Badanie zgodności z określonym rozkładem

H_0 : Cecha X ma rozkład F

F jest dowolnym rozkładem prawdopodobieństwa

Test chi–kwadrat zgodności

F jest rozkładem ciągłym

Test Kołmogorowa

F jest rozkładem normalnym

Test Shapiro–Wilka

Test Chi–kwadrat zgodności
(poziom istotności α)

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

Statystyka testowa

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

$$n_i^t = Np_i^t,\ N = \sum_{i=1}^{k} n_i,$$

$$p_i^t = P_F\{X \text{ przyjęła wartość z klasy } i\}$$

Wartość krytyczna $\chi^2(\alpha; k-u-1)$ (u jest liczbą nieznanych parametrów hipotetycznego rozkładu F)

Wniosek. Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; k-u-1)$, to hipotezę H_0 odrzucamy

Przykład. Pracodawca przypuszcza, że liczba pracowników nieobecnych w różne dni tygodnia nie jest taka sama. W tym celu w ciągu pewnego okresu czasu zebrał następujące dane

Dzień	n_i
Poniedziałek	200
Wtorek	160
Środa	140
Czwartek	140
Piątek	100

Populacja:

Cecha X:
dzień nieobecności pracownika

Założenie:
cecha przyjmuje wartości będące nazwami dni tygodnia (cecha jakościowa)

Formalizacja:
Liczbę pracowników nieobecnych w kolejne dni tygodnia można przedstawić jako odesetek załogi. Odsetki te można interpretować jako prawdopodobieństwo nieobecności pracownika w danym dniu tygodnia. Jeżeli ilość pracowników nieobecnych w kolejne dni tygodnia jest „mniej więcej” taka sama, to można ten fakt sformalizować jako identyczne prawdopodobieństwo nieobecności pracownika w poszczególne dni tygodnia. Tak więc, weryfikowana będzie hipoteza

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

Technika statystyczna:
test chi–kwadrat zgodności
poziom istotności $\alpha = 0.05$

Obliczenia

Dzień	n_i	p_i^t	n_i^t	$(n_i - n_i^t)^2/n_i^t$
Poniedziałek	200	1/5	148	$\frac{(200-148)^2}{148}$=18.270
Wtorek	160	1/5	148	$\frac{(160-148)^2}{148}$= 0.973
Środa	140	1/5	148	$\frac{(140-148)^2}{148}$= 0.432
Czwartek	140	1/5	148	$\frac{(140-148)^2}{148}$= 0.432
Piątek	100	1/5	148	$\frac{(100-148)^2}{148}$=15.676
	740			χ^2_{emp}=35.676

Wartość krytyczna

$$\chi^2(\alpha; k-u-1) = \chi^2(0.05; 5-0-1) = 9.4877$$

Odpowiedź: hipotezę odrzucamy

Wniosek:
Odrzucamy hipotezę o równomiernym rozkładzie nieobecności w tygodniu. Zatem przypuszczenie pracodawcy można uznać za uzasadnione

Przykład. Na pewnej uczelni badano strukturę miesięcznych dochodów (na głowę) w rodzinach studentów. W tym celu wylosowano grupę 192 studentów i zanotowano miesięczne dochody w ich rodzinach. Uzyskano następujące wyniki (w setkach złotych):

x_i	x_{i+1}	n_i
poniżej 6		6
6	7	11
7	8	18
8	9	27
9	10	32
10	11	35
11	12	24
12	13	20
13	14	13
powyżej 14		6

Czy można, że rozkładów dochodów w rodzinach studenckich jest normalny?

Populacja:
studenci pewnej uczelni

Cecha X:
miesięczne dochody na głowę w rodzinach studentów

Założenie:
cecha ciągła

Formalizacja:

$$H_0 : \text{Cecha } X \text{ ma rozkład normalny } N(\mu, \sigma^2)$$

Technika statystyczna:
test chi–kwadrat zgodności
poziom istotności $\alpha = 0.05$

Obliczenia
Szereg ma $k = 10$ klas
Do całkowitego określenia hipotetycznego rozkładu brakuje dwóch parametrów, czyli $u = 2$
Wartość krytyczna
$\chi^2(\alpha; k - u - 1) = \chi^2(0.05; 10 - 2 - 1) = 14.0671$

Wyznaczenie wartości statystyki χ^2_{emp}

Wyznaczenie prawdopodobieństw teoretycznych

$$p_i^t = P\{x_i < X < x_{i+1}\} = F\left(\frac{x_{i+1}-\mu}{\sigma}\right) - F\left(\frac{x_i-\mu}{\sigma}\right)$$

Z próby wyznaczamy $\bar{x} = 10.09, s^2 = 4.81$

$$p_i^t = F\left(\frac{x_{i+1}-\bar{x}}{s}\right) - F\left(\frac{x_i-\bar{x}}{s}\right) = F(z_{i+1}) - F(z_i)$$

x_i	x_{i+1}	z_i	z_{i+1}	$F(z_i)$	$F(z_{i+1})$	p_i^t
poniżej 6		$-\infty$	-1.82	0.0000	0.0344	0.0344
6	7	-1.82	-1.36	0.0344	0.0869	0.0525
7	8	-1.36	-0.91	0.0869	0.1814	0.0943
8	9	-0.91	-0.45	0.1814	0.3264	0.1450
9	10	-0.45	0.00	0.3264	0.5000	0.1736
10	11	0.00	0.45	0.5000	0.6736	0.1736
11	12	0.45	0.91	0.6736	0.8186	0.1450
12	13	0.91	1.36	0.8186	0.9131	0.0943
13	14	1.36	1.82	0.9131	0.9656	0.0525
powyżej 14		1.82	∞	0.9656	1.0000	0.0344

Wyznaczenie wartości statystyki testowej

x_i	x_{i+1}	n_i	p_i^t	n_i^t	$(n_i - n_i^t)^2/n_i^t$
poniżej 6		6	0.0344	6.53	0.36
6	7	11	0.0525	10.18	0.07
7	8	18	0.0943	18.05	0.00
8	9	27	0.1450	27.84	0.02
9	10	32	0.1736	33.41	0.06
10	11	35	0.1736	33.41	0.08
11	12	24	0.1450	27.84	0.53
12	13	20	0.0943	18.05	0.21
13	14	13	0.0525	10.18	0.78
powyżej 14		6	0.0344	6.53	0.03
		192			**2.14**

Odpowiedź. Nie odrzucamy hipotezy

Wniosek. Możemy uznać, że miesięczne dochody na głowę w rodzinach studentów mają rozkład normalny $N(10.09, 4.81)$