matematyki zwana statystyką matematyczną nadaje temu pewną miarę, to znaczy pozwala zmierzyć to odchylenie. Zrozumiemy to na serii przykładów. W każdym z nich mamy dwie grupy (M i K, na przykład mężczyźni i kobiety, albo mieszkańcy Mławy i Kalisza, albo zwolennicy klubów piłkarskich Milan Milanówek i Korona Kraków, albo ci, którzy piją Mepsi-colę albo Koka-kolę,.....). Badamy, czy pewne dwie cechy (oznaczone umownie A i B, na przykład grupy krwi A, B, albo inteligencja: A wysoka, B niska, albo.... przykłady można kontynuować) występują równie często wśród M i K. W poniższych przykładach rozpatrujemy próbki 200-osobowe, 100 M i 100 K. Próbki te nie muszą być równej wielkości; nie powinny jednak różnić się znacznie.

Musimy teraz

- 1) Wybrać poziom istotności, to znaczy dopuszczalny margines błędu, czyli ustalić prawdopodobieństwo, z jakim wygłaszane przez nas twierdzenia będą prawdziwe.
- 2) uwzględnić liczbę stopni swobody df . Ogólnie, jest ona równa liczbie niezależnych parametrów. W przypadku tablicy dwudzielnej mamy $df = 1$.
- 3) Wyznaczyć wartość krytyczną testu.

Obserw.	M	k	Oczek.	m	k	
A	60	40	100	50	50	chi-2 8
B	40	60	100	50	50	
	100	100	200	100	100	

Obserw.	m	k	Oczek.	m	k	
A	55	45	100	50	50	chi2 2
B	45	55	100	50	50	
	100	100	200	100	100	

Obserw.	m	k	Oczek.	m	k	
A	60	40	100	50	50	chi-2 8
B	40	60	100	50	50	
	100	100	200	100	100	

Obszew.	m	k		Oczek.	m	k	
A	550	450	1000	500	500	1000	chi-2
B	450	550	1000	500	500	1000	
	1000	1000	2000	1000	1000	2000	20

Obszew.	m	k		Oczek.	m	k	
A	51	49	100	50	50	100	chi-2
B	49	51	100	50	50	100	
	100	100	200	100	100	200	0,08

Obszew.	m	k		Oczek.	m	k	
A	510	490	1000	500	500	1000	chi-2
B	490	510	1000	500	500	1000	
	1000	1000	2000	1000	1000	2000	0,8

Obszew.	m	k		Oczek.	m	k	
A	5100	4900	10000	5000	5000	10000	chi-2
B	4900	5100	10000	5000	5000	10000	
	10000	10000	20000	10000	10000	20000	8

Z tablic rozkładu chi-kwadrat odczytujemy, że dla jednego stopnia swobody wartość krytyczna wynosi:

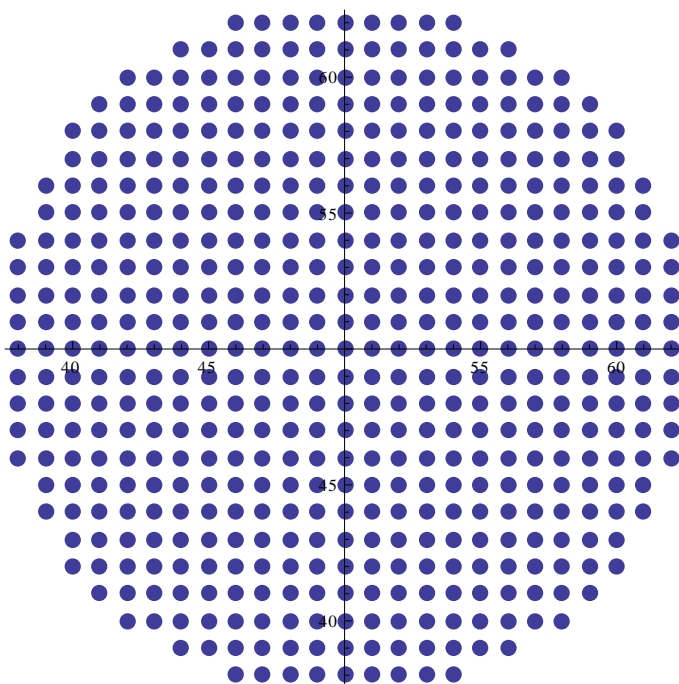
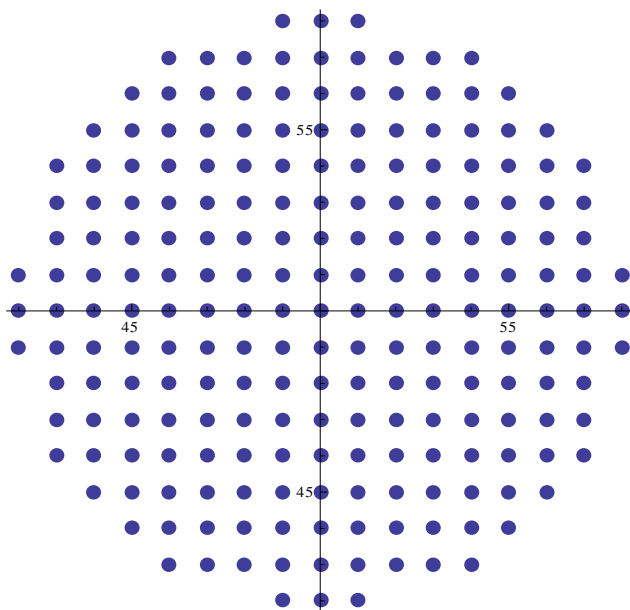
poziom	0,2	0,10	0,05	0,02	0,01
χ krytyczne	1,642	2,706	3,841	5,412	6,635

Jeżeli zatem pracujemy na poziomie istotności 0,1, to odrzucamy hipotezę o zgodności rozkładu empirycznego z teoretycznym, jeżeli wyliczona wielkość chi-kwadrat jest większa od 0,10.

Zadanie. Rozpatrujemy sytuację jak wyżej, to znaczy próbkę 100 M i 100 K badaną pod kątem dwóch cech A i B. Ustalmy poziom istotności $p=0,10$. Przyjmijmy, że w grupie M rozkład cech A i B jest $50+m$ do $50-m$, w grupie K rozkład jest $50+n$ do $50-n$. Nie mamy wątpliwości, że gdy m, n są małymi liczbami, zgodność rozkładu teoretycznego i doświadczalnego jest bardzo dobra. Dla dużych m , zgodność jest niewielka, poniżej postulowanego prawdopodobieństwa. Wyliczmy, jaki warunek mają spełniać m, n , żeby można było jeszcze mówić o zgodności. Powinno być

$$\chi^2 = \frac{m^2}{50} + \frac{m^2}{50} + \frac{n^2}{50} + \frac{n^2}{50} = \frac{m^2+n^2}{25} < 2,706,$$

czyli $m^2 + n^2 < 67,65$. Widzimy to na wykresie



Dla poziomu istotności 0,01 mamy znacznie szersze spektrum wartości dopuszczalnych (rys. powyżej).

2. Test zgodności z rozkładem. Zaczniemy od zgodności z rozkładem jednostajnym. Stosowny przykład wyjaśni sprawę lepiej niż teoria. Chcemy się przekonać, czy każdego dnia tygodnia roboczego w sklepie sprzedaje się tyle samo pieczywa. Obserwacje dają:

Dzień	Pon	Wtorek	Środa	Czwartek	Piątek
Sprzedanych kg	175	210	225	186	204
Wielkości teoret.	200	200	200	200	200

Obliczamy wielkość chi-kwadrat:

$$\chi^2 = \sum_{i=1}^n \frac{(\text{obserw} - \text{oczekiw})^2}{\text{oczekiw}}$$

W naszym przypadku wynosi ona $\frac{625+100+625+196+16}{200} = 7,81$. Oto fragment tablic rozkładu chi-kwadrat

df	P	0,2	0,15	0,1	0,05	0,025	0,01	0,001		
						0,02	0,005			
1		1.64237	2.07225	2.70554	3.84146	5.02389	5.41189	6.63490	7.87944	10.8276
2		3.21888	3.79424	4.60517	5.99146	7.37776	7.82405	9.21034	10.5966	13.8155
3		4.64163	5.31705	6.25139	7.81473	9.34840	9.83741	11.3449	12.8382	16.2663
4	→	5.98862	6.74488	7.77944	9.48773	11.1433	11.6678	13.2767	14.8603	18.4668
5		7.28928	8.11520	9.23636	11.0705	12.8325	13.3882	15.0863	16.7496	20.5150
6		8.55806	9.44610	10.6446	12.5916	14.4494	15.0332	16.8119	18.5476	22.4578
7		9.80325	10.7479	12.0170	14.0671	16.0128	16.6224	18.4753	20.2777	24.3219
8		11.0301	12.0271	13.3616	15.5073	17.5345	18.1682	20.0902	21.9550	26.1245
9		12.2421	13.2880	14.6837	16.9190	19.0228	19.6790	21.6660	23.5893	27.8772
10		13.4420	14.5339	15.9872	18.3070	20.4832	21.1608	23.2093	25.1882	29.5883
11		14.6314	15.7671	17.2750	19.6751	21.9201	22.6179	24.7250	26.7569	31.2641
12		15.8120	16.9893	18.5493	21.0261	23.3367	24.0540	26.2170	28.2995	32.9095
13		16.9848	18.2020	19.8119	22.3620	24.7356	25.4715	27.6882	29.8195	34.5282
14		18.1508	19.4062	21.0641	23.6848	26.1189	26.8728	29.1412	31.3194	36.1232
15		19.3107	20.6030	22.3071	24.9958	27.4884	28.2595	30.5779	32.8013	37.6973

Zatem na poziomie istotności 0,10 mamy prawo odrzucić hipotezę zerową, ale na poziomie 0,05 nie ma podstaw do jej odrzucenia.

Zadanie 1 (z książki Stanley Gregory, „Metody statystyki w geografii”, Warszawa, 1976). Zbadać, czy prawdziwa jest hipoteza, że nie ma istotnej różnicy między rodzajem terenu a typem zabudowania.

Rodzaj terenu	Liczba gospodarstw	Procent, jaki ten rodzaj terenu zajmuje na całym obszarze
Obszar zalewowy	10	10
Taras	100	35
Strome zbocze	2	10
Płaskowyż wapienny	38	25
Płaskowyż piaskowcowy	50	20

Zadanie 2. Badano, czy jest różnica między kobietami a mężczyznami, jeśli chodzi o wybór konkretnej marki butów. Wybrano losowo 100 osób. Wynik badania są w tabeli:

	M	K	Razem
Zwraca uwagę na markę	36	34	70
Nie zwraca uwagi	24	6	30
Ogółem	60	40	100

Sprawdź hipotezę, że nie ma różnic między M i K jeśli chodzi o przywiązanie do konkretnej marki butów. (Wynik obliczeń: $\chi^2 = 7,143$).

Zadanie 3. Sieć handlowa „Ogród ponad wszystko” otrzymuje kosiarki do trawy od dwóch producentów. Przeprowadzono ankietę wśród klientów, jak są zadowoleni z nabytych kosiarek. Wyniki obrazuje tabela:

	Klient niezadowolony, ale używa	Klient zadowolony	Kosiarka do zwrotu	Razem
Producent A	28	50	36	114
Producent B	10	32	20	62
Ogółem	38	82	56	178

Zbadać, czy dostawca i jakość kosiarek są niezależne.

Zadanie 4. W ramach ogólnopolskiej akcji budowy obiektów sportowych, w mieście X zbadano preferencje różnych sportów wśród mieszkańców. Wyniki ujmuję tabela.

Wiek, dyscyplina	18-30	31-40	41-50	>50	Razem
Tenis	25	50	75	100	250
Piłka nożna	100	80	30	10	220
Tenis stołowy	5	25	25	30	85
Biegi terenowe	20	30	40	35	125
Ogółem	150	185	170	175	680

Opracuj te wyniki. Sformułuj hipotezę zerową. Wyznacz liczbę stopni swobody. Na poziomie istotności 0,10 zbadaj hipotezę, czy preferencje sportowe są zależne od wieku. Sformułuj rezultat badań.

Zadanie 4. W zakładach Forda badano, czy usterki w wyprodukowanych samochodach zależą od tego, w jakiej części tygodnia zostały montowane samochody. Wyniki ujmują tabela. Przeprowadź stosowne opracowanie statystyczne.

	Z usterkami	Bez usterek
Poniedziałek-wtorek	67	726
Środa-czwartek	33	575
Piątek	61	363
Ogółem		

3. Zgodność z rozkładem Poissona. Znowu zilustrujemy teorię na przykładzie. Ruch turystyczny w schronisku „Daleka Pustelnia.” Testujemy, czy jest ruch turystyczny w schronisku „Daleka Pustelnia” jest zgodny z rozkładem Poissona. Obserwacje pokazują, że w ciągu jednej godziny, między 13 a 14 przychodzi tyle turystów:

Liczba turystów	Częstości
0	3
1	15
2	23
3	20
4	12
5	10
6	7
7 i więcej	5
Suma	95

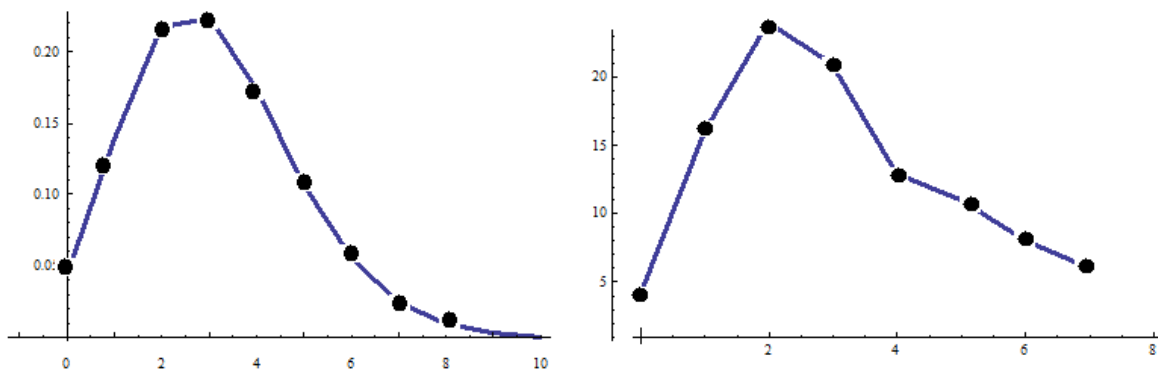
Obliczamy stąd średnią = 3,1. Testujemy hipotezę, że dane są zgodne z rozkładem Poissona dla średniej 3,1.

Tablicujemy rozkład Poissona dla średniej 3,1, obliczając kolejne wartości $3,1^k e^{-3,1} / k!$

k	Wartość $3,1^k e^{-3,1} / k!$	Liczba obok razy $N = 95$ (jest to wartość oczekiwana)	Wartość obserwowana	Składniki chi-kwadrat
0	0,450	4,275	3	0,380
1	0,1397	13,272	15	0,225
2	0,2165	20,568	23	0,288

3	0,2237	21,252	20	0,074
4	0,1734	16,473	12	1,215
5	0,1075	10,213	10	0,004
6	0,0555	5,273	7	0,566
7	0,0387	3,677	5	0,476
		Suma	$N = 95$	$\chi^2 = 3,228$

Patrzmy do tablic. Dla 6 stopni swobody wyznaczona wielkości chi-kwadrat jest mniejsza od parametru krytycznego nawet dla poziomu istotności 0,2.



Rozkład Poissona dla $\lambda = 3,1$ i rozkład obserwowany

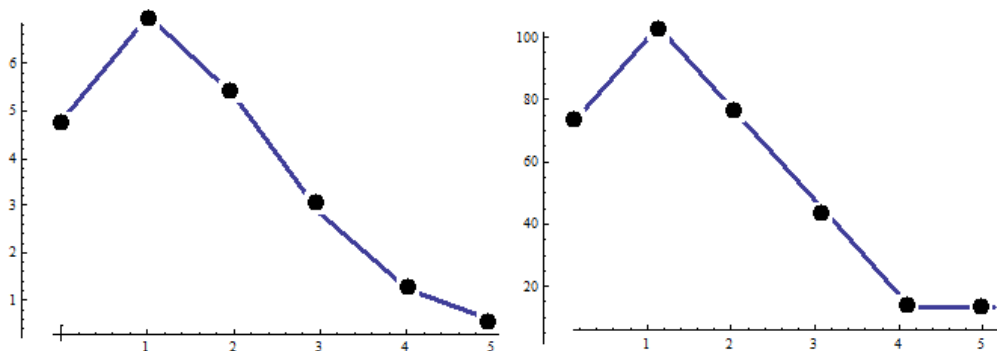
Zilustrujemy rzecz na wykresie: po lewej stronie mamy krzywą teoretyczną, po prawej – dane rzeczywiste (obserwowane). Zadanie polega na odpowiedzi na pytanie, czy te krzywe różnią się w „dopuszczalny” sposób.

Zadanie. Badano sportowców, jak często przytrafiają się im kontuzje. W wylosowanej próbie 300 sportowców, zbadano, ile każdy z nich miał poważnych kontuzji w ostatnich 3 latach. Oto tabela. Na poziomie istotności 0,10 zbadać, czy rozkład ten jest rozkładem Poissona.

Liczba kontuzji	0	1	2	3	4	5
Liczba zawodników	70	100	70	40	10	10

Szkic rozwiązania. Testujemy zgodność z rozkładem Poissona o średniej

$$\frac{70 \cdot 0 + 100 \cdot 1 + 70 \cdot 2 + 40 \cdot 3 + 10 \cdot 4 + 10 \cdot 5}{300} = 1,5$$



Zadanie o sportowcach.

Po lewej dane teoretyczne, po prawej dane obserwowane (empiryczne).

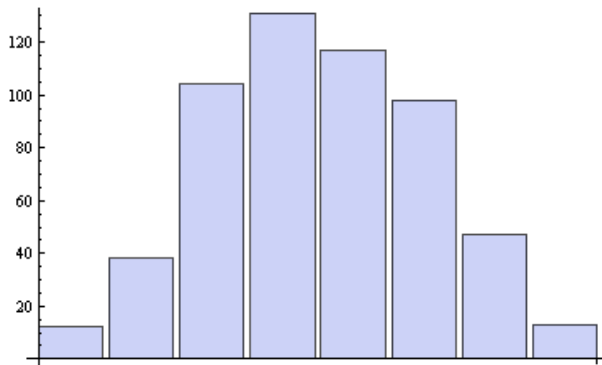
Czy krzywe te różnią się w sposób "dopuszczalny"?

4. Zgodność z rozkładem normalnym. To trudne zadanie, dotyczy bowiem zgodności z rozkładem ciągłym. Znowu omówimy test na przykładzie. Dane są prawdziwe.

Na Uniwersytecie Notre Dame w stanie Indiana w USA przeprowadzono badania, o której godzinie przychodzą do pracy pracownicy. Potrzebne to było do analizy konieczności przebudowy ulicy dojazdowej. Wyniki obrazuje tabela.

	Liczba pracowników
Między 6 a 6:30	12
Między 6:31 a 7:00	38
Między 7:01 a 7:30	104
Między 7:31 a 8:00	131
Między 8:01 a 8:30	117
Między 8:31 a 9:00	98
Między 9:01 a 9:30	47
Między 9:31 a 10:00	13
Łącznie	560

Wykres słupkowy sugeruje, że zmienna losowa, podająca liczbę pracowników przyjeżdżających do pracy o określonych porach ma rozkład normalny:



Zbadamy tę hipotezę. Musimy najpierw obliczyć średnią z próbki i odchylenie standardowe.

Średnia wynosi 8,00:

$$\frac{(12 * 6.25 + 38 * 6.75 + 104 * 7.25 + 131 * 7.75 + 117 * 8.25 + 98 * 8.75 + 47 * 9.25 + 13 * 9.75)}{560}$$

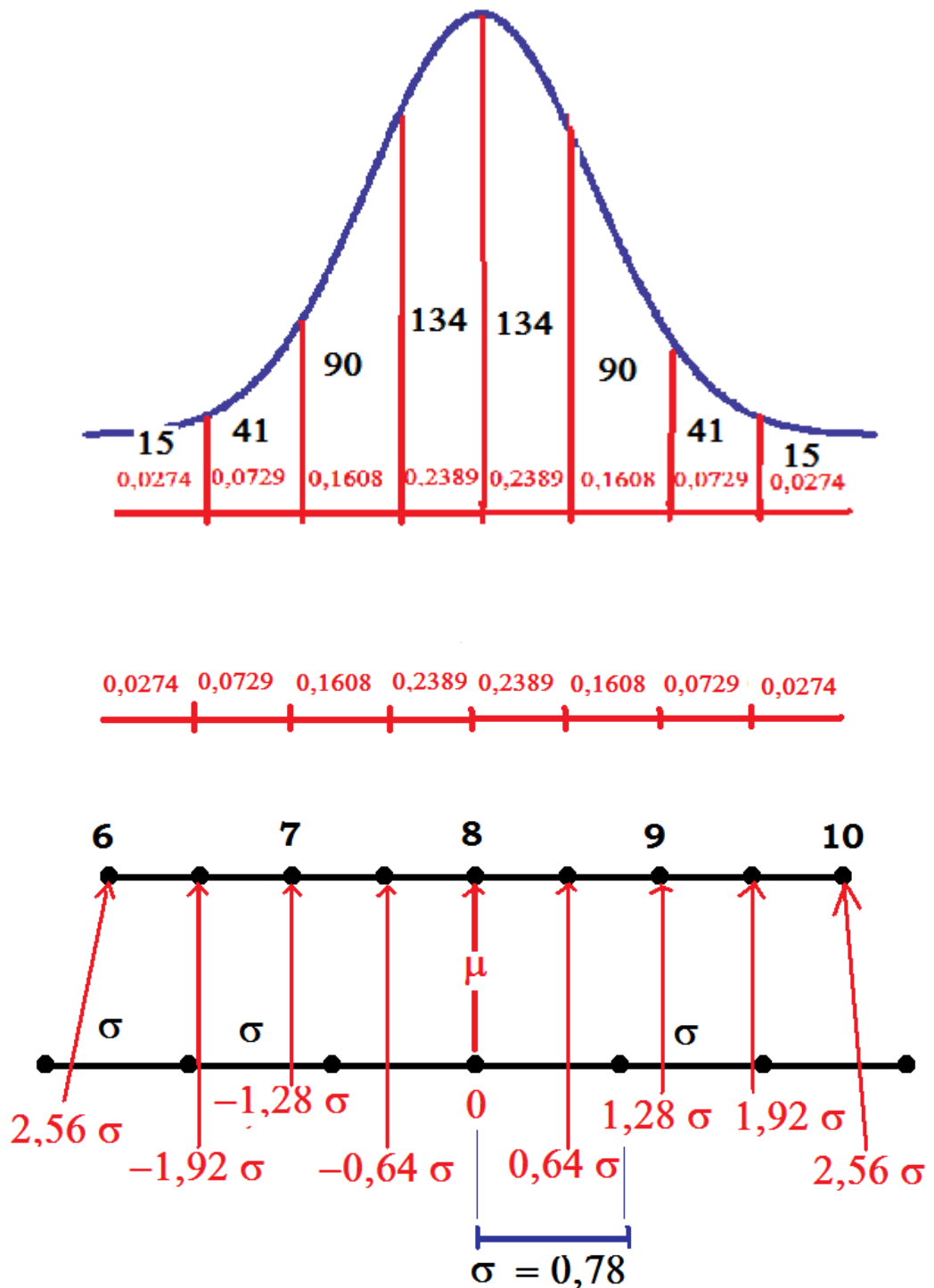
zaś odchylenie standardowe 0,78. Obliczamy je tak. Wariancja wynosi

$$\frac{(12 * (6.25 - 8)^2 + 38 * (6.75 - 8)^2 + 104 * (7.25 - 8)^2 + 131 * (7.75 - 8)^2 + 117 * (8.25 - 8)^2 + 98 * (8.75 - 8)^2 + 47 * (9.25 - 8)^2 + 13 * (9.75 - 8)^2)}{559}$$

tj. 0,605546, a odchylenie standardowe jest pierwiastkiem tej wielkości, czyli (w przybliżeniu) 0,78. Testujemy zatem hipotezę, że mamy do czynienia z rozkładem normalnym o średniej 8 i odchyleniu standardowym 0,78.

Liczba stopni swobody wynosi 5; z tablic rozkładu chi-kwadrat wnioskujemy, że odrzucimy hipotezę (na poziomie istotności 0,05), gdy χ^2 będzie większe niż 11,07.

Wyliczenie wartości oczekiwanych:



Mamy zatem

Obserwowane	12	38	104	131	117	98	47	13
Oczekiwane	15	41	90	134	134	90	41	15
(obserw. - oczek.) ² / oczek.	0,6	0,22	2,18	0,07	2,16	0,71	0,88	0,27

$$\chi^2 = 7,09$$

Jest mniejsze niż wartość krytyczna 11,07. HIPOTEZĘ PRZYJMujemy!!!!