Statystyczne testy nieparametryczne

Testami nieparametrycznymi nazywamy testy służące do weryfikacji hipotez nieparametrycznych, tj hipotez niedotyczących wartości nieznanych parametrów populacji (choć czasem pojęcie to oznacza hipotezy nie zakładające rozkładu Normalnego dla populacji). Ze względu na różnorodność hipotez nieparametrycznych, klasę testów nieparametrycznych można podzielić na następujące podklasy:

- testy zgodności (z pewnym rozkładem teoretycznym), w tym testy normalności,
- testy jednorodności, czyli zgodności dwóch (lub więcej) rozkładów,
- testy niezależności,
- inne testy, w tym np. testy weryfikujące hipotezę, ze próba ma charakter losowy.

Badanie niezależności rozkładu dwu cech

- Cecha (X, Y) ma dwuwymiarowy, nieznany rozkład

⇓

Test Chi-Kwadrat niezależności

- Cecha (X, Y) ma dwuwymiarowy rozkład ciągły

⇓

Współczynnik korelacji rangowej Spearmana

Współczynnik korelacji rangowej Kendalla

- Czy kolejność obserwacji w próbie jest losowa?

⇓

test serii

ZALEŻNOŚĆ MIĘDZY CECHAMI JAKOŚCIOWYMI /SKATEGORYZOWANYMI/

X, Y – cechy obserwowane

Próba: $(X_1, Y_1), \ldots (X_k, Y_m)$

klasy cechy X	**klasy cechy Y**			
	1	**2**	**. . .**	**m**
1	n_{11}	n_{12}	. . .	n_{1m}
2	n_{21}	n_{22}	. . .	n_{2m}
⋮	⋮	⋮		⋮
k	n_{k1}	n_{k2}	. . .	n_{km}

H_0: Cechy X i Y są niezależne

H_1: Cechy X i Y są zależne

Test Chi-kwadrat (c^2) niezależności

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

n_{ij} – liczba obserwacji realizujących i–tą wartość cechy X i j-tą wartość Y

n^t_{ij} – teoretyczna liczba obserwacji realizujących i–tą wartość cechy X i j-tą wartość Y (wg. rozkładów brzegowych dla każdej z tych dwu cech)

$$n^t_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{N}, \qquad N = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij},$$

$$n_{i\bullet} = \sum_{j=1}^{m} n_{ij}, \qquad n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$$

Jeśli $c^2_{emp} > c^2_{kryt}$, to hipotezę H_0 odrzucamy.

$c^2_{kryt} = c^2_{a,v}$, gdzie $v = (k-1)\bullet(m-1)$

Przykład:

Badano dwie właściwości wędliny: związanie (słabo związana, związana, dobrze związana) oraz smakowitość (dostateczna, dobra, bardzo dobra). Analizę przeprowadzono dla 60 batonów wędliny. Wyniki były następujące:

X- smakowitość	**Y_1-słabo związana**	**Y_2- związana**	**Y_3- dobrze związana**	**$n_{i.}$**
X_1- dostateczna	9	5	3	**17**
X_2-dobra	4	12	6	**22**
X_3-b. dobra	1	6	14	**21**
$n_{.j}$	**14**	**23**	**23**	**60**

Hipoteza badawcza i statystyczna brzmi:

H_0 : Smakowitość wędliny (X) nie zależy od stopnia związania (Y) tj. cechy te są niezależne

Konstruujemy funkcję testową opartą na rozkładzie chi-kwadrat.

Dane (**liczebności**) **teoretyczne** N^t_{ij} (jeśli cechy są niezależne czyli P(A∩B)=P(A)*P(B), to rozkłady brzegowe P(A) i P(B) wyznaczają prawdopodobieństwo podklas P(A∩B)):

X-smakowitość	**Y_1-słabo związana**	**Y_2-związana**	**Y_3-dobrze związana**	**$n_{i.}$**
X_1-dostateczna	4	6,5	6,5	**17**
X_2-dobra	5	8,5	8,5	**22**
X_3-b. dobra	5	8	8	**21**
$n_{.j}$	**14**	**23**	**23**	**60**

$$n^t_{11}=\frac{17\cdot 14}{60}=3.97,\quad n^t_{12}=\frac{17\cdot 23}{60}=6.52,\ldots$$

$$n_{33}^{t} = \frac{21 \cdot 23}{60} = 8.05$$

$$c_{emp}^2 = \frac{(9-3.97)^2}{3.97} + \frac{(5-6.52)^2}{6.52} + \ldots + \frac{(14-8.05)^2}{8.05}$$

$$= 19.2$$

$$c_{kryt}^2 = c_{a,v}^2 = c_{0.05,4}^2 = 9.49$$

Ponieważ $c_{emp}^2 > c_{kryt}^2$, to hipotezę H_0 odrzucamy.

Wyniki pozwalają stwierdzić, że smakowitość badanej wędliny zależy od związania jej składników. Dla podniesienia walorów sensorycznych tej wędliny należy tak prowadzić proces technologiczny, aby uzyskać możliwie największe jej związanie

Można wyznaczyć także współczynnik kontyngencji ***P***, który przyjmuje wartość zero, gdy występuje całkowita niezależność cech.

$$P=\sqrt{\frac{c^2}{N+c^2}}$$

Dla naszego przykładu

$$P=\sqrt{\frac{c^2}{N+c^2}}=\sqrt{\frac{19.2}{60+19.2}}=0.492$$

co świadczy o dużej sile związku między rozważanymi cechami.

TESTY ZGODNOŚCI

Hipotezy tego typu dotyczą zgodności rozkładu empirycznego z rozkładem określonym przez hipotezę lub zgodności (jednorodności) rozkładów pewnej cechy w kilku populacjach bez określania, o jaki rozkład chodzi. Z tego też powodu testy służące do weryfikacji takich hipotez nazywamy **testami zgodności (jednorodności)**.

Do najczęściej stosowanych testów zgodności należą:

- $\chi 2$ (chi-kwadrat) Pearsona
- λ (lambda) Kołmogorowa-Smirnowa
- *w* Shapiro-Wilka

Niech hipotezą zerową będzie przypuszczenie, że cecha *X* ma w populacji rozkład określony dystrybuantą $F_0(x)$:

$$H_0 : F(x) = F_0(x) \text{ wobec } H_1 : F(x) \neq F_0(x)$$

Statystyka

$$c^2 = \sum_{j=1}^{k} \frac{(n_j - n_j^t)^2}{n_j^t}$$

przy prawdziwości H_0 ma asymptotyczny rozkład c^2 z liczbą stopni swobody $v=k-u-1$.

Wielkość $n_j^t = np_j$ jest teoretyczną (to znaczy, obliczoną przy założeniu prawdziwości testowanej hipotezy H_0) liczebnością w *j*–tym przedziale, *k* jest liczbą przedziałów klasowych, a *u* **liczbą parametrów populacyjnych, szacowanych z próby**.

Wartość empiryczną statystyki

$$c^2_{emp} = \sum_j \frac{(n_j - n_j^t)^2}{n_j^t}$$

porównujemy z wartością krytyczną oczytaną z tablic statystycznych

$$c^2_{a,v=k-u-1}$$

wnioskując analogicznie jak w pozostałych hipotezach.

Elementem kluczowym przy wykorzystaniu statystyki Chi-kwadrat jest wielkość

$$p_j^t = P(x \in (x_{1j}; x_{2j}))$$

która jest teoretycznym (to znaczy, obserwowanym przy założeniu prawdziwości testowanej hipotezy H_0) prawdopodobieństwem wystąpienia obserwacji w j–tym przedziale.

Przykład: Pracodawca przypuszcza, że liczba pracowników nieobecnych w różne dni tygodnia nie jest taka sama.

W celu sprawdzenia swojego przypuszczenia obserwował, przez pewien okres, liczby pracowników nieobecnych w kolejnych dniach tygodnia. Wyniki obserwacji zawiera tabela:

dzień tygodnia	**liczba nieobecnych**
poniedziałek	200
wtorek	160
środa	140
czwartek	140
piątek	100

Badaną cechą X jest dzień, w którym pracownik był nieobecny w pracy. Jest to cecha jakościowa o wartościach: poniedziałek, wtorek, … , piątek.

Hipoteza badawcza, że absencja pracownika jest zależna od dnia tygodnia pracy, może być zapisana ‘przez negację’, to znaczy sugerujemy **brak preferencji w opuszczaniu dni.**

Zapis statystyczny tego przypuszczenia pracodawcy ma postać hipotezy:

H : cecha X ma rozkład:

Pon.	**Wtk.**	**Śro.**	**Czw.**	**Ptk.**
1/5	**1/5**	**1/5**	**1/5**	**1/5**

Do weryfikacji badanej hipotezy stosujemy test chi–kwadrat zgodności, przyjmując $\alpha = 0.05$.

Pomocnicze obliczenia funkcji testowej zawiera tabela:

	$\mathbf{n_i}$	$\mathbf{n_i}$ teoret.
Pon	200	148
Wtk	160	148
Śro	140	148
Czw	140	148
Ptk	100	148
suma	740	740

Wartość statystyki jest wyznaczona według formuły:

$$c^2_{emp} = \sum_j \frac{(n_j - n^t_j)^2}{n^t_j} =$$

$$\frac{(200-148)^2}{148} + \ldots + \frac{(100-148)^2}{148} = 35{,}68$$

Ponieważ wartość krytyczna

$$c^2_{a,v=k-u-1} = \; c^2_{0.05,5-0-1} = c^2_{0.05,4} = 9.49$$

zachodzi relacja $c^2_{emp} > c^2_{0.05,4}$, czyli hipotezę o zgodności z określonym rozkładem odrzucamy.

Oznacza to, że przypuszczenie pracodawcy o nierównomiernym rozkładzie absencji w zakładzie pracy można uznać za uzasadnione.

Test c2 zgodności (jednorodności) kilku rozkładów

Obserwujemy tę samą cechę w kilku populacjach. Interesuje nas odpowiedź na pytanie, czy rozkłady te są takie same (**co pociąga za sobą równość wszystkich parametrów**).

Jeżeli dystrybuantę danej cechy w *i*–tej populacji oznaczymy jako F_i, to hipoteza zerowa ma postać:

$$H_0 : F_1 = F_2 = ... = F_k$$

Zastosowanie testu χ^2 wymaga zestawienia próby w postaci tabeli dwukierunkowej.

W jednym kierunku umieszczamy poziomy danej cechy, w drugim populacje.

numer populacji	**klasy cechy X**			
	$\mathbf{X_1}$	$\mathbf{X_2}$	. . .	$\mathbf{X_r}$
1	n_{11}	n_{12}	. . .	n_{1r}
2	n_{21}	n_{22}	. . .	n_{2r}
⋮	⋮	⋮		⋮
k	n_{k1}	n_{k2}	. . .	n_{kr}

Statystyka testowa ma postać:

$$c^2_{emp} = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{\left(n_{ij} - n^t_{ij}\right)^2}{n^t_{ij}}$$

gdzie $n^t_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$

n_{ij} – oznacza liczbę obserwacji reprezentujących i–tą populację i j–tą klasę cechy X.

Z indeksem górnym *t*, jest to odpowiednia **liczebność teoretyczna**.

Przy prawdziwości H_0 statystyka ta ma rozkład χ^2 Pearsona z liczbą stopni swobody $v=(k-1)(r-1)$.

Wnioskowanie przebiega analogicznie jak przy innych hipotezach.

przykład:

We wszystkich 10 sklepach pewnej sieci sklepów jest takie samo zapotrzebowanie na mleko o różnej zawartości tłuszczu.

sklep	0,5%	2,0%	3,2%
1	34	36	28
2	48	42	46
3	15	18	10
4	61	45	51
5	37	29	46
6	18	16	25
7	39	28	35
8	42	18	31
9	41	38	29
10	19	26	16

Wartość statystyki C^2_{emp} wynosi 23,8. Powoduje to, iż krytyczny poziom istotności (P-value) wynosi 16%. Nie ma podstaw, aby na poziomie istotności 5% odrzucić hipotezę o takim samym rozkładzie preferencji odnośnie zawartości tłuszczu we wszystkich 10 sklepach