Testy dla dwóch prób w rodzinie rozkładów normalnych

dr Mariusz Grządziel

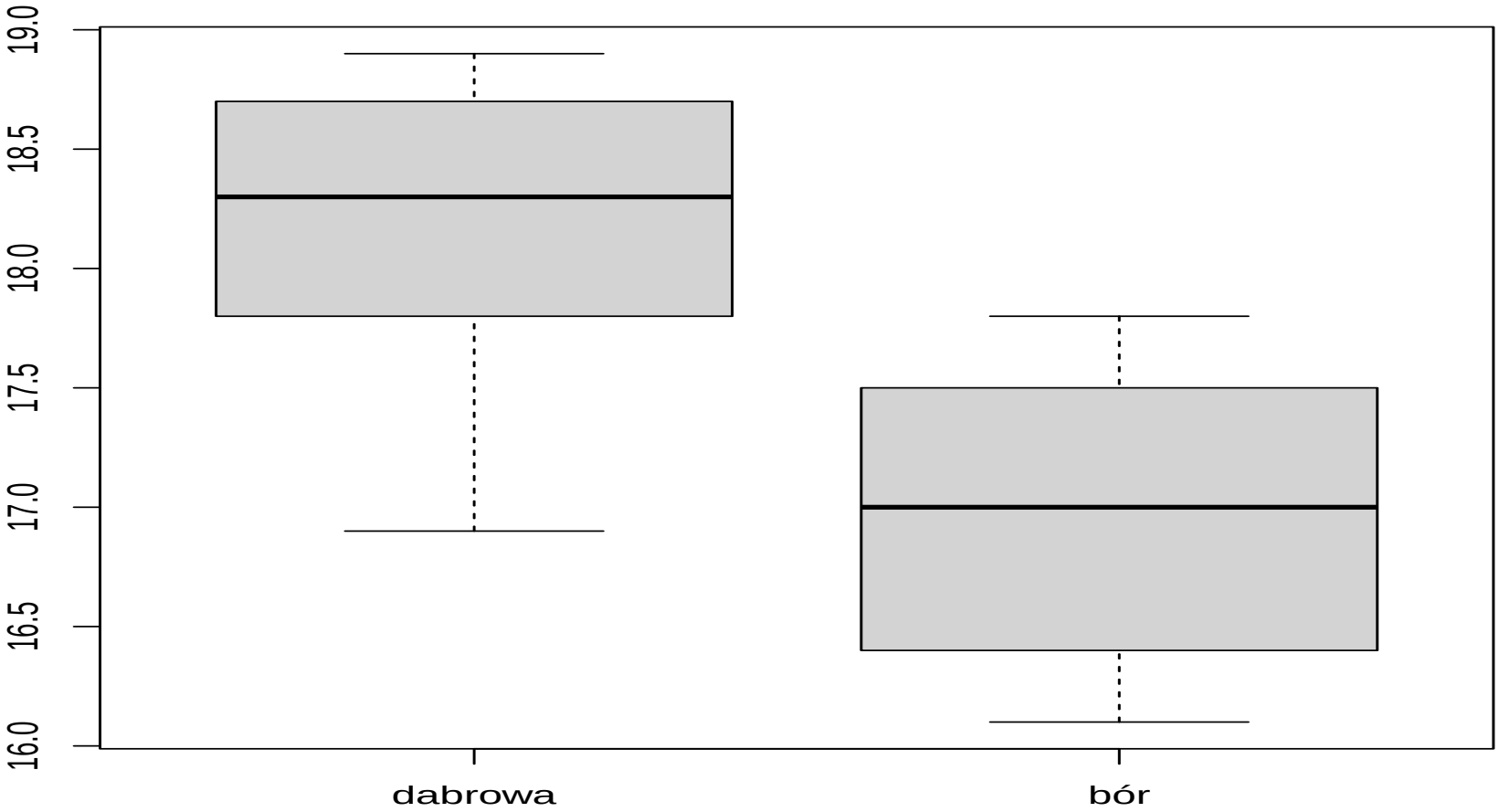
Katedra Matematyki, Uniwersytet Przyrodniczy we Wrocławiu

r. akad. 2022/2023

Przykład

Rozważamy dane (podobne do danych z przykładu 7.2 z książki A. Łomnickiego)
$n_1 = 9$ poletek w dąbrowie, $n_2 = 10$ poletek w borze
Liczby pająków (na poletkach) w dąbrowie : 40, 57, 37, 61, 52, 49, 50, 68, 48;
Liczby pająków (na poletkach) w borze: 46, 52, 39, 34, 46, 31, 37, 26, 52, 22.
Chcielibyśmy zweryfikować hipotezę o równości liczby pająków (przypadających na jednostkę pola powierzchni) w borze i dąbrowie.

Wykres ramkowy



Formalizacja problemu

Zakładamy,że $x_1, x_2, \ldots, x_{n_1}$, liczby pająków na poletkach w dąbrowie, są realizacjami próby prostej $X_1, X_2, \ldots, X_{n_1}$,

$$X_1, X_2, \ldots, X_{n_1}, \quad X_i \sim N(\mu_1, \sigma_1), \quad i = 1, 2, \ldots, n_1$$

oraz że $y_1, y_2, \ldots, y_{n_2}$ są realizacjami próby prostej

$$Y_1, Y_2, \ldots, Y_{n_2}, \quad Y_i \sim N(\mu_2, \sigma_2), \quad i = 1, 2, \ldots, n_2.$$

W naszym przykładzie :

$$x_1 = 40, \ldots, y_1 = 46, \ldots.$$

Chcemy zweryfikować hipotezę

$$H_0 : \mu_1 = \mu_2 \textit{ przeciwko hipotezie } H_1 : \mu_1 \neq \mu_2$$

Założenie normalności

```
x=c(40, 57, 37, 61, 52, 49, 50, 68, 48) # dąbrowa

y=c(46, 52, 39, 34, 46, 31, 37, 26, 52, 22) # bór

> shapiro.test(x)

Shapiro-Wilk normality test

data:  x
W = 0.9734, p-value = 0.9222

>
> shapiro.test(y)

Shapiro-Wilk normality test

data:  y
W = 0.94698, p-value = 0.6329
```

Weryfikacja hipotezy o równości wariancji

Przy założeniu, że: $\sigma_1 = \sigma_2$ (warunek "jednorodności wariancji" jest spełniony) moglibyśmy zastosować test Studenta dla jednorodnych wariancji dla prównania dwóch średnich. Ma on wiele zalet (z punktu widzenia teorii testowania hipotez). Dlatego pierwszym krokiem powinno być sprawdzenie hipotezy o równości wariancji:

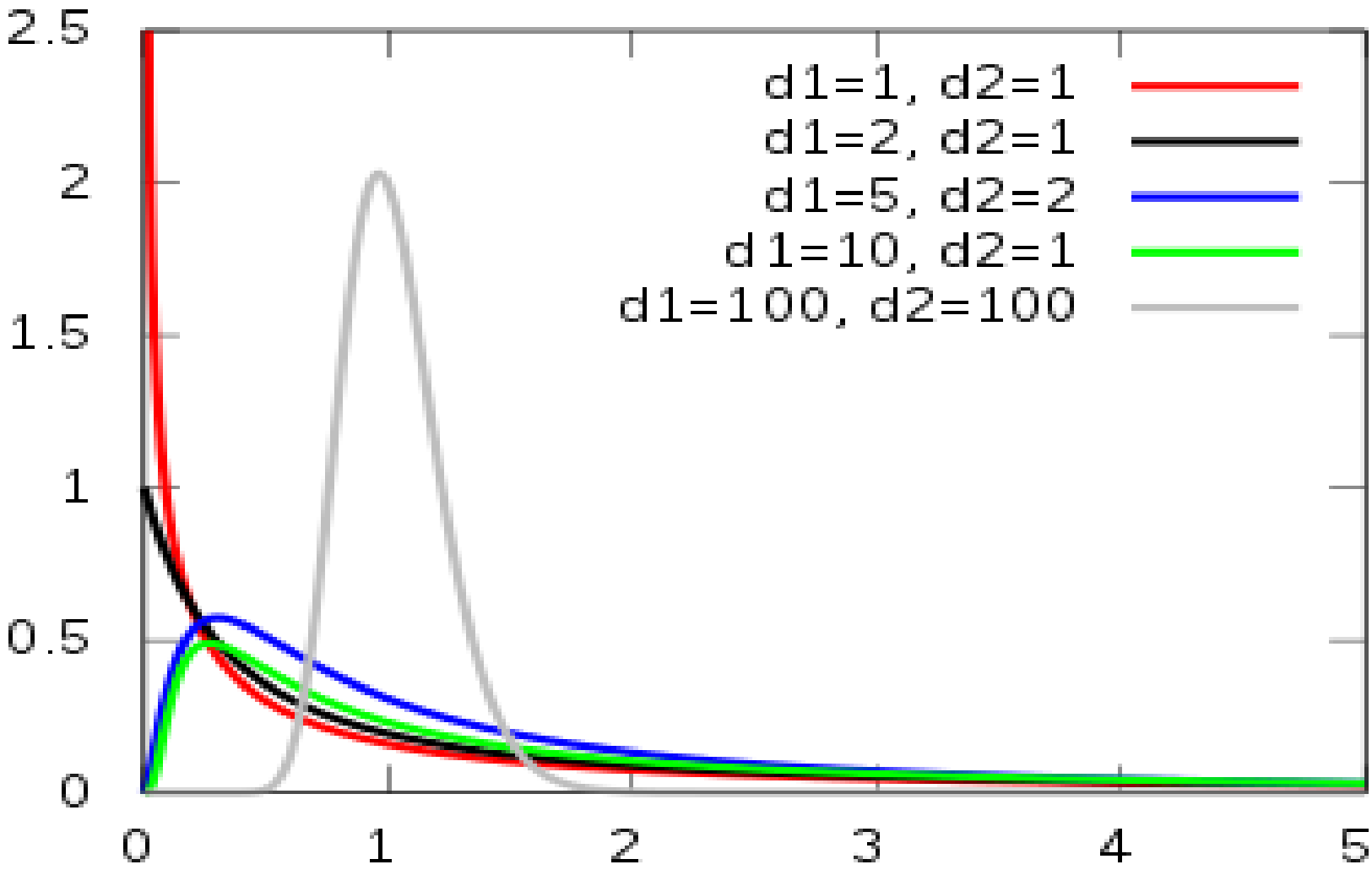
$$H_0^j : \sigma_1^2 = \sigma_2^2 \text{ przeciwko } H_1^j : \sigma_1^2 \neq \sigma_2^2.$$

Odpowiedną statystyką testową dle weryfikacji H_0^j przeciwko H_1^j jest

$$F = \frac{\mathcal{S}_1^2}{\mathcal{S}_2^2}$$

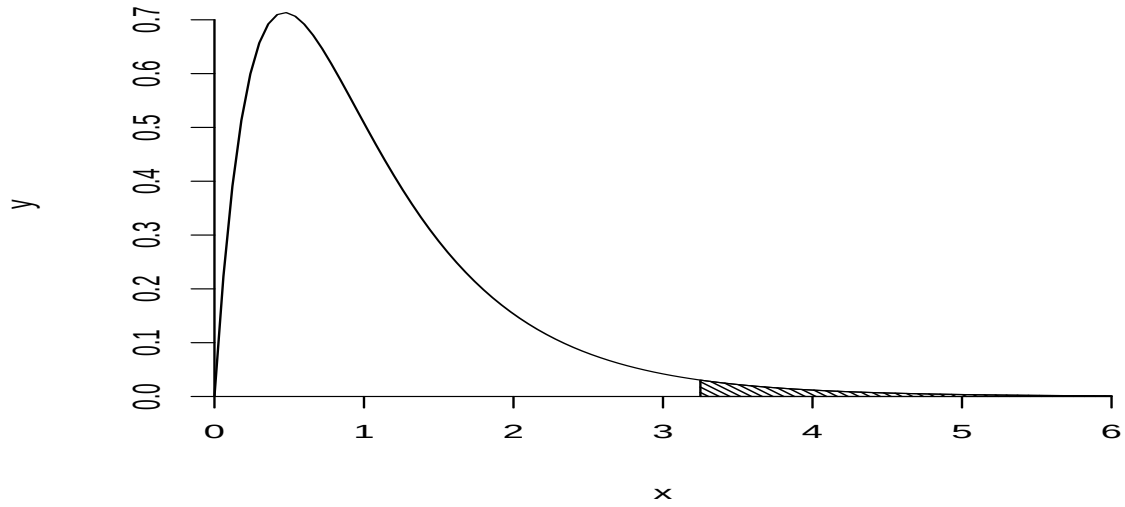
gdzie $\mathcal{S}_1^2$ jest wariancją próbkową dla pierwszej próby, $\mathcal{S}_2^2$ jest wariancją próbkową dla drugiej próby. Można pokazać, że F ma rozkład F_{n_1-1,n_2-1} (rozkład F-Snedecora z $n_1 - 1$ i $n_2 - 1$ stopniami swobody).

Gęstości rozkładów F Snedecora



Źródło: Artykuł „Rozkład F-Snedecora” w Wikipedii; lic. CC BY-SA 4.0 (Uzn. Autorstwa — Na Tych Samych Warunkach)

Rozkład F-Snedecora z liczbami st. sw. 4 i 30



Rysunek: Wykres gęstości rozkładu rozkładu $F_{4,30}$; obszar zakreskowany odpowiada wartościom większym niż $F_{0,975;4;30} = 3{,}25$, kwantyl rzędu 0,975 tego rozkładu.

Funkcja gęstości rozkładu F_{n_1,n_2}: patrz Koronacki i Mielniczuk (2001, str. 208).

Hipoteza o równości wariancji — obszar krytyczny

Obszar krytyczny, dla poziomu istotności α, ma postać

$$[0, F_{\alpha/2;n_1,n_2}] \cup [F_{1-\alpha/2,n_1,n_2}, \infty).$$

Weryfikacja

$$H_0^j : \sigma_1^2 = \sigma_2^2 \text{ przeciwko } H_1^j : \sigma_1^2 \neq \sigma_2^2.$$

może być również oparta na statystyce

$$F' = \frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)};$$

obszar krytyczny (odpowiadający F') ma postać:

$$[F_{1-\alpha/2,n_1,n_2}, \infty)$$

Przykład — obliczenia

$$s_1^2 = 94{,}5; \quad s_2^2 = 91{,}61111;$$

widzimy, że $s_1^2 \geqslant s_2^2$;
f', realizacja statystyki testowej F', jest równa

$$f' = \frac{s_1^2}{s_2^2} = 1{,}031534.$$

Wartość kwantyla rzędu 0,975 rozkładu $f_{8,9}$ (oznaczmy ją przez $f_{0,975,8,9}$) jest równa

$$f_{0,975,8,9} = 4{,}101956,$$

wartość statystyki testowej nie należy do obszaru krytycznego $[4{,}101956, \infty)$, więc nie ma podstaw do odrzucenia hipotezy H_0^j.

Hipoteza o równości wariancji: środowisko R

```
x=c(40, 57, 37, 61, 52, 49, 50, 68, 48)
y=c(46, 52, 39, 34, 46, 31, 37, 36, 52, 22)
var.test(x,y)

   F test to compare two variances

data:  x and y
F = 1.0315,
num df = 8, denom df = 9, p-value = 0.9541
alternative hypothesis: true ratio of variances
is not equal to 1
95 percent confidence interval:
0.2514738 4.4946352
sample estimates:
ratio of variances
1.031534
```

Testowanie hipotezy o równości średnich — przypadek jednorodnych wariancji

Jeśli $X_1, X_2, \ldots, X_{n_1}$ jest próbą prostą, $X_i \sim N(\mu_1, \sigma)$, a $Y_1, Y_2, \ldots, Y_{n_2}$ jest również próbą prostą , $Y_i \sim N(\mu_2, \sigma)$, (zauważ, że wariancje obu rozkładów, z których pochodzą próby, są równe!), wtedy można do weryfikacji hipotezy

$$H_0 : \mu_1 = \mu_2 \quad \text{przeciwko} \quad H_1 : \mu_1 \neq \mu_2$$

można zastosować "klasyczny" test t-Studenta dla dwóch prób. Statystyka testowa ma postać:

$$T = \frac{\bar{X} - \bar{Y}}{S_{X,Y}},$$

gdzie

$$\mathcal{S}_{X,Y} = \sqrt{\frac{(n_1 - 1)\mathcal{S}_1^2 + (n_2 - 1)\mathcal{S}_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

$\mathcal{S}_1^2$ oznacza wariancję zmiennych X_i, $i = 1, \ldots, n_1$, $\mathcal{S}_2^2$ oznacza wariancję zmiennych Y_i, $i = 1, \ldots, n_2$; T ma rozkład t-Studenta z liczbą stopni swobody $n_1 + n_2 - 2$.

Obszar krytyczny, dla poziomu istotności $\alpha \in (0, 1)$ ma postać

$$(-\infty, -t_{1-\alpha/2, n_1+n_2-2}] \cup [t_{1-\alpha/2, n_1+n_2-2}, \infty).$$

Przyjmujemy $\alpha = 0{,}05$ (poziom istotności testowania hipotezy);

$$
\begin{aligned}
n_1 &= 9, \quad n_2 = 10 \\
\bar{x} &= 51{,}33333 \quad \bar{y} = 39{,}5 \\
s_1^2 &= 94{,}5, \quad s_2^2 = 91{,}61111 \\
s_{x,y}^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \times \frac{n_1+n_2}{n_1 n_2} = \\
&= \frac{8 \times 94{,}5 + 9 \times 91{,}61111}{9+10-2} \times \frac{9+10}{9 \times 10} = \\
&= 19{,}62712 \\
s_{x,y} &= \sqrt{19{,}62712} = 4{,}43025 \\
t &= \frac{\bar{x}-\bar{y}}{s_{x,y}} = \frac{51{,}33333 - 39{,}5}{4{,}43025} \approx 2{,}67103.
\end{aligned}
$$

$t_{0{,}975;17} = 2{,}109816$, więc wartość statystyki testowej należy do obszaru krytycznego. Należy odrzucić hipotezę o równości liczebności pająków na jednostkę powierzchni w borze i dąbrowie.

Obliczenia w środowisku R

Powyższe obliczenia można wykonać przy użyciu programu R w następujący sposób:

```
> x
[1] 40 57 37 61 52 49 50 68 48
> y
[1] 46 52 39 34 46 31 37 36 52 22
```

```
> t.test(x,y,var.equal=TRUE)

 Two Sample t-test

 data:  x and y
 t = 2.671, df = 17, p-value = 0.01612
 alternative hypothesis:
 true difference in means is not equal to 0
 95 percent confidence interval:
 2.486321 21.180346
 sample estimates:
 mean of x mean of y
 51.33333  39.50000
```

Wersje testu t dla przypadku niejednorodnych wariancji

Jeśli nie można przyjąć założenia o jednorodności wariancji, wtedy należy użyć wersji testu t dla przypadku niejednorodnych wariancji takich jak:
- test oparty na statystyce Cochrana-Coxa (por. Łomnicki, Rozdz. 7)
- test oparty na aproksymacji Welcha (dostępny w pakiecie R)
Należy pamiętać, że korzystanie z powyższych testów wymaga założenia normalności porównywanych populacji.

Obliczenia przy braku założenia dotyczącego równości wariancji

```
> t.test(x,y,var.equal=FALSE)

Welch Two Sample t-test

data:  x and y
t = 2.6687, df = 16.73, p-value = 0.01636
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
2.466698 21.199968
sample estimates:
mean of x mean of y
51.33333  39.50000
```

Obliczenia przy braku założenia dotyczącego równości wariancji — komentarz

Wartość statystyki testowej t- obliczana jest trochę inaczej niż w „przypadku równych wariancji”;
liczba stopni swobody— nie jest liczbą całkowitą!

Hipotezy alternatywne jednostronne

Testowanie w przypadku alternatywy jednostronnej

$$H_0 : \mu_1 = \mu_2 \quad \text{przeciwko} \quad H_1' : \mu_1 > \mu_2$$

wymaga użycia opcji alternative="greater" w procedurze t.test chcąc testować

$$H_0 : \mu_1 = \mu_2 \quad \text{przeciwko} \quad H_1'' : \mu_1 < \mu_2$$

należy wykorzystać skorzystać z opcji alternative="less" tej procedury, por. rozdz. 5 książki Grzegorzewskiego i in.

Obliczenia przy użyciu pakietu Statistica

Zapisujemy dane do arkusza Lasy2.xls — na przykład tak:

pajaki	typ
40	d
57	d
37	d
61	d
52	d
49	d
50	d
68	d
48	d
46	b
52	b
39	b
34	b
46	b
31	b
37	b
36	b
52	b
22	b

Obliczenia wykonane przy użyciu pakietu Statistica.
Statystyka/Statystyki podstawowe/Test t dla dla prób

niezależnych (wzgl. grup) należy wybrać:

zmienną zależną: pajaki
zmienną grupującą: typ

w zakładce Opcje: zaznaczyć:
Test z niezal. estymacją wariancji
PU dla ocen (z poziomem ufności 95%) : otrzymamy wtedy przedział ufności dla $\mu_1 - \mu_2$.

Testy t; Grupująca: typ (Arkusz1 w S5.stw)
Grupa 1: d
Grupa 2 b

	Średnia	Średnia	t	df	p	t oddz.	df	p
Zmienna								
pajaki	51,33333	39,50000	2,671030	17	0,016120	2,668720	16,72957	0,016356

Nważnych	Nważnych	Odch.std	Odch.std	iloraz F	p	Średnia 1	Ufność	Ufność
9	10	9,721111	9,571369	1,031534	0,954099	11,83333	2,466698	21,19997

Testy nieparametryczne

Gdy nie można przyjąć założenia o normalności rozkładów populacji, z których pochodzą dane: pomocne mogą być testy nieparametrycze takie, jak test Wilcoxona (por. książkę T. Góreckiego, rozdz. 6.3.11)

Polecana literatura

T. Górecki, Podstawy statystyki z przykładami w R, BTC, Legionowo 2011.

P. Grzegorzewski i in., Wnioskowanie statystyczne z wykorzystaniem środowiska R, Warszawa 2014, Rozdział 5.

J. Koronacki i J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT 2001, str. 229-237.

A. Łomnicki, Wprowadzenie do statystyki dla przyrodników, Wyd. 5, Warszawa 2014, Rozdział 7.