Test χ^2 zgodności rozkładów

Poprawność wielu schematów wnioskowania statystycznego zależy od tego, czy założona postać rozkładu prawdopodobieństwa, któremu podlegają wyniki eksperymentu, odpowiada rzeczywistości.

Zgodność dwóch rozkładów, na przykład założonego i doświadczalnego, można sprawdzić przy pomocy testu χ^2.

χ^2 jest nazwą zmiennej losowej o następującym rozkładzie gęstości prawdopodobieństwa

$$p_\chi(x;\nu) = \frac{x^{\nu/2-1} \cdot e^{-x/2}}{\Gamma(\nu/2) \cdot 2^{\nu/2}}$$

gdzie $\Gamma(a)$ jest funkcją gamma, a ν jest liczbą stopni swobody rozkładu χ^2.

Funkcja gamma jest uogólnieniem silni. Dla potrzeb rozkładu χ^2 należy znać wartości $\Gamma(a)$ dla naturalnych i połówkowych wartości argumentu

$$\Gamma(a) = (a-1)\Gamma(a-1)$$
$$\Gamma(1) = 1, \ \ \Gamma(n) = (n-1)!$$
$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

Na przykład

$$\Gamma(\tfrac{5}{2}) = \tfrac{3}{2} \cdot \Gamma(\tfrac{3}{2}) = \tfrac{3}{2} \cdot \tfrac{1}{2} \cdot \Gamma(\tfrac{1}{2}) = \tfrac{3}{2} \cdot \tfrac{1}{2} \cdot \sqrt{\pi} \approx 1{,}329...$$

W statystyce dowodzi się, że jeżeli niezależne zmienne x_i mają rozkłady prawdopodobieństwa normalne $N(0,1)$, to zmienna

$$\chi^2 = \sum_{i=1}^{n} x_i^2$$

ma rozkład $p_\chi(\chi^2;n)$.

Testowanie zgodności wartości doświadczalnych z założoną postacią rozkładu prawdopodobieństwa wygląda następująco. Z danych konstruujemy histogram (szereg rozdzielczy), zawierający m przedziałów o końcach $a_0, a_1, a_2, ...a_m$. Liczebności kolejnych przedziałów wynoszą

n_i i $\sum_{i=1}^{m} n_i = n$, gdzie n jest całkowitą liczbą wartości.

Oczekiwana liczba wartości w $i-$tym przedziale histogramu wynosi nP_i, gdzie

$$P_i = P\left(x \in \left(a_{i-1}, a_i\right\rangle\right) = \int_{a_{i-1}}^{a_i} p(x)dx$$

Zmienna $\sum_{i=1}^{m} \frac{(n_i - nP_i)^2}{nP_i}$ ma (asymptotycznie przy $n \to \infty$) rozkład χ^2 o $\nu = n-1$ stopniach swobody. Jeżeli rozkład $p(x)$ ma r parametrów, których estymatory wyznaczamy z analizowanego zestawu wartości, to liczba stopni swobody zmniejsza się do $\nu = n - r - 1$.

Wartości dystrybuanty rozkładu χ^2 można znaleźć w tablicach lub korzystając z odpowiednich pakietów programów statystycznych (w tym również z arkusza kalkulacyjnego Excel).

Wartość oczekiwana rozkładu χ^2 jest równa liczbie stopni swobody

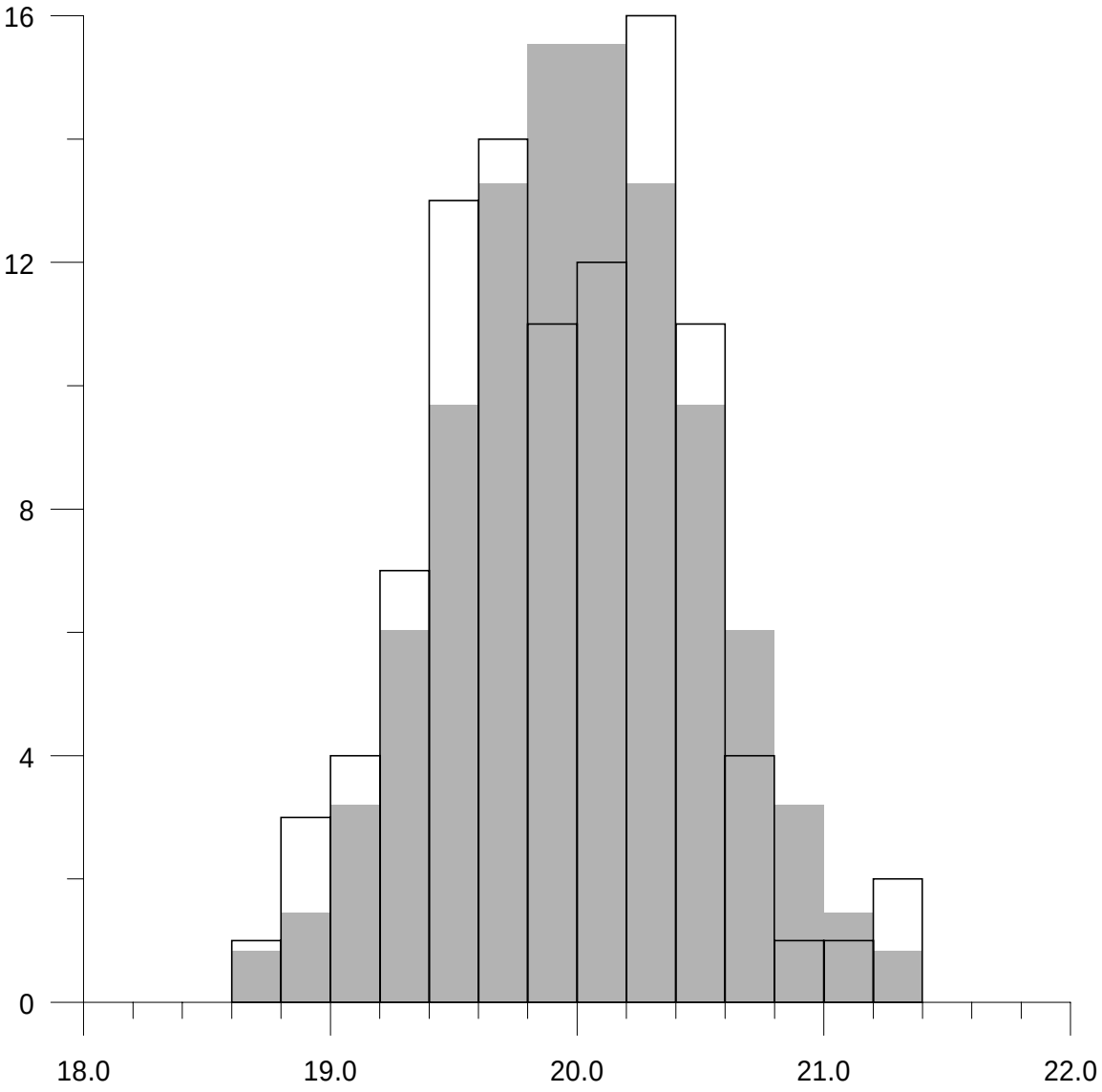
$$E(\chi^2) = \nu$$

a wariancja

$$V(\chi^2) = 2\nu\,,\ \left(\sigma(\chi^2) = \sqrt{2\nu}\right)$$

Przykład

Wyniki 100 pomiarów grupujemy w szereg rozdzielczy i tworzymy histogram (na rysunku słupki rysowane ciągłą linią). Szare słupki pokazują wartości oczekiwane histogramu, obliczone dla rozkładu normalnego $N(20{,}00;\,0{,}50)$ (założony rozkład populacji, z której pochodzą wyniki). Wysokości skrajnych słupków obliczone są dla przedziałów otwartych, w ten sposób pola obu histogramów są jednakowe i równe liczbie wszystkich wyników.



Przykład analizy χ^2 danych doświadczalnych

Wartość środkowa przedziału	Częstość obserwowana	Dla populacji $\mu = 20{,}00$ $\sigma = 0{,}50$			Dla próby $\mu' = 19{,}94$ $\sigma' = 0{,}53$		
	h_j	y_j	σ_j	$(y_j-h_j)/\sigma_j$	y_j'	σ_j'	$(y_j'-h_j)/\sigma_j'$
18,7 *	1	0,82	0,91	-0,20	1,59	1,26	0,47
18,9	3	1,46	1,21	-1,28	2,24	1,50	-0,50
19,1	4	3,20	1,79	-0,44	4,34	2,08	0,16
19,3	7	6,03	2,46	-0,40	7,29	2,70	0,11
19,5	13	9,68	3,11	-1,07	10,64	3,26	-0,72
19,7	14	13,27	3,64	-0,20	13,50	3,67	-0,14
19,9	11	15,54	3,94	1,15	14,89	3,86	1,01
20,1	12	15,54	3,94	0,90	14,28	3,78	0,60
20,3	16	13,27	3,64	-0,75	11,90	3,45	-1,19
20,5	11	9,68	3,11	-0,42	8,62	2,94	-0,81
20,7	4	6,03	2,46	0,83	5,43	2,33	0,61
20,9	1	3,20	1,79	1,23	2,97	1,72	1,14
21,1	1	1,46	1,21	0,38	1,41	1,19	0,35
21,3 *	2	0,82	0,91	-1,30	0,88	0,94	-1,19
			χ_0^2	10,13		χ_0^2	7,72
			v	13		v	11
			χ_v^2	0,78		χ_v^2	0,70
		$P(\chi^2 \geq \chi_0^2; v)$		0,68	$P(\chi^2 \geq \chi_0^2; v)$		0,74

W praktyce należy zadbać o to, żeby wszystkie przedziały były odpowiednio liczne, to znaczy nie zawierały mniej niż 5-10 wartości. Przedziały histogramu nie muszą być jednakowej szerokości.

W powyższym przykładzie skrajne przykłady powinny być połączone w większe.

Wartości χ^2 obliczone w tym przykładzie są trochę mniejsze od oczekiwanych (odpowiednio 13 i 11). Dyspersje tych wartości są jednak względnie duże (odpowiednio 5,1 i 4,7) i dlatego odpowiednie prawdopodobieństwa są dość bliskie 50%. Bardzo małe wartości prawdopodobieństwa (np., poniżej 3%) oznaczałyby zbyt duże różnice między rozkładami i mogłyby być podstawą do odrzucenia hipotezy o postaci rozkładu.
W przypadku bardzo dużych wartości prawdopodobieństwa (bliskich 100%), co oznaczałoby bardzo dobrą zgodność wyników z zakładanym rozkładem, można podejrzewać brak losowości wyników.