Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

STATYSTYKA MATEMATYCZNA

z pakietem R

VI. Weryfikacja hipotez, cz. II

Przemysław Grzegorzewski
Konstancja Bobecka-Wesołowska
Marek Gągolewski

Spis treści

1 Wprowadzenie

1.1 Testy dla mediany

Do weryfikacji hipotez dotyczących mediany w populacjach o rozkładach typu ciągłego możemy użyć w `R` funkcji `wilcox.test()`. W przypadku jednej próby za jej pomocą przeprowadzany jest test znakowanych rang Wilcoxona. W przypadku dwóch prób — test Wilcoxona, zwany też testem Manna-Whitneya-Wilcoxona.

1.1.1 Test dla pojedynczej próby

Wektor `x`, zawierający wartości próby, podajemy jako pierwszy agument funkcji *wilcox.test*. Jako drugi — podajemy hipotetyczną wartość mediany (parametr `mu`).

Dla przykładu, chcemy zweryfikować hipotezę $H : \mathrm{Med} = 0$ przeciwko $K : \mathrm{Med} \neq 0$:

```
x <- rcauchy(100)
wilcox.test(x, mu=0)
```

Jeśli próba nie pochodzi z rozkładu, który jest symetryczny, to do weryfikacji hipotez dotyczących wartości mediany zamiast testu znakowanych rang używamy testu znaków, samodzielnie pisząc odpowiedni program.

1.1.2 Test dla dwóch prób

Test Manna-Whitneya-Wilcoxona służy do weryfikacji hipotez o równości median w dwóch populacjach. Stosujemy go w przypadku dwóch prób niezależnych. Wektory `x` i `y`, zawierające wartości dwóch prób, podajemy jako pierwszy i drugi agument funkcji `wilcox.test()`.

Dla ilustracji, chcemy zweryfikować hipotezę $H : \mathrm{Med}_1 = \mathrm{Med}_2$ przeciwko $K : \mathrm{Med}_1 \neq \mathrm{Med}_2$:

```
x <- rcauchy(100, 0, 1)
y <- rcauchy(100, 1, 1)
wilcox.test(x, y)
```

W każdym z powyższych przypadków można określić, czy hipoteza alterantywna ma być dwustronna czy jednostronna, podając jako argument funkcji `wilcox.test()` parametr `alternative="twosided"` (wartość domyślna), `"less"` lub `"greater"`.

1.2 Wykres kwantylowy

Do oceny normalności można wykorzystać wykres kwantylowy, który w `R` tworzony jest za pomocą funkcji `qqnorm`. Na wykresie tym porównujemy kwantyle obliczone dla próbki z kwantylami teoretycznymi rozkładu normalnego. Wektor `x`, zawierający wartości próbki, podajemy jako argument tej funkcji. Za pomocą wywołania metody `qqline(x)`, można do wykresu kwantylowego dorysować prostą przechodzącą przez kwantyle rozkładu teoretycznego. Jeśli nasza próba pochodzi z rozkładu normalego, to punkty na wykresie będą układać się wzdłuż tej prostej.

Na przykład:

```
x <- rnorm(100);
qqnorm(x)
qqline(x, col="red")
```

1.3 Testy zgodności i jednorodności

Do weryfikacji hipotez o zgodności rozkładu badanej cechy z interesującym nas rozkładem, można korzystać w `R` z testu chi-kwadrat (funkcja `chisq.test()`) lub testu Kołmogorowa (funkcja `ks.test()`), przy czym testu Kołmogorowa używamy w przypadku prób pochodzących z populacji o rozkładach typu ciągłego. Fukcja `ks.test()` służy też do przeprowadzania testu Kołmogorowa-Smirnowa o identyczności rozkładów badanej cechy w dwóch populacjach o rozkładach typu ciągłego. W przypadku $k > 2$ populacji o rozkładach typu ciągłego używa się w tym celu testu Kruskala-Wallisa (funkcja `kruskal.test()`). Do porównania rozkładów w wielu populacjach o rozkładach dyskretnych, można użyc funkcji `prop.test()`, za pomocą której przeprowadzany jest test jednorodności chi-kwadrat.

1.3.1 Test chi-kwadrat Pearsona

Test zgodności chi-kwadrat przeprowadza się dla prób pogrupowanych w szereg rozdzielczy. Jako pierwszy argument funkcji `chisq.test()`, podajemy wektor zawierający liczności poszczególnych klas szeregu. Drugim argumentem tej fukcji jest wektor `p`, zawierający wartości hipotetycznych prawdopodobieństw.

Np. chcemy zbadać, czy moneta nie jest sfałszowana. Rzucamy nią 1000 razy i liczymy ile razy wypadł orzeł, a ile razy reszka. Powiedzmy, że otrzymaliśmy 512 orłów i 488 reszek. Test zgodności z rozkładem jednostajnym przeprowadzamy następująco:

```
licznosci <- c(512,488)
pstwa <- (1/2, 1/2)
chisq.test(licznosci, p=pstwa)
```

1.3.2 Test Kołmogorowa

W teście zgodności Kołmogorowa porównuje się wartość dystrybuanty empirycznej, zbudowanej na podstawie próby, z dystrybuantą teoretyczną. Jako pierwszy argument funkcji `ks.test()`, podajemy wektor `x`, zawierający wartości próbki. Drugim argumentem jest nazwa funkcji wyznaczającej dystrybuantę rozkładu, z którym chcemy badać zgodność, np. `"punif"` — gdy testujemy zgodność z rozkładem jednostajnym lub `"pnorm"` — gdy testujemy zgodność z rozkładem normalnym. Jako kolejne argumenty podajemy parametry interesującego nas rozkładu.

Np. badamy zgodność z rozkładem $\mathrm{U}([0,1])$:

```
x <- runif(100)
ks.test(x, "punif")
```

Zbadajmy teraz zgodność z rozkładem normalnym $\mathrm{N}(0{,}5, 1)$:

```
ks.test(x, "pnorm", 0.5, 1)
```

A teraz z wykładniczym $\mathrm{Exp}(2)$:

```
ks.test(x, "pexp", 2)
```

1.3.3 Testy normalności

Do weryfikacji hipotezy o normalności rozkładu badanej cechy można korzystać z obu wymienionych wyżej testów. Jednak istnieją testy zaprojektowane specjalnie w celu badania zgodności z rozkładem normalnym. Jak wspomnieliśmy w poprzedniej części, w `R` dostępne są następujące testy z tej klasy: test Shapiro-Wilka (zalecany; funkcja `shapiro.test()`), test Cramera-von Misesa (`cvm.test()`), test Andersona-Darlinga (`ad.test()`), test Lillieforsa (`lillie.test()`), test chi-kwadrat Pearsona (`pearson.test()`), test Shapiro-Francii (`sf.test()`).

Wektor `x`, zawierający wartości próbki, dla której chcemy przeprowadzić test zgodności z rozkładem normalnym, podajemy jako pierwszy argument wyżej wymienionych funkcji.

1.3.4 Test Kołmogorowa-Smirnowa

W teście Kołmogorowa-Smirnowa porównuje się wartości dystrybuant empirycznych, zbudowanych na podstawie obu prób. Jako pierwszy argument funkcji `ks.test()` podajemy wektor `x`, zawierający wartości pierwszej próbki, a jako drugi — wektor `y`, zawierający wartości drugiej próbki.

1.3.5 Test Kruskala-Wallisa

Test Kruskala-Wallisa służy do weryfikacji hipotezy o identyczności rozkładów badanej cechy w $k > 2$ populacjach. Jako argumenty funkcji `kruskal.test()` podajemy kolejno wektory $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{k}$, zawierające wartości odpowiednich prób.

Dla ilustracji:

```
x <- runif(100)
y <- rexp(100)
z <- rnorm(100)
kruskal.test(x,y,z)
```

1.3.6 Test jednorodności chi-kwadrat

Test jednorodności chi-kwadrat służy do weryfikacji hipotezy o jednakowym rozkładzie badanej cechy w wielu populacjach, w przypadku dykretnym. Przeprowadzany jest na podstawie tablicy częstości pojawień się interesujących nas obserwacji w kolejnych próbach. Jako argumenty fukcji `prop.test()` podajemy kolejno wektory `x` i `y`, przy czym `x` zawiera częstości pojawień się obserwacji w próbach, a `y` — liczności prób.

Np. chcemy porównać prawdopodobieństwa wyrzucenia orła dla trzech monet. Przypuśćmy, że na 100 rzutów pierwszą monetą orzeł wypadł 56 razy, na 95 rzutów drugą — 46 razy, a na 120 rzutów trzecią — 64 razy. Mamy więc:

```
x <- c(56,46,64)
y <- c(100,95,120)
prop.test(x,y)
```

1.4 Test niezależności

Do weryfikacji hipotezy o niezależności dwóch cech populacji służy np. test niezależności chi-kwadrat przeprowadzany na podstawie tablicy korelacyjnej (macierzy kontyngencji).

W `R` dostępny jest on w funkcji `chisq.test()`. Jako argument tej funkcji podajemy odpowiednią macierz kontyngencji.

Gdy liczebności w komórkach macierzy kontyngencji nie są wysokie (< 10), można użyć dokładnego testu Fishera, dostępnego w funkcji `fisher.test()`.

2 Zadania rozwiązane

Zadanie 6.1. Porównaj wykresy kwantylowe dla rozkładu normalnego wygenerowane dla $n = 200$-elementowych próbek z różnych rozkładów. Jak zmienia się kształt wykresu w zależności od typu rozkładu?

Rozwiązanie.
Stworzymy wykresy dla następujących rozkładów.

a) normalnego $\mathrm{N}(0, 1)$,

b) Cauchy'ego $\mathrm{C}(0, 1)$,

c) jednostajnego $\mathrm{U}([0, 1])$,

d) Laplace'a $\mathrm{La}(0, 1)$,

e) Wykładniczego $\mathrm{Exp}(1)$,

f) Ujemnego wykładniczego $\mathrm{NegExp}(1)$.

Rozkład Laplace'a $\mathrm{La}(a, b)$, $b > 0$ dany jest dystrybuantą

$$F(x) = 0{,}5 + 0{,}5\,\mathrm{sgn}\,(x - a)\left(1 - \exp\left(-\frac{|x - a|}{b}\right)\right). \tag{1}$$

Łatwo pokazać, że jeżeli $U \sim \mathrm{U}([-0{,}5, 0{,}5])$, wtedy $a - b\,\mathrm{sgn}(U)\,\ln(1 - 2|U|) \sim \mathrm{La}(a, b)$. Kurtoza zmiennej losowej z rozkładu Laplace'a wynosi 3.

Rozkład $\mathrm{NegExp}(\lambda)$, $\lambda > 0$, dany jest dystrybuantą

$$F(x) = \mathbb{I}_{x\leqslant 0}\left(1 - \exp\left(\lambda x\right)\right). \tag{2}$$

Wygenerujmy próbki z zadanych rozkładów...

```
> n <- 200;
> a <- rnorm(n);
> b <- rcauchy(n);
> c <- runif(n);
> U <- runif(n, -0.5, 0.5);
> d <- -sign(U)*log(1-2*abs(U));
> e <- rexp(n, 1);
> f <- -rexp(n, 1);
```

...i narysujmy stosowne wykresy kwantylowe dla rozkładu normalnego.

```
> par(mfrow=c(3,2)); # 3x2 podwykresy
> qqnorm(a, main="N(0,1)");
> qqline(a);
> qqnorm(b, main="C(0,1)");
> qqline(b);
> qqnorm(c, main="U([0,1])");
> qqline(c);
> qqnorm(d, main="La(0,1)");
> qqline(d);
> qqnorm(e, main="Exp(1)");
> qqline(e);
> qqnorm(f, main="NegExp(1)");
> qqline(f);
```

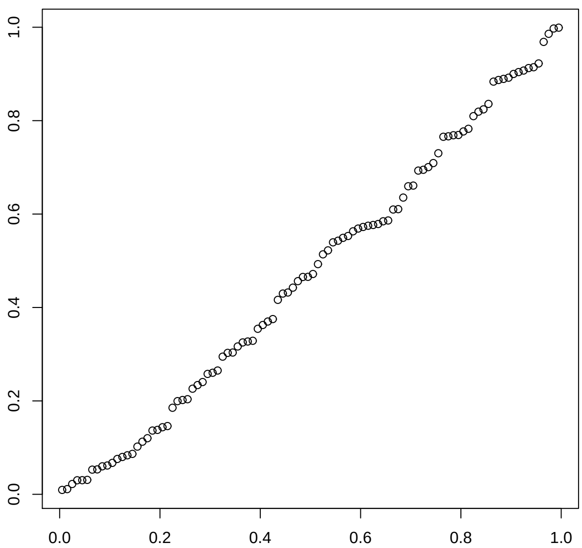
Wnioski pozostawiamy Czytelnikowi (zob. rys. 1).

⊡

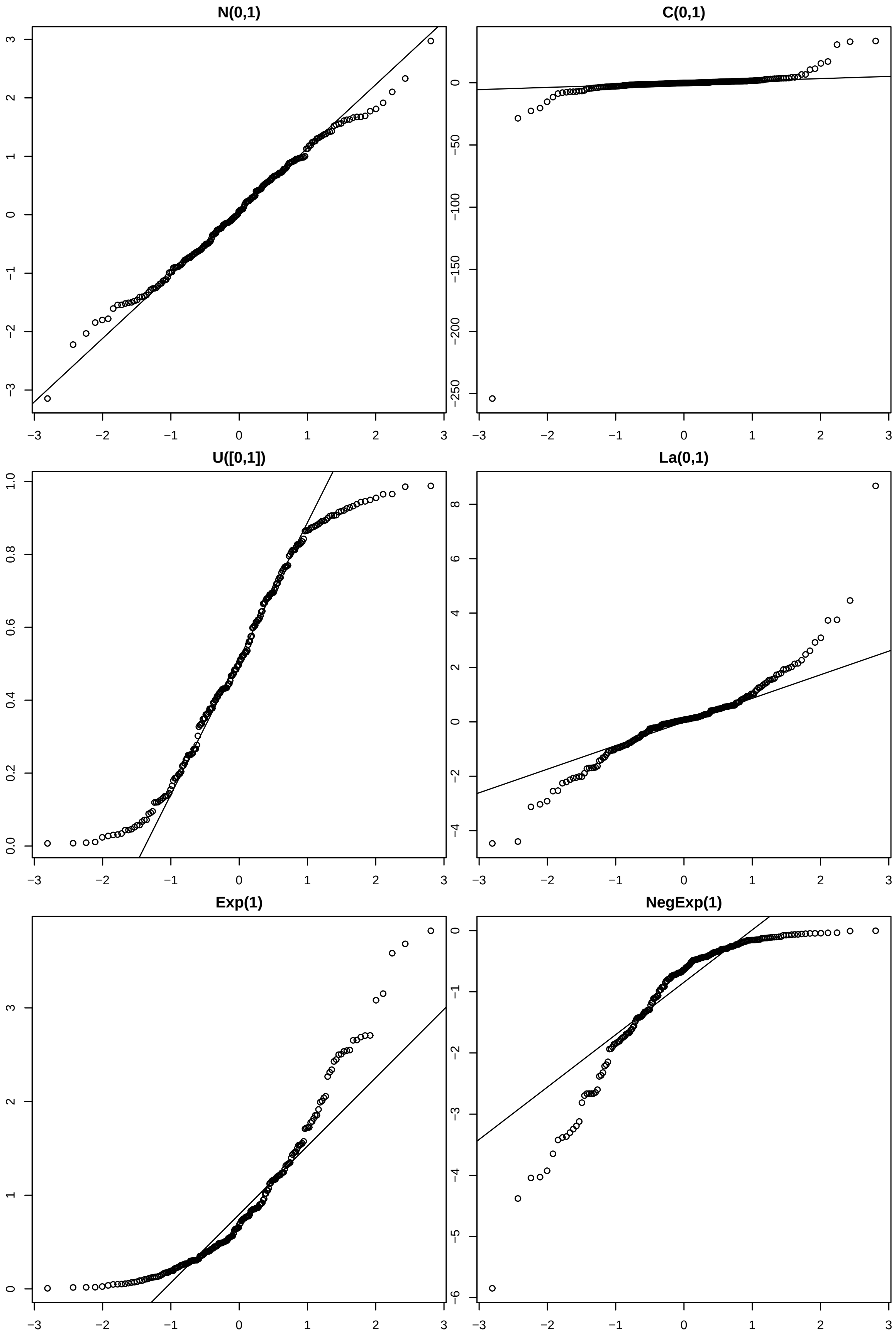
Zadanie 6.2. Utwórz wykresy kwantylowe do porównywania losowych prób z rozkładu jednostajnego i rozkładu wykładniczego z kwantylami teoretycznymi odpowiadających im rozkładów.

Rozwiązanie.

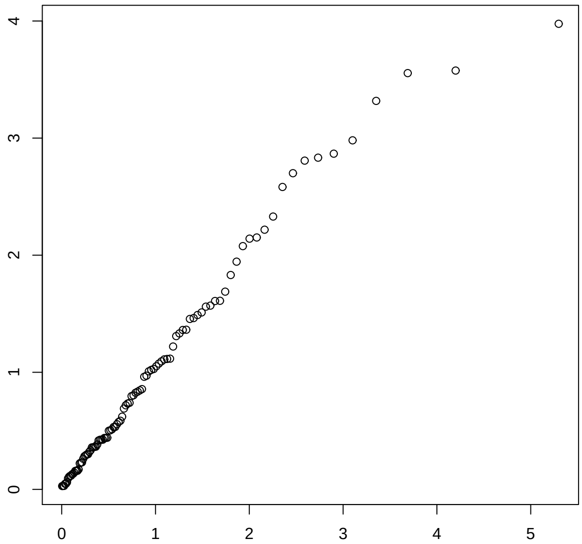
```
>  u <- runif(100);              # losujemy próbę
> qu <- qunif(ppoints(100));     # kwantyle teoretyczne
> qqplot(qu, u);
```




```
>  e <- rexp(100);               # losujemy próbę
> qe <- qexp(ppoints(100));      # kwantyle teoretyczne]
> qqplot(qe, e);
```



Rysunek 1:



□

Zadanie 6.3. Badania grupy krwi 200 osób dały następujące wyniki: grupę O miały 73 osoby, grupę A — 74 osoby, grupę B — 34 osoby, natomiast grupę AB miało 19 osób.

a) Czy na podstawie tych wyników można przyjąć hipotezę o równomiernym rozkładzie wszystkich grup krwi? Przyjmij poziom istotności $\alpha = 0{,}05$.

b) Zweryfikuj hipotezę, że grupa krwi O występuje średnio u 36,7% ludzi, grupa A — u 37,1%, B — u 18,6%, natomiast grupa AB występuje u 7,6% ogółu ludzi. Przyjmij poziom istotności $\alpha = 0{,}05$.

Rozwiązanie.
Sprawdźmy, czy można przyjąć hipotezę H_0 o równomiernym rozkładzie badanej cechy.

```
> krew <- c(73, 74, 34, 19) # wyniki badań
> praw <- rep(0.25, 4)      # testowane prawdopodobieństwa
> chisq.test(krew, p=praw)  # test chi-kwadrat

        Chi-squared test for given probabilities

data:  krew
X-squared = 46.44, df = 3, p-value = 4.572e-10
```

Zatem na poziomie istotności 0,05 odrzucamy H_0 (test chi-kwadrat, $\chi^2 = 46{,}44$, p-value $<$ 0,0001), zatem rozkład grupy krwi nie jest równomierny.

Sprawdźmy zatem zadany rozkład cech.

```
> praw2 <- c(0.367, 0.371, 0.186, 0.076)
> chisq.test(krew, p=praw2)

        Chi-squared test for given probabilities

data:  krew
X-squared = 1.228, df = 3, p-value = 0.7463
```

Interpretację wyników pozostawiamy Czytelnikowi.

$\boxdot$

Zadanie 6.4. Badano, ile zadań rozwiązują studenci w czasie egzaminu ze statystyki. Poniższa tabela wyniki badania przeprowadzonego w grupie 120 losowo wybranych studentów:

Liczba rozwiązanych zadań	0	1	2	3	4
Liczba studentów	10	32	46	26	6

Na poziomie istotności 0,05 zweryfikuj hipotezę, że liczba rozwiązanych zadań ma rozkład dwumianowy $\mathrm{Bin}(4, 0{,}5)$.

Rozwiązanie.

```
> licznosci <- c(10, 32, 46, 26, 6)
> pstwa     <- dbinom(0:4, 4, 0.5) # Pr(X=i); X~Bin(4, 0.5), i=0,1,...,4
> pstwa

[1] 0.0625 0.2500 0.3750 0.2500 0.0625


> chisq.test(licznosci, p=pstwa)

        Chi-squared test for given probabilities

data:  licznosci
X-squared = 1.8222, df = 4, p-value = 0.7684
```

$\boxdot$

Zadanie 6.5. Badano, ile zadań, w ramach przygotowań do egzaminu z rachunku prawdopodobieństwa, rozwiązują w domu studenci, którzy nie zaliczają tego przedmiotu. Poniższa tabelka przedstawia wyniki badania przeprowadzonego w grupie 120 losowo wybranych studentów:

Liczba rozwiązanych zadań	0	1	2	$\geqslant 3$
Liczba studentów	64	30	18	8

Na poziomie istotności 0,05 zweryfikuj hipotezę, że liczba rozwiązanych zadań ma rozkład Poissona.

Rozwiązanie.

```
> zadania <- c(64, 30, 18, 8)
> zadania / sum(zadania)

[1] 0.53333333 0.25000000 0.15000000 0.06666667
```

Estymator ENW parametru w rozkładzie Poissona dla próby $\mathbf{X}$ jest postaci

$$\hat{\lambda} = \bar{\mathbf{X}}. \tag{3}$$

```
> (lambda <- sum(zadania*(0:3))/sum(zadania))

[1] 0.75


> pstwa <- dpois(0:2, lambda);
> pstwa <- c(pstwa, 1-sum(pstwa));
> pstwa

[1] 0.47236655 0.35427491 0.13285309 0.04050544


> chisq.test(zadania, p=pstwa)

        Chi-squared test for given probabilities

data:  zadania
X-squared = 6.9204, df = 3, p-value = 0.07448
```

Zwróćmy uwagę, że formalnie rzecz biorąc, stosowany test nie uwzględnia zmniejszenia o 1 stopni swobody (`df`, ang. *degrees of freedom*), spowodowanego estymacją parametru z próby.

```
> qchisq(0.95, 3) # lewa granica obszaru krytycznego dla df=3

[1] 7.814728


> qchisq(0.95, 2) # lewa granica obszaru krytycznego dla df=2

[1] 5.991465
```

Zatem w tym wypadku raczej jesteśmy skłonni odrzucić hipotezę zerową, głoszącą, że badana cecha ma rozkład Poissona z parametrem 0,75, gdyż wartość statystyki testowej $\chi^2 = 6{,}9204$ wpada do obszaru krytycznego $K_{0{,}05} = [5{,}9915, +\infty)$.
$\boxdot$

Zadanie 6.6. Wygeneruj $n = 20$-elementową próbę losową z rozkładu Cauchy'ego $\mathrm{C}(m_0, 1)$. Za pomocą testu znaków zweryfikuj hipotezę, że mediana rozkładu, z którego pochodzi próba, wynosi $m_0 = 4$. Porównaj otrzymany rezultat z wynikiem testu Wilcoxona.

Rozwiązanie.
Generujemy próbę:

```
> n  <- 20;
> m0 <- 4
> x  <- rcauchy(n, m0)
> median(x)

[1] 4.636915
```

Przeprowadzamy test znaków: Niech $\mathbf{X} = (X_1, .., X_n)$ będzie próbą z rozkładu ciągłego w otoczeniu mediany. Test do weryfikacji hipotezy $H : \mathrm{Med}_{\mathbf{X}} = m_0$ przeciw alternatywie $K : \mathrm{Med}_{\mathbf{X}} \neq m_0$ jest postaci. Statystyka testowa:

$$R = \sum_{i=1}^{n} \mathbb{I}_{X_i > m_0}. \tag{4}$$

Obszar krytyczny:

$$K_\alpha = \left[0, r_{\alpha/2}\right] \cup \left[n - r_{\alpha/2}, n\right], \tag{5}$$

gdzie $r_{\alpha/2}$ jest kwantylem rzędu $\alpha/2$ rozkładu $Bin(n, 0{,}5)$.

W naszym zadaniu: $n = 20$, $\alpha = 0{,}05$, $m_0 = 4$. Wyznaczamy wartość statystyki testowej:

```
> (Z <- sum(x>m0))

[1] 13


> alfa <- 0.05;
> (r <- qbinom(alfa*0.5, n, 0.5));

[1] 6
```

Zatem obszar krytyczny jest następujący: $K_{0,05} = [0, 6] \cup [14, 20]$. $Z \notin K_{0,05}$, więc H nie odrzucamy.

Dla porównania, zastosujmy test rangowanych znaków Wilcoxona:

```
> wilcox.test(x, mu=4)

        Wilcoxon signed rank test

data:  x
V = 125, p-value = 0.4749
alternative hypothesis: true location is not equal to 4
```

$\boxdot$

Zadanie 6.7. Zmierzono czas trwania siedmiu rozmów telefonicznych i otrzymano następujące dane (w minutach): 2.5, 1.8, 6.0, 0.5, 8.75, 1.2, 3.75. Na poziomie istotności $\alpha = 0{,}01$ zweryfikuj hipotezę, że czas trwania rozmowy ma rozkład wykładniczy o wartości średniej 4 minuty.

Rozwiązanie.
Do weryfikacji hipotezy, że czas trwania rozmowy ma rozkład wykładniczy o wartości średniej 4 minuty (tzn. z parametrem $\lambda = 1/4$), wykorzystamy test zgodności Kołmogorowa:

```
> telefony <- c(2.5, 1.8, 6, 0.5, 8.75, 1.2, 3.75)
> ks.test(telefony, "pexp", 0.25)
```

```
        One-sample Kolmogorov-Smirnov test

data:  telefony
D = 0.1175, p-value = 0.9997
alternative hypothesis: two-sided
```

⊡

Zadanie 6.8. Na podstawie danych dotyczących parametrów kilku wybranych marek samochodów (plik `samochody.csv`), zweryfikuj hipotezę o jednakowym rozkładzie zużycia paliwa przez samochody produkowane w USA i w Japonii (wykorzystaj zmienne `mpg` i `producent`). Przyjmij poziom istotności $\alpha = 0{,}05$.

Rozwiązanie.

```
> auta <- read.csv("http://www.ibspan.waw.pl/~pgrzeg/stat_lab/samochody.csv",
     head=T, dec=",", sep=";");
> head(auta) # czy ok?

   mpg cylindry moc przysp rok waga producent      marka      model cena
1 43.1        4  48   21.5  78 1985         2 Volkswagen Rabbit Dl  2400
2 36.1        4  66   14.4  78 1800         1 Ford       Fiesta     1900
3 32.8        4  52   19.4  78 1985         3 Mazda      GLC Deluxe 2200
4 39.4        4  70   18.6  78 2070         3 Datsun     B210 GX    2725
5 36.1        4  60   16.4  78 1800         3 Honda      Civic CVCC 2250
6 19.9        8 110   15.5  78 3365         1 Oldsmobile Cutlass    3300
    legenda
1 America=1
2  Europe=2
3  Japan =3
4
5
6
```

Do weryfikacji hipotezy o jednakowym rozkładzie zużycia paliwa przez samochody produkowane w USA i w Japonii użyjemy testu Kołmogorowa-Smirnowa:

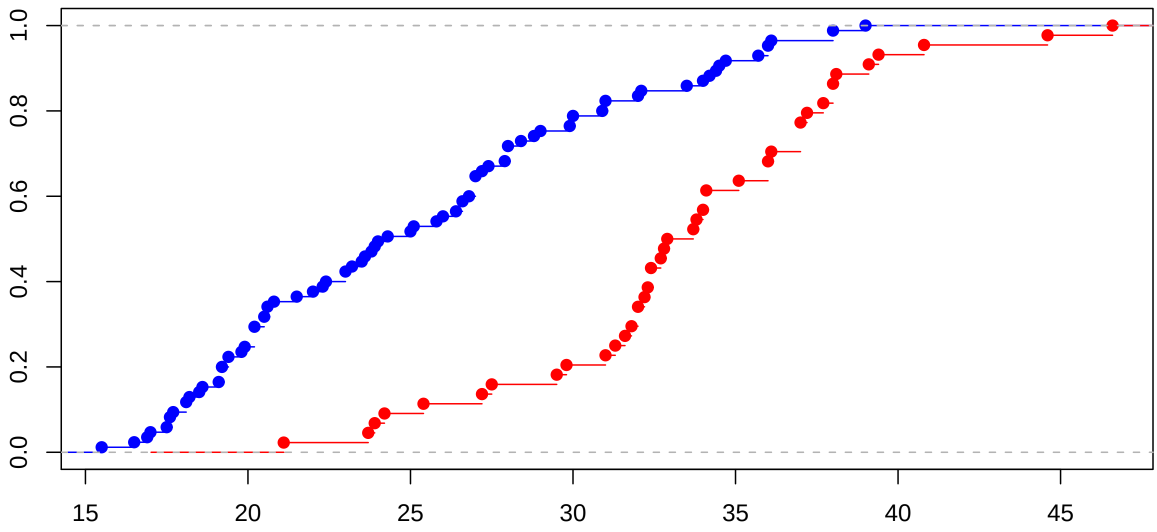
```
> mpga <- auta$mpg[auta$producent == 1]
> mpgj <- auta$mpg[auta$producent == 3]
> ks.test(mpga, mpgj)

        Two-sample Kolmogorov-Smirnov test

data:  mpga and mpgj
D = 0.5963, p-value = 2.229e-09
alternative hypothesis: two-sided
```

Porównajmy jeszcze wykresy dystrybuant empirycznych dla obu prób:

```
> plot(ecdf(mpga), xlim=range(c(mpga,mpgj)), main="", col="blue")
> plot(ecdf(mpgj), add=T, col="red")
```



Nadto, przeprowadźmy test sum rang Wilcoxona (Manna-Whitney'a):

```
> wilcox.test(mpga, mpgj)

        Wilcoxon rank sum test with continuity correction

data:  mpga and mpgj
W = 607, p-value = 3.547e-10
alternative hypothesis: true location shift is not equal to 0
```

⊡

Zadanie 6.9. W celu zbadania, czy nowy rodzaj paliwa lotniczego ma istotny wpływ na zasięg lotu pewnego samolotu sportowego, wykonano 10 pomiarów zasięgu dla samolotów napędzanych stosowanym dotąd paliwem oraz 10 pomiarów dla samolotów zasilanych nowym paliwem. Otrzymano następujące wyniki (w km):

Stosowane dotąd paliwo	1039, 1168, 1008, 1035, 1035, 1025, 1059, 1012, 1212, 1039
Nowy rodzaj paliwa	1096, 1161, 1210, 1088, 1154, 1111, 1103, 1094, 1059, 1177

Czy na podstawie tych danych można stwierdzić, że nowy rodzaj paliwa lotniczego ma istotny wpływ na wzrost przeciętnego zasięgu samolotu? Przyjmij poziom istotności 0,05.

Rozwiązanie.
Jesteśmy oczywiście zainteresowani doborem najmocniejszego testu. Zweryfikujmy hipotezy o normalności obydwu rozkładów:

```
> stare <- c(1039, 1168, 1008, 1035, 1035, 1025, 1059, 1012, 1212, 1039);
> nowe  <- c(1096, 1161, 1210, 1088, 1154, 1111, 1103, 1094, 1059, 1177);
> shapiro.test(stare);

        Shapiro-Wilk normality test

data:  stare
W = 0.7223, p-value = 0.001641
```

```
> shapiro.test(nowe);

        Shapiro-Wilk normality test

data:  nowe
W = 0.9355, p-value = 0.5043
```

Jako że jedna z prób nie spełnia tego założenia, nie możemy skorzystać z testu t dla średnich w dwóch próbach niezależnych. Pozostaje nam użyć znanego testu nieparametrycznego dla median, odpowiednio, m_1 i m_2:

```
> mean(stare); mean(nowe);

[1] 1063.2

[1] 1125.3


> wilcox.test(stare, nowe, alternative="less")

        Wilcoxon rank sum test with continuity correction

data:  stare and nowe
W = 18.5, p-value = 0.009487
alternative hypothesis: true location shift is less than 0
```

Test sum rang Wilcoxona sugeruje odrzucenie ($\alpha = 0{,}05$) hipotezy zerowej $H : m_1 = m_2$ na rzecz $K : m_1 < m_2$ ($W = 18{,}5$, p-value $= 0{,}00949$).

$\boxdot$

Zadanie 6.10. W celu porównania trzech metod nauki stenografii, przeprowadzono sprawdzian na losowych próbach osób szkolonych poszczególnymi metodami. Otrzymano następujące wyniki:

Metoda A	147, 188, 162, 144, 157, 179, 165, 180
Metoda B	153, 161, 157, 155, 163, 160, 154
Metoda C	173, 152, 194, 186, 166, 194, 178, 192, 186

Zbadaj, czy te trzy metody są tak samo efektywne. Przyjmij poziom istotności $\alpha = 0{,}05$.

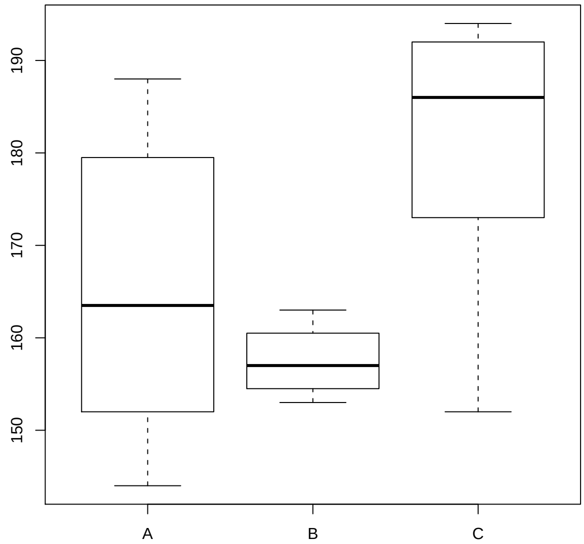
Rozwiązanie.
Do weryfikacji hipotezy o jednakowej efektywności metod A, B i C użyjemy testu sum rang Kruskala-Wallisa:

```
> A <- c(147, 188, 162, 144, 157, 179, 165, 180);
> B <- c(153, 161, 157, 155, 163, 160, 154);
> C <- c(173, 152, 194, 186, 166, 194, 178, 192, 186);
> kruskal.test(list(A,B,C))

        Kruskal-Wallis rank sum test

data:  list(A, B, C)
Kruskal-Wallis chi-squared = 7.7436, df = 2, p-value = 0.02082
```

```
> boxplot(list(A,B,C), names=c("A", "B", "C"))
```



⊡

Zadanie 6.11. W celu zbadania, czy istnieje związek pomiędzy dochodem i posiadanym wykształceniem, przeprowadzono badanie na 450 osobowej próbie losowej i otrzymano następujące wyniki:

	Roczny dochód (tys. zł.)		
	< 120	$120 - 250$	> 250
Wykształcenie wyższe	80	115	55
Brak ukończonych studiów	95	70	35

Zweryfikuj odpowiednią hipotezę na poziomie istotności $\alpha = 0{,}01$.

Rozwiązanie.
Do weryfikacji hipotezy o niezależności interesujących nas cech użyjemy testu niezależności chi-kwadrat. Wymaga on zbudowania macierz kontyngencji:

```
> ww <- c(80, 115, 55)
> bw <- c(95, 70, 35)
> ct <- rbind(ww, bw) # funkcja rbind tworzy tabelę łącząc wiersze
> ct

   [,1] [,2] [,3]
ww   80  115   55
bw   95   70   35


> chisq.test(ct)

        Pearson's Chi-squared test

data:  ct
X-squared = 11.2596, df = 2, p-value = 0.003589
```

Zatem możemy przyjąć, że wykształcenie i dochód nie są niezależne ($\alpha = 0{,}01$).
$\boxdot$

Zadanie 6.12. Na podstawie danych dotyczących parametrów kilku wybranych marek samochodów (plik `samochody.csv`) stwierdź, czy istnieją istotne różnice w mocy silników samochodów produkowanych w USA, w Europie i w Japonii (wykorzystaj zmienne moc i producent). Przyjmij poziom istotności $\alpha = 0{,}05$.

Rozwiązanie.

```
> auta <- read.csv("http://www.ibspan.waw.pl/~pgrzeg/stat_lab/samochody.csv",
     head=T, dec=",", sep=";");
```

Stwórzmy tabelę o kolumnach: `moc` i `producent`:

```
> ms <- data.frame(moc=auta$moc, prod=auta$producent)
```

Obliczymy medianę dla próby, której wartości mamy w kolumnie `moc`, a następnie policzymy, ile samochodów amerykańskich, europejskich i japońskich ma moc większą niż ta mediana (zapiszemy te informacje w wektorze `pm`). Obliczymy też, ile samochodów amerykańskich, europejskich i japońskich było w naszej próbie (wektor `n123`).

```
> ms <- na.omit(ms) # usuwamy rekordy z~brakującymi informacjami
> (med  <- median(ms[,1]))

[1] 85


> ms <- cbind(ms, wiekszy=(ms$moc>med));  # dodajemy kolumnę wskazującą, czy moc>mediana
> head(ms); # podgląd

  moc prod wiekszy
1  48    2   FALSE
2  66    1   FALSE
3  52    3   FALSE
4  70    3   FALSE
5  60    3   FALSE
6 110    1    TRUE


> pm   <- rep(0,3);
> for (i in 1:3) pm[i] = sum(ms$wiekszy[ms$prod==i]);
> pm

[1] 54  6 13


> (n123 <- table(ms$prod))  # zliczamy samochody wg regionu pochodzenia

 1  2  3
83 24 44
```

Teraz możemy przeprowadzić test jednorodności chi-kwadrat:

```
> prop.test(pm,n123)
```

```
        3-sample test for equality of proportions without continuity
        correction

data:  pm out of n123
X-squared = 20.7509, df = 2, p-value = 3.119e-05
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.6506024 0.2500000 0.2954545
```

Możemy również użyć testu niezależności chi-kwadrat. W tym celu trzeba stworzyć odpowiednią tablicę kontyngencji:

```
>  lm <- n123-pm
> (zz <- rbind(pm, lm))

    1  2  3
pm 54  6 13
lm 29 18 31


> chisq.test(zz)

        Pearson's Chi-squared test

data:  zz
X-squared = 20.7509, df = 2, p-value = 3.119e-05
```

⊡

Zadanie 6.13. Spośród studentów czterech wydziałów, na których pan Iksiński wykłada najciekawszy przedmiot świata[1], pobrano próbki losowe i zliczono studentów (zwanych dalej „szczęśliwcami"), którym udało się zdać egzamin z tego przedmiotu. Wyniki zamieszczono w poniższej tabeli:

Wydział	Liczność próbki	Liczba szczęśliwców
Nauk niepotrzebnych	206	61
Mniemanologii stosowanej	164	34
Nauk ciekawych	98	38
Nauk przydatnych	102	35

Czy w świetle zebranych danych można stwierdzić, że występują istotne różnice między odsetkami osób na poszczególnych wydziałach, które zdały statystykę? Przyjmij poziom istotności $\alpha = 0{,}05$.

Rozwiązanie.

Stostujemy test jednorodności chi-kwadrat.

```
> prop.test(c(61,34,38,35), c(206,164,98,102))
```

[1]Osobom niezorientowanym w sprawach tego świata wyjaśniamy, że mowa tu oczywiście o statystyce matematycznej.

```
        4-sample test for equality of proportions without continuity
        correction

data:  c(61, 34, 38, 35) out of c(206, 164, 98, 102)
X-squared = 11.2601, df = 3, p-value = 0.0104
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.2961165 0.2073171 0.3877551 0.3431373
```

⊡

3 Zadania do rozwiązania

Zadanie 6.14. Dla 200 próbek betonu przeprowadzono badanie wytrzymałości na ściskanie i otrzymano wyniki (w MPa):

Wytrzymałość	Liczba próbek
19 – 20	10
20 – 21	26
21 – 22	56
22 – 23	64
23 – 24	30
24 – 25	14

Zweryfikuj hipotezę głoszącą, że wytrzymałość na ściskanie ma rozkład normalny. Przyjmij poziom istotności $\alpha = 0{,}05$.

Zadanie 6.15. Na podstawie danych zawartych w pliku `samochody.csv`, zweryfikuj przypuszczenie, że rozkład przyspieszenia samochodów o wadze 2500–3000 funtów jest normalny (wykorzystaj zmienne `przysp` i `waga`). Czy można twierdzić, że przeciętne przyspieszenie tych samochodów przekracza 15? Przyjmij poziom istotności 0,01.

Zadanie 6.16. Psycholog pracujący w poradni rodzinnej zebrał dane dotyczące powodów kryzysów małżeńskich, które wymieniane były przez przychodzące do poradni pary. Dane te, zamieszczone w poniższej tabeli, pokazują źródła kryzysu postrzegane przez każde z małżonków.

Żona \ Mąż	Pieniądze	Dzieci	Zainteresowania	Inne
Pieniądze	86	31	132	19
Dzieci	17	64	43	13
Zainteresowania	54	39	132	33
Inne	30	17	37	54

Czy na podstawie zebranych danych można stwierdzić, że istnieje zależność poglądów mężów i żon co do przyczyn kryzysu w ich małżeństwach? Przyjmij poziom istotności $\alpha = 0{,}05$.

Zadanie 6.17. Badano istnienie związku między ciśnieniem krwi a nadwagą. W poniższej tabeli zebano dane na temat losowo wybranej grupy osób:

	Ciśn. ++	Ciśn. OK
Nadwaga	57	18
Brak nadwagi	24	91

Czy na podstawie tych danych można stwierdzić istnienie takiej zależności? Przyjmij poziom istotności 0,05.

Zadanie 6.18. Badano, czy istnieje zależność między zawodami ojców i ich dorosłych synów. W tym celu zbadano losowo wybraną grupę ojców i synów. Otrzymano następujące wyniki:

Ojciec \ Syn	Polityk	Prawnik	Lekarz
Polityk	33	48	17
Prawnik	21	38	12
Lekarz	7	8	68

Na podstawie powyższych danych stwierdzić, czy istnieje taka zależność. Przyjmij poziom istotności testu 0,01.

4 Wskazówki i odpowiedzi

Wskazówka do zadania 6.14. Wyestymuj parametry μ i σ rozkładu $\mathrm{N}(\mu, \sigma)$ z próby. Użyj testu zgodności chi-kwadrat. Rozpatrz następujące klasy wartości wytrzymałości: $(-\infty, 20], (20, 21], (21, 22], \ldots, (25, +\infty)$. Jak wyznaczyć np. $P(20 < X \leqslant 21)$, gdzie $X \sim \mathrm{N}(\mu, \sigma)$?