Test nieparametryczne

Dotyczą:

- porównania dwóch grup danych typu ilościowego, gdy ich rozkład jest zdecydowanie różny od normalnego,
- porównania dwóch grup danych typu porządkowego,
- rodzaju samego rozkładu zmiennej losowej,
- losowości próby,
- niezależności zmiennych.

Hipotezę zerową odrzucamy, jeżeli wynik testu należy do **obszaru krytycznego** i wówczas przyjmujemy hipotezę alternatywną (H_A). W przeciwnym przypadku, nie ma podstaw do odrzucenia H_0.

Podstawowym kryterium odrzucenia weryfikowanej hipotezy, w programie STATISTICA jest nierówność: $\mathbf{p < \alpha,}$ gdzie **p** jest obliczonym przez program prawdopodobieństwem testowym, a α to założony poziom istotności.

1) Porównanie dwóch prób niezależnych:
   - dla zmiennych mierzalnych: H_0: $m=m_0$
   - dla zmiennych w skali porządkowej:
     H_0: występowanie badanej cechy w różnych próbach nie różni się istotnie

     a) Test U Manna – Whitneya
     b) Test serii Walda – Wolfowitza
     c) Test Kołmogorowa - Smirnowa

2) Testy nieparametryczne dla prób zależnych (ta sama grupa dwukrotnie badana)
   H_0: występowanie badanej cechy w różnych próbach nie różni się istotnie

   a) Test znaków (oparty na znakach różnic pomiędzy kolejnymi parami wyników)
   b) Test kolejności par Wilcoxona

3) Porównywanie zmiennych jakościowych
   a) Dwie próby zależne z wynikami dychotomicznymi
      - Test McNemary,
        H_0: jedna cecha nie ma istotnego wpływu na drugą

b) Więcej prób zależnych z wynikami dychotomicznymi
   - Test Q Cochrana
   - Test Kruskala - Wallisa

     H_0: jedna cecha nie ma istotnego wpływu na drugą

c) Test niezależności chi kwadrat

   H_0: zmienne X i Y są niezależne

4) Testy zgodności
   a) Test chi kwadrat Pearsona
   b) Test Kołmogorowa - Smirnowa

      H_0: zmienna X ma rozkład R

Test McNemary

Test ten służy do określania istotności różnic w wynikach, które zaszły pod wpływem jakiegoś działania

H_0: jedna cecha (oddziaływanie) nie ma istotnego wpływu na drugą

		Po działaniu		Suma
		–	**+**	
Przed działaniem	**+**	**A**	**B**	**A+B**
	–	**C**	**D**	**C+D**
Suma		**A+C**	**B+D**	**N**

A oznacza liczbę osób, u których w wyniku zastosowanego działania, poziom badanej cechy zmienił się (z + na −) lub cecha + zmieniała się na cechę −
D oznacza liczbę osób, u których w wyniku zastosowanego działania, poziom badanej cechy zmienił się (z − na +) lub cecha − zmieniała się na cechę +
B i **C** liczba osób, u których w wyniku zastosowanego działania, poziom badanej cechy nie zmienił się (z + na + lub z − na −)

Test	Obszar krytyczny	Obszar przyjęcia hipotezy
$\chi^2 = \frac{\left(\lvert A-D\rvert -1\right)^2}{A+D}$	$\left(\chi^2_\alpha, +\infty\right)$	$\left\langle 0, \chi^2_\alpha \right\rangle$

liczbę χ^2_α odczytujemy z tablic rozkładu χ^2 (chi kwadrat) dla 1 stopnia swobody i danego poziomu istotności α.

Przykład 1

Przebadano 195 pacjentów na występowanie pewnych bakterii. Stwierdzono ich występowanie u 103 osób. Po zastosowaniu leczenia przeprowadzono ponownie badania. Bakterie wykryto u 47 osób, z czego 39 to pacjenci, u których wcześniej tez wykryto bakterie. Czy można stwierdzić, że leczenie ma istotny wpływ na zmniejszenie się liczby osób zakażonych bakteriami?

Przykład 1

Przebadano 195 pacjentów na występowanie pewnych bakterii. Stwierdzono ich występowanie u 103 osób. Po zastosowaniu leczenia przeprowadzono ponownie badania. Bakterie wykryto u 47 osób, z czego 39 to pacjenci, u których wcześniej tez wykryto bakterie. Czy można stwierdzić, że leczenie ma istotny wpływ na zmniejszenie się liczby osób zakażonych bakteriami?

		Po leczeniu		Suma
		–	**+**	
Przed leczeniem	**+**	**A**	**B**	**A+B**
	–	**C**	**D**	**C+D**
Suma		**A+C**	**B+D**	**N**

Przykład 1

Przebadano 195 pacjentów na występowanie pewnych bakterii. Stwierdzono ich występowanie u 103 osób. Po zastosowaniu leczenia przeprowadzono ponownie badania. Bakterie wykryto u 47 osób, z czego 39 to pacjenci, u których wcześniej tez wykryto bakterie. Czy można stwierdzić, że leczenie ma istotny wpływ na zmniejszenie się liczby osób zakażonych bakteriami?

		Po leczeniu		Suma
		–	**+**	
Przed leczeniem	**+**	**64**	**39**	**103**
	–	**84**	**8**	**92**
Suma		**148**	**47**	**195**

$$\chi^2 = \frac{\left(|A-D|-1\right)^2}{A+D} = \frac{\left(|64-8|-1\right)^2}{64+8} = 42,01$$

Liczba χ^2_α dla 1 stopnia swobody i poziomu istotności α = 0,01 wynosi 6,64, czyli statystyka testowa należy do obszaru krytycznego $\left(6,64;+\infty\right)$. Odrzucamy więc hipotezę zerową i stwierdzamy, że leczenie ma istotny wpływ na liczbę osób zakażonych bakteriami.

Przykład 2

Przebadano 195 pacjentów na występowanie pewnych bakterii. Stwierdzono ich występowanie u 103 osób. Po zastosowaniu leczenia przeprowadzono ponownie badania. Bakterie wykryto u 82 osób, z czego 39 to pacjenci, u których wcześniej tez wykryto bakterie. Czy można stwierdzić, że leczenie ma istotny wpływ na zmniejszenie się liczby osób zakażonych bakteriami?

		Po leczeniu		Suma
		–	**+**	
Przed leczeniem	**+**	**60**	**43**	**103**
	–	**53**	**39**	**92**
Suma		**113**	**82**	**195**

$$\chi^2 = \frac{\left(|A-D|-1\right)^2}{A+D} \quad \text{czyli} \quad \chi^2=400/99=4{,}1$$

Liczba χ^2_α dla 1 stopnia swobody i poziomu istotności α = 0,01 wynosi 6,64, czyli statystyka testowa należy do obszaru przyjęcia hipotezy zerowej (0, 6.64⟩. Brak więc podstaw by odrzucić hipotezę zerową i stwierdzamy, że leczenie nie ma istotnego wpływu na liczbę osób zakażonych bakteriami.

Test niezależności chi kwadrat

Jeżeli przedmiotem badania jest populacja ze względu na występowanie dwóch cech X i Y, to w celu stwierdzenia niezależności tych cech stosujemy test niezależności chi kwadrat. Jest on oparty o tak zwaną tablicę niezależności. Tablica ta zawiera tyle wierszy ile jest wariantów cechy X i tyle kolumn ile jest wariantów cechy Y.
Niech k oznacza liczbę wariantów cechy X, a r liczbę wariantów cechy Y. Wtedy tablica niezależności wygląda następująco:

Y / X	y_1	y_2	...	y_r	
x_1	n_{11}	n_{12}	...	n_{1r}	$\sum_{j=1}^{r} n_{1j}$
x_2	n_{21}	n_{22}	...	n_{2r}	$\sum_{j=1}^{r} n_{2j}$
...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kr}	$\sum_{j=1}^{r} n_{kj}$
	$\sum_{i=1}^{k} n_{i1}$	$\sum_{i=1}^{k} n_{i2}$	...	$\sum_{i=1}^{k} n_{ir}$	

Test	Obszar krytyczny	Obszar przyjęcia hipotezy
$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(\frac{n_{ij}^2}{\widehat{n}_{ij}}\right) - n$	$\left(\chi_{\alpha}^2, +\infty\right)$	$\left\langle 0, \chi_{\alpha}^2 \right\rangle$

gdzie

n to liczność próby ,

n_{ij} to zaobserwowane liczności z tabeli niezależności,

$\widehat{n}_{ij}$ to teoretyczne liczności wystąpienia odpowiednich wariantów, gdyby zmienne *X* i *Y* były niezależne.

Teoretyczne liczności oblicza się według wzoru: $\widehat{n}_{ij} = \frac{1}{n} \cdot \sum_{i=1}^{k} n_{ij} \cdot \sum_{j=1}^{r} n_{ij}$

Dla zadanego poziomu istotności α, z tablic rozkładu χ^2 z $\left(r-1\right)\left(k-1\right)$stopniami swobody odczytujemy liczbę χ_{α}^2

Przykład

Przy nowym podziale studentów na grupy, postanowiono zbadać zależność między oceną semestralną z języka angielskiego, a oceną semestralną z matematyki. Poniższa tablica zawiera liczebności studentów, którzy uzyskali dane oceny z angielskiego i z matematyki.

Mat / *J. Ang*	2	3	4	5	$\sum_{j=1}^{r} n_{kj}$
2	8	5	0	0	13
3	7	57	34	3	101
4	1	17	27	4	47
5	0	20	12	3	39
$\sum_{i=1}^{k} n_{ir}$	16	99	71	14	**200**

Zweryfikujemy hipotezę o niezależności ocen z języka angielskiego i z matematyki, na poziomie istotności 0,1.

Testem jest w tym wypadku statystyka: $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(\frac{n_{ij}^2}{\widehat{n}_{ij}}\right) - n$, gdzie n = 200, a n_{ij} to zaobserwowane

liczności z tabeli niezależności .

Teoretyczne liczności obliczamy według wzoru: $\widehat{n}_{ij} = \frac{1}{n} \cdot \sum_{i=1}^{k} n_{ij} \cdot \sum_{j=1}^{r} n_{ij}$

$$\widehat{n}_{11} = \frac{1}{200} \cdot 13 \cdot 16 = 1{,}04$$

$$\widehat{n}_{12} = \frac{1}{200} \cdot 13 \cdot 99 = 6{,}44$$

Obliczamy tak wszystkie liczebności teoretyczne i obliczamy statystykę testową. Dla ułatwienia dalszych obliczeń, liczebności teoretyczne można również umieścić w tabeli:

Mat / *J. Ang*	2	3	4	5	$\sum_{j=1}^{r} n_{kj}$
2	1,04	6,44	4,62	0,91	13
3	8,08	50	35,86	7,07	101
4	3,76	23,27	16,69	3,29	47
5	3,12	19,31	13,85	2,73	39
$\sum_{i=1}^{k} n_{ir}$	16	99	71	14	**200**

Wartość statystyki testowej wynosi $\chi^2 = 274{,}03$.

Jako, że liczba wariantów cechy X (J. Ang.) jest równa k=4 i że liczba wariantów cechy Y (Mat) jest równa r=4, stąd wartość krytyczną χ^2_α odczytujemy z tablic rozkładu χ^2 dla 9 stopni swobody ($(r-1)(k-1)$) i α=0,1.

$\chi^2_\alpha = 14{,}684$ czyli obszarem krytycznym jest przedział (14,684; ∞).

Wartość statystyki testowej należy do tego przedziału, należy więc zdecydowanie odrzucić hipotezę o niezależności ocen z języka angielskiego i z matematyki.

1. Zbadać niezależność cech X – miejsce zamieszkania studenta i Y – wynik egzaminu, na podstawie poniższej tabeli:

Y \ *X*	Akademik	Stancja	Dom
zdał	10	25	30
Nie zdał	40	25	20

2. Zbadać skuteczność szkolenia, jeżeli wiadomo, że przed szkoleniem test zaliczyło 80 na 150 studentów. A po szkoleniu test zaliczyło 100 studentów, pośród których 70 zaliczyło go przed szkoleniem.

3. Zbadać niezależność cech X – dzienna liczba godzin przy komputerze i Y – roczna liczba przeczytanych książek, na podstawie poniższej tabeli:

Y \ *X*	0 - 2	2 - 5	Ponad 5
0 - 5	20	6	3
6 – 10	10	2	0
Ponad 10	5	0	0

4. Zbadać skuteczność środka na odchudzanie, jeżeli wiadomo, że przed jego zastosowaniem 50 na 120 osób straciło na wadze. A po jego zastosowaniu schudło 60 osób, pośród których 40 straciło na wadze wcześniej bez stosowania badanego środka.

1. Zbadać niezależność cech X i Y , na podstawie poniższej tabeli:

Y \ *X*	A	B
1	2	5
2	4	3
3	0	2

2. Zbadać skuteczność szkolenia, jeżeli wiadomo, że przed szkoleniem test zaliczyło 90 na 170 studentów. A po szkoleniu test zaliczyło 140 studentów, pośród których 80 zaliczyło go przed szkoleniem.

3. Zbadać niezależność cech X – dzienna liczba godzin przy komputerze i Y – roczna liczba przeczytanych książek, na podstawie poniższej tabeli:

Y \ *X*	0 - 2	2 - 5	Ponad 5
0 - 5	20	6	3
6 – 10	10	2	0
Ponad 10	5	0	0

4.Zbadać skuteczność środka na odchudzanie, jeżeli wiadomo, że przed jego zastosowaniem 60 na 180 osób straciło na wadze. A po jego zastosowaniu schudło 70 osób, pośród których 50 straciło na wadze wcześniej bez stosowania badanego środka.

Test zgodności chi kwadrat (Pearsona)

Hipoteza H_0 jest hipotezą orzekającą, że dystrybuanta zmiennej losowej *X* ma postać $F(x)$, a hipotezą alternatywną jest hipoteza, która stwierdza, że rozkład zmiennej *X* ma dystrybuantę różną od $F(x)$.

Zakładamy, że zmienna losowa *X* ma rozkład o nieznanej dystrybuancie $F(x)$. Dysponujemy *n* elementową próbą losową o wartościach $x_1, x_2, ..., x_n$. Zbiór możliwych wartości zmiennej losowej *X* dzielimy na *r* rozłącznych podzbiorów J_k, $k = 1, 2, ..., r$ za pomocą liczb
$-\infty = a_0 < a_1 < ... < a_r = \infty$.

Niech $p_k \left(p_k > 0\right)$ oznacza prawdopodobieństwo, że zmienna losowa *X* przyjmuje wartość z przedziału J_k, tzn.

$$p_k = P\left(X \in J_k\right) = F\left(a_k\right) - F\left(a_{k-1}\right),\ k = 1, 2, ..., r$$

gdzie $F(x)$ jest hipotetyczną dystrybuantą.

Test	Obszar krytyczny	Obszar przyjęcia hipotezy
$\chi^2 = \sum_{k=1}^{r} \frac{\left(N_k - np_k\right)^2}{np_k}$	$\left(\chi^2_\alpha, +\infty\right)$	$\left\langle 0, \chi^2_\alpha \right\rangle$

Liczba np_k jest oczekiwaną liczbą obserwacji n elementowej próbki według założonego rozkładu, które powinny znaleźć się w przedziale J_k,

natomiast N_k jest zmienną losową o wartościach n_k będących liczbą obserwacji, które znalazły się w przedziale J_k.

Dla zadanego poziomu istotności α, z tablic rozkładu χ^2 z ($k - m - 1$) stopniami swobody odczytujemy liczbę χ^2_α (m oznacza liczbę estymowanych parametrów hipotetycznego rozkładu).

Empiryczny współczynnik korelacji i regresja liniowa

Niech (x_1, y_1), (x_2, y_2), …, (x_n, y_n) będą realizacjami zmiennej losowej dwuwymiarowej $(X\ Y)$. Empirycznym współczynnikiem korelacji nazywamy liczbę:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{X} \cdot \overline{Y}}{n \cdot S_X \cdot S_Y}$$

gdzie

$$S_X = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2} \quad \text{i} \quad S_Y = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y}\right)^2}$$

Powyższy współczynnik jest miernikiem siły związku prostoliniowego między dwoma cechami mierzalnymi X i Y.

Bezpośrednio z pojęciem korelacji wiąże się zagadnienie ***regresji.*** Polega ono na znalezieniu takiej linii o równaniu $y = f(x)$, aby suma kwadratów różnic pomiędzy wartościami zaobserwowanymi y_i i obliczonymi $f(x_i)$ była najmniejsza (***metoda najmniejszych kwadratów***). Najprostszą i najczęściej używaną funkcją w regresji jest funkcja liniowa. Mówimy wtedy o regresji linowej. Wtedy zależność między zmiennymi X i Y jest opisana funkcją liniową:

$$y = a \cdot x + b,$$

gdzie

$$a = r \cdot \frac{S_Y}{S_X} \ \text{ i } \ b = \overline{Y} - a \cdot \overline{X}$$

Przykład

Mierzono współzależność między ciśnieniem, a temperaturą dla 10 elementowej próby losowej urządzeń pewnego typu. Wyniki pomiarów przedstawiono w poniższej tabeli

Ciśnienie [hPa]	17	19	20	21	22	24	26	27	27	30
Temperatura [° C]	19	20	23	21	23	23	26	25	26	34

Wyznaczymy najpierw współczynnik korelacji między ciśnieniem, a temperaturą, a następnie równanie regresji liniowej dla tych dwóch zmiennych. Jako zmienną X wzięto ciśnienie, a temperatura to zmienna Y. Stąd parametry poszczególnych zmiennych wynoszą:

$\overline{X} = 23{,}3$, $\overline{Y} = 24$, $S_X = 3{,}95$, $S_Y = 4{,}02$, a suma $\sum_{i=1}^{n} x_i y_i$ wynosi 5735. Wstawiając te wartości do wzoru na empiryczny współczynnik korelacji

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{X} \cdot \overline{Y}}{n \cdot S_X \cdot S_Y}$$

otrzymujemy r = 0,899.

Współczynnik *a* w równaniu regresji liniowej $y = a \cdot x + b$ wynosi: $a = r \cdot \frac{S_Y}{S_X} = 0{,}899 \cdot \frac{4{,}02}{3{,}95} \approx 0{,}915$. Natomiast współczynnik *b* jest równy 2,655.

Stąd równanie regresji ma postać:

$$y = 0{,}915 \cdot x + 2{,}655.$$

Tego typu równania można wykorzystywać do wyznaczania wartości zmiennej Y czyli temperatury. Na przykład dla ciśnienia równego 25 hPa, obliczona wartość temperatury wynosi 25,53.