Grzegorz KOŃCZAK

Wizualizacja wyników badań naukowych

Wnioskowanie statystyczne i analizy on-line

Wydawnictwo Uniwersytetu Ekonomicznego
w Katowicach

Grzegorz Kończak

Wizualizacja wyników badań naukowych

Wnioskowanie statystyczne i analizy on-line

Katowice 2026

Praca naukowa

Komitet redakcyjny

Tomasz Ingram (przewodniczący), Beata Kwiecień (sekretarz),
Agata Austen, Monika Kulikowska-Pawlak, Małgorzata Pańkowska,
Aleksandra Pethe, Jacek Pietrucha, Anna Skórska, Marzena Strojek-Filus,
Edyta Szafranek-Stefaniuk, Tomasz Wachowicz, Ewa Ziemba

Recenzent

Jacek Białek

Redakcja i korekta językowa

Karolina Koluch

Skład tekstu

Marzena Safian

Projekt okładki

Grzegorz Kończak

Ilustracje na okładce dostarczone przez Autora

ISBN 978-83-7875-988-1

doi.org/10.22367/uekat.9788378759881

WYDAWNICTWO UNIWERSYTETU EKONOMICZNEGO W KATOWICACH
ul. 1 Maja 50, 40-287 Katowice, tel.: +48 32 257-76-33
www.wydawnictwo.ue.katowice.pl, e-mail: wydawnictwo@ue.katowice.pl
Facebook: @wydawnictwouekatowice

Spis treści

*Jeden **obraz** bywa wart więcej niż tysiąc **słów**.*

Przysłowie chińskie

*Jeden **wykres** bywa wart więcej niż tysiąc **liczb**.*

Parafraza przysłowia

Wprowadzenie

Metody graficzne są kojarzone przede wszystkim z zagadnieniami statystyki opisowej. W tradycyjnym ujęciu pełnią one funkcję pomocniczą – służą do przedstawienia rozkładów, relacji między zmiennymi czy wykrywania obserwacji odstających. Ich rola ogranicza się często do wstępnej analizy danych, która poprzedza właściwe wnioskowanie statystyczne. W statystyce opisowej wykresy (np. histogramy, wykresy pudełkowe, wykresy rozrzutu) ułatwiają interpretację nawet bardzo złożonych zależności w zbiorach danych, czyniąc wyniki bardziej przystępnymi i zrozumiałymi dla odbiorcy. W tym sensie wykresy są traktowane jako narzędzie eksploracji umożliwiające intuicyjne zrozumienie struktury danych.

Wnioskowanie statystyczne odgrywa jednak kluczową rolę w badaniach naukowych, ponieważ umożliwia uogólnianie wyników uzyskanych z próby na całą populację oraz szacowanie niepewności tych uogólnień. Ta fundamentalna właściwość czyni je niezastąpionym narzędziem w niemal wszystkich dyscyplinach nauki – od medycyny i fizyki, przez psychologię, aż po ekonomię i finanse. Kluczowe znaczenie wnioskowania statystycznego polega również na jego roli w zapewnianiu powtarzalności i weryfikowalności badań naukowych. Poprzez formalizację procedur testowania hipotez i szacowania niepewności metody statystyczne umożliwiają innym badaczom ocenę wiarygodności przedstawionych wyników oraz replikację badań. To z kolei przyczynia się do akumulacji wiedzy naukowej i eliminowania błędnych wniosków. Metody wnioskowania statystycznego nie tylko umożliwiają uogólnianie wyników badań, ale też stały się fundamentalnym narzędziem wspierającym rozwój współczesnej nauki, zapewniając jej rygor metodologiczny, obiektywność i możliwość empirycznej weryfikacji teorii naukowych.

Wraz z rozwojem metod wizualizacji danych, a także coraz większą dostępnością narzędzi obliczeniowych metody graficzne zaczynają odgrywać znacznie ważniejszą rolę również w kontekście wnioskowania statystycznego. Coraz częściej są one wykorzystywane nie tylko w celu zilustrowania wyników, lecz również jako integralny element procesu podejmowania decyzji statystycznych. W szczególności pozwalają zobrazować ocenę istotności statystycznej, wiarygodność szacunków czy porównanie modeli – w sposób bardziej przejrzysty i przystępny niż tradycyjne wskaźniki liczbowe.

Znaczenie metod graficznych we wnioskowaniu statystycznym rośnie, zwłaszcza w kontekście eksploracyjnej analizy danych oraz komunikacji wyników testów statystycznych. Graficzne przedstawienie rozkładów statystyk testowych,

p-wartości czy przedziałów ufności pozwala lepiej zrozumieć mechanizmy losowe, ocenić założenia modeli oraz zilustrować efekty badanych zjawisk. Coraz częściej wizualizacje wspierają interpretację wyników testów permutacyjnych czy bootstrapowych, ułatwiając zarówno eksplorację, jak i prezentację wniosków.

W niniejszej monografii podjęto próbę zwrócenia uwagi na znaczną i coraz większą rolę metod graficznych w zagadnieniach wnioskowania statystycznego. Szczególną uwagę poświęcono ich zastosowaniu w permutacyjnym wnioskowaniu statystycznym opartym na modelu Fishera-Pitmana. Podejście to stanowi interesującą i coraz szerzej akceptowaną alternatywę wobec klasycznego modelu populacyjnego, znanego z ujęcia Neymana-Pearsona.

Permutacyjne testy statystyczne, będące podstawą tego modelu, pozwalają na ocenę istotności bez konieczności przyjmowania założeń o rozkładzie zmiennej losowej w populacji. Dzięki temu są one szczególnie użyteczne w analizach danych o nietypowej strukturze lub w badaniach, w których klasyczne założenia nie są spełnione. Co istotne, permutacyjne metody dają się w sposób naturalny przedstawić i zinterpretować graficznie – np. poprzez rozkład permutacyjny statystyki testowej czy wizualizację p-wartości na podstawie próbkowania permutacyjnego.

Opracowywanie nowych metod badawczych stanowi istotny element pracy naukowej. Ważną rolę w ich analizie i prezentacji odgrywa wizualizacja. Zastosowanie tego narzędzia przedstawiono na przykładzie autorskiego wielowymiarowego testu pustych cel. Takie wizualne podejście nie tylko zwiększa transparentność procesu wnioskowania, ale także ułatwia jego komunikację – zarówno w kontekście dydaktycznym, jak i w prezentacji wyników badań naukowych. W efekcie metody graficzne zyskują na znaczeniu nie tylko jako narzędzie wspomagające interpretację, lecz także jako pełnoprawna część procesu analizy statystycznej.

Monografia ma na celu ukazanie rosnącego znaczenia metod graficznych w kontekście wnioskowania statystycznego, a nie jedynie w ramach klasycznej statystyki opisowej. Zmierza to do wykazania, że wizualizacja danych może pełnić funkcję nie tylko ilustracyjną, lecz także analityczną i inferencyjną, wspierając proces podejmowania decyzji w badaniach statystycznych. Ponadto w pracy ukazano potencjał wizualizacji w ułatwianiu interpretacji i komunikacji wyników wnioskowania statystycznego, zwłaszcza w kontekście metod symulacyjnych, w tym testów permutacyjnych. Wskazano również, że graficzne przedstawienie wyników testów permutacyjnych (np. rozkładów permutacyjnych statystyk testowych, wizualizacji p-wartości) sprzyja transparentności i interpretowalności wyników analiz statystycznych.

W niniejszej monografii wyróżniono sześć rozdziałów. W rozdziale pierwszym przytoczono wybrane historyczne fakty dotyczące wnioskowania staty-

stycznego oraz graficznych prezentacji historycznych związanych z opisem niepewności.

W rozdziale drugim przedstawiono wybrane typy wykresów, które mogą być wykorzystywane w prezentacji wyników wnioskowania statystycznego. Uwzględniono m.in. funkcje rozkładu prawdopodobieństwa, gęstości teoretyczne i empiryczne, wykresy Q-Q plot i P-P plot, a także wykresy z wynikami wnioskowania statystycznego, w tym testów permutacyjnych.

W rozdziale trzecim skoncentrowano się na wizualizacji rozkładów teoretycznych zmiennych losowych oraz prezentacji oszacowań takich rozkładów. Uwzględniono najczęściej stosowane rozkłady teoretyczne, takie jak rozkład normalny, t-Studenta i chi-kwadrat.

W kolejnym rozdziale omówiono sposoby prezentacji wyników wnioskowania statystycznego z wykorzystaniem wybranych bibliotek programu **R**. Wskazano m.in. na wybrane funkcje pakietów: **`gginference`**, **`ggpval`**, **`ggpubr`**, **`ggsignif`**, **`ggstatsplot`**, **`ggdist`** i **`infer`**. Zwrócono również uwagę na potrzebę inicjatywy badacza w poszukiwaniu form wizualizacji „nietypowych" zagadnień związanych z wnioskowaniem statystycznym. W tym przypadku przedstawiono przykłady wizualizacji dotyczących testu pustych cel.

Rozdział piąty to charakterystyka wnioskowania w modelu permutacyjnym. Wnioskowanie to opiera się na modelu Fishera-Pitmana, który cechuje się istotnymi własnościami, a bardzo rzadko jest stosowany w naukach ekonomicznych. Wnioskowanie permutacyjne wspomagano pakietami **`wPerm`**, **`coin`**, **`perm`**, **`resample`**, **`exactRankTests`** i **`infer`**.

Rozdział szósty ukazuje możliwości przeprowadzania analiz statystycznych on-line. W procedurach zaprezentowanych w tym rozdziale odwołano się do danych dostępnych w Banku Danych Lokalnych. Przedstawiono sposoby pobierania danych on-line, metody wizualizacji tych danych oraz realizację wnioskowania permutacyjnego.

W niniejszej monografii przyjęto określoną konwencję zapisu i oznaczeń, których zestawienie przedstawiono poniżej. Kody źródłowe w programie **R** oraz odpowiadające im rezultaty wykonania poleceń są prezentowane w następujący sposób:

```
data(mtcars)
head(mtcars)
```

```
#                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
# Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
# Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
# Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
# Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
# Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
# Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

W pracy odwołano się do różnych zbiorów dostępnych w **R**. Zbiory są oznaczane jako **mtcars**, **iris**. Szczegółowy opis zbiorów wykorzystanych w analizach można znaleźć np. w Wickham (2009), Albert i Rizzo (2012), Kończak (2024a), Kończak i Kosińska (2026). W tekście zastosowano następujące wyróżnienia: nazwy pakietów programu R zapisano jako **`ggplot2`**, zmienne jako *cyl*, natomiast funkcje jako *`ggplot()`*.

Dla skutecznego korzystania z treści jest wymagana znajomość podstaw obsługi środowiska programistycznego **R** oraz systemu RStudio (2026). Niezbędna jest także znajomość podstaw wizualizacji danych, w szczególności pakietu **`ggplot2`** (Wickham, 2009; Kończak, 2024b). Materiały pomocnicze, w tym kody procedur w **R** zamieszczono na stronie internetowej Kończak (2026).

1. Rys historyczny metod wnioskowania statystycznego i wizualizacji niepewności

Świata nie da się zrozumieć bez liczb.
Ale świata nie da się zrozumieć wyłącznie za pomocą liczb.

Hans Rosling (Rosling i in., 2018, s. 176)

Za panowania Henryka II (1154–1189) w Mennicy Królewskiej w Londynie ustanowiono ceremonię *Trial of the Pyx*. Termin „Pyx" w staroangielskim oznacza „skrzynię" lub „puszkę" (Aczel, 2000, s. 670). Ceremonia ta miała wymiar religijny i była związana z oceną jakości złotych monet wykonywanych w mennicy królewskiej. Codziennie jedna z monet ze złota (czasami srebrne monety) wyprodukowanych w mennicy była losowo wybierana (próbkowanie) i umieszczana w skrzyni. Co trzy lub cztery lata otwierano skrzynię, zliczano monety i ustalano ich łączną wagę. Wynikową wagę zestawiono z ustalonym standardem, uwzględniając liczbę monet i określony zakres tolerancji. Dzięki tej ceremonii król mógł nadzorować wykorzystanie królewskiego złota. Gdyby monety wykazywały zbyt dużą wagę, oznaczałoby to nadmierne wyczerpanie złota ze skarbca królewskiego. I odwrotnie, gdyby monety miały niedobór wagi, waluta traciłaby na wartości, a osoba z obsługi mogłaby potencjalnie skorzystać z powierzonego złota. Po pomyślnym zakończeniu próby organizowano wielką ucztę. W przypadku negatywnego wyniku mistrz mennicy mógł trafić do więzienia, a nawet zostać skazany na śmierć. Od 1699 do 1727 roku mistrzem mennicy był wybitny naukowiec Isaac Newton. W tym okresie jedna z weryfikacji zakończyła się niepomyślnie, ale w tych czasach kary nie były już tak surowe, jak we wcześniejszych wiekach (Aczel, 2000, s. 670). Ceremonia *Trial of the Pyx* stanowi niezwykłą ilustrację zastosowania metodologii probabilistycznych i zasad statystycznych setki lat przed formalnym ustanowieniem podstaw rachunku prawdopodobieństwa. W latach dwudziestych XX wieku Walter A. Shewhart zaproponował karty kontrolne jako narzędzie sterowania jakością procesów produkcyjnych (Iwasiewicz i Paszek, 2000; Kończak, 2000, 2007). Idea kart kontrolnych przypomina opisane powyżej rozwiązanie. Karty kontrolne stanowią niezwykłe powiązanie wnioskowania statystycznego i graficznej prezentacji.

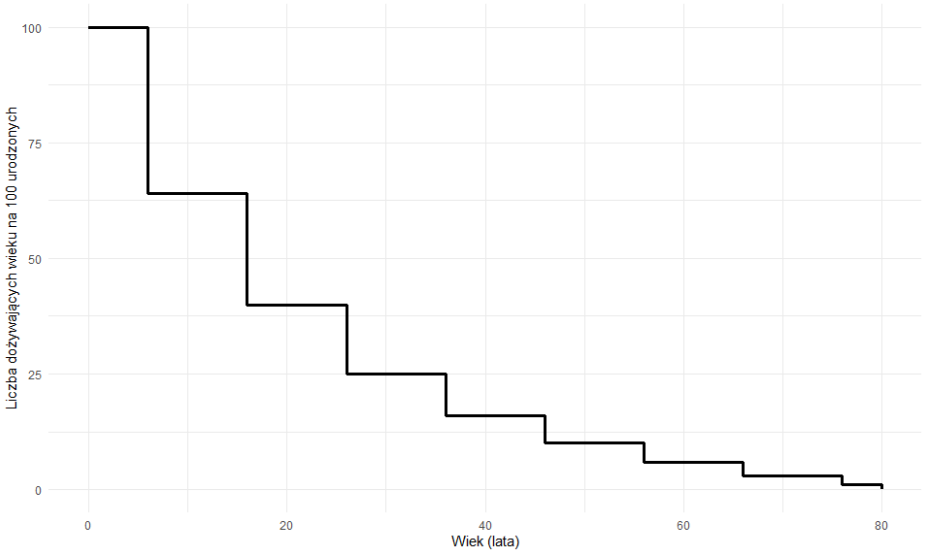
W 1662 roku John Graunt, uznawany za twórcę demografii i statystycznych badań populacji, opublikował książkę *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality* (Bernstein, 1997, s. 70). Zaobserwował, że umieralność w miastach jest wyższa niż na wsi, że więcej rodzi się chłopców niż dziewczynek. Zauważył jednak, że długotrwała równowaga płci była zachowana, ponieważ umieralność mężczyzn jest większa niż kobiet. Opracował także tabelę umieralności w kategoriach długości życia dla wieku od 6 do 76 lat w odstępach dziesięcioletnich (zob. tabela 1.1).

Tabela 1.1. Liczba dożywających danego wieku dla 100 urodzonych

Wiek (lata)	Liczba dożywających
0	100
6	64
16	40
26	25
36	16
46	10
56	6
66	3
76	1
80	0

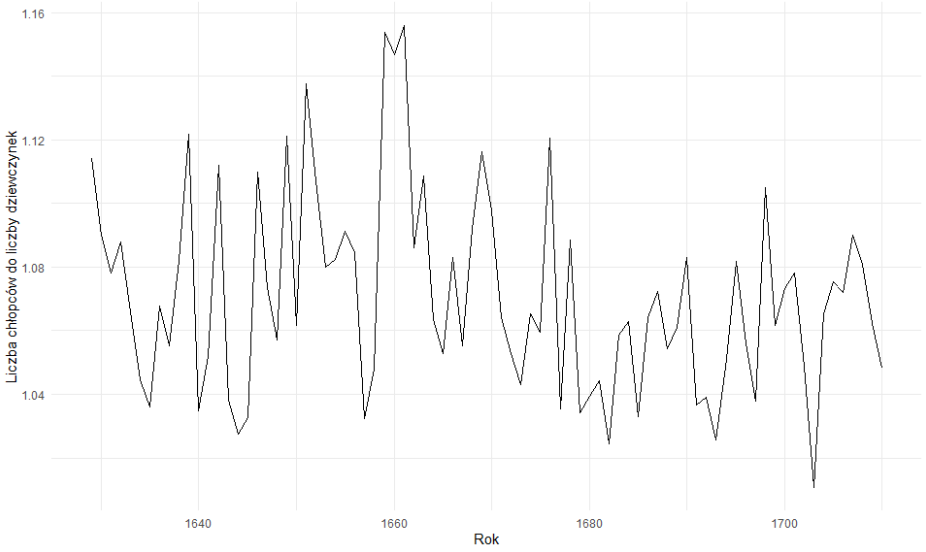
Źródło: Opracowanie własne na podstawie: Bernstein (1997).

W tabeli 1.1 przedstawiono wyniki analiz przeprowadzonych przez Johna Graunta w 1662 roku, a na rys. 1.1 wyniki te zostały zobrazowane graficznie. O ile John Graunt formułował pewne tezy, to jednak nie pojawiał się w jego zapisach formalny zapis hipotez statystycznych. Pierwszy test hipotezy zerowej, dotyczącej proporcji liczby urodzeń chłopców i dziewczynek w Londynie, pojawia się w pracy Johna Arbuthnota z 1710 roku. John Arbuthnot, szkocki fizyk i matematyk, w 1710 roku w artykule *An Argument for Divine Providence* (Arbuthnot, 1710) przeanalizował rejestry urodzeń w Londynie dla każdego z 82 lat od 1629 do 1710 roku oraz stosunek liczby urodzeń chłopców do liczby urodzeń dziewczynek. W wyniku przeprowadzonych badań stwierdził, że w każdym z analizowanych lat liczba urodzeń chłopców w Londynie była wyższa niż liczba urodzeń dziewczynek. Ta publikacja z 1710 roku jest prawdopodobnie najwcześniejszym rozważaniem o formie badawczej przypominającej hipotezę statystyczną (Zar, 2010; Baszczyńska, 2016).



Rys. 1.1. Wizualizacja przeżywalności na podstawie danych z tablicy Graunta

Źródło: Opracowanie własne na podstawie danych z tabeli 1.1.



Rys. 1.2. Stosunek liczby urodzonych chłopców do liczby urodzonych dziewczynek w Londynie (1629–1710)

Rysunek 1.2 przedstawia stosunek liczby urodzonych chłopców do liczby urodzonych dziewczynek w Londynie w latach 1629–1710. Widoczne jest, że każdego roku analizowany wskaźnik przyjmuje wartości większe od 1,00, a dość często większe nawet od 1,10.

Adolphe Quetelet zaproponował szerokie, systematyczne stosowanie statystyki jako metody badawczej do nauk społecznych. Zbierał dane o społeczeństwie i określał rozkład częstości występowania poszczególnych odmian wartości zjawiska (Tarka i Olszewska, 2018). Quetelet w 1844 roku zadziwił statystyków, porównując rozkład wzrostu tych, którzy usłuchali wezwania do poboru, z aktualnym rozkładem wzrostu w populacji generalnej (Rao, 1994, s. 55). Na tej podstawie obliczył, że ponad 2000 mężczyzn uchyliło się od poboru, pozorując, że ich wzrost jest niższy od minimalnego, kwalifikującego do poboru. To bez wątpienia interesujący pokaz praktycznego zastosowania metod wnioskowania statystycznego.

Porównanie liczby urodzeń chłopców i dziewczynek, podobnie jak wcześniej zrealizował to Arbuthnot, ale dla wielu miast Europy, w 1778 roku przeprowadził Pierre-Simon Laplace. Laplace zebrał i przeanalizował dane z wielu miast europejskich, aby sprawdzić, czy ta nierównowaga jest zjawiskiem powszechnym czy też specyficznym dla Londynu. Na podstawie dostępnych danych Laplace obliczył, że prawdopodobieństwo urodzenia chłopca jest większe niż $\frac{1}{2}$ (Laplace, 1820). Jego analiza była jednym z pierwszych formalnych zastosowań statystyki do wnioskowania o populacji na podstawie danych z próby, a jej wniosek był sformułowany w kategoriach ścisłych prawdopodobieństw. Jednocześnie po raz pierwszy sformalizował ideę hipotezy zerowej.

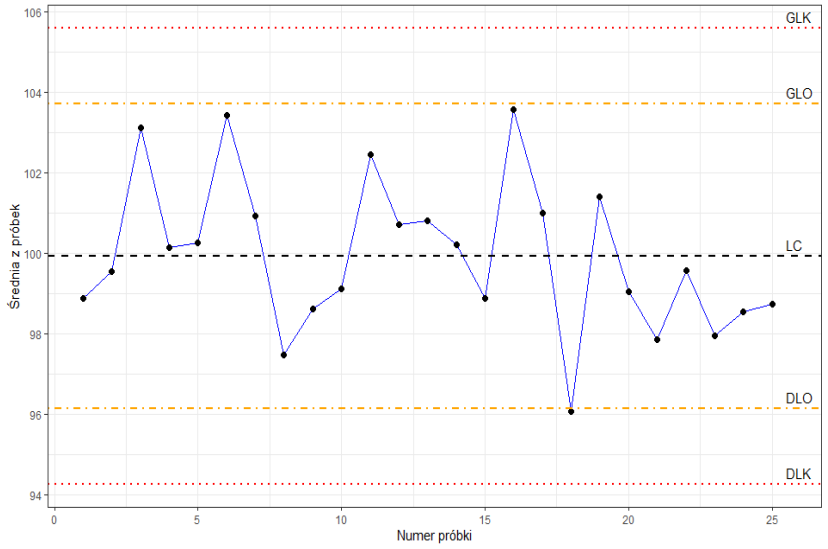
Historycznie za pierwszy formalnie opracowany test statystyczny uznaje się test chi-kwadrat, zaproponowany przez Karla Pearsona w 1900 roku (Pearson, 1900; Plackett, 1983). Test ten jest wykorzystywany do oceny zgodności rozkładu empirycznego z rozkładem teoretycznym oraz do badania niezależności zmiennych losowych w tablicach kontyngencji. Był to przełomowy moment w rozwoju statystyki, ponieważ umożliwił matematyczne weryfikowanie hipotez na podstawie danych z próby, a nie tylko ich opis.

W 1908 roku William Sealy Gosset – statystyk i piwowar z zakładów Guinness w Dublinie – opublikował pod pseudonimem „Student" pracę, w której przedstawił nową metodę wnioskowania dla niewielkich prób pomiarowych. Rezultaty tej pracy do dziś są powszechnie znane jako test t-Studenta i rozkład t-Studenta.

W 1920 roku Jarl Waldemar Lindeberg oraz Paul Lévy sformułowali centralne twierdzenie graniczne w nowoczesnej postaci. Twierdzenie głosiło, że suma dużej liczby niezależnych zmiennych losowych, nawet o różnych rozkładach, ma w przybliżeniu rozkład normalny – pod warunkiem spełnienia dość ogólnych warunków dotyczących momentów tych zmiennych.

Jerzy Spława-Neyman wykorzystał metody permutacyjne w swojej rozprawie doktorskiej z 1923 roku poświęconej badaniom nad doświadczeniami w rolnictwie (Spława-Neyman, 1923). Publikacja ta długo pozostawała nieznana światowej społeczności statystyków, ponieważ na język angielski została przetłumaczona dopiero w 1990 roku (Lehmann i Romano, 2005, s. 210; Bednarski, 2013, s. 13; Berry i in., 2014, s. 14). Jest to pierwsze historyczne zastosowanie modelu permutacyjnego na potrzeby wnioskowania statystycznego.

Zaproponowane przez Waltera A. Shewharta w 1924 roku karty kontrolne stały się kluczowym narzędziem w sterowaniu jakością procesów produkcyjnych. Karty kontrolne to faktycznie sekwencje testów statystycznych, ich siła tkwi przede wszystkim w przejrzystej prezentacji graficznej, umożliwiającej szybkie wykrywanie odchyleń (Iwasiewicz i Paszek, 2000; Kończak, 2000, 2007). Transformują one rygorystyczne, matematyczne procedury testowania hipotez w intuicyjny język wizualny (rys. 1.3) zrozumiały dla każdego operatora procesu. Wykreślone granice kontrolne nie są jedynie liniami pomocniczymi, lecz geometryczną reprezentacją przedziałów ufności, co sprawia, że każdy punkt naniesiony na wykres staje się natychmiastowym sprawdzianem stabilności procesu produkcyjnego, niewymagającym od użytkownika wykonywania skomplikowanych obliczeń. Dzięki tej syntezie statystyka przestaje być abstrakcyjną teorią liczb, a staje się dynamicznym obrazem, gdzie wyjście poza granice kontrolne lub nielosowy układ punktów jest w istocie graficznym sygnałem odrzucenia hipotezy o prawidłowym przebiegu procesu. Niezwykle skuteczne w następnych dziesięcioleciach karty kontrone poniekąd miały swój pierwowzór we wcześniej wspomnianej tradycji *Trial of the Pyx*.



Rys. 1.3. Karta kontrolna $\bar{X}$

Kolejnym ważnym etapem na drodze popularyzacji i formalizacji metod wnioskowania statystycznego było wprowadzenie poziomu istotności α (zazwyczaj 0,05; 0,01 lub 0,10) oraz testu F (analiza wariancji, ANOVA) przez Ronalda Aylmera Fishera w 1925 roku. W tym samym czasie Fisher (1925) przedstawił ideę testów dokładnych i testów permutacyjnych. Idea ta została szeroko opisana z przykładem zastosowania w eksperymencie *The Lady Tasting Tea* (Salsburg, 2001). Berry i in. (2014, s. 3) wskazują, że znaczny wkład w rozwój teorii testów permutacyjnych wnieśli także Robert Charles Geary (1927) oraz Thomas Eden i Frank Yates (1933). Kluczowe dla sformalizowania testowania hipotez były prace Jerzego Neymana i Egona Pearsona (1928a, 1928b), w których zostały wprowadzone pojęcia błędów I i II rodzaju, mocy testu oraz poziomu istotności (α), a także kompleksowe powiązanie tych pojęć. Model populacyjny wnioskowania statystycznego przedstawiony przez Neymana i Pearsona stał się kluczowym modelem, który jest nauczany praktycznie na wszystkich kursach statystyki matematycznej.

Sformalizowanie zaproponowanego wcześniej modelu permutacyjnego przedstawił w latach 1937–1938 Edwin J. G. Pitman (1937a, 1937b, 1938). Mimo znacznej atrakcyjności tej formy wnioskowania statystycznego w tamtych czasach nie mógł on znaleźć szerokiego praktycznego zastosowania ze względu na brak odpowiednich możliwości obliczeniowych.

Jerzy Neyman (1937) wprowadził pojęcie przedziałów ufności dla parametrów populacji. Jego praca była kluczowa w rozwoju teorii estymacji i wnioskowania statystycznego. Przedziały ufności umożliwiają wizualizację i komunikację niepewności pomiarów, co jest szczególnie ważne w naukach empirycznych.

W 1947 roku Abraham Wald wprowadził testy sekwencyjne. Tradycyjne testy statystyczne wymagają z góry ustalonej wielkości próby. Dopiero po zebraniu wszystkich danych przeprowadza się test i podejmuje decyzję. Wald zaproponował test sekwencyjny, w którym dane są analizowane w miarę ich napływania, a decyzja o przyjęciu lub odrzuceniu hipotezy może być podjęta w dowolnym momencie, gdy zgromadzone argumenty są wystarczająco silne (Wald, 1947).

W 1958 roku Edward L. Kaplan i Paul Meier opublikowali pracę, która wprowadziła do statystyki estymator Kaplana-Meiera – metodę do estymacji funkcji przeżycia (ang. survival function) w obecności cenzurowanych danych. Jest to obecnie ważne narzędzie w analizie danych przeżycia, szczególnie w medycynie, biologii i innych dziedzinach, gdzie badane zdarzenia mogą być niepełnie zaobserwowane (Kaplan i Meier, 1958).

Problem jednoczesnego testowania wielu hipotez (testowanie wielokrotne) rozważał John W. Tukey (1977). Przedstawił też w 1970 roku konstrukcję wykresów pudełkowych, które z czasem stały się powszechnym narzędziem do wizua-

lizacji zmienności i niepewności. We wnioskowaniu statystycznym szczególnie pomocne mogą być wykresy pudełkowe z wcięciami (boxplot with notched). Wykresy pudełkowe z nacięciami wykorzystują „nacięcie" lub zwężenie pudełka wokół mediany. Wycięcia są przydatne w dostarczaniu wskazówek dotyczących istotności różnicy median; jeśli wycięcia dwóch pudełek nie nakładają się na siebie, stanowi to dowód statystycznie istotnej różnicy między medianami.

Powyższy przegląd ukazuje, że formalizacja statystyki rozwijała się równolegle z potrzebami cywilizacyjnymi, a jej związki z praktyką – od mennictwa po przemysł, rolnictwo, medycynę i psychologię – pokazują uniwersalność i znaczenie metod statystycznych w rozumieniu współczesnego świata. W tabeli 1.2 przedstawiono wybrane fakty o kluczowym znaczeniu z prac nad rozwojem teorii wnioskowania statystycznego i wizualizacji niepewności. Część z tych możliwości zostanie przedstawiona w kolejnych rozdziałach.

Tabela 1.2. Chronologia najważniejszych testów statystycznych i wizualizacji niepewności

Rok	Test/statystyka/pojęcie	Twórca/autor
1662	Analiza urodzeń w Londynie	John Graunt
1710	Pierwszy test hipotezy zerowej	John Arbuthnot
1778	Formalizacja hipotezy zerowej	Pierre-Simon Laplace
1900	Test chi-kwadrat	Karl Pearson
1908	Test *t*-Studenta	William S. Gosset („Student")
1920	Centralne twierdzenie graniczne	Waldemar Lindeberg, Paul Lévy
1923	Zastosowanie metod permutacyjnych	Jerzy Spława-Neyman
1924	Wprowadzenie kart kontrolnych	Walter A. Shewhart
1925	Testy F, ANOVA, poziom istotności	Ronald A. Fisher
1925	Testy dokładne, testy permutacyjne	Ronald A. Fisher
1933	Formalizacja testowania hipotez	Jerzy Neyman, Egon Pearson
1937	Formalizacja idei testów permutacyjnych	Edwin J. G. Pitman
1937	Przedziały ufności	Jerzy Neyman
1947	Testy sekwencyjne	Abraham Wald
1955	Test Wilcoxona	Frank Wilcoxon
1956	Test Kruskala-Wallisa	William Kruskal, W. Allen Wallis
1958	Estymator Kaplana-Meiera	Edward L. Kaplan, Paul Meier
1970	Wykresy pudełkowe (boxplot)	John W. Tukey
1977	Testy wielokrotnych porównań	John W. Tukey
1979	Metody bootstrapowe	Bradley Efron
1999	Gramatyka grafiki	Leland Wilkinson
2007	Pakiet `ggplot2`	Hadley Wickham

Źródło: Opracowanie własne na podstawie: Lange i Banasiński (1970); Kotz i in. (2006); Kończak i Złotoś (2026).

Rozwój techniki obliczeniowej i grafiki cyfrowej otworzył nowe możliwości w wizualizacji niepewności. Możliwe stało się konstruowanie złożonych i dynamicznych wizualizacji. Pojawiły się możliwości konstrukcji animacji, interaktywnych dashboardów i bardziej zaawansowane techniki graficzne, aby skuteczniej przekazywać informacje o niepewności.

W wizualizacji danych milowym krokiem było zaproponowanie przez Lelanda Wilkinsona (Wilkinson i Wills, 2005) idei gramatyki grafiki (*Grammar of Graphics*) w 1999 roku, a następnie zrealizowanie jej w środowisku **R** w 2007 roku w pakiecie **ggplot2** przez Hadleya Wickhama (2009). W kolejnych latach opracowano wiele bibliotek dla programu **R** pozwalających na konstrukcję złożonych wizualizacji, w szczególności na potrzeby wnioskowania statystycznego. Wizualizacje wyników analiz on-line i wnioskowania statystycznego prezentowane w kolejnych rozdziałach wykorzystują wspomnianą ideę i pakiet **ggplot2**.

2. Charakterystyka wybranych metod graficznych stosowanych we wnioskowaniu statystycznym

> *Doskonałość graficzna polega na przekazywaniu złożonych idei w sposób jasny, precyzyjny i skuteczny.*
>
> Edward R. Tufte (2013, s. 51)

Wykresy statystyczne są głównie wykorzystywane w charakterystyce wstępnej zbioru danych, a także przy prezentacji wyników analiz statystycznych. Dobrze dobrana grafika statystyczna może także wspomagać różne etapy wnioskowania statystycznego.

Wybrane rodzaje metod wizualizacji danych przedstawiono w Kończak (2024a, s. 35–71). W tej części monografii skoncentrowano się na przedstawieniu zwięzłej charakterystyki wybranych typów wykresów związanych z rozkładami zmiennych losowych, estymacją nieparametryczną, konstrukcją przedziałów ufności i weryfikacją hipotez statystycznych, także dla wnioskowania permutacyjnego.

W niniejszym rozdziale przedstawiono zwięzłą charakterystykę następujących typów prezentacji graficznych:

- wykres rozkładu prawdopodobieństwa (zmienna losowa skokowa);
- wykres gęstości prawdopodobieństwa (zmienna losowa ciągła);
- wykres dystrybuanty;
- wykresy przedziałów ufności, wstęga ufności;
- wykresy ocen estymacji nieparametrycznej;
- histogram z dopasowaną funkcją gęstości;
- wykresy rozrzutu z gęstościami brzegowymi;
- wykresy dla przedziałów ufności w analizie regresji;
- wykresy dla przedziałów prognoz;
- Q-Q plot;
- P-P plot;
- wykresy dla testów permutacyjnych.

Przed prezentacją charakterystyki wskazanych powyżej rodzajów wykresów zostaną przedstawione kluczowe określenia (Trzpiot i Kończak, 2002; Zeliaś i in., 2002):

Rozkład prawdopodobieństwa zmiennej losowej dyskretnej *X* może być zadany następująco:

$$P(X = x_i) = p_i \text{ dla } i = 1,2, \dots, k \text{ lub } i = 1,2, \dots \tag{2.1}$$

gdzie $p_i > 0$ oraz $\sum_{i=1}^{n} p_i = 1 \; lub \; \sum_{i=1}^{\infty} p_i = 1$.

Gęstość prawdopodobieństwa zmiennej losowej ciągłej to funkcja $f: \mathbb{R} \rightarrow$ $\rightarrow [0, \infty)$ spełniająca następujące warunki:

1. $f(x) \geq 0$, dla $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f\,(x)dx = 1$.

Dystrybuanta zmiennej losowej *X* to funkcja $F: \mathbb{R} \rightarrow [0,1]$ zadana następującym wzorem:

$$F(x) = P(X < x) \text{ dla } x \in \mathbb{R} \tag{2.2}$$

Tak określona funkcja $F(x)$ zawsze spełnia następujące warunki:

1. $0 \leq F(x) \leq 1$.
2. $F(x)$ jest funkcją niemalejącą.
3. $F(x)$ jest funkcją lewostronnie ciągłą.
4. $\lim_{x \to -\infty} F(x) = 0$ oraz $\lim_{x \to \infty} F(x) = 1$.

Należy zaznaczyć, że w niektórych źródłach lub kontekstach definicja dystrybuanty wykorzystuje nierówność słabą ($F(x) = P(X \leq x)$). Różnica w tych określeniach jest nieistotna dla rozkładów ciągłych, ale istotna dla rozkładów dyskretnych.

Przedziałem ufności dla nieznanego parametru θ na poziomie ufności $1 - \alpha$ jest przedział losowy $[L, U]$, którego krańce są statystykami z próby, spełniający warunek, że prawdopodobieństwo pokrycia prawdziwej wartości parametru jest co najmniej równe $1 - \alpha$, czyli:

$$P(L \leq \theta \leq U) \geq 1 - \alpha \tag{2.3}$$

Estymacja nieparametryczna zajmuje się szacowaniem właściwości rozkładu prawdopodobieństwa (np. jego gęstości lub dystrybuanty) bez wcześniejszych założeń o jego konkretnej, parametrycznej formie. Głównym atutem tej formy estymacji jest uniwersalność, ponieważ model jest dopasowywany bezpośrednio do obserwacji, a nie do z góry przyjętej teorii. Dla nieznanej funkcji gęstości $f(x)$ nieparametryczną ocenę można uzyskać, wykorzystując estymator jądrowy postaci (Domański i Pruska, 2000; Baszczyńska, 2016):

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \tag{2.4}$$

gdzie:
$K(x)$ – funkcja jądra,
$h > 0$ – parametr wygładzania (szerokość okna),
n – liczebność próby.

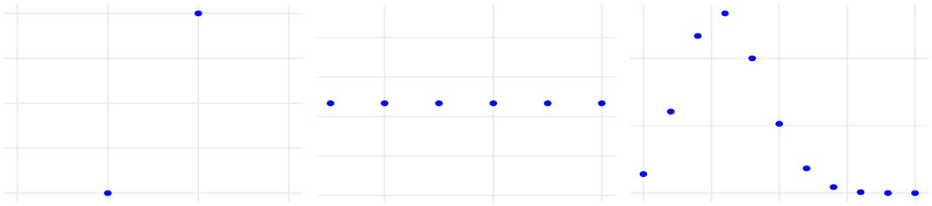
Testy permutacyjne to alternatywne podejście do wnioskowania statystycznego opartego na klasycznym podejściu Neymana-Pearsona (Kończak, 2016, 2020b). W testach permutacyjnych kluczową rolę odgrywa Achieved Significance Level (*ASL*) zadany następująco:

$$ASL \approx \frac{card\{i:T_i \geq T_0\}}{N} \tag{2.5}$$

gdzie:
T_0 – wartość statystyki testowej na podstawie próby,
T_i – wartości statystyki testowej dla i-tej permutacji ($i = 1, 2, ..., N$),
N – liczba permutacji zbioru danych.

Aby go wyznaczyć, konstruuje się empiryczny rozkład statystyki testowej, symulując dane przy założeniu prawdziwości hipotezy zerowej (H_0). Wizualizacja tego rozkładu pozwala zobaczyć, jaki odsetek wygenerowanych wartości (*ASL*) jest co najmniej tak skrajny, jak obserwowana wartość statystyki testowej. W literaturze *ASL* często jest określane jako *p*-wartość, tak jak w modelu klasycznym.

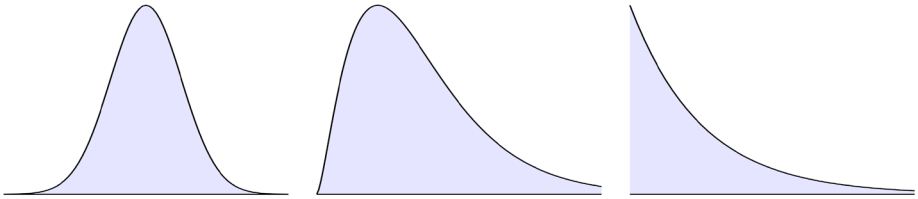
2.1. Wykresy rozkładu prawdopodobieństwa



Rozkład prawdopodobieństwa zmiennej losowej skokowej (dyskretnej) to funkcja, która przypisuje każdej możliwej wartości tej zmiennej odpowiednie prawdopodobieństwo jej wystąpienia. Suma wszystkich prawdopodobieństw rozkładu zmiennej skokowej jest równa 1. Rozkład ten opisuje, z jakim prawdopodobieństwem zmienna losowa przyjmie konkretną wartość spośród wszystkich możliwych wyników w danym eksperymencie losowym. Na tej podstawie można określić, jak często poszczególne zdarzenia pojawiają się w całej przestrzeni zdarzeń. Rozkład prawdopodobieństwa stanowi podstawę do analizowania i przewidywania zachowania zmiennej losowej oraz podejmowania decyzji na podstawie jej możliwych wyników.

Wykresy rozkładu prawdopodobieństwa dla zmiennej losowej skokowej są bardzo pomocnym narzędziem do wizualizacji, które pozwala szybko określić, jak rozkładają się wartości tej zmiennej oraz jakie jest prawdopodobieństwo ich wystąpienia. Dzięki takim wykresom można łatwo dostrzec, które wartości są najczęstsze i mają największe prawdopodobieństwo, a które pojawiają się rzadziej i są stosunkowo mało prawdopodobne. Takie wykresy ułatwiają także identyfikację zakresu zmienności zmiennej losowej, pokazując, na jakim obszarze wartości skupiają się dane i jak rozkłada się prawdopodobieństwo pomiędzy nimi. Pozwala to lepiej zrozumieć charakterystykę zmiennej, zwłaszcza w kontekście analizy ryzyka, podejmowania decyzji czy modelowania statystycznego. Wizualizacja rozkładu prawdopodobieństwa dla zmiennej skokowej umożliwia szybkie wyciąganie wniosków i wspiera interpretację wyników, co jest szczególnie istotne przy prezentacji danych osobom podejmującym decyzje na podstawie analiz statystycznych. Rozkład ten opisuje, z jakim prawdopodobieństwem zmienna losowa przyjmuje poszczególne wartości w danym eksperymencie losowym. Na jego podstawie można analizować zachowanie zmiennej losowej, obliczać miary położenia i zróżnicowania oraz formułować wnioski probabilistyczne i statystyczne.

2.2. Wykresy gęstości prawdopodobieństwa

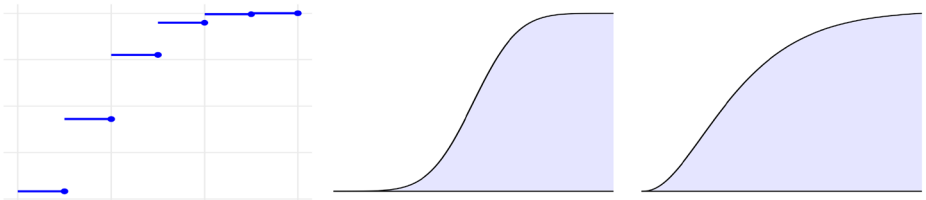


Gęstość prawdopodobieństwa to funkcja matematyczna, która opisuje, jak rozkłada się prawdopodobieństwo wartości zmiennej losowej ciągłej na całym jej zakresie. Pole pod całą krzywą funkcji gęstości jest równe 1. W przeciwieństwie do zmiennych dyskretnych, gdzie prawdopodobieństwo jest przypisane konkretnym, pojedynczym wartościom, zmienne ciągłe mogą przyjmować nieskończenie wiele wartości w określonym przedziale, przez co prawdopodobieństwo wystąpienia dokładnie jednej, konkretnej wartości zawsze jest równe zero. Gęstość prawdopodobieństwa wskazuje, jak bardzo prawdopodobieństwo jest „skoncentrowane" w różnych częściach zakresu zmiennej. Dzięki temu można za jej pomocą obliczyć prawdopodobieństwo, że zmienna losowa przyjmie wartość z określonego przedziału, a miernikiem jest pole obszaru pod krzywą funkcji gęstości w tym przedziale. Pole obszaru oblicza się za pomocą całki oznaczonej z funkcji gęstości w zadanym przedziale. Funkcja gęstości prawdopodobieństwa nigdy nie może przyjmować wartości ujemnych ($f(x) \geq 0$). Wynika to wprost z aksjomatów prawdopodobieństwa, które nie dopuszczają istnienia ujemnej szansy na zajście jakiegokolwiek zdarzenia.

Funkcja gęstości pełni rolę narzędzia do analizy i opisu rozkładu zmiennych ciągłych, umożliwiając zrozumienie, gdzie w przestrzeni zdarzeń wartości skupia się największa część prawdopodobieństwa oraz jak rozkład ten może wyglądać w praktyce. Jest to kluczowe dla dalszych analiz statystycznych, w tym estymacji, testowania hipotez i modelowania statystycznego.

Wykresy gęstości prawdopodobieństwa dla zmiennej losowej ciągłej pozwalają na szybką ocenę zakresu zmienności danej zmiennej, wskazanie zakresów wartości najbardziej i stosunkowo mało prawdopodobnych.

2.3. Wykresy dystrybuanty



Dystrybuanta to podstawowa funkcja w teorii prawdopodobieństwa i statystyce, która przypisuje każdej liczbie rzeczywistej x wartość określającą prawdopodobieństwo, że zmienna losowa X przyjmie wartość mniejszą od x. Dystrybuanta zawiera pełną informację o tym, jak rozkładają się wartości danej zmiennej losowej. Jest to funkcja niemalejąca i przyjmuje wartości od 0 do 1. Dystrybuanta jest niezbędnym narzędziem do opisu rozkładu zmiennych losowych zarówno dyskretnych, jak i ciągłych, ponieważ umożliwia wyznaczenie prawdopodobieństw realizacji dla różnych przedziałów wartości zmiennej. W przypadku zmiennych losowych dyskretnych dystrybuanta ma charakter skokowy – wartości prawdopodobieństw rosną skokowo w punktach odpowiadających możliwym realizacjom zmiennej. Dla zmiennych losowych ciągłych dystrybuanta jest funkcją ciągłą, której pochodna (jeśli istnieje) jest funkcją gęstości prawdopodobieństwa.

Dystrybuanta znajduje szerokie zastosowanie w analizie statystycznej, m.in. do obliczania prawdopodobieństw zdarzeń, testowania hipotez czy też budowania modeli probabilistycznych. Dzięki niej można również łatwo porównywać różne rozkłady, badać własności zmiennych losowych oraz wykonywać symulacje. Jest fundamentem wielu narzędzi statystycznych i probabilistycznych, stanowiąc swego rodzaju pomost między surowymi danymi a ich probabilistyczną interpretacją. Można porównać dystrybuanty dwóch zmiennych, by ocenić, która zmienna przyjmuje częściej określone wartości, która charakteryzuje się większym rozproszeniem. Dystrybuanta odgrywa kluczową rolę w generowaniu wartości losowych, w tym np. metodą odwracania dystrybuanty. Dystrybuanta umożliwia także wyznaczanie kwantyli, takich jak mediana czy kwartyle, które służą do opisu położenia i rozproszenia rozkładu. Z tych powodów dystrybuanta odgrywa ważną rolę w symulacjach komputerowych.

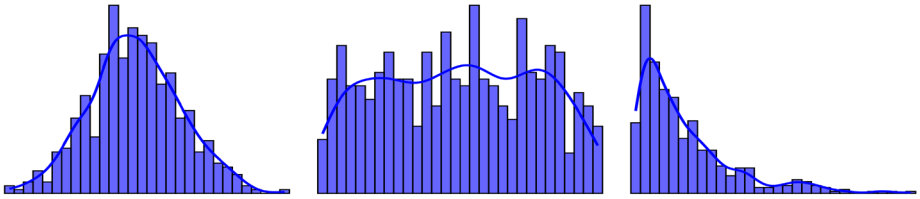
2.4. Wykresy dla estymacji nieparametrycznej

Estymacja nieparametryczna to zbiór metod statystycznych, które pozwalają na oszacowanie funkcji rozkładu prawdopodobieństwa lub innych charakterystyk populacji bez przyjmowania założeń dotyczących konkretnego typu rozkładu (np. normalnego, wykładniczego). Oznacza to, że w odróżnieniu od metod parametrycznych, które opierają się na założeniu, że dane pochodzą z określonego typu rozkładu i są w pełni opisane przez niewielką liczbę parametrów, metody nieparametryczne są znacznie bardziej elastyczne i dostosowują się bezpośrednio do struktury obserwowanych danych. Elastyczność ta ma jednak swoją cenę – metody nieparametryczne zazwyczaj wymagają znacznie większych prób badawczych niż ich parametryczne odpowiedniki. W przypadku zbyt małej liczby obserwacji wygenerowana ocena rozkładu może być niewystarczająca.

Estymacja nieparametryczna funkcji gęstości pozwala na ocenę charakterystyk populacji. W szczególności można oszacować wartość oczekiwaną, wariancję zmiennej losowej, a także ocenić asymetrię rozkładu. Metody te nie wymagają wcześniejszego określenia, z jakiego rozkładu mogą pochodzić dane. Estymacja nieparametryczna cechuje się znaczną odpornością na błędy wynikające z niewłaściwego doboru modelu, ponieważ nie wymaga wcześniejszej specyfikacji rozkładu. W praktyce oznacza to, że metody te są szczególnie przydatne, gdy nie ma wystarczającej wiedzy na temat rozkładu populacji lub gdy istnieje podejrzenie, że klasyczne modele parametryczne mogą nie oddawać dobrze struktury danych.

Estymacja nieparametryczna funkcji gęstości pozwala na uzyskanie gładkiej krzywej opisującej rozkład zmiennej losowej, uwzględniając np. wielomodalność czy asymetrię rozkładu. Dzięki temu możliwe jest dokładniejsze zrozumienie i opisanie danych, co ma kluczowe znaczenie w analizie statystycznej, podejmowaniu decyzji oraz budowaniu modeli predykcyjnych.

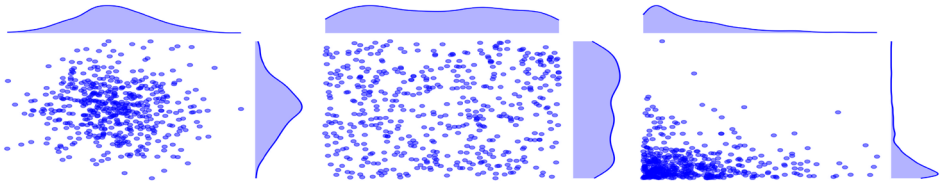
2.5. Histogram z dopasowaną funkcją gęstości



Histogram z dopasowaną gęstością jest graficznym narzędziem służącym do oceny nieznanej postaci rozkładu analizowanej zmiennej ciągłej na podstawie próby. Łączy on w sobie dwie formy przedstawiania informacji: z jednej strony pokazuje histogram, czyli sposób rozmieszczenia wartości empirycznych w zbiorze danych, a z drugiej nakłada na niego funkcję gęstości. Słupki histogramu obrazują, jak często dane wartości pojawiają się w określonych przedziałach. Każdy słupek odpowiada konkretnemu zakresowi wartości zmiennej, a jego wysokość pokazuje liczbę obserwacji przeliczoną na gęstość, co pozwala porównywać histogram z ciągłą oceną funkcji gęstości.

Na ten wykres nakłada się krzywą gęstości, która może pochodzić z różnych źródeł. Może to być funkcja estymowana nieparametrycznie, najczęściej przy użyciu estymatora jądrowego, który dopasowuje się elastycznie do struktury danych i pozwala uchwycić nieregularności, takie jak wielomodalność czy asymetria. Może to być również funkcja gęstości wynikająca z założonego rozkładu parametrycznego, np. normalnego, którego parametry (średnia i odchylenie standardowe) są zadane lub estymowane na podstawie próby. Połączenie histogramu i dopasowanej gęstości pozwala w prosty i intuicyjny sposób ocenić, jak dobrze rozkład teoretyczny jest dopasowany do danych rzeczywistych. Umożliwia też wstępną diagnozę nietypowości w danych, takich jak obecność skrajnych wartości, asymetrii czy wielu szczytów w rozkładzie. Dzięki temu stanowi jedno z podstawowych narzędzi w rękach analityka danych, szczególnie na etapie poznawania i opisu zbioru danych przed zastosowaniem bardziej formalnych metod statystycznych. Taka wizualna weryfikacja dopasowania stanowi zazwyczaj wstęp do przeprowadzenia ścisłych testów statystycznych, takich jak test Shapiro-Wilka czy Kołmogorowa-Smirnowa.

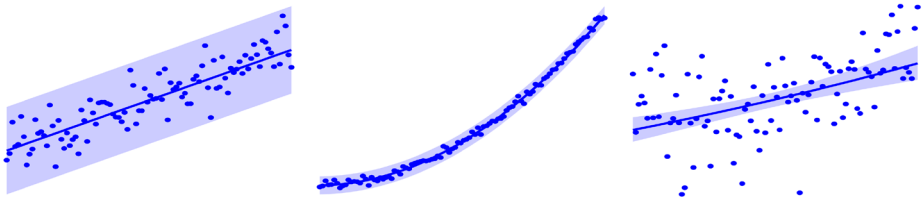
2.6. Wykresy rozrzutu z szacowanymi gęstościami brzegowymi



Wykresy rozrzutu z szacowanymi gęstościami brzegowymi stanowią rozbudowane narzędzie wizualizacji danych dwuwymiarowych. Ich główną funkcją jest umożliwienie jednoczesnego oglądu zależności pomiędzy dwiema zmiennymi losowymi oraz charakterystyki każdej z nich z osobna. Centralnym elementem takiej wizualizacji jest wykres rozrzutu, który obrazuje rozmieszczenie punktów odpowiadających obserwacjom dwóch zmiennych. Na jego podstawie można ocenić ogólną formę zależności między zmiennymi, np. liniową, nieliniową czy też brak jakiejkolwiek zależności. Uzupełnieniem wykresu rozrzutu są dwa wykresy gęstości umieszczane zazwyczaj na marginesach – nad i po prawej stronie wykresu głównego. Przedstawiają one tzw. gęstości brzegowe, czyli szacowane rozkłady prawdopodobieństwa dla każdej zmiennej z osobna. Gęstości te można oszacować za pomocą metod nieparametrycznych, takich jak estymator jądrowy, co pozwala uchwycić dokładniejszą strukturę rozkładu niż w przypadku histogramów. Dzięki temu, nawet jeśli dane nie pochodzą z rozkładu normalnego, możliwe jest realistyczne przedstawienie ich rozkładu.

Taka forma wizualizacji pozwala nie tylko na wykrywanie zależności pomiędzy zmiennymi, ale także na zrozumienie rozkładów każdej z nich. Jest to szczególnie użyteczne w analizie eksploracyjnej danych, gdy celem jest zbudowanie intuicji o strukturze danych, zanim zostaną zastosowane formalne metody statystyczne. Można dzięki temu łatwo dostrzec asymetrię, wielomodalność czy obecność ogonów w rozkładzie, a jednocześnie zaobserwować, czy i w jaki sposób zmienne są ze sobą powiązane. Takie zestawienie jest niezwykle pomocne przy identyfikacji dwuwymiarowych wartości odstających (tzw. outliers). Często zdarza się bowiem, że obserwacja wydaje się zupełnie typowa na wykresach brzegowych, ale na centralnym wykresie rozrzutu wyraźnie wyłamuje się z ogólnego trendu. Wykrycie takich anomalii na wczesnym etapie chroni badacza przed wyciąganiem błędnych wniosków.

2.7. Wykresy przedziału ufności dla funkcji regresji

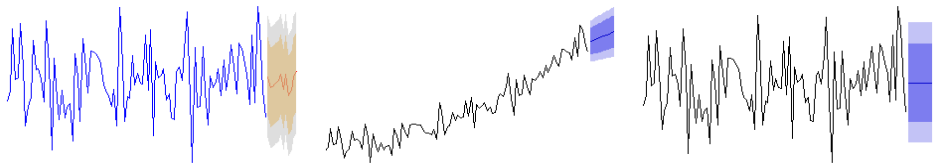


Wykresy ilustrujące estymację parametrów wraz z niepewnością odgrywają istotną rolę w statystycznej analizie danych, ponieważ nie tylko pokazują oszacowane wartości, takie jak średnia czy wartość oczekiwana, ale również informują o stopniu niepewności, jaki towarzyszył wyznaczeniu tych wartości. Najczęściej stosowaną formą przedstawienia niepewności są przedziały ufności, które otaczają wartość estymowaną, wskazując, jak bardzo są wiarygodne przy danym poziomie ufności.

Przedział ufności dla średniej odpowiedzi (funkcji regresji) dla ustalonego poziomu ufności $1 - \alpha$ (zwykle 95%) pokazuje, jaką precyzję ma oszacowanie. Przy wielokrotnym losowaniu prób z tej samej populacji około $(1 - \alpha)100\%$ tak skonstruowanych przedziałów zawierałoby prawdziwą wartość średniej dla danego poziomu zmiennej niezależnej. Oznacza to, że jeśli przy wielokrotnym pobieraniu próbki z tej samej populacji i dla każdej z nich zostałby wyznaczony przedział ufności, to około 95% z tych przedziałów obejmowałoby prawdziwą wartość parametru. Przedziały te mają zatem charakter probabilistyczny i są wyrazem niepewności wynikającej z losowości próby.

W kontekście analizy regresji przedziały ufności pokazują, z jakim prawdopodobieństwem przedział ufności dla zmiennej zależnej (średniej odpowiedzi) obejmuje przeciętną wartość funkcji regresji dla konkretnej wartości zmiennej niezależnej. W zależności od stopnia dopasowania modelu i liczby obserwacji te przedziały mogą być węższe lub szersze – węższe, gdy jest dużo danych i model dobrze opisuje zależność, oraz szersze, gdy dane są rozproszone lub próba niewielka. Należy jednak odróżnić przedziały ufności od przedziałów przewidywania. Te drugie informują o tym, gdzie może się znaleźć pojedyncza, nowa obserwacja przy danym poziomie zmiennej niezależnej. Przedziały przewidywania są z reguły szersze, ponieważ uwzględniają nie tylko niepewność estymacji średniej, ale także zmienność losową wokół niej.

2.8. Wykresy dla przedziałów prognozy (przewidywania)

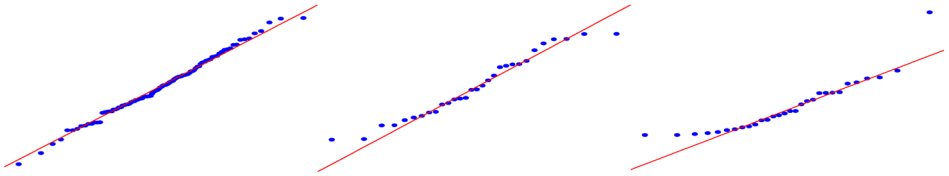


Przedziały ufności i przedziały prognozy (przewidywania) odgrywają fundamentalną rolę w analizie danych, zwłaszcza wtedy, gdy celem jest modelowanie zależności między zmiennymi lub prognozowanie przyszłych wartości. Są one nieodzownym elementem takich dziedzin, jak analiza regresji, modelowanie zjawisk ekonomicznych czy prognozowanie szeregów czasowych, ponieważ umożliwiają nie tylko szacowanie wartości, ale również wyrażenie niepewności związanej z tym szacunkiem.

Przedział prognozy różni się od przedziału ufności zakresem oraz interpretacją. O ile przedział ufności odnosi się do nieznanego parametru populacyjnego, takiego jak wartość oczekiwana zmiennej zależnej dla określonego poziomu zmiennej niezależnej, o tyle przedział prognozy odnosi się do nowej, pojedynczej obserwacji.

W kontekście praktycznym przedziały przewidywania są szczególnie istotne, gdy celem analizy jest prognozowanie. Na przykład w analizie sprzedaży, prognozując wynik dla kolejnego miesiąca, interesująca jest nie tylko średnia wartość prognozy, lecz również to, jak szeroki może być rozrzut tej wartości – i właśnie to obrazuje przedział przewidywania. Dzięki temu użytkownik ma świadomość, że nawet przy dobrze dopasowanym modelu rzeczywiste wyniki mogą odbiegać od prognozy i wie, jak duża może być ta niepewność. Tego rodzaju informacje są kluczowe dla podejmowania decyzji w warunkach niepewności i mają zastosowanie zarówno w kontekście indywidualnych przypadków, jak i przy formułowaniu strategii opartych na danych statystycznych. Szerokość tych przedziałów jest doskonałym miernikiem jakości samego procesu modelowania. Jeśli przedział przewidywania jest na tyle rozległy, że obejmuje niemal wszystkie skrajne scenariusze, może to stanowić sygnał o braku kluczowych zmiennych objaśniających w modelu. Wymusza to na analityku ponowną weryfikację założeń i ewentualne wzbogacenie algorytmu o dodatkowe informacje, które pomogą zredukować ten margines błędu.

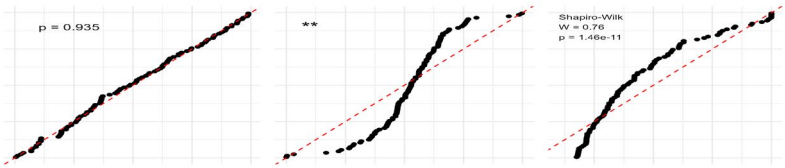
2.9. Q-Q plot



Wykres kwantyl-kwantyl (Q-Q plot) to narzędzie graficzne wykorzystywane w statystyce do oceny zgodności rozkładu danych empirycznych z określonym rozkładem teoretycznym. Jest to jedna z podstawowych metod diagnostycznych pozwalających ocenić, na ile dane odpowiadają założeniom statystycznym dotyczącym rozkładu, np. w kontekście normalności, co ma istotne znaczenie przy doborze odpowiednich metod analizy. Ideą wykresu Q-Q jest porównanie wartości kwantyli z dwóch rozkładów: empirycznego, pochodzącego z danych, oraz teoretycznego, z którym jest zestawiany. Kwantyl empiryczny to taka wartość z próby, poniżej której znajduje się zadany odsetek danych. Kwantyl teoretyczny to odpowiadająca temu samemu poziomowi prawdopodobieństwa wartość obliczona z przyjętego rozkładu teoretycznego. Dla każdego punktu w próbie wyznacza się odpowiadający mu kwantyl teoretyczny i empiryczny, po czym obie wartości przedstawia się jako współrzędne punktu na płaszczyźnie. Oś pozioma wykresu reprezentuje kwantyle rozkładu teoretycznego, natomiast oś pionowa kwantyle rozkładu empirycznego. Jeśli dane rzeczywiście pochodzą z rozkładu, który został przyjęty jako odniesienie, to punkty na wykresie powinny się układać wzdłuż linii prostej. Odchylenia od tej linii świadczą o różnicach między rozkładami. Jeśli punkty na końcach wykresu odchodzą od linii prostej, może to wskazywać na różnice w ogonach rozkładów. Dłuższe ogony w rozkładzie empirycznym niż w teoretycznym sugerują większą liczebność skrajnych wartości w danych.

Wykres Q-Q ma zastosowanie głównie jako metoda diagnostyczna, szczególnie w analizie normalności rozkładu reszt w modelach regresyjnych. Pozwala na szybkie, wizualne zidentyfikowanie przypadków, w których dane odbiegają od przyjętych założeń, co może prowadzić do decyzji o zastosowaniu innych metod statystycznych, bardziej odpornych na takie odstępstwa lub opartych na innych rozkładach. Jest to zatem narzędzie proste w interpretacji, a jednocześnie bardzo pomocne, zwłaszcza na etapie wstępnej analizy danych lub w procesie diagnostyki modeli statystycznych. Q-Q plot jest zwykle czuły na różnice w ogonach rozkładu.

2.10. P-P plot

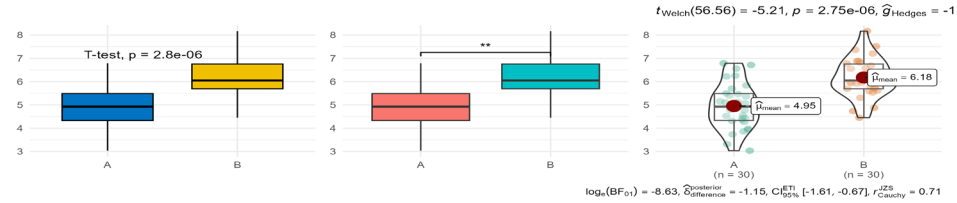


Wykres prawdopodobieństwo-prawdopodobieństwo, znany jako P-P plot, jest narzędziem graficznym wykorzystywanym do oceny zgodności rozkładu empirycznego z rozkładem teoretycznym. W przeciwieństwie do wykresu kwantyl-kwantyl (Q-Q plot), gdzie porównuje się bezpośrednio wartości kwantyli, wykres P-P opiera się na porównaniu wartości dystrybuant, czyli skumulowanych prawdopodobieństw obu rozkładów.

Na osi poziomej wykresu znajdują się wartości dystrybuanty rozkładu teoretycznego, natomiast na osi pionowej umieszcza się odpowiadające im wartości dystrybuanty empirycznej, wyliczone na podstawie danych. Każdy punkt na wykresie przedstawia parę wartości skumulowanego prawdopodobieństwa z rozkładu teoretycznego oraz z danych empirycznych. Jeśli rozkład empiryczny jest zgodny z rozkładem teoretycznym, to punkty na wykresie ułożą się blisko linii prostej, przecinającej wykres pod kątem 45 stopni. Odchylenia punktów od tej linii wskazują na różnice między rozkładami, co może pomóc w identyfikacji obszarów, gdzie rozkład empiryczny odbiega od teoretycznego. Wykres P-P jest szczególnie przydatny przy analizie rozkładów ciągłych i jest często stosowany jako narzędzie diagnostyczne do sprawdzania założeń dotyczących rozkładu danych przed zastosowaniem metod statystycznych opartych na tych założeniach (np. rozkład normalny).

Różnica pomiędzy wykresem P-P plot a wykresem Q-Q plot polega na tym, że pierwszy porównuje wartości dystrybuant, a drugi wartości kwantyli. W praktyce oznacza to, że wykres P-P plot bardziej skupia się na porównaniu skumulowanych prawdopodobieństw, podczas gdy wykres Q-Q plot pozwala ocenić dopasowanie rozkładów przez bezpośrednią analizę odpowiednich wartości danych. Dzięki temu wykres P-P plot stanowi uzupełnienie wykresu Q-Q plot i razem tworzą zestaw narzędzi umożliwiających kompleksową ocenę zgodności rozkładów w analizie statystycznej. Konstrukcja wykresu P-P plot prowadzi do jego większej czułości na różnice w środkowej części rozkładu.

2.11. Wykresy z wynikami wnioskowania statystycznego

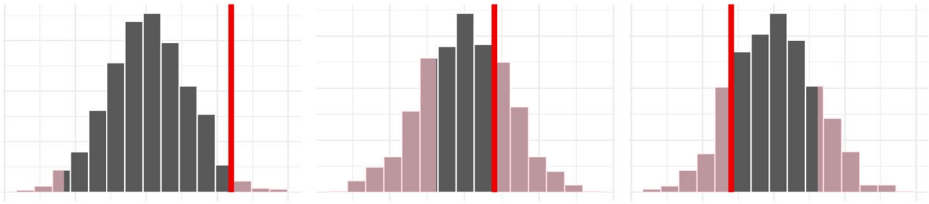


Wiele pakietów statystycznych umożliwia przedstawienie w oknie graficznym, poza samą prezentacją wykresu, także różnych charakterystyk liczbowych dotyczących przeprowadzonego wnioskowania statystycznego. Może to obejmować m.in. wartości statystyk testowych, przedziały ufności, współczynniki regresji, *p*-wartości, wskaźniki dopasowania modeli czy efekty średnie. Zamieszczenie takich wyników bezpośrednio na wykresie nie zastąpi oczywiście właściwie przedstawionych wyników w formie tabelarycznej, jednak pozwala na szybką ocenę różnic w rozkładach populacji, wielkości efektu lub występujących zależności pomiędzy zmiennymi.

Do najważniejszych pakietów programu **R** pozwalających zamieszczać na wykresach podstawowe rezultaty wnioskowania statystycznego należą:
`ggplot2` – pozwala na nakładanie elementów statystycznych za pomocą takich warstw, jak stat_summary, geom_smooth, geom_errorbar, geom_pointrange itp.;
`ggpubr` – upraszcza tworzenie wykresów publikacyjnej jakości i dodawanie wyników testów statystycznych (np. *t*-test, ANOVA) bezpośrednio na wykresach;
`ggsignif` – umożliwia dodawanie znaczników istotności statystycznej między grupami (np. gwiazdki *, **, ***) wraz z łukami i *p*-wartościami;
`ggstatsplot` – automatyzuje generowanie wykresów z wynikami testów statystycznych (np. *t*-testy, testy nieparametryczne, korelacje), integrując analizę i wizualizację.

Dzięki tym pakietom użytkownik może łączyć przejrzystość graficzną z informacyjną wyników statystycznych, co jest szczególnie przydatne w eksploracyjnej analizie danych, raportowaniu wyników oraz tworzeniu wizualizacji do publikacji naukowych. W przypadku jakiejkolwiek zmiany lub aktualizacji zbioru danych zarówno warstwa wizualna, jak i nałożone na nią miary statystyczne aktualizują się w pełni automatycznie.

2.12. Wykresy dla wnioskowania statystycznego w modelu permutacyjnym



W testach permutacyjnych wykresy graficzne pełnią istotną rolę w wizualizacji wyników analizy statystycznej. Stanowią one nie tylko pomocnicze narzędzie w ocenie istotności statystycznej, ale często są kluczowym elementem interpretacji rezultatów. Przede wszystkim wykresy te są wykorzystywane do przedstawienia rozkładu permutacyjnego statystyki testowej, uzyskanego na podstawie wielu permutacji danych, co umożliwia wizualne porównanie z wartością empiryczną tej statystyki obliczoną na podstawie oryginalnych danych. Typowym przykładem takiego wykresu jest histogram lub wykres gęstości rozkładu permutacyjnego z zaznaczoną na nim wartością obserwowaną. Pozwala to w sposób graficzny oszacować p-wartość jako proporcję permutacji, dla których wartość statystyki była co najmniej tak ekstremalna, jak obserwowana. W przypadku testów dwustronnych na wykresie zazwyczaj oznacza się dwa ogony rozkładu permutacyjnego oraz obszary istotności.

Wizualizacja wyników testów permutacyjnych może również przybierać bardziej złożoną formę, np. rozkładu statystyki testowej w zależności od różnych warunków eksperymentalnych, permutacyjnych map cieplnych (heatmap) czy wykresów permutacyjnych wartości statystyki w przestrzeni parametrów (np. w analizach wielowymiarowych). Wykresy te są nie tylko intuicyjnym narzędziem dla statystyków, ale także znacząco ułatwiają komunikację wyników analizy dla odbiorców nieposiadających zaawansowanego przygotowania matematycznego. Dzięki graficznemu przedstawieniu możliwe jest szybkie uchwycenie istotnych różnic, odchyleń od losowości oraz stopnia dopasowania modelu do danych. W przeciwieństwie do klasycznych metod parametrycznych opartych na teoretycznych krzywych w testach permutacyjnych rozkład odniesienia jest generowany bezpośrednio z pobranej próby. Taka empiryczna forma weryfikacji hipotez sprawia, że wyciągnięte wnioski budzą większe zaufanie, zwłaszcza w przypadku danych o nietypowej strukturze.

3. Rozkłady teoretyczne i ich nieparametryczna estymacja – wybrane biblioteki programu R

Wszystkie modele są błędne,
ale niektóre są użyteczne.

George E. P. Box (1979, s. 202)

Współczesna analiza danych statystycznych wykracza poza same obliczenia numeryczne, kładąc coraz większy nacisk na zaawansowaną wizualizację, która jest kluczowa dla zrozumienia natury badanych zjawisk i weryfikacji założeń modelowych. Graficzna prezentacja wyników wnioskowania pozwala nie tylko na skuteczniejszą interpretację skomplikowanych zależności, ale także ułatwia komunikację wniosków odbiorcom o różnym poziomie zaawansowania. Choć w środowisku **R** standardem stał się pakiet **`ggplot2`**, specyficzne zagadnienia statystyczne – takie jak wizualizacja obszarów krytycznych, operacje na zmiennych losowych czy estymacja jądrowa – często wymagają zastosowania wyspecjalizowanych rozszerzeń i dodatkowych bibliotek.

Niniejszy rozdział koncentruje się na przeglądzie i praktycznym zastosowaniu wybranych pakietów środowiska **R**, które wspomagają prezentację teoretycznych rozkładów prawdopodobieństwa oraz oszacowań ich parametrów. Omówione zostaną narzędzia do automatyzacji wykreślania funkcji gęstości i dystrybuanty dla zmiennych losowych o zadanym rozkładzie prawdopodobieństwa (m.in. **`ggfortify`**, **`ggnormalviolin`**), pakiety umożliwiające pracę na mieszankach i funkcjach zmiennych losowych (**`convdistr`**), a także metody nieparametrycznej estymacji gęstości jedno- i wielowymiarowej z wykorzystaniem bibliotek takich jak **`KernSmooth`**, **`MASS`** czy **`lattice`**. Celem rozdziału jest zaprezentowanie zestawu narzędzi pozwalających na kompleksową wizualizację probabilistycznych aspektów analizy danych.

3.1. Rozkłady teoretyczne i ich estymacja – wybrane biblioteki programu R

Kończak (2024a) wskazuje na różnorodne biblioteki pozwalające rozszerzyć możliwości pakietu graficznego **ggplot2**. W tej części skupiono się na pakietach, które mogą być zastosowane w graficznych prezentacjach rezultatów wnioskowania statystycznego. Wykaz pakietów, które zostały wykorzystane w tym rozdziale, przedstawiono w tabeli 3.1.

Tabela 3.1. Wybrane biblioteki programu R wspomagające prezentację rozkładów zmiennych losowych

Biblioteka	Opis
ggfortify	Automatyzacja wykreślania wybranych wykresów rozkładów teoretycznych
ggnormalviolin	Konstrukcja normalnych wykresów skrzypcowych (określanych także jako wiolinowe) z określonymi średnimi i odchyleniami standardowymi
convdistr	Łączenie rozkładów prawdopodobieństwa przy użyciu funkcji generatora liczb losowych każdego rozkładu
ggplot2	Ogólne środowisko graficzne oparte na *Grammar of Graphics*; tworzenie elastycznych wykresów rozkładów (histogramy, gęstości, wykresy Q-Q, wykresy rozrzutu z gęstościami brzegowymi itp.)
lattice	System grafiki trellis do wielowymiarowych wizualizacji, w tym prezentacji rozkładów warunkowych i porównywania gęstości w wielu grupach
MASS	Zbiór funkcji i danych do klasycznej książki Venables i Ripley (2002); generowanie zmiennych z różnych rozkładów, dopasowanie modeli i rozkładów, m.in. funkcje pomocne przy analizie rozkładów empirycznych
KernSmooth	Implementacja jądrowej estymacji gęstości i regresji; umożliwia wygładzoną, nieparametryczną estymację gęstości oraz jej wizualizację
see	Pakiet do atrakcyjnej wizualizacji wyników modeli statystycznych, rozkładów a posteriori oraz przedziałów niepewności
data.table	Wysokowydajne narzędzie do przetwarzania i agregacji danych ułatwiające przygotowanie dużych zbiorów danych do wizualizacji rozkładów i estymacji gęstości
ggmosaic	Rozszerzenie **ggplot2** do tworzenia wykresów mozaikowych (mosaic plots) prezentujących rozkłady wielowymiarowych zmiennych jakościowych
tidyverse	Zestaw spójnych pakietów (m.in. **dplyr**, **tidyr**, **readr**) do pracy z „uporządkowanymi" danymi ułatwiający przygotowanie danych do wizualizacji w **ggplot2** oraz do estymacji gęstości

Źródło: Opracowanie własne na podstawie: CRAN (2026).

Na wstępie konieczne jest załadowanie bibliotek, które będą wykorzystywane w tym rozdziale. Z zakresu wnioskowania statystycznego są to biblioteki:

```
library(ggfortify)
library(ggnormalviolin)
library(convdistr)
library(ggplot2)
library(lattice)
library(ggpubr)
library(MASS)
```

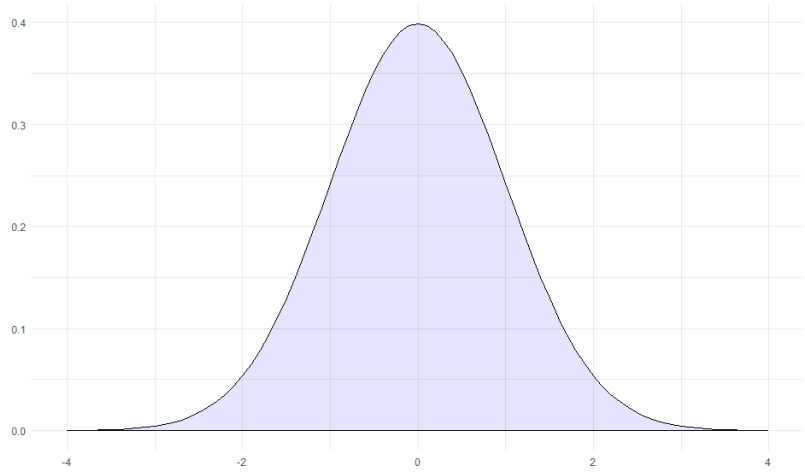
```
library(KernSmooth)
library(see)
library(data.table)
library(ggmosaic)
library(tidyverse)
library(gridExtra)
```

3.2. Rozkłady teoretyczne zmiennych losowych – gęstość i dystrybuanta

Pakiet **ggfortify** to narzędzie pozwalające na konstrukcję wykresów z wynikami dla różnych analiz statystycznych. Pakiet umożliwia m.in. przedstawienie wykresów dla modeli liniowych, metod klasyfikacji oraz szeregów czasowych z uwzględnieniem prognoz. W tej części uwaga zostanie jednak skoncentrowana wyłącznie na wykreślaniu funkcji gęstości i dystrybuanty rozkładów teoretycznych wybranych zmiennych losowych. Rozkłady teoretyczne stanowią podstawę konstrukcji przedziałów ufności, jak również parametrycznych testów statystycznych. Do najczęściej wykorzystywanych rozkładów teoretycznych we wnioskowaniu statystycznym należy zaliczyć m.in. rozkład normalny, rozkład t-Studenta oraz rozkład chi-kwadrat.

Wykres gęstości rozkładu normalnego standardowego uzyskuje się za pomocą następującego kodu, którego wynik realizacji przedstawiono na rys. 3.1.

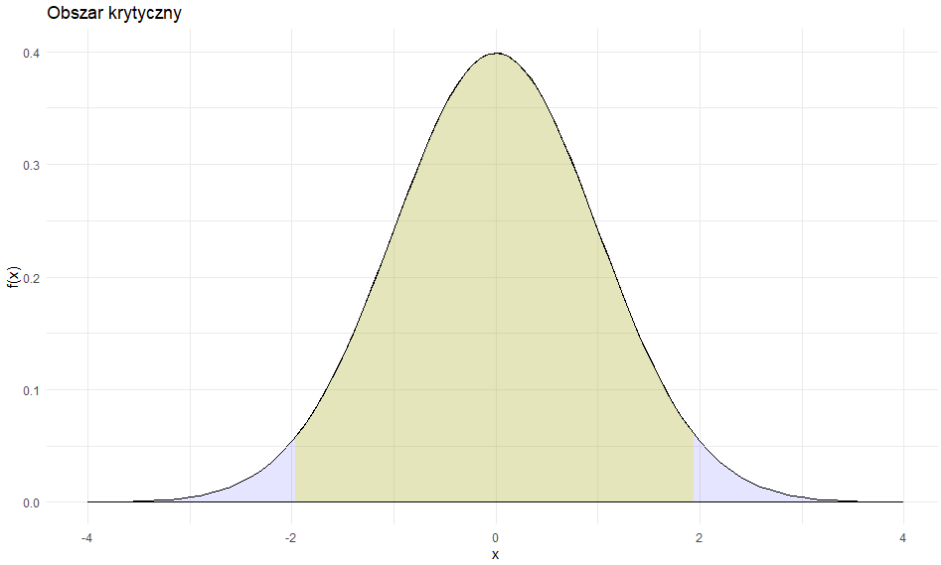
```
ggdistribution(dnorm, seq(-4, 4, 0.1), mean = 0, sd = 1,linetype=1,fill = 'blue',alpha=0.1)
+ theme_minimal()
```



Rys. 3.1. Gęstość rozkładu normalnego standardowego

Podczas prezentowania rezultatu testowania hipotez statystycznych często przydatna jest prezentacja wartości statystyki testowej na tle odpowiedniego obszaru krytycznego. Pakiet **ggfortify** umożliwia dla określonego rozkładu i przyjętego poziomu istotności α zaznaczenie na wykresie obszaru krytycznego. Taką możliwość w przypadku rozkładu normalnego standardowego przedstawia poniższy kod.

```
alpha=0.05
q=qnorm(1-alpha/2)
r1=ggdistribution(dnorm, seq(-q, q, 0.1), mean = 0, sd = 1,fill = 'yellow',alpha=0.3)
r2=ggdistribution(dnorm, seq(-4, 4, 0.1), mean = 0, sd = 1,linetype=1,fill = 'blue',alpha=
0.1,p=r1)
r2+xlim(-4,4)+ylim(0,0.4)+labs(title="Obszar krytyczny",  x="x", y="f(x)")+theme_min
imal()
```

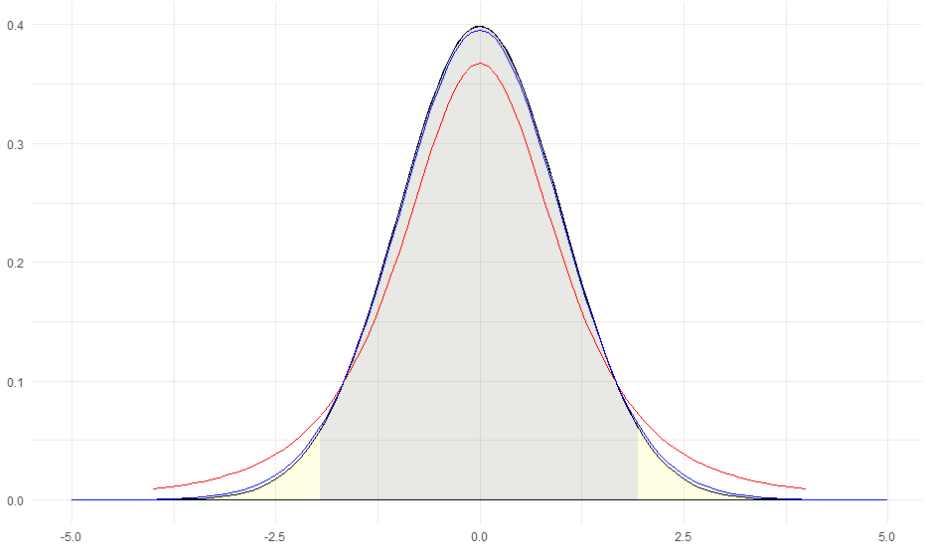


Rys. 3.2. Gęstość rozkładu normalnego standardowego z wyróżnionym obszarem krytycznym

W wyniku realizacji powyższego kodu otrzymuje się rys. 3.2. Wykres przedstawia dwustronny obszar krytyczny dla poziomu istotności $\alpha = 0{,}05$ dla zmiennej losowej o rozkładzie normalnym standardowym.

Pakiet **ggfortify** umożliwia przedstawienie na jednym wykresie kilku funkcji gęstości. Poniższy kod prezentuje umieszczenie na wykresie gęstości rozkładu normalnego standardowego oraz dwóch funkcji gęstości rozkładu *t*-Studenta dla stopni swobody 3 (linia czerwona) i 30 (linia niebieska).

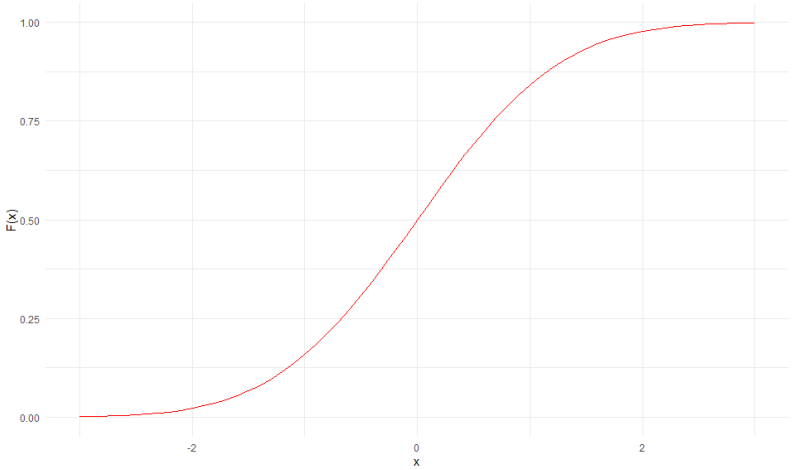
```
r1=ggdistribution(dnorm, seq(-q, q, 0.1), mean = 0, sd = 1,fill = 'blue',alpha=0.1)
r2=ggdistribution(dnorm, seq(-4, 4, 0.1), mean = 0, sd = 1,linetype=1,fill = 'yellow',alph
a=0.1,p=r1)
r3=ggdistribution(dt, seq(-4, 4, 0.1), df=3,colour='red',p=r2)
ggdistribution(dt, seq(-5, 5, 0.1), df=30,colour='blue',p=r3)+theme_minimal()
```



Rys. 3.3. Gęstości rozkładu normalnego standardowego i rozkładów *t*-Studenta o 3 i 30 stopniach swobody

Na rysunku 3.3 przedstawiono gęstość rozkładu normalnego oraz gęstości 2 zmiennych losowych o rozkładzie t-Studenta z liczbą stopni swobody 3 oraz 30. Zauważalne jest, że gęstość rozkładu t-Studenta dla $df = 30$ stopni swobody prawie pokrywa się z gęstością rozkładu normalnego standardowego. Dla wykreślenia dystrybuanty zmiennej losowej o rozkładzie normalnym standardowym należy wykonać następujący kod:

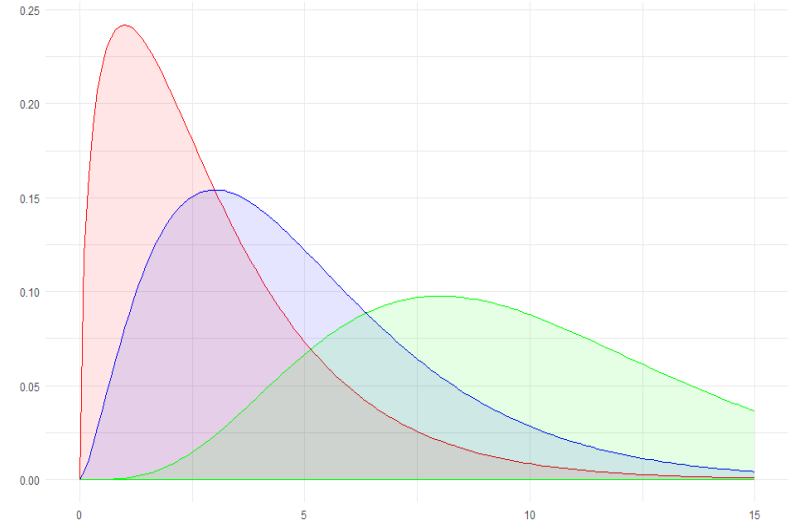
```
ggdistribution(pnorm, seq(-3, 3, 0.1), mean = 0, sd = 1, colour = 'red',xlab='x',ylab='F(x)')+
theme_minimal()
```



Rys. 3.4. Dystrybuanta rozkładu normalnego standardowego

Na rysunku 3.4 przedstawiono dystrybuantę rozkładu normalnego standardowego. Poniższy kod pozwala na jednym wykresie zamieścić funkcje gęstości trzech zmiennych losowych o rozkładzie chi-kwadrat odpowiednio z liczbą stopni swobody $df = 3$, 5 oraz 10. Wynik realizacji kodu został przedstawiony na rys. 3.5.

```
r1=ggdistribution(dchisq, seq(0, 15, 0.1), df=3,colour='red',fill='red',alpha=0.1)
r2=ggdistribution(dchisq, seq(0, 15, 0.1), df=5,colour='blue',p=r1,fill='blue',alpha=0.1)
ggdistribution(dchisq, seq(0, 15, 0.1), df=10,colour='green',p=r2,fill='green',alpha=0.1)+
theme_minimal()
```

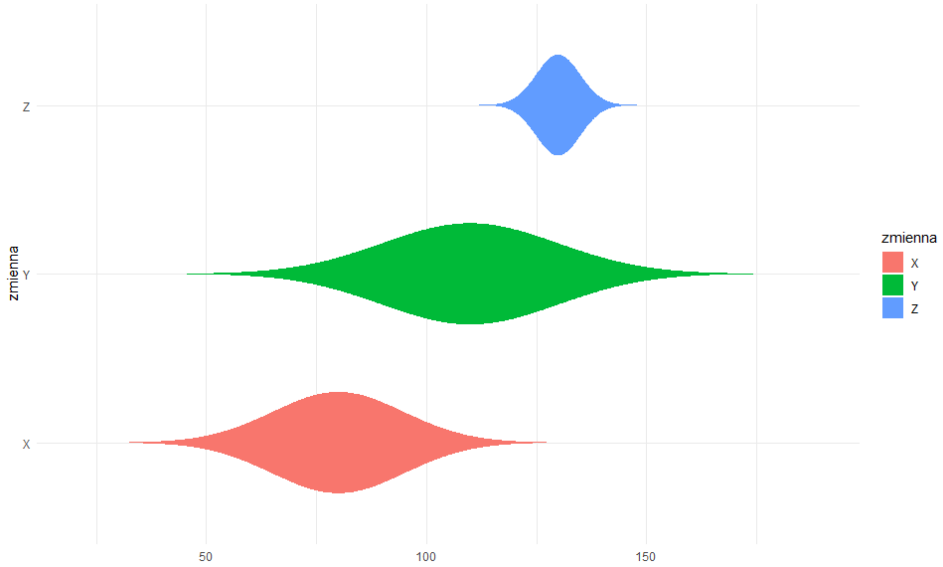


Rys. 3.5. Gęstości rozkładów chi-kwadrat o 3, 5 i 10 stopniach swobody

3.3. Rozkłady teoretyczne zmiennych losowych – wykresy skrzypcowe

Nieco inne podejście do prezentacji graficznej teoretycznych rozkładów prawdopodobieństwa umożliwia pakiet **`ggnormalviolin`**. Do prezentacji graficznej są wykorzystywane wykresy skrzypcowe pozwalające w sposób czytelny porównać charakterystyki kilku rozkładów. Taki przykład z wykresami gęstości trzech zmiennych losowych o rozkładzie normalnym z różnymi parametrami prezentuje poniższy kod.

```
d <- data.frame(
  zmienna = c("X", "Y", "Z"),
  zmienna_mean = c(80, 110, 130),
  zmienna_sd = c(15,  20, 5))
p <- ggplot(data = d,
            aes(x = zmienna,
                mu = zmienna_mean,
                sigma = zmienna_sd,
                fill = zmienna)) +
  theme(legend.position = "none")+
  coord_flip()
p + geom_normalviolin()+theme_minimal()
```

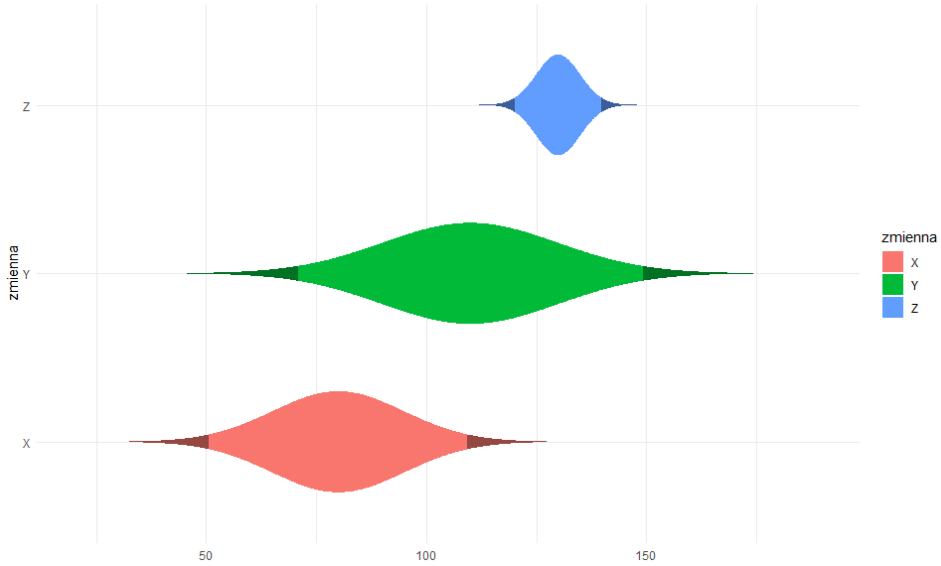


Rys. 3.6. Gęstość trzech zmiennych o rozkładach normalnych z wartościami oczekiwanymi: 80, 110 i 130 oraz odchyleniami standardowymi: 15, 20 i 5

W efekcie realizacji tego kodu uzyskuje się wykres jak na rys. 3.6. Na tym wykresie za pomocą graficznej reprezentacji w postaci wiolin przedstawiono gęstości trzech zmiennych losowych o rozkładach normalnych z wartościami oczekiwanymi kolejno μ = 80, 110 i 130 oraz odpowiednio odchyleniami standardowymi σ = 15, 20 i 5.

Podobnie jak w przypadku wcześniej omawianego pakietu **ggfortify**, tak i w tym przypadku jest możliwość zaznaczenia obszarów krytycznych. Przykład taki dla wcześniej ujętych zmiennych losowych przy przyjęciu poziomu istotności α = 0,05 realizuje poniższy kod, a wynik jego realizacji przedstawiono na rys. 3.7.

```
p + geom_normalviolin(p_tail = 0.05)+theme_minimal()
```



Rys. 3.7. Gęstość trzech zmiennych o rozkładach normalnych z wartościami oczekiwanymi: 80, 110 i 130 oraz odchyleniami standardowymi: 15, 20 i 5 z zaznaczonymi dwustronnymi obszarami krytycznymi dla poziomu istotności α = 0,05

3.4. Wizualizacja weryfikacji normalności rozkładu zmiennej

Poniższe rozważania zostaną przeprowadzone z wykorzystaniem danych ze zbioru **iris** (Fisher, 1936; Kończak, 2024a). Zbiór danych **iris** to jeden z najbardziej znanych i najczęściej używanych zbiorów danych w dziedzinie statystyki. Został

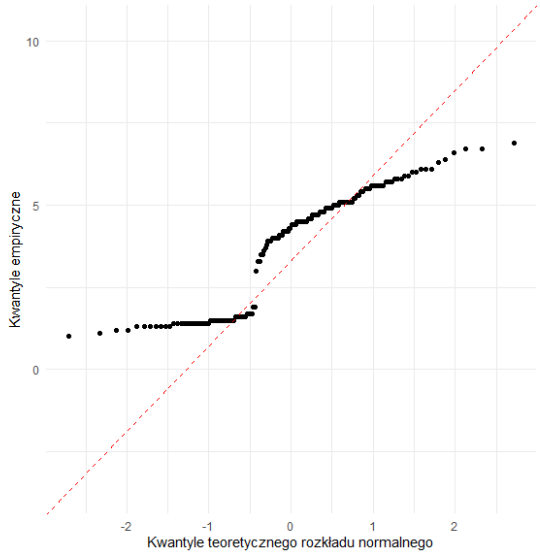
on wprowadzony przez brytyjskiego statystyka i biologa Ronalda Aylmera Fishera w pracy z 1936 roku (Fisher, 1936) jako przykład w zastosowaniu analizy dyskryminacyjnej.

Zbiór składa się ze 150 obserwacji i 5 następujących zmiennych:

- *Sepal.Length*: Długość kielicha (działki kielicha) w centymetrach.
- *Sepal.Width*: Szerokość kielicha (działki kielicha) w centymetrach.
- *Petal.Length*: Długość płatka w centymetrach.
- *Petal.Width*: Szerokość płatka w centymetrach.
- *Species*: Gatunek irysa (zmienna jakościowa o trzech wariantach: setosa, versicolor i virginica).

W pierwszym kroku zostanie sprawdzona normalność rozkładu długości płatków kielicha wraz z graficzną wizualizacją w formie wykresu Q-Q plot.

```
ggplot(iris, aes(sample = Petal.Length)) +
  stat_qq() +
  stat_qq_line(color = "red", linetype = "dashed") +
  labs(x = "Kwantyle teoretycznego rozkładu normalnego",
    y = "Kwantyle empiryczne") +
  theme_minimal() +
  coord_fixed(ratio = .4)
```

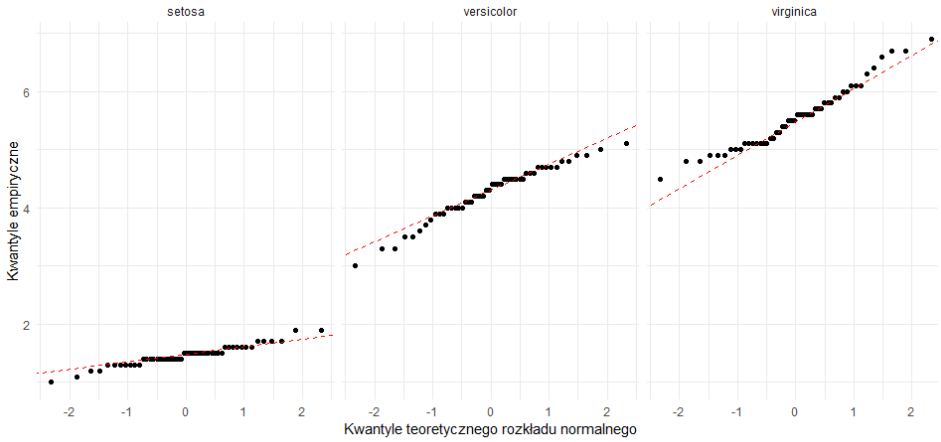


Rys. 3.8. Rozkład długości płatka (Petal.Length) a rozkład normalny

Na rysunku 3.8 badana zmienna ma inny rozkład niż normalny. W analizowanym przypadku wyróżnia się trzy gatunki kwiatu iris. W takiej sytuacji niezasadne jest sprawdzanie normalności rozkładu charakterystyk dla gatunków łącz-

nie, ale dla każdego z gatunków osobno. Zrealizowano to w formie graficznej następująco:

```
ggplot(iris, aes(sample = Petal.Length)) +
  stat_qq() +
  stat_qq_line(color = "red", linetype = "dashed") +
  facet_wrap(~ Species, ncol = 3) +
  labs( x = "Kwantyle teoretycznego rozkładu normalnego", y = "Kwantyle empiryczne") +
  theme_minimal() +
  coord_fixed(ratio = 1)
```



Rys. 3.9. Rozkład długości płatka (Petal.Length) a rozkład normalny według gatunku

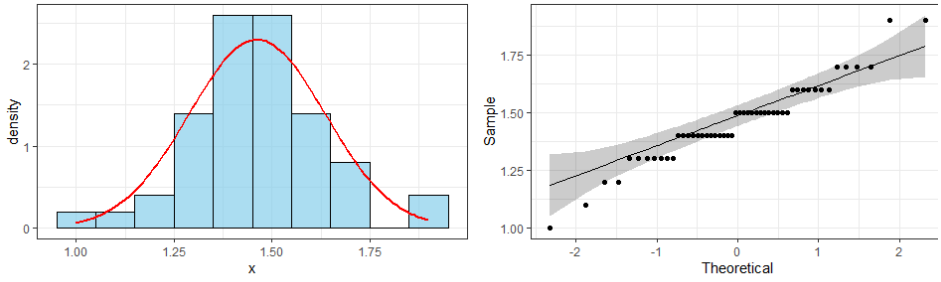
Na rysunku 3.9 przedstawiono trzy wykresy Q-Q plot dla zmiennej *Petal.Length* dla każdego z gatunków z osobna. Widoczne jest, że we wszystkich tych przypadkach punkty na wykresie są położone w pobliżu linii teoretycznej. Dla uzyskania formalnego potwierdzenia wskazane jest przeprowadzenie testu normalności.

```
shapiro.test(iris$Petal.Length[iris$Species=='setosa'])$p.value
# [1] 0.05481147
shapiro.test(iris$Petal.Length[iris$Species=='versicolor'])$p.value
# [1] 0.1584778
shapiro.test(iris$Petal.Length[iris$Species=='virginica'])$p.value
# [1] 0.1097754
```

Przeprowadzenie testu normalności we wszystkich przypadkach, przy przyjęciu poziomu istotności $\alpha = 0{,}05$, prowadzi do stwierdzenia braku podstaw do odrzucenia hipotezy głoszącej normalność rozkładu zmiennej *Petal.Length*.

Dla testowania normalności i odpowiedniej wizualizacji niekiedy wskazane jest przedstawienie wyników na kilku wykresach. Poniższy kod określa funkcję *diag_nor()* i prezentuje wyniki na wykresie.

```
diag_nor <- function(dane) {
  shapiro_test <- shapiro.test(dane)
  p1 <- ggplot(data.frame(x = dane), aes(x = x)) +
    geom_histogram(aes(y = after_stat(density)),
                   bins = 10,
                   color = "black",
                   fill = "skyblue",
                   alpha = 0.7) +
    stat_function(fun = dnorm,
                  args = list(mean = mean(dane), sd = sd(dane)),
                  color = "red", size = 1) +
    theme_bw()
  p2 <- ggqqplot(dane) +
    theme_bw()
  wyniki <- data.frame(
    Test = "Shapiro-Wilk",
    Statystyka = round(shapiro_test$statistic, 4),
    p_wartość = round(shapiro_test$p.value, 4),
    Wniosek = ifelse(shapiro_test$p.value > 0.05,
                     "Rozkład normalny",
                     "Brak normalności")
  )
  p3 <- gridExtra::tableGrob(wyniki)
  gridExtra::grid.arrange(p1, p2, p3, ncol = 2,
               layout_matrix = rbind(c(1, 2), c(3, 3)))
}
dane <- iris %>%
  filter(Species == "setosa") %>%
  select(Petal.Length) %>%
  unlist(use.names = FALSE)
diag_nor(dane)
```



	Test	Statystyka	p_wartość	Wniosek
W	Shapiro-Wilk	0.955	0.0548	Rozkład normalny

Rys. 3.10. Diagnoza normalności rozkładu długości płatka (Petal.Length)

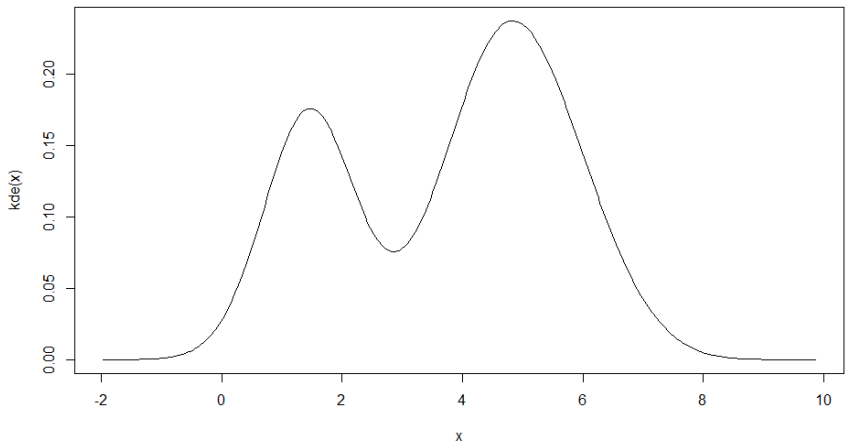
Funkcja *diag_nor()* przeprowadza kompleksową diagnostykę normalności rozkładu zmiennej numerycznej, tworząc panel trzech elementów wizualizacyjnych. Generuje histogram z nałożoną teoretyczną krzywą rozkładu normalnego (górny lewy), wykres kwantyl-kwantyl (Q-Q plot) porównujący rozkład empiryczny z teoretycznym (górny prawy) oraz tabelę z wynikami testu Shapiro-Wilka zawierającą statystykę testową, *p*-wartość i wniosek o normalności rozkładu (dolny pas). W przykładzie użycia funkcja analizuje długość płatków (*Petal.Length*) dla gatunku *setosa* z zestawu danych **iris** (por. rys. 3.10).

3.5. Nieparametryczna estymacja gęstości

Jednym z ważniejszych zagadnień nieparametrycznego wnioskowania statystycznego jest estymacja funkcji gęstości. Zagadnienie jest szeroko opisywane w literaturze (por. np. Domański i Pruska, 2000; Baszczyńska, 2016).

W pierwszej kolejności estymacja nieparametryczna długości płatka kwiatu iris zostanie wykonana z pakietem **KernSmooth**.

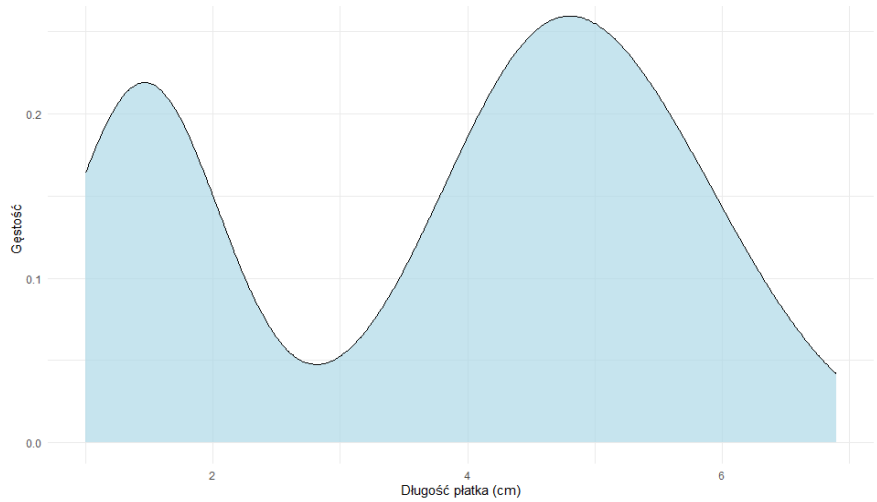
```
x <- iris$Petal.Length
plot(bkde(x), type = "l", xlab='x',ylab='kde(x)')
```



Rys. 3.11. Estymacja gęstości z pakietem KernSmooth. Gęstość długości płatka (Petal.Length)

Na rysunku 3.11 przedstawiono rezultat powyższego kodu. Wykres przedstawia estymację gęstości zmiennej *Petal.Length*. Podobne możliwości jak pakiet **`KernSmooth`** w zakresie graficznej prezentacji wyników estymacji nieparametrycznej zapewnia pakiet **`ggplot2`**. Dla tych samych danych graficzna prezentacja zrealizowana w oparciu o ten pakiet jest następująca:

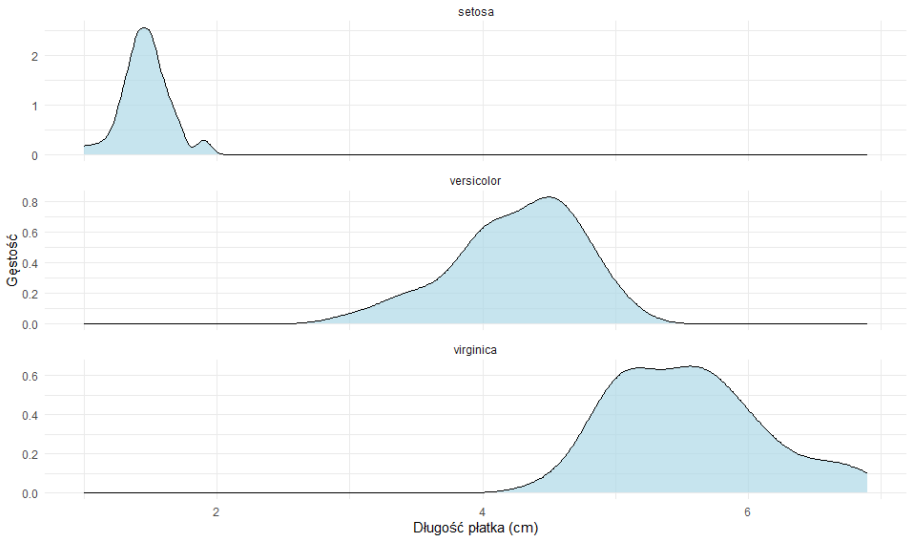
```
ggplot(iris, aes(x = Petal.Length)) +
  geom_density(fill = "lightblue", alpha = 0.7) + # Dodajmy trochę estetyki
  labs( x = "Długość płatka (cm)", y = "Gęstość"  ) +
  theme_minimal()
```



Rys. 3.12. Estymacja długości płatka (Petal.Length)

Na rysunku 3.12 przedstawiono rezultat powyższego kodu. Wykres przedstawia estymację gęstości zmiennej *Petal.Length*. W analizowanym przypadku wskazane jest przeprowadzenie estymacji gęstości zmiennej dla poszczególnych gatunków z osobna. Takie zadanie realizuje poniższy kod.

```
ggplot(iris, aes(x = Petal.Length)) +
  geom_density(fill = "lightblue", alpha = 0.7) +
  facet_wrap(~ Species, ncol=1,scales = "free_y") +
  labs(x = "Długość płatka (cm)", y = "Gęstość" ) +
  theme_minimal()
```

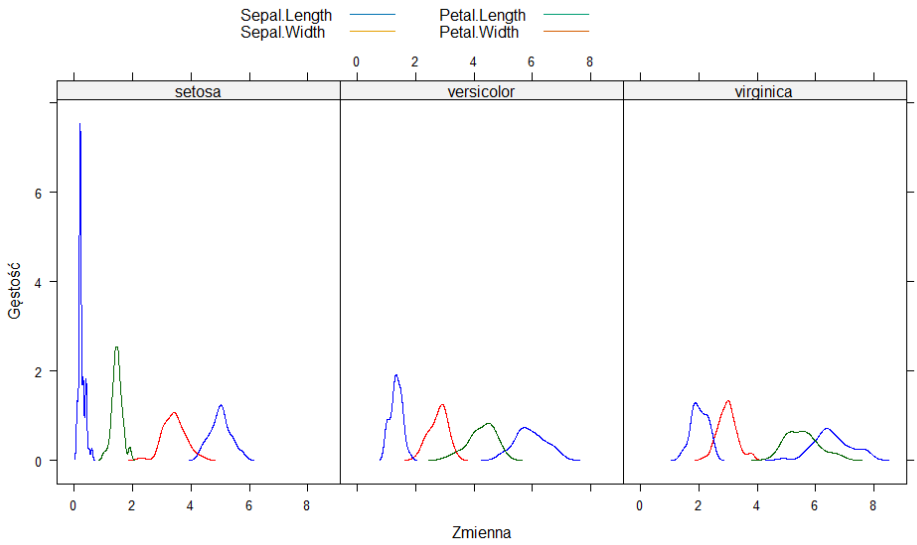


Rys. 3.13. Estymacja gęstości długości płatka (Petal.Length) według gatunku

Na rysunku 3.13 przedstawiono oszacowania gęstości zmiennej *Petal.Length* według gatunku.

Uzyskanie w bardzo wygodny sposób jedno- i dwuwymiarowych ocen gęstości i ich wizualizacji umożliwia także pakiet **`lattice`**. W przeciwieństwie do **`ggplot2`** nie opiera się on na idei gramatyki grafiki, lecz na zestawie predefiniowanych szablonów wykresów i mechanizmach ich modyfikacji. Poniższy kod ilustruje wykorzystanie funkcji z pakietu **`lattice`** do uzyskania jednowymiarowych ocen gęstości dla wszystkich czterech zmiennych liczbowych ze zbioru **iris**, przedstawionych osobno dla każdego gatunku.

```
densityplot(~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width | Species,
            data = iris,
            auto.key = list(columns = 3),
            plot.points = FALSE,
            col = c("blue", "red", "darkgreen"),
            xlab = "Zmienna")
```

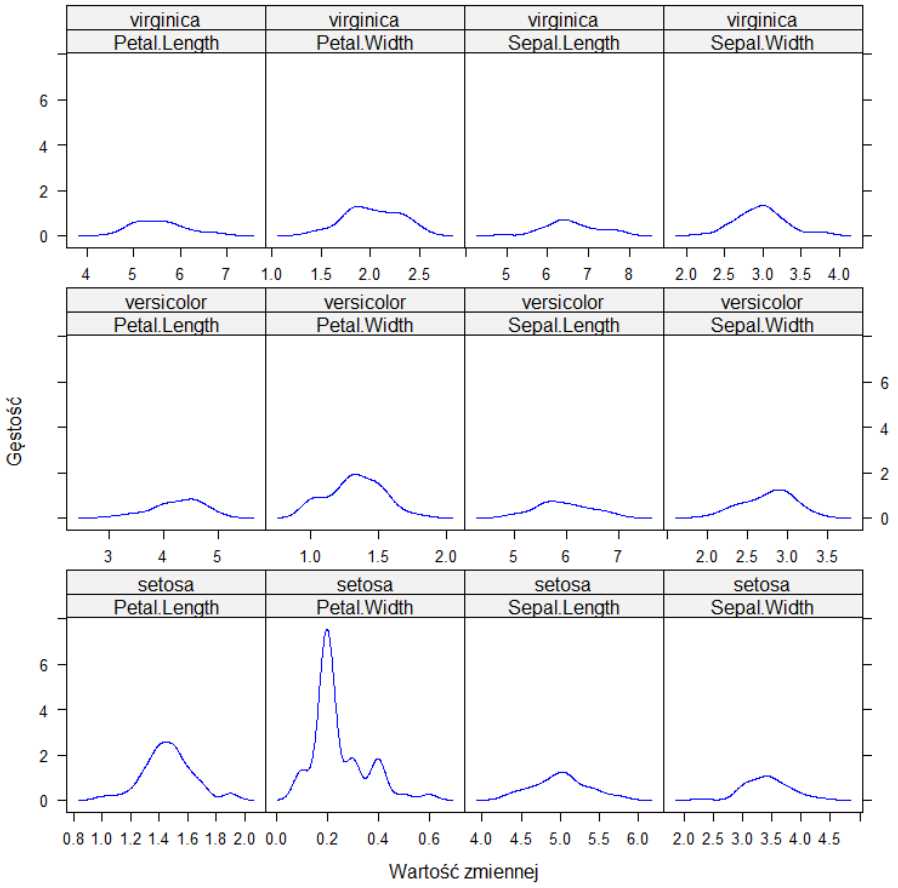


Rys. 3.14. Estymacja gęstości zmiennych numerycznych zbioru iris według gatunku

Na rysunku 3.14 przedstawiono estymację gęstości zmiennych *Petal.Length*, *Petal.Width*, *Sepal.Length* i *Sepal.Width* ze zbioru **iris** według gatunku.

Nieco inną formę prezentacji tych samych oszacowań przedstawia poniższy kod, którego rezultat prezentuje rys. 3.15.

```
iris_long <- data.frame(
  value = unlist(iris[, 1:4]),
  variable = rep(names(iris[, 1:4]), each = nrow(iris)),
  Species = rep(iris$Species, times = ncol(iris[, 1:4])))
densityplot(~ value | variable * Species,
            data = iris_long,
            auto.key = list(columns = 3),
            plot.points = FALSE,
            col = c("blue", "red", "darkgreen"),
            xlab = "Wartość zmiennej",
            scales = list(x = list(relation = "free")))
```

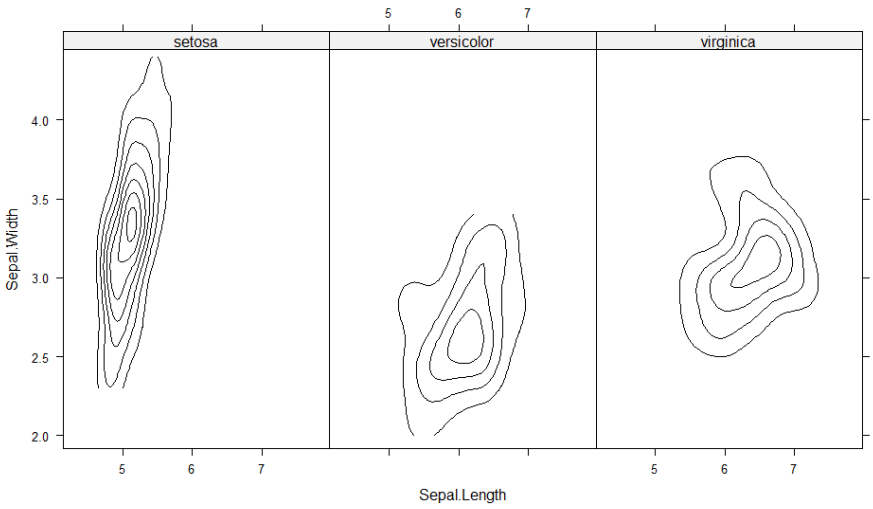


Rys. 3.15. Estymacja gęstości zmiennych numerycznych zbioru iris według zmiennej i gatunku

Pakiet **`lattice`** umożliwia także uzyskanie ocen dla dwuwymiarowych zmiennych. Przykład takiego rozwiązania przedstawia poniższy kod.

```
get_density_data <- function(data, x_var, y_var, species_name) {
  x <- data[[x_var]]
  y <- data[[y_var]]
  if (length(x) < 2 || length(y) < 2) {
    return(NULL)
  }
  dens <- MASS::kde2d(x, y, n = 50)
  data.frame(
    x = rep(dens$x, each = length(dens$y)),
    y = rep(dens$y, times = length(dens$x)),
    z = as.vector(dens$z),
    Species = species_name)
}
```

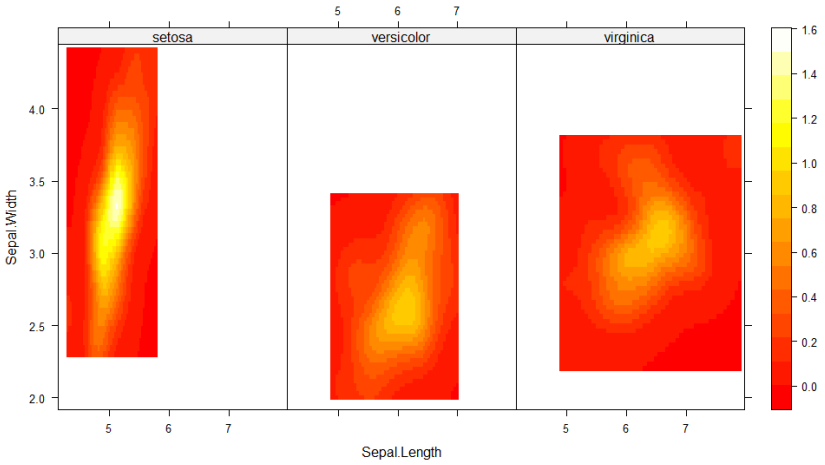
```
density_list <- lapply(levels(iris$Species), function(s) {
  subset_iris <- subset(iris, Species == s)
  get_density_data(subset_iris, "Sepal.Length", "Sepal.Width", s)
})
density_df <- do.call(rbind, density_list)
contourplot(z ~ x * y | Species,
            data = density_df,
            cuts = 10,
            labels = FALSE,
            col.regions = topo.colors,
            xlab = "Sepal.Length",
            ylab = "Sepal.Width",
            panel = function(x, y, z, subscripts, ..., B = NULL, A = NULL) {
              current_species_name <- levels(density_df$Species)[panel.number()]
              panel.contourplot(x, y, z, subscripts, ...)
              points_data_filtered <- subset(iris, Species == current_species_name)
              panel.points(points_data_filtered$Sepal.Length, points_data_filtered$Sepal.Wid
th,
                           col = "grey", pch = 16, cex = 0.6)
            })
```



Rys. 3.16. Estymacja dwuwymiarowej gęstości (Sepal.Length i Sepal.Width)

Na rysunku 3.16 przedstawiono dwuwymiarowe oceny gęstości zmiennych *Sepal.Length* i *Sepal.Width* według gatunków. Nieco inną formę graficzną tych samych ocen prezentuje kolejny kod, którego wynik przedstawia rys. 3.17.

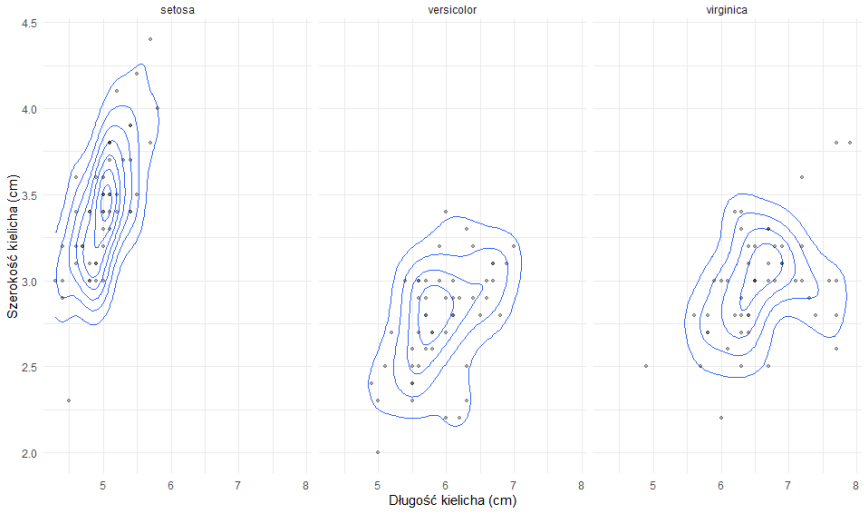
```
levelplot(z ~ x * y | Species,
          data = density_df,
          col.regions = heat.colors(100),
          xlab = "Sepal.Length",
          ylab = "Sepal.Width")
```



Rys. 3.17. Mapa ciepła – estymacja dwuwymiarowej gęstości (Sepal.Length i Sepal.Width)

Wykorzystanie pakietu **`KernSmooth`** dla estymacji gęstości dwuwymiarowych przedstawia poniższy kod.

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_density_2d() +
  geom_point(alpha = 0.3, size = 1) +
  facet_wrap(~ Species) +
  labs(x = "Długość kielicha (cm)", y = "Szerokość kielicha (cm)" ) +
  theme_minimal()
```



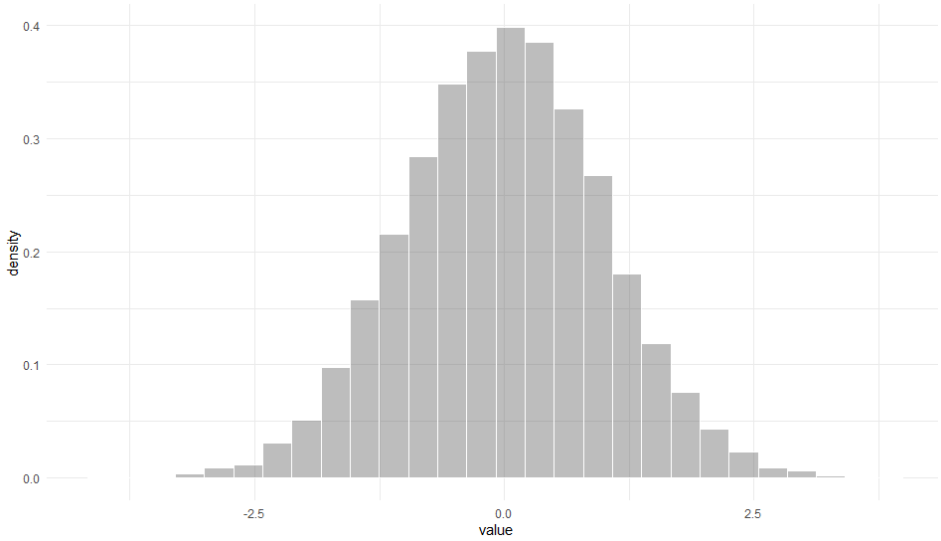
Rys. 3.18. Estymacja gęstości dwuwymiarowej (długość i szerokość kielicha) według gatunku

Na rysunku 3.18 przedstawiono wizualizację estymacji gęstości rozkładu dwuwymiarowego dla zmiennych długość (*Sepal.Length*) i szerokość kielicha (*Sepal.Width*) według gatunku.

3.6. Funkcje i mieszanki zmiennych losowych

W badaniach naukowych bardzo często należy się odwołać nie do zmiennej losowej, której rozkład jest znany, ale do kombinacji kilku zmiennych losowych lub do ich mieszanek. Często prowadzi to do bardzo złożonych rozważań teoretycznych, a w wielu przypadkach uzyskanie postaci rozkładu zmiennej losowej na drodze rozważań teoretycznych jest wręcz niemożliwe. W takim przypadku pomocny może być pakiet **`convdistr`**, który wykorzystując metody symulacyjne, pozwala na przedstawienie graficznego obrazu kombinacji lub mieszanek zmiennych losowych. Pakiet ten umożliwia również generowanie obserwacji losowych z takiego rozkładu, co ma duże znaczenie w analizach symulacyjnych. Dla uzyskania przybliżonego graficznego obrazu zmiennej losowej o rozkładzie normalnym standardowym można wykorzystać następujące komendy:

```
x <- new  NORMAL(0,1)
ggDISTRIBUTION(x)+theme_minimal()
```

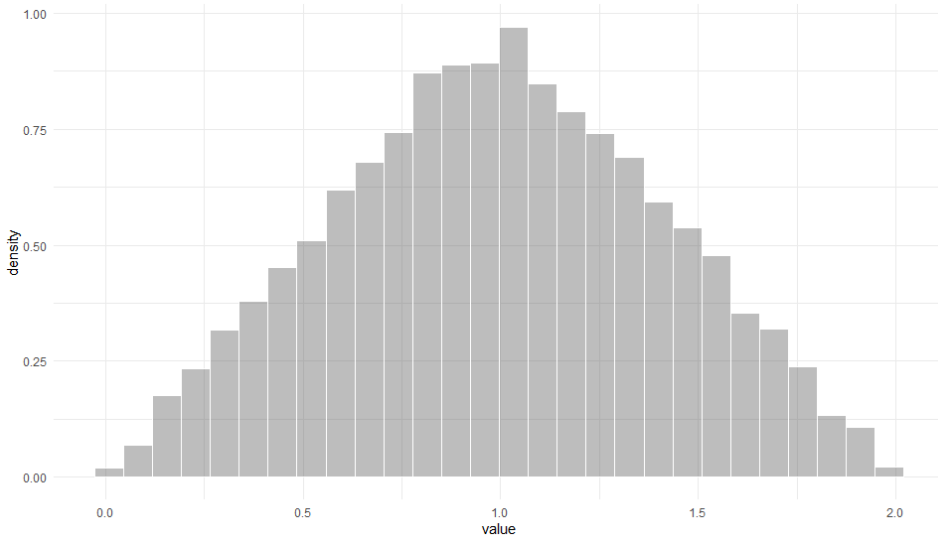


Rys. 3.19. Estymacja gęstości rozkładu normalnego standardowego

Na rysunku 3.19 przedstawiono wizualizację estymacji gęstości zmiennej losowej o rozkładzie normalnym standardowym.

Pakiet **convdistr** umożliwia bardzo wygodną ocenę rozkładów funkcji zmiennych losowych. Rozkład sumy dwóch zmiennych losowych o rozkładach jednostajnych na przedziale (0, 1) jest rozkładem trójkątnym określonym na przedziale (0, 2). Symulacyjną estymację takiego rozkładu przedstawia rys. 3.20, który uzyskano po wykonaniu następującego kodu:

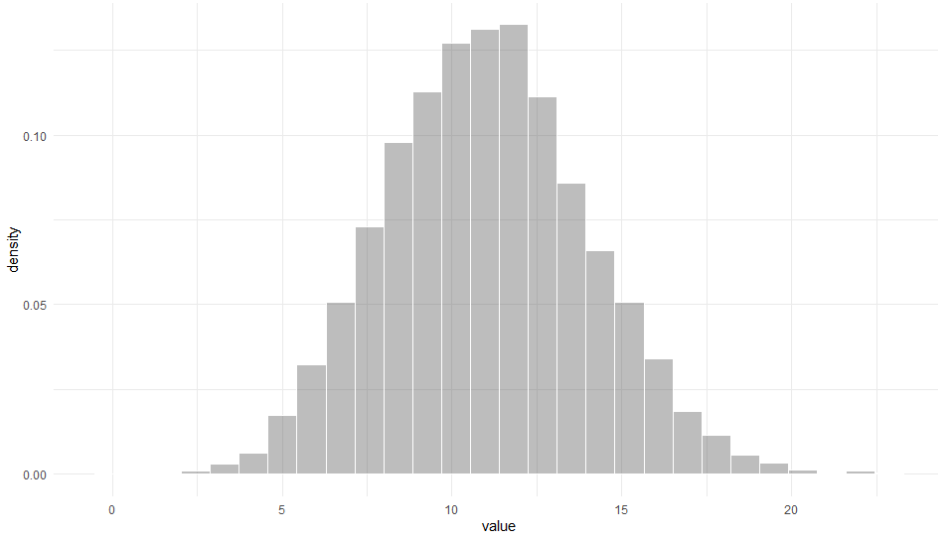
```
d1 <- new_UNIFORM(0,1)
d2 <- new_UNIFORM(0,1)
dsum <- new_SUM(list(d1,d2))
ggDISTRIBUTION(dsum)+theme_minimal()
```



Rys. 3.20. Estymacja gęstości sumy dwóch zmiennych losowych o rozkładzie jednostajnym

Dla uzyskania rozkładu sumy trzech zmiennych losowych o rozkładach normalnym, jednostajnym i Poissona należy wykonać poniższy kod, którego rezultat przedstawia rys. 3.21.

```
d1 <- new_NORMAL(1,1)
d2 <- new_UNIFORM(2,8)
d3 <- new_POISSON(5)
dsum <- new_SUM(list(d1,d2,d3))
ggDISTRIBUTION(dsum)+theme_minimal()
```



Rys. 3.21. Estymacja gęstości sumy zmiennych losowych o rozkładach normalnym, jednostajnym i Poissona

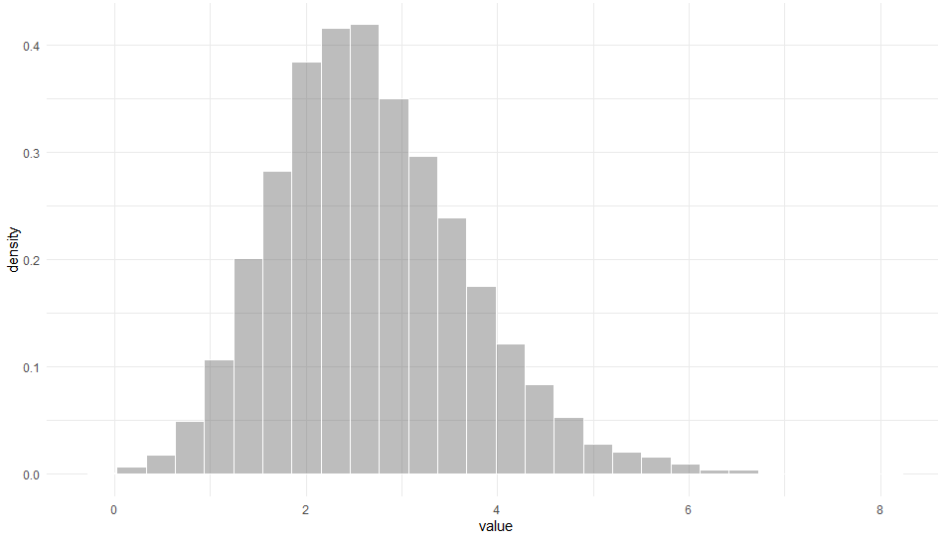
Dla uzyskania graficznego obrazu zmiennej losowej Y będącej sumą zmiennych losowych o rozkładzie normalnym (X_1) oraz zmiennej losowej będącej iloczynem zmiennych losowych o rozkładach Poissona (X_2) oraz beta (X_3) odpowiednio z parametrami jak poniżej:

$$Y = X1 + X2 * X3$$

gdzie $X_1 \sim N(1;\ 0{,}5)$, $X_2 \sim Poisson(5)$, $X_3 \sim Beta(10{,}20)$

należy wykonać następujący kod:

```
X1 <- new_NORMAL(1,0.5)
X2 <- new_POISSON(5)
X3 <- new_BETA(10,20)
res <- X1 + X2 * X3
ggDISTRIBUTION(res) + ggtitle("X1 + X2 * X3")+theme_minimal()
```

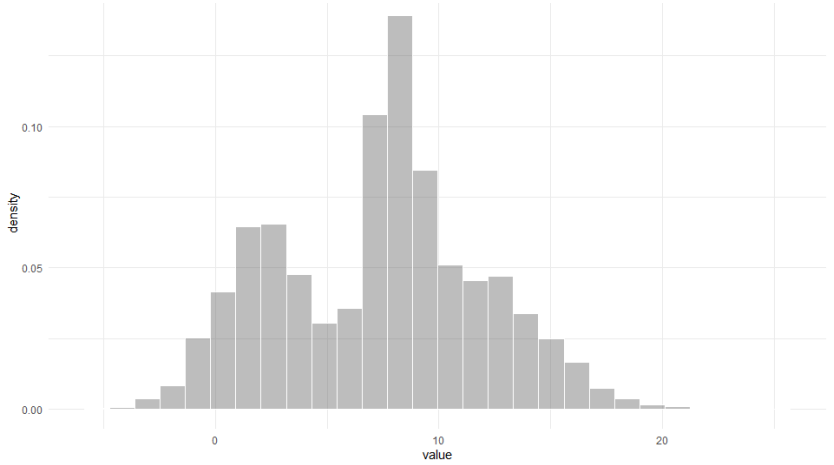


Rys. 3.22. Estymacja gęstości sumy zmiennych losowych o rozkładzie normalnym oraz iloczynu zmiennych losowych o rozkładach Poissona i beta

W wyniku realizacji powyższego kodu otrzymuje się graficzną prezentację estymacji gęstości zmiennej losowej Y (zob. rys. 3.22).

Pakiet **`convdistr`** umożliwia także uzyskanie wizualizacji mieszanek rozkładów zmiennych losowych. Dla uzyskania mieszanki trzech rozkładów zmiennych losowych o rozkładach normalnych o różnych wartościach oczekiwanych i wariancjach należy wykonać następujący kod:

```
d1 <- new_NORMAL(2,2)
d2 <- new_NORMAL(8,1)
d3 <- new_NORMAL(12,3)
dmix <- new_MIXTURE(list(d1,d2,d3))
ggDISTRIBUTION(dmix)+ggtitle("")+theme_minimal()
```

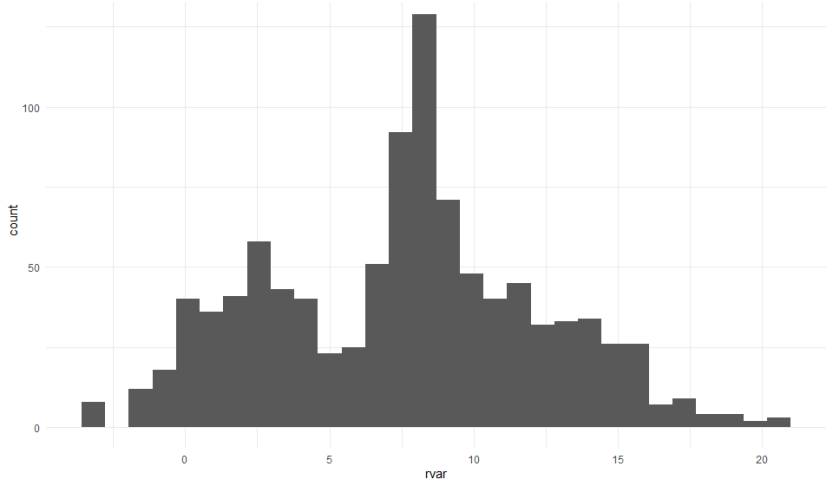


Rys. 3.23. Estymacja gęstości mieszanki trzech zmiennych losowych o rozkładach normalnych

Rysunek 3.23 przedstawia wynik realizacji powyższego kodu. Jest na nim zaprezentowana gęstość mieszanki trzech zmiennych losowych o rozkładach normalnych z różnymi parametrami.

Ogromną zaletą rozważanego pakietu jest możliwość uzyskania losowych wartości z otrzymanych rozkładów zmiennych losowych – funkcji lub mieszanek. Uzyskanie 1000 losowych wartości dla ostatniej mieszanki i przedstawienie wykresu jest realizowane następująco:

```
dane <- rfunc(dmix, 1000)
ggplot(dane,aes(rvar))+geom_histogram()+theme_minimal()
```



Rys. 3.24. Histogram 1000 losowych wartości z rozkładu mieszanki zmiennych d1, d2 i d3

4. Wizualizacja wyników wnioskowania statystycznego – wybrane biblioteki programu R

> *Mogą być nie tak odległe czasy, w których zrozumie się,*
> *że do całkowitego wtajemniczenia dojrzałego obywatela […],*
> *zdolność obliczania, myślenia w kategoriach średnich, maksimów i minimów*
> *jest tak konieczna, jak teraz zdolność czytania i pisania.*
>
> Herbert G. Wells (1903, s. 136)

W tym rozdziale przedstawiono wybrane biblioteki programu **R**, które wspierają graficzną prezentację wyników wnioskowania statystycznego. Omówione pakiety rozszerzają możliwości **`ggplot2`**, umożliwiając wizualizację rezultatów testów (wartości statystyk, obszary krytyczne, *p*-wartości, przedziały ufności), a także łączenie klasycznych procedur testowych z nowoczesnymi metodami symulacyjnymi. Szczególny nacisk położono na to, jak wyniki testowania hipotez można w sposób czytelny i zwięzły zaprezentować na wykresach.

W kolejnych podrozdziałach zaprezentowano zastosowania wybranych pakietów programu **R** do wizualizacji wyników testów: chi-kwadrat (zgodności i niezależności), testów dla proporcji, współczynnika korelacji, testów *t* (dla prób niezależnych i zależnych), testów dla wariancji oraz analizy wariancji (ANOVA). Zwieńczeniem rozdziału jest przykład „wizualizacji wnioskowania" w pełni zdefiniowanej przez użytkownika – implementacja testu pustych cel jako testu wielowymiarowej normalności wraz z graficzną prezentacją rozkładu statystyki, obszaru krytycznego i decyzji testowej.

4.1. Wnioskowanie statystyczne – wybrane biblioteki programu R

W pracy Kończaka (2024a) wskazano na różnorodne biblioteki pozwalające rozszerzyć możliwości pakietu graficznego **`ggplot2`**. W tej części skupiono się na pakietach, które mogą być zastosowane w graficznych prezentacjach rezulta-

tów wnioskowania statystycznego. Wykaz pakietów, które zostały wykorzystane w tym rozdziale, przedstawiono w tabeli 4.1. Zaprezentowane w niej pakiety rozszerzeń w różny sposób wspomagają przeprowadzenie wnioskowania statystycznego. Pakiety **`gginference`**, **`ggpval`**, **`ggpubr`**, **`ggsignif`**, **`ggstatsplot`**, **`ggdist`** i **`infer`** umożliwiają przeprowadzenie wnioskowania statystycznego, wzbogacając prezentacje o wyświetlenie wyników testowania hipotez na wykresie. Mogą to być dla przykładu wartości statystyk testowych, zaznaczone obszary krytyczne i wskazane p-wartości.

Tabela 4.1. Wybrane biblioteki programu R wspomagające prezentację wyników wnioskowania statystycznego

Biblioteka	Opis
`gginference`	Wizualizacja wyników wybranych testów statystycznych
`ggpval`	Przeprowadzanie testów statystycznych i wprowadzanie p-wartości na wykres
`ggpubr`	Funkcje do tworzenia i dostosowywania gotowych do publikacji wykresów opartych na **`ggplot2`**
`ggsignif`	Wskazanie na wykresie, czy pomiędzy grupami istnieją istotne statystycznie różnice
`ggstatsplot`	Konstrukcja wykresów ze szczegółami wyników testów statystycznych
`ggdist`	Zestaw funkcji zaprojektowanych specjalnie do wizualizacji rozkładów i niepewności
`infer`	Pakiet do przeprowadzania testów statystycznych zgodnie ze składnią **`tidyverse`**

Źródło: Opracowanie własne na podstawie: CRAN (2026).

Na wstępie konieczne jest załadowanie bibliotek, które będą wykorzystywane w tym rozdziale:

```
library(gginference)
library(ggpval)
library(ggpubr)
library(ggsignif)
library(ggstatsplot)
library(ggdist)
library(infer)
library(ggplot2)
library(tidyverse)
library(data.table)
library(ggmosaic)
```

4.2. Wizualizacja wyników wnioskowania statystycznego

Wnioskowanie statystyczne to dział statystyki zajmujący się metodami umożliwiającymi uogólnianie wyników uzyskanych na podstawie próby na całą populację. Obejmuje ono zarówno szacowanie nieznanych parametrów populacji

(np. średniej, wariancji, współczynnika korelacji), jak i testowanie hipotez dotyczących tej populacji. Kluczową cechą wnioskowania statystycznego jest uwzględnianie niepewności i błędów wynikających z tego, że badania przeprowadza się na próbie, a nie na całej populacji. W tym podrozdziale przedstawiono możliwości wizualizacji wyników klasycznych testów statystycznych w oparciu o wybrane biblioteki programu **R**.

4.2.1. Podstawowe określenia związane z wnioskowaniem statystycznym

W niniejszym podrozdziale przedstawiono kluczowe pojęcia związane z wnioskowaniem statystycznym (por. np. Hellwig, 1998; Zeliaś i in., 2002; Wywiał, 2012).

Populacja (zbiorowość generalna, uniwersum) to cały zbiór jednostek, których dotyczy wnioskowanie. Mogą to być np. wszyscy mieszkańcy kraju, wszyscy uczniowie określonej szkoły, wszyscy potencjalni klienci firmy. Zwykle oznacza się ją ogólnie jako zbiór wszystkich możliwych realizacji zmiennej losowej X o pewnym nieznanym rozkładzie (np. z nieznaną średnią μ i wariancją σ^2).

Próba to podzbiór populacji, faktycznie obserwowane dane, na podstawie których wnioskuje się o populacji. Zwykle zakłada się, że są to obserwacje niezależne o identycznych rozkładach (i.i.d.), co może być zapisane: $X_1, X_2, \ldots, X_n \sim F$, niezależne, gdzie n to liczebność próby, a F to nieznany rozkład (dystrybuanta) populacji.

Wnioskowanie statystyczne opiera się na weryfikacji hipotez, czyli formalnych przypuszczeń dotyczących populacji. Formułowane są dwie wykluczające się hipotezy:

Hipoteza zerowa (H_0) to stwierdzenie, które domyślnie przyjmuje się za prawdziwe. Zazwyczaj odpowiada brakowi efektu, różnicy lub zależności. W procesie wnioskowania dąży się do odrzucenia tej hipotezy na podstawie danych z próby.

Hipoteza alternatywna (H_1) to operacyjne sformułowanie problemu badawczego, które precyzuje, jaki efekt (wraz z jego ewentualnym kierunkiem) ma zostać wykryty. Często jest to formalne zaprzeczenie hipotezy zerowej. Opisuje stan, który badacz chce potwierdzić, jak np. istnienie określonego efektu, różnicy lub zależności.

Hipotezy dotyczące parametrów populacji zwykle są zapisywane w postaci:

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0, H_1: \theta > \theta_0 \text{ lub } H_1: \theta < \theta_0$$

W procesie weryfikacji hipotezy możliwe jest popełnienie błędów I lub II rodzaju.

Błąd I rodzaju polega na odrzuceniu hipotezy zerowej, która w rzeczywistości jest prawdziwa. Jest to tzw. fałszywy alarm, gdy dochodzi się do wniosku, że istnieje efekt, podczas gdy w rzeczywistości go nie ma. Prawdopodobieństwo popełnienia błędu I rodzaju jest zwykle oznaczane przez α.

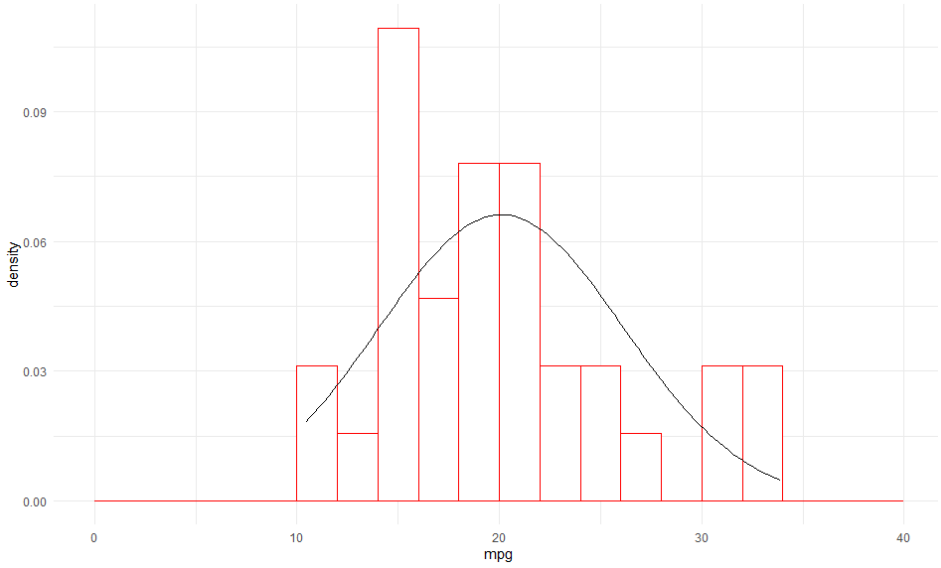
Błąd II rodzaju polega na stwierdzeniu braku podstaw do odrzucenia hipotezy zerowej, która w rzeczywistości była fałszywa. Jest to „przeoczenie" istniejącego efektu, które może być spowodowane np. zbyt małą liczebnością próby lub niedostateczną jej reprezentatywnością. Prawdopodobieństwo popełnienia błędu II rodzaju jest zwykle oznaczane przez β.

Moc testu to prawdopodobieństwo prawidłowego odrzucenia hipotezy zerowej, gdy jest ona fałszywa. Jest to zdolność testu do wykrywania efektu, który faktycznie istnieje. W praktyce dąży się do tego, aby testy miały jak największą moc. Przy przyjętych oznaczeniach moc testu to $1 - \beta$.

4.2.2. Wizualizacja w ggplot2

Bardzo często przed wykonaniem testu statystycznego konieczne jest sprawdzenie wymaganych założeń, a najczęściej jest to wymaganie dotyczące normalności rozkładu badanej zmiennej losowej. Wymóg ten pojawia się przy stosowaniu wielu testów parametrycznych. W związku z powyższym w pierwszym etapie jest weryfikowana stosowna hipoteza. Przykład graficznego wglądu w proces weryfikacji hipotezy głoszącej, że zmienna *mpg* (zbiór **mtcars**) ma rozkład normalny, realizuje następujący kod:

```
ggplot(data=mtcars, mapping=aes(mpg)) +
  geom_histogram(mapping=aes(y =after_stat(density)),
                 breaks = seq(0, 40, by = 2),
                 colour = "red",
                 fill = "white") +
  stat_function(fun = dnorm, args = list(mean = mean(mtcars$mpg), sd = sd(mtcars$mpg
)))+theme_minimal()
```



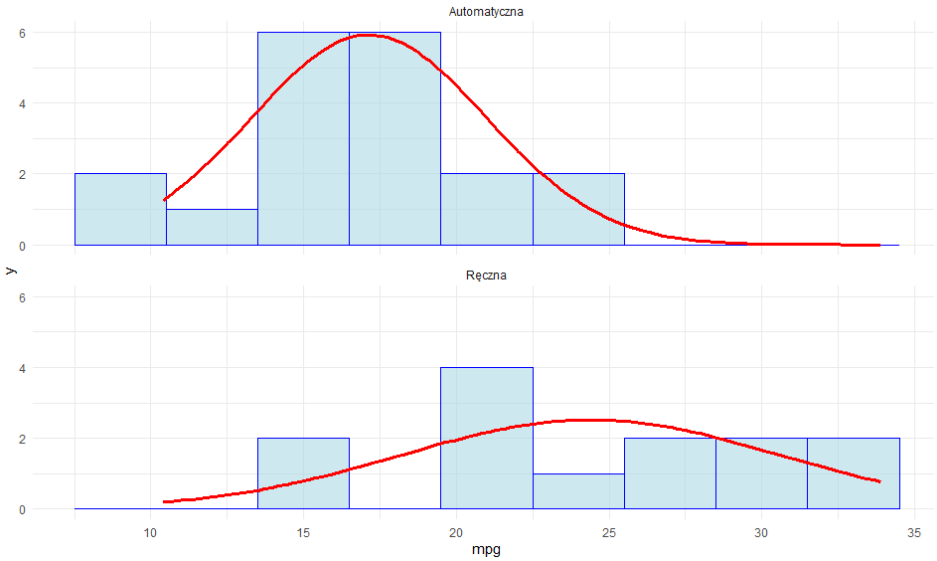
Rys. 4.1. Estymacja gęstości i krzywa rozkładu normalnego dla zmiennej mpg

Wynik realizacji powyższego kodu został przedstawiony na rys. 4.1. Poza histogramem, który jest oszacowaniem teoretycznego rozkładu gęstości badanej zmiennej, umieszczono na nim krzywą gęstości rozkładu normalnego z parametrami, które zostały oszacowane na podstawie danych.

W badaniach statystycznych bardzo często zachodzi konieczność porównania dwóch populacji na podstawie pobranych prób losowych. Zwykle interesujące jest porównanie gęstości rozkładów, których estymatorem są odpowiednie histogramy. Wstępną analizę graficzną rozkładów dla dwóch grup umożliwia następujący kod:

```
df_mtcars <- mtcars
df_mtcars$group <- factor(x=df_mtcars$am, labels = c("Automatyczna", "Ręczna"))
bw=3

grid_data <- df_mtcars %>%
  group_by(group) %>%
  summarise(
    mean = mean(mpg),
    sd = sd(mpg),
    n = n(),
    .groups = "drop") %>%
  mutate(
    mpg = map(group, ~ seq(min(df_mtcars$mpg), max(df_mtcars$mpg), length.out = 100)),
    y = pmap(list(mpg, mean, sd, n), function(x, m, s, n) dnorm(x, m, s) * bw * n)) %>%
  unnest(c(mpg, y))
```

```
ggplot() +
  geom_histogram(data = df_mtcars, aes(x = mpg), binwidth = bw, colour = "blue", fill =
"lightblue", alpha = 0.6) +
  geom_line(data = grid_data, aes(x = mpg, y = y), color = "red", linewidth = 1.2) +
  facet_wrap(~ group, ncol = 1) +
  theme_minimal()
```



Rys. 4.2. Histogramy i gęstości dwóch rozkładów

Wynik realizacji powyższego kodu został przedstawiony na rys. 4.2. Zauważalne jest, że wartość oczekiwana *mpg* jest mniejsza dla samochodów z automatyczną skrzynią biegów, a jednocześnie odchylenie standardowe *mpg* w tej grupie jest większe.

4.2.3. Pakiet gginference

Przedstawiona wcześniej biblioteka **`gginference`** pozwala nie tylko na wykreślenie teoretycznych postaci gęstości lub dystrybuant zmiennych losowych, ale również na przeprowadzenie testów statystycznych i zaprezentowanie odpowiednich wyników w formie graficznej. W tej części zostanie wskazanych kilka przykładów takich realizacji z wykorzystaniem tej biblioteki.

We wnioskowaniu statystycznym stosuje się różne testy odwołujące się do statystyki chi-kwadrat. Do najprostszych testów opartych na tej statystyce jest zaliczany test zgodności dla proporcji.

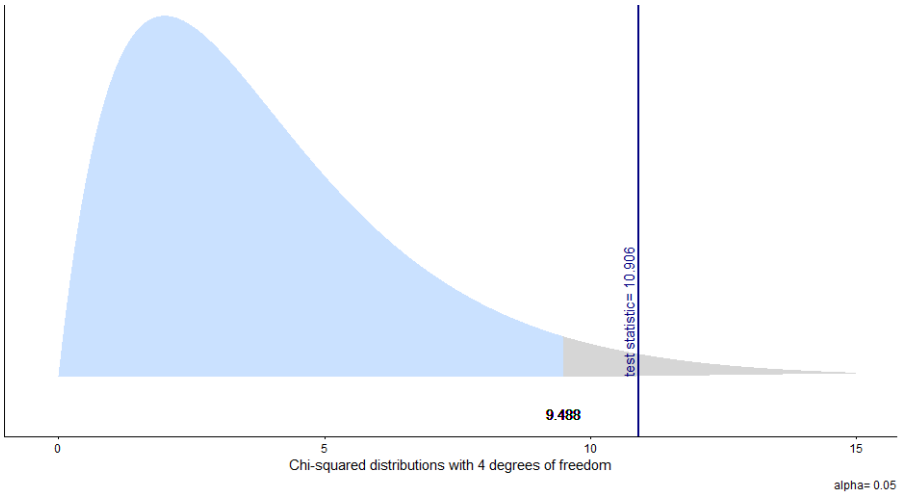
W badaniu ankietowym przeprowadzonym na próbie 320 osób, traktowanej jako próba losowa z populacji potencjalnych nabywców samochodów, pytano o preferowany kolor samochodu. Uzyskane liczebności odpowiedzi w poszczególnych kategoriach przedstawia tabela 4.2.

Tabela 4.2. Preferencje kolorów samochodów i liczba wskazań

Kolor samochodu	Srebrny	Czarny	Biały	Niebieski	Czerwony
Liczba wskazań	82	70	65	55	48

Test zgodności dla proporcji pozwala zweryfikować hipotezę, że prawdziwe (populacyjne) udziały wskazań poszczególnych kolorów mają określone, z góry zadane wartości. Na podstawie powyższych danych jest weryfikowana hipoteza H_0 głosząca, że wszystkie kolory samochodów są tak samo preferowane, wobec hipotezy alternatywnej H_1 głoszącej, że kolory nie są jednakowo preferowane.

```
x <- c(Srebrny = 82, Czarny = 70, Biały = 65, Niebieski = 55, Czerwony = 48)
ggchisqtest(chisq.test(x))
```

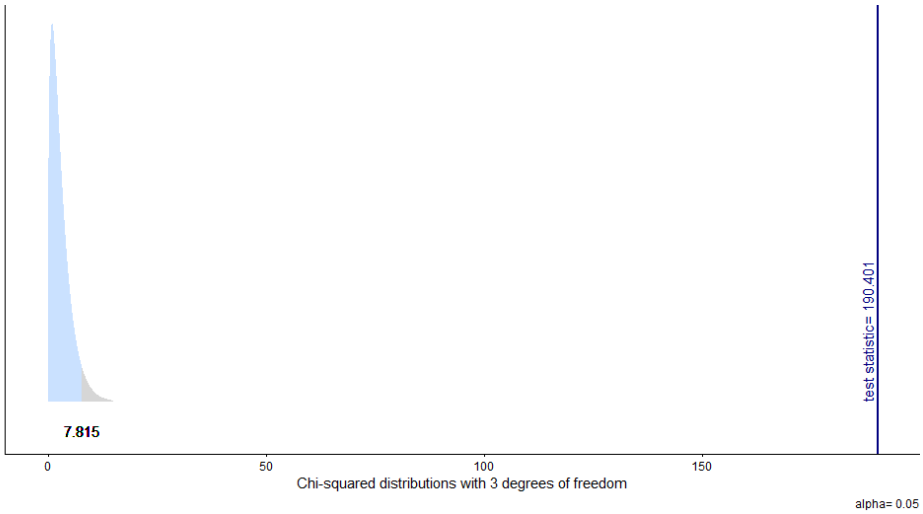


Rys. 4.3. Wizualizacja wyników testu chi-kwadrat dla proporcji z gginference. Preferencje kolorów samochodu

Rezultat realizacji powyższych komend przedstawiono na rys. 4.3. Na wykresie pionowa linia wskazuje wartość obserwowanej statystyki testowej ($\chi^2 = 10{,}906$). Przedstawiono także wartość krytyczną ($\chi^2_{4;\,0{,}05} = 9{,}488$) i obszar krytyczny dla tego testu (szare pole). Graficzna prezentacja wyników przy przyjętym poziomie istotności $\alpha = 0{,}05$ prowadzi do odrzucenia hipotezy H_0, a więc można twierdzić, że kolory samochodów nie cieszą się jednakowym zainteresowaniem badanych.

Kolejnym bardzo często wykorzystywanym testem w badaniach naukowych jest test niezależności chi-kwadrat. Test ten pozwala na weryfikację hipotezy o niezależności dwóch zmiennych jakościowych, dla których zostały wyróżnione pewne kategorie. W poniżej przedstawionym przykładzie weryfikacji została poddana hipoteza H_0 głosząca, że przeżycie katastrofy statku Titanic (zbiór danych **Titanic**) nie zależało od klasy, którą podróżowała dana osoba. Weryfikacji tej hipotezy służy następujący kod:

```
tab_Titanic <- margin.table(Titanic, margin = c(1, 4))
ggchisqtest(chisq.test(tab_Titanic))
```

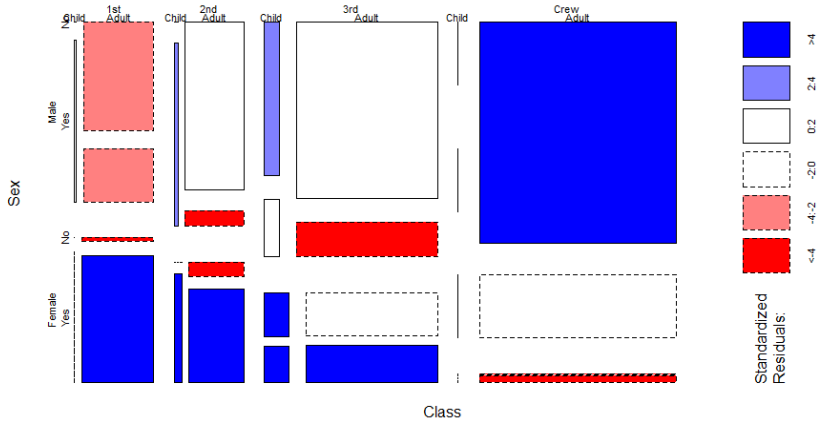


Rys. 4.4. Wizualizacja wyników testu chi-kwadrat niezależności z gginference. Przeżycie katastrofy statku Titanic

Na rysunku 4.4 przedstawiono wizualizację wyników weryfikacji hipotezy głoszącej, że klasa, w której przebywał pasażer statku Titanic, nie miała wpływu na przeżycie katastrofy. Zaznaczony na tym rysunku obszar krytyczny i wartość krytyczna $\chi^2_{3;\,0,05} = 7{,}815$ oraz wartość statystyki testowej $\chi^2 = 190{,}401$ prowadzą do odrzucenia hipotezy o niezależności dwóch rozważanych zmiennych przy poziomie istotności $\alpha = 0{,}05$. Bardzo duża wartość statystyki testowej świadczy przeciw tej hipotezie. Można twierdzić, że klasa miała wpływ na przeżycie katastrofy.

Do wizualizacji wyników testu na podstawie danych z tablic wielodzielczych można wykorzystać także funkcję *mosaicplot()* dostępną w podstawowej wersji programu **R**.

```
plot(Titanic,shade=TRUE, main='')
```



Rys. 4.5. Wizualizacja wyników testu chi-kwadrat niezależności z pakietem mosaic. Przeżycie katastrofy statku Titanic

Wykres mozaikowy (zob. rys. 4.5) przedstawia zależności pomiędzy klasą podróży, płcią oraz wiekiem pasażera a przeżyciem katastrofy. Każdy prostokąt na wykresie odpowiada określonej grupie pasażerów, a jego wielkość odzwierciedla liczebność tej grupy. Kolorowanie (shade=TRUE) wskazuje na istotność statystyczną odchyleń od oczekiwanej liczebności – intensywniejsze kolory oznaczają większe różnice.

Najwyższy odsetek przeżywalności dotyczył kobiet i dzieci, zwłaszcza podróżujących pierwszą i drugą klasą. Mężczyźni, szczególnie z niższych klas, mieli znacznie mniejsze szanse na przeżycie. Klasa podróży miała istotny wpływ na przeżywalność – pasażerowie pierwszej klasy przeżywali częściej niż pasażerowie trzeciej klasy. Wiek również odgrywał rolę – dzieci miały większe szanse na przeżycie niż dorośli, zwłaszcza w wyższych klasach. Wykres mozaikowy wyraźnie pokazuje, że przeżywalność na Titanicu była silnie związana z płcią, wiekiem i klasą podróży pasażerów.

Jednym z najczęściej stawianych pytań w badaniach statystycznych jest pytanie o występowanie zależności pomiędzy zmiennymi. Jeżeli dotyczy to zmiennych mierzonych na skalach mocnych, to zazwyczaj wykorzystuje się test dla współczynnika korelacji liniowej Pearsona. Przez skale mocne rozumie się skale ilościowe (przedziałowe i ilorazowe), w których jest zachowany zarówno porządek wartości, jak i znaczenie różnic (a w przypadku skali ilorazowej także punkt zerowy). Skale słabe to skale nominalne i porządkowe. W skali nominalnej rozróżnia się jedynie kategorie bez porządku (np. płeć, kolor, rodzaj paliwa), a w skali porządkowej znany jest porządek kategorii, ale różnice między nimi nie mają liczbowego znaczenia (np. oceny w ankiecie: „zły”, „średni”, „dobry”). W przypadku skal słabych stosuje się inne metody analizy zależności, takie jak testy nieparametryczne czy współczynniki korelacji rang.

Weryfikowana hipoteza H_0 głosząca, że wartość współczynnika korelacji liniowej w populacji (ρ) wynosi 0, ma postać:

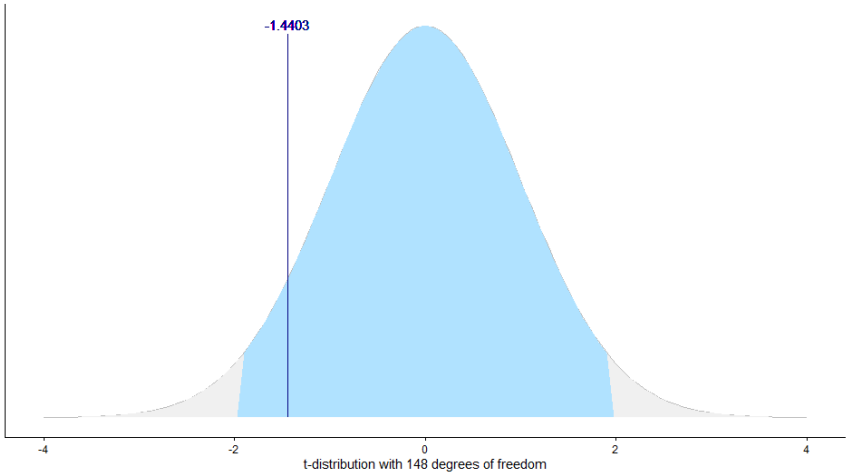
$$H_0: \rho = 0$$

wobec hipotezy alternatywnej:

$$H_1: \rho \neq 0$$

Przykład zastosowania takiego testu dla danych o szerokości i długości kielicha kwiatu iris na podstawie danych ze zbioru **iris** przedstawia poniższy kod.

```
corr_test <- cor.test(iris$Sepal.Length, iris$Sepal.Width)
ggcortest(corr_test)
```



Rys. 4.6. Wizualizacja wyników testu dla współczynnika korelacji z gginference. Długość i szerokość płatka

Na rysunku 4.6 przedstawiono graficznie wyniki testu. Wizualizacja wskazuje na brak podstaw do odrzucenia hipotezy H_0 głoszącej, że nie ma zależności pomiędzy długością i szerokością kielicha kwiatu iris (dane ze zbioru **iris**).

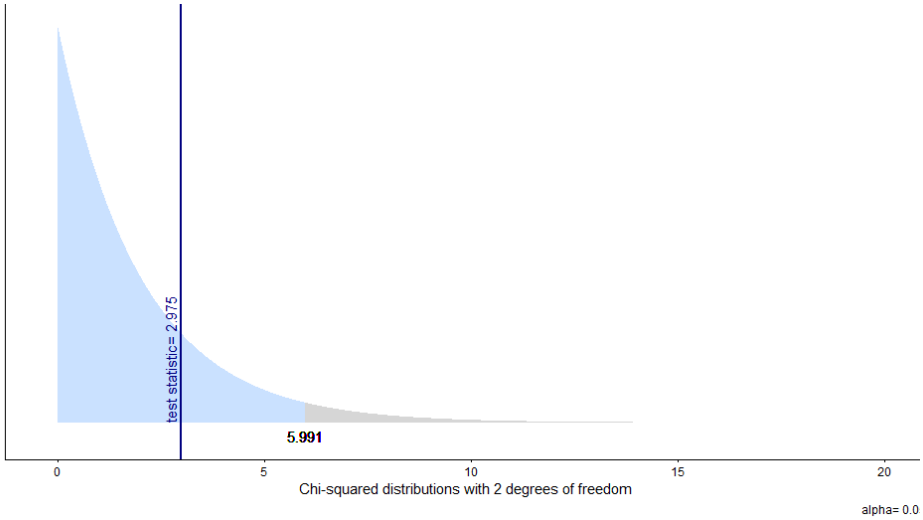
Kolejny przykład przedstawia zastosowanie testu dla równości proporcji w trzech populacjach. Weryfikowana hipoteza głosi, że proporcje w trzech populacjach (p_1, p_2 i p_3) są jednakowe i może być zapisana następująco:

$$H_0: p_1 = p_2 = p_3$$

wobec hipotezy alternatywnej:

$$H_1: \text{nie wszystkie proporcje są jednakowe}$$

```
m <- c(52, 25, 50)
n <- c(80, 36, 64)
pr_test <- prop.test(m, n)
ggproptest(pr_test)
```

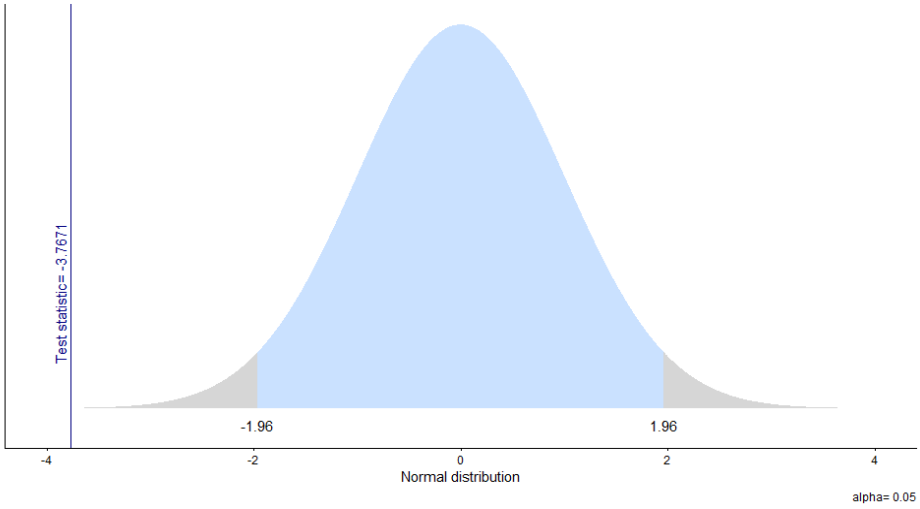


Rys. 4.7. Wizualizacja wyników testu równości proporcji z gginference. Równość trzech proporcji

Wynik realizacji testu dla równości proporcji został przedstawiony na rys. 4.7. Widoczne jest, że wartość statystyki testowej ($\chi^2 = 2{,}975$) jest znacznie mniejsza od wartości krytycznej $\left(\chi^2_{2;\,0{,}05} = 5{,}991\right)$, co prowadzi do stwierdzenia braku podstaw do odrzucenia hipotezy zerowej głoszącej, że proporcje w trzech populacjach są jednakowe.

Do najczęściej stosowanych testów statystycznych należy zaliczyć test t-Studenta dla równości dwóch wartości oczekiwanych. Przykład zastosowania takiego testu, pozwalającego na weryfikację hipotezy głoszącej, że przeciętna liczba mil przejechanych na galonie paliwa jest taka sama w samochodach z automatyczną i ręczną skrzynią biegów (zbiór danych **mtcars**), przedstawia kolejny kod:

```
t_test <- t.test(mtcars$mpg ~ mtcars$am)
ggttest(t_test)
```

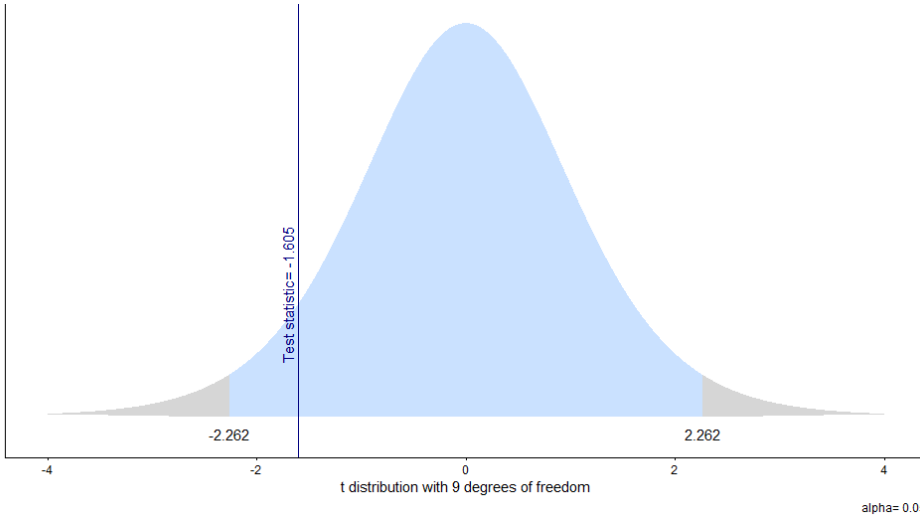


Rys. 4.8. Wizualizacja wyników testu t-Studenta z gginference. Równość wartości oczekiwanych dla grup niezależnych

Wynik tego testu został przedstawiony na rys. 4.8. Wartość statystyki testowej ($t = -3{,}761$) znajduje się w obszarze krytycznym, a więc można twierdzić, że liczba mil przejechanych na galonie paliwa w samochodach z ręczną i automatyczną skrzynią biegów nie jest jednakowa.

Inną formą testu t-Studenta dla dwóch prób jest test dla danych sparowanych (skojarzonych). Przykład zastosowania takiego testu przedstawia poniższy kod.

```
t_test2 <- t.test(x = rnorm(10), y = rnorm(10,1,1), paired=TRUE)
ggttest(t_test2)
```

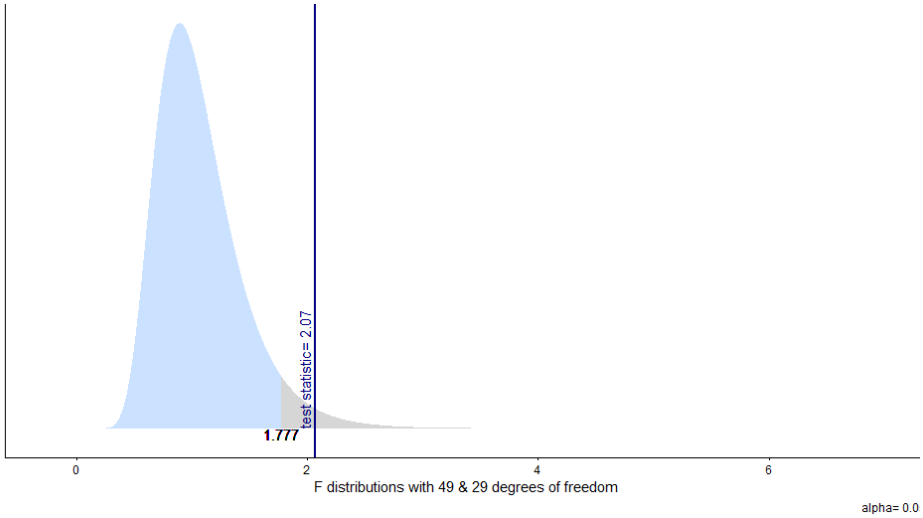


Rys. 4.9. Wizualizacja wyników testu t-Studenta z gginference. Równość wartości oczekiwanych dla grup zależnych

Wynik realizacji testu t-Studenta dla danych sparowanych został przedstawiony na rys. 4.9. Wartość statystyki testowej ($t = -1{,}605$) nie znajduje się w obszarze krytycznym, a więc brak podstaw do odrzucenia hipotezy H_0.

Niekiedy podczas prowadzenia wnioskowania statystycznego zachodzi potrzeba sprawdzenia równości wariancji w dwóch populacjach. Przykład realizacji tego testu dla umownych danych przedstawia następujący kod:

```
x <- rnorm(50, mean = 0, sd = 2)
y <- rnorm(30, mean = 1, sd = 1)
var_test <- var.test(x, y)
ggvartest(var_test)
```

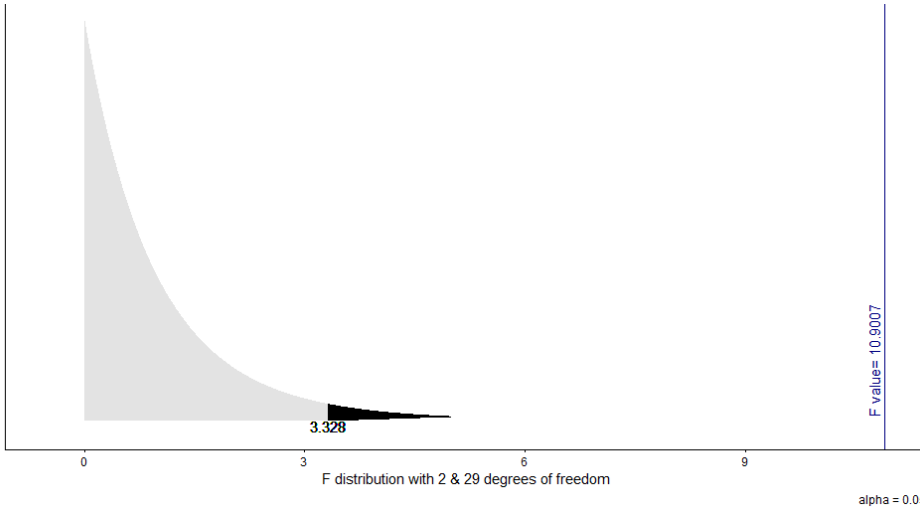


Rys. 4.10. Wizualizacja wyników testu F z gginference. Jednorodność wariancji

Wynik realizacji kodu dla weryfikacji hipotezy o równości dwóch wariancji został przedstawiony na rys. 4.10. Wartość statystyki testowej ($F = 2{,}07$) znajduje się w obszarze krytycznym. W związku z tym można twierdzić, że wariancje w obu populacjach nie są jednakowe.

Dotychczas przedstawione testy pozwalały na porównanie dwóch populacji. W badaniach statystycznych takie porównania często odnoszą się do większej liczby populacji. Porównanie wartości oczekiwanych dla trzech lub większej liczby populacji można przeprowadzić z wykorzystaniem testu analizy wariancji (ANOVA). Przykład kodu realizującego taki test, pozwalającego weryfikować hipotezę głoszącą, że przeciętna liczba przejechanych mil na jednym galonie paliwa (*mpg*) jest jednakowa w trzech grupach samochodów wyróżnionych ze względu na liczbę biegów (*gear*), przedstawiono poniżej:

```
mt_aov <- aov(mpg~factor(gear), data = mtcars)
ggaov(mt_aov, colaccept = "grey89", colreject = "black")
```



Rys. 4.11. Wizualizacja wyników testu z gginference. Test ANOVA

Wynik realizacji powyższego kodu został przedstawiony na rys. 4.11. Wartość statystyki testowej ($F = 10{,}9007$) znajduje się w obszarze krytycznym i na tej podstawie można stwierdzić, że przeciętna liczba mil przejechanych na jednym galonie paliwa (*mpg*) nie jest jednakowa dla samochodów o różnej liczbie biegów (*gear*).

4.2.4. Pakiet ggpval

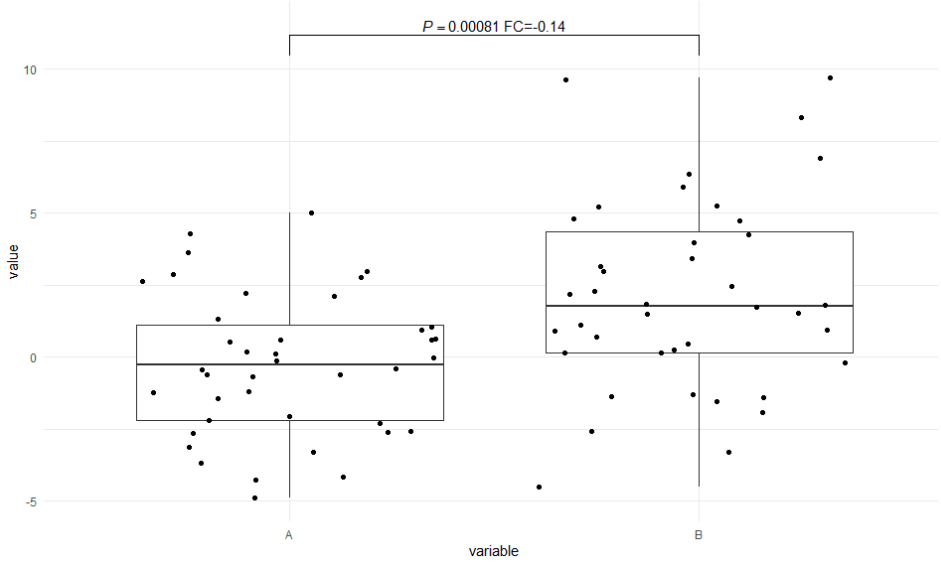
Kolejną biblioteką umożliwiającą nie tylko przeprowadzenie testów statystycznych, ale również graficzną prezentację wyników takich testów jest biblioteka **`ggpval`**. Biblioteka ta pozwala na czytelną wizualizację występujących różnic pomiędzy porównywanymi populacjami.

Do najczęściej stosowanych testów statystycznych należy test t-Studenta dla dwóch prób. Pakiet **`ggpval`** umożliwia wizualizację wyników tego testu. Możliwości te zostaną przedstawione na przykładzie danych umownych, które są utworzone w następujący sposób:

```
A <- rnorm(40, 0, 3)
B <- rnorm(40, 2, 4)
G <- rep(c("G1", "G2"), each = 20)
dt <- data.table::data.table(A, B, G)
dt <- data.table::melt(dt, id.vars = "G")
```

Prezentację graficzną dwóch zbiorów wraz z zaznaczeniem różnic przedstawia poniższy kod.

```
p <- ggplot(dt, aes(variable, value)) +
  geom_boxplot() +
  geom_jitter()
add_pval(p, pairs = list(c(1, 2)), fold_change=TRUE)+theme_minimal()
```

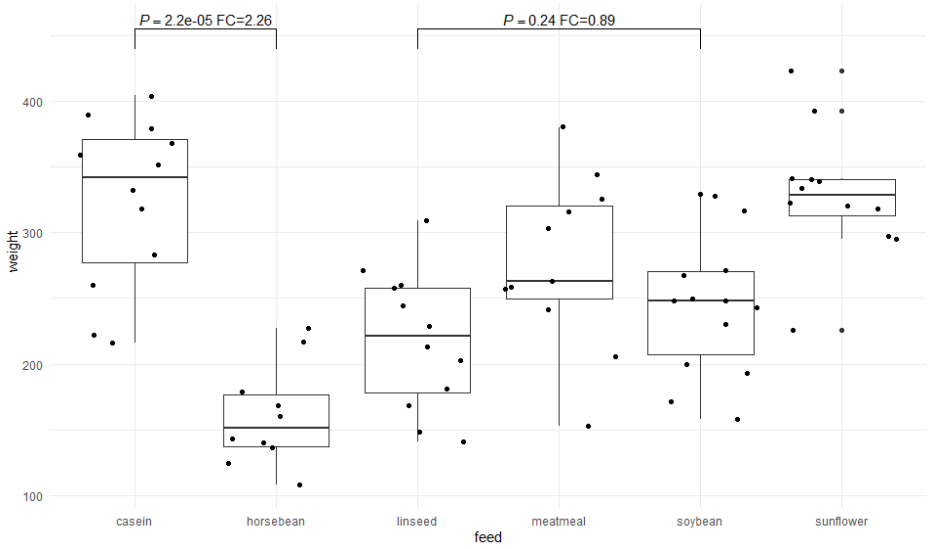


Rys. 4.12. Wizualizacja wyników testu z ggpval. Porównanie dwóch średnich

Na rysunku 4.12 przedstawiono wykres pudełkowy pozwalający na porównanie rozkładu dwóch prób. Na wykresie dodatkowo wskazano p-wartość, która potwierdza, że na poziomie istotności $\alpha = 0{,}05$ należy odrzucić hipotezę H_0, a więc można twierdzić, że wartości oczekiwane w populacjach nie są jednakowe.

W poprzednim przykładzie porównano dwie próby. Poniższy kod prowadzi do wizualizacji porównania parami dla 6 prób. Wynik został przedstawiony na rys. 4.13. Na wykresie zostały zamieszczone p-wartości dla porównań parami zmiennych 1 i 2 oraz 3 i 5. Przyjmując poziom istotności $\alpha = 0{,}05$, w pierwszym przypadku różnice są istotne, natomiast w drugim nie są istotne, co zostało zilustrowane na rys. 4.13.

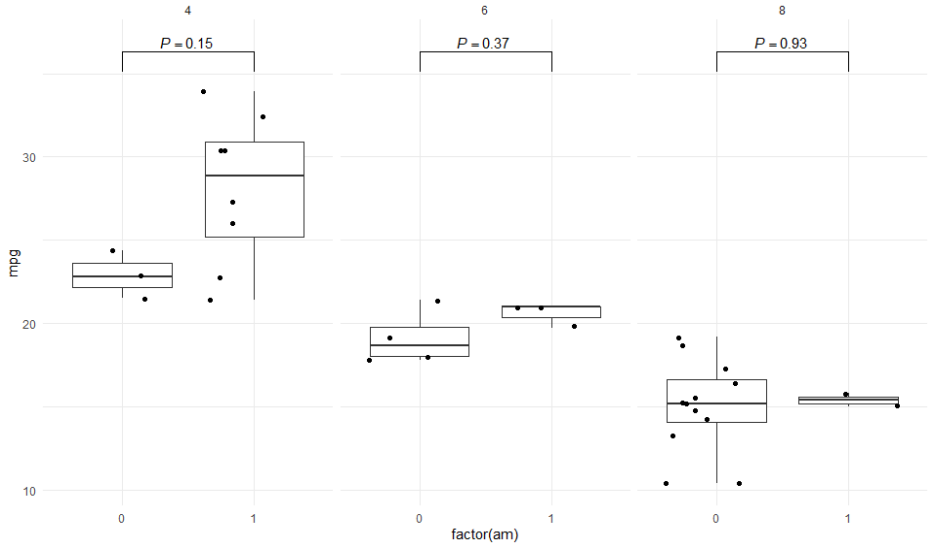
```
p <- ggplot(chickwts, aes(feed, weight)) +
  geom_boxplot() +
  geom_jitter()
add_pval(p, pairs = list(c(1, 2),c(3,5)), fold_change=TRUE)+theme_minimal()
```



Rys. 4.13. Wizualizacja wyników testu z ggpval. Wpływ różnych rodzajów paszy na masę kurcząt – porównanie dla wybranych par

Na rysunku 4.14 przedstawiono porównanie trzech grup w innej formie – z wykorzystaniem paneli. Dla porównywanych zmiennych na wykresach zostały zamieszczone *p*-wartości.

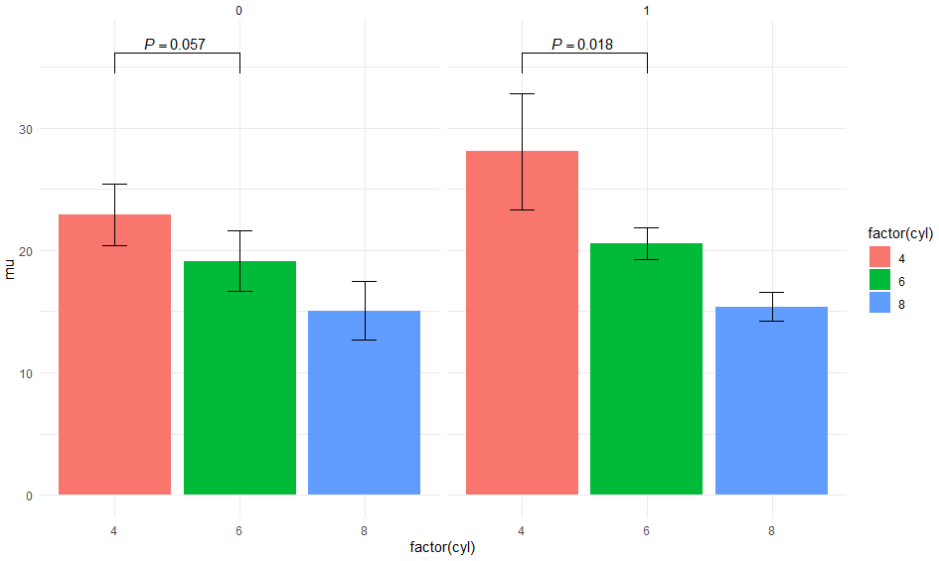
```
p <- ggplot(mtcars, aes(factor(am), mpg)) +
  geom_boxplot() +
  geom_jitter() +
  facet_wrap(~cyl)
add_pval(p, pairs = list(c(1, 2)))+theme_minimal()
```



Rys. 4.14. Wizualizacja wyników testu z ggpval. Wpływ typu skrzyni biegów na zużycie paliwa w zależności od liczby cylindrów

Na rysunku 4.15 przedstawiono porównania, nie jak poprzednio w formie wykresów pudełkowych, ale w formie wykresów słupkowych. Podobnie jak w poprzednich przykładach, wskazano *p*-wartości dla wybranych par zmiennych.

```
mt= data.table::as.data.table(mtcars)
mt[, mu := mean(mpg), by = c("am", "cyl")]
mt[, se := sd(mpg) / sqrt(.N), by = c("am", "cyl")]
p_bar <- ggplot(mt, aes(x=factor(cyl), y=mu, fill = factor(cyl))) +
  geom_bar(stat = "identity", position = 'dodge') +
  geom_errorbar(aes(ymin=mu-3*se, ymax=mu+3*se),width = .2) +
  facet_wrap(~am)
add_pval(p_bar, pairs = list(c(1, 2)), response = 'mpg')+theme_minimal()
```



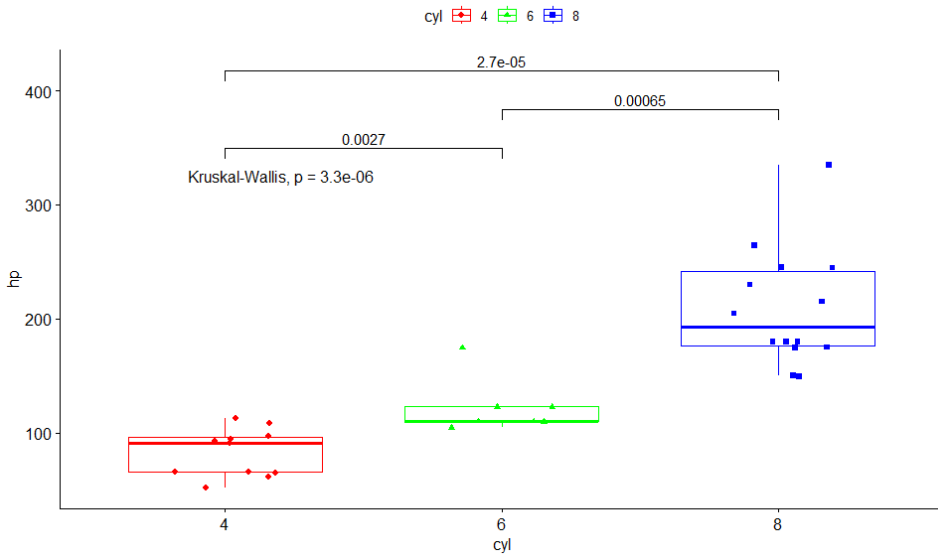
Rys. 4.15. Wizualizacja wyników testu z ggpval. Średnie zużycie paliwa (mpg) w zależności od liczby cylindrów i typu skrzyni biegów

4.2.5. Pakiet ggpubr

Kolejną biblioteką, która może być wykorzystana do graficznej prezentacji wyników testowania hipotez, jest biblioteka **ggpubr**. Możliwości tej biblioteki są zbliżone do wcześniej opisanej biblioteki **ggpval**. Funkcje tej biblioteki pozwalają w prosty sposób wzbogacić tradycyjne wykresy o wyniki testów statystycznych.

W poniższym przykładzie została przedstawiona graficzna prezentacja mocy silnika (*hp*) dla trzech grup wyróżnionych ze względu na liczbę cylindrów (*cyl*) na podstawie danych pochodzących ze zbioru **mtcars**.

```
p <- ggboxplot(mtcars, x = "cyl", y = "hp",
          color = "cyl", palette =c("red", "green", "blue"),
          add = "jitter", shape = "cyl")
porownanie <- list( c("4", "6"), c("6", "8"), c("4", "8") )
p + stat_compare_means(comparisons = porownanie)+
  stat_compare_means(label.y = 320)
```

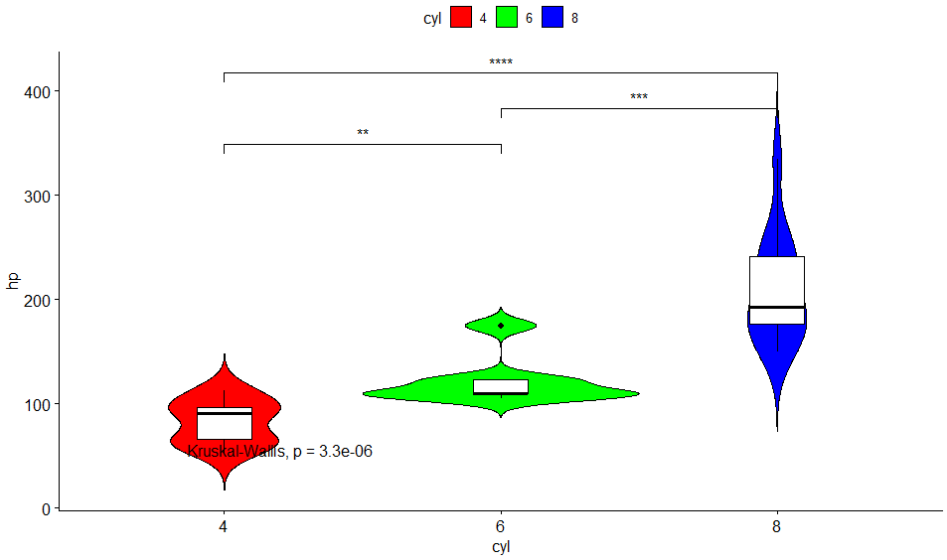


Rys. 4.16. Wykresy pudełkowe z wynikami testu z ggpubr. Moc silnika według liczby cylindrów

Rysunek 4.16 przedstawia wykres pudełkowy dla zmiennej *hp* ze zbioru **mtcars**. Na wykresie zostały wskazane *p*-wartości dla porównań wartości oczekiwanych *hp* dla trzech wyróżnionych grup ze względu na liczbę cylindrów.

Inną formę prezentacji, nie w formie wykresów pudełkowych, a skrzypcowych z pudełkami, wyników wnioskowania dla tych samych zmiennych przedstawia rys. 4.17.

```
ggviolin(mtcars, x = "cyl", y = "hp", fill = "cyl",
       palette = c("red", "green", "blue"),
       add = "boxplot", add.params = list(fill = "white"))+
 stat_compare_means(comparisons = porownanie, label = "p.signif")+
 stat_compare_means(label.y = 50)
```



Rys. 4.17. Wykresy skrzypcowe z wynikami testu z ggpubr. Moc silnika według liczby cylindrów

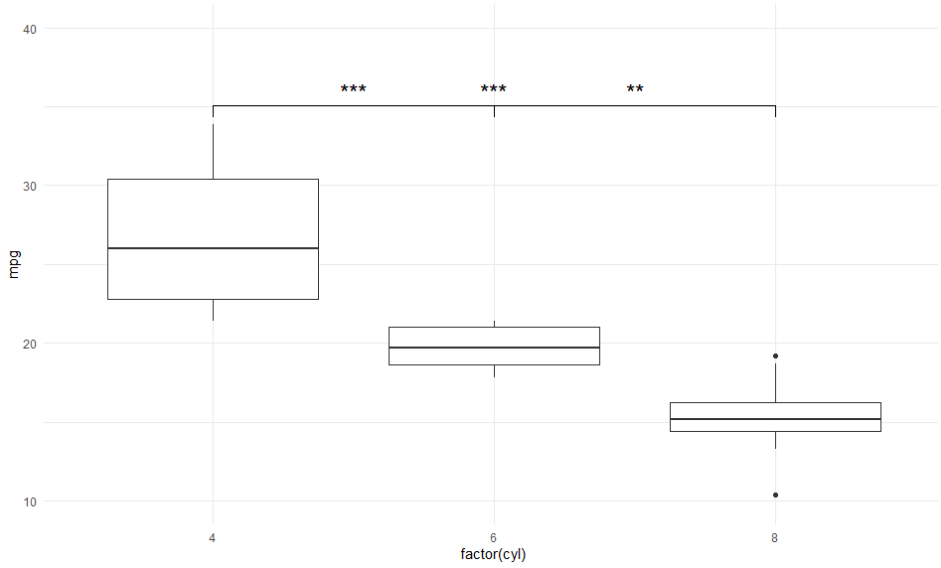
Na rysunku 4.17 poza wiolinami i pudełkami podano istotność dla testu Kruskala-Wallisa, a także symbolicznie oznaczono istotności różnic pomiędzy zmiennymi w wyróżnionych grupach.

4.2.6. Pakiet ggsignif

Kolejnym pakietem umożliwiającym wizualizację wyników wnioskowania statystycznego jest pakiet **`ggsignif`**. Umożliwia on dodawanie oznaczeń istotności statystycznej (np. *p*-wartości, gwiazdek, linii łączących grupy) bezpośrednio na wykresach. Jest szczególnie przydatny przy prezentacji wyników testów statystycznych na wykresach typu boxplot, barplot, violin plot itp.

Na rysunku 4.18 w sposób czytelny zaprezentowano różnice w zużyciu paliwa pomiędzy samochodami 4-, 6- i 8-cylindrowymi oraz informację, które z tych różnic, i na jakim poziomie, są statystycznie istotne.

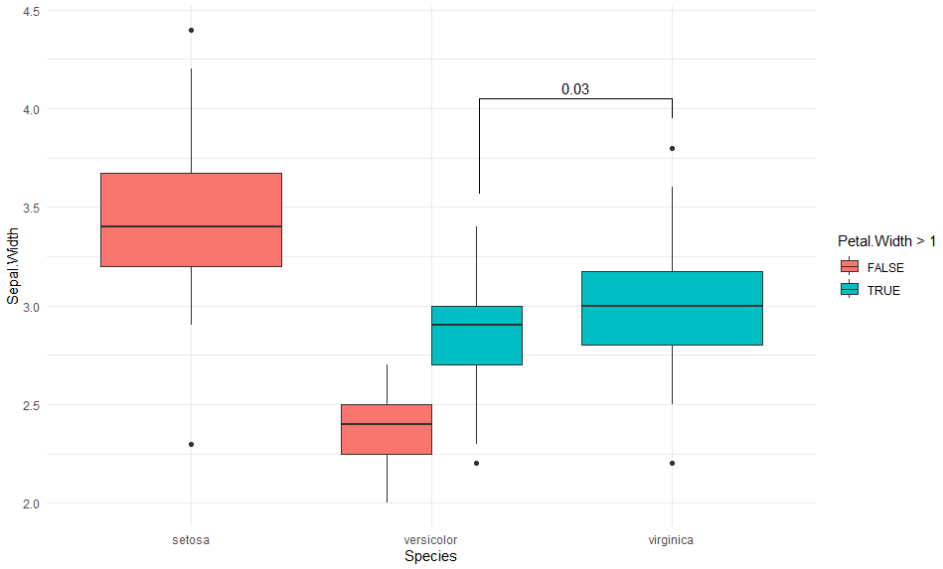
```
ggplot(mtcars, aes(factor(cyl), mpg)) +
  geom_boxplot() +
  geom_signif(
    comparisons = list(c("4", "6"), c("6", "8"),c("4", "8")),
    map_signif_level = TRUE, textsize = 6) +
    ylim(10, 40)+theme_minimal()
```



Rys. 4.18. Wykresy pudełkowe z wynikami testu z ggsignif. Liczba mil przejechanych na galonie paliwa według liczby cylindrów

Rysunek 4.19 łączy wizualną prezentację rozkładów (boxplot) z oceną statystyczną jednej konkretnej pary gatunków, uwzględniając dodatkowo informację o „szerokości płatka" za pomocą koloru.

```
test <- t.test(
  iris[iris$Petal.Width > 1 & iris$Species == "versicolor", "Sepal.Width"],
  iris[iris$Species == "virginica", "Sepal.Width"])$p.value

ggplot(iris, aes(x = Species, y = Sepal.Width, fill = Petal.Width > 1)) +
  geom_boxplot(position = "dodge") +
  geom_signif(
    annotation = formatC(test, digits = 1),
    y_position = 4.05, xmin = 2.2, xmax = 3,
    tip_length = c(0.2, 0.04))+theme_minimal()
```

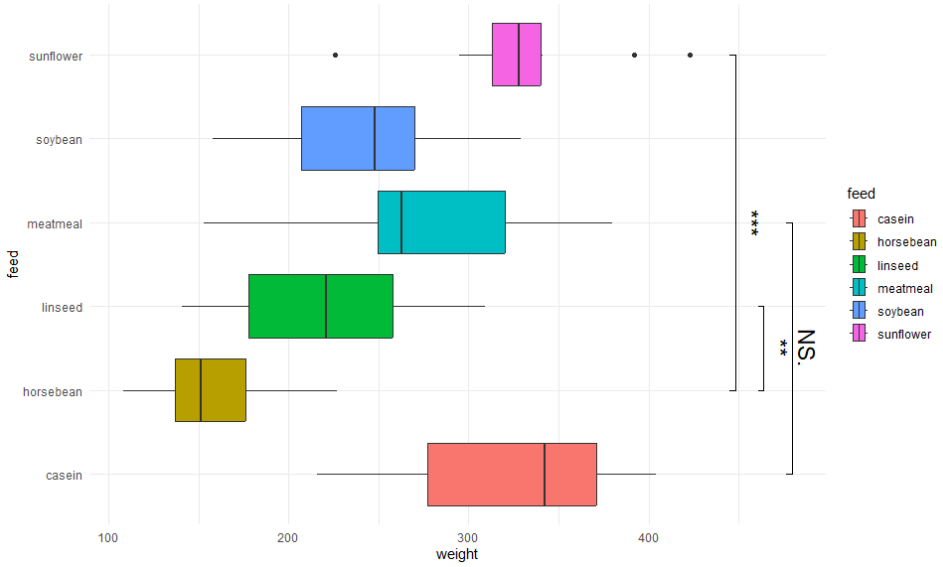


Rys. 4.19. Wizualizacja wyników testu z ggsignif. Szerokość działki kielicha (Sepal.Width) w zależności od gatunku i szerokości płatka

Na rysunku 4.19 przedstawiono porównanie szerokości działki kielicha dla trzech gatunków z wyróżnieniem szerokości płatka.

Kolejny kod i rys. 4.20 przedstawiają porównanie masy piskląt w zależności od rodzaju paszy. Na wykresie wskazano istotne różnice dla wybranych par paszy.

```
ggplot(data = chickwts,mapping = aes(x = weight,y = feed,fill=feed)) +
  geom_boxplot(orientation = "y") +
  geom_signif(
    comparisons = list(c("horsebean", "sunflower"), c("horsebean", "linseed"), c("casein",
"meatmeal")),
    map_signif_level = TRUE,
    textsize = 6,
    margin_top = 0.08,
    step_increase = 0.05,
    tip_length = 0.01,
    orientation = "y")+theme_minimal()
```



Rys. 4.20. Wizualizacja wyników testu z ggsignif. Masa piskląt w zależności od rodzaju paszy – porównanie wybranych par

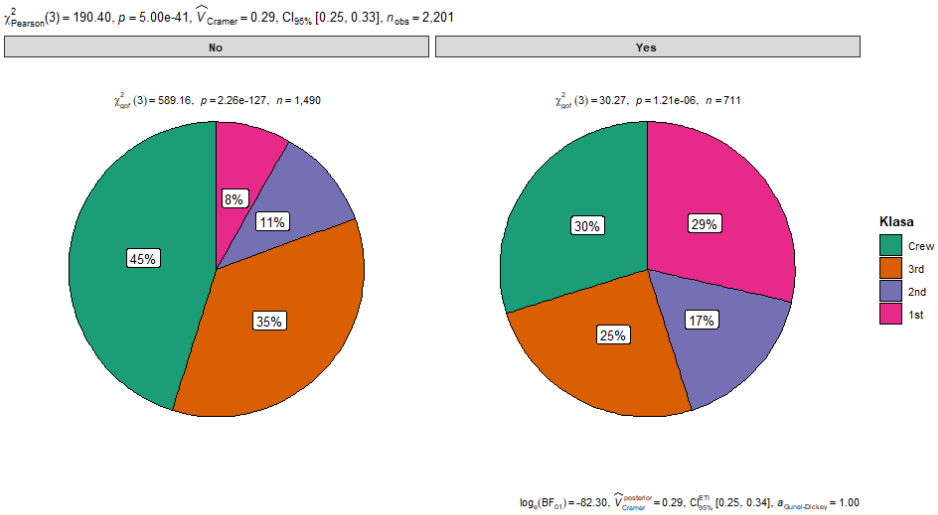
Wykres na rys. 4.20 potwierdza, że rodzaj paszy wpływa znacząco na masę kurcząt: casein i sunflower prowadzą do najwyższej wagi, a horsebean do wyraźnie najniższej. Dla wybranych par rodzajów paszy wskazano istotne różnice.

4.2.7. Pakiet ggstatsplot

Pakiet **`ggstatsplot`** to narzędzie do wizualizacji danych w programie **R**, które łączy prezentację graficzną z przedstawieniem wyników testów statystycznych. Pakiet automatycznie dobiera i wykonuje odpowiednie testy statystyczne (np. t-test, ANOVA, testy nieparametryczne, testy dla danych jakościowych) w zależności od typu danych i liczby grup.

Poniższy kod poddaje analizie przeżywalność katastrofy statku Titanic.

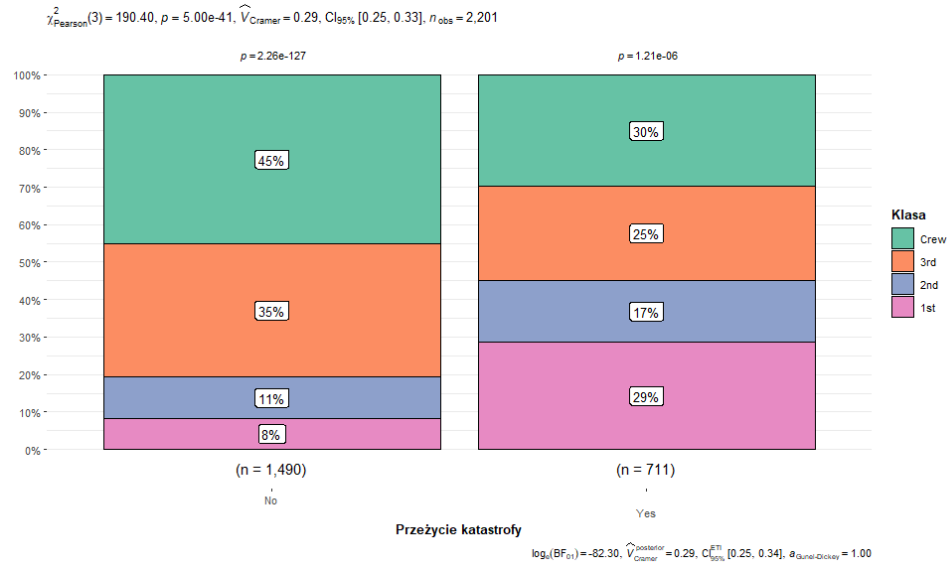
```
ggpiestats(
  data        = titanic,
  x           = Class,
  y           = Survived,
  legend.title = "Klasa")
```



Rys. 4.21. Wizualizacja wyników testu z ggstatsplot. Przeżycie katastrofy statku Titanic – wykres kołowy

Rysunek 4.21 przedstawia udział pasażerów, którzy przeżyli i nie przeżyli katastrofy Titanica, w podziale na klasy podróży (1st, 2nd, 3rd, Crew). Oba wykresy kołowe pokazują proporcje pasażerów według klas. Pasażerowie pierwszej klasy mieli najwyższy odsetek przeżycia – większość osób z tej grupy przeżyła katastrofę. Pasażerowie trzeciej klasy mieli najniższy odsetek przeżycia – większość osób z tej grupy nie przeżyła. Wyniki testu statystycznego (np. test chi-kwadrat) prezentowane na wykresie wskazują, że różnice w przeżywalności pomiędzy klasami są statystycznie istotne (p-wartość < 0,05).

```
ggbarstats(
  data             = titanic,
  x                = Class,
  y                = Survived,
  xlab             = "Przeżycie katastrofy",
  legend.title     = "Klasa",
  ggplot.component = list(ggplot2::scale_x_discrete(guide = ggplot2::guide_axis(n.dodge
= 2))),
  palette          = "Set2")
```

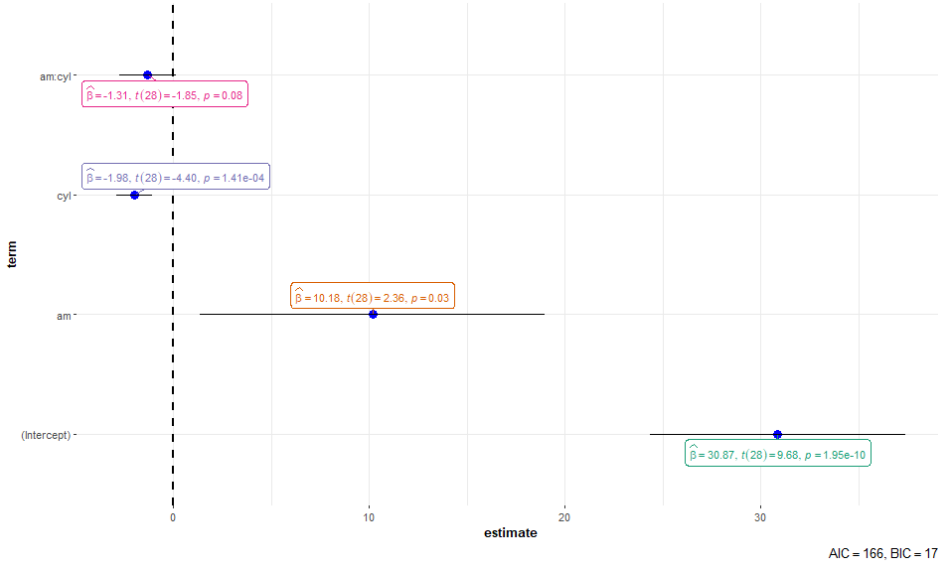


Rys. 4.22. Wizualizacja wyników testu z ggstatsplot. Przeżycie katastrofy statku Titanic – wykres słupkowy

Wykres słupkowy (rys. 4.22) przedstawia liczbę pasażerów, którzy przeżyli i nie przeżyli katastrofy Titanica, w podziale na klasy podróży (1st, 2nd, 3rd, Crew). Każdy słupek reprezentuje liczbę osób w danej klasie, a kolory odpowiadają poszczególnym klasom według legendy. Pasażerowie pierwszej klasy mieli najwyższy odsetek przeżycia – w tej grupie liczba osób, które przeżyły, jest zbliżona do liczby tych, które nie przeżyły, a nawet może być wyższa. Pasażerowie trzeciej klasy mieli najniższy odsetek przeżycia – w tej grupie zdecydowanie przeważają osoby, które nie przeżyły katastrofy. Na wykresie widoczne są także wyniki testu chi-kwadrat niezależności, które potwierdzają, że różnice w przeżywalności pomiędzy klasami są istotne statystycznie (p-wartość < 0,05).

Kolejne dwa wykresy dotyczą wizualizacji wyników testowania dla danych z wcześniej wspomnianego zbioru **mtcars**.

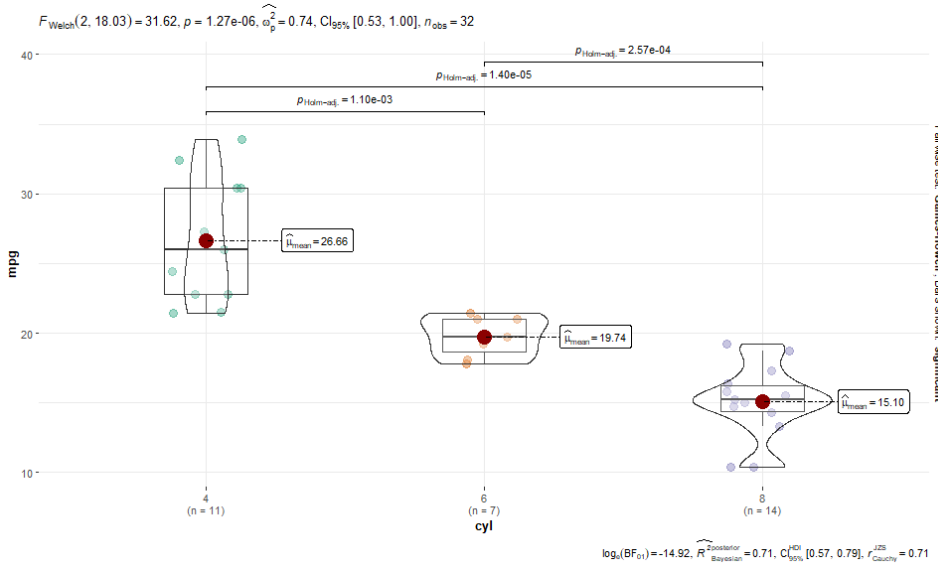
```
model <- stats::lm(formula = mpg ~ am * cyl, data = mtcars)
ggcoefstats(model)
```



Rys. 4.23. Współczynniki regresji dla modelu mpg w zależności od typu skrzyni i liczby cylindrów

Wykres na rys. 4.23 pokazuje punkty reprezentujące oszacowane wartości współczynników regresji. Te współczynniki określają ilościowo wpływ każdej zmiennej predykcyjnej na zmienną zależną (*mpg*). Paski błędu wokół punktów reprezentują przedziały ufności (standardowo 95%) dla współczynników. Te przedziały wskazują zakres, w którym z dużą pewnością znajduje się rzeczywista wartość współczynnika regresji. Wykres wykorzystuje kolory do wskazania istotności statystycznej każdego współczynnika. Współczynniki z *p-wartościami* poniżej wybranego poziomu istotności (np. $\alpha = 0{,}05$) są uważane za statystycznie istotne, co oznacza, że jest mało prawdopodobne, aby obserwowany efekt wystąpił przypadkowo.

```
ggbetweenstats(data  = mtcars,x = cyl, y = mpg )
```



Rys. 4.24. Wizualizacja wyników testu z ggstatsplot. Liczba cylindrów i liczba mil przejechanych na galonie paliwa

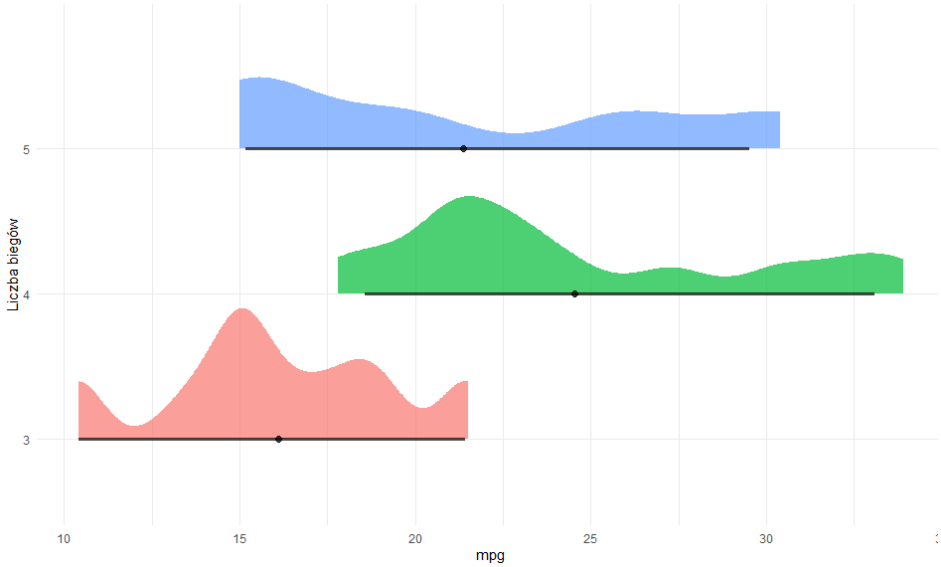
Rysunek 4.24 przedstawia porównanie średniego zużycia paliwa (*mpg*) w samochodach z różną liczbą cylindrów (4, 6 i 8) w zbiorze danych **mtcars**. Każda grupa na osi OX odpowiada innej liczbie cylindrów, a na osi OY są przedstawione wartości *mpg* dla poszczególnych samochodów. Samochody z 4 cylindrami mają najwyższe średnie wartości *mpg* (czyli są najbardziej ekonomiczne). Samochody z 8 cylindrami mają najniższe wartości *mpg*, co oznacza, że zużywają najwięcej paliwa. Na wykresie są prezentowane wyniki testu ANOVA i porównania parami. Różnice pomiędzy grupami są istotne statystycznie (p-wartości < 0,05).

4.2.8. Pakiet ggdist

Pakiet **`ggdist`** umożliwia konstrukcję elastycznych i bogatych w informacje wizualizacji, które pokazują rozkłady danych oraz niepewność estymacji. Poza prezentacją gęstości umożliwia pokazanie np. średnich i przedziałów ufności.

```
mtcars$gear <- as.factor(mtcars$gear)
ggplot(mtcars, aes(x = mpg, y = gear, fill = gear)) +
  stat_halfeye(
    adjust = 0.4,
    .width = c(0.90),
    point_interval = mean_qi,
    alpha = 0.7) +
```

```
  labs(
    x = "mpg",
    y = "Liczba biegów") +
  theme_minimal() +
  theme(legend.position = "none")
```

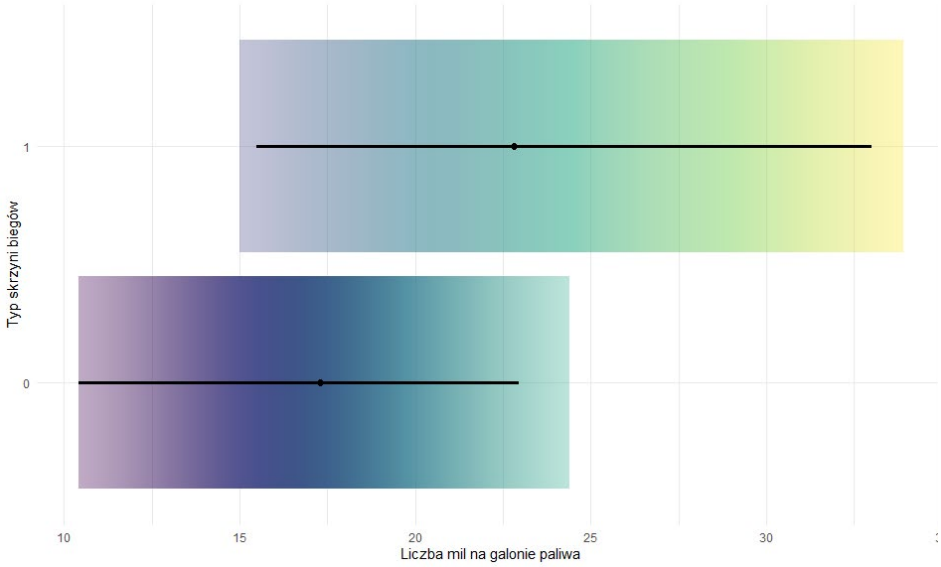


Rys. 4.25. Wizualizacja gęstości i przedziałów ufności dla zmiennej mpg według liczby biegów

Na rysunku 4.25 pokazano jednocześnie oszacowanie gęstości, średnią oraz przedziały ufności zmiennej *mpg* dla samochodów z różną liczbą biegów (*gear*).

Poniższy kod pozwala porównać rozkład *mpg* dla samochodów z automatyczną i ręczną skrzynią biegów. Używa gradientu kolorów do pokazania gęstości prawdopodobieństwa, co jest alternatywą dla klasycznego wykresu skrzypcowego.

```
ggplot(mtcars, aes(x = mpg, y = factor(am), fill = after_stat(x))) +
  stat_gradientinterval(
    show.legend = FALSE ,
    .width = c(0.90)) +
  scale_fill_viridis_c() +
  labs(
    x = "Liczba mil na galonie paliwa",
    y = "Typ skrzyni biegów") +
  theme_minimal()
```



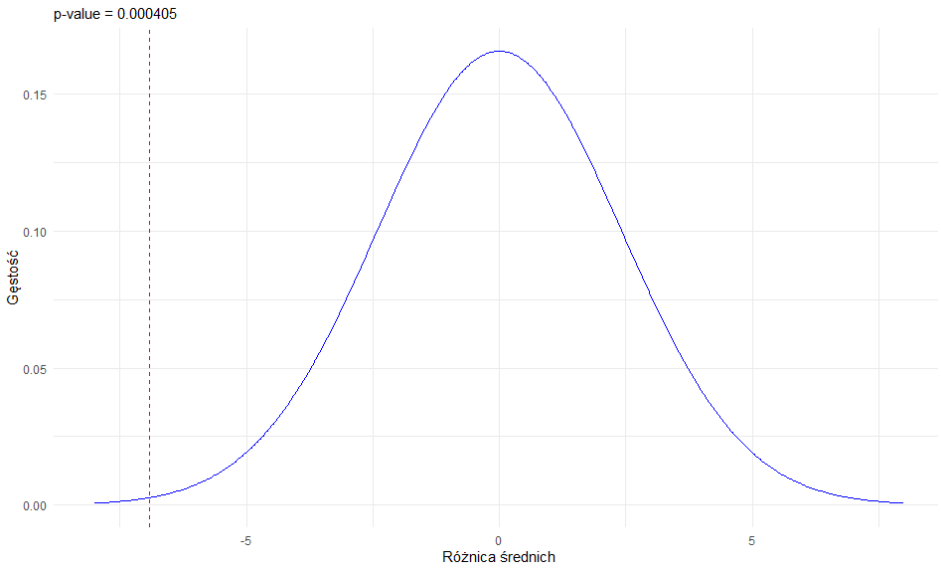
Rys. 4.26. Rozkład mpg według rodzaju skrzyni biegów

Na podstawie rys. 4.26 można zauważyć, że samochody z ręczną skrzynią biegów mają generalnie przejechaną większą liczbę mil na galonie paliwa (*mpg*). Dla obu wariantów skrzyni biegów na wykresie zostały zaznaczone średnia wartość *mpg* i przedział ufności (poziom ufności $1 - \alpha = 0{,}90$).

4.2.9. Pakiet infer

Pakiet **`infer`** w **R** został skonstruowany z myślą o prostym i intuicyjnym przeprowadzaniu wnioskowania statystycznego, szczególnie za pomocą metod symulacyjnych, takich jak testy permutacyjne czy bootstrap. Dzięki zastosowaniu składni zgodnej z **`tidyverse`** pakiet **`infer`** pozwala użytkownikom na łatwe definiowanie hipotez, generowanie rozkładów statystyk testowych oraz wyznaczenie p-wartości i przedziałów ufności. Pakiet ten ułatwia stosowanie nowoczesnych nieklasycznych metod statystycznych, oferując czytelne i elastyczne podejście do analizy danych. Pakiet ten umożliwia także przeprowadzenie klasycznego testowania i wizualizację wyników. Przykład takiego testu, porównującego *mpg* dla samochodów z 4 i 6 cylindrami, przedstawia poniższy kod, natomiast wizualizację wyników zamieszczono na rys. 4.27.

```r
df <- mtcars %>%
  filter(cyl %in% c(4, 6)) %>%
  mutate(cyl = factor(cyl))
obs_stat <- df %>%
  specify(mpg ~ cyl) %>%
  calculate(stat = "diff in means", order = c("6", "4"))
t_test <- t.test(mpg ~ cyl, data = df)

n1 <- sum(df$cyl == 6)
n2 <- sum(df$cyl == 4)
s2 <- var(df$mpg)  # lwspólna wariancja z obu grup
se <- sqrt(s2 * (1/n1 + 1/n2))
null_dist <- tibble(
  stat = seq(obs_stat$stat - 4*se, obs_stat$stat + 8*se, length.out = 1000)) %>%
  mutate(density = dnorm(stat, mean = 0, sd = se))

ggplot(null_dist, aes(x = stat, y = density)) +
  geom_line(color = "blue") +
  xlim(-8,8)+
  geom_vline(xintercept = obs_stat$stat, color = "red", linetype = "dashed") +
  labs(subtitle = paste0("p-value = ", signif(t_test$p.value, 3)),x = "Różnica średnich",
y = "Gęstość") +
  theme_minimal()
```



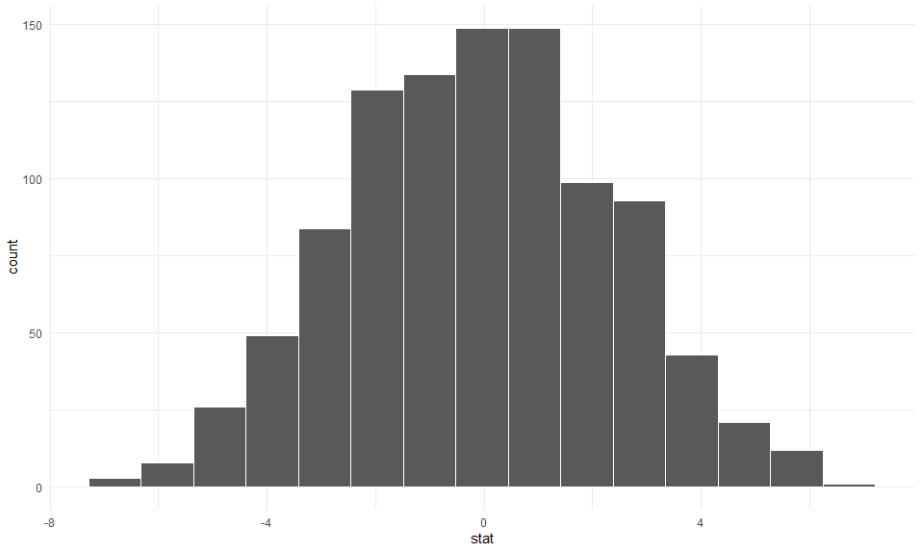
Rys. 4.27. Wizualizacja testu t – różnica mpg dla samochodów o liczbie cylindrów 6 i 4

Rezultat realizacji powyższego kodu przedstawiono na rys. 4.27. Na wykresie zaprezentowano gęstość rozkładu różnicy średnich przy założeniu słuszności H_0. Zaznaczono także wartość statystyki testowej i odpowiadającą jej p-wartość.

W omawianym przypadku należy odrzucić hipotezę H_0, a więc można twierdzić, że przeciętna liczba mil przejechanych na galonie paliwa dla samochodów z 4 i 6 cylindrami nie jest jednakowa.

Z wykorzystaniem pakietu **infer** można uzyskać symulacyjne przybliżenie rozkładu teoretycznego przy założeniu hipotezy H_0 (rys. 4.28).

```
null_dist <- df %>%
  specify(mpg ~ cyl) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("6", "4"))

visualize(null_dist) + labs(title = NULL)+theme_minimal()
```



Rys. 4.28. Symulacyjne przybliżenie rozkładu teoretycznego przy założeniu hipotezy H_0

Możliwości pakietu **infer** zostaną wykorzystane w kolejnym rozdziale do weryfikacji i wizualizacji testów permutacyjnych.

4.3. Wizualizacja wyników wnioskowania – implementacja użytkownika

W poprzednim podrozdziale przedstawiono zastosowanie różnych bibliotek programu **R** do wizualizacji wyników wnioskowania statystycznego. Praktyka badań naukowych prowadzi często do konieczności samodzielnego opracowania

procedur wnioskowania i niekiedy też odpowiednich ich wizualizacji. Przykład takiego rozwiązania w zastosowaniu do testu pustych cel zostanie przedstawiony w tym podrozdziale.

4.3.1. Test pustych cel – idea, założenia i hipotezy

Test pustych cel (ang. *empty cells test*) jest jednym z testów zgodności. Za pomocą tego testu można weryfikować hipotezę dotyczącą postaci rozkładu zmiennej losowej. Test pustych cel to statystyczna metoda weryfikacji losowości wywodząca się z klasycznego problemu rozmieszczeń (ang. *occupancy problem*). W problemie tym badane jest, czy n elementów zostało rozmieszczonych w m celach w sposób całkowicie losowy. Ideę testu oraz własności przedstawiają m.in. David (1950); Hellwig (1965) oraz Domański (2011; 2012). Test pustych cel pozwala na bardzo różnorodne zastosowania. Modyfikacje testu dla przypadków jedno- i dwuwymiarowego przedstawił Kończak (2008; 2009). Z kolei zastosowanie testu pustych cel dla weryfikacji wielowymiarowej normalności przedstawił Kończak (2025).

Cały obszar zmienności zmiennej losowej dzieli się na m rozłącznych cel i dla n-elementowej próby zlicza się liczbę elementów w każdej celi. Statystyką testową jest liczba cel, które po zakończeniu procesu pozostają puste (K_n). Liczba ta jest porównywana z wartością krytyczną. Domański i Pruska (2000) przedstawili interpolowane wartości krytyczne dla statystyki testu pustych cel. Weryfikacji jest poddawana hipoteza dotycząca postaci rozkładu. Kluczowe założenia dotyczące testu pustych cel można zapisać następująco:

1. Dane: liczba elementów n oraz stała liczba cel m.
2. Niezależność: każdy element trafia do celu niezależnie od pozostałych elementów.
3. Jednakowe prawdopodobieństwo: prawdopodobieństwo trafienia każdego elementu do dowolnej z m cel jest jednakowe i wynosi: $p = \frac{1}{m}$.
4. Brak ograniczeń pojemności: każda cela może pomieścić dowolną liczbę elementów.

Niech będzie pobrana próba n-elementowa z populacji o dystrybuancie $F(x)$.

Weryfikowana jest hipoteza o zgodności rozkładu z pewnym ustalonym rozkładem teoretycznym $F_0(x)$:

$$H_0: F(x) = F_0(x)$$
$$H_1: F(x) \neq F_0(x)$$

gdzie F_0 jest daną dystrybuantą.

Obszar zmienności zmiennej losowej X jest dzielony na m rozłącznych cel $M_i (i = 1,2, \dots, m)$, takich, że:

$$\mathbb{X} = \bigcup_{i=1}^{m} M_i \tag{4.1}$$

gdzie $\mathbb{X}$ jest obszarem zmiennosci zmiennej losowej X oraz:

$$M_i \cap M_j = \emptyset \text{ dla } i, j \in \{1,2, \dots, m\}, i \neq j$$
$$P(x \in M_i) = \frac{1}{m} \text{ dla } i \in \{1,2, \dots, m\} \tag{4.2}$$

jeśli hipoteza H_0 jest prawdziwa.

Statystyka testowa (liczba pustych cel) ma następująca postać:

$$K = card\{j: m_j = 0\} \tag{4.3}$$

Dla przyjętego poziomu istotności α obszar krytyczny jest następujący:

$$\mathbf{K} = \{k: k \geq K_{n,\alpha}\} \tag{4.4}$$

gdzie $K_{n,\alpha}$ są wartościami krytycznymi uzyskanymi z tablic (Domański i Pruska, 2000) lub uzyskanymi symulacyjnie.

4.3.2. Test pustych cel – implementacja testu normalności

Test pustych cel to ogólna idea. Na tej podstawie mogą być konstruowane specyficzne testy statystyczne. Kończak (2005) przedstawił modyfikację testu na przypadek wyznaczania nie tylko pustych cel, ale także cel jednoelementowych. Z kolei Kończak (2025) przedstawia konstrukcję wielowymiarowego testu normalności z zadanymi parametrami.

Poniżej przedstawiono konstrukcję testu pustych cel dla weryfikacji hipotezy o normalności rozkładu. Przed zastosowaniem testu zostaną określone dwie funkcje. Funkcja `ect_Kn()` służy do obliczania wartości statystyki testowej (K_n), czyli liczby pustych cel.

```
ect_Kn <- function(data, mean_val, sd_val, num_bins) {
  p_vals <- pnorm(data, mean = mean_val, sd = sd_val)
  p_vals <- pmax(1e-10, pmin(1 - 1e-10, p_vals))
  bins <- ceiling(p_vals * num_bins)
  return(num_bins - length(unique(bins)))
}
```

Funkcja `ect_norm()` realizuje procedurę testu zgodności z rozkładem normalnym. Dla zadanego poziomu istotności α w oparciu o symulacje komputerowe wyznacza wartość krytyczną ($K_{n,\alpha}$) oraz p-wartość. Na podstawie R losowych

prób z rozkładu normalnego konstruowany jest rozkład odniesienia statystyki przy nieznanych parametrach. Wynikiem jest decyzja statystyczna oparta na porównaniu z symulacjami oraz wizualizacja graficzna wyniku na tle histogramu.

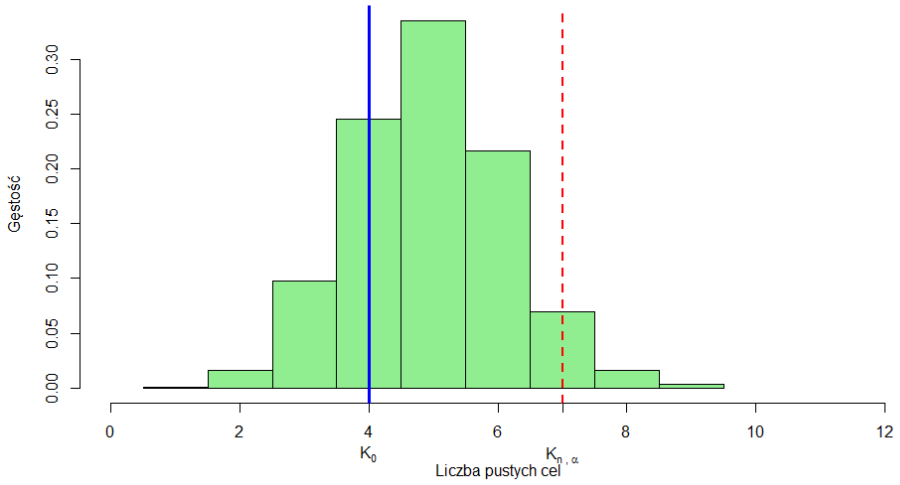
```
ect_norm <- function(x, alpha=0.05, R = 2000, plot = TRUE) {
 m <- n <- length(x)
 mu_obs <- mean(x, na.rm = TRUE)
 sd_obs <- sd(x, na.rm = TRUE)
 K0 <- ect_Kn(x, mu_obs, sd_obs, m)
 K_sim <- numeric(R)
 for (i in 1:R) {
  x_sim <- rnorm(n, mean = mu_obs, sd = sd_obs)
  sim_mu <- mean(x_sim)
  sim_sd <- sd(x_sim)
  K_sim[i] <- ect_Kn(x_sim, sim_mu, sim_sd, m)
 }
 p_value <- (sum(K_sim >= K0) + 1) / (R + 1)
 if (plot) {
  K_crit <- quantile(K_sim, 1 - alpha)
  hist_res <- hist(K_sim, breaks = seq(min(K_sim)-0.5, max(K_sim)+0.5, by=1), plot =
FALSE)
  col_hist <- if (p_value < 0.05) "mistyrose" else "lightgreen"
  plot(hist_res, col = col_hist,  freq = FALSE, xlim = c(0,12),
       main = "", xlab = "Liczba pustych cel", ylab="Gęstość")
  abline(v = K_crit, col = "red", lwd = 2, lty = 2)
  abline(v = K0, col = "blue", lwd = 3)
  axis(side = 1, at = K_crit, labels = expression(K[n ~ "," ~ alpha]), font = 2, padj = 1.5)
  axis(side = 1, at = K0, labels = expression(K[0]), font = 2, padj = 1.5)
 }
 return(list(n_empty = K0, p.value = p_value, estimated_params = list(mean=mu_obs,
sd=sd_obs)))
}
```

Poniżej przedstawiono zastosowanie wprowadzonego powyżej testu pustych cel dla weryfikacji hipotezy o normalności rozkładu analizowanej zmiennej. Przeprowadzono weryfikację hipotezy dla dwóch prób o liczebnościach $n_1 = n_2 = 15$. Pierwsza z tych prób została wygenerowana z rozkładu normalnego, a druga z rozkładu wykładniczego. Tylko w pierwszym z tych przypadków należy oczekiwać, że hipoteza H_0 nie zostanie odrzucona. Na rysunkach 4.29 i 4.30 przedstawiono graficznie wyniki testowania hipotezy o normalności rozkładu dla tych dwóch rozkładów.

Na wykresach zaznaczono liczbę pustych cel (K_0) (kolor niebieski) oraz wyznaczoną symulacyjnie wartość krytyczną $(K_{n,\alpha})$ (czerwona linia przerywana). Dodatkowo kolorem słupków rozróżniono przypadki braku podstaw do odrzucenia hipotezy H_0 (kolor jasnozielony) i odrzucenia tej hipotezy (kolor jasnoczerwony).

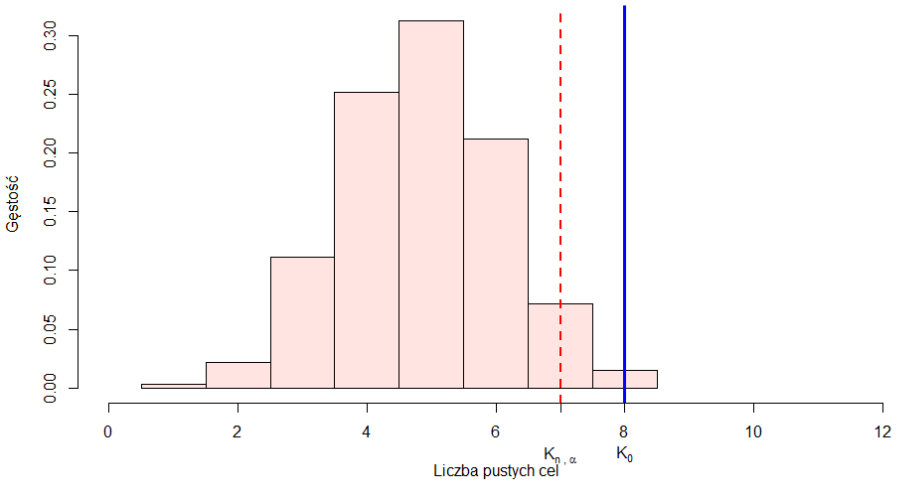
Na rysunkach 4.29 i 4.30 przedstawiono najważniejsze wyniki przeprowadzonego testu, a dodatkowo decyzja została zakodowana w kolorze słupków wykresu, co może być pomocne zwłaszcza w praktycznych zastosowaniach biznesowych.

```
dane1 <- rnorm(15, mean = 10, sd = 2)
dane2 <- rexp(15,rate=2)

ect_norm(dane1)$p.value
ect_norm(dane2)$p.value
# [1] 0.8850575
```



Rys. 4.29. Wizualizacja implementacji testu pustych cel dla weryfikacji hipotezy o normalności rozkładu – dane generowane z rozkładu normalnego

```
# [1] 0.01549225
```



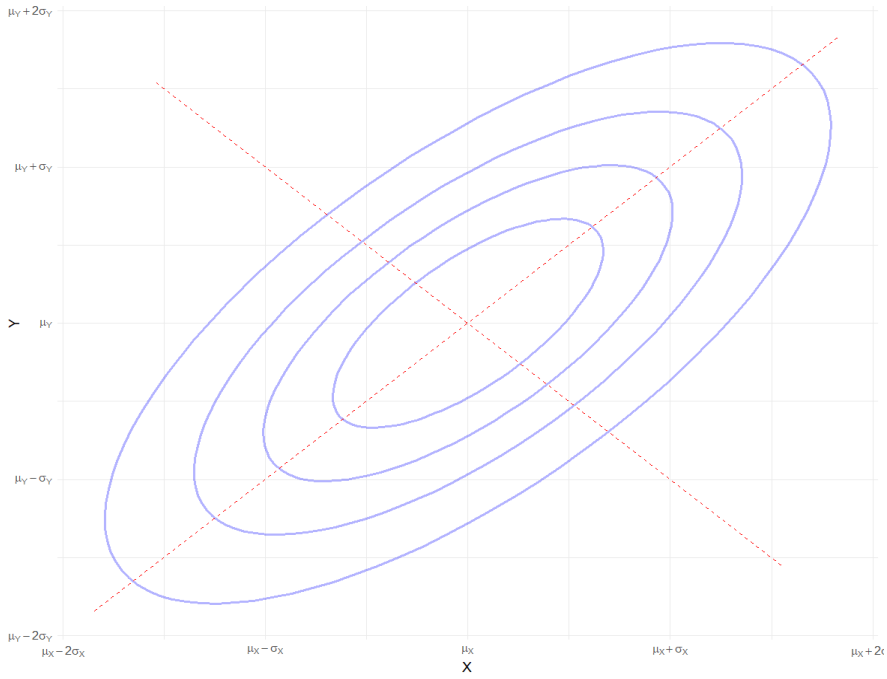
Rys. 4.30. Wizualizacja implementacji testu pustych cel dla weryfikacji hipotezy o normalności rozkładu – dane generowane z rozkładu wykładniczego

4.3.3. Dwuwymiarowy test normalności – wizualizacja wyników

Kończak (2025) przedstawił propozycję wielomiarowego testu normalności opartego na idei pustych cel. Weryfikowana jest hipoteza o p-wymiarowym rozkładzie normalnym z zadanymi parametrami: wektorem wartości oczekiwanych $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ i macierzą kowariancji $\boldsymbol{\Sigma}$. Poniżej zostanie przytoczona idea testu dla rozkładu dwuwymiarowego.

Konstrukcja testu opiera się na partycjonowaniu przestrzeni dwuwymiarowej na rozłączne obszary (cele) o jednakowym prawdopodobieństwie teoretycznym. Podział ten przebiega dwuetapowo:

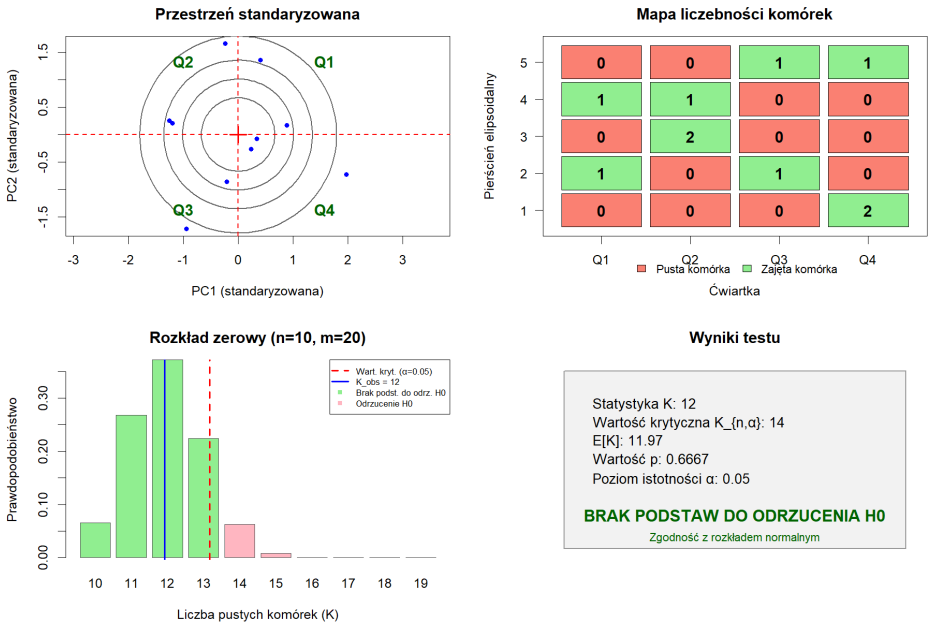
1. Przestrzeń jest dzielona na k koncentrycznych pierścieni eliptycznych, dobranych tak, aby każdy z nich gromadził równą masę prawdopodobieństwa $\left(\frac{1}{k}\right)$ zgodnie z teoretycznym rozkładem normalnym.
2. Każdy pierścień jest następnie dzielony na cztery sektory przez główne osie elipsoidy wyznaczone przez wektory własne macierzy kowariancji. W rezultacie otrzymuje się $m = 4k$ cel, do których, przy założeniu hipotezy zerowej, obserwacje trafiają z jednakowym prawdopodobieństwem. Idea konstrukcji cel została przedstawiona na rys. 4.31.



Rys. 4.31. Idea konstrukcji cel dla przypadku dwuwymiarowej normalności

Źródło: Opracowanie własne na podstawie: Kończak (2025).

W ogólnym przypadku p-wymiarowym każdy z pasów eliptycznych jest dzielony na 2^p obszarów. Wyniki testu normalności zawierają w szczególności następujące informacje: wartość statystyki testowej (K), wartość krytyczną $\left(K_{n,\alpha}\right)$, p-wartość oraz decyzję. Wszystkie te wyniki warto przedstawić w formie wizualizacji jak np. na rys. 4.32.



Rys. 4.32. Wizualizacja wyników dwuwymiarowego testu normalności

Źródło: Opracowanie własne.

Rysunek 4.32 przedstawia graficzną ilustrację wyniku testu wielowymiarowej normalności opartego na liczbie pustych cel. Grafika pokazuje relację między wartością obliczoną z danych K a rozkładem teoretycznym (symulacyjnym) przy założeniu prawdziwości hipotezy zerowej. Całość składa się z czterech paneli tworzących kompleksową wizualizację wyników testu dwuwymiarowej normalności. W kolejnych panelach przedstawiono następujące informacje.

Panel 1:

Wykres przedstawia obserwacje po transformacji do przestrzeni standaryzowanej, w której elipsoidy stają się okręgami. Wykres umożliwia ocenę, czy obserwacje rozkładają się równomiernie w celach. Występowanie skupień jest argumentem przeciw hipotezie H_0.

Panel 2:

Wykres typu mapa ciepła pokazuje liczbę obserwacji w każdej celi w formie macierzy. Wykres przedstawia macierz, gdzie wiersze odpowiadają kolejnym elipsoidom, a kolumny ćwiartkom $(Q1 - Q4)$. Każdy prostokąt reprezentuje jedną celę. Wewnątrz każdego pola jest podana liczba obserwacji, które wpadły do danej celi. Cele puste zostały wyróżnione kolorem czerwonym.

Panel 3:

Wykres słupkowy przedstawia rozkład empiryczny statystyki testowej (liczby pustych cel K) przy założeniu, że hipoteza zerowa jest prawdziwa. Został on wyznaczony symulacyjnie. Kolor histogramu jest wizualnym wskaźnikiem decyzji testowej. Kolor jasnozielony oznacza brak podstaw do odrzucenia hipotezy zerowej. Różowy oznacza odrzucenie hipotezy zerowej. Czerwona linia przerywana wskazuje interpolowaną wartość krytyczną $(K_{n,\alpha})$ dla ustalonego poziomu istotności α. Jest to granica obszaru krytycznego. W tym teście obszar krytyczny znajduje się na prawo od tej linii. Jeśli liczba pustych cel przekroczy tę linię, oznacza to, że należy odrzucić hipotezę H_0. Niebieska linia ciągła wskazuje wartość statystyki testowej (K) obliczoną dla badanej próby.

Panel 4:

Jest to panel tekstowy, który przedstawia kluczowe wyniki testu w jednym miejscu.
Statystyka K: Liczba pustych cel znaleziona w danych.
Wartość krytyczna $K_{n,\alpha}$: Próg liczby pustych cel, powyżej którego przy ustalonym poziomie istotności α należy odrzucić hipotezę H_0.
$E[K]$: Oczekiwana liczba pustych cel.
p-wartość: Prawdopodobieństwo uzyskania takiego lub gorszego wyniku przy założeniu H_0.
Decyzja: Komunikat (“ODRZUĆ H0” czerwoną czcionką lub “BRAK PODSTAW DO ODRZUCENIA H0” zieloną czcionką).

5. Model permutacyjny Fishera-Pitmana – alternatywa dla klasycznego wnioskowania w modelu populacyjnym Neymana-Pearsona

> *[…] hipoteza mogłaby zostać obalona*
> *przez jeden jedyny przypadek niepowodzenia,*
> *lecz nigdy nie mogłaby zostać potwierdzona*
> *przez żadną skończoną liczbę eksperymentów.*
>
> Ronald A. Fisher (1935, s. 16)

Testy permutacyjne są stosunkowo rzadko stosowane w badaniach ekonomicznych. Trudno zrozumieć taką sytuację, ponieważ mają one wiele zalet w stosunku do klasycznych testów parametrycznych i nieparametrycznych. Testy parametryczne wymagają spełnienia założeń dotyczących postaci analizowanego rozkładu – najczęściej jest to rozkład normalny. W praktyce badań ekonomicznych często trudno o spełnienie tego wymogu. Eden i Yates (1933) analizowali zastosowanie testu *z* Fishera na rzeczywistych danych o rozkładzie skośnym, wskazując na ograniczenia tego testu. Miało to istotne znaczenie dla praktyki eksperymentów rolniczych i statystyki stosowanej. Jednak rozkłady skośne występują nie tylko w danych przyrodniczych, ale także często w różnych dyscyplinach naukowych, w szczególności w zagadnieniach dotyczących ekonomii i finansów. Klasyczne testy nieparametryczne charakteryzują się zwykle mniejszą mocą, a w związku z tym do skutecznego wnioskowania są wymagane próby o większej liczebności. Testy permutacyjne, które także są zaliczane do metod nieparametrycznych, opierają się na modelu Fishera-Pitmana, który znacznie różni się od modelu Neymana-Pearsona wykorzystywanego w klasycznym wnioskowaniu statystycznym. Testy permutacyjne są wolne od większości założeń towarzyszących zastosowaniom testów parametrycznych.

Testy permutacyjne znajdują szerokie zastosowanie w analizie statystycznej, szczególnie w sytuacjach, gdy nie są spełnione założenia klasycznych metod parametrycznych lub gdy badacz dysponuje próbami o małych liczebnościach. Metody te pozwalają na precyzyjną weryfikację istotności różnic w strukturach populacji bez konieczności odwoływania się do rozkładów asymptotycznych, co czyni je kluczowym narzędziem przy ograniczonych zbiorach danych (Kończak i Kosińska, 2023). Szczególnie przydatne są w analizach wyników eksperymen-

tów losowych (Złotoś, 2019). Są one efektywnie wykorzystywane w analizie danych jakościowych, umożliwiając testowanie złożonych hipotez kierunkowych (Kończak, 2012; Polko i Kończak, 2016; Bonnini i Borghesi, 2025) oraz przeprowadzanie testów dla danych w wielowymiarowych tablicach wielodzielczych (Kończak i Chmielińska, 2013; Kończak i Kosińska, 2025). Ponadto podejście to sprawdza się w badaniu istnienia i siły współzależności między zmiennymi, oferując odporną alternatywę dla tradycyjnych testów asocjacji (Kończak, 2020a). Metody permutacyjne mogą być także z powodzeniem stosowane w analizie danych pochodzących z eksperymentów ekonomicznych (Złotoś, 2025).

W badaniach ekonomicznych i społecznych metody te są cenione za możliwość kompleksowego porównywania populacji wielowymiarowych (Polko-Zając, 2021) oraz weryfikację jednorodności danych, co jest istotne dla zapewnienia jakości wnioskowania (Polko i Kończak, 2016). Testy permutacyjne wykazują dużą użyteczność przy porównywaniu parametrów położenia w grupach o niejednorodnych wariancjach (Mielke i Berry, 1994), a także pozwalają na testowanie równoważności (*equivalence*) z wykorzystaniem procedur typu intersection-union (Arboretti i in., 2018). Ich elastyczność jest widoczna w praktycznych zastosowaniach socjoekonomicznych, takich jak ocena wpływu pandemii na aktywność w wydarzeniach kulturalnych, gdzie mogą służyć do analizy zmian w strukturach odpowiedzi respondentów (Kosińska, 2023).

W rozdziale przedstawiono kluczowe różnice pomiędzy modelami wnioskowania populacyjnym (Neymana-Pearsona) i permutacyjnym (Fishera-Pitmana). Szczególną uwagę zwrócono na bardzo rzadko stosowany, zwłaszcza w naukach ekonomicznych, model permutacyjny. Przedstawiono zasady wnioskowania w tym modelu, wybrane pakiety środowiska **R** umożliwiające weryfikację hipotez oraz metody wizualizacji uzyskanych wyników.

5.1. Dwa modele wnioskowania statystycznego

Model populacyjny Neymana-Pearsona

Model wnioskowania statystycznego został zaproponowany przez Jerzego Neymana i Egona Pearsona (1933). Wnioskowanie w tym modelu opiera się na lemacie Neymana-Pearsona, który stanowi matematyczną podstawę konstrukcji testów statystycznych dla prostych hipotez. Lemat ten wskazuje, że dla zadanego poziomu istotności α (poziomu błędu I rodzaju) test oparty na ilorazie wiarygodności (likelihood ratio) jest testem o największej mocy spośród wszystkich testów tego samego poziomu istotności.

W modelu populacyjnym określa się hipotezy H_0 (hipoteza zerowa) i H_1 (hipoteza alternatywna). W tym modelu przyjmuje się, że próba została pobrana losowo z określonej populacji o zadanym rozkładzie. Rozróżnia się dwa typy błędów – I rodzaju (odrzucenie hipotezy H_0, gdy jest ona prawdziwa) i II rodzaju (brak decyzji o odrzuceniu hipotezy H_0, gdy jest ona fałszywa).

> ⓘ Model permutacyjny Fishera-Pitmana
>
> Model wnioskowania statystycznego został wprowadzony przez Ronalda Aylmera Fishera (1925) i sformalizowany przez Edwina Jamesa George Pitmana (1937a, 1937b). Nieco wcześniej model permutacyjny, który jest podstawą wnioskowania permutacyjnego, został zastosowany do analizy wyników doświadczeń polowych przez Jerzego Spławę-Neymana (1923). Model opiera się na wykorzystaniu permutacji etykiet obserwacji. Nie wymaga założenia dotyczącego postaci (typu) rozkładu, może być także stosowany dla prób nielosowych.

Model permutacyjny wnioskowania statystycznego charakteryzuje się wieloma zaletami w stosunku do klasycznego modelu populacyjnego Neymana-Pearsona. Jest bardziej elastyczny i może być stosowany w sytuacjach, w których klasyczne modele parametryczne są nieodpowiednie. Wnioskowanie w modelu permutacyjnym opiera się wyłącznie na pobranych danych (Berry i in., 2014; Kończak, 2016), nie zakłada się postaci rozkładu badanej zmiennej. Pozwala w szczególności na przeprowadzenie wnioskowania statystycznego nawet w sytuacjach, gdy (Brombin i Salmaso, 2013; Berry i in., 2014; Kończak, 2020b):

- nie jest znana postać rozkładu analizowanych zmiennych,
- pobrano próby o niewielkich liczebnościach,
- próby nie są losowe.

Choć losowość próby nie jest wymagana do przeprowadzenia testów permutacyjnych, to jednak jej brak uniemożliwia uogólnienie wyników na całą populację. Uzyskane wyniki mogą być odniesione wyłącznie do analizowanych zbiorów danych. Ogólne zasady wnioskowania w modelu permutacyjnym przedstawiają m.in. Good (2006b), Good (2013), Bonnini Corain i in. (2014), Kończak (2016) i Berry i in. (2021).

W naukach ekonomicznych bardzo często trudno utrzymać założenie o normalności badanych zmiennych. Przykłady zastosowań testów permutacyjnych w naukach ekonomicznych przedstawiają m.in. Moore (2011), Złotoś (2020) i Kosińska (2025). Zastosowania testów permutacyjnych dla danych wielowymiarowych i złożonych przedstawiają m.in. Pesarin (2001), Pesarin i Salmaso (2010)

oraz Polko-Zając (2020). W dalszej części tego rozdziału zostaną omówione zasady stosowania testów permutacyjnych i wskazane odpowiednie narzędzia.

5.2. Wybrane pakiety środowiska R wspomagające wnioskowanie permutacyjne

W środowisku **R** jest wiele pakietów pozwalających na przeprowadzenie wnioskowania statystycznego w modelu permutacyjnym. Dla przeprowadzenia testów permutacyjnych niezbędne jest załadowanie odpowiednich pakietów. Realizuje to poniższy kod.

```
library(wPerm)
library(coin)
library(perm)
library(resample)
library(exactRankTests)
library(mosaic)
library(infer)
library(ggplot2)
library(dplyr)
library(tidyverse)
```

W poniższych analizach uwaga zostanie skoncentrowana na funkcjach pakietu **`wPerm`**. Pozostałe pakiety zostaną w sposób zwięzły przedstawione w dalszej części.

5.2.1. Pakiet wPerm

Pakiet **`wPerm`** (Weiss, 2015) w środowisku **R** pozwala na przeprowadzanie ważonych testów permutacyjnych. Testy te są szczególnie przydatne w analizie danych pochodzących z badań społecznych, w których często stosuje się złożone schematy doboru próby, a każda jednostka otrzymuje określoną wagę odzwierciedlającą jej reprezentatywność w populacji. Uwzględnienie wag w procedurze testowania pozwala na uzyskanie wyników o większej zgodności z rzeczywistą strukturą populacji, co pozwala na zmniejszenie błędów wnioskowania statystycznego. Pakiet **`wPerm`** zapewnia także standardową wizualizację wyników testowania permutacyjnego.

Test dla dwóch prób niezależnych

Test dla wartości oczekiwanych oraz median w dwóch populacjach zostanie przedstawiony na przykładzie danych generowanych z dwóch rozkładów chi-kwadrat o różnych wartościach oczekiwanych. Hipotezy można zapisać następująco:

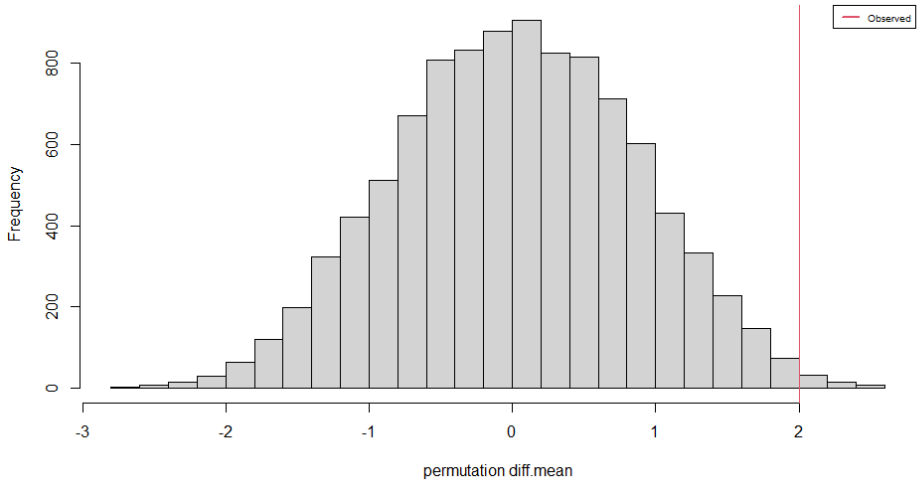
- dla wartości oczekiwanych:

$$H_0: \mu_X = \mu_Y$$
$$H_1: \mu_X \neq \mu_Y$$

- dla median:

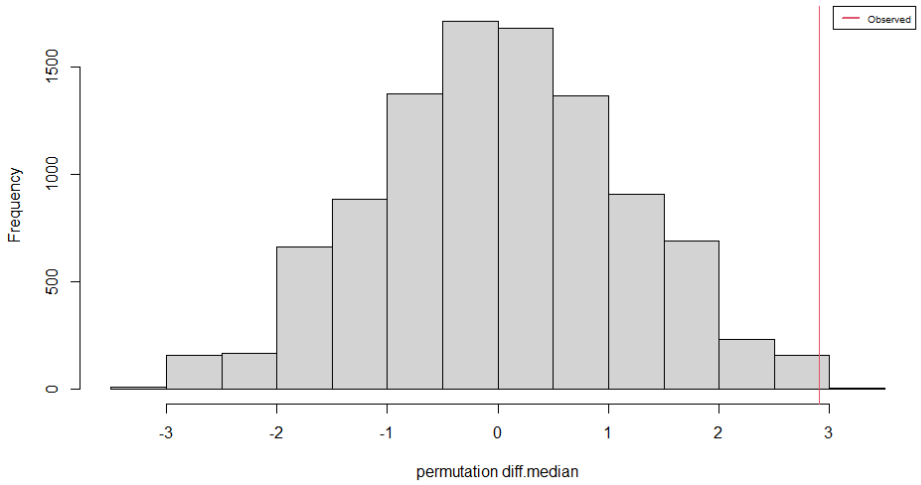
$$H_0: Me_X = Me_Y$$
$$H_1: Me_X \neq Me_Y$$

```
set.seed(123)
dane <- data.frame(
  x = c(rchisq(10, df=1), rchisq(10, df = 5)),
  grupa = rep(c("kontrola", "eksperyment"), each = 10))
perm.ind.loc(x=dane$x, y=dane$grupa, parameter=mean)
perm.ind.loc(x=dane$x, y=dane$grupa, parameter=median)
#
#
#  RESULTS OF PERMUTATION INDEPENDENT TWO-SAMPLE LOCATION TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY Variable       Pop.1    Pop.2 n.1 n.2 Statistic
Observed
#  STATISTICS        x eksperyment kontrola  10  10 diff.mean
2.005029
#
#  HYPOTHESIS      Null Alternative P.value
#        TEST identical     shifted  0.0102
```



Rys. 5.1. Wyniki testu permutacyjnego dla równości wartości oczekiwanych dla prób niezależnych. Histogram rozkładu permutacyjnego

```
#
#
#  RESULTS OF PERMUTATION INDEPENDENT TWO-SAMPLE LOCATION TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY Variable       Pop.1    Pop.2 n.1 n.2   Statistic
Observed
#  STATISTICS        x eksperyment kontrola  10  10 diff.median
2.910802
#
#  HYPOTHESIS      Null Alternative P.value
#        TEST identical     shifted   0.004
```



Rys. 5.2. Wyniki testu permutacyjnego dla równości median dla prób niezależnych. Histogram rozkładu permutacyjnego

W powyższym kodzie wykorzystano funkcję *perm.ind.loc()* z pakietu **wPerm**, która realizuje test permutacyjny, sprawdzając istotność różnic między grupami odpowiednio dla średniej (mean) i mediany (median). Wyniki tych testów pozwalają ocenić, czy zaobserwowane różnice pomiędzy grupami są statystycznie istotne. Nie jest wymagane założenie normalności badanych zmiennych, co sprawia, że jest to skuteczne narzędzie w analizie różnic międzygrupowych. Dla uzyskania wyłącznie p-wartości można wykonać komendy:

```
perm.ind.loc(dane$x, dane$grupa, mean)$p.value
# [1] 0.0098
perm.ind.loc(dane$x, dane$grupa, median)$p.value
# [1] 0.0022
```

Na rysunkach 5.1 i 5.2 przedstawiono wizualizację wyników testu permutacyjnego dla porównania odpowiednio: wartości przeciętnych i median. Na wykresach zostały zaznaczone wartości statystyki testowej. Wyniki zamieszczone pod wykresami w obu przypadkach prowadzą do odrzucenia hipotezy H_0.

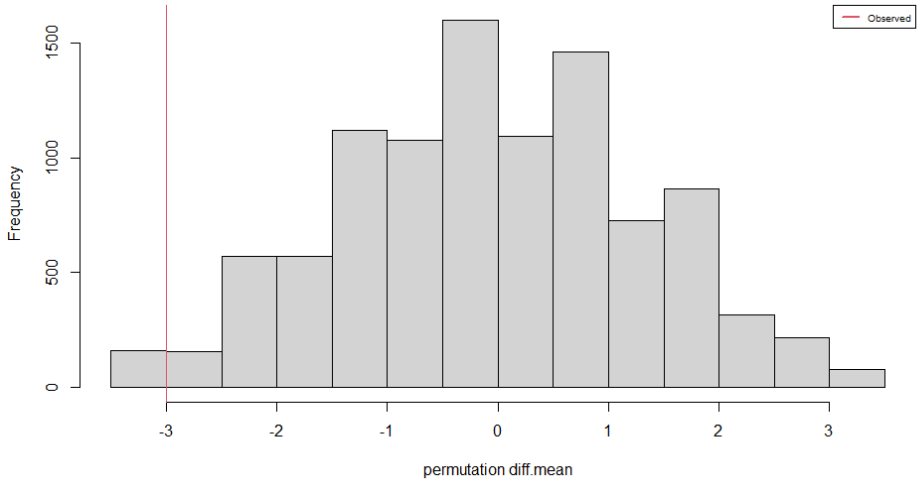
Niekiedy jest przeprowadzane porównanie charakterystyk (np. wartości oczekiwanych) w dwóch grupach dla prób sparowanych. Jest tak np. wówczas, gdy dla tej samej jednostki obserwacji są dokonywane dwa pomiary.

Porównano wyniki egzaminacyjne 10 studentów (x – liczba punktów przed kursem, y – liczba punktów po kursie), a ich rezultaty przedstawiono w poniższym kodzie. Siedmiu poprawiło swoje wyniki, dwóch uzyskało identyczne wyniki, a jeden nieznacznie obniżył wynik. Celem jest sprawdzenie, czy przeprowadzenie kursu poprawiło przeciętne wyniki studentów.

Hipotezy przyjmują następującą postać:

$$H_0: \mu_X = \mu_Y$$
$$H_1: \mu_X \neq \mu_Y$$

```
x <- c(65, 70, 72, 60, 75, 80, 68, 73, 62, 78)
y <- c(70, 70, 80, 60, 73, 85, 75, 74, 66, 80)

perm.paired.loc(x=x, y=y, parameter=mean, alternative='less')
#
#
#   RESULTS OF PERMUTATION PAIRED LOCATION TEST
#   BASED ON 9999 REPLICATIONS
#
#      SUMMARY Pop.1 Pop.2  n Statistic Observed
#   STATISTICS     x     y 10 diff.mean       -3
#
#   HYPOTHESIS      Null  Alternative P.value
#         TEST identical shifted.left  0.0159
```



Rys. 5.3. Wyniki testu permutacyjnego dla prób zależnych. Histogram rozkładu permutacyjnego

Powyższy kod prezentuje zastosowanie testu permutacyjnego dla prób zależnych (sparowanych), gdzie porównuje się dwie serie pomiarów uzyskane dla tych samych jednostek, np. w dwóch okresach czasowych. Funkcja *perm.paired.loc()* umożliwia ocenę, czy średnia różnica pomiędzy parami obserwacji jest istotnie mniejsza od zera (alternative='less'). W omawianym przypadku test potwierdził, że przeciętne rezultaty po zakończeniu kursu są lepsze niż przed kursem (p-wartość < α). Wizualizacja wyników testu została przedstawiona na rys. 5.3.

Test permutacyjny dla prób zależnych jest szczególnie zalecany, gdy nie można założyć normalności rozkładu różnic lub gdy liczba obserwacji jest niewielka, a klasyczny test t dla prób zależnych mógłby prowadzić do błędnych wniosków.

Test niezależności chi-kwadrat stosuje się do weryfikacji hipotezy o braku zależności między dwiema zmiennymi jakościowymi. W poniższym kodzie wykorzystano funkcję `perm.ind.test()` realizującą permutacyjną wersję testu niezależności chi-kwadrat. Zastosowanie permutacyjnego testu niezależności nie wymaga spełnienia założenia o minimalnych licznościach oczekiwanych w komórkach minimum 5, jak to jest w klasycznym teście niezależności chi-kwadrat.

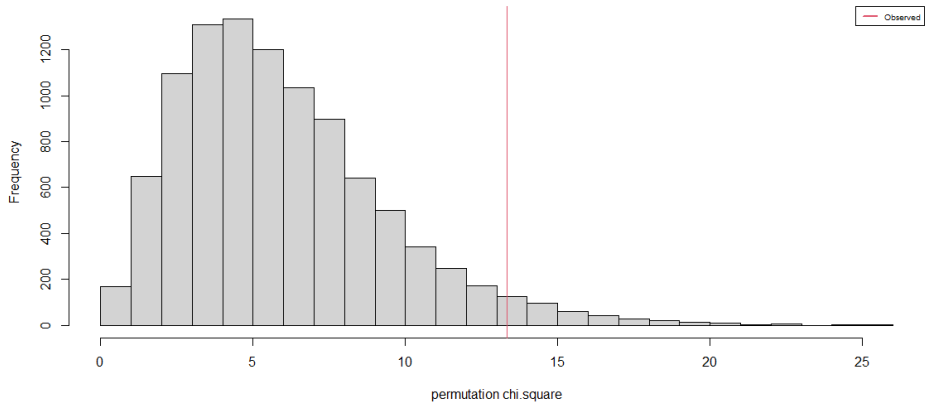
Zastosowanie testu zostanie przedstawione na przykładzie zbioru danych **relig.and.ed** z pakietu `wPerm`. Celem jest sprawdzenie, czy istnieje statystycznie istotny związek pomiędzy dwiema zmiennymi jakościowymi: poziomem religijności (*RELIGIOUSITY*) oraz poziomem wykształcenia (*EDUCATION*).

H_0: Poziom religijności i poziom wykształcenia są niezależne w badanej populacji.

Hipoteza ta oznacza, że rozkład religijności jest taki sam niezależnie od poziomu wykształcenia, i odwrotnie.

H_1: istnieje związek (zależność) między poziomem religijności a poziomem wykształcenia.

```
data("relig.and.ed")
tabela_relig_ed <- xtabs(COUNT ~ RELIGIOUSITY + EDUCATION, data = relig.and.ed)
tabela_relig_ed
#                EDUCATION
# RELIGIOUSITY    Basic Secondary Advanced
#   Religious        77       149       78
#   Not religious    23        56       36
#   Atheist           8        24       29
#   Don't know        6        15        8
perm.ind.test(x=relig.and.ed, type = "flat", var.names = c("RELIGIOUSITY", "EDUCATION"),
R = 9999)
#
#
#  RESULTS OF PERMUTATION INDEPENDENCE TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY   Variable.1 Variable.2   n  Statistic Observed
#  STATISTICS RELIGIOUSITY  EDUCATION 509 chi.square 13.32231
#
#  HYPOTHESIS          Null Alternative P.value
#        TEST nonassociated  associated  0.0363
```

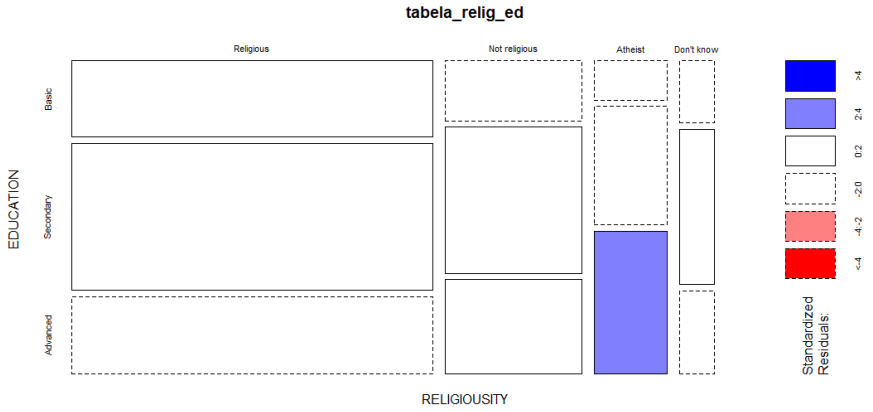


Rys. 5.4. Wizualizacja wyników testu permutacyjnego niezależności. Histogram rozkładu permutacyjnego

Powyższy wynik oznacza, że na podstawie tych danych można stwierdzić (por. rys. 5.4), że istnieje statystycznie istotny związek (asocjacja) między poziomem religijności a poziomem wykształcenia w badanej próbie (p-wartość < α).

Zależności pomiędzy zmiennymi jakościowymi można zwizualizować, odwołując się do wykresów mozaikowych (*mosaic plot*). Poniższy kod konstruuje wykres mozaikowy (rys. 5.5), który wizualizuje dane tabelaryczne dwu- lub wielowymiarowe (Friendly, 1994; Kończak, 2025).

```
plot(tabela_relig_ed , shade=TRUE)
```



Rys. 5.5. Wizualizacja wyników testu permutacyjnego niezależności

Kolor prostokątów na wykresie inny niż biały wskazuje na występowanie statystycznie istotnych zależności pomiędzy zmiennymi. Niebieski prostokąt przedstawia liczbę osób o poziomie edukacji “Advanced” deklarujących się jako

ateiści (*Atheist*). Osób tych jest więcej niż wskazują na to liczebności oczekiwane przy założeniu niezależności badanych zmiennych.

Zbiór **self** zawiera dane, w postaci tablicy kontyngencji, dotyczące samooceny niezależnych losowych próbek widzących i niewidomych nastolatków z Indii. Dla danych ze wskazanego zbioru zostanie przeprowadzony test jednorodności struktur. Zostanie zweryfikowana hipoteza głosząca, że samoocena w wyróżnionych grupach nastolatków ma tę samą strukturę.

Hipotezy mają następującą postać:

H_0: Samoocena w grupach nastolatków ma tę samą strukturę

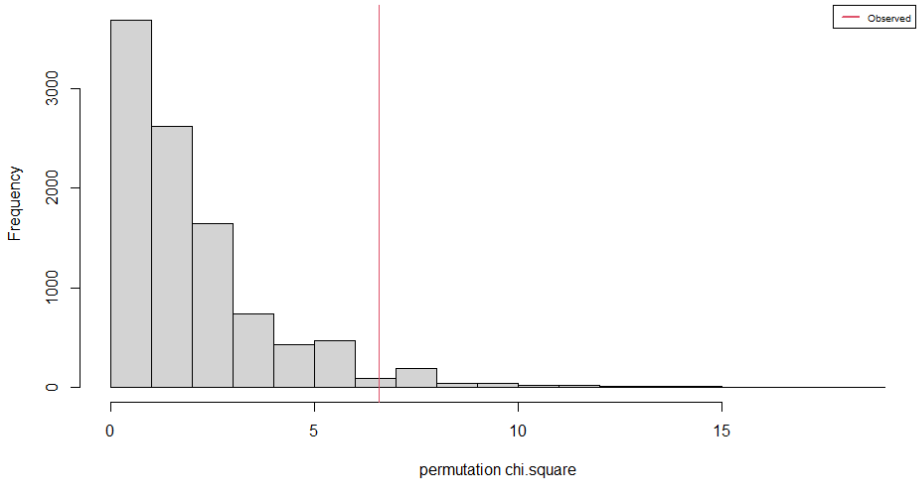
H_1: Struktury samooceny w grupach nie są jednakowe.

Strukturę zbioru **self** można pokazać następująco:

```
data(self)
self
#   SIGHTEDNESS High Moderate Low
# 1     Sighted   13       73  14
# 2       Blind    3       40  17
```

Przeprowadzenie testu jednorodności struktur realizuje funkcja *perm.hom.test()*.

```
perm.hom.test(x=self, type="cont")
#
#
#  RESULTS OF PERMUTATION HOMOGENEITY TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY   n  Statistic Observed
#  STATISTICS 160 chi.square 6.589323
#
#  HYPOTHESIS        Null    Alternative P.value
#       TEST homogeneous nonhomogeneous  0.0364
```



Rys. 5.6. Wizualizacja wyników testu chi-kwadrat jednorodności. Histogram rozkładu permutacyjnego

Wyniki testu wskazują, że przyjmując poziom istotności $\alpha = 0{,}05$, należy odrzucić hipotezę H_0 o jednorodności struktur (por. rys. 5.6), a więc można twierdzić, że struktury samooceny nastolatków w badanych grupach nie są jednakowe.

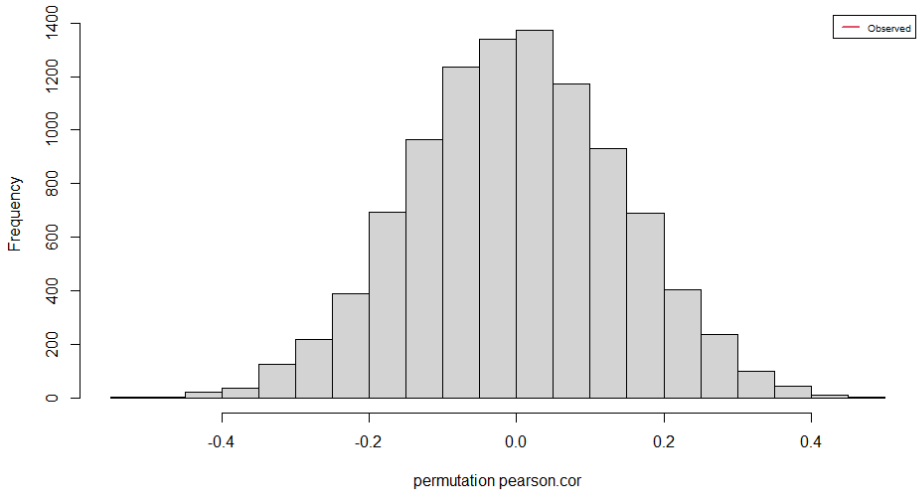
Testowanie współczynnika korelacji polega na weryfikacji, czy obserwowana zależność pomiędzy dwiema zmiennymi liczbowymi jest statystycznie istotna. Klasyczne testy istotności współczynnika korelacji, takie jak test t dla współczynnika Pearsona, zakładają m.in. normalność rozkładu badanych zmiennych i liniowość zależności. Gdy te założenia nie są spełnione lub próbka jest niewielka, alternatywą są testy permutacyjne, które nie wymagają tak restrykcyjnych założeń. Test permutacyjny polega na wielokrotnym losowym permutowaniu jednej z analizowanych zmiennych i wyznaczaniu rozkładu współczynnika korelacji Pearsona, Spearmana lub Kendalla przy założeniu słusznosci hipotezy zerowej o braku związku. Funkcja `perm.relation()` z pakietu **`wPerm`** umożliwia przeprowadzenie takiego testu zarówno dla klasycznego współczynnika Pearsona, jak i nieparametrycznych współczynników rangowych Spearmana i Kendalla, pozwalając na elastyczne dopasowanie testu do charakteru danych i postawionej hipotezy (np. alternatywa jednostronna "greater"). Testy permutacyjne są szczególnie zalecane, gdy dane są obciążone wartościami odstającymi, nie mają rozkładu normalnego lub zależność nie jest liniowa, a także w sytuacjach, gdy liczebność próby jest mała i klasyczne testy mogą zawodzić.

Dla danych ze zbioru **cars** zostanie zweryfikowana hipoteza o niezależności zmiennych *speed* oraz *dist*, wobec hipotezy alternatywnej, że występuje zależność dodatnia. Hipotezy można zapisać następująco:

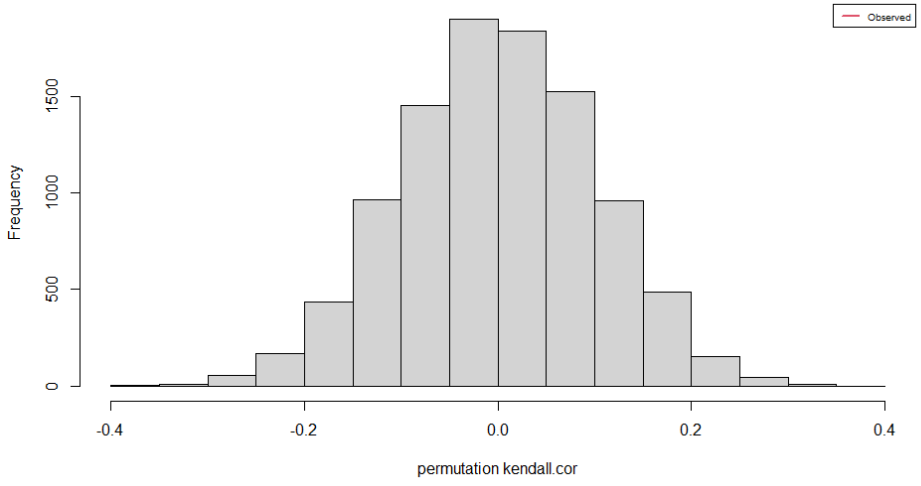
$$H_0: \rho = 0$$
$$H_1: \rho > 0$$

Poniższy kod przedstawia weryfikację hipotezy dla współczynnika korelacji liniowej Pearsona, rang Spearmana oraz rang Kendalla.

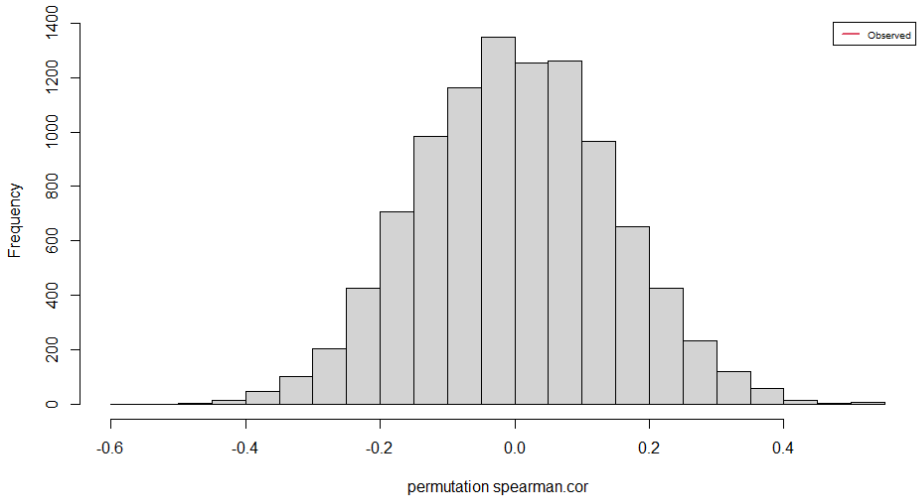
```
attach(cars)
perm.relation(x=speed, y=dist, alternative = "greater")
#
#
#  RESULTS OF PERMUTATION RELATIONSHIP TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY Variable.1 Variable.2  n   Statistic  Observed
#  STATISTICS      speed       dist 50 pearson.cor 0.8068949
#
#  HYPOTHESIS        Null  Alternative   P.value
#       TEST no.relation pos.relation P < 0.001
perm.relation(speed, dist, "kendall", "greater")
#
#
#  RESULTS OF PERMUTATION RELATIONSHIP TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY Variable.1 Variable.2  n   Statistic  Observed
#  STATISTICS      speed       dist 50 kendall.cor 0.6689901
#
#  HYPOTHESIS        Null  Alternative   P.value
#       TEST no.relation pos.relation P < 0.001
perm.relation(speed, dist, "spearman", "greater")
#
#
#  RESULTS OF PERMUTATION RELATIONSHIP TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY Variable.1 Variable.2  n    Statistic  Observed
#  STATISTICS      speed       dist 50 spearman.cor 0.8303568
#
#  HYPOTHESIS        Null  Alternative   P.value
#       TEST no.relation pos.relation P < 0.001
```



Rys. 5.7. Wizualizacja wyników testu dla współczynnika korelacji liniowej Pearsona. Histogram rozkładu permutacyjnego



Rys. 5.8. Wizualizacja wyników testu dla współczynnika korelacji rang Kendalla. Histogram rozkładu permutacyjnego



Rys. 5.9. Wizualizacja wyników testu dla współczynnika korelacji rang Spearmana. Histogram rozkładu permutacyjnego

Wyniki pozwalają ocenić, czy obserwowana korelacja jest istotnie większa od zera, bez konieczności spełniania restrykcyjnych założeń klasycznych testów parametrycznych. Na rysunkach 5.7–5.9 przedstawiono rozkłady empiryczne współczynników korelacji odpowiednio: liniowej Pearsona, rang Spearmana i rang Kendalla. We wszystkich tych przypadkach wartość współczynnika z pierwotnych danych znajduje się poza wyświetlanym zakresem na OX. We wszystkich trzech przypadkach należy odrzucić hipotezę H_0 na rzecz hipotezy alternatywnej.

5.2.2. Pakiet coin

Pakiet **`coin`** (Hothorn i in., 2006) to narzędzie do przeprowadzania nieparametrycznych testów permutacyjnych. Został zaprojektowany w celu zapewnienia ujednoliconego podejścia do szerokiej gamy tych testów dla różnych typów danych.

Funkcja *`oneway_test()`* z pakietu **`coin`** służy do nieparametrycznego testowania różnic między kilkoma niezależnymi grupami (układ jednoczynnikowy) jako alternatywa dla klasycznego jednoczynnikowego ANOVA. Jest to pomocne narzędzie w analizie permutacyjnej, pozwalające na precyzyjne wnioskowanie bez konieczności odwoływania się do teoretycznych rozkładów.

Warunkowy rozkład zerowy statystyki testowej może być przybliżony przez jej rozkład asymptotyczny („asymptotic” – parametr domyślny) lub poprzez ponowne próbkowanie Monte Carlo („approximate”). Dla jednowymiarowych te-

stów dla dwóch prób możliwe jest ustawienie parametru „exact" powodujące wyznaczenie dokładnego rozkładu.

Kolejne dwa przykłady kodu pozwalają na przeprowadzenie testów permutacyjnych dla porównania wartości oczekiwanych na podstawie:

- dwóch prób,
- trzech lub większej liczby prób.

```
oneway_test(mpg~factor(am) , data = mtcars, alternative='less')
#
#   Asymptotic Two-Sample Fisher-Pitman Permutation Test
#
# data:  mpg by factor(am) (0, 1)
# Z = -3.3397, p-value = 0.0004193
# alternative hypothesis: true mu is less than 0
```

Wyniki testu wskazują, że należy odrzucić hipotezę H_0, a więc można twierdzić, że przeciętna liczba mil przejechanych na galonie paliwa (*mpg*) dla samochodów z automatyczną skrzynią biegów jest mniejsza niż dla samochodów z ręczną skrzynią biegów.

Dla porównania danych z trzech lub większej liczby prób można wykorzystać tę samą funkcję, co przedstawiono poniżej.

```
oneway_test(mpg~factor(cyl) , data = mtcars)
#
#   Asymptotic K-Sample Fisher-Pitman Permutation Test
#
# data:  mpg by factor(cyl) (4, 6, 8)
# chi-squared = 22.706, df = 2, p-value = 1.173e-05
```

Wyniki testu wskazują, że należy odrzucić hipotezę H_0, a więc można twierdzić, że przeciętna liczba mil przejechanych na galonie paliwa (*mpg*) w trzech grupach wyróżnionych ze względu na liczbę cylindrów nie jest jednakowa.

5.2.3. Pakiet perm

Pakiet **perm** (Fay i Shaw, 2010) w **R** jest przeznaczony do przeprowadzania dokładnych (exact) lub asymptotycznych testów permutacyjnych. Jego głównym celem jest umożliwienie użytkownikowi wykonywania testów permutacyjnych dla dwóch prób, wielu prób oraz testów trendu, z szerokim wyborem metod wyznaczania *p*-wartości.

Dla porównania dwóch lub większej liczby prób niezależnych należy skorzystać odpowiednio z funkcji `permTS()` oraz `permKS()`. Poniższy kod przedstawia porównanie przeciętnej liczby mil przejechanej na galonie paliwa (*mpg*) w samochodach z automatyczną i ręczną skrzynią biegów (dwie próby) oraz dla 4, 6 i 8 cylindrów (trzy próby).

```
permTS(mpg~am,data=mtcars,alternative="less",method="pclt")
#
#   Permutation Test using Asymptotic Approximation
#
# data:  mpg by am
# Z = -3.3397, p-value = 0.0004193
# alternative hypothesis: true mean am=0 - mean am=1 is less
than 0
# sample estimates:
# mean am=0 - mean am=1
#             -7.244939
permKS(mpg~cyl,data=mtcars,method="pclt")
#
#   K-Sample Asymptotic Permutation Test
#
# data:  mpg by cyl
# Chi Square = 22.706, df = 2, p-value = 1.173e-05
```

W obu przedstawionych przypadkach należy odrzucić hipotezę H_0 na rzecz hipotezy alternatywnej, a więc w wyróżnionych grupach przeciętna liczba mil przejechanych na galonie paliwa nie jest jednakowa.

5.2.4. Pakiet resample

Pakiet **`resample`** (Hesterberg, 2022) jest przeznaczony m.in. do przeprowadzenia testów bootstrapowych i permutacyjnych. Pakiet oferuje łatwą i czytelną składnię. Jest to narzędzie, które ułatwia szacowanie rozkładów próbkowania statystyk i testowanie hipotez bez sztywnych założeń o rozkładzie danych, co jest często wymagane w klasycznych testach parametrycznych.

Dla sprawdzenia, czy przeciętna liczba mil przejechanych na galonie paliwa (*mpg*) dla samochodów z ręczną i automatyczną skrzynią biegów jest jednakowa, wobec hipotezy alternatywnej głoszącej, że liczba mil przejechanych na galonie paliwa samochodami wyposażonymi w automatyczną skrzynię biegów jest mniejsza niż w samochodach wyposażonych w ręczną skrzynię biegów, w pakiecie **`resample`** należy wykonać test:

```
x1=mtcars$mpg[mtcars$am==0]
x2=mtcars$mpg[mtcars$am==1]
permutationTest2(data=x1, data2 = x2, statistic=mean, alternative='less',R=999)
# Call:
# permutationTest2(data = x1, statistic = mean, data2 = x2,
R = 999,
#     alternative = "less")
# Replications: 999
# Two samples, sample sizes are 19 13
#
# Summary Statistics for the difference between samples 1 and 2:
#               Observed       Mean Alternative PValue
# mean: x1-x2 -7.244939 0.02931474        less  0.001
```

Otrzymane wyniki prowadzą do odrzucenia hipotezy H_0, a więc można twierdzić, że przeciętna liczba mil przejechanych na galonie paliwa dla samochodów z automatyczną skrzynią biegów jest mniejsza niż dla samochodów z ręczną skrzynią biegów.

5.2.5. Pakiet exactRankTests

Pakiet **`exactRankTests`** (Hothorn i Hornik, 2022) pozwala na obliczanie dokładnych rozkładów dla testów rangowych i permutacyjnych. Umożliwia uzyskanie precyzyjnych p-wartości, zwłaszcza w sytuacjach, gdy standardowe testy parametryczne nie mogą być zastosowane ze względu na naruszenia założeń (np. braku normalności rozkładu danych). Dostępna w pakiecie funkcja *`perm.test()`* pozwala na weryfikację hipotezy o równości wartości oczekiwanych dla danych niezależnych i sparowanych. W tym podrozdziale przedstawiono dwa takie przykłady.

Poniższy kod prowadzi do weryfikacji hipotezy o równości wartości oczekiwanych liczby mil przejechanych na galonie paliwa w samochodach z ręczną i automatyczną skrzynią biegów.

```
x=mtcars$mpg[mtcars$am==0]
y=mtcars$mpg[mtcars$am==1]
perm.test( y, x, conf.int=TRUE, exact=TRUE)
#
#   2-sample Permutation Test (scores mapped into 1:(m+n) using
rounded
#    scores)
#
# data:  y and x
# T = 246, p-value = 0.0003442
# alternative hypothesis: true mu is not equal to 0
# 95 percent confidence interval:
#   4.80000 14.71429
```

Należy odrzucić hipotezę H_0 głoszącą, że przeciętna liczba mil przejechanych na galonie paliwa jest jednakowa dla samochodów z ręczną i automatyczną skrzynią biegów. Poniższy kod pozwala na weryfikację podobnej hipotezy dla danych sparowanych.

```
x <- c(65, 70, 72, 60, 75, 80, 68, 73, 62, 78)
y <- c(70, 70, 80, 60, 73, 85, 75, 74, 66, 80)
perm.test(x=x, y=y, paired=TRUE, exact=TRUE)
#
#   1-sample Permutation Test
#
# data:  x and y
# T = 2, p-value = 0.03125
# alternative hypothesis: true mu is not equal to 0
```

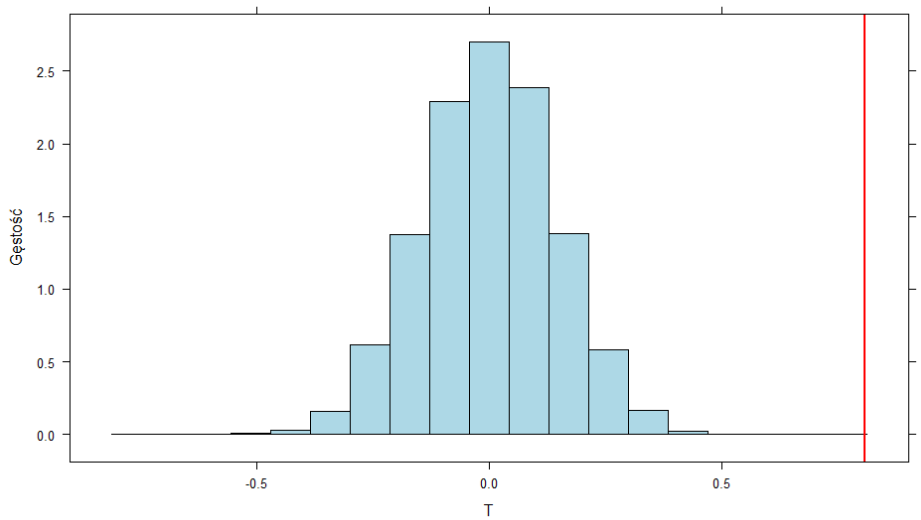
Ponieważ p-wartość $< \alpha$, więc należy odrzucić hipotezę H_0 o równości wartości oczekiwanych, a zatem można twierdzić, że wartości oczekiwane w badanych populacjach nie są jednakowe.

5.2.6. Pakiet mosaic

Pakiet **`mosaic`** to biblioteka języka **R** zaprojektowana do przeprowadzania analizy danych w sposób spójny, prosty i „formułowy" (Pruim i in., 2017). Jej ideą jest ujednolicenie interfejsu wielu operacji statystycznych (opis, wizualizacja danych, wnioskowanie statystyczne, symulacje) w jeden schemat oparty na notacji formułowej $y \sim x$.

Poniższy kod przedstawia jednostronny test permutacyjny dla współczynnika korelacji liniowej Pearsona. Badaniu poddano istotność zależności między prędkością a drogą hamowania na podstawie danych ze zbioru **cars**. Jako statystykę testową T przyjęto współczynnik korelacji liniowej. Na wstępie obliczana jest wartość z rzeczywistych danych (T_0), po czym generowany jest empiryczny rozkład zerowy poprzez 10 000-krotne losowe przetasowanie wartości zmiennej *dist*, co symulowało sytuację braku zależności między zmiennymi. Na tej podstawie wyznaczono p-wartość ASL (*achieved significance level*, por. Efron i Tibshirani, 1993) jako frakcję permutacji dających wynik równy lub wyższy od obserwowanego. Wyniki zilustrowano histogramem rozkładu symulacyjnego, na którym czerwoną linią zaznaczono wartość statystyki obserwowanej, aby wizualnie ocenić jej nietypowość na tle rozkładu otrzymanego przy założeniu hipotezy H_0 (rys. 5.10).

```
library(mosaic)
T0 <- cor(cars$speed, cars$dist)
T <- do(10000) * cor(cars$speed, shuffle(cars$dist))
ASL <- mean(T >= T0)
cat("n =", nrow(cars), "| r =", round(T0, 3), "| p =", ASL, "\n")
# n = 50 | r = 0.807 | p = 0
histogram(~ T,
        xlim=c(-.9,0.9),
        xlab = "T",
        v = T0,
        col = "lightblue" ,
        par.settings = list(add.line = list(col = "red", lwd = 2)))
```



Rys. 5.10. Wartość statystyki testowej T_0 (czerwona linia) i symulacyjny rozkład statystyki T przy założeniu hipotezy H_0

Uzyskany rezultat prowadzi do odrzucenia hipotezy H_0. Można twierdzić, że występuje zależność pomiędzy prędkością a odległością do zatrzymania samochodu.

5.2.7. Pakiet infer

Pakiet **`infer`** (Couch i in., 2021) umożliwia przeprowadzanie testów statystycznych opartych na permutacjach i resamplingu w ramach spójnego interfejsu. Pozwala na testowanie hipotez dotyczących związków między zmiennymi, różnic pomiędzy grupami czy dla wartości parametru populacji, bez konieczności przyjmowania silnych założeń teoretycznych. Pakiet ten szczególnie dobrze sprawdza

się w sytuacjach, gdy użytkownik chce samodzielnie określić sposób losowania przy założeniu słuszności H_0, co daje dużą kontrolę nad całym procesem wnioskowania statystycznego.

Poniżej przedstawiono realizację testu par w oparciu o pakiet **`infer`**. Na przykładzie umownych danych przeprowadzono test permutacyjny dla równości wartości oczekiwanych, a następnie wyznaczono rozkład bootstrapowy różnic i bootstrapowy przedział ufności dla różnicy.

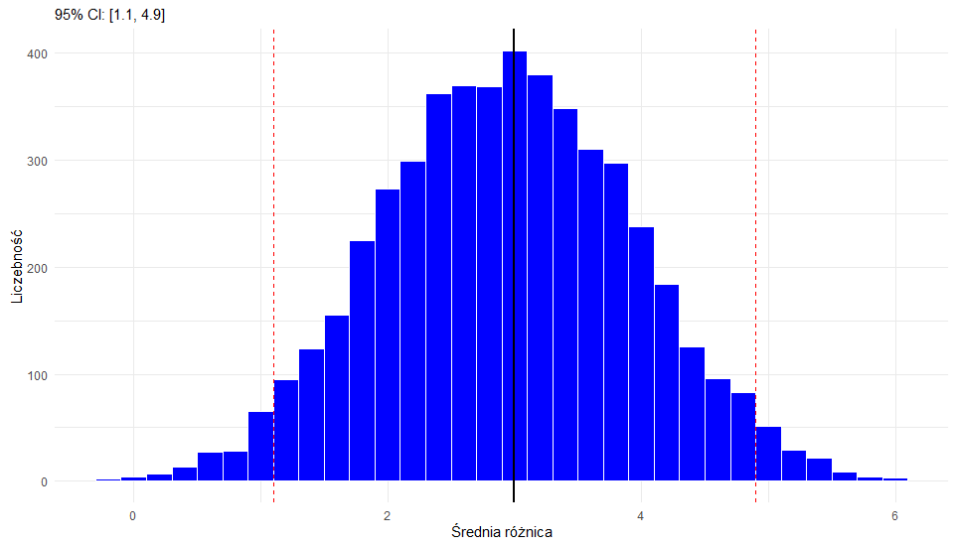
```
library(infer)
library(ggplot2)
x1 <- c(65, 70, 72, 60, 75, 80, 68, 73, 62, 78)
x2 <- c(70, 70, 80, 60, 73, 85, 75, 74, 66, 80)
dane <- data.frame(
  id = 1:length(x1),
  g1 = x1,
  g2 = x2
) %>%
  tidyr::pivot_longer(cols = -id, names_to = "grupa", values_to = "x")

wyniki <- dane %>%
  specify(x ~ grupa) %>%        # Definicja zmiennych
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("g1", "g2"))

T0 <- dane %>%
  specify(x ~ grupa) %>%
  calculate(stat = "diff in means", order = c("g1", "g2"))
ASL <- get_p_value(wyniki, obs_stat = T0, direction = "both")

paste0("Obserwowana różnica średnich:", T0$stat)
# [1] "Obserwowana różnica średnich:-3"
paste0("ASL: ", ASL)
# [1] "ASL: 0.416"
```

Metody permutacje są wykorzystywane głównie do weryfikacji hipotez statystycznych, natomiast metody bootstrapowe do budowy przedziałów ufności dla parametrów populacji i oszacowania wariancji estymatorów.

Pakiet **`infer`** umożliwia także wyznaczanie przedziałów ufności. Do konstrukcji przedziałów ufności jest wykorzystywana metoda bootstrap. Może to być zrealizowane następująco:

```
różnice <- x2 - x1
dane <- data.frame(diff = różnice)
```

```
boot_diffs <- dane %>%
  specify(response = diff) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "mean")

ci <- boot_diffs %>%
  get_confidence_interval(level = 0.95, type = "percentile")

ggplot(boot_diffs, aes(x = stat)) +
  geom_histogram(binwidth = 0.2, fill = "blue", color = "white") +
  geom_vline(xintercept = ci$lower_ci, color = "red", linetype = "dashed") +
  geom_vline(xintercept = ci$upper_ci, color = "red", linetype = "dashed") +
  geom_vline(xintercept = mean(różnice), color = "black", linewidth = 1) +
  labs(title = "",subtitle = paste0("95% CI: [", round(ci$lower_ci, 2), ", ", round(ci$upper
_ci, 2), "]"),x = "Średnia różnica", y = "Liczebność" ) +
  theme_minimal()
```



Rys. 5.11. Rozkład bootstrapowy średniej różnicy dla danych skojarzonych z zaznaczonym 95-procentowym przedziałem ufności

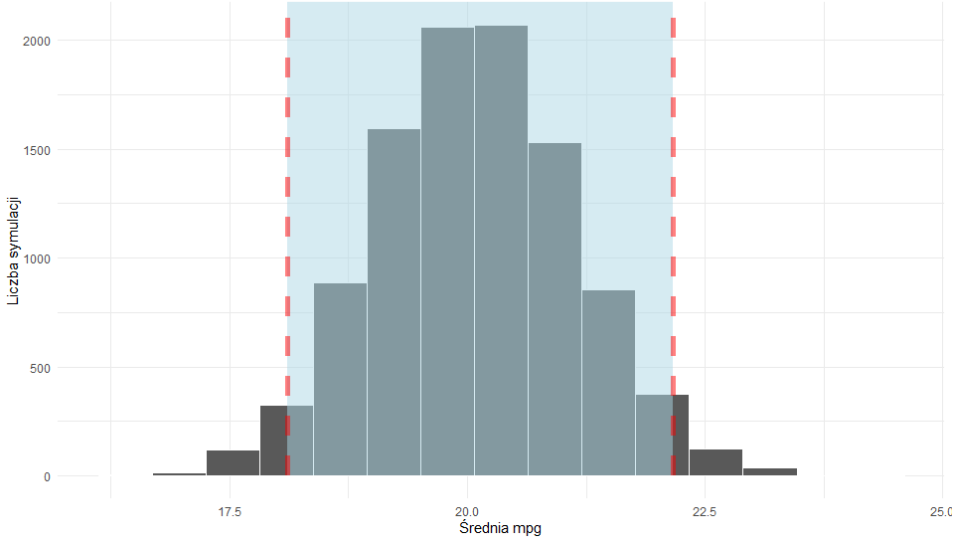
Rysunek 5.11 przedstawia bootstrapowy rozkład średniej różnicy dla danych skojarzonych. Na wykresie zaznaczono także przedział ufności dla różnicy średnich wyznaczony metodą bootstrapową oraz średnią różnicę z prób bootstrapowych.

Poniższy kod prowadzi do wyznaczenia przedziału ufności dla wartości oczekiwanej dla zmiennej *mpg* (zbiór **mtcars**).

```r
boot_dist_mpg <- mtcars %>%
  specify(response = mpg) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

ci_mpg <- boot_dist_mpg %>%
  get_confidence_interval(level = 0.95)
print(ci_mpg)
# # A tibble: 1 × 2
#   lower_ci upper_ci
#      <dbl>    <dbl>
# 1     18.1     22.2
```

Dla zwizualizowania wyniku można wykonać następujący kod:

```r
p <- boot_dist_mpg %>%
  visualize() +
  shade_confidence_interval(endpoints = ci_mpg, fill = "lightblue", color = "red", lty=2,
alpha = 0.5)

p + labs(title = "", x = "Średnia mpg", y = "Liczba symulacji") +
  theme_minimal()
```



Rys. 5.12. Rozkład bootstrapowy średniej mpg z zaznaczonym 95-procentowym przedziałem ufności

```r
mtcars_df <- mtcars %>%
  mutate(am = factor(am, levels = c(0, 1), labels = c("Automatyczna", "Ręczna")))

mpg_diff_obs <- mtcars_df %>%
```

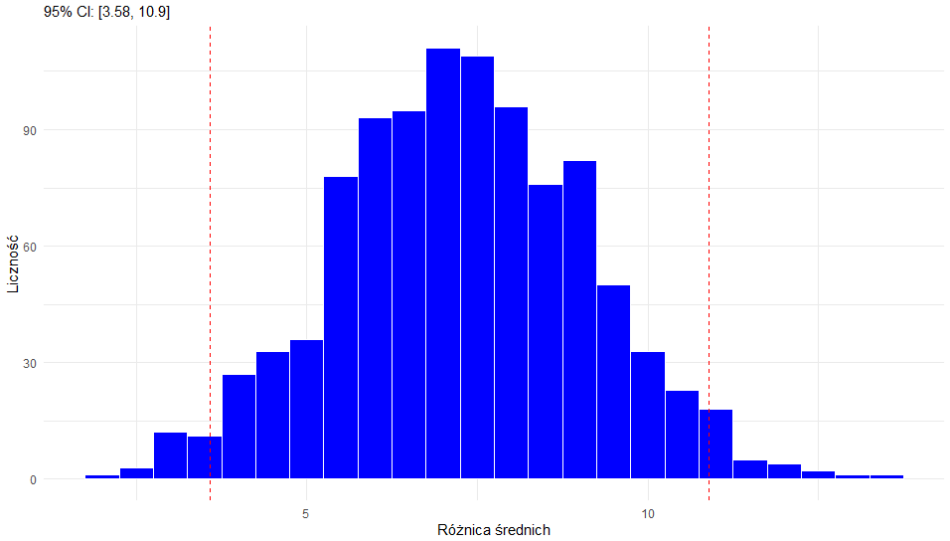
```
  specify(mpg ~ am) %>%
  calculate(stat = "diff in means", order = c("Ręczna", "Automatyczna"))

null_distribution_mpg <- mtcars_df %>%
  specify(mpg ~ am) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("Ręczna", "Automatyczna"))

p_value_mpg_independent <- null_distribution_mpg %>%
  get_p_value(obs_stat = mpg_diff_obs, direction = "both")

conf_int_mpg <- mtcars_df %>%
  specify(mpg ~ am) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("Ręczna", "Automatyczna")) %>%
  get_confidence_interval(level = 0.95, point_estimate = mpg_diff_obs)

paste0("Obserwowana różnica średnich (Ręczna - Automatyczna)dla mpg:", mpg_diff_obs$stat)
# [1] "Obserwowana różnica średnich (Ręczna - Automatyczna) dla mpg:7.24493927125506"
paste0("Empiryczna p-wartość (dwustronna):", p_value_mpg_independent$p_value)
# [1] "Empiryczna p-wartość (dwustronna):2e-04"
paste0("95% przedział ufności dla różnicy średnich (bootstrap):", conf_int_mpg$lower_ci,
"do", conf_int_mpg$upper_ci)
# [1] "95% przedział ufności dla różnicy średnich (bootstrap):3.63488214285714do10.9
820252912759"
```

Pakiet **`infer`** umożliwia czytelną wizualizację przedziału ufności dla różnicy wartości przeciętnych. Rozkład jest szacowany symulacyjnie w oparciu o próbkowanie bootstrapowe.

Należy sprawdzić, czy istnieje różnica w liczbie mil przejechanych na galonie paliwa (*mpg*) między samochodami z automatyczną a ręczną skrzynią biegów. Realizują to poniższe komendy.

```
mtcars2 <- mtcars %>%
  mutate(am = factor(am, labels = c("automat", "manual")))

bootstrap_dist <- mtcars2 %>%
  specify(mpg ~ am) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("manual", "automat"))

ci <- bootstrap_dist %>%
  get_confidence_interval(level = 0.95, type = "percentile")

ggplot(bootstrap_dist, aes(x = stat)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "white") +
```

```
  geom_vline(xintercept = ci$lower_ci, color = "red", linetype = "dashed") +
  geom_vline(xintercept = ci$upper_ci, color = "red", linetype = "dashed") +
  labs(title = "", subtitle = paste0("95% CI: [", round(ci$lower_ci, 2), ", ", round(ci$uppe
r_ci, 2), "]"), x = "Różnica średnich", y = "Liczność") +
  theme_minimal()
```



Rys. 5.13. Rozkład bootstrapowy różnicy średnich mpg z zaznaczonym 95-procentowym przedziałem ufności

Przedział ufności o końcach 3,58 i 10,9 (rys. 5.13) z prawdopodobieństwem 0,95 obejmuje nieznaną rzeczywistą średnią liczbę mil przejechanych na galonie paliwa.

6. Analizy on-line i wnioskowanie statystyczne w modelu permutacyjnym

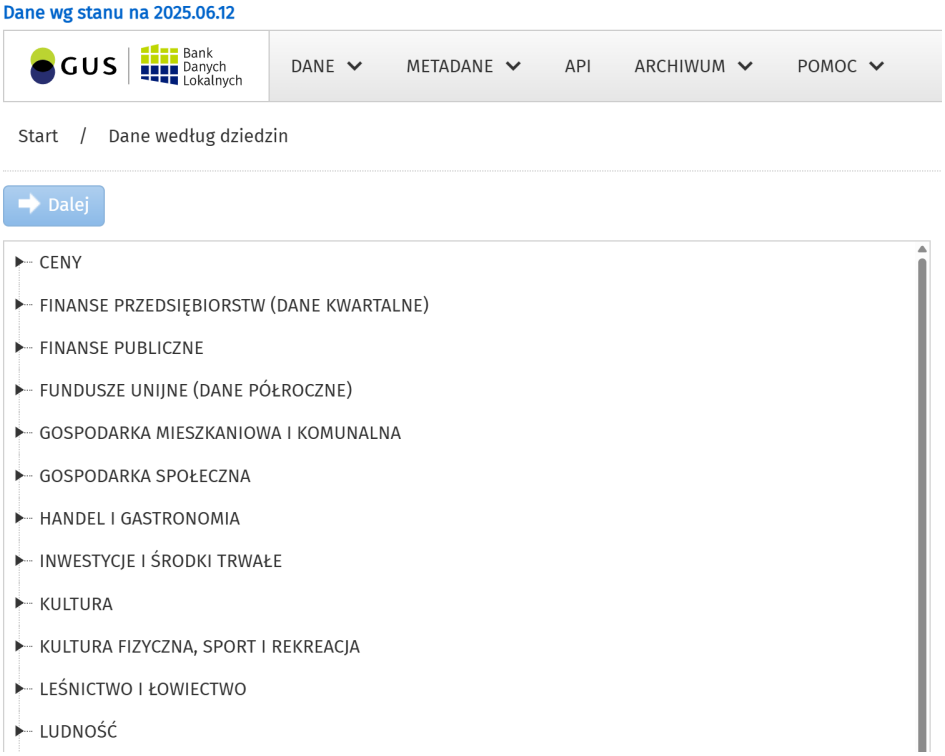
> *Grafika może opowiadać fascynującą historię,*
> *ale może też zachęcać do poszerzania horyzontów w sposób,*
> *którego jej twórcy nie przewidzieli.*
>
> Alberto Cairo (2016, s. 7)

W rozdziale przedstawiono możliwości przeprowadzania analiz on-line na podstawie danych pozyskanych dynamicznie z Banku Danych Lokalnych (BDL, 2025). Bank Danych Lokalnych stanowi największą w Polsce bazę danych o gospodarce, społeczeństwie i środowisku. W bazie tej jest dostępnych, przede wszystkim w układzie terytorialnym, ponad 40 tys. cech statystycznych pogrupowanych tematycznie. Najstarsze dane pochodzą z 1995 roku. Dostępne są m.in. dane i wskaźniki opisujące miejscowości statystyczne, gminy, powiaty, województwa i Polskę, a także jednostki zgodne z nomenklaturą NUTS: podregiony i regiony (Bąk i in., 2024). Ponadto użytkownicy mają dostęp do zasobów informacyjnych, jak wybrane statystyki roczne i krótkookresowe, które są na bieżąco aktualizowane i uzupełniane. Dodatkowo można uzyskać informacje: Portret terytorium, Ranking, Statystyczne vademecum samorządowca, a także Wskaźniki zrównoważonego rozwoju dostępne w Dziedzinowych Bazach Wiedzy – Statystyka wielodziedzinowa – Przekroje Terytorialne BDL. Szczegółowe informacje o funkcjach dostępnych w pakiecie **`bdl`** podają Szpadel i Kania (2023). Opis możliwości dostępu do Banku Danych Lokalnych z poziomu programu **R** przedstawia API_BDL (2025). Po wybraniu odpowiedniej zmiennej można uzyskać szczegółowe informacje i pobrać odpowiednie dane w postaci pliku *csv*, *xls* lub *xlsx*.

6.1. Bank Danych Lokalnych i załadowanie wymaganych bibliotek

Pakiet **bdl** w środowisku **R** umożliwia programowy dostęp do danych zgromadzonych w Banku Danych Lokalnych. Zamiast pozyskiwania danych ze strony internetowej BDL (por. rys. 6.1), użytkownik może się odwołać do funkcji pakietu

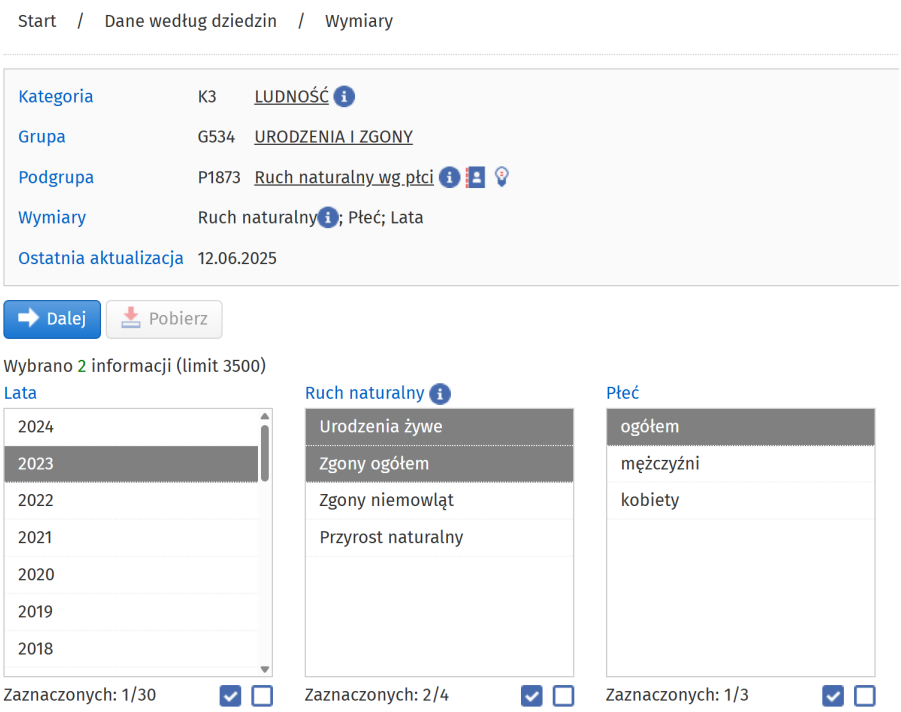
bdl w celu wyszukiwania, filtrowania i pobierania interesujących go danych bezpośrednio z poziomu programu **R**.



Rys. 6.1. Pobranie danych z Banku Danych Lokalnych

Źródło: BDL (2025).

Dane mogą być pobierane po wskazaniu jedynie kodu danej zmiennej. Niezbędny kod można uzyskać, wybierając odpowiednią zmienną na portalu Banku Danych Lokalnych i odwołując się do numeru grupy. Na rysunku 6.2 można dostrzec, że dla zmiennych urodzenia żywe na 1000 ludności i zgony na 1000 ludności odpowiednie symbole są następujące: Kategoria K3 (Ludność), Grupa G534 (Urodzenia i zgony) oraz Podgrupa P3428 (Urodzenia żywe, zgony i przyrost naturalny na 1000 ludności).



Rys. 6.2. Pozyskanie danych z Banku Danych Lokalnych – kategoria, grupa i podgrupa

Źródło: BDL (2025).

Przed przystąpieniem do pozyskania danych z BDL i przeprowadzeniem analiz niezbędne jest załadowanie odpowiednich bibliotek, co realizuje poniższy kod.

```
library(ggplot2)
library(bdl)
library(ggrepel)
library(ggforce)
library(ggExtra)
library(hrbrthemes)
library(ggthemes)
library(GGally)
library(ggcorrplot)
library(ggridges)
library(geofacet)
library(maps)
library(see)
library(patchwork)
library(cowplot)
library(scales)
library(lemon)
library(tidyverse)
library(forecast)
```

6.2. Pozyskanie i wstępna analiza danych z BDL

Po załadowaniu biblioteki **bdl** wprowadzenie w programie **R** komendy *get_variables*("P3428") pozwala uzyskać *id* danej zmiennej. W przypadku danych dla urodzeń żywych na 1000 ludności jest to *id*=450540 (por. rys. 6.2).

W podgrupie P3428 są dostępne następujące zmienne:

- liczba urodzeń żywych na 1000 ludności,
- liczba zgonów na 1000 ludności,
- przyrost naturalny na 1000 ludności.

Pozyskanie informacji o 'id' zmiennych z podgrupy P3428 jest realizowane następująco:

```
get_variables('P3428')
# # A tibble: 3 × 6
#       id subjectId n1                          level measureUnitId measureUnitName
#    <int> <chr>     <chr>                       <int>         <int> <chr>
# 1 450540 P3428     urodzenia żywe na 1000 l…     6             1 -
# 2 450541 P3428     zgony na 1000 ludności        6             1 -
# 3 450551 P3428     przyrost naturalny na 10…     6             1 -
```

W tabeli 6.1 wskazano 'id' oraz nazwy zmiennych, które zostaną pobrane z poziomu **R** i wykorzystane w dalszych analizach.

Tabela 6.1. Zmienne i ich kody z BDL wykorzystane w analizach

id	Zmienna
72305	Liczba ludności
450540	Liczba urodzeń żywych na 1000 ludności
450541	Liczba zgonów na 1000 ludności
450551	Przyrost naturalny na 1000 ludności
450543	Małżeństwa na 1000 ludności
1616556	Rozwody na 10 000 ludności

Pozyskanie wskazanych w tabeli 6.1 zmiennych w układzie dla województw (unitLevel=2) realizuje następujący kod:

```
bdl_woj <- c(
  "Liczba_lud"      = "72305",
  "Urodzenia_1000"  = "450540",
  "Zgony_1000"      = "450541",
  "Przyrost_n_1000" = "450551",
  "Małżeństwa_1000" = "450543",
  "Rozwody_10000"   = "1616556"
)

raw_woj <- get_data_by_variable(unname(bdl_woj), unitLevel = 2)
```

```
woj <- raw_woj %>%
  mutate(year = as.numeric(year)) %>%
  filter(year > 2001) %>%
  rename_with(~ names(bdl_woj)[which(bdl_woj == sub("val_", "", .x))],
              starts_with("val_")) %>%
  select(Id = id,
         Województwo = name,
         Rok = year,
         all_of(names(bdl_woj)))
head(woj)
```

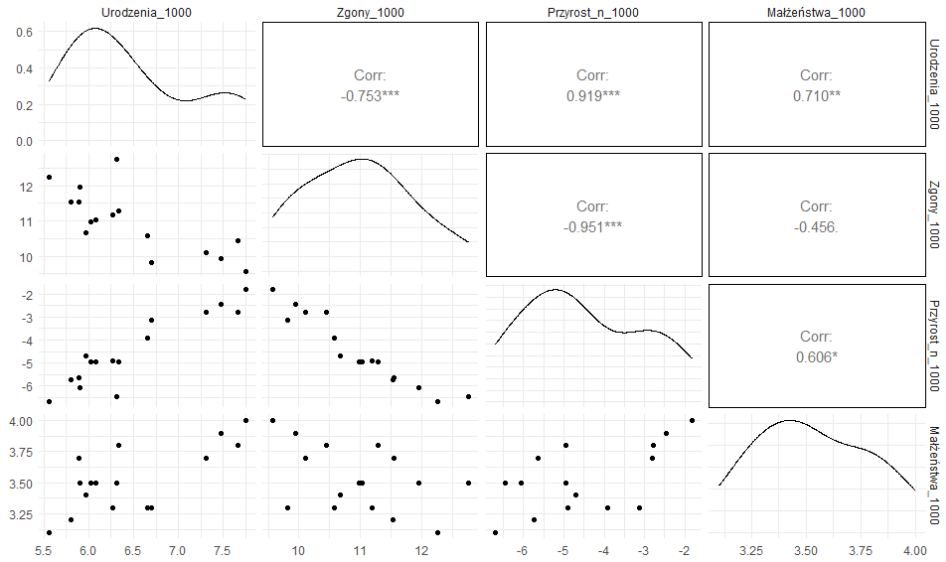
```
# # A tibble: 6 × 9
#   Id      Województwo   Rok Liczba_lud Urodzenia_1000 Zgony_1000 Przyrost_n_1000
#   <chr>   <chr>       <dbl>      <int>          <dbl>      <dbl>           <dbl>
# 1 011200… MAŁOPOLSKIE  2002    3237217          10.0        8.64            1.4
# 2 011200… MAŁOPOLSKIE  2003    3252949           9.86       8.88            0.98
# 3 011200… MAŁOPOLSKIE  2004    3260201           9.87       8.71            1.16
# 4 011200… MAŁOPOLSKIE  2005    3266187          10          8.92            1.08
# 5 011200… MAŁOPOLSKIE  2006    3271206          10.0        8.82            1.22
# 6 011200… MAŁOPOLSKIE  2007    3279036          10.4        9.03            1.42
# # ℹ 2 more variables: Małżeństwa_1000 <dbl>, Rozwody_10000 <dbl>
```

Dla przeprowadzenia analiz określono dwa podzbiory pozyskanego zbioru **woj**. Zbiór **woj_od2002** zawiera zmienne dla lat 2002–2024, natomiast zbiór **woj_2024** to wyłącznie dane dla 2024 roku.

```
woj_od2002 <- woj %>% filter(Rok > 2001)
woj_2024   <- woj %>% filter(Rok == 2024)
```

Wykorzystując drugi z podanych zbiorów, można przedstawić zależności pomiędzy wybranymi zmiennymi w następujący sposób:

```
GGally::ggpairs(woj_2024[,5:8])
```

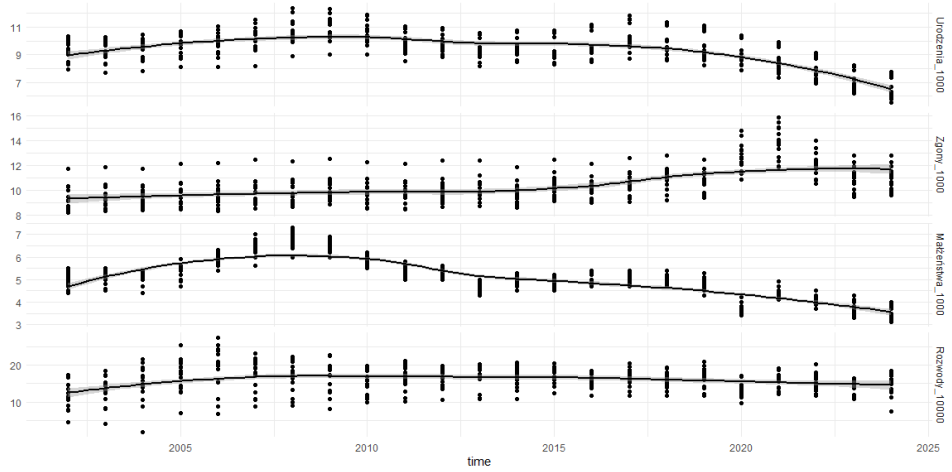


Rys. 6.3. Liczba urodzeń żywych, liczba zgonów i liczba małżeństw na 1000 ludności oraz liczba rozwodów na 10 000 ludności w 2024 roku

Rysunek 6.3 przedstawia wzajemne powiązania i rozkłady zmiennych: urodzenia żywe na 1000 ludności, zgony na 1000 ludności, małżeństwa na 1000 ludności oraz rozwody na 10 000 ludności w województwach w 2024 roku. Na wykresie dodatkowo zamieszczono wartości współczynnika korelacji pomiędzy analizowanymi zmiennymi. Najsilniejsza zależność dodatnia występuje pomiędzy liczbą urodzeń żywych a przyrostem naturalnym ($r = 0{,}919$), a ujemna pomiędzy liczbą zgonów a przyrostem naturalnym ($r = -0{,}951$).

Na rysunku 6.4 przedstawiono zmiany analizowanych zmiennych w latach 2002–2024. Zauważalny jest spadek liczby urodzeń żywych na 1000 ludności i wzrost liczby zgonów na 1000 ludności od 2015 roku. Widoczny jest także trend spadkowy liczby małżeństw na 1000 ludności od 2008 roku.

```
woj %>%
  forecast::ggts("Rok", c("Urodzenia_1000", "Zgony_1000", "Małżeństwa_1000",
"Rozwody_10000"))
```



Rys. 6.4. Urodzenia żywe, zgony i małżeństwa na 1000 ludności oraz rozwody na 10 000 ludności w województwach w latach 2002–2024

W dalszej części poza danymi dla województw będą wykorzystywane dane dla powiatów (unitLevel=5) dla tych samych zmiennych. Dodatkowo utworzono zmienną *mz*, która pozwala rozróżnić powiaty (powiat) i miasta na prawach powiatu (m.powiat). Realizuje to następujący kod:

```
bdl_pow <- c(
  "Liczba_lud"      = "72305",
  "Urodzenia_1000"  = "450540",
  "Zgony_1000"      = "450541",
  "Przyrost_n_1000" = "450551",
  "Małżeństwa_1000" = "450543",
  "Rozwody_10000"   = "1616556"
)

raw_pow <- get_data_by_variable(unname(bdl_pow), unitLevel = 5)
pow <- raw_pow %>%
  mutate(year = as.numeric(year)) %>%
  select(Id = id,
         Powiat = name,
         Rok = year,
         all_of(setNames(paste0("val_", bdl_pow), names(bdl_pow)))) %>%
  mutate(
    mz = if_else(stringr::str_detect(Powiat, "m\\."), "m.powiat", "powiat")
  )
head(pow)
# # A tibble: 6 × 10
#   Id           Powiat    Rok Liczba_lud Urodzenia_1000 Zgony_1000 Przyrost_n_1000
#   <chr>        <chr>   <dbl>      <int>          <dbl>      <dbl>           <dbl>
# 1 011212001000 Powia…  1995      96318             NA
# 2 011212001000 Powia…  1996      96904             NA
```

```
# 3 011212001000 Powia…  1997      97185             NA
# 4 011212001000 Powia…  1998      97516             NA
# 5 011212001000 Powia…  1999      97048             NA
# 6 011212001000 Powia…  2000      97328             NA
# # i 3 more variables: Małżeństwa_1000 <dbl>, Rozwody_10000 <dbl>, mz <chr>
```

Pobrane dane zawierają kolumnę (zmienną) *Id*, ale nie są to nazwy, a jedynie kody województw. Dla wprowadzenia nazw można wykonać rekodowanie zmiennej *województwo* jak w poniższym kodzie.

```
pow <- pow %>%
 mutate(
   Województwo = substr(Id, 1, 4),
   Województwo = recode(Województwo,
    '0714' = 'MAZOWIECKIE',
    '0620' = 'PODLASKIE',
    '0618' = 'PODKARPACKIE',
    '0606' = 'LUBELSKIE',
    '0526' = 'ŚWIĘTOKRZYSKIE',
    '0510' = 'ŁÓDZKIE',
    '0428' = 'WARMIŃSKO-MAZURSKIE',
    '0422' = 'POMORSKIE',
    '0404' = 'KUJAWSKO-POMORSKIE',
    '0316' = 'OPOLSKIE',
    '0302' = 'DOLNOŚLĄSKIE',
    '0232' = 'ZACHODNIOPOMORSKIE',
    '0230' = 'WIELKOPOLSKIE',
    '0208' = 'LUBUSKIE',
    '0124' = 'ŚLĄSKIE',
    '0112' = 'MAŁOPOLSKIE'
   )
 )
```

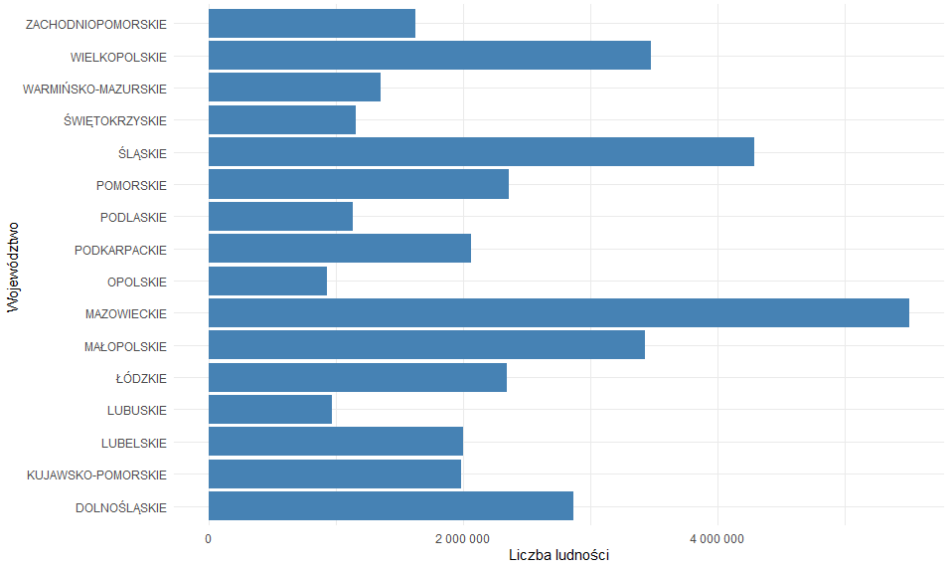
Na potrzeby dalszych analiz na podstawie danych ze zbioru **pow** zostaną utworzone podzbiory **pow_od2002** (dane za lata 2002–2024), **pow_2009** i **pow_2024** (dane wyłącznie za lata odpowiednio: 2009 i 2024) oraz **pow_2y** (dane z lat 2019 i 2024).

```
pow_od2002 <- filter(pow, Rok > 2001)
pow_2009   <- filter(pow, Rok == 2009)
pow_2024   <- filter(pow, Rok == 2024)
pow_2y     <- filter(pow, Rok %in% c(2019, 2024))
```

6.3. Ludność według województw

Przed przystąpieniem do analiz dotyczących liczby urodzeń żywych na 1000 ludności oraz liczby zgonów na 1000 ludności zostanie przedstawiony rozkład liczby ludności w województwach w 2024 roku.

```
woj_2024 %>%
 ggplot(aes(Województwo,Liczba_lud)) +
 geom_col(fill = "steelblue") +
 labs(x='Województwo ',y='Liczba ludności')+
 scale_y_continuous(scales::labels = label_number())+
 coord_flip()+
 theme_minimal()
```



Rys. 6.5. Liczba ludności w województwach w 2024 roku

Rysunek 6.5 prezentuje liczbę ludności polskich województw. Najliczniejsze z nich to mazowieckie, śląskie, małopolskie i wielkopolskie, natomiast najmniej mieszkańców odnotowuje się w opolskim, lubuskim, podlaskim i świętokrzyskim.

Dla zobrazowania struktury wieku ludności często jest wykorzystywany wykres zwany piramidą wieku. Dla skonstruowania takiego wykresu niezbędne jest pobranie danych o liczbie ludności według płci i grup wieku. Pobranie danych do piramidy wieku może być zrealizowane następująco:

```
nazwy <- c(
  'Rok', '0-4', '5-9', '10-14', '15-19', '20-24', '25-29', '30-34', '35-39', '40-44',
  '45-49', '50-54', '55-59', '60-64', '65-69', '70-74', '75-79', '80-84', '85 i więcej'
)
pw_k <- get_data_by_variable(
  c(72296, 72297, 72298, 72299, 47738, 47696, 47695, 47716, 47698,
    47727, 47723, 47702, 47693, 72241, 76014, 76015, 76016, 76017),
  unitLevel = 0
)
pw_m <- get_data_by_variable(
  c(72301, 72302, 72303, 72304, 47711, 47736, 47724, 47712, 47725,
    47728, 47706, 47715, 47721, 72243, 76018, 76019, 76020, 76021),
  unitLevel = 0
)
```

Po pobraniu danych niezbędne jest wprowadzenie odpowiednich nazw kolumn.

```
names(pw_k)[3:21] <- nazwy
names(pw_m)[3:21] <- nazwy

pw_k <- pw_k[8:30, 3:21]
pw_m <- pw_m[8:30, 3:21]

pw_k$płeć <- 'kobieta'
pw_m$płeć <- 'mężczyzna'

pw <- rbind(pw_k, pw_m)
pw$Rok <- as.numeric(pw$Rok)
pw$płeć <- factor(pw$płeć, levels = c('mężczyzna', 'kobieta'))
```

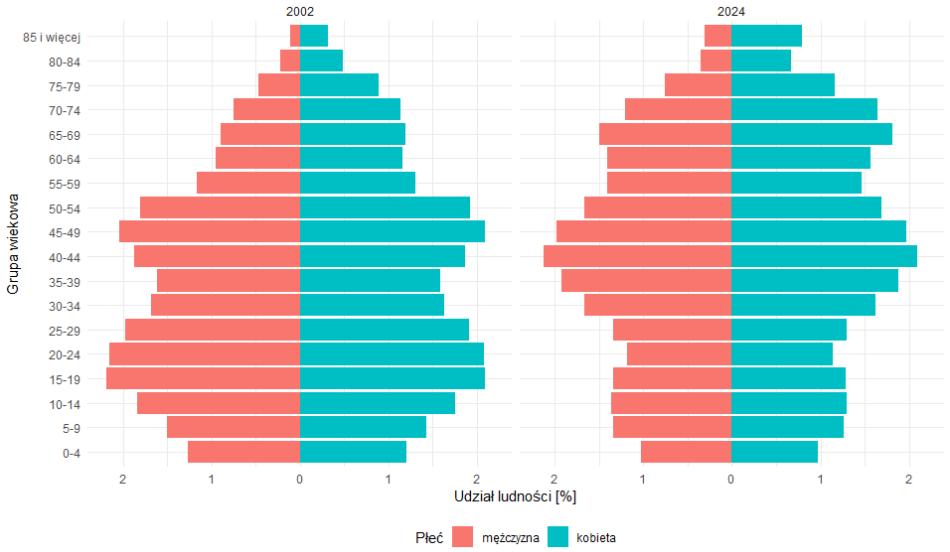
Na potrzeby konstrukcji piramidy wieku format danych zostanie zmieniony na długi. Dodatkowo typ zmiennej *płeć* zostanie zdefiniowany jako czynnikowy (factor).

```
pw_long= pw %>%
  tidyr::pivot_longer(!c(Rok,płeć), names_to = "Wiek", values_to = "val")
pw_long$Wiek=factor(pw_long$Wiek,levels=c('0-4','5-9', '10-14', '15-19', '20-24','25-29'
,'30-34','35-39','40-44','45-49','50-54','55-59','60-64','65-69','70-74','75-79','80-84','85 i wi
ęcej'))
```

Na podstawie utworzonego zbioru konstrukcja piramidy wieku dla lat 2002 i 2024 może być wykonana następująco:

```
pw_long %>%
  filter(Rok %in% c(2002, 2024)) %>%
  mutate(
    ludn = sum(val),
```

```
  val_pct = ifelse(płeć == "mężczyzna", -val / ludn * 100, val / ludn * 100)
) %>%
ggplot(aes(x = val_pct, y = Wiek, fill = płeć)) +
geom_col() +
lemon::scale_x_symmetric(labels = abs) +
facet_wrap(~Rok) +
labs(
  x = "Udział ludności [%]",
  y = "Grupa wiekowa",
  fill = "Płeć"
) +
theme_minimal() +
theme(legend.position = 'bottom')
```



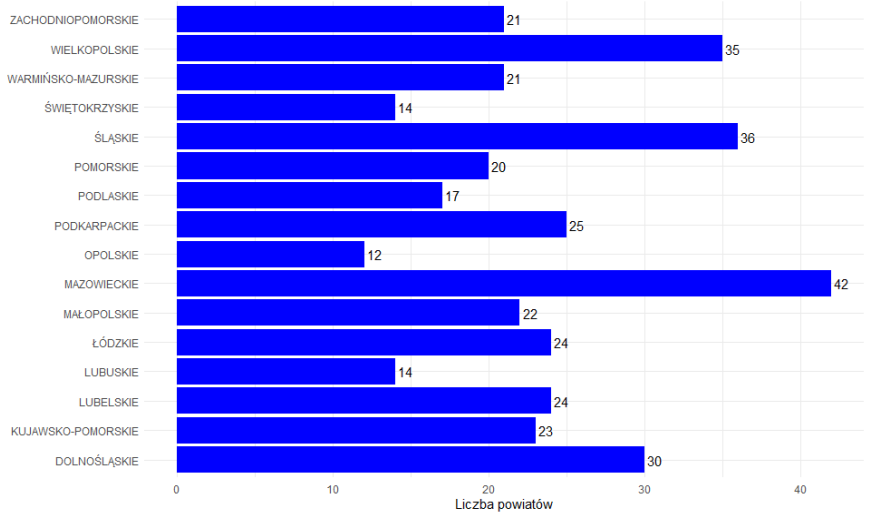
Rys. 6.6. Struktura wieku według płci w 2002 i 2024 roku

Podczas porównywania piramid wieku dla lat 2002 i 2024 (rys. 6.6) widoczne jest znaczne zwiększenie liczby ludności w najwyższych grupach wiekowych (85 lat i więcej oraz 80–84). Jednocześnie zauważalne jest zmniejszenie się liczby ludności w grupach wiekowych do 30 lat.

6.4. Liczba i typy powiatów

W niniejszym podrozdziale przedstawiono wizualizacje analizowanych zmiennych według powiatów. Na wstępie graficznie zilustrowano liczbę powiatów według województw.

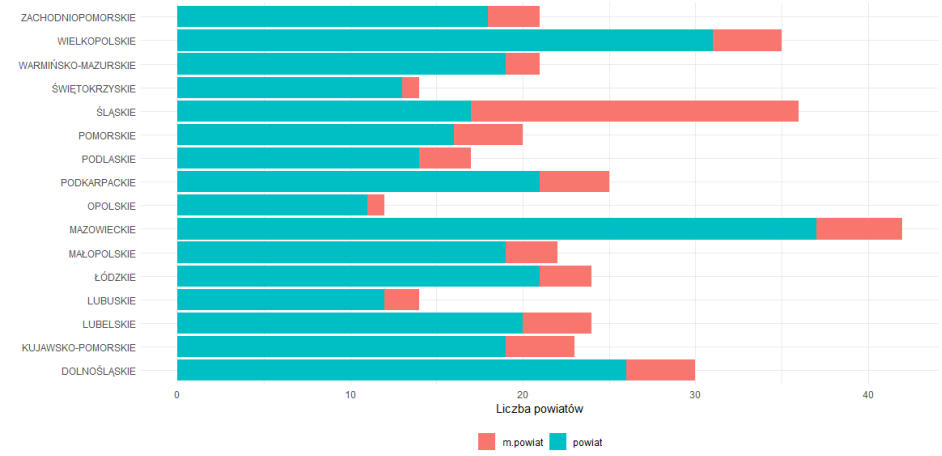
```
pow_2024 %>%
  ggplot(aes(x = Województwo)) +
  geom_bar(fill = 'blue') +
  geom_text(stat = "count", aes(label = after_stat(count)), hjust = -0.2) +
  labs(y = 'Liczba powiatów', x = NULL) +
  coord_flip() +
  theme_minimal() +
  theme(legend.position = 'none')
```



Rys. 6.7. Liczba powiatów według województw w 2024 roku

Rysunek 6.7 obrazuje liczbę powiatów w poszczególnych województwach. Najwięcej powiatów znajduje się w województwach mazowieckim (42), śląskim (36) i wielkopolskim (35), a najmniej – w opolskim (12), lubuskim (14) i podlaskim (17). W dalszej części wykorzystano rozróżnienie na powiaty i miasta na prawach powiatów. Dodatkowe rozróżnienie ze względu na typ powiatu (powiat i miasto na prawach powiatu) uwzględnia poniższy kod, którego wynik prezentuje rys. 6.8.

```
pow_2024 %>%
  ggplot(aes(x = Województwo, fill = mz)) +
  geom_bar() +
  labs(
    x = NULL,
    y = 'Liczba powiatów',
    fill = NULL
  ) +
  coord_flip() +
  theme_minimal() +
  theme(legend.position = 'bottom')
```



Rys. 6.8. Liczba powiatów i miast na prawach powiatu według województw w 2024 roku

Porównania wybranych charakterystyk dla powiatów i miast na prawach powiatów zostaną w dalszej części przeprowadzone dla trzech województw: małopolskiego, mazowieckiego i śląskiego. W wyniku realizacji poniższego kodu uzyskuje się liczbę powiatów i miast na prawach powiatu w tych trzech województwach.

```
pow %>%
  filter(
    Rok == 2024,
    Województwo %in% c("ŚLĄSKIE", "MAŁOPOLSKIE", "MAZOWIECKIE")
  ) %>%
  select(Województwo, mz) %>%
  table()
#                mz
#  Województwo  m.powiat powiat
#    MAŁOPOLSKIE         3     19
#    MAZOWIECKIE         5     37
#    ŚLĄSKIE            19     17
```

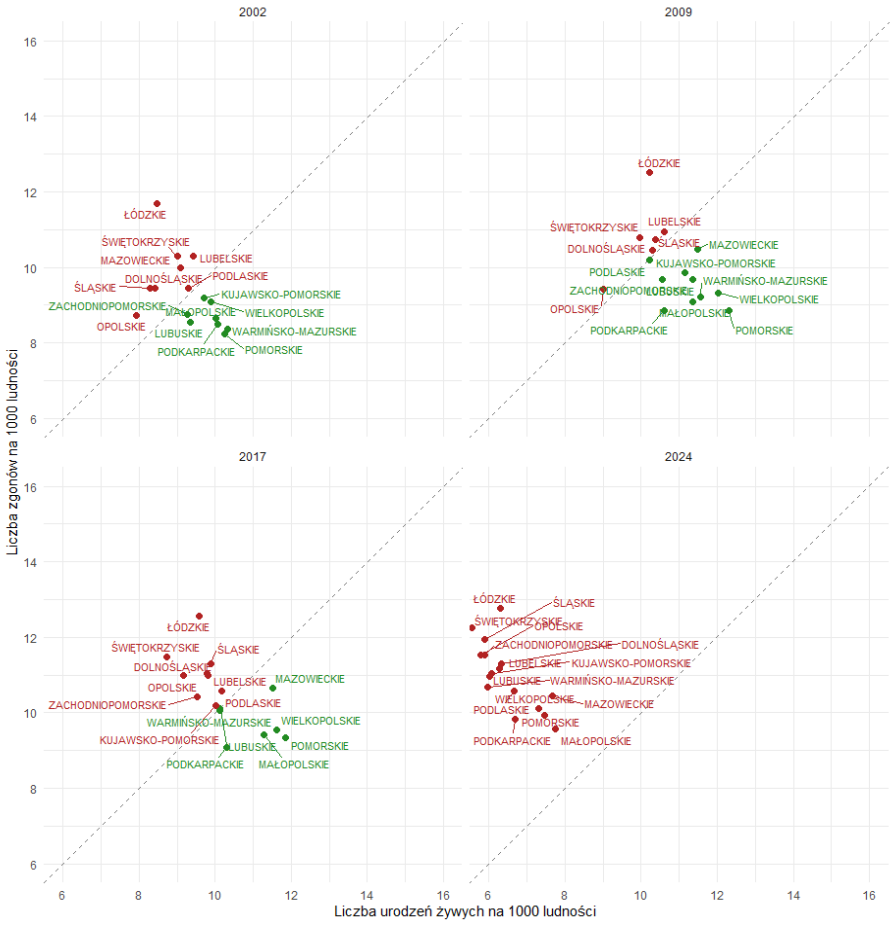
6.5. Urodzenia żywe i zgony na 1000 ludności

6.5.1. Charakterystyka zależności w województwach i powiatach

Na rysunku 6.9 przedstawiono liczbę urodzeń żywych i liczbę zgonów w województwach w latach 2002, 2009, 2017 i 2024. Na wykresach zaznaczono przekątną. Punkty oznaczające województwa, w których liczba urodzeń żywych na

1000 ludności jest większa niż liczba zgonów na 1000 ludności, znajdują się poniżej przekątnej i zostały oznaczone kolorem zielonym. O ile w 2009 roku w większości województw liczba urodzeń żywych przewyższała liczbę zgonów, to w 2024 roku we wszystkich województwach liczba zgonów była większa od liczby urodzeń żywych.

```
woj %>%
  filter(Rok %in% c(2002, 2009, 2017, 2024)) %>%
  mutate(przyrost_dodatni = Urodzenia_1000 > Zgony_1000) %>%
  ggplot(aes(
    x = Urodzenia_1000,
    y = Zgony_1000,
    label = Województwo,
    color = przyrost_dodatni
  )) +
  geom_abline(slope = 1, intercept = 0, lty = 2, color = "gray50") +
  geom_point(size = 2) +
  ggrepel::geom_text_repel(size = 3, max.overlaps = 16) +
  scale_color_manual(values = c("TRUE" = "forestgreen", "FALSE" = "firebrick")) +
  coord_fixed(ratio = 1, xlim = c(6, 16), ylim = c(6, 16)) +
  labs(
    x = 'Liczba urodzeń żywych na 1000 ludności',
    y = 'Liczba zgonów na 1000 ludności'
  ) +
  facet_wrap(~Rok) +
  theme_minimal() +
  theme(legend.position = 'none')
```

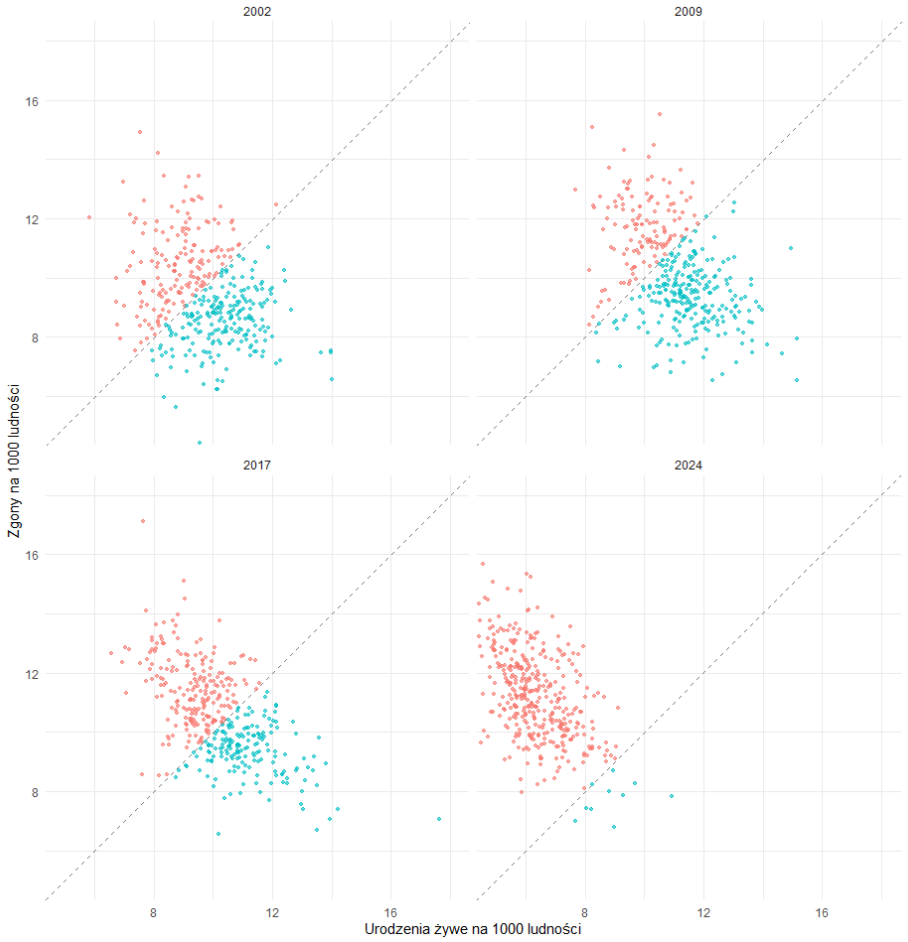


Rys. 6.9. Liczba urodzeń żywych na 1000 ludności i liczba zgonów na 1000 ludności w województwach w latach 2002, 2009, 2017 i 2024

Rysunek 6.10 przedstawia podobne ujęcie jak rys. 6.9, ale nie dla województw, a dla powiatów. Podobnie jak poprzednio, w 2009 roku w większości powiatów liczba urodzeń żywych była większa od liczby zgonów, natomiast w 2024 roku tylko w 10 powiatach w całej Polsce liczba urodzeń żywych przewyższała liczbę zgonów.

```
pow %>%
  filter(Rok %in% c(2002, 2009, 2017, 2024)) %>%
  ggplot(aes(
    x = Urodzenia_1000,
    y = Zgony_1000,
    color = (Urodzenia_1000 > Zgony_1000)
  )) +
```

```
geom_abline(slope = 1, intercept = 0, lty = 2, color = "gray50") +
geom_point(size = 1, alpha = 0.6) +
coord_fixed(ratio = 1, xlim = c(5, 18), ylim = c(5, 18)) +
labs(
  x = 'Urodzenia żywe na 1000 ludności',
  y = 'Zgony na 1000 ludności'
) +
facet_wrap(~Rok) +
theme_minimal() +
theme(legend.position = 'none')
```



Rys. 6.10. Urodzenia żywe i zgony na 1000 ludności w powiatach w latach 2002, 2009, 2017 i 2024

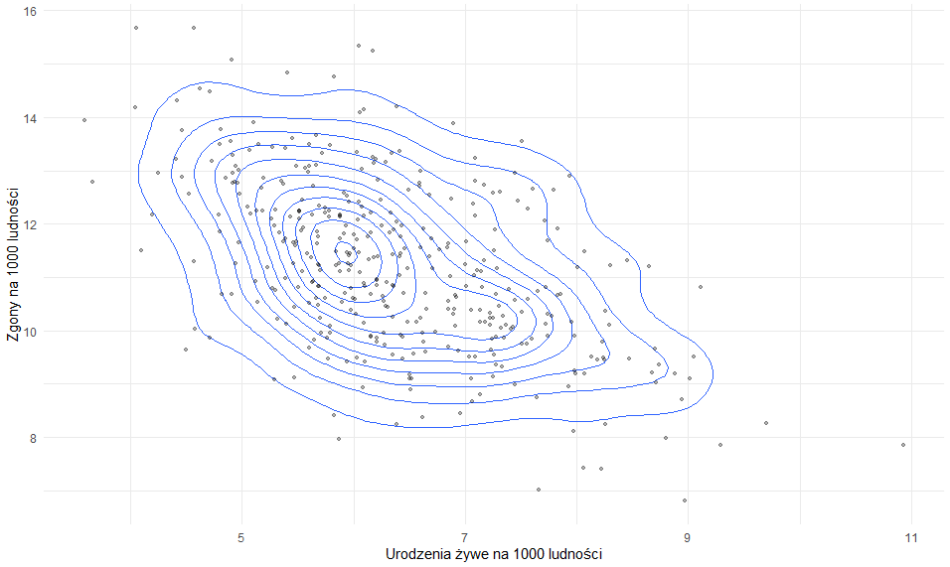
Na rysunku 6.11 przedstawiono wykresy pudełkowe przyrostu naturalnego na 1000 ludności w powiatach w latach 2009 i 2024. Zauważalne jest, że o ile w 2009 roku we wszystkich województwach były powiaty o dodatnim przyroście naturalnym, o tyle w 2024 roku w aż 11 województwach nie było żadnego powiatu o dodatnim przyroście naturalnym.

```
pow %>%
  filter(Rok %in% c(2009, 2024)) %>%
  ggplot(aes(x = Województwo, y = Przyrost_n_1000, fill = factor(Rok))) +
  geom_boxplot() +
  labs(
    x = NULL,
    y = 'Przyrost naturalny na 1000 ludności',
    fill = 'Rok',
    title = 'Przyrost naturalny w powiatach według województw'
  ) +
  coord_flip() +
  theme_minimal() +
  theme(legend.position = 'bottom')
```

Rys. 6.11. Przyrost naturalny na 1000 ludności w powiatach według województw w latach 2009 i 2024

Zastosowanie nieparametrycznej estymacji gęstości dwuwymiarowego rozkładu (X, Y), gdzie:
X – urodzenia żywe na 1000 ludności w powiatach w 2024 roku,
Y – zgony na 1000 ludności w powiatach w 2024 roku,
pozwala na wskazanie najczęściej współwystępujących wartości obu zmiennych.

```
pow %>%
  filter(Rok == 2024) %>%
  ggplot(aes(x = Urodzenia_1000, y = Zgony_1000)) +
  geom_density_2d() +
  geom_point(alpha = 0.3, size = 1) +
  labs(
    x = "Urodzenia żywe na 1000 ludności",
    y = "Zgony na 1000 ludności"
  ) +
  theme_minimal()
```



Rys. 6.12. Urodzenia żywe i zgony na 1000 ludności w powiatach w 2024 roku

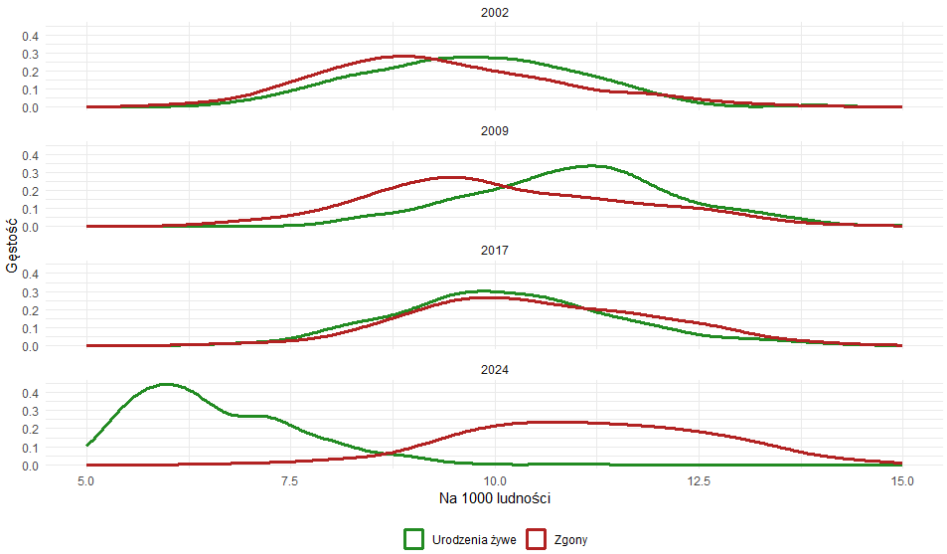
Na rysunku 6.12 widoczne jest, że największe wartości gęstości (dominanta dwuwymiarowa) znajdują się w okolicach liczby urodzeń żywych na 1000 ludności wynoszącej 5,9 oraz liczby zgonów na 1000 ludności równej 11,8. Fakt ten pokazuje, że w znacznej liczbie powiatów liczba zgonów niemal dwukrotnie przekracza liczbę urodzin.

6.5.2. Charakterystyka wskaźników w wybranych latach

Dla uzyskania ocen gęstości obu badanych zmiennych w powiatach w poszczególnych latach można zrealizować kod:

```
pow %>%
  filter(Rok %in% c(2002, 2009, 2017, 2024)) %>%
  select(Rok, Urodzenia_1000, Zgony_1000) %>%
```

```
  tidyr::pivot_longer(cols = c(Urodzenia_1000, Zgony_1000), names_to = "Zmienna",
values_to = "Wartość") %>%
  mutate(Zmienna = case_match(Zmienna,
    "Urodzenia_1000" ~ "Urodzenia żywe",
    "Zgony_1000" ~ "Zgony"
  )) %>%
  ggplot(aes(x = Wartość, color = Zmienna)) +
  geom_density(linewidth = 1.2) +
  xlim(5, 15) +
  scale_color_manual(values = c("Urodzenia żywe" = "forestgreen", "Zgony" = "firebrick")) +
  labs(x = 'Na 1000 ludności', y = 'Gęstość', color = NULL) +
  facet_wrap(~Rok, nrow = 4) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

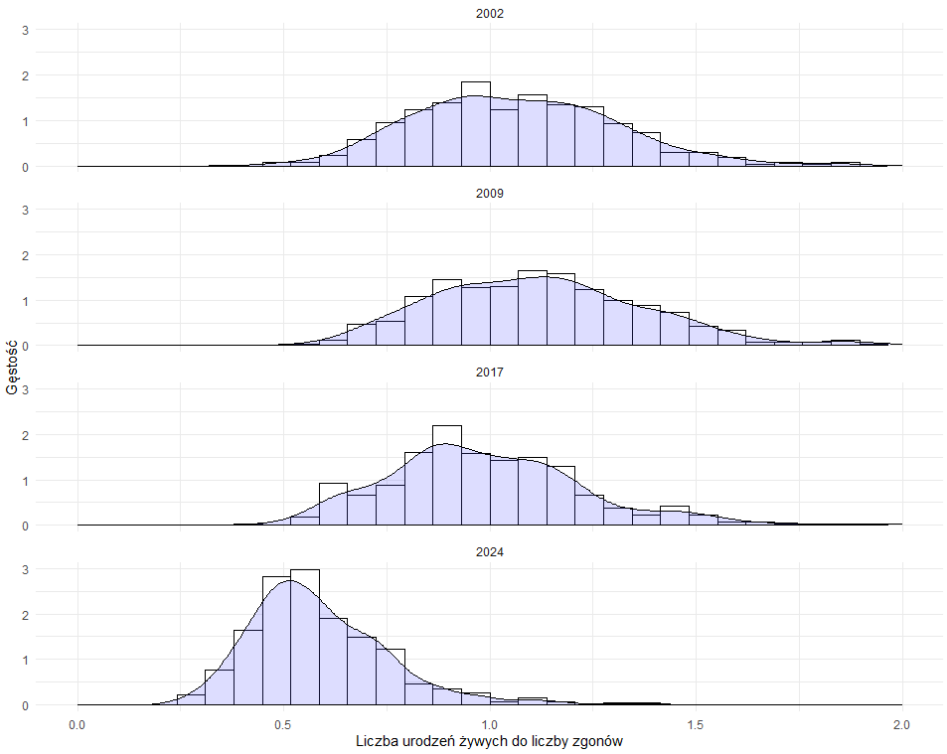


Rys. 6.13. Urodzenia żywe i zgony na 1000 ludności w powiatach w latach 2002, 2009, 2017 i 2024

Rysunek 6.13 przedstawia oszacowania rozkładów liczby urodzeń żywych na 1000 ludności i liczby zgonów na 1000 ludności w powiatach w latach 2002, 2009, 2017 i 2024. W przypadku urodzeń żywych na 1000 ludności rozkład dla 2024 roku wyróżnia się znacznym przesunięciem w lewo wartości najczęstszych (zmniejszenie się wartości) w stosunku do wszystkich pozostałych rozkładów. W przypadku liczby zgonów na 1000 ludności dla 2024 roku widoczne jest przesunięcie się w prawo wartości najczęstszych rozkładu (wzrost wartości) liczby zgonów.

Rysunek 6.14 przedstawia rozkład ilorazu liczby urodzeń żywych do liczby zgonów. O ile w latach 2002, 2009 i 2017 najczęściej występujące wartości są bliskie 1,0, o tyle w 2024 roku wartości najczęściej występujące wynoszą około 0,5. Proporcja ta jest dobrze widoczna także na rys. 6.15, który przedstawia rozkłady tego współczynnika w układzie wojewódzkim w 2024 roku.

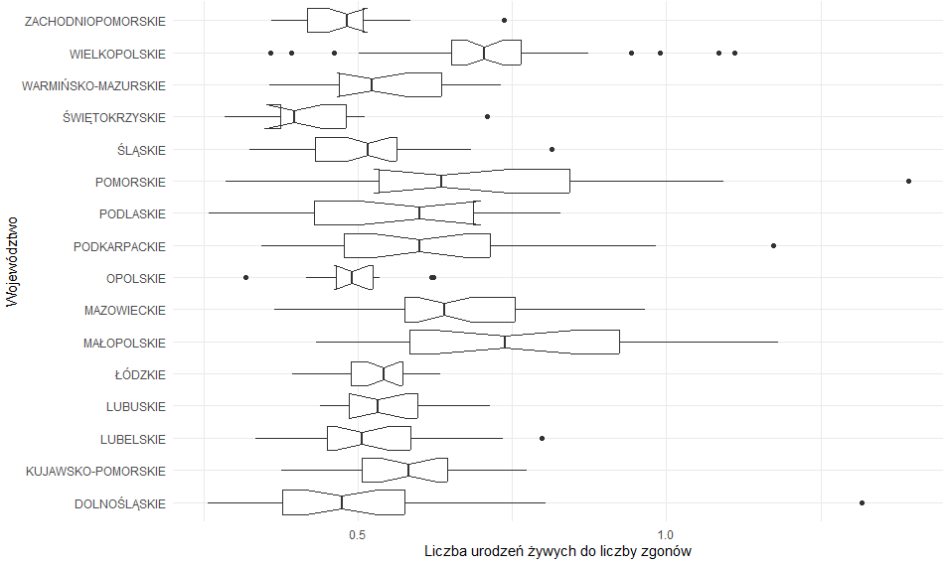
```
pow %>%
  filter(Rok %in% c(2002, 2009, 2017, 2024)) %>%
  mutate(stosunek = Urodzenia_1000 / Zgony_1000) %>%
  ggplot(aes(x = stosunek)) +
  geom_histogram(aes(y = after_stat(density)), colour = "black", fill = "white") +
  geom_density(alpha = 0.2, fill = "#5555FF") +
  xlim(0, 2) +
  labs(x = 'Liczba urodzeń żywych do liczby zgonów', y = 'Gęstość') +
  facet_wrap(~Rok, ncol = 1) +
  theme_minimal()
```



Rys. 6.14. Stosunek liczby urodzeń żywych do liczby zgonów w powiatach w latach 2002, 2009, 2017 i 2024

Poniższy kod prowadzi do przedstawienia na wykresie liczby urodzeń żywych do liczby zgonów w powiatach w 2024 roku.

```
ggplot(pow_2024, aes(x = Województwo, y = Urodzenia_1000/Zgony_1000))+
  geom_boxplot(notch = TRUE) +
  labs(y='Liczba urodzeń żywych do liczby zgonów') +
  coord_flip()+theme_minimal()
```

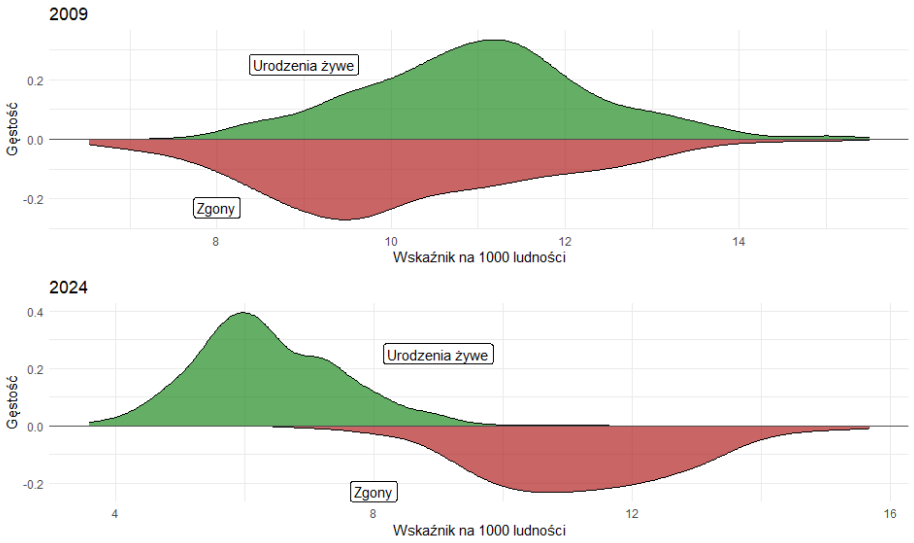


Rys. 6.15. Stosunek liczby urodzeń żywych do liczby zgonów w powiatach według województw w 2024 roku

Na rysunku 6.15 przedstawiono wykresy pudełkowe z nacięciami (notched boxplots) dla ilorazu liczby urodzeń żywych do liczby zgonów w powiatach w 2024 roku. Wykresy te poza typowymi dla wykresów pudełkowych informacjami o wartości mediany, kwartyli pierwszego i trzeciego, wartości maksymalnej i minimalnej dodatkowo przekazują informacje o przedziałach ufności dla mediany. Daje to wskazówkę co do występowania ewentualnych istotnych różnic median. Na rysunku 6.15 widoczne są istotne różnice median, np. pomiędzy województwami małopolskim i świętokrzyskim. Obserwacje na podstawie wykresów powinny zostać potwierdzone zastosowaniem właściwych testów statystycznych.

W nieco innej formie graficznej (rys. 6.16) porównano rozkład urodzeń żywych na 1000 ludności i liczby zgonów na 1000 ludności w powiatach w latach 2009 i 2024.

```
wykres_lustrzany <- function(dane, rok) {
  ggplot(dane) +
    geom_density(
      aes(x = Urodzenia_1000, y = after_stat(density)),
      fill = "forestgreen", alpha = 0.7
    ) +
    geom_density(
      aes(x = Zgony_1000, y = -after_stat(density)),
      fill = "firebrick", alpha = 0.7
    ) +
    geom_hline(yintercept = 0, color = "gray30") +
    annotate("label", x = 9, y = 0.25, label = "Urodzenia żywe") +
    annotate("label", x = 8, y = -0.23, label = "Zgony") +
    labs(
      x = "Wskaźnik na 1000 ludności",
      y = "Gęstość",
      title = as.character(rok)
    ) +
    theme_minimal()
}
p1 <- wykres_lustrzany(pow_2009, 2009)
p2 <- wykres_lustrzany(pow_2024, 2024)
p1 / p2
```



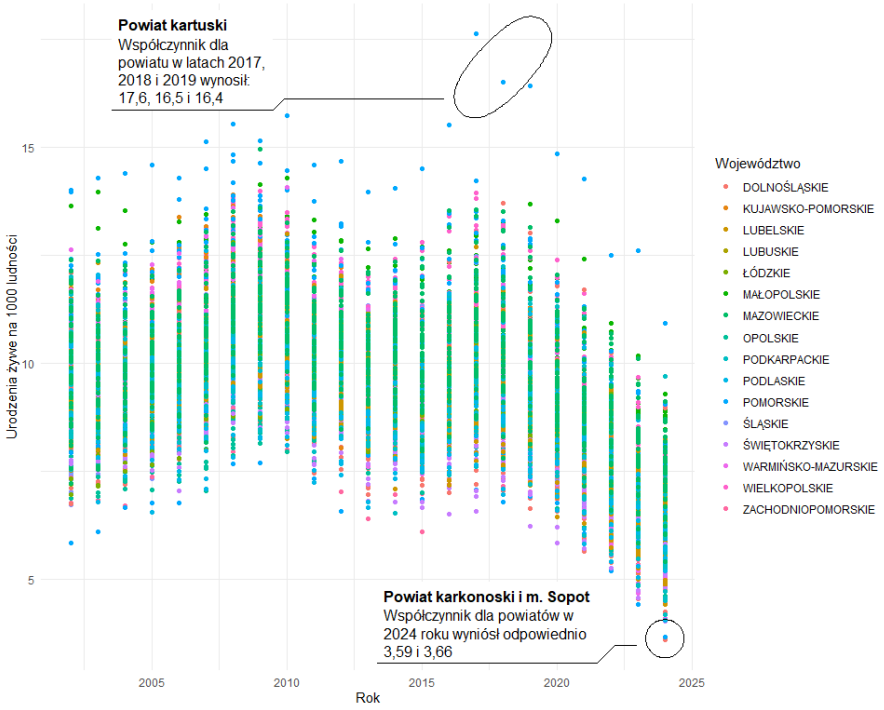
Rys. 6.16. Relacja liczby urodzeń żywych do liczby zgonów w powiatach w latach 2009 i 2024

Charakterystyczne jest to, że o ile w 2009 roku zakres zmienności obu zmiennych (liczba urodzeń żywych na 1000 ludności i liczba zgonów na 1000 ludności) był zbliżony, o tyle w 2024 roku są to już zdecydowanie inne zakresy.

6.5.3. Zmiany wskaźników w latach 2002–2024

Kolejne wykresy pozwolą ukazać zmiany analizowanych współczynników w latach 2002–2024 w powiatach.

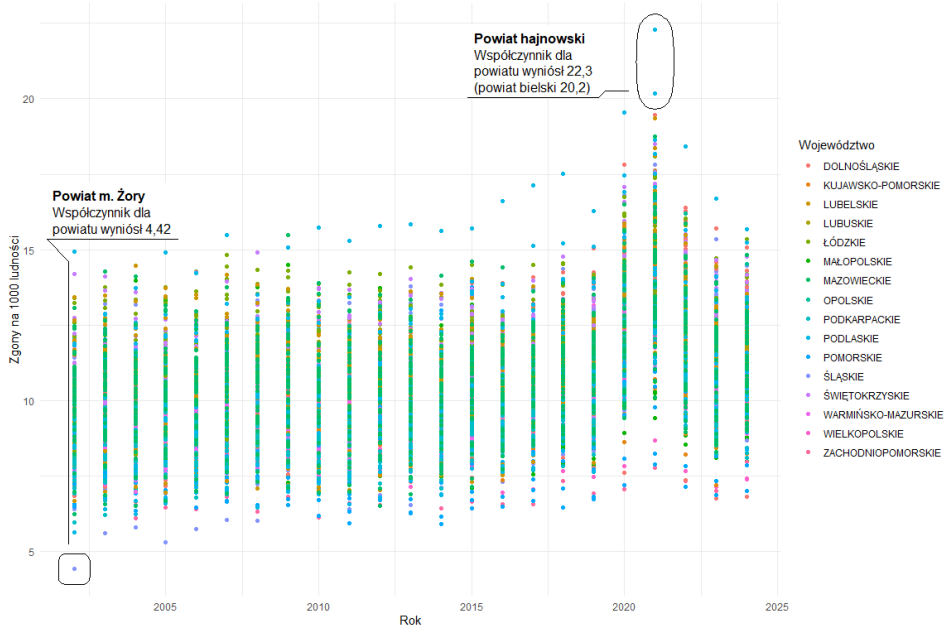
```
pow_od2002 <- pow_od2002 %>%
  drop_na() %>%
  mutate(Województwo = as.factor(Województwo))
ggplot(pow_od2002, aes(x = Rok, y = Urodzenia_1000)) +
  geom_point(aes(color = Województwo)) +
  ggforce::geom_mark_ellipse(aes(
    filter = Urodzenia_1000 > 16.0,
    label = 'Powiat kartuski',
    description = 'Współczynnik dla powiatu w latach 2017, 2018 i 2019 wynosił: 17,6,
16,5 i 16,4'
  )) +
  ggforce::geom_mark_circle(aes(
    filter = Urodzenia_1000 < 4.0,
    label = 'Powiat karkonoski i m. Sopot',
    description = 'Współczynnik dla powiatów w 2024 roku wyniósł odpowiednio 3,59 i 3,66'
  )) +
  labs(
    x = 'Rok',
    y = 'Urodzenia żywe na 1000 ludności',
    color = 'Województwo'
  ) +
  theme_minimal()
```



Rys. 6.17. Urodzenia żywe na 1000 ludności w powiatach według województw w latach 2002–2024

Na rysunku 6.17 przedstawiono liczbę urodzeń żywych na 1000 ludności w powiatach w latach 2002–2024. Największe wartości w tym okresie występowały w powiecie kartuskim (17,6, 16,5 i 16,4 odpowiednio w latach 2017, 2018 i 2019). Najmniejsze wartości wystąpiły w ostatnim z badanych lat – w 2024 roku – i miały miejsce w powiecie karkonoskim (3,59) oraz w mieście na prawach powiatu Sopot (3,66).

```
ggplot(pow_od2002, aes(x = Rok, y = Zgony_1000)) +
  geom_point(aes(color = Województwo)) +
  labs(y = 'Zgony na 1000 ludności') +
  ggforce::geom_mark_ellipse(
    aes(filter = Zgony_1000 > 20,
        label = 'Powiat hajnowski',
        description = 'Współczynnik dla powiatu wyniósł 22,3 (powiat bielski 20,2)')
  ) +
  ggforce::geom_mark_rect(
    aes(filter = Zgony_1000 < 5,
        label = 'Powiat m. Żory',
        description = 'Współczynnik dla powiatu wyniósł 4,42')
  ) +
  theme_minimal()
```

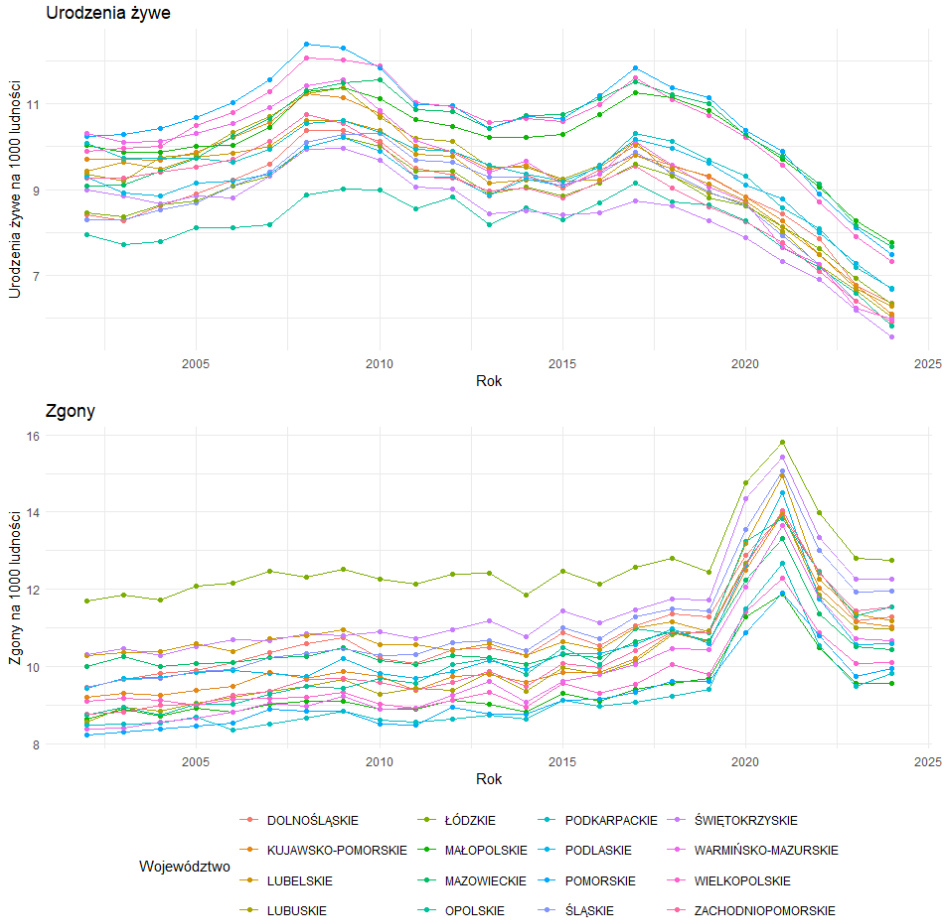


Rys. 6.18. Zgony na 1000 ludności w powiatach według województw w latach 2002–2024

Na rysunku 6.18 przedstawiono liczbę zgonów na 1000 ludności w powiatach w latach 2002–2024. Największe wartości w tym okresie występowały w 2021 roku w powiatach (hajnowskim – 22,3 i bielskim – 20,2). Najniższą wartość odnotowano w 2002 roku w Żorach (mieście na prawach powiatu) i wyniosła ona 4,42. Zmiany w liczbie urodzeń żywych na 1000 ludności i w liczbie zgonów na 1000 ludności w ujęciu wojewódzkim zostały przedstawione na rys. 6.19. Na obu tych wykresach są widoczne systematyczne zmiany we wszystkich województwach w ostatnich latach (2017–2024). W przypadku liczby urodzeń żywych na 1000 ludności jest to systematyczny spadek, a w przypadku liczby zgonów na 1000 ludności – znaczny wzrost w latach 2019–2021 (wpływ pandemii COVID-19), a następnie spadek w latach 2021–2023.

```
p1 <- ggplot(woj_od2002, aes(x = Rok, y = Urodzenia_1000,color=Województwo))+
  geom_point()+
  labs(y='Urodzenia żywe na 1000 ludności',title='Urodzenia żywe na 1000 ludności
w województwach latach 2002 - 2024') +
  geom_line()+
  labs(y='Urodzenia żywe na 1000 ludności',title='Urodzenia żywe') +
  theme(legend.position="none")

p2=ggplot(woj_od2002, aes(x = Rok, y = Zgony_1000,color=Województwo))+
  geom_point()+
```

```
  labs(y='Zgony na 1000 ludności',title='Zgony')  +
  geom_line()+
  theme(legend.position="bottom") +
  labs(y='Zgony na 1000 ludności',title='Zgony')

p1/p2 + theme_minimal()+theme(legend.position='bottom')
```



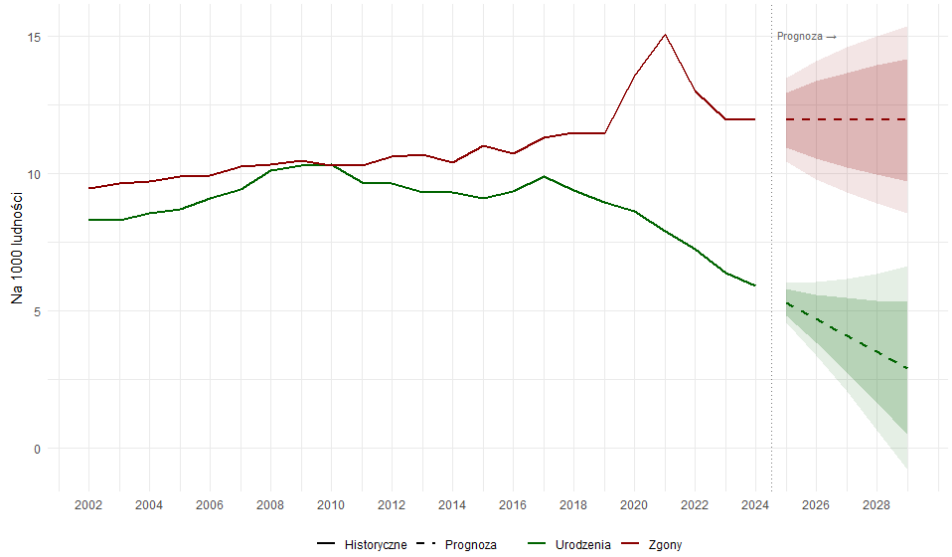
Rys. 6.19. Urodzenia żywe i zgony na 1000 ludności według województw w latach 2002–2024

Na podstawie informacji z lat 2002–2024 możliwe jest postawienie prognozy odnośnie do wartości analizowanych współczynników. Realizuje to następujący kod.

```
slask <- woj_od2002 %>%
  filter(Województwo == 'ŚLĄSKIE') %>%
  select(Rok, Urodzenia_1000, Zgony_1000) %>%
  na.omit()
```

```
prog_arima <- function(x, h = 5) {
  p <- forecast(auto.arima(ts(x, start = 2002)), h = h)
  data.frame(Rok = 2025:(2024+h), Wartość = as.numeric(p$mean),
             D80 = p$lower[,1], G80 = p$upper[,1],
             D95 = p$lower[,2], G95 = p$upper[,2], Typ = "Prognoza")
}

df_comb_slask <- slask %>%
  tidyr::pivot_longer(-Rok, names_to = "Zmienna", values_to = "Wartość") %>%
  mutate(Typ = "Historyczne", D80 = NA, G80 = NA, D95 = NA, G95 = NA) %>%
  bind_rows(
    prog_arima(slask$Urodzenia_1000) %>% mutate(Zmienna = "Urodzenia_1000"),
    prog_arima(slask$Zgony_1000) %>% mutate(Zmienna = "Zgony_1000")
  )
```

Uzyskane prognozy zostały przedstawione na rys. 6.20.



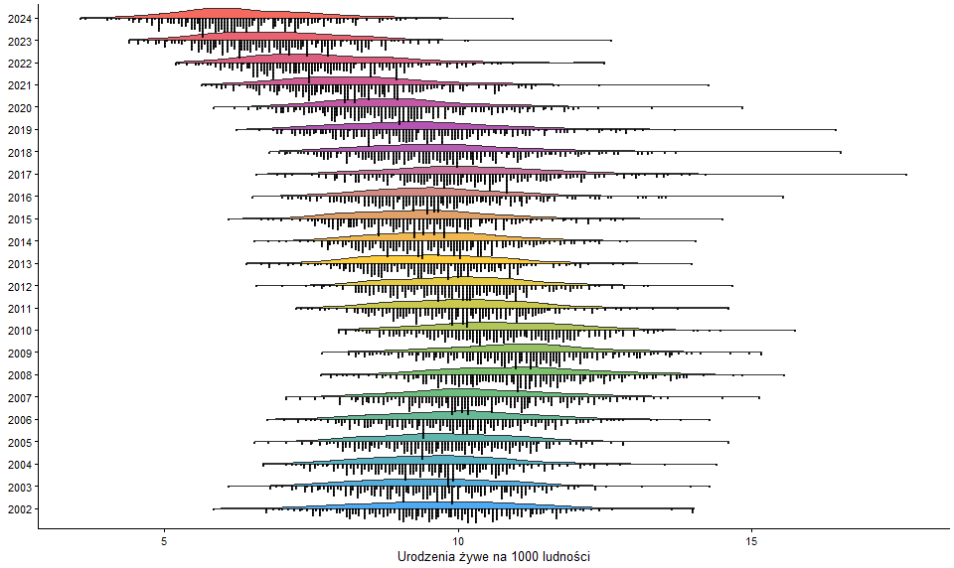
Rys. 6.20. Prognoza liczby urodzeń i zgonów dla województwa śląskiego (na 1000 ludności) na lata 2025–2029

Zgodnie z prognozą w latach 2025–2029 w województwie śląskim należy się spodziewać stabilizacji liczby zgonów na 1000 ludności oraz dalszego zmniejszania się liczby urodzeń żywych na 1000 ludności.

Dla ukazania zmian w rozkładach badanych zmiennych w latach 2002–2024 można się odwołać do wykresu skrzypcowego z punktami.

```
ggplot(pow_od2002, aes(x = factor(Rok), y = Urodzenia_1000, fill = factor(Rok))) +
  see::geom_violindot(
    fill_dots = "black",
```

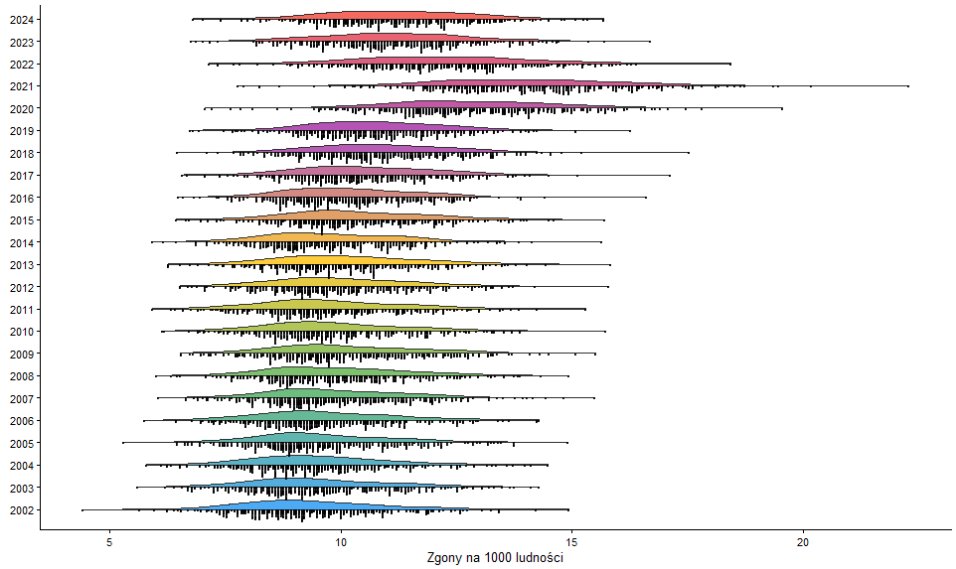
```
  size_dots = 0.5,
  alpha = 0.8,
  dots_alpha = 0.6
) +
labs(
  y = "Urodzenia żywe na 1000 ludności",
  x = NULL,
  fill = "Rok"
) +
ggthemes::scale_fill_material_d() +
coord_flip() +
  cowplot::theme_cowplot(10) +
theme(
  legend.position = "none",
  axis.text.y = element_text(size = 8)
)
```



Rys. 6.21. Liczba urodzeń żywych na 1000 ludności w powiatach według województw w latach 2002–2024

```
ggplot(pow_od2002, aes(x = factor(Rok), y = Zgony_1000, fill = factor(Rok))) +
  see::geom_violindot(
  fill_dots = "black",
  size_dots = 0.5,
  alpha = 0.8,
  dots_alpha = 0.6
) +
labs(
  y = "Zgony na 1000 ludności",
```

```
  x = NULL,
  fill = "Rok"
) +
ggthemes::scale_fill_material_d() +
coord_flip() +
cowplot::theme_cowplot(10) +
theme(
  legend.position = "none",
  axis.text.y = element_text(size = 8)
)
```



Rys. 6.22. Liczba zgonów na 1000 ludności w powiatach według województw w latach 2002–2024

Na rysunkach 6.21 oraz 6.22 przedstawiono odpowiednio liczbę urodzeń żywych na 1000 ludności oraz liczbę zgonów na 1000 ludności w latach 2002–2024. W przypadku zgonów widoczny jest stosunkowo stabilny rozkład wartości, z wyjątkiem wyraźnego wzrostu w okresie epidemii COVID-19. Natomiast dla urodzeń w ostatnich latach obserwuje się systematyczne przesuwanie się rozkładu w kierunku niższych wartości, co wskazuje na spadek liczby urodzeń w przeliczeniu na 1000 mieszkańców.

6.5.4. Charakterystyka wskaźników w ujęciu przestrzennym w 2024 roku

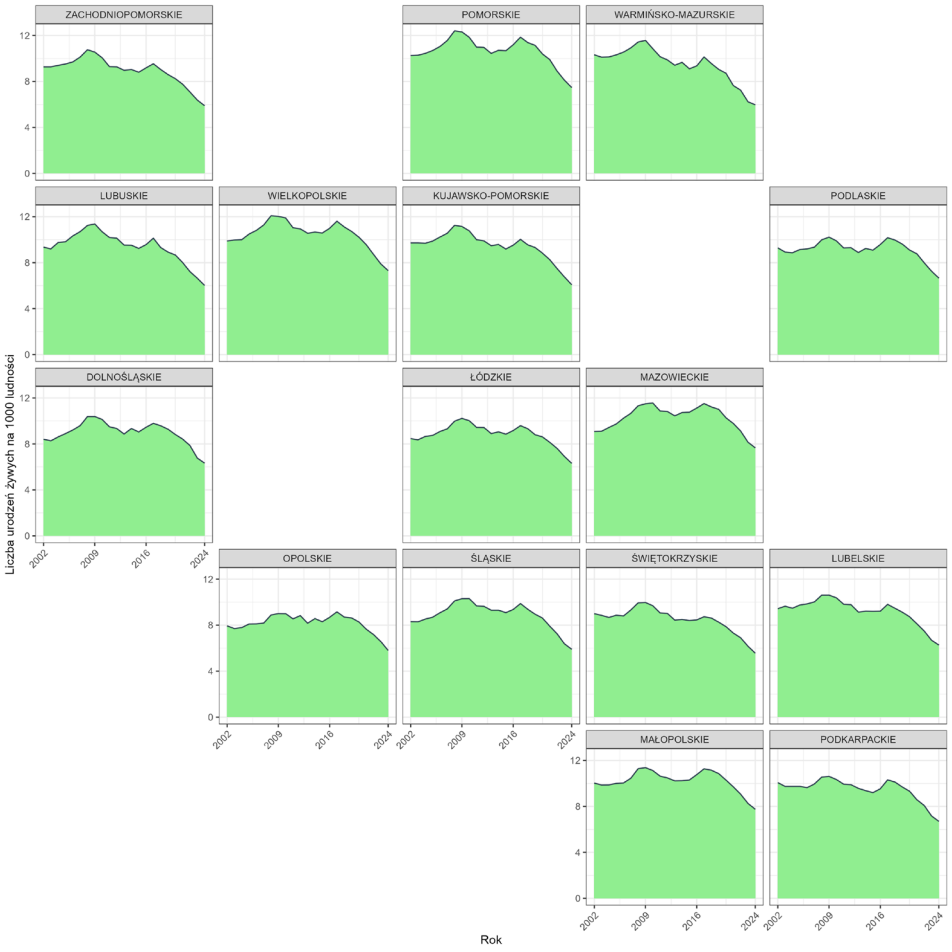
W programie **R** możliwe jest przedstawienie wykresów wybranych zmiennych dla województw z zachowaniem informacji o ich układzie przestrzennym. Można to zrealizować np. z wykorzystaniem biblioteki **`geofacet`**. W pierwszej kolejności zostanie określona siatka województw dla Polski.

```
grid_pl <- data.frame(
  name = c("POMORSKIE", "ZACHODNIOPOMORSKIE", "WARMIŃSKO-MAZURSKIE", "KUJAWSKO-POMORSKIE",
           "LUBUSKIE", "WIELKOPOLSKIE", "PODLASKIE", "MAZOWIECKIE", "ŁÓDZKIE",
           "DOLNOŚLĄSKIE", "OPOLSKIE", "ŚWIĘTOKRZYSKIE", "LUBELSKIE", "ŚLĄSKIE",
           "MAŁOPOLSKIE", "PODKARPACKIE"),
  code = c("PM", "ZP", "WM", "KP", "LB", "WP", "PD", "MZ", "LD",
           "DS", "OP", "SK", "LU", "SL", "MA", "PK"),
  row = c(1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5),
  col = c(3, 1, 4, 3, 1, 2, 5, 4, 3, 1, 2, 4, 5, 3, 4, 5)
)

woj_od2002 <- woj_od2002 %>%
  left_join(grid_pl %>% select(name, code),
            by = c("Województwo" = "name"))
```

Kolejne kody prowadzą do uzyskania wykresów zgodnie z siatką województw dla analizowanych zmiennych.

Na rysunku 6.23 przedstawiono zmiany w czasie liczby urodzeń żywych na 1000 ludności w województwach.

```
ggplot(woj_od2002, aes(x = Rok, y = Urodzenia_1000)) +
  geom_area(fill = "lightgreen") +
  geom_line(color = "#2c3e50") +
  facet_geo(~ Województwo, grid = grid_pl, label = "Województwo") +
  scale_x_continuous(breaks = c(2002, 2009, 2016, 2024), guide = guide_axis(angle = 45)) +
  labs(x = "Rok", y = "Liczba urodzeń żywych na 1000 ludności") +
  theme_bw()
```

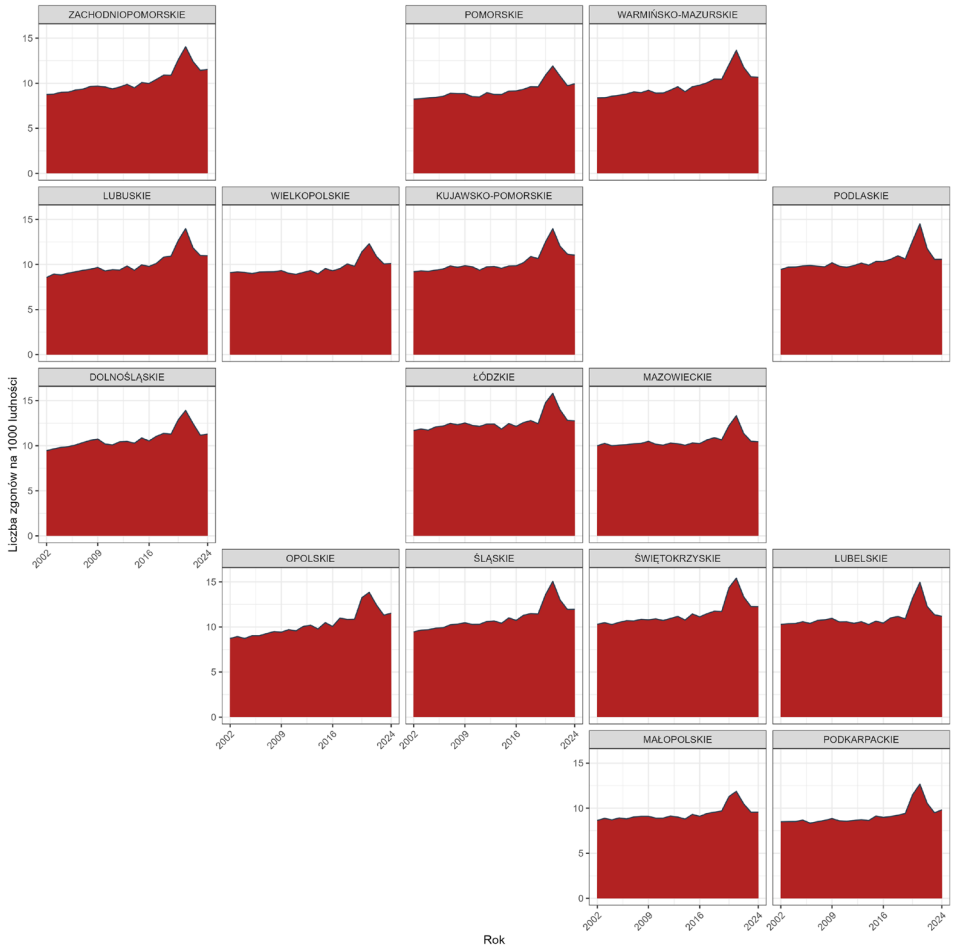


Rys. 6.23. Liczba urodzeń żywych na 1000 ludności w powiatach według województw w latach 2002–2024

We wszystkich województwach zauważalny jest spadek liczby urodzeń żywych na 1000 ludności w ostatnich latach.

Na rysunku 6.24 przedstawiono zmiany w czasie liczby zgonów na 1000 ludności w województwach.

```
ggplot(woj_od2002, aes(x = Rok, y = Zgony_1000)) +
  geom_area(fill = "firebrick") +
  geom_line(color = "#2c3e50") +
  facet_geo(~ Województwo, grid = grid_pl, label = "Województwo") +
  scale_x_continuous(breaks = c(2002, 2009, 2016, 2024), guide = guide_axis(angle =
45)) +
  labs(x = "Rok", y = "Liczba zgonów na 1000 ludności") +
  theme_bw()
```

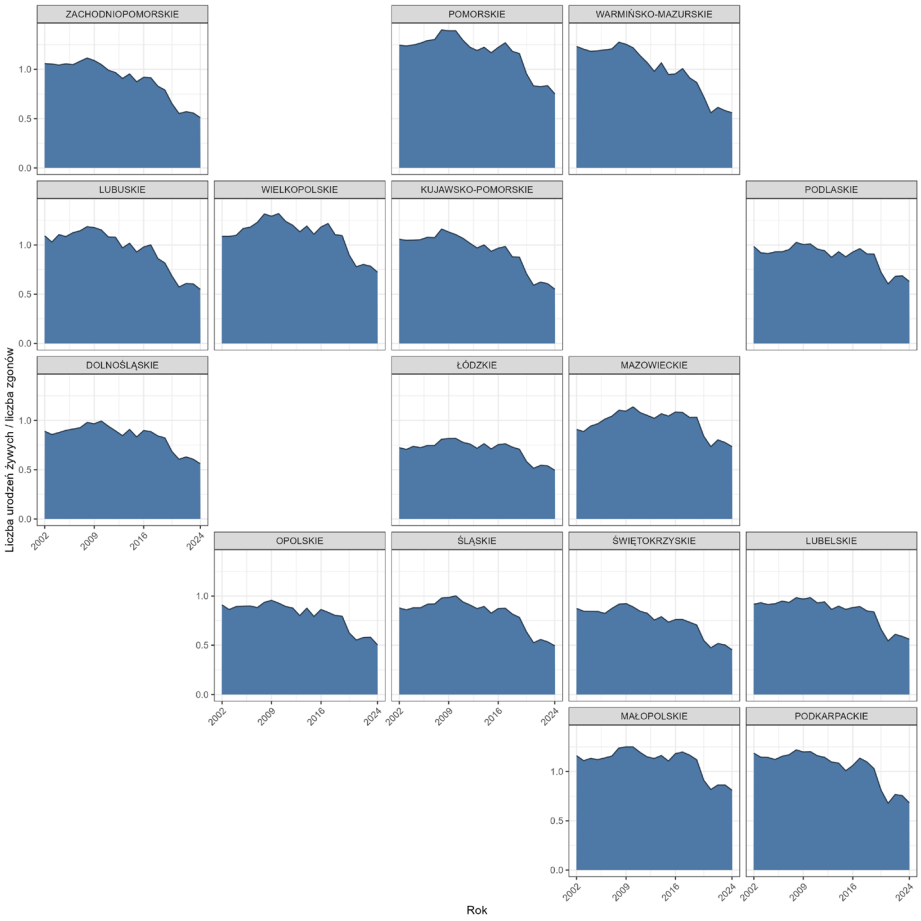


Rys. 6.24. Liczba zgonów na 1000 ludności w latach 2002–2024

Zauważalna jest systematyczna tendencja wzrostowa liczby zgonów na 1000 ludności w ostatnich latach. Znaczny wzrost liczby zgonów był obserwowany w okresie pandemii COVID-19 (lata 2020–2021).

Na rysunku 6.25 przedstawiono zmiany w czasie ilorazu liczby urodzeń żywych do liczby zgonów w województwach.

```
ggplot(woj_od2002, aes(x = Rok, y=Urodzenia_1000/Zgony_1000)) +
  geom_area(fill = "#4e79a7") +
  geom_line(color = "#2c3e50") +
  facet_geo(~ Województwo, grid = grid_pl, label = "Województwo") +
  scale_x_continuous(breaks = c(2002, 2009, 2016, 2024), guide = guide_axis(angle =
45)) +
  labs(x = "Rok", y = "Liczba urodzeń żywych / liczba zgonów") +
  theme_bw()
```

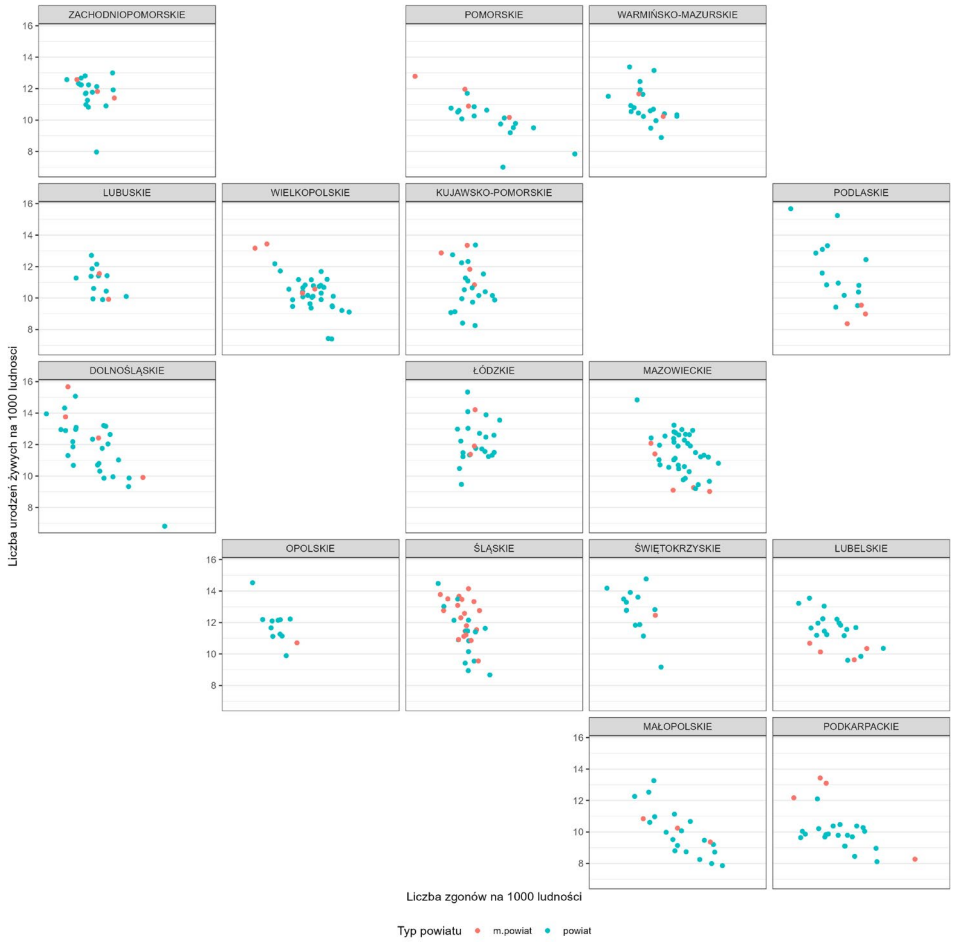


Rys. 6.25. Stosunek liczby urodzeń żywych do liczby zgonów w województwach w latach 2002–2024

Stosunek liczby urodzeń żywych do liczby zgonów w województwach systematycznie się zmniejsza, a w ostatnich latach jest to już blisko 0,5 (dwa razy więcej zgonów niż urodzeń).

Na rysunku 6.26 przedstawiono liczbę urodzeń żywych i zgonów na 1000 ludności w województwach.

```
ggplot(pow_2024, aes(x = Urodzenia_1000, y = Zgony_1000, colour = mz)) +
  geom_point() +
  facet_geo(~Województwo, grid = grid_pl, label = "Województwo") +
  scale_x_continuous(breaks = c(2002, 2009, 2016, 2024), guide = guide_axis(angle =
45)) +
  labs(x = "Liczba zgonów na 1000 ludności", y = "Liczba urodzeń żywych na 1000
ludnosci", color="Typ powiatu") +
  theme_bw() +
  theme(legend.position='bottom')
```



Rys. 6.26. Liczba urodzeń żywych i liczba zgonów na 1000 ludności w powiatach według województw i typu powiatu w 2024 roku

Na rysunku 6.27 przedstawiono liczbę małżeństw na 1000 ludności i liczbę rozwodów na 10 000 ludności w województwach.

```
ggplot(pow_2024, aes(x = Małżeństwa_1000, y = Rozwody_10000,, colour = mz)) +
  geom_point() +
  facet_geo(~Województwo, grid = grid_pl, label = "Województwo") +
  scale_x_continuous(breaks = c(2002, 2009, 2016, 2024), guide = guide_axis(angle =
45)) +
  labs(x = "Liczba małżeństw na 1000 ludności", y = "Liczba rozwodów na 10000
ludności") +
  theme_bw() +
  theme(legend.position='bottom')
```

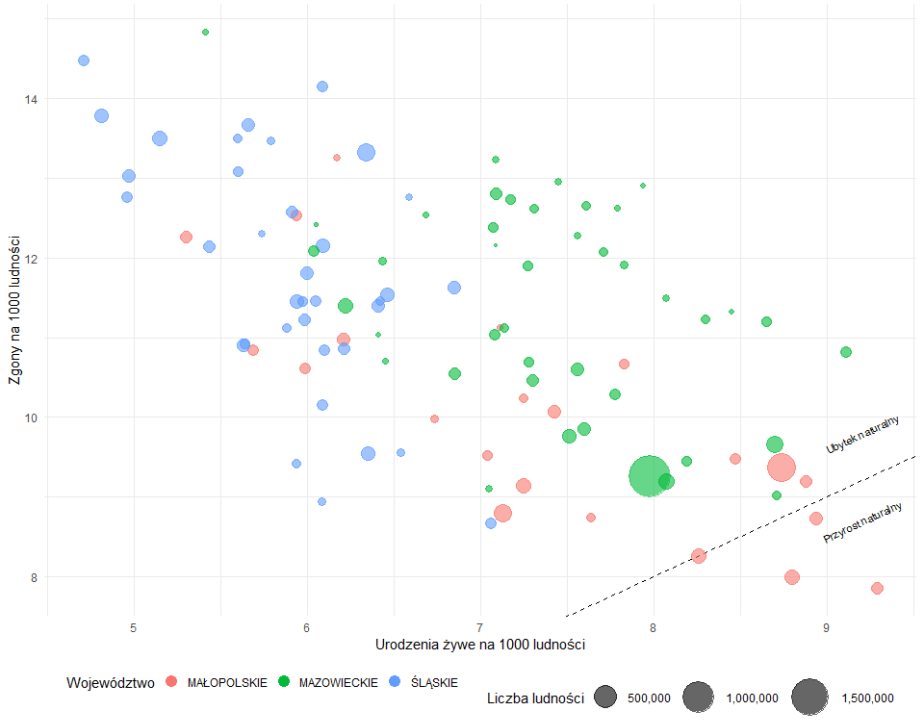
Rys. 6.27. Liczba małżeństw na 1000 ludności i liczba rozwodów na 10 000 ludności w powiatach według województw i typu powiatu w 2024 roku

6.6. Charakterystyka wskaźników w powiatach w wybranych województwach

W tej części analizie zostaną poddane powiaty z trzech województw: mazowieckiego, małopolskiego i śląskiego. Przed wykonaniem analiz wygodne jest wyodrębnienie odpowiednich danych, które realizuje poniższy kod.

```
pow3 <- pow %>% filter( Województwo %in% c('ŚLĄSKIE','MAŁOPOLSKIE',
'MAZOWIECKIE'))
pow3_2024 <- pow3 %>% filter(Rok==2024)
p <- ggplot(pow3_2024, aes(x = Urodzenia_1000, y = Zgony_1000, color = Województwo))
+
```

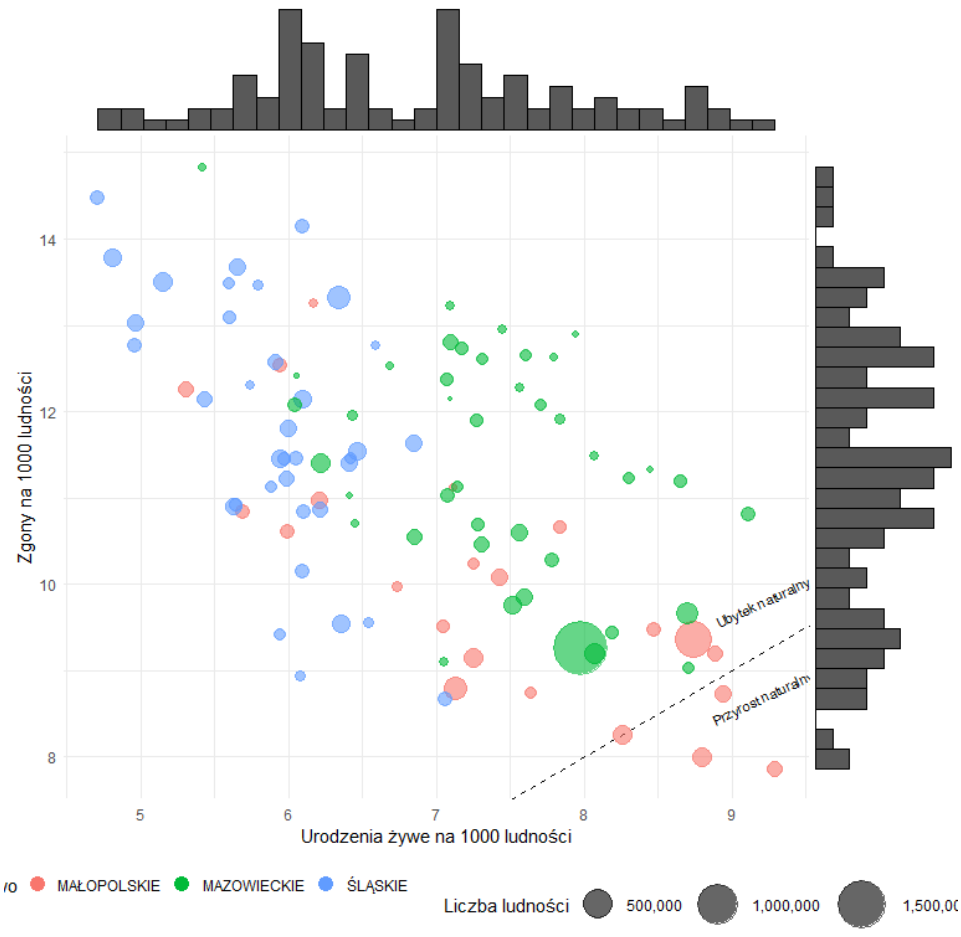
```
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "black") +
  annotate("text", x = 9.2, y = 9.8, label = "Ubytek naturalny", angle = 25, color = "black",
size = 3) +
  annotate("text", x = 9.2, y = 8.7, label = "Przyrost naturalny", angle = 25, color = "black",
size = 3) +
  geom_point(aes(size = Liczba_lud), alpha = 0.6, stroke = 0.5) +
  scale_size(range = c(1, 15), labels = scales::comma_format()) +
  labs(
    x = 'Urodzenia żywe na 1000 ludności',
    y = 'Zgony na 1000 ludności',
    size = 'Liczba ludności',
    color = "Województwo"
  ) +
  theme_minimal() +
  theme(
    legend.position = 'bottom',
    plot.title = element_text(face = "bold"),
    legend.key.size = unit(1, "lines")
  ) +
  guides(color = guide_legend(override.aes = list(size = 4, alpha = 1)))

p
```



Rys. 6.28. Liczba urodzeń żywych i liczba zgonów na 1000 ludności w powiatach wybranych województw w 2024 roku

Na rysunku 6.28 przedstawiono zależność pomiędzy liczbą urodzeń żywych a liczbą zgonów (obie wartości w przeliczeniu na 1000 ludności) w badanych województwach w 2024 roku. Każdy punkt odpowiada jednemu powiatowi, a jego kolor oznacza przynależność do konkretnego województwa. Wielkość koła odzwierciedla liczbę ludności danego powiatu – im większa liczba mieszkańców, tym większy rozmiar koła na wykresie. W ten sposób można jednocześnie porównać natężenie urodzeń i zgonów oraz skalę populacji w różnych powiatach. Jedynie w województwie małopolskim w 2024 roku wystąpiły powiaty, w których liczba urodzeń żywych na 1000 ludności była większa niż liczba zgonów na 1000 ludności (przyrost naturalny). W województwie śląskim i mazowieckim wszystkie powiaty charakteryzowały się ubytkiem naturalnym.

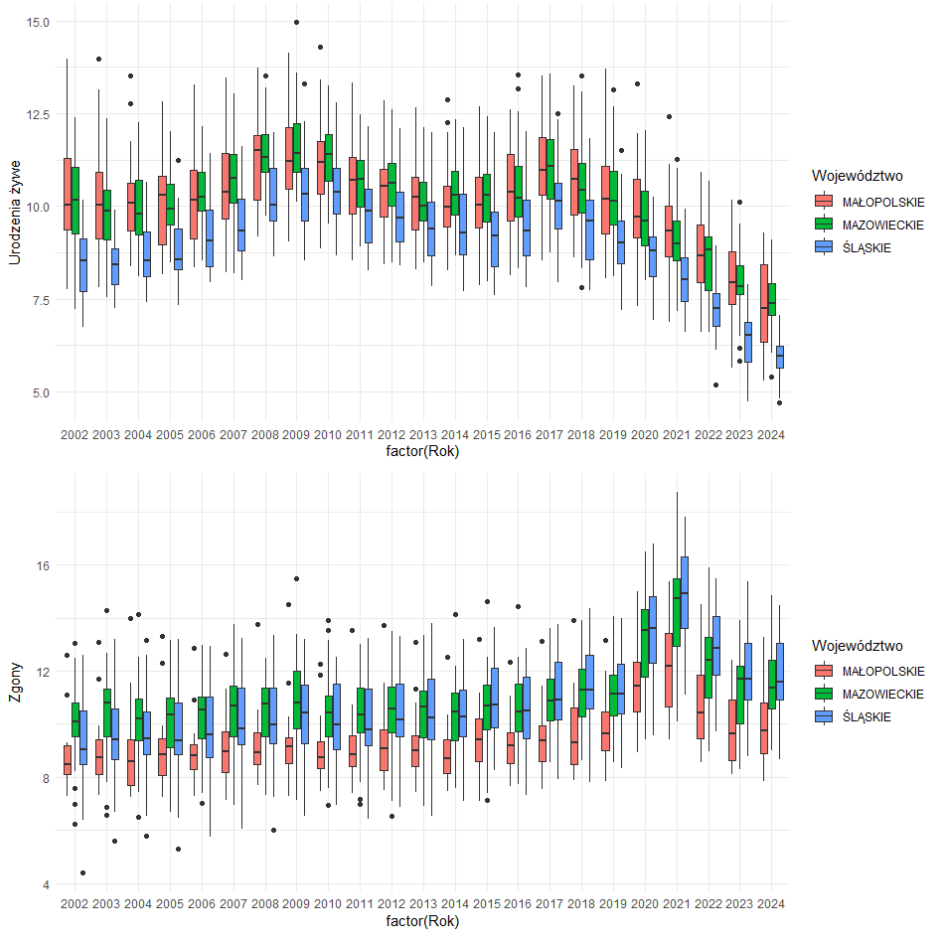
```
ggExtra::ggMarginal(p, type = "histogram")
```



Rys. 6.29. Liczba urodzeń żywych i liczba zgonów w powiatach według województw w 2024 roku

Rysunek 6.29 przedstawia podobne informacje jak rys. 6.28, ale dodatkowo uwzględnia rozkłady brzegowe obu zmiennych w postaci histogramów.

```
p1=pow3 %>% filter (Rok>2001) %>%
  ggplot( aes(x=factor(Rok), y=Urodzenia_1000, fill=Województwo)) +
  labs(y='Urodzenia żywe') +
  geom_boxplot()

p2=pow3 %>% filter (Rok>2001) %>%
  ggplot( aes(x=factor(Rok), y=Zgony_1000, fill=Województwo)) +
  labs(y='Zgony') +
  geom_boxplot()

p1/p2
```

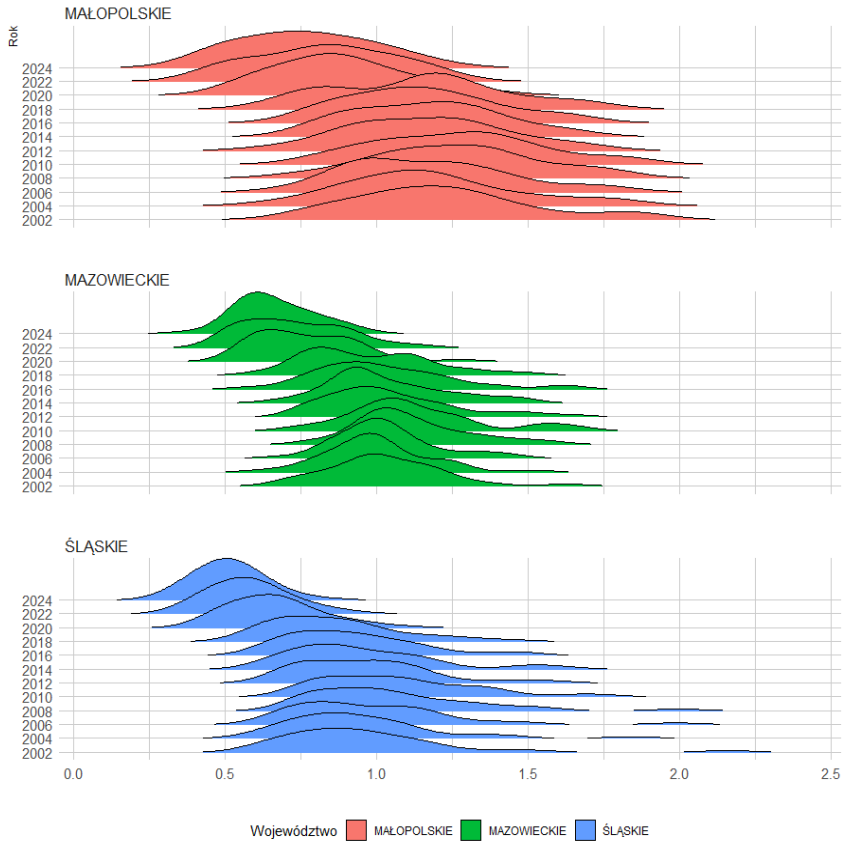


Rys. 6.30. Liczba urodzeń żywych i liczba zgonów na 1000 ludności w powiatach w wybranych województwach w latach 2002–2024

Na rysunku 6.30 przedstawiono zestawienie zmian liczby urodzeń żywych oraz liczby zgonów (obie wartości w przeliczeniu na 1000 ludności) w powiatach wybranych województw w latach 2002–2024. Górny wykres prezentuje rozkład liczby urodzeń żywych na 1000 mieszkańców, natomiast dolny wykres – rozkład liczby zgonów na 1000 mieszkańców. Każde pudełko (boxplot) przedstawia rozkład wartości w powiatach danego województwa w określonym roku. Kolory odpowiadają poszczególnym województwom.

Wykresy na rys. 6.30 umożliwiają porównanie zmian w rozkładzie liczby urodzeń i zgonów w czasie oraz pomiędzy województwami. Można zaobserwować, że w województwie śląskim mediany liczby urodzeń żywych na 1000 ludności są mniejsze niż w obu pozostałych województwach we wszystkich latach.

```
pow3 %>%  filter (Rok>2001 & Rok == 2*floor(Rok/2)) %>%
  ggplot( aes(y=factor(Rok), x=Urodzenia_1000/Zgony_1000,  fill=Województwo)) +
  ggridges::geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
  labs(x='Liczba urodzin do liczby zgonów',y='Rok') +
  ggridges::theme_ridges() +
  theme(
    legend.position="none",
    panel.spacing = unit(0.1, "lines"),
    strip.text.x = element_text(size = 8)    ) +
 hrbrthemes::theme_ipsum()+
  xlab("") +
  theme(legend.position ='bottom' )+
  facet_wrap(~Województwo,ncol=1)
```



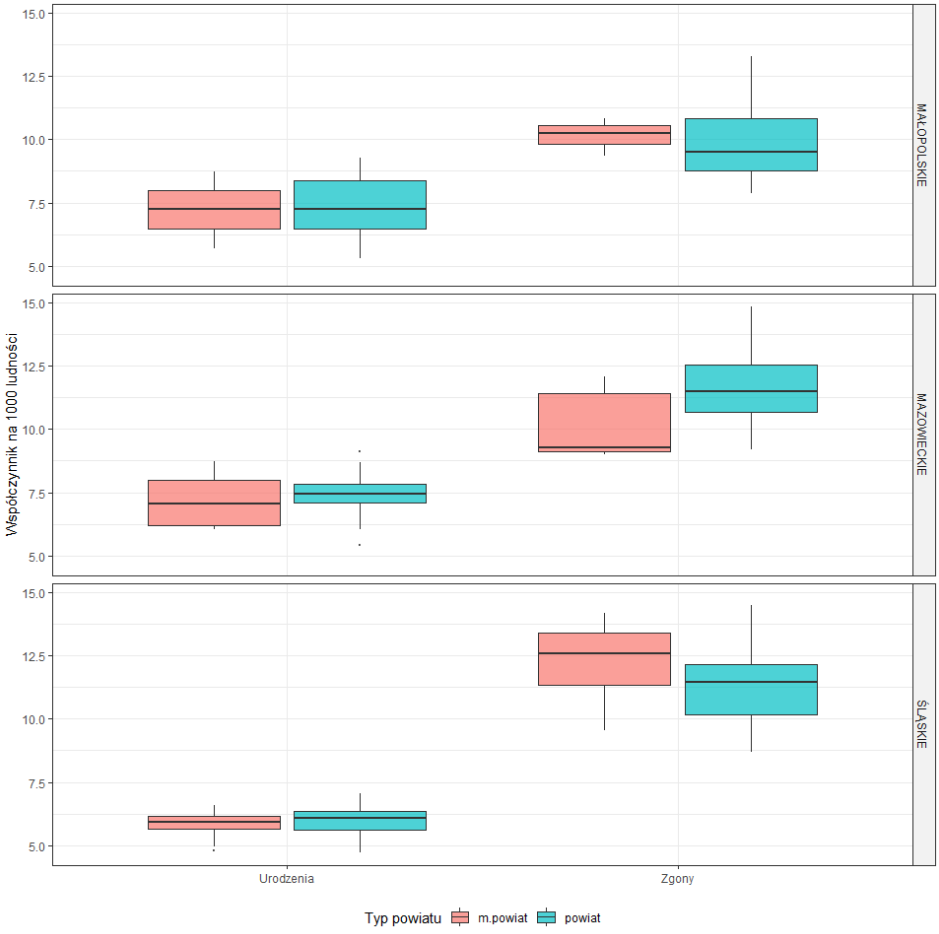
Rys. 6.31. Wskaźnik liczby urodzeń do liczby zgonów w powiatach wybranych województw w latach 2002–2024

Na rysunku 6.31 przedstawiono rozkład ilorazu liczby urodzeń do liczby zgonów w powiatach wybranych województw od 2002 do 2024 roku. Wykresy grzbietowe (*ridge plots*) pokazują, jak kształtował się ten wskaźnik w powiatach poszczególnych województw i latach. Wykres umożliwia obserwację zmian rozkładu tego wskaźnika w czasie oraz porównanie sytuacji demograficznej pomiędzy województwami. Dobrze widoczne są niskie wartości w ostatnich latach w województwie śląskim, a relatywnie większe w małopolskim.

Dla zaobserwowania ewentualnych różnic badanych wskaźników w powiatach i miastach na prawach powiatu można wykonać kod:

```
pow_long <- pow3_2024 %>%
  tidyr::pivot_longer(cols = c(Urodzenia_1000, Zgony_1000),
            names_to = "Wskaźnik",
            values_to = "Wartość") %>%
  mutate(Wskaźnik = recode(Wskaźnik,
                    "Urodzenia_1000" = "Urodzenia",
```

```
                    "Zgony_1000" = "Zgony"))

ggplot(pow_long, aes(x = Wskaźnik, y = Wartość, fill = mz)) +
  facet_grid(Województwo ~ .) +
  geom_boxplot(alpha = 0.7, outlier.size = 0.5) +
  labs(
    y = "Współczynnik na 1000 ludności",
    x = NULL,
    fill = "Typ powiatu"
  ) +
  theme_bw() +
  theme(
    legend.position = "bottom",
    strip.background = element_rect(fill = "grey95")
  )
```



Rys. 6.32. Liczba urodzeń żywych i liczba zgonów w powiatach według województw w 2024 roku

Na rysunku 6.32 przedstawiono rozkład liczby urodzeń żywych (górny wykres) oraz liczby zgonów (dolny wykres) na 1000 mieszkańców według typu powiatu wybranych województw w 2024 roku. Każdy wykres został podzielony na panele odpowiadające poszczególnym województwom (*facet_wrap*). Taka prezentacja umożliwia porównanie rozkładów liczby urodzeń i zgonów pomiędzy powiatami i miastami na prawach powiatu w obrębie każdego województwa oraz identyfikację powiatów o wartościach skrajnych lub odstających.

6.7. Wnioskowanie statystyczne w modelu permutacyjnym

W poprzednim rozdziale przedstawiono zasady wnioskowania statystycznego w modelu permutacyjnym. Poniżej przedstawiono rezultaty takiego wnioskowania dla pozyskanych powyżej danych z Banku Danych Lokalnych. Testy permutacyjne stosunkowo rzadko są wykorzystywane w analizach demograficznych. Orwat-Acedańska i Kończak (2025) wykorzystują testy permutacyjne dla grup niezależnych i zależnych do oceny statystycznej istotności różnic we współczynnikach dzietności ogółem w województwach. Z kolei Holland i in. (2025) stosują wnioskowanie permutacyjne do analizy zgonów spowodowanych zapaleniem płuc.

Dla przeprowadzenia testów permutacyjnych należy załadować odpowiednie biblioteki. W analizach zostaną wykorzystane następujące biblioteki: **coin**, **wPerm** i **infer**. We wszystkich rozważanych hipotezach przyjęto poziom istotności $\alpha = 0{,}05$.

```
library(coin)
library(wPerm)
library(infer)
```

6.7.1. Porównanie dwóch grup niezależnych

W pierwszej kolejności zostanie przeprowadzona weryfikacja hipotezy o równości wartości oczekiwanych liczby urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego w 2024 roku.

Przeprowadzenie takiego wnioskowania w klasycznym modelu (Neymana-Pearsona) nie jest możliwe, ponieważ dostępne są wszystkie dane dla wszystkich powiatów z obu województw (Sokołowski, 2004). Model populacyjny wymaga precyzyjnego określenia populacji oraz odpowiedniego doboru próby. W modelu

permutacyjnym możliwe jest porównanie dwóch zbiorów (nie prób) i ewentualne wskazanie statystycznie istotnych różnic.

Hipotezy można zapisać w następującej postaci:

$$H_0: \mu_M = \mu_Ś$$
$$H_1: \mu_M \neq \mu_Ś$$

Przed przeprowadzeniem testu należy odpowiednio przygotować dane. Utworzony zostanie zbiór **ur_1000**, a jest to realizowane za pomocą poniższego kodu.

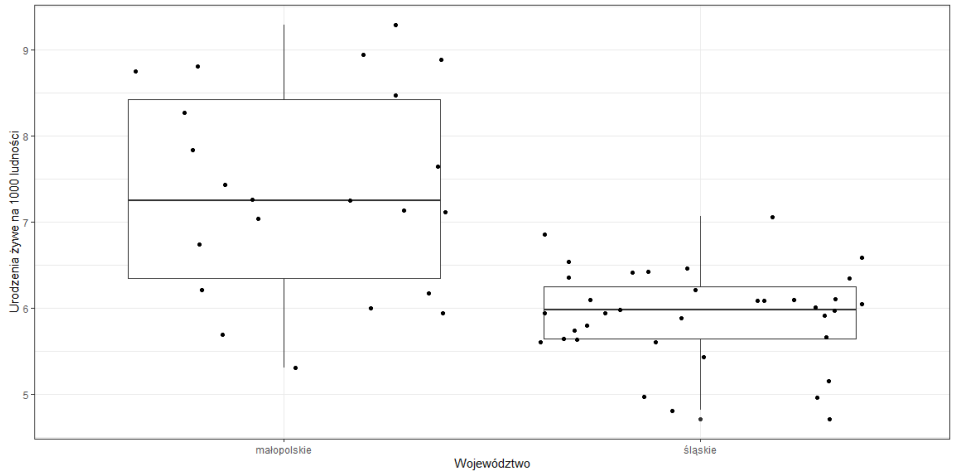
```
mal <- pow3_2024 %>% filter(Województwo == 'MAŁOPOLSKIE')
sla <- pow3_2024 %>% filter(Województwo == 'ŚLĄSKIE')

ur_1000 <- data.frame(
  w_ur = c(mal$Urodzenia_1000, sla$Urodzenia_1000),
  woj = factor(
    rep(c('małopolskie', 'śląskie'),
        times = c(nrow(mal), nrow(sla)))
  )
)
```

Fragment otrzymanego zbioru danych przedstawiono poniżej:

```
ur_1000[c(1:3,23:25),]
#     w_ur          woj
# 1   8.47 małopolskie
# 2   7.13 małopolskie
# 3   6.17 małopolskie
# 23  6.35     śląskie
# 24  6.09     śląskie
# 25  6.85     śląskie
```

Na rysunku 6.33 przedstawiono liczbę urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego w 2024 roku. Zauważalne jest, że w powiatach województwa małopolskiego średnia liczba urodzeń żywych na 1000 ludności jest większa niż w powiatach województwa śląskiego. Pozostaje odpowiedź na pytanie, czy ta różnica jest statystycznie istotna.

```
ggplot(ur_1000, aes(x = woj, y = w_ur)) +
  geom_boxplot() +
  geom_jitter() +
  labs(
    x = 'Województwo',
    y = 'Urodzenia żywe na 1000 ludności'
  )+
  theme_bw()
```



Rys. 6.33. Liczba urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego w 2024 roku

Test można zrealizować z wykorzystaniem funkcji *oneway_test()* z pakietu **coin**. Funkcja ta pozwala na przeprowadzenie testu permutacyjnego porównania wartości oczekiwanych dla dwóch lub większej liczby grup. Realizuje to następująca komenda:

```
oneway_test(w_ur ~ woj, data = ur_1000)
#
#   Asymptotic Two-Sample Fisher-Pitman Permutation Test
#
# data:  w_ur by woj (małopolskie, śląskie)
# Z = 4.9017, p-value = 9.499e-07
# alternative hypothesis: true mu is not equal to 0
```

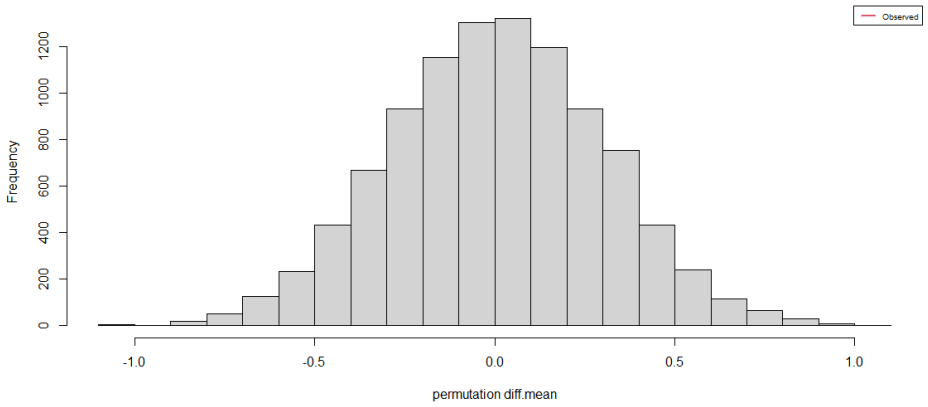
Otrzymana p-wartość wskazuje na statystycznie istotne różnice pomiędzy średnią liczbą urodzeń na 1000 ludności w powiatach obu województw. Potwierdzenie tego rezultatu można uzyskać, stosując test dokładny (Kończak, 2016).

```
oneway_test(w_ur ~ woj, data = ur_1000, distribution = "exact")
#
#   Exact Two-Sample Fisher-Pitman Permutation Test
#
# data:  w_ur by woj (małopolskie, śląskie)
# Z = 4.9017, p-value = 4.774e-08
# alternative hypothesis: true mu is not equal to 0
```

Do porównania liczby urodzeń żywych na 1000 ludności w powiatach dwóch województw można także wykorzystać funkcję *perm.ind.loc()* z pakietu **wPerm**. Poniżej przedstawiono porównania wartości oczekiwanych i median.

- Porównanie wartości oczekiwanych współczynników urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego

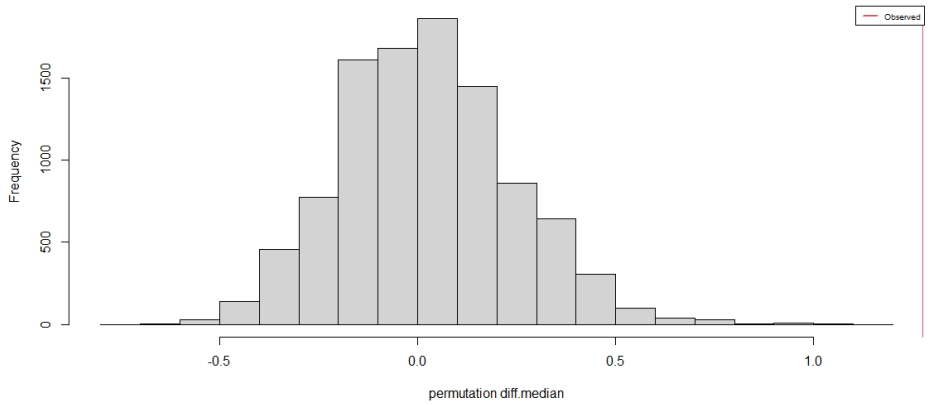
```
perm.ind.loc(ur_1000$w_ur, ur_1000$woj, mean)
#
#
#  RESULTS OF PERMUTATION INDEPENDENT TWO-SAMPLE LOCATION TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY Variable       Pop.1   Pop.2 n.1 n.2 Statistic Observed
#  STATISTICS     w_ur małopolskie śląskie  22  36 diff.mean 1.449949
#
#  HYPOTHESIS      Null Alternative   P.value
#        TEST identical     shifted P < 0.001
```



Rys. 6.34. Wyniki testu permutacyjnego dla równości wartości oczekiwanych współczynników urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego

- Porównanie median współczynników urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego

```
perm.ind.loc(ur_1000$w_ur, ur_1000$woj, median)
#
#
#  RESULTS OF PERMUTATION INDEPENDENT TWO-SAMPLE LOCATION TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY Variable       Pop.1   Pop.2 n.1 n.2   Statistic Observed
#  STATISTICS     w ur małopolskie śląskie  22  36 diff.median    1.275
#
#  HYPOTHESIS      Null Alternative   P.value
#        TEST identical     shifted P < 0.001
```

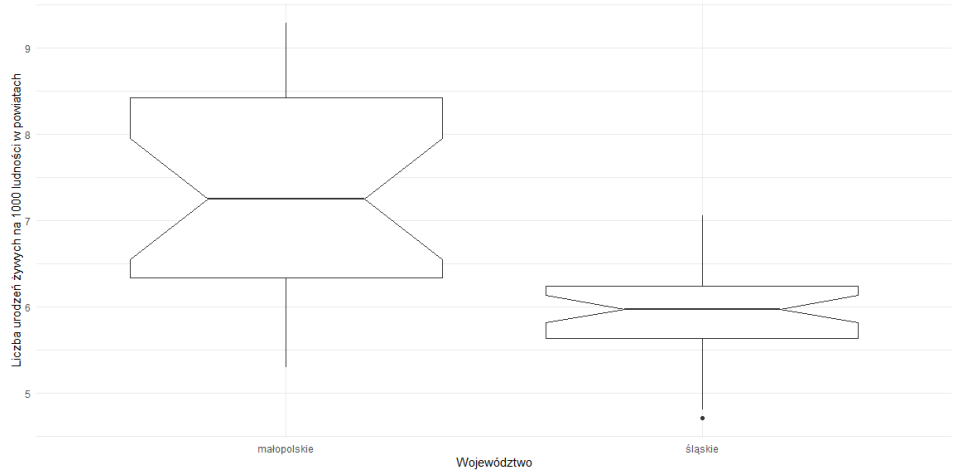


Rys. 6.35. Wyniki testu permutacyjnego dla równości median współczynników urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego

Na rysunkach 6.34 i 6.35 przedstawiono empiryczny rozkład statystyk testowych, odpowiednio różnic wartości oczekiwanych oraz median, dla dwóch testów. W obu przypadkach należy odrzucić hipotezę o równości, odpowiednio wartości oczekiwanych oraz median.

Na rysunku 6.36 przedstawiono wykresy pudełkowe z nacięciami współczynnika urodzeń żywych na 1000 ludności w powiatach województw małopolskiego i śląskiego.

```
ggplot(ur_1000, aes(woj, w_ur)) +
  geom_boxplot(notch = TRUE) +
  labs(x = "Województwo", y = "Liczba urodzeń żywych na 1000 ludności w powiatach")
```



Rys. 6.36. Liczba urodzeń żywych na 1000 ludności w powiatach według województw

Wykresy pudełkowe z nacięciami wykorzystują „nacięcie” pudełka wokół mediany. Wycięcia są przydatne w dostarczaniu wskazówek dotyczących istotności różnicy median. Wcięcia pudełek na rys. 6.36 wskazują, że występują statystycznie istotne różnice median dla województw małopolskiego i śląskiego, co potwierdził przeprowadzony powyżej test dla równości median.

Wcześniej przedstawiono rozkłady urodzeń żywych na 1000 ludności w powiatach i miastach na prawach powiatów (por. rys. 6.32). Dla odpowiedzi na pytanie, czy pomiędzy powiatami i miastami na prawach powiatu występują statystycznie istotne różnice współczynników urodzeń żywych na 1000 ludności, należy zweryfikować hipotezę:

$$H_0: \mu_{powiat} = \mu_{miasto.pow}$$

wobec hipotezy alternatywnej:

$$H_1: \mu_{powiat} \neq \mu_{miasto.pow}$$

Dla przeprowadzenia testu równości wartości oczekiwanych w tych grupach można wykorzystać komendę:

```
perm.ind.loc(pow_2y$Urodzenia_1000, pow_2y$mz, mean)
```

```
#
#
#  RESULTS OF PERMUTATION INDEPENDENT TWO-SAMPLE LOCATION TEST
#  BASED ON 9999 REPLICATIONS
#
#     SUMMARY       Variable    Pop.1  Pop.2 n.1 n.2 Statistic   Observed
#  STATISTICS Urodzenia_1000 m.powiat powiat 132 628 diff.mean -0.2562372
#
#  HYPOTHESIS      Null Alternative P.value
#        TEST identical     shifted   0.176
```

Otrzymany rezultat (p-wartość = 0,372) prowadzi do stwierdzenia braku podstaw do odrzucenia hipotezy H_0.

6.7.2. Porównanie trzech grup

Weryfikacja powyżej przedstawionych hipotez dotyczyła porównania dwóch populacji. Jeśli porównywana jest większa liczba populacji, to należy się odwołać do permutacyjnej wersji testu ANOVA. Porównane zostaną rozkłady współczynników urodzeń żywych na 1000 ludności w powiatach województw małopolskiego, śląskiego i mazowieckiego. Na wstępie zostaną wyodrębnione odpowiednie obserwacje.

```
ur_1000 <- pow_2024 %>%
  dplyr::select(Województwo, Urodzenia_1000) %>%
  filter(Województwo %in% c('MAŁOPOLSKIE','ŚLĄSKIE','MAZOWIECKIE'))
ur_1000$Województwo = factor(ur_1000$Województwo)
```

Fragment otrzymanego zbioru danych przedstawiono poniżej:

```
ur_1000[c(1:3,23:25,59:61),]
# # A tibble: 9 × 2
#   Województwo Urodzenia_1000
#   <fct>                <dbl>
# 1 MAŁOPOLSKIE           8.47
# 2 MAŁOPOLSKIE           7.13
# 3 MAŁOPOLSKIE           6.17
# 4 ŚLĄSKIE               6.35
# 5 ŚLĄSKIE               6.09
# 6 ŚLĄSKIE               6.85
# 7 MAZOWIECKIE           7.98
# 8 MAZOWIECKIE           6.85
# 9 MAZOWIECKIE           7.56
```

Hipoteza H_0 głosząca identyczność rozkładów badanej zmiennej w trzech wskazanych województwach i hipoteza alternatywna H_1 mogą być zapisane:

$$H_0: \mu_{małopolskie} = \mu_{śląskie} = \mu_{mazowieckie}$$
$$H_1: \sim H_0$$

Weryfikacja hipotezy H_0 może być przeprowadzona następująco:

```
oneway_test(Urodzenia_1000 ~ Województwo, data = ur_1000)
#
#   Asymptotic K-Sample Fisher-Pitman Permutation Test
#
# data:  Urodzenia_1000 by
#    Województwo (MAŁOPOLSKIE, MAZOWIECKIE, ŚLĄSKIE)
# chi-squared = 42.915, df = 2, p-value = 4.798e-10
```

Otrzymana p-wartość prowadzi do odrzucenia hipotezy H_0, a więc można twierdzić, że wartości oczekiwane liczby urodzeń żywych na 1000 ludności w powiatach trzech województw nie są jednakowe.

6.7.3. Porównanie dwóch grup zależnych

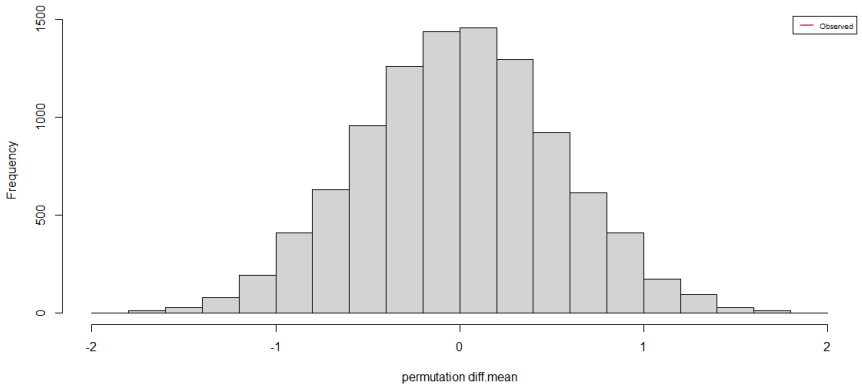
Dla porównania wartości oczekiwanych liczby urodzeń żywych na 1000 ludności w powiatach województwa śląskiego w dwóch latach 2019 i 2024 należy

się odwołać do testu dla danych skojarzonych. Hipotezy w tym przypadku można zapisać następująco:

$$H_0: \mu_{2019} = \mu_{2024}$$
$$H_1: \mu_{2019} \neq \mu_{2024}$$

Realizację tego testu dla wskazanego przypadku prezentuje poniższy kod.

```
p_19_24w <- pow_2y %>% dplyr::select(Powiat,Rok,Urodzenia_1000, Województwo)
%>%
  filter(Województwo=="ŚLĄSKIE") %>%
  tidyr::pivot_wider(
    names_from = Rok,
    values_from = Urodzenia_1000,
    values_fn = mean)
x=p_19_24w$'2019'
y=p_19_24w$'2024'

perm.paired.loc(x, y, mean)
#
#
#   RESULTS OF PERMUTATION PAIRED LOCATION TEST
#   BASED ON 9999 REPLICATIONS
#
#      SUMMARY Pop.1 Pop.2  n Statistic Observed
#   STATISTICS     x     y 36 diff.mean 3.110278
#
#   HYPOTHESIS      Null Alternative   P.value
#         TEST identical     shifted P < 0.001
```



Rys. 6.37. Wyniki testu permutacyjnego dla 2 prób zależnych

Empiryczny rozkład statystyki testowej przedstawia rys. 6.37. Otrzymana p-wartość prowadzi do odrzucenia hipotezy H_0, a więc można twierdzić, że wartości oczekiwane liczby urodzeń żywych na 1000 ludności w powiatach województwa śląskiego w dwóch latach 2019 i 2024 nie są jednakowe.

6.7.4. Testowanie jednorodności struktur

Test permutacyjny może być wykorzystany do testowania jednorodności struktur. W takim przypadku dane powinny być zamieszczone w tablicy wielodzielczej.

Hipotezy w teście jednorodności struktur można zapisać następująco:

H_0: Struktury w grupach kwintylowych w trzech województwach są jednakowe

H_1: Struktury w grupach kwintylowych w trzech województwach są różne

Konstrukcja odpowiedniej tablicy wielodzielczej przebiega następująco:

```
pow3_2024 <- pow3 %>% filter(Rok==2024)

qu=quantile(pow3_2024$Urodzenia_1000, probs = c(0.2, 0.4, 0.6,0.8), na.rm = TRUE)
pow3_2024 <- pow3_2024 %>%
  mutate(
    Urodz_q4 = cut(Urodzenia_1000,
                   breaks = c(-Inf, qu[1], qu[2], qu[3], qu[4], Inf),
                   labels = c("q1", "q2", "q3", "q4","q5"),
                   right = TRUE)
  )
tab=table(pow3_2024$Województwo, pow3_2024$Urodz_q4)
```

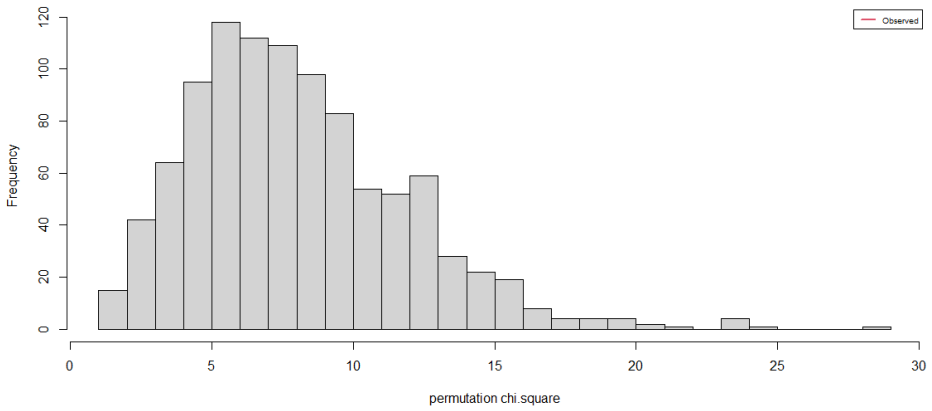
Tabela 6.2. Rozkład urodzeń żywych na 1000 ludności w grupach kwintylowych w powiatach w 2024 roku według województw

Województwo	Q1	Q2	Q3	Q4	Q5
Małopolskie	3	3	3	5	8
Mazowieckie	1	4	10	15	12
Śląskie	17	13	6	0	0

W tabeli przedstawiono rozkład liczby powiatów ze względu na liczbę urodzeń żywych na 1000 ludności w grupach kwintylowych. Zauważalne jest, że w zdecydowanej większości powiatów województwa śląskiego powiaty znajdują się w dwóch pierwszych wyróżnionych grupach (mała liczba urodzeń żywych na 1000 ludności). Natomiast w województwie mazowieckim zdecydowana większość powiatów znajduje się w grupach od trzeciej do piątej (relatywnie wysoka liczba urodzeń żywych na 1000 ludności). To spostrzeżenie sugeruje wniosek, że struktury nie są jednakowe w badanych województwach. Dla potwierdzenia wskazanego spostrzeżenia niezbędne jest przeprowadzenie testu statystycznego.

Dla tak określonej tablicy nie jest jednak możliwe zastosowanie klasycznego testu jednorodności chi-kwadrat (Domański, 1979) ze względu na występowanie liczebności oczekiwanych mniejszych od 5. W takim przypadku dobrym rozwiązaniem jest zastosowanie testu permutacyjnego. To zadanie realizuje poniższy kod.

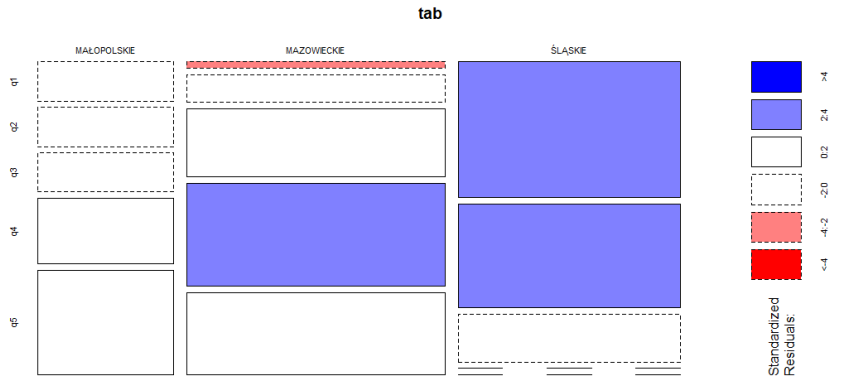
```
tab.df=as.data.frame(tab)
tab.df
#           Var1 Var2 Freq
# 1   MAŁOPOLSKIE   q1    3
# 2   MAZOWIECKIE   q1    1
# 3       ŚLĄSKIE   q1   17
# 4   MAŁOPOLSKIE   q2    3
# 5   MAZOWIECKIE   q2    4
# 6       ŚLĄSKIE   q2   13
# 7   MAŁOPOLSKIE   q3    3
# 8   MAZOWIECKIE   q3   10
# 9       ŚLĄSKIE   q3    6
# 10  MAŁOPOLSKIE   q4    5
# 11  MAZOWIECKIE   q4   15
# 12      ŚLĄSKIE   q4    0
# 13  MAŁOPOLSKIE   q5    8
# 14  MAZOWIECKIE   q5   12
# 15      ŚLĄSKIE   q5    0
perm.hom.test(tab.df, "flat", "Self-concept", 999)
#
#
#   RESULTS OF PERMUTATION HOMOGENEITY TEST
#   BASED ON 999 REPLICATIONS
#
#      SUMMARY     Variable   n  Statistic Observed
#   STATISTICS Self-concept 100 chi.square 51.81482
#
#   HYPOTHESIS        Null    Alternative P.value
#         TEST homogeneous nonhomogeneous   0.001
```



Rys. 6.38. Wyniki testu permutacyjnego jednorodności struktur – rozkład empiryczny i wartość obserwowana statystyki

Na rysunku 6.38 przedstawiono wyniki testu permutacyjnego dla jednorodności struktur. W efekcie należy odrzucić hipotezę H_0. Można twierdzić, że struktury nie są jednakowe. Potwierdzenie otrzymanego wniosku daje rys. 6.39, gdzie przedstawiono wykres mozaikowy na podstawie danych z tabeli 6.2.

```
plot(tab , shade=TRUE)
```



Rys. 6.39. Wizualizacja wyników testu permutacyjnego niezależności

Powyższy kod konstruuje wykres mozaikowy (mosaic plot), który wizualizuje dane tabelaryczne dwu- lub wielowymiarowe (Friendly, 1994; Kończak, 2025). Zauważalne jest, że w województwie śląskim jest więcej powiatów w grupach kwintylowych (qu1 i qu2) o małej liczbie urodzeń żywych na 1000 ludności niż wskazują liczebności oczekiwane. Natomiast w województwie mazowieckim zbyt mało jest powiatów w grupie qu1, a zbyt dużo w grupie qu4.

6.7.5. Test dla współczynnika korelacji

Dla zbadania statystycznej istotności zależności pomiędzy analizowanymi wskaźnikami może być przeprowadzony test dla współczynnika korelacji. Weryfikowana będzie hipoteza H_0 głosząca, że pomiędzy liczbą urodzeń żywych na 1000 ludności a liczbą zgonów na 1000 ludności w powiatach w 2024 roku nie ma zależności. Formalnie może to być zapisane:

$$H_0: \rho = 0$$
$$H_1: \rho < 0$$

Zastosowany zostanie test permutacyjny *perm.relation* z pakietu **`wPerm`**.

```
woj_ur_zg=woj %>% filter(Rok == 2024)
cor(woj_ur_zg$Urodzenia_1000, woj_ur_zg$Zgony_1000)
# [1] -0.7530437
```

W wyniku obliczeń należy stwierdzić brak podstaw do odrzucenia hipotezy H_0.

Taki test dla współczynników korelacji rang Spearmana i Kendalla można wykonać następująco:

```
perm.relation(woj_ur_zg$Urodzenia_1000, woj_ur_zg$Zgony_1000, "kendall", "less")
perm.relation(woj_ur_zg$Urodzenia_1000, woj_ur_zg$Zgony_1000, "spearman", "less")
```

Zakończenie

Przeprowadzone w niniejszej monografii rozważania wskazują na konieczność szerszego wykorzystania wizualizacji danych na potrzeby wnioskowania statystycznego. Wizualizacja danych przestaje pełnić funkcję wyłącznie opisową, ograniczoną do wstępnej eksploracji czy prezentacji wyników, a staje się ważnym narzędziem procesu wnioskowania statystycznego. W dobie powszechnej dostępności zaawansowanych narzędzi obliczeniowych graficzna reprezentacja rozkładów teoretycznych i empirycznych, obszarów krytycznych czy p-wartości pozwala na głębsze zrozumienie mechanizmów losowych oraz ocenę założeń modelowych. Wizualizacja pozwala przekształcić wskaźniki liczbowe w intuicyjne obrazy, co nie tylko zwiększa transparentność procesu decyzyjnego, ale jest kluczowe dla zapewnienia rygoru metodologicznego i komunikowalności wyników badań naukowych. Zaprezentowane rozwiązania wskazują, że odpowiednio zaprojektowane wykresy mogą obrazować niepewność, istotność statystyczną oraz strukturę rozkładów testowych w sposób zrozumiały i trafny – nawet dla naukowców, którzy nie zajmują się analizą danych.

Szczególny nacisk w pracy położono na alternatywne podejścia do wnioskowania, zestawiając klasyczny model populacyjny Neymana-Pearsona z modelem permutacyjnym Fishera-Pitmana. Wskazano, że metody permutacyjne, wolne od restrykcyjnych założeń o rozkładzie zmiennej w populacji, stanowią ważne narzędzie analityczne, cały czas niedoceniane w badaniach ekonomicznych i społecznych. Zaprezentowane przykłady – w tym autorska implementacja i wizualizacja testu pustych cel (Empty Cells Test) – pokazują, że nowoczesne wnioskowanie wymaga poszukiwania nowych, niestandardowych form wizualizacji. Graficzne ujęcie rozkładów permutacyjnych statystyk testowych pozwala badaczom na bezpośrednią ocenę istotności i wiarygodności hipotez w sposób, który jest trudny do osiągnięcia przy użyciu wyłącznie tradycyjnych metod tabelarycznych.

Wymiar praktyczny przedstawionych rozwiązań koncentrował się na wykorzystaniu środowiska **R** jako narzędzia wspierającego nowoczesną analizę danych. Przegląd bibliotek takich jak **`gginference`**, **`ggstatsplot`**, **`coin`**, **`ggpubr`** czy **`infer`** potwierdził, że współczesny badacz dysponuje narzędziami umożliwiającymi pełną automatyzację procesu analitycznego. Przedstawione procedury analizy on-line, integrujące pobieranie danych z Banku Danych Lokalnych (BDL), ich przetwarzanie oraz permutacyjną weryfikację hipotez, wyznaczają możliwości rozwoju zastosowań statystyki w badaniach ekonomicznych.

Rozwiązania przedstawione w monografii wskazują, że metody graficzne wspomagające wnioskowanie statystyczne nie są obecnie jedynie narzędziem prezentacyjnym, lecz mogą odgrywać kluczową rolę w całym procesie analitycznym – od eksploracji danych, przez testowanie hipotez, aż po komunikację wyników.

Bibliografia

Wydawnictwa zwarte i ciągłe

Aczel, A. D. (2000). *Statystyka w zarządzaniu*. Wydawnictwo Naukowe PWN.

Albert, J., Rizzo, M. (2012). *R by example: Concepts to code*. Springer. https://doi.org/10.1007/978-1-4614-1365-3

Arboretti, R., Carrozzo, E., Pesarin, F., Salmaso, L. (2018). *Testing for equivalence: An intersection-union permutation solution.* arXiv. https://doi.org/10.48550/arXiv.1802.01877

Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London, 27*(1710–1712), 186–190.

Baszczyńska, A. (2016). *Parametr wygładzania w estymacji jądrowej funkcji gęstości dla zmiennych losowych w badaniach ekonomicznych* (I). Wydawnictwo Uniwersytetu Łódzkiego.

Bąk, A., Dehnel, G., Dudek, A., Gatnar, E., Kania, K., Walesiak, M., Wawrowski, Ł., Zaborski, A. (2024). *Analiza danych z Banku Danych Lokalnych z wykorzystaniem programu R*. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu. https://doi.org/10.15611/2024.45.1

Bednarski, T. (2013). Rola Jerzego Spławy-Neymana w kształtowaniu metod statystycznej analizy przyczynowości. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 309*, 11–18.

Bernstein, P. L. (1997). *Przeciw bogom: niezwykłe dzieje ryzyka*. WIG-Press.

Berry, K. J., Johnston, J. E., Mielke, P. W. (2019). *A primer of permutation statistical methods*. Springer International Publishing.

Berry, K. J., Johnston, J. E., Mielke, P. W. Jr. (2014). *A chronicle of permutation statistical methods*. Springer International Publishing.

Berry, K. J., Kvamme, K. L., Johnston, J. E., Mielke, P. W. (2021). *Permutation statistical methods with R*. Springer International Publishing. https://doi.org/10.1007/978-3-030-74361-1

Bonnini, S., Borghesi, M. (2025). Multivariate two-sample permutation test with directional alternative for categorical data. *Statistics in Transition New Series, 26*(3), 181–194. https://doi.org/10.59139/stattrans-2025-033

Bonnini, S., Corain, L., Marozzi, M., Salmaso, L. (2014). *Nonparametric hypothesis testing. Rank and permutation methods with applications in R*. John Wiley & Sons.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. W: R. L. Launer, G. N. Wilkinson (eds.), *Robustness in statistics* (s. 201–236). Academic Press.

Brombin, Ch., Salmaso, L. (2013). *Permutation tests in shape analysis*. Springer Science +Business Media.

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders.

Couch, S. P., Bray, A. P., Ismay, C., Chasnovski, E., Baumer, B. S., Çetinkaya-Rundel, M. (2021). infer: An R package for tidyverse-friendly statistical inference. *Journal of Open Source Software, 6*(65), 3661. https://doi.org/10.21105/joss.03661

David, F. N. (1950). Two combinatorial tests whether a sample has come from a given population. *Biometrika, 37*, 97–110.

Domański, C. (1979). *Statystyczne testy nieparametryczne*. Państwowe Wydawnictwo Ekonomiczne.

Domański, C. (2011). Własności testu Davida-Hellwiga. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 165*, 40–49.

Domański, C. (2012). Statistical tests based on empty cells. *Acta Universitatis Lodziensis. Folia Oeconomica, 269*, 39–47.

Domański, C., Pruska, K. (2000). *Nieklasyczne metody statystyczne*. Polskie Wydawnictwo Ekonomiczne.

Eden, T., Yates, F. (1933). On the validity of Fisher's z test when applied to an actual example of non-normal data. *The Journal of Agricultural Science, 23*, 6–17.

Efron, B., Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Fay, M. P., Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software, 36*(2), 1–34. https://doi.org/10.18637/jss.v036.i02

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments*. Hafner Press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188.

Friendly, M. (1994). Mosaic display for multi-way contingency tables. *Journal of the American Statistical Association, 89*(425), 190–200.

Geary, R. C. (1927). Some properties of correlation and regression in a limited universe. *Metron Rivista Internazionale di Statistica, 7*, 83–119.

Good, P. I. (2006a). *Permutation, parametric, and bootstrap tests of hypotheses.* (T. 64). Springer Science+Business Media, Inc. http://link.springer.com/10.1007/s00184-006-0088-1

Good, P. I. (2006b). *Resampling methods: A practical guide to data analysis*. Birkhäuser.

Good, P. I. (2013). *Introduction to statistics through resampling methods and R: Good/introduction to statistics through resampling methods and R*. John Wiley & Sons. https://doi.org/10.1002/9781118497593

Hellwig, Z. (1965). Test zgodności dla małej próby. *Przegląd Statystyczny, 12*, 99–112.

Hellwig, Z. (1998). *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*. Wydawnictwo Naukowe PWN.

Hesterberg, T. (2022). *resample: Resampling functions*. https://doi.org/10.32614/CRAN.package.resample

Holland, E., Jabbar, A. B. A., Asghar, M. S., Asghar, N., Mistry, K., Mirza, M., Tauseef, A. (2025). Demographic and regional trends of pneumonia mortality in the United States, 1999 to 2022. *Scientific Reports, 15*(1), 10103. https://doi.org/10.1038/s41598-025-94715-6

Hothorn, T., Hornik, K. (2022). *exactRankTests: Exact distributions for rank and permutation tests*. https://doi.org/10.32614/CRAN.package.exactRankTests

Hothorn, T., Hornik, K., Wiel, M. A. van de, Zeileis, A. (2006). A Lego system for conditional inference. *The American Statistician, 60*(3), 257–263. https://doi.org/10.1198/000313006X118430

Iwasiewicz, A., Paszek, Z. (2000). *Statystyka z elementami statystycznych metod sterowania jakością*. Wydawnictwo Akademii Ekonomicznej w Krakowie.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Kończak, G. (2000). *Wykorzystanie kart kontrolnych w sterowaniu jakością w toku produkcji*. Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego.

Kończak, G. (2005). On the modification of David-Hellwig test. W: *Innovation in classification, data science and information systems: Proceedings of the 27th Annual Conference of the Gesellschaft für Klassifikation e. V., Brandenburg University of Technology, Cottbus, March 12–14, 2003* (s. 138–145). Springer.

Kończak, G. (2007). *Metody statystyczne w sterowaniu jakością produkcji*. Wydawnictwo Akademii Ekonomicznej im. Karola Adamieckiego.

Kończak, G. (2008). On the multivariate goodness-of-fit test. *Compstat 2008: Proceedings in Computational Statistics*, 751–758.

Kończak, G. (2009). On the modification of the empty cells test. *Acta Universitatis Lodziensis. Folia Oeconomica, 228*, 129–137.

Kończak, G. (2012). On testing multi-directional hypotheses in categorical data analysis. *20th International Conference on Computational Statistics (COMPSTAT 2012). Conference proceedings*, 427–436.

Kończak, G. (2016). Teoretyczne podstawy zastosowania testów permutacyjnych. *Prace Komisji Naukowych, 39–40*, 88–91.

Kończak, G. (2020a). Applications of permutation methods in the analysis of associations. *Argumenta Oeconomica Cracoviensia, 1*(22), 31–45. https://doi.org/10.15678/AOC.2020.2203

Kończak, G. (2020b). *Nieklasyczne metody statystyczne w badaniach ekonomicznych*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Kończak, G. (2024a). *Wizualizacja wyników badań naukowych. Zasady, metody i narzędzia*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Kończak, G. (2025). On a new goodness-of-fit test for multivariate normality with fixed parameters based on the David-Hellwig test idea. *Statistics in Transition New Series, 26*(4), 151–166. https://doi.org/10.59139/stattrans-2025-044

Kończak, G., Chmielińska, M. (2013). Zastosowanie metod symulacyjnych w analizie wielowymiarowych tablic wielodzielczych. *Studia Ekonomiczne, 133*, 107–118.

Kończak, G., Kosińska, M. (2023). O testowaniu istotności różnic w strukturach populacji na podstawie prób o małych liczebnościach. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, 3*(1001), 145–160. https://doi.org/10.15678/ZNUEK.2023.1001.0308

Kończak, G., Kosińska, M. (2025). Use of permutation methods in testing the independence of variables in multidimensional contingency tables. *Wiadomości Statystyczne. The Polish Statistician, 70*(5), 1–15.

Kończak, G., Kosińska, M. (2026). *Wprowadzenie do analizy danych w środowisku R*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Kończak, G., Złotoś, M. (2026). *Wprowadzenie do rachunku prawdopodobieństwa*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Kosińska, M. (2023). Wpływ pandemii COVID-19 na aktywność w wydarzeniach kulturalnych. *Studia Ekonomiczne. Gospodarka – Społeczeństwo – Środowisko*, *2*(12), 61–73.

Kosińska, M. (2025). Nierówności dochodowe w państwach Unii Europejskiej. *Zeszyty Naukowe Akademii Górnośląskiej, 35*(11), 113–121.

Kotz, S., Balakrishnan, N., Read, C. B., Vidakovic, B. (eds.). (2006). *Encyclopedia of statistical sciences*. Wiley-InterScience.

Lange, O., Banasiński, A. (1970). *Teoria statystyki*. Państwowe Wydawnictwo Ekonomiczne.

Laplace, P. S. (1820). *The analytic theory of probabilities. Book II*. Courcier.

Lehmann, E. L., Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Science +Business Media.

Mielke, P. W., Berry, K. J. (1994). Permutation tests for common locations among samples with unequal variances. *Journal of Educational and Behavioral Statistics, 19*(3), 217. https://doi.org/10.2307/1165295

Moore, D. S. (2011). *The practice of statistics for business and economics*. W. H. Freeman and Company.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*(767), 333–380.

Neyman, J., Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*, 175–240.

Neyman, J., Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A*, 263–294.

Neyman, J., Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*(694–706), 289–337.

Orwat-Acedańska, A., Kończak, G. (2025). Istotność różnic rozkładów współczynników dzietności w powiatach Polski w latach 2002–2023. *Zeszyty Naukowe Akademii Górnośląskiej, 33*(9/2025), 197–211.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50*(302), 157–175.

Pesarin, F. (2001). *Multivariate permutation tests: With applications in biostatistics*. John Wiley & Sons.

Pesarin, F., Salmaso, L. (2010). *Permutation tests for complex data*. John Wiley & Sons.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society, 4*, 119–130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations: II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society, 4*, 225–232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika, 29*, 322–335.

Plackett, R. L. (1983). Karl Pearson and the chi-squared test. *International Statistical Review, 51*, 59–72.

Polko, D., Kończak, G. (2016). On using permutation tests in the data homogeneity analysis. W: *Knowledge – Economy – Society. Selected challenges for statistics in contemporary management sciences* (s. 79–87). Foundation of the Cracow University of Economics.

Polko-Zając, D. (2020). A comparative study of the power of parametric and permutation tests for a multidimensional two-sample location problem. *Argumenta Oeconomica Cracoviensia, 23*, 69–79.

Polko-Zając, D. (2021). *Metody porównywania populacji w badaniach ekonomicznych*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Pruim, R., Kaplan, D. T., Horton, N. J. (2017). The mosaic package: Helping students to 'think with data' using R. *The R Journal, 9*(1), 77–102. https://journal.r-project.org/archive/2017/RJ-2017-024/index.html

Rao, C. R. (1994). *Statystyka i prawda*. Wydawnictwo Naukowe PWN.

Rosling, H., Rosling, O., Rosling Rönnlund, A. (2018). *Factfulness: Ten reasons we're wrong about the world – and why things are better than you think*. Flatiron Books.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Henry Holt and Company.

Sokołowski, A. (2004). O niewłaściwym stosowaniu metod statystycznych. W: *Statystyka i Data Mining w badaniach naukowych* (s. 5–14). StatSoft Polska.

Spława-Neyman, J. (1923). Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych. *Roczniki Nauk Rolniczych, 10*, 1–51.

Szpadel, M., Kania, K. (2023). *bdl: Interface and tools for 'BDL' API*. https://CRAN.R-project.org/package=bdl

Tarka, D., Olszewska, A. M. (2018). *Elementy statystyki: Opis statystyczny*. Oficyna Wydawnicza Politechniki Białostockiej.

Trzpiot, G., Kończak, G. (2002) *Statystyka dla studentów ekonomii*. Górnośląska Wyższa Szkoła Handlowa.

Tufte, E. R. (2013). *The visual display of quantitative information* (2nd ed.). Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Pearson Education.

Venables, W. N., Ripley, B. D. (2002). *Modern applied statistics with S*. Springer Science +Business Media.

Wald, A. (1947). *Sequential analysis*. John Wiley.

Weiss, N. A. (2015). *wPerm: Permutation tests*. https://doi.org/10.32614/CRAN.package.wPerm

Wells, H. G. (1903). *Mankind in the making*. Macmillan and Co.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer. https://doi.org/10.1007/978-0-387-98141-3

Wilkinson, L., Wills, G. (2005). *The grammar of graphics* (2nd ed.). Springer.

Wywiał, J. (2012). *Wprowadzenie do wnioskowania statystycznego*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Zar, J. H. (2010). *Biostatistical analysis* (5th ed.). Prentice-Hall/Pearson.

Zeliaś, A., Pawełek, B., Wanat, S. (2002). *Metody statystyczne: Zadania i sprawdziany*. Polskie Wydawnictwo Ekonomiczne.

Złotoś, M. (2019). On the use of permutation tests in the analysis of the factorial design of experiment results. *Acta Universitatis Lodziensis. Folia Oeconomica, 4*(343), 123–136.

Złotoś, M. (2020). On the use of permutation tests in the significance testing of response surface function parameters. *Argumenta Oeconomica Cracoviensia, 22*, 21–29. https://doi.org/10.15678/AOC.2020.2202

Złotoś, M. (2025). *Wybrane metody planowania eksperymentów w badaniach ekonomicznych*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach. https://doi.org/10.22367/uekat.9788378759607

Netografia

API_BDL. (2025). *API Bank Danych Lokalnych (BDL) – Dokumentacja*. https://api.stat.gov.pl/Home/BdlApi (dostęp: 25.01.2026).

BDL. (2025). *Bank Danych Lokalnych*. https://stat.gov.pl/bdl (dostęp: 25.01.2026).

CRAN. (2026). http://www.r-project.org (dostęp: 25.01.2026).

Kończak (2024b). *Wizualizacja wyników badań naukowych.* https://stat.ue.katowice.pl/wwbn (dostęp: 25.01.2026).

Kończak (2026). *Wizualizacja wyników badań naukowych*. https://stat.ue.katowice.pl/wwbn2 (dostęp: 25.01.2026).

RStudio. (2026). https://www.rstudio.com/ (dostęp: 25.01.2026).

Spis rysunków

Spis tabel

Metody graficzne są kojarzone przede wszystkim z zagadnieniami statystyki opisowej. W tradycyjnym ujęciu pełnią one funkcję pomocniczą – służą do przedstawienia rozkładów, relacji między zmiennymi czy wykrywania obserwacji odstających. Ich rola ogranicza się często do wstępnej analizy danych, która poprzedza właściwe wnioskowanie statystyczne. W tym sensie wykresy są traktowane jako narzędzie eksploracji umożliwiające intuicyjne zrozumienie struktury danych. Wnioskowanie statystyczne odgrywa jednak kluczową rolę w badaniach naukowych, ponieważ umożliwia uogólnianie wyników uzyskanych z próby na całą populację oraz szacowanie niepewności tych uogólnień. Ta fundamentalna właściwość czyni je niezastąpionym narzędziem w niemal wszystkich dyscyplinach nauki – od medycyny i fizyki, przez psychologię, aż po ekonomię i finanse. Kluczowe znaczenie wnioskowania statystycznego polega również na jego roli w zapewnianiu powtarzalności i weryfikowalności badań naukowych. Poprzez formalizację procedur testowania hipotez i szacowania niepewności metody statystyczne umożliwiają innym badaczom ocenę wiarygodności przedstawionych wyników oraz replikację badań. To z kolei przyczynia się do akumulacji wiedzy naukowej i eliminowania błędnych wniosków.

Monografia ma na celu ukazanie rosnącego znaczenia metod graficznych w kontekście wnioskowania statystycznego, a nie jedynie w ramach klasycznej statystyki opisowej. Zmierza to do wykazania, że wizualizacja danych może pełnić funkcję nie tylko ilustracyjną, lecz także analityczną i inferencyjną, wspierając proces podejmowania decyzji w badaniach statystycznych. Ponadto w pracy ukazano potencjał wizualizacji w ułatwianiu interpretacji i komunikacji wyników wnioskowania statystycznego, zwłaszcza w kontekście metod symulacyjnych, w tym testów permutacyjnych. Wskazano również, że graficzne przedstawienie wyników testów permutacyjnych (np. rozkładów permutacyjnych statystyk testowych, wizualizacji p-wartości) sprzyja transparentności i interpretowalności wyników analiz statystycznych.

Prof. dr hab. Grzegorz Kończak jest kierownikiem Katedry Statystyki, Ekonometrii i Matematyki na Wydziale Zarządzania Uniwersytetu Ekonomicznego w Katowicach. Obszar jego zainteresowań naukowych jest związany z wykorzystaniem metod symulacji komputerowej w badaniach ekonomicznych, metodami graficznej prezentacji i analizy danych, a także z zagadnieniami statystycznej kontroli jakości. Jest autorem lub współautorem ponad 100 publikacji naukowych, w tym 16 książek oraz kilkudziesięciu prac niepublikowanych. Uczestniczył w ponad 80 konferencjach i seminariach naukowych krajowych i zagranicznych.

0000-0002-4696-8215

ISBN 978-83-7875-988-1

Uniwersytet Ekonomiczny w Katowicach