



Rozkłady zmiennych. Testy statystyczne i ich zastosowanie

Hipoteza statystyczna

Dowolne przypuszczenie co do rozkładu populacji generalnej

Prawdziwość tego przypuszczenia jest oceniana na podstawie wyników próby losowej

Hipotezę, która podlega weryfikacji to hipoteza zerowa (H_0) a jej przeciwieństwo to hipoteza alternatywna (H_1)

$H_0 : \mu_1 = \mu_2$ dwie średnie z populacji nie różnią się istotnie

$H_1 : \mu_1 < \mu_2$ dwie średnie z populacji różnią się istotnie

Poziom istotności

Maksymalne ryzyko błędu jakie badacz jest skłonny zaakceptować -
prawdopodobieństwo odrzucenia hipotezy zerowej gdy jest ona prawdziwa

Prawdopodobieństwo P-value

Krytyczny (graniczny) poziom istotności; prawdopodobieństwo testowe

Najmniejszy poziom istotności przy którym dla zaobserwowanej wartości statystyki testowej odrzucilibyśmy hipotezę zerową

Hipotezę zerową odrzucamy, gdy wyliczone prawdopodobieństwo testowe okaże się nie większe od przyjętego przez nas poziomu istotności (**zwykle 0,05**)

Effect size

Obecnie coraz częściej odchodzi się od klasycznych założeń statystycznych w wykrywaniu zależności pomiędzy zmiennymi oraz ich porównywaniu, np.:

- liczba prób nie mniejsza niż 30
- $P < 0.05$
- Wielkość współczynnika korelacji r , czy determinacji R^2

Biol. Rev. (2007), **82**, pp. 591–605.
doi:10.1111/j.1469-185X.2007.00027.x

591

Effect size, confidence interval and statistical significance: a practical guide for biologists

Shinichi Nakagawa^{1,*} and Innes C. Cuthill²

Moving to a World Beyond “ $p < 0.05$ ”

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “ $p < 0.05$ ”, *The American Statistician*, 73:sup1, 1-19, DOI: [10.1080/00031305.2019.1583913](https://doi.org/10.1080/00031305.2019.1583913)

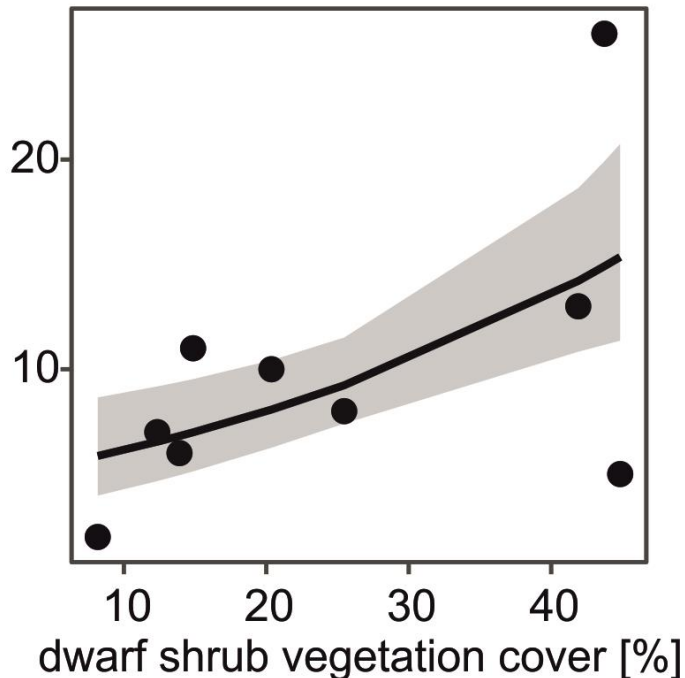
To link to this article: <https://doi.org/10.1080/00031305.2019.1583913>

Dlaczego?

Ponieważ często dysponując mniejszymi zbiorami danych (np. w sytuacjach, gdzie pobór prób jest ekstremalnie trudny), już wtedy można zaobserwować jakąś tendencję interpretowalną pod względem ekologicznym

predictor	model parameters			
	estimate	SE	z	Pr(> z)
(intercept)	1.669	0.386	4.314	<0.001
dwarf shrub vegetation cover [%]	2.240	1.175	1.906	0.056

c)

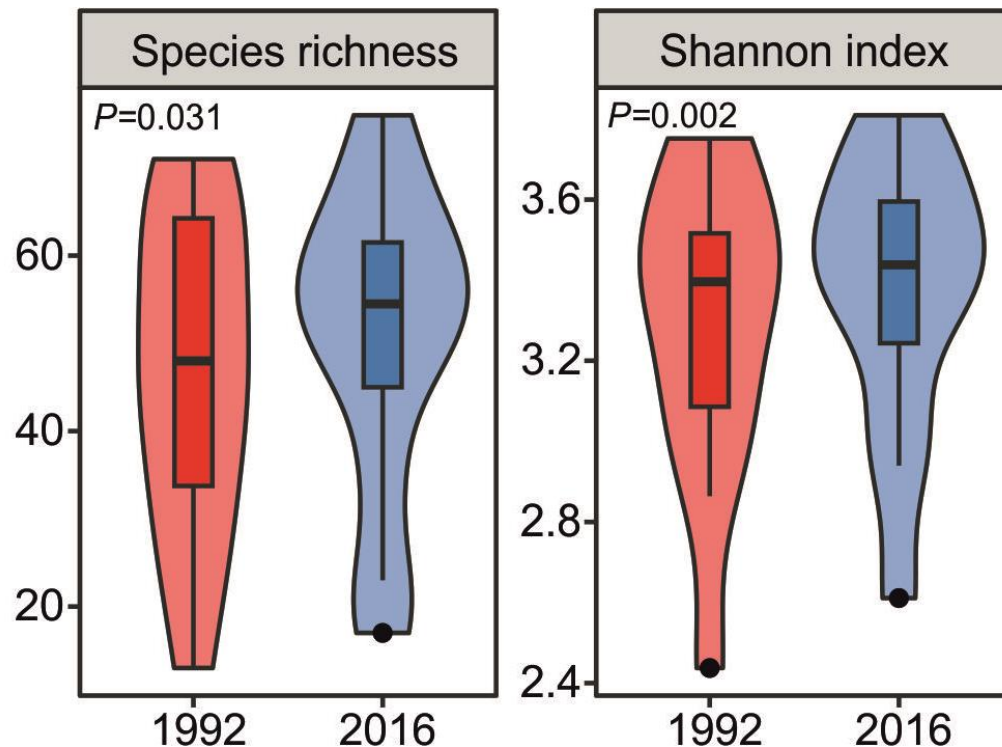


Model zależności liczebności cietrzewia w zależności od pokrycia stanowiska przez zarośla kosówki (Adamowicz i in., w przyg.)

Z drugiej strony, posiadając większy zbiór danych można:

- a) Albo zaobserwować brak istotnych różnic ($P > 0.05$) przy stosunkowo wielkich różnicach pomiędzy średnimi z prób
- b) Albo zaobserwować istotną różnicę ($P > 0.00000001$) przy nikłych różnicach pomiędzy średnimi z prób (w przypadku, gdy liczebność prób jest ogromna)

Dlatego bardziej informatywne jest podanie wielkości różnic oraz ich wyjaśnienie w sensie ekologicznym, gdyż będzie to mniej obciążone artefaktami związanymi z wielkością próby, co może prowadzić do sformułowania nieprawdziwych wniosków



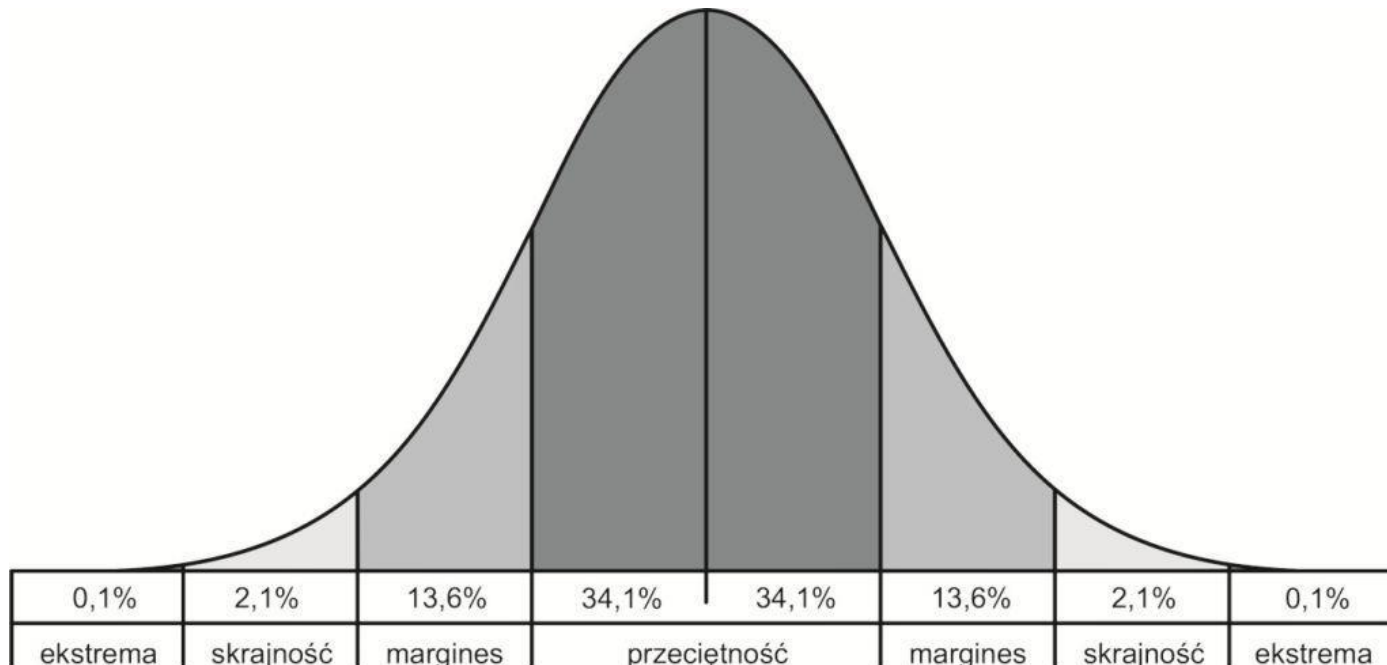
Zmiany różnorodności
taksonomicznej mszaków
epifitycznych BPN po około 20
latach od pierwszych badań
(1992 vs 2016)

Normalność rozkładu

Rozkład zbliżony do normalnego jest jednym z najważniejszych rozkładów w biologii

Rozwiązanie wielu zagadnień statystycznych jest "prostsze", jeśli analizowana cecha ma rozkład normalny

Wiele analiz statystycznych i testów wymaga założenia o normalności rozważanej zmiennej (testy t-Studenta, analiza wariancji, regresja itd.)



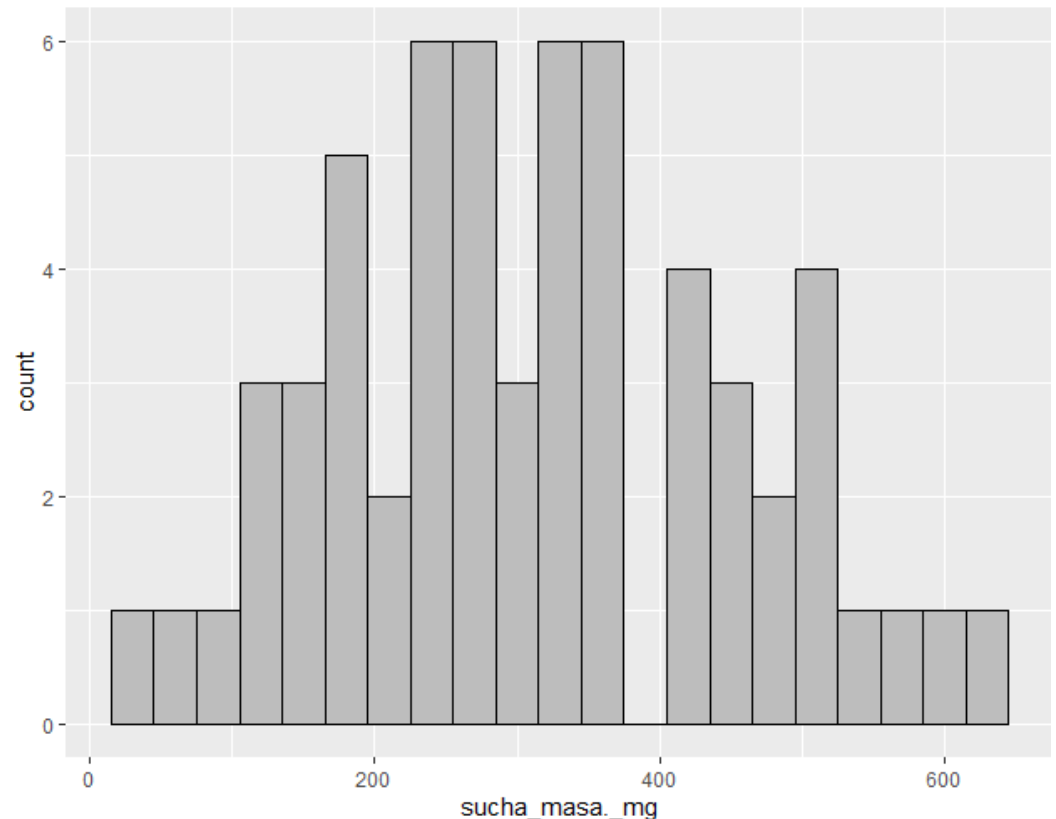
Ocena normalności rozkładu

Histogram

Pytanie: czy średnia sucha masa liści herbaty w przeliczeniu na jeden krzew herbaciany reprezentuje rozkład normalny?

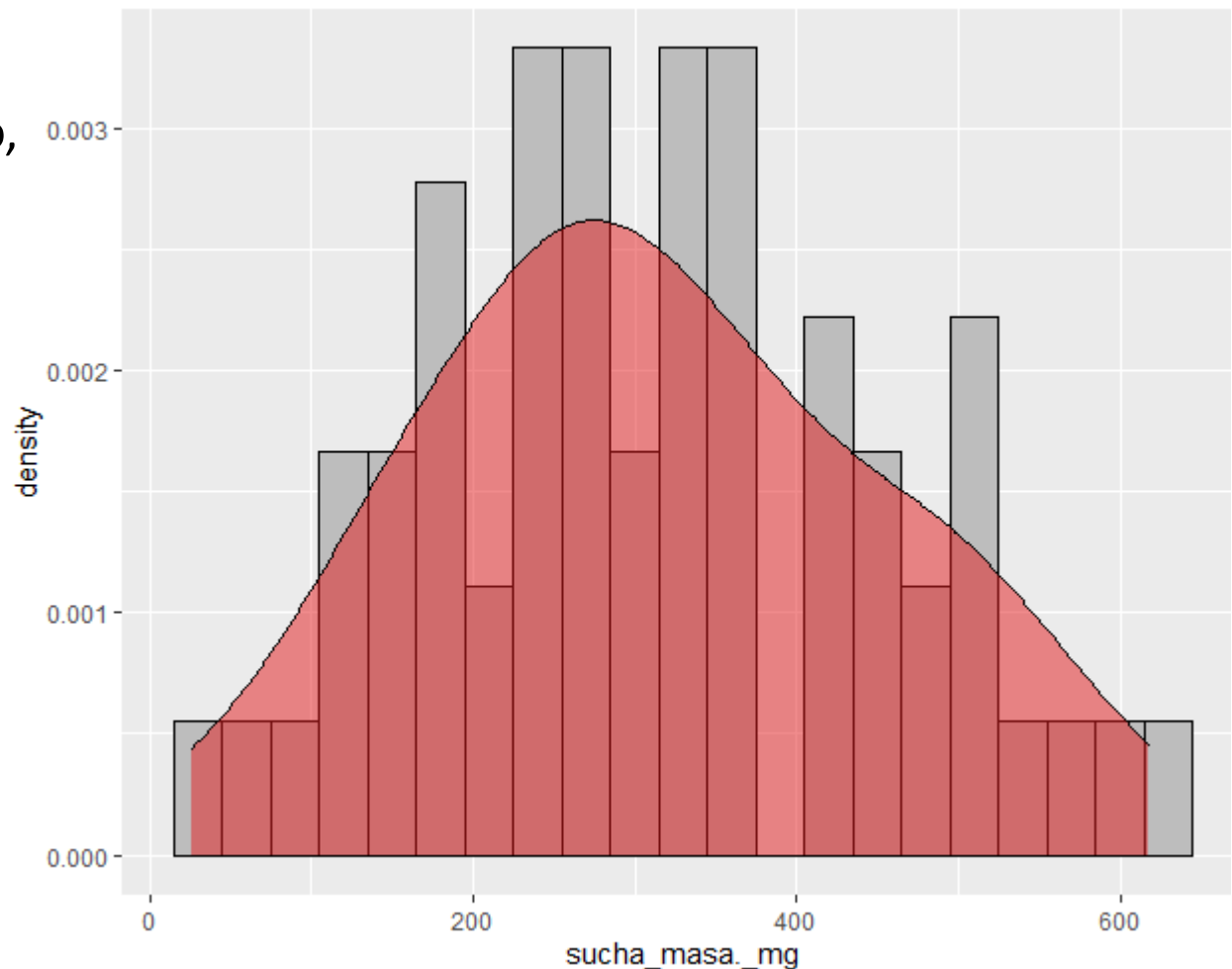
```
ggplot(kofeina.data, aes(x=sucha_masa._mg))+  
  geom_histogram(aes(y=..count..), binwidth=30, colour="black", fill="#bdbdbd")
```

Tutaj może mało widoczne,
Więc użyjmy jądrowego
estymatora gęstości, by lepiej
pokazać rozkład danych



```
ggplot(kofeina.data, aes(x=sucha_masa._mg))+  
  geom_histogram(aes(y=..density..), binwidth=30, colour="black", fill="#bdbdbd")+  
  geom_density(alpha=.5, fill="#e31a1c")
```

Jądrowy estymator gęstości
mierzy prawdopodobieństwo,
że losowo wybrana ze zbioru
danych
obserwacja będzie miała
wartość z danego przedziału



Testy pozwalające na ocenę normalności rozkładu:

- test Kołmogorova-Smirnova

```
fBasics::ksnormTest()
```

- test W Shapiro-Wilka (preferowany ze względu na dużą moc)

```
stats::shapiroTest()
```

Współcześnie mało kto używa tych testów do sprawdzania normalności rozkładu.

Częściej stosuje się metody wizualizacji danych w postaci histogramów.

Ponadto, oceny rozkładu zmiennych można dokonać intuicyjnie, znając strukturę danych

Wskazówka: nie ma idealnie normalnych rozkładów, więc jeśli mamy zmienną ciągłą, a priori możemy założyć, że mamy rozkład normalny

Testy statystyczne

Służą do badania istotności różnic pomiędzy próbami

Rozkład normalny

Zakładamy, że zbliżony

Inny (np. Poisson, beta, itd.)

Testy parametryczne

Testy nieparametryczne

Test t Studenta dla par
niezwiązanych

Test t Studenta dla par
związanych

ANOVA

Test Chi kwadrat

Test Manna-Whitneya dla par
niezwiązanych

Test Manna-Whitneya dla par
związanych

Test Kruskala-Wallisa

Testy parametryczne

Casus: zmiany w różnorodności porostów epifitycznych w Rezerwacie Ścisłym Białowieskiego Parku Narodowego w:
Czasie: time=='h' (1980s) oraz time=='n' (2010s)
Pomiędzy typami lasów: habitat=='conif' (iglasty) oraz habitat=='decid' (liściasty)

```
> porosty
```

	habitat	time	EIV_N	Rich	Shan
1	decid	h	3.125000	28	3.245232
2	decid	n	3.158730	40	3.589339
3	decid	h	2.921569	33	3.404548
4	decid	n	3.253968	42	3.633877
5	decid	h	2.925000	32	3.394398
6	decid	n	3.225806	43	3.645540
7	decid	h	3.134615	36	3.486709
8	decid	n	3.350877	40	3.606988
9	decid	h	3.226415	36	3.499831
10	decid	n	3.094340	41	3.615386
11	decid	h	3.058824	40	3.606320
12	decid	n	3.116667	44	3.678743

Science of the Total Environment 643 (2018) 468–478



Contents lists available at [ScienceDirect](#)

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv



Changes in the epiphytic lichen biota of Białowieża Primeval Forest are not explained by climate warming

Anna Łubek ^{a,*}, Martin Kukwa ^b, Bogdan Jaroszewicz ^c, Patryk Czortek ^c



Test t Studenta dla par niewiązanych

Stosowany, gdy obserwacje z próby **A** (habitat=='decid') nie odpowiadają obserwacjom z próby **B** (habitat=='conif')

Liczba obserwacji z próby **A** może być równa liczbie obserwacji z próby **B** lub różna od liczby obserwacji z próby **B**

Zakładamy, że Shan reprezentuje rozkład zbliżony do normalnego

```
> porosty
  habitat time      EIV_N Rich      Shan
1    decid  h 3.125000   28 3.245232
2    decid  n 3.158730   40 3.589339
3    decid  h 2.921569   33 3.404548
4    decid  n 3.253968   42 3.633877
5    decid  h 2.925000   32 3.394398
6    decid  n 3.225806   43 3.645540
7    decid  h 3.134615   36 3.486709
8    decid  n 3.350877   40 3.606988
9    decid  h 3.226415   36 3.499831
10   decid  n 3.094340   41 3.615386
```

...

```
92   decid  n 3.195122   30 3.296836
93   conif  h 2.931034   19 2.840565
94   conif  n 3.232558   25 3.082018
95   conif  h 3.846154    8 1.951260
96   conif  n 3.581395   26 3.163942
97   conif  h 3.285714   19 2.858006
98   conif  n 3.388889   35 3.452254
99   conif  h 2.818182   20 2.898746
100  conif  n 3.184615   43 3.656239
101  decid  h 3.272727   27 3.210176
```

```
> summary(porosty$habitat)
conif decid
  100   188
```

H0: Średnia różnorodność gatunkowa (Shan) bioty porostów epifitycznych nie różni się pomiędzy borami a lasami liściastymi

H1: Średnia różnorodność gatunkowa (Shan) bioty porostów epifitycznych różni się istotnie pomiędzy dwoma typami lasów

```
t.test(porosty$Shan[porosty$habitat=="conif"],  
      porosty$Shan[porosty$habitat=="decid"],  
      paired=FALSE)
```

```
welch Two Sample t-test
```

```
data: porosty$Rich[porosty$habitat == "conif"] and porosty$Rich[porosty$habitat == "decid"]  
t = -9.7105, df = 162.14, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -16.33801 -10.81604  
sample estimates:  
mean of x mean of y  
 24.29000  37.86702
```


Test t Studenta dla par wiązanych

Stosowany, gdy obserwacje z próby **A** odpowiadają obserwacjom z próby **B**

Liczba obserwacji z próby **A** równa liczbie obserwacji z próby **B**

```
> porosty
  habitat time      EIV_N Rich      Shan
1    decid  h 3.125000   28 3.245232
2    decid  n 3.158730   40 3.589339
3    decid  h 2.921569   33 3.404548
4    decid  n 3.253968   42 3.633877
5    decid  h 2.925000   32 3.394398
6    decid  n 3.225806   43 3.645540
7    decid  h 3.134615   36 3.486709
8    decid  n 3.350877   40 3.606988
9    decid  h 3.226415   36 3.499831
10   decid  n 3.094340   41 3.615386
```

...

```
92   decid  n 3.195122   30 3.296836
93   conif  h 2.931034   19 2.840565
94   conif  n 3.232558   25 3.082018
95   conif  h 3.846154    8 1.951260
96   conif  n 3.581395   26 3.163942
97   conif  h 3.285714   19 2.858006
98   conif  n 3.388889   35 3.452254
99   conif  h 2.818182   20 2.898746
100  conif  n 3.184615   43 3.656239
101  decid  h 3.272727   27 3.210176
```

```
> summary(porosty$time)
  h    n
144 144
```

H0: Średnia różnorodność gatunkowa (Shan) epifitów nie różni się pomiędzy dwoma terminami badań h (1990s) i n (2010s)

H1: Średnia różnorodność gatunkowa (Shan) epifitów różni się pomiędzy dwoma terminami badań h i n

```
t.test(porosty$Shan[porosty$time=="h"], porosty$Shan[porosty$time=="n"],  
paired=TRUE)
```

Inna forma zapisu:

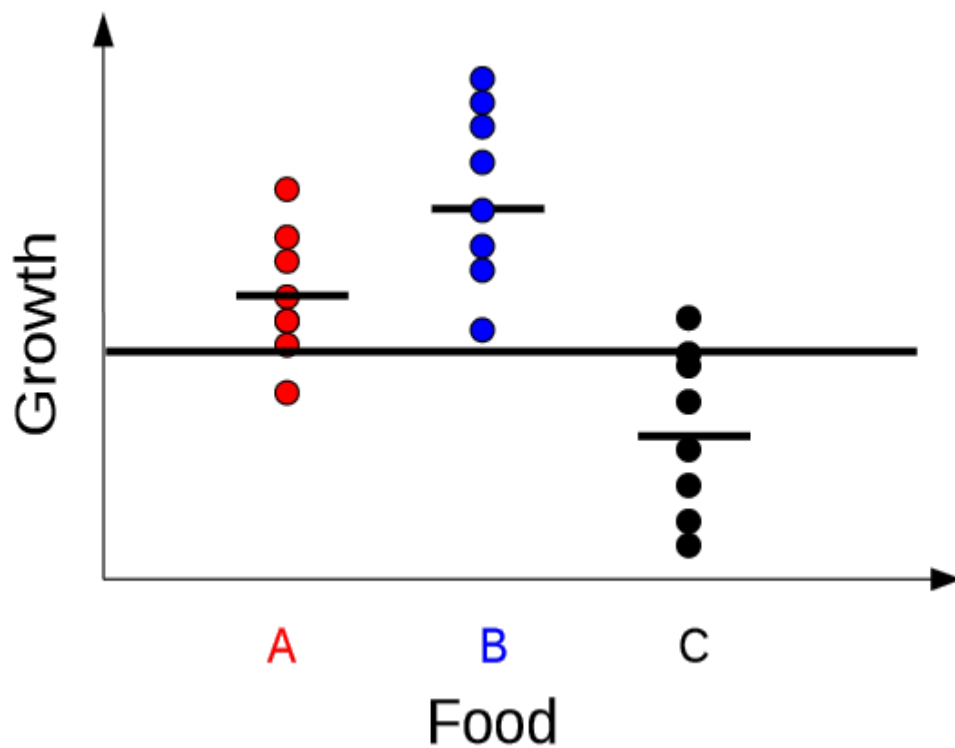
```
t.test(Shan, time, data=porosty, paired=TRUE)
```

```
Paired t-test  
data: porosty$Shan[porosty$time == "h"] and porosty$Shan[porosty$time == "n"]  
t = -14.72, df = 143, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-0.4350827 -0.3320635  
sample estimates:  
mean of the differences  
-0.3835731
```

Jednoczynnikowa ANOVA

Growth = zmienna objaśniana

Food = zmienna objaśniająca (kategoryczna)



Food	Growth
A	51.16
A	46.24
A	48.79
etc	etc
B	56.19
B	50.83
B	49.83
etc	etc
C	49.26
C	42.19
C	40.08
etc	etc

Hipotezy

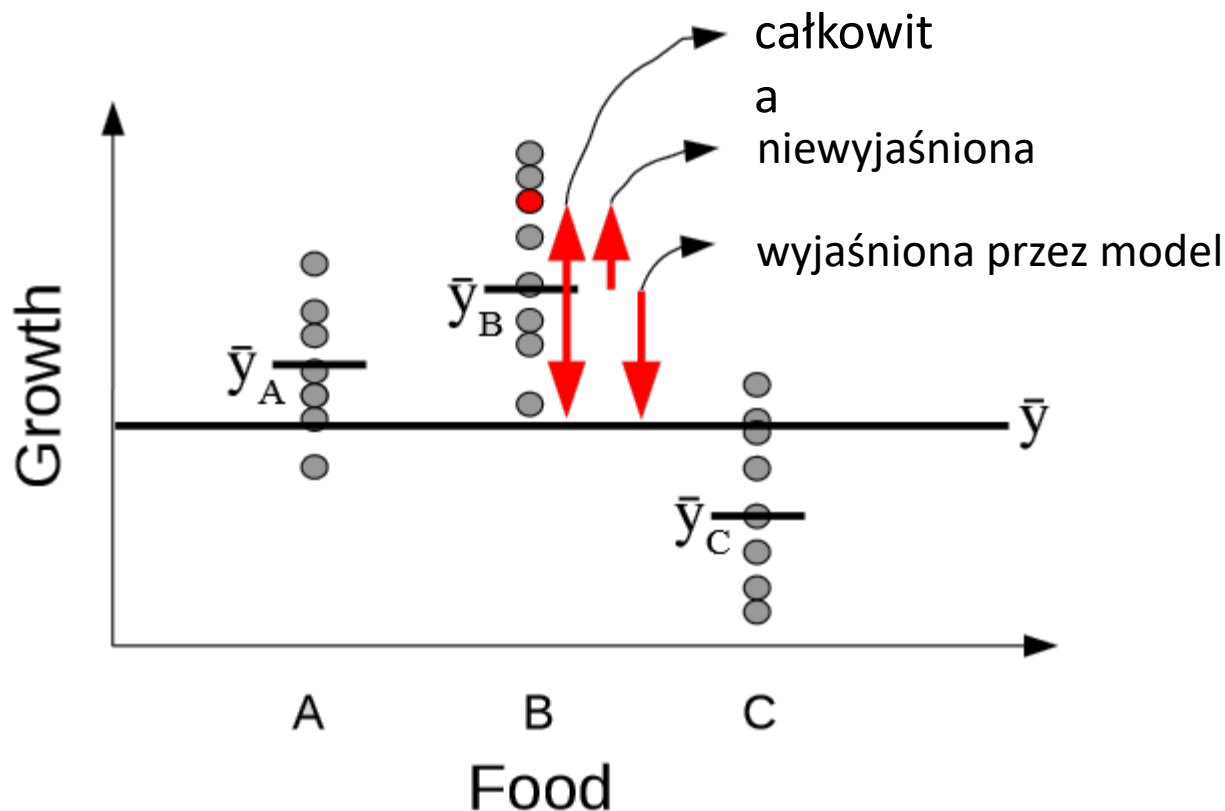
H0: $\mu_{\text{Food A}} = \mu_{\text{Food B}} = \mu_{\text{Food C}}$

H1: Przynajmniej dwie średnie różnią się

Statystyka testowa F

Bardzo duża wartość F oznacza, że wyjaśniona wariancja (pomiędzy grupami) znacznie przewyższa niewyjaśnioną wariancję (w obrębie grup)

$F = \text{"wyjaśniona wariancja"} / \text{"niewyjaśniona wariancja"}$



H0: średni plon dwóch odmian jęczmienia nie różni się

H1: średni plon dwóch odmian jęczmienia istotnie różni się od siebie

```
aov(barley$Yield~barley$Barley)
```

Albo:

```
aov(Yield~Barley, data=barley)
```

```
call:
```

```
  aov(formula = barley$Yield ~ barley$Barley)
```

```
Terms:
```

	barley\$Barley	Residuals
Sum of Squares	0.00678462	0.13723077
Deg. of Freedom	1	24

```
Residual standard error: 0.07561712
```

```
Estimated effects may be unbalanced
```

```
summary(aov(barley$Yield~barley$Barley))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
barley\$Barley	1	0.00678	0.006785	1.187	0.287
Residuals	24	0.13723	0.005718		

Wniosek: średni plon dwóch odmian jęczmienia nie różni się

Przyjmujemy H0, odrzucamy H1

Dwuczynnikowa ANOVA

H0: Ani typ pokarmu, ani jego widzialność nie wpływa na wzrost dorsza

Cladocera



granulki



Cladocera



granulki



Cladocera



granulki



= woda czysta



= woda mętna

Własności zmiennych

Zmienna objaśniana jest ciągła (Growth)

Zmienne objaśniające są kategoryczne
(Food, Visibility)

dorsz1

	Growth	Food	visibility
1	495	mysids	turbid
2	501	mysids	turbid
3	483	mysids	turbid
4	490	mysids	turbid
5	482	mysids	turbid
6	462	mysids	turbid
7	497	mysids	turbid
8	498	mysids	turbid
9	501	mysids	turbid
10	491	mysids	turbid
11	504	mysids	clear
12	528	mysids	clear
13	509	mysids	clear
14	511	mysids	clear
15	525	mysids	clear
16	514	mysids	clear
17	526	mysids	clear
18	518	mysids	clear
19	504	mysids	clear
20	505	mysids	clear
21	527	pellets	turbid
22	526	pellets	turbid
23	519	pellets	turbid
24	525	pellets	turbid

musimy wiedzieć, co chcemy osiągnąć, tzn. jeśli zakładamy addytywność i interakcję między predyktorami, wtedy dajemy znak mnożenia między nimi

```
aov(Growth~Visibility*Food)
```

Jeśli zakładamy tylko addytywność, wtedy dajemy Visibility+Food

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
B1 dorsz1\$Visibility	1	1716	1716	14.10	0.000613	***
B2 dorsz1\$Food	1	6101	6101	50.12	2.58e-08	***
B3 dorsz1\$Visibility:dorsz1\$Food	1	1277	1277	10.49	0.002582	**
Residuals	36	4382	122			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Interakcja **B3** jest istotna. Na tej podstawie stwierdzamy, że zarówno widzialność pokarmu (**B1**), jak i jego typ (**B2**) istotnie wpływają na wzrost dorsza

Test post-hoc Tukeya

Po zrobieniu ANOVy z dwoma poziomami predyktora (np. dwa punkty w czasie) sytuacja jest prosta

A co, jeśli nasz predyktor ma kilka poziomów?

ANOVA dam nam tylko odpowiedź, że są istotne różnice między grupami, ale nie powie nam pomiędzy którymi

Tej odpowiedzi udzieli nam test post-hoc Tukeya (jeden z wielu, ale najczęściej stosowany; inne: LSD Fishera, Duncana)

W R dwie wersje:

pakiet `agricolae` i funkcja `HSD.test`

pakiet `multcomp` i funkcja `glht`

Casus: różnice w różnorodności funkcjonalnej porostów (wskaźnik FDis – dyspersja funkcjonalna) w Puszczy Białowieskiej w zależności od typu lasu

3 poziomy katerycznego predyktora: conif, mixed_decid, moist

wersja z pakietu agricolae

```
an13<-aov(FDis~forest, data=pred.pr)
HSD.test(an13, 'forest',console = T)
```

```
> HSD.test(an13, 'forest',console = T)

Study: an13 ~ "forest"

HSD Test for FDis

Mean Square Error:  0.5434269

forest,  means

      FDis      std  r      Min      Max
conif    28.30547 1.0981871 50 25.95167 30.33885
mixed_decid 29.76826 0.4446079 54 28.38835 30.50749
moist     29.67715 0.4252187 40 28.68260 30.43563

Alpha: 0.05 ; DF Error: 141
Critical Value of Studentized Range: 3.349881

Groups according to probability of means differences and alpha level( 0.05 )

Treatments with the same letter are not significantly different.

      FDis groups
mixed_decid 29.76826      a
moist       29.67715      a
conif       28.30547      b
```

```
summary(glht(an13, mcp(type='Tukey')))
```

Nie działa, wyskakuje błąd:

```
> summary(glht(an13, mcp(type='Tukey')))  
Error in mcp2matrix(model, linfct = linfct) :  
  variable(s) 'type' have been specified in 'linfct' but cannot be found in 'model'!
```

Trzeba atrybut „type” zmienić na nazwę zmiennej z ramki danych, pod którą kryje się predyktor, w zależności od którego zrobiliśmy ANOVę:

```
summary(glht(an13, mcp('forest'='Tukey')))
```

Co nam wypływa?

```
> summary(glht(an13, mcp('forest'='Tukey')))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = FDis ~ forest, data = pred.pr)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)	
mixed_decid - conif == 0	1.46279	0.14468	10.111	<1e-04	***
moist - conif == 0	1.37168	0.15638	8.772	<1e-04	***
moist - mixed_decid == 0	-0.09111	0.15378	-0.592	0.824	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Adjusted p values reported -- single-step method)

Pokazuje, pomiędzy którymi grupami różnice są istotne, ale nie pokazuje literek

Można kazać R dodać literki przy użyciu funkcji cld

```
cld(glht(an1, mcp('forest'='Tukey')))
```

```
> cld(glht(an13, mcp('forest'='Tukey')))
```

conif	mixed_decid	moist
"a"	"b"	"b"

Testy nieparametryczne

Jakie inne rozkłady mogą reprezentować dane?

Rozkład Poissona:

rozkład dyskretny, używany do analizy liczebności, np.

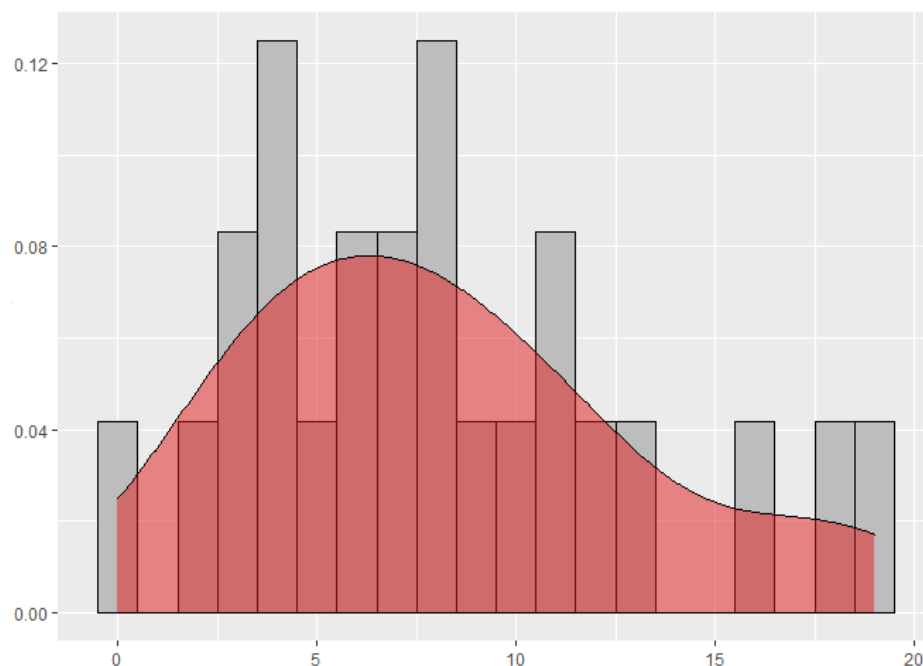
liczba osobników

liczba gatunków itd.

liczba morfotypów grzybów związanych z rozkładem biomasy *Solidago canadensis* (Kisło i in, w przyg.)



```
ggplot(dane, aes(x=rich.morfotypen))+  
geom_histogram(aes(y=..density..),  
binwidth=1, colour="black",  
fill="#bdbdbd")+  
  geom_density(alpha=.5,  
fill="#e31a1c")
```

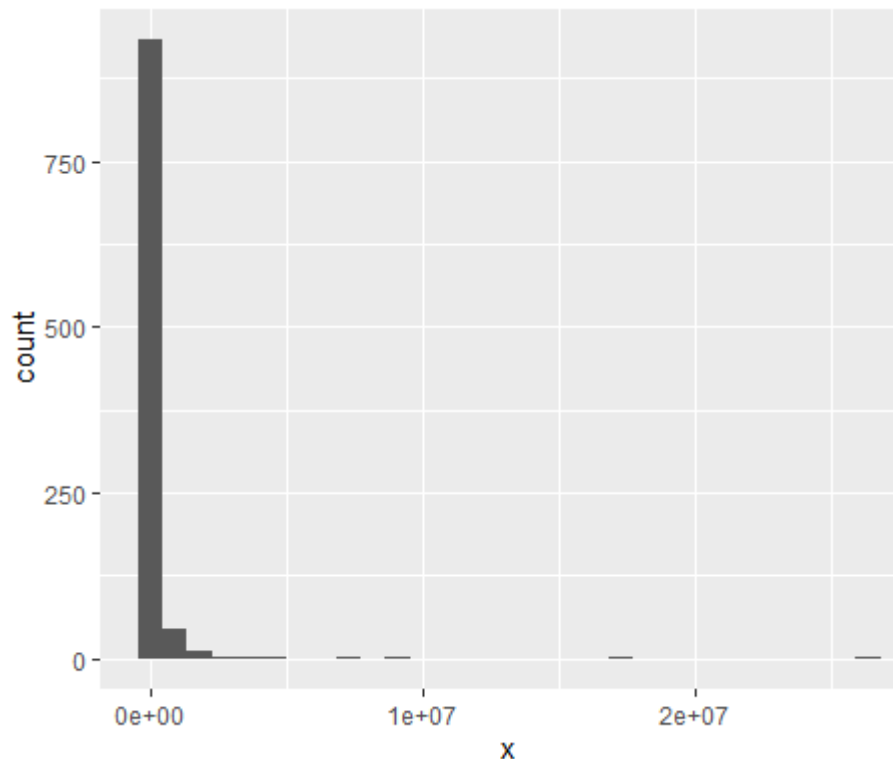


Rozkład log-normalny

`rlnorm()` jest funkcją generującą liczby pseudolosowe z rozkładu log-normalnego o parametrach, które możemy sami określić:

1000, 10, 2 – mówimy R, że ma wylosować 1000 liczb pseudolosowych o średniej 10 i odchyleniu standardowym 2

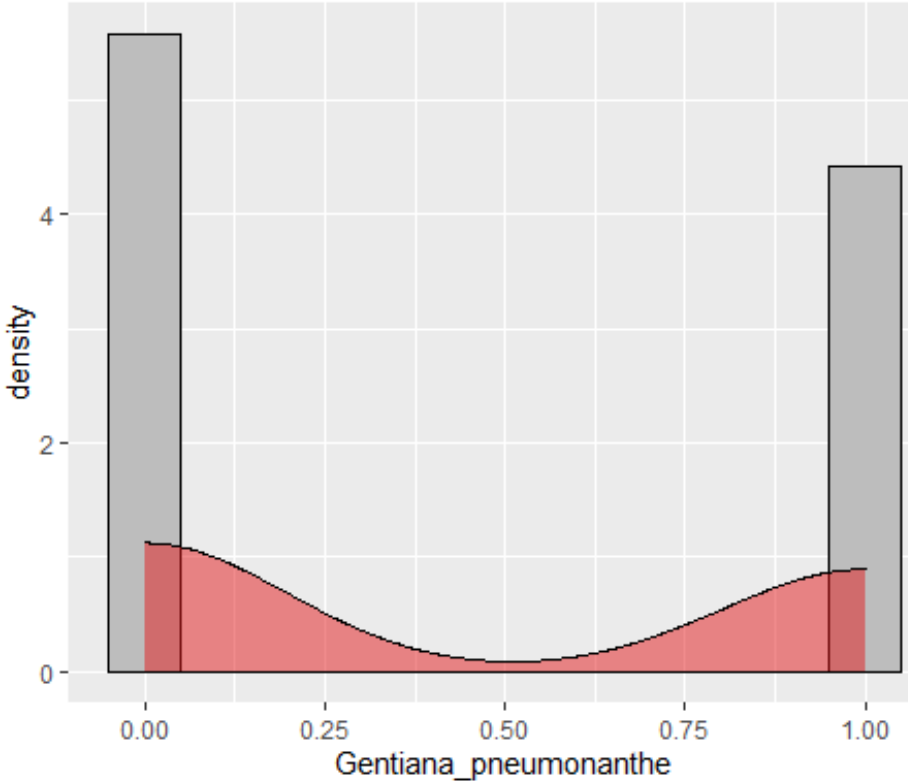
```
rozklad<-data.frame(x=rlnorm(1000,10,2))  
ggplot(rozklad, aes(x=x))+geom_histogram()
```



Rozkład dwumianowy:

obserwacje przyjmują tylko wartości binarne (jest/nie ma; 1/0)
pozwala odpowiedzieć na pytania:
-czy dany gatunek występuje w danym środowisku, czy nie
-czy przeżywa w danych warunkach, czy nie itd.

np. występowanie *Gentiana pneumonanthe* na łąkach trzęślicowych w zależności od
cech roślinności w otoczeniu



Received: 31 March 2020 | Revised: 9 December 2020 | Accepted: 16 December 2020
DOI: 10.1111/jvs.12983

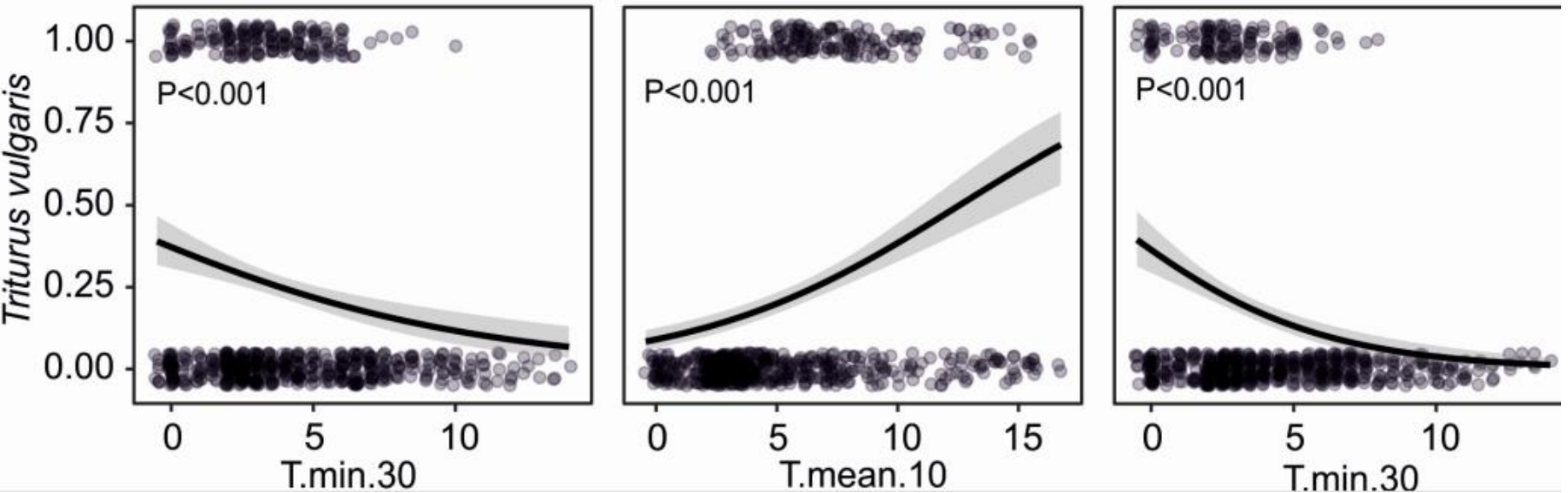
RESEARCH ARTICLE

Journal of Vegetation Science

Niche differentiation, competition or habitat filtering?
Mechanisms explaining co-occurrence of plant species on wet
meadows of high conservation value

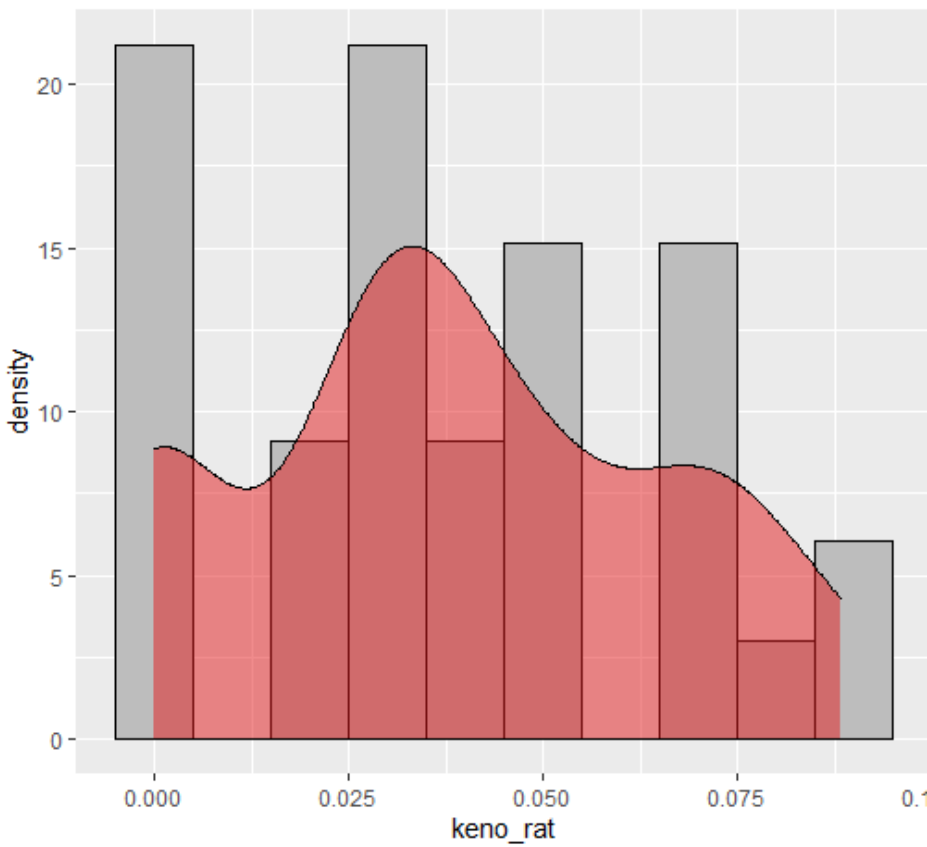
Patryk Czortek¹ | Anna Orczewska² | Marcin K. Dyderski³

Rozkład dwumianowy jest bardzo przydatny w szacowaniu prawdopodobieństwa wystąpienia danej obserwacji w zależności od wybranego czynnika, np. występowanie *Triturus vulgaris* w zależności od temperatury



Rozkład beta

Służy do analizy proporcji, np. udziału obcych gatunków roślin (keno_rat) w parkach miejskich w zależności od gęstości sieci rzecznej w otoczeniu



Urban Forestry & Urban Greening 47 (2020) 126525



Contents lists available at [ScienceDirect](#)

Urban Forestry & Urban Greening

journal homepage: www.elsevier.com/locate/ufug



Surrounding landscape influences functional diversity of plant species in urban parks

Patryk Czortek^{a,*}, Remigiusz Pielech^b

^a Białowieża Geobotanical Station, Faculty of Biology, University of Warsaw, ul. Sportowa 19, 17-230 Białowieża, Poland

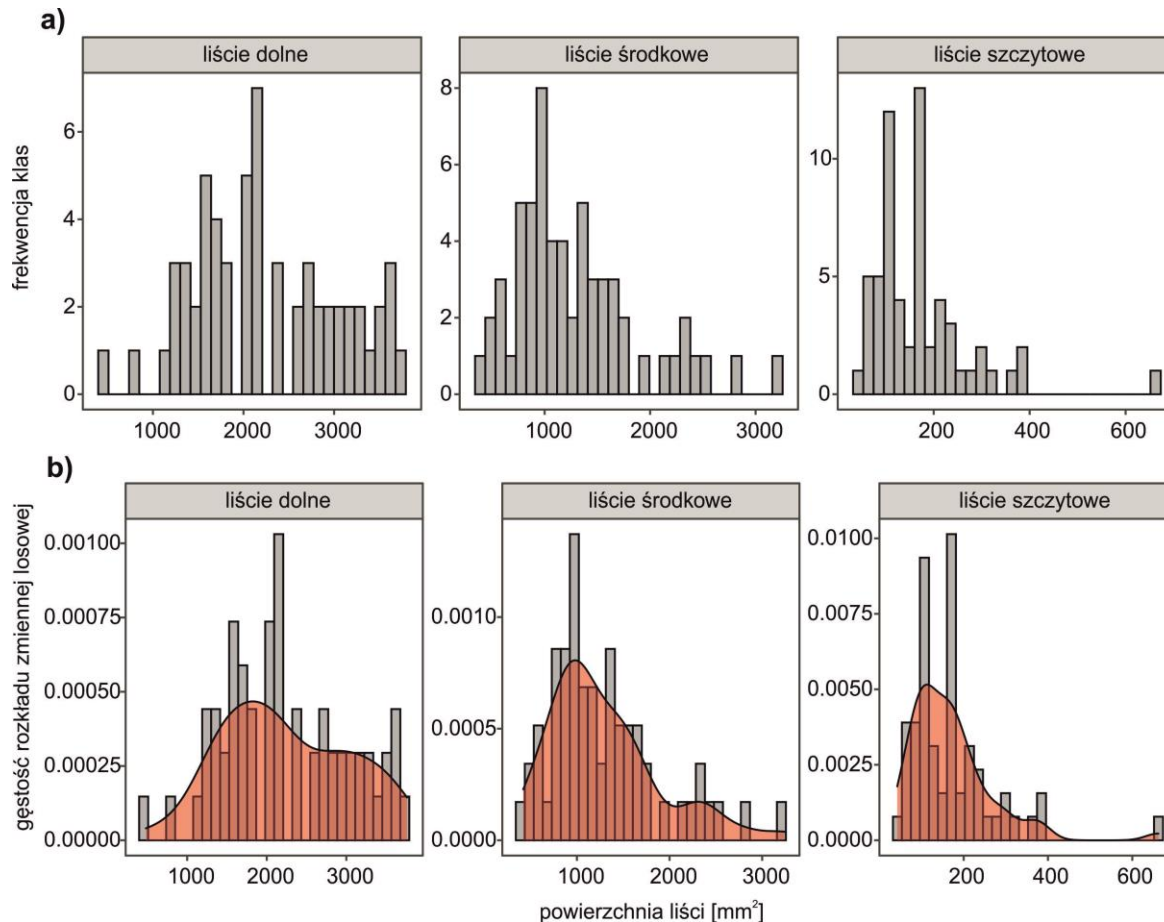
^b Department of Forest Biodiversity, University of Agriculture in Kraków, al. 29 Listopada 46, 31-425 Kraków, Poland



Jak informację o rozkładach pokazywać w pracach?

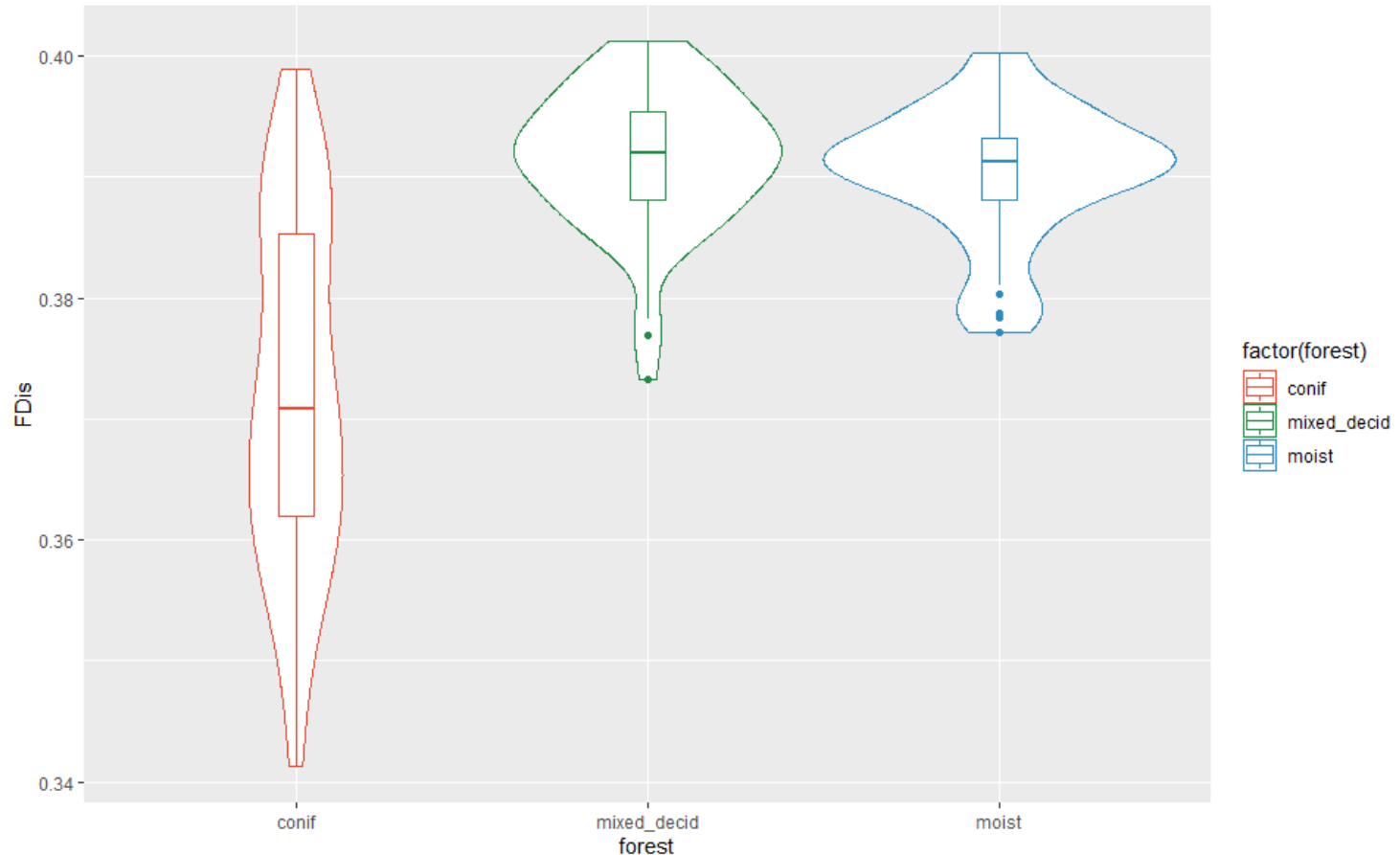
W Appendixie – tabela z parametrami testów Shapiro-Wilka/ Kołmogorova-Smirnova (ale komu się chce to robić?)

W Appendixie – obrazki z rozkładami



Albo w tekście manuskryptu jako dodatkowa informacja na obrazkach
wykres strunowy (violin plot) jako dodatek do wykresu pudełkowego (boxplot)
Struna z każdej strony pudełka to jądrowy estymator gęstości - info o rozkładzie

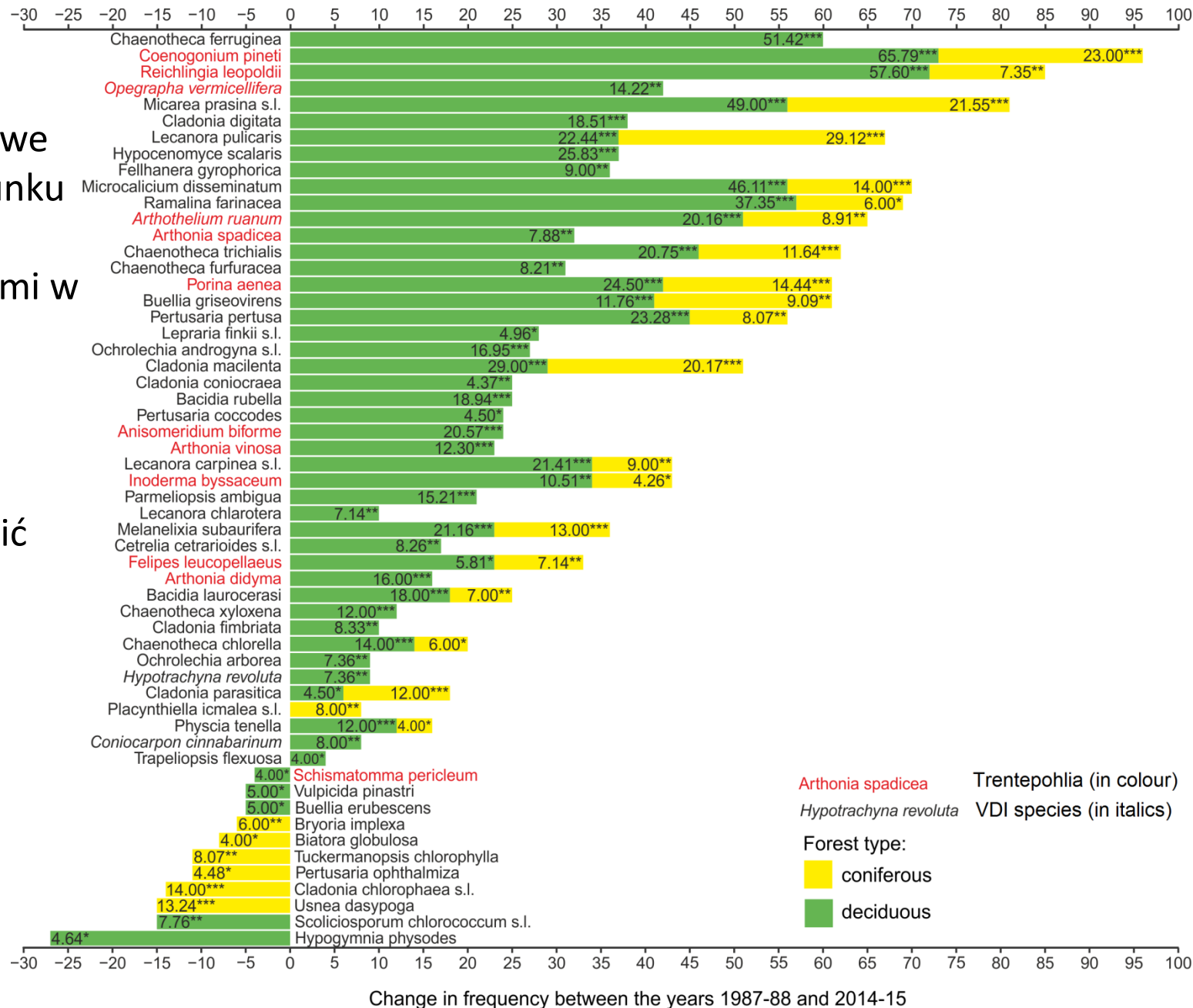
```
ggplot(traity.lasy, aes(x=forest, y=FDIs, col=factor(forest)))+  
  scale_colour_manual(values=c("#e34a33", "#238b45", "#2b8cbe"))+  
  geom_violin(aes(x=forest, y=FDIs, col=factor(forest)), width=1)+  
  geom_boxplot(aes(x=forest, y=FDIs, col=factor(forest)), width=0.1)
```



Test Chi kwadrat (tylko dla par niewiązanych)

Znakomity do badania różnic we frekwencji gatunku np. pomiędzy dwoma punktami w czasie, dwoma siedliskami itd.

Można tak, ale można też zrobić GLM z binomialem




```
> freq.epiphytes
  species freq.dec.old freq.dec.new
1  Ram.far         15         72
2  Ino.bys         38         72
3  Ope.niv         84         73
4  Cha.tri         28         74
5  Cha.fur         43         74
6  Prt.coc         52         76
7  Aly.var         72         76
8  Coe.pin          4         77
9  Lcr.arg         72         79
10 Rei.leo          9         81
```



Coenogonium pineti

```
chisq.test(freq.epiphytes[,c(2:3)][8,])
```

chi-squared test for given
probabilities

```
data:  freq.epiphytes[, c(2:3)][8, ]
X-squared = 65.79, df = 1,
p-value = 5.016e-16
```


Test Manna-Whitneya dla par niewiązanych

```
> cover.clearcut
[1] 7 28 19 29 2 7 24 5 30 14 18 10 28 8 11 6 54 34 29 37 32 31 13 37 12 22 19
[28] 11 31 17
> cover.forest
[1] 35 50 33 31 32 25 36 54 39 43 41 3 39 44 44 27 4 39 55 33 21 22 36 30 40 51 53
[28] 3 23 1 49
```

H0: pokrycie gatunków leśnych nie różni się pomiędzy lasem a zrębem zupełnym

H1: pokrycie gatunków leśnych różni się pomiędzy lasem a zrębem zupełnym

```
wilcox.test(cover.clearcut, cover.forest, paired=FALSE)
```

```
wilcoxon rank sum test with continuity correction

data: cover.clearcut and cover.forest
W = 229, p-value = 0.0006779
alternative hypothesis: true location shift is not equal to 0


Warning message:
In wilcox.test.default(cover.clearcut, cover.forest, paired = FALSE) :
nie można obliczyć dokładnej wartości prawdopodobieństwa z powtórzonymi wartościami
```

Zakładamy, że % pokrycie gatunków leśnych nie reprezentuje rozkładu normalnego (rozkład beta)

Biodiversity and Conservation
<https://doi.org/10.1007/s10531-019-01795-8>

ORIGINAL PAPER

The impact of salvage logging on herb layer species composition and plant community recovery in Białowieża Forest

Anna Orczewska¹  · Patryk Czortek² · Bogdan Jaroszewicz²



Test Manna-Whitneya dla par wiązanych

H0: bogactwo gatunkowe wyleżysk nie różni się pomiędzy dwoma punktami w czasie

H1: bogactwo gatunkowe wyleżysk różni się pomiędzy dwoma punktami w czasie

Znowu zakładamy rozkład dyskretny (Poisson)

wylezyska

```
> wylezyska
      rich time
58k      18    k
58n      28    n
67k      20    k
67n      17    n
32k      23    k
32n      25    n
85k      22    k
85n      22    n
8k       20    k
8n       21    n
30k      20    k
30n      18    n
122k     25    k
122n     26    n
100k     20    k
100n     33    n
107k     19    k
107n     33    n
45k      23    k
```

```
wilcox.test(wylezyska$rich[wylezyska$time=='k'],  
wylezyska$rich[wylezyska$time=='n'], paired=TRUE)
```

```
wilcoxon signed rank test with continuity correction
```

```
data: wylezyska$rich[wylezyska$time == "k"] and wylezyska$rich[wylezyska$time == "n"]  
V = 8.5, p-value = 0.01067  
alternative hypothesis: true location shift is not equal to 0
```



A co jeśli nasza zmienna grupująca ma więcej niż dwa poziomy?

Test post-hoc Tukeya robimy, gdy zakładamy normalność rozkładu

Dla danych reprezentujących inne rozkłady alternatywa w postaci testu Kruskala-Wallisa

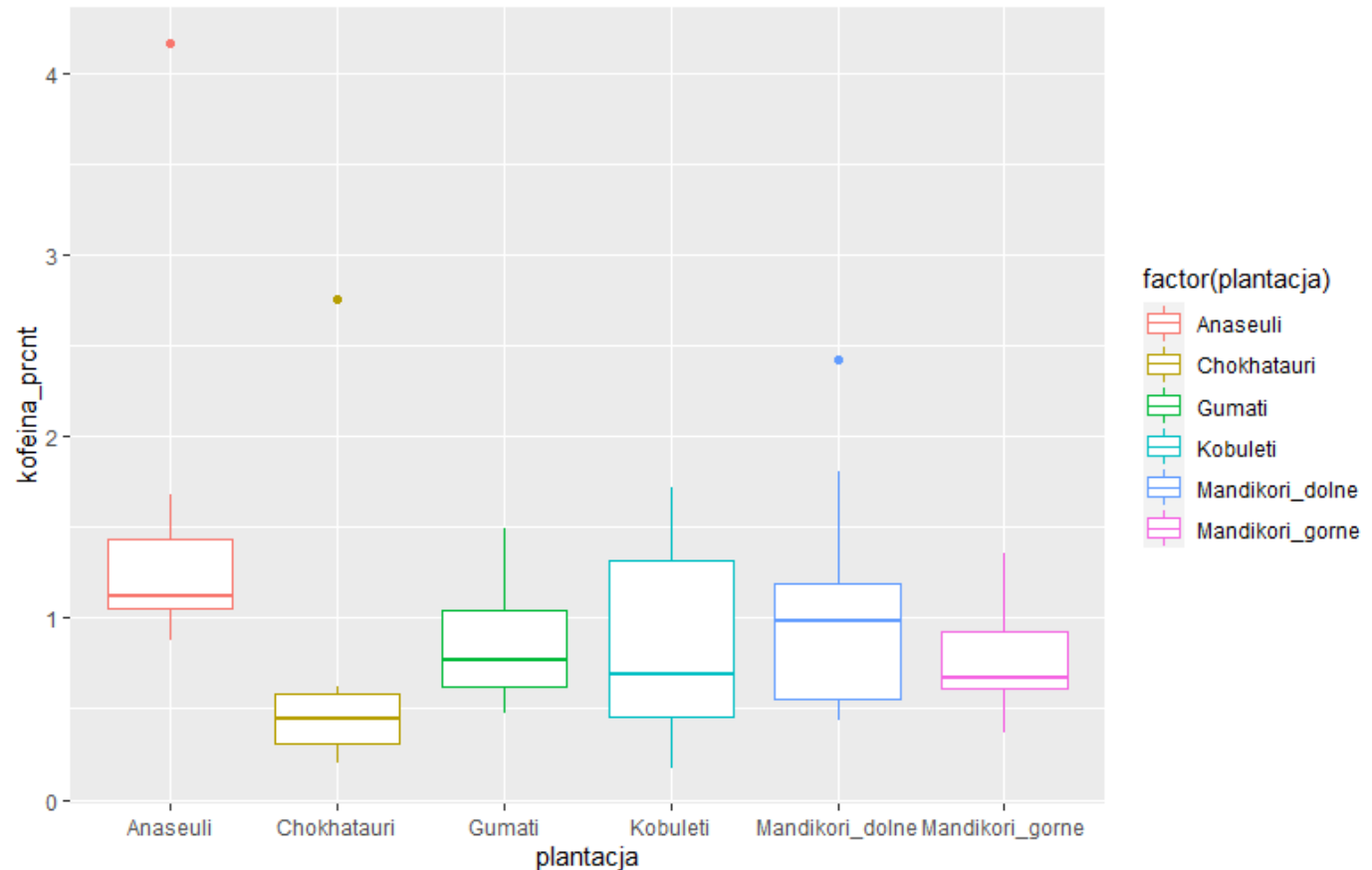
(pakiet `agricolae` i funkcja `kruskal`)

Test post-hoc Kruskala Wallisa

Casus: Czy % zawartość kofeiny w liściach herbaty różni się w zależności od plantacji (n=6), z których był zbierany materiał?

```
ggplot(kofeina.data, aes(x=plantacja, y=kofeina_prct, col=factor(plantacja)))+  
  geom_boxplot(aes(x=plantacja, y=kofeina_prct, col=factor(plantacja)))
```

Zakładamy, że
zmienna
objaśniana nie
reprezentuje
rozkładu zbliżonego
do normalnego (%
zawartość kofeiny
w liściach –
proporcja i wartości
w zakresie od 0 do
1, czyli znowu beta



Implementacja w R:

```
library(agricolae)
```

```
(kruskal(kofeina.data$kofeina_prcnt,kofeina.data$plantacja, console=TRUE))
```

```
Treatments with the same letter are not significantly different.
```

```
      kofeina.data$kofeina_prcnt groups
Anaseuli                45.7      a
Mandikori_dolne         34.9     ab
Gumati                  31.5      b
Kobuleti                 28.2     bc
Mandikori_gorne         27.3     bc
Chokhatauri             15.4      c
```

```
$statistics
```

```
      chisq Df      p.chisq  t.value      MSD
16.22754    5 0.006223586 2.004879 13.93601
```

```
$parameters
```

```
      test p.adjusted      name.t ntr alpha
Kruskal-wallis      none kofeina.data$plantacja  6 0.05
```

```
$means
```

```
      kofeina.data.kofeina_prcnt rank      std  r      Min      Max      Q25      Q50      Q75
Anaseuli                1.4859681 45.7 0.9709982 10 0.8807143 4.160526 1.0482920 1.1252414 1.4309748
Chokhatauri             0.6391805 15.4 0.7597323 10 0.2000000 2.758442 0.3042013 0.4419466 0.5855836
Gumati                  0.8515180 31.5 0.3227989 10 0.4759207 1.489774 0.6168676 0.7654962 1.0403884
Kobuleti                 0.8683580 28.2 0.5629786 10 0.1757225 1.719512 0.4518889 0.6890466 1.3175000
Mandikori_dolne         1.0640055 34.9 0.6358308 10 0.4342003 2.424851 0.5510269 0.9876085 1.1850083
Mandikori_gorne         0.7649301 27.3 0.3021265 10 0.3710526 1.356000 0.6103379 0.6660609 0.9242852
```

```
$comparison
```

```
NULL
```

```
$groups
```

```
      kofeina.data$kofeina_prcnt groups
Anaseuli                45.7      a
Mandikori_dolne         34.9     ab
Gumati                  31.5      b
Kobuleti                 28.2     bc
Mandikori_gorne         27.3     bc
Chokhatauri             15.4      c
```

```
attr(,"class")
```

```
[1] "group"
```

W praktyce nie jest tak różowo...

Często pracujemy z wieloma zmiennymi jednocześnie, a predyktor jest tylko jeden, np. czas, typ lasu, typ siedliska, itd.

Najpewniej będzie tak, że różne zmienne objaśniane będą reprezentować różne rozkłady...

Co wtedy robić?

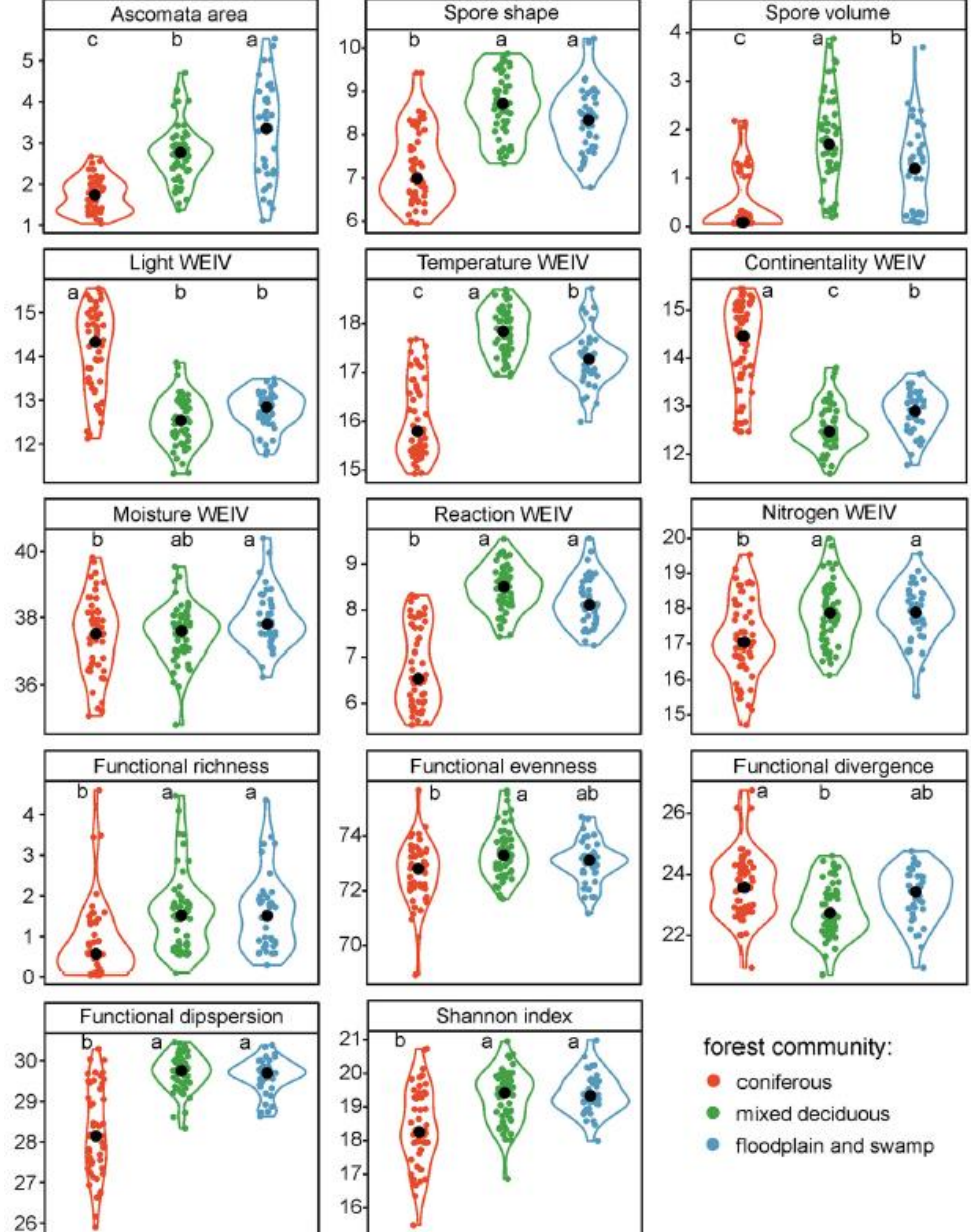
Jaki typ rozkładu założyć?

Czy dla każdej zmiennej osobno, tak jak to wynika z rozkładu gęstości prawdopodobieństwa?

A może przyjąć jeden typ rozkładu dla całości?

Casus: różnice w różnorodności funkcjonalnej porostów w Puszczy Białowieskiej w zależności od typu lasu

zarzut recenzenta po popatrzeniu na violin ploty – jaki typ rozkładu dobraliście skoro większość zmiennych objaśnianych mniej lub bardziej odbiega od klasycznego rozkładu normalnego?



Forest Ecology and Management 475 (2020) 118434

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco



Identifying mechanisms shaping lichen functional diversity in a primeval forest

Anna Łubek^{a,*}, Martin Kukwa^b, Bogdan Jaroszewicz^c, Patryk Czortek^c

^a The Jan Kochanowski University in Kielce, Institute of Biology, Division of Environmental Biology, Uniwersytecka 7, PL-25-406 Kielce, Poland
^b University of Gdańsk, Faculty of Biology, Department of Plant Taxonomy and Nature Conservation, Wita Stwosza 59, PL-80-308 Gdańsk, Poland
^c University of Warsaw, Faculty of Biology, Białowieża Geobotanical Station, Sportowa 19, PL-17-230 Białowieża, Poland

Odpowiedź na ten zarzut wyglądała następująco:

Założyliśmy wszędzie rozkład normalny - Dlaczego?

Bo gdybyśmy każdą zmienną potraktowali indywidualnie, trudno by było porównywać wyniki między sobą

Byłyby trudności z poprawną diagnozą niektórych dziwnych rozkładów, które reprezentują niektóre zmienne objaśniane

Wniosek – czasem prostota jest lepsza niż wybrzydzenie

Czasem tracimy na jakości, ale zyskujemy generalizację i lepszą „przenośność”/ekstrapolację wyników



Identifying mechanisms shaping lichen functional diversity in a primeval forest

Anna Łubek^{a,*}, Martin Kukwa^b, Bogdan Jaroszewicz^c, Patryk Czortek^c



Casus: cechy funkcjonalne liści herbaty z plantacji w Gruzji (region Batumi)

Wysokość roślin – trochę przypomina rozkład beta, trochę Poisson

%Zawartość kofeiny – trochę log-normal, trochę Poisson, trochę beta

Sucha masa i powierzchnia liści – tu jest w miarę ok, mamy krzywą Gaussa

SLA – chyba log-normal?

Co mamy?

Przy SLA wpływ outliera

Przy zawartości kofeiny

„ciężki ogon”

Co można zrobić?

Usunąć outliera i
założyć, że wszędzie nie
ma rozkładu Gaussa

Usunąć outliera i
transformować dane w taki
sposób, aby był rozkład stał
się bardziej zbliżony do
normalnego

