6. TESTOWANIE HIPOTEZ NIEPARAMETRYCZNYCH

Test zgodności χ^2

```
[h p stat] = chi2gof(x, ...)
```

gdzie: `x`, `h`, `p`, `stat` – jak przy testach parametrycznych, *funkcja niedostępna w Statistics Toolbox 5.0.1.*

Test Kołmogorowa – Smirnowa z poprawką Lillieforsa (weryfikacja hipotezy o normalności rozkładu)

```
[h p stat] = lillietest(x, alfa)
```

gdzie: `x`, `h`, `p`, `stat`, `alfa` – jw.

Test Kołmogorowa – Smirnowa

```
[h p stat] = kstest(x, cdf, alfa, ha)
```

gdzie: `x`, `h`, `p`, `stat`, `alfa` jw.

`ha` – określa postać hipotezy alternatywnej, dozwolone wartości: `'unequal'` – jest wartością domyślną, `'smaller'` i `'larger'` (testy obustronne, lewo i prawo stronne);

`cdf` – dwukolumnowa macierz zawierająca wartości dystrybuanty teoretycznej, pierwsza kolumna zawiera wartości z dziedziny dystrybuanty (jeśli to możliwe to powinny to być wartości z próby), druga kolumna zawiera odpowiadające im wartości dystrybuanty,

parametr można podać jako macierz pustą `[]`, wtedy funkcja weryfikuje hipotezę o zgodności rozkładu z rozkładem normalnym standaryzowanym.

Test niezależności χ^2

```
[tabela chi2 p grupy] = crosstab(x1,x2)
```

gdzie:

`x1`, x2 – wektory zawierające liczby lub znaki reprezentujące wyniki badania, wektor `x1` wypełniony jest wartościami pierwszej analizowanej cechy, wektor `x2` wypełniony jest wartościami cechy drugiej;

`tabela` – tabela krzyżowa z liczebnościami unikalnych kombinacji cechy pierwszej i drugiej;

`chi2` – wartość statystyki testowej χ^2, `p` – wartość *p-value*; `grupy` – tablica komórkowa zawierająca w pierwszej kolumnie unikalne wartości cechy pierwszej, w drugiej kolumnie – wartości cechy drugiej.

Sposób konstrukcji wektorów `x1` i `x2` wyjaśnia następujący przykład: przeprowadzono badania na grupie 5 studentów zapisując dla każdego z nich płeć (‘`k`’ ‘`m`’) i kierunek studiów (‘’`m`’ – matematyka, 'f' – fizyka', 'c' – chemia'). W badanej grupie były 2 studentki matematyki, 1 studentka fizyki, 1 student fizyki i 1 student chemii. Wektory `x1` i `x2` powinny zawierać wyniki badania, wektor `x1` – informacje o płci, wektor `x2` o kierunku studiów. W tym przypadku wektory te należałoby zdefiniować jako: `x1 = ['k';'k';'k';'m';'m']; x2 = ['m';'m';'f';'f';'c'];` tabela krzyżowa zwracana prze funkcję `crosstab` przyjęłaby następująco postać:

Przykład 1.

Zweryfikować na poziomie istotności α = 0.01 hipotezę, że rozkład długości detalu jest rozkładem normalnym. Wyniki pomiarów długości zostały zapisane w funkcji *dane*.

Długość	Liczność
[19, 19.5]	1
[19.5 20]	6
[20 20.5]	18
[20.5 21]	29
[21 21.5]	26
[21.5 22]	12
[22 22.5]	6
[22.5 23]	2

```
alfa = 0.01; x = dane; n=length(x);
xs=mean(x), s=std(x), xd=min(x),xg=max(x),
xs = 20.9622, s = 0.6940,
xd = 19.3295, xg = 22.9906

% liczebności w przedz. klasowych
prz=19:0.5:23; ni=histc(x,prz)'
ni = 1 6 18 29 26 12 6 2 0

% liczebności po poprawce
prz=[-Inf 20:0.5:22 Inf];
ni=histc(x,prz)'; ni=ni(1:length(ni)-1)
ni = 7 18 29 26 12 8

% prawdopodobieństwo teoretyczne
v1=prz(1:length(prz)-1); v2=prz(2:length(prz));
pi = normcdf(v2,xs,s) - normcdf(v1,xs,s)
pi = 0.0828 0.1699 0.2690 0.2591 0.1518 0.0674
% liczebność teoretyczna
npi=n*pi;

% statystyka testowa
chi2_n = sum(((ni-n*pi).^2) ./ (n*pi))
chi2_n = 1.3228
% obszar krytyczny
r=length(ni);
chi2_a = chi2inv(1-alfa, r-2-1)
chi2_a = 11.3439

% p-value
p = 1-chi2cdf(chi2_n, r-2-1)
p = 0.7237
```

wartość statystyki poza obszarem krytycznym: → nie można odrzucić hipotezy H_0

α < p–value → nie można odrzucić hipotezy H_0

Przykład 2.

Wykonano 5 pomiarów długości detalu: -1.0, 0.1, -0.6, 0.5, 1.5. Zweryfikować na poziomie istotności $\alpha = 0.01$ hipotezę, że rozkład długości jest rozkładem normalnym standaryzowanym.

```
X = [-1.0; 0.1; -0.6; 0.5; 1.5]; alfa = 0.01;
```

wersja 1.	*wersja 2.*
	`cdf = [X normcdf(X, 0, 1 )];`
`[h p stat] = kstest(X, [], alfa)`	`[h p stat] = kstest(X, cdf, alfa)`
`h = 0`	`h = 0`
`p = 0.9986`	`p = 0.9986`
`stat = 0.1587`	`stat = 0.1587`

Uwagi.

W wersji 1. drugi parametr w wywołaniu funkcji jest pusty (równy `[]`). Takie wywołanie można wykorzystać jedynie w przypadku gdy weryfikowana jest hipoteza o zgodności rozkładu z rozkładem normalnym standaryzowanym. W wersji 2. w drugim parametrze przekazywana jest, przygotowana wcześniej, macierz zawierająca wartości dystrybuanty teoretycznej. W pierwszej kolumnie macierzy umieszcza się wartości z próby a w drugiej odpowiadające im wartości dystrybuanty teoretycznej – czyli w tym przypadku wartości dystrybuanty rozkładu normalnego standaryzowanego.
W obydwu przypadkach wynik działania funkcji `kstest` jest identyczny, wartość `h = 0` wskazuje, że nie można odrzucić hipotezy o zgodności rozkładów, podobnie wartość p-value `p = 0.9986` większa od założonego poziomu istotności nie pozwala na odrzucenie hipotezy o zgodności rozkładów. Wartość `stat = 0.1587` określa maksymalną odległość pomiędzy dystrybuantami.

Dystrybuantę empiryczną wykreśla w MATLAB-ie polecenie:

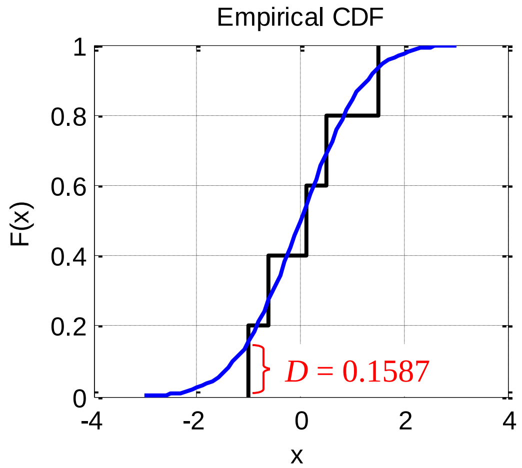
`cdfplot(x),`

gdzie: `x` to wektor zawierający wartości z próby.

Dystrybuantę teoretyczną należy wykreślić posługując się poleceniem `plot`, przygotowując wcześniej odpowiedni zbiór punktów.

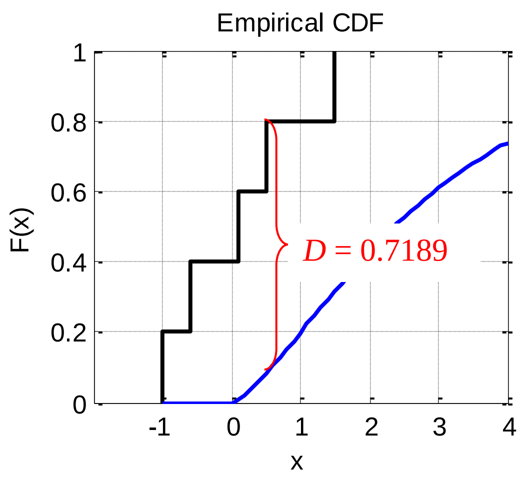
```
cdfplot(X)
x = -3:0.1:3;
y = normcdf(x, 0, 1);
hold on; plot(x,y); hold off
```

Wyznaczona w teście maksymalna odległość pomiędzy dystrybuantami występuje w punkcie `x = -1`.



Dla porównania, dla tych samych wyników pomiarów wykonany został test zgodności rozkładu z rozkładem χ^2 o 3 stopniach swobody.

```
cdf = [X, chi2cdf(X, 3)];
[h p stat]=kstest(X, cdf, 0.01)
h = 1, p = 0.0051, stat = 0.7189

cdfplot(X)
x = -1:0.1:10;
y = chi2cdf(x, 3);
hold on; plot(x,y); hold off
```
.



W tym przypadku hipotezę należy odrzucić, wskazują na to zwracane przez funkcję `kstest` zmienne: `h` i `p`. Wyznaczona w teście maksymalna odległość pomiędzy dystrybuantami występuje w punkcie `x = 0.5`.

Przykład 3.

Pewien produkt może być wytwarzany dwiema różnymi metodami. W celu sprawdzenia czy trwałość zależy od użytej metody produkcji przeprowadzono badania dwóch partii produktów. Ocenę trwałości przeprowadzano poprzez wyznaczenie liczby dni pracy do chwili uszkodzenia produktu, wyniki oceny zostały zebrane w poniższej tabeli. Zweryfikować na poziomie istotności $\alpha = 0.01$ hipotezę, że trwałość produktu zależy od użytej metody produkcji.

Trwałość w dniach	Metoda 1	Metoda 2
40-50	10	10
50-60	30	110
60-70	20	20

Wyniki zilustrować histogramem przedstawiającym zebrane wyniki trwałości produktu dla dwóch badanych metod produkcji.

```
nij = [10 10; 30 110; 20 20]; alfa = 0.01; bar(nij, 'grouped')
```

wersja 1.	*wersja 2.*

`ni = sum(nij, 2), nj = sum(nij, 1)`
`ni =`
`20`
`140`
`40`
`nj = 60   140`
`n = sum(ni)`
`n = 200`
	`d = [repmat('11',nij(1,1),1);`
	`repmat('12',nij(1,2),1);`
	`repmat('21',nij(2,1),1);`
	`repmat('22',nij(2,2),1);`
	`repmat('31',nij(3,1),1);`
	`repmat('32',nij(3,2),1)];`

```
npij = ni * nj / n
npij =
     6    14
    42    98
    12    28
chi2n=sum(sum((nij-npij).^2./npij))
chi2n = 16.3265

r = length(pi), c = length(pj)
r = 3, c = 2
v = (r-1)*(c-1)
v = 2

chi2a = chi2inv(1-alfa, v)
chi2a = 9.2103
p = 1 - chi2cdf(chi2n, v)
p = 2.8493e-004
```

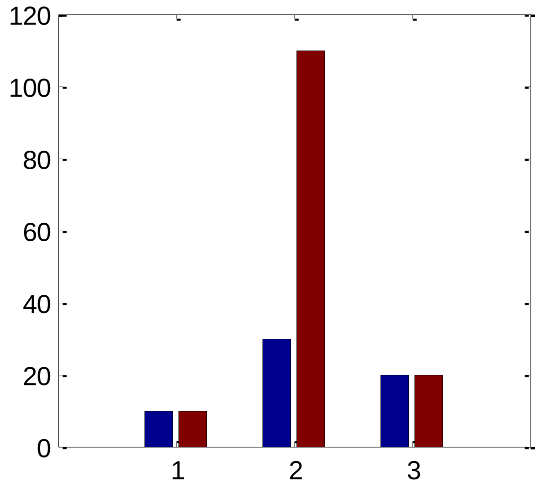
wartość statystyki w obszarze krytycznym: → hipotezę H_0 trzeba odrzucić

```
[tbl,chi2n,p,g] = ...
            crosstab(d(:,1),d(:,2))
tbl =
    10    10
    30   110
    20    20

chi2n = 16.3265

p = 2.8493e-004

g =
    '1'    '1'
    '2'    '2'
      '     []
```

α > p–value → hipotezę H_0 trzeba odrzucić

Uwaga

W wersji 2. dla potrzeb funkcji `crosstab` wygenerowana została pomocnicza tablica `d` z danymi odpowiadającymi liczebnościom empirycznym zawartym w tablicy `nij`. W tablicy tej do zakodowania różnych trwałości produktu zostały użyte kody ‘1’, ‘2’ i ‘3’ a do zakodowania dwóch stosowanych metod produkcji kody ‘1’ i ‘2’. Poprawność utworzonej tablicy potwierdza macierz `tbl` zwracana przez funkcję – zawiera ona liczebności empiryczne `nij`.



Gdyby w zadaniu zamiast liczebności empirycznych podano właściwe dane opisujące trwałości przebadanych produktów to dla potrzeb funkcji `crosstab` należałoby przygotować wektory z zakodowanymi wartościami obydwu badanych cech. Poniżej przedstawiono przykładowe rozwiązanie powyższego problemu, przy założeniu, że do zakodowania trwałości i metody produkcji użyte zostały te same kody co w wersji 2. rozwiązania (`m1` i `m2` to wektory zawierające trwałość produktów wyprodukowanych odpowiednio metodą pierwszą i drugą).

```
M1 = m1; M2 = m2;
M1((m1>=40) & (m1<50)) = 1;
M1((m1>=50) & (m1<60)) = 2;
M1((m1>=60) & (m1<70)) = 3;
M2((m2>=40) & (m2<50)) = 1;
M2((m2>=50) & (m2<60)) = 2;
M2((m2>=60) & (m2<70)) = 3;
```

wektor M1 zawiera kody trwałości produktów wyprodukowanych metodą pierwszą M2 – metodą drugą

```
d1 = [M1; M2];
d2 = [repmat(1,length(m1),1); repmat(2,length(m2),1)];
```

wektor `d1` zawiera kody trwałości wszystkich produktów – na początku wektora umieszczono dane produktów wyprodukowanych metodą pierwszą, na końcu – metodą drugą
wektor `d2` zawiera kody użytych metod produkcji – tyle jedynek ile było elementów w wektorze `m1` i tyle dwójek ile było elementów w wektorze `m2`

```
[tbl,chi2n,p,g] = crosstab(d1, d2)
tbl =
   10    10
   30   110
   20    20

chi2n = 16.3265

p = 2.8493e-004

g =
   '1'    '1'
   '2'    '2'
   '3'     []
```