Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

Statystyka I i II

Piotr Kościelniak

Kraków, 2022

Spis treści

1 Wstęp **3**
- 1.1 Definicje ogólne . . . 3
- 1.2 Rodzaje zmiennych i skale pomiarowe . . . 4
- 1.3 Rys historyczny . . . 5
- 1.4 R . . . 5
- 1.5 Uwagi końcowe . . . 6

2 Statystyka opisowa **7**
- 2.1 Miary liczbowe . . . 7
- 2.2 Podstawowe metody graficzne . . . 12

3 Estymacja punktowa **17**
- 3.1 Metoda największej wiarygodności . . . 18
- 3.2 Metoda momentów . . . 23
- 3.3 Własności teoretyczne estymatorów . . . 25

4 Przedziały ufności **29**
- 4.1 Rozkład t-Studenta . . . 30
- 4.2 Przedziały ufności dla średniej . . . 30
- 4.3 Rozkład χ^2 i Beta . . . 33
- 4.4 Przedział ufności dla wariancji . . . 33

5 Testowanie hipotez **35**
- 5.1 Ogólna teoria . . . 35

6 Przegląd podstawowych testów parametrycznych **38**
- 6.1 Podstawowe testy parametryczne . . . 38

7 Przegląd podstawowych testów nieparametrycznych **45**
- 7.1 Testy zgodności i niezależności . . . 45
- 7.2 Test znaków . . . 48
- 7.3 Testy serii i rangowe . . . 49
- 7.4 Inne testy rangowe i testy jednorodności . . . 54
- 7.5 Paradoks Simpsona . . . 58
- 7.6 Współczynnik τ Kendalla . . . 61

8 Testy normalności **62**

9 Wybrane zagadnienia **67**
- 9.1 Estymacja punktowa nieparametryczna . . . 67
- 9.2 Metody komputerowe . . . 70

10 Wstęp do metod bayesowskich **77**
- 10.1 Rozkład Beta . . . 77
- 10.2 Wnioskowanie bayesowskie . . . 78

11 Wielowymiarowy rozkład normalny **82**
- 11.1 Macierze dodatnio i nieujemnie określone . . . 82

11.2 Wektory losowe . . . 83
11.3 Funkcja generująca momenty . . . 84
11.4 Wielowymiarowy rozkład normalny . . . 84
11.5 Projekcje ortogonalne . . . 86

12 Modele liniowe **90**
12.1 Teoria ogólna . . . 90

13 Regresja liniowa **92**
13.1 Uwagi ogólne o modelowaniu statystycznym . . . 92
13.2 Postać modelu regresji liniowej . . . 92
13.3 Estymacja parametrów . . . 93
13.4 Przedziały ufności dla β oraz dla predykacji . . . 94
13.5 Testowanie hipotez . . . 95
13.6 Implementacja w R . . . 95
13.7 Współczynnik R^2 . . . 97
13.8 Diagnostyka modelu . . . 97
13.9 Transformacje zmiennych . . . 104
13.10 Wybór (selekcja) zmiennych do modelu . . . 106
13.11 Analiza kowariancji – ANCOVA . . . 109

14 Analiza wariancji – ANOVA **112**
14.1 ANOVA jednoczynnikowa . . . 112
14.2 ANOVA dwuczynnikowa . . . 115

Bibliografia **120**

Indeks **121**

Rozdział 1

Wstęp

Statystyka zajmuje się zbieraniem, analizowaniem, interpretowaniem oraz prezentacją danych (statystycznych). Można też powiedzieć, że statystyka to zbiór metod, które mają znajdować użyteczne informacje zawarte w danych.

W poniższych podrozdziałach bardziej precyzyjnie powiemy co to są dane, jakie są ich rodzaje oraz omówimy podstawowe pojęcia dotyczące badań statystycznych.

1.1 Definicje ogólne

Pierwszym krokiem w badaniu statystycznym jest określenie jakie obiekty badamy. Zbiór tych obiektów nazywamy *populacją* (*populacją generalną, zbiorowością statystyczną*[1]). Oznaczmy ten zbiór przez Ω. Formalnie wszystkie elementy populacji muszą mieć jakąś cechę wspólną (własność, która definiuje ten zbiór). Populacja oczywiście nie musi składać się z obiektów żywych (co sugeruje nazwa). Może to być zbiór wszystkich obywateli Polski, ale też zbiór wszystkich mieszkań w Krakowie. Populacja może być skończona lub też nieskończona. Jeden element populacji nazywamy *osobnikiem* (*jednostką statystyczną*[2]).

Drugim krokiem jest zdefiniowanie własności osobników, którą badamy. Własność tę nazywamy *cechą* (*zmienną*[3]). Naturalnie musimy założyć, że cecha jest mierzalna (obserwowalna), tzn. dla danego osobnika jesteśmy w stanie określić (lub zmierzyć) wartość tej cechy. Zakładamy też, że wartościami cech są liczby rzeczywiste. Wtedy cechę X możemy traktować jako funkcję $X : \Omega \to \mathbb{R}$, a $X(\omega)$ będzie wartością tej cechy dla osobnika $\omega \in \Omega$ (wartość ta nazywana też jest *obserwacją*). Możemy też badać kilka cech naraz (powiedzmy k) i wtedy formalnie cecha X będzie funkcją $X : \Omega \to \mathbb{R}^k$. Celem statystyki jest badanie własności cechy X na populacji Ω. W tym momencie możemy badania statystyczne podzielić na dwa główne rodzaje: badania *pełne* (*całkowite*[4]) oraz badania *częściowe* (*niepełne*[5]).

W badaniu pełnym wyznaczamy wartość badanej cechy dla **wszystkich** osobników z populacji. W ten sposób mamy pełną informacje o badanej zmiennej. Przykładami takich badań są powszechne spisy statystyczne, ewidencje urodzeń i zgonów, inwentaryzacje majątku, przymusowa sprawozdawczość dla urzędów statystycznych, czy finansowych. Problemem w tego typu badaniach jest to, że, wyłączając wyżej wymienione przykłady, trudno jest je przeprowadzić (może z wyjątkiem sytuacji, gdy populacja jest niewielka). Głównym powodem jest to, że takie badania są kosztowne i czasochłonne. Badania pełne są też oczywiście bezsensowne, jeśli sam pomiar cechy powoduje zniszczenie (czy zabicie) badanego osobnika.

Badania częściowe obejmują zaś tylko część osobników. Podzbiór badanych osobników $\{\omega_1, \ldots, \omega_n\} \subset \Omega$ nazywamy *próbką* (*próbą*[6]). Wartości badanej cechy na osobnikach z próby $X_1, \ldots, X_n$, gdzie $X_i = X(\omega_i)$, także nazywamy próbą. Generalnie będziemy zakładać, że *liczność próby* n jest istotnie mniejsza niż liczba wszystkich osobników. Teraz statystykę możemy podzielić na dwa rodzaje: *statystykę opisową*[7] oraz *statystykę matematyczną* (*wnioskowanie statystyczne*[8]).

Statystyka opisowa podaje własności próby (głównie za pomocą różnych charakterystyk liczbowych i metod graficznych) bez wnikania z jakiej populacji ta próba pochodzi. Tylko w sytuacji, gdy próba obejmuje wszystkich

[1] ang. *population.*
[2] ang. *individual, unit.*
[3] ang. *variable.*
[4] ang. *census, complete enumeration.*
[5] ang. *partial enumeration.*
[6] ang. *sample.*
[7] ang. *descriptive statistics.*
[8] ang. *inferential statistics.*

osobników (czyli w badaniu pełnym) statystyka opisowa da nam informacje o badanej zmiennej na całej populacji.

Statystyka matematyczna stara się zaś uogólnić informacje zawarte w próbie na całą populację. Bardziej obrazowo, stara się powiedzieć coś o całości znając własności części populacji. Wnioskować o całej populacji mając niepełną informację. Oczywiście nie można tego zrobić bez dodatkowych założeń. Widać, że kluczowy jest tutaj sposób wybrania próby. Na przykład powiedzmy, że badamy populację 18-latków w Polsce ze względu na wzrost, a do próby wzięliśmy 100-tu najwyższych. Trudno w tej sytuacji podejrzewać, że wnioski o całej populacji oparte o tę próbę będą prawdziwe. Dlatego chcemy, aby próba była *reprezentatywna*, tzn. o podobnych własnościach (strukturze) co cała populacja. Robimy to, losując osobników do próby i otrzymując w ten sposób *próbę losową*. Gdy każdy osobnik ma taką samą szansę bycia wylosowanym, mówimy, że mamy *próbę (losową) prostą*. Możemy też wtedy potraktować badaną cechę $X : \Omega \to \mathbb{R}$ jako zmienną losową, a całe badanie jako eksperyment losowy i użyć całego aparatu rachunku prawdopodobieństwa do jego opisu. W szczególności do oszacowania prawdopodobieństwa popełnienia błędu przy wnioskowaniu o całej populacji. Punktem wyjścia jest teraz zdefiniowanie rodziny rozkładów $\mathcal{P} = \{\mathrm{P}_\theta \,|\theta \in \Theta\}$ do której potencjalnie należy rozkład badanej cechy P_X. W tym kontekście charakterystyki liczbowe próby będziemy nazywali *statystykami*, a charakterystyki liczbowe cechy na całej populacji *parametrami*. Wnioskowanie statystyczne można wtedy sprowadzić do problemu, co można powiedzieć o parametrach na podstawie znanych statystyk i jaki błąd przy tym popełniamy.

Metody wnioskowania możemy też podzielić na *parametryczne*, gdy zbiór parametrów Θ zawiera się w przestrzeni skończenie wymiarowej, oraz *nieparametryczne* w przeciwnym wypadku. Do tych spraw wrócimy dokładniej w późniejszych rozdziałach, w szczególności pojawią się bardziej precyzyjne definicje.

Kończąc ten podrozdział należy podkreślić, że wnioskowanie statystyczne zawsze jest obarczone ryzykiem błędu. Nawet jeśli próba jest reprezentatywna, to błąd ma prawo się pojawić (wprost przeciwnie jego brak może być dziwny i podejrzany). Cała sztuka polega na tym, aby prawdopodobieństwo tego błędu było małe, a wnioski statystycznie odpowiednio interpretować (w szczególności być świadomym istnienia tego błędu).

1.2 Rodzaje zmiennych i skale pomiarowe

Założyliśmy, że wartość badanej cechy jest liczbą rzeczywistą. Jak wiadomo, na liczbach rzeczywistych możemy przeprowadzać wiele operacji (dodawanie, mnożenie, porównywanie, porządkowanie, etc). W zależności od tego jakie operacje na liczbach w sensowny sposób przenoszą się na własności badanej cechy, będziemy mieć do czynienia z różnymi *skalami pomiarowymi*. Formalną definicję skal pomiarowych można znaleźć na przykład w [23], my podamy jednak definicję bardziej opisową (intuicyjną).

W skali *nominalnej*[9] liczba jest tylko nazwą wartości cechy. Co za tym idzie możemy jedynie porównywać wartości cech, tzn. czy są równe, czy różne (mówiąc bardziej matematycznie, możemy tylko używać relacji równości). Rozważmy najprostszy przykład cechę płeć. Możemy zakodować wartość cechy „kobieta" jako 1, a „mężczyzna" jako 0. Jedyne co możemy robić, to porównywać te liczby, czyli na przykład jak dwóch osobników ma wartość tej cechy 1, to wiemy, że są kobietami, a jak są różne, to osobnicy mają różną płeć. Z nierówności liczb $1 > 0$ nic nie wynika dla badanych osobników. W skali tej nie ma jednostki pomiarowej, a przeskalowanie (czyli operacja zamiany wartości pomiarów tak, aby własności skali się nie zmieniły) odbywa się za pomocą dowolnej bijekcji (w powyższym przykładzie równie dobrze byłoby zakodować „kobietę" jako 5, a „mężczyznę" jako 0).

Skala *porządkowa*[10] dodatkowo umożliwia uporządkowanie (inaczej *rangowanie*) obserwacji (możemy używać relacji porządku liczb rzeczywistych). Skala ta także nie ma jednostki pomiarowej ani naturalnego zera. Przykładem jest skala ocen w szkole: „niedostateczny" – 1, „mierny" – 2, aż do „celujący" – 6. Z relacji $3 < 5$ wynika, że uczeń z oceną 5 jest lepszy niż uczeń z oceną 3. Nie ma tu naturalnego zera, a przeskalowanie dokonujemy przez funkcję rosnącą (jeśli chcemy zachować porządek). Więc własności tej skali się nie zmienią, jeśli liczby $\{1, \ldots, 6\}$ zmienimy na $\{-100, -50, 0, 1, 2, 10\}$. W skali tej możemy więc używać określeń typu „większy-mniejszy", „lepszy-gorszy", etc. Nie możemy natomiast interpretować różnic. W powyższym przykładzie nie możemy powiedzieć, czy że są takie same różnice w wiedzy miedzy uczniami z ocenami 5 i 4, a uczniami z ocenami 2 i 1 (można to też uzasadnić tym, że po przeskalowaniu te różnice się mogą po prostu zmienić).

W skali *przedziałowej*[11] dodatkowo możemy porównywać różnice obserwacji, tzn. badać o ile różnią się osobnicy względem badanej cechy. W skali tej jest ustalona (arbitralnie) jednostka oraz umowne zero. Przeskalowanie odbywa się za pomocą funkcji afinicznej. Przykładem jest pomiar temperatury w stopniach Celsjusza (a przez przeskalowanie w stopniach Fahrenheita). Przykładowo jeśli w Krakowie rano było 10°C, a w Warszawie 20°C, to jest sens powiedzieć,

[9] ang. *nominal scale.*
[10] ang. *ordinal scale.*
[11] ang. *interval scale.*

że w Warszawie było cieplej (porządkowanie) oraz, że w Warszawie było o 10°C cieplej (badanie różnicy). Nie możemy natomiast powiedzieć, że w Krakowie było 2 razy zimniej niż w Warszawie (po pierwsze, przeskalowanie może zmienić stosunek tych wartości, a po drugie co byśmy powiedzieli, jakby te temperatury wynosiły 0°C i −10°C?).

Ostatnią skalą jest skala *ilorazowa*[12]. W skali tej dodatkowo możemy badać ilorazy wartości cech. Skala ta ma absolutne zero, a skalowanie odbywa się za pomocą funkcji liniowej. Przykładami są waga (w kilogramach, można przeskalować na tony), wzrost (w centymetrach), etc. W tej sytuacji stwierdzenie, że pies ważący 5kg jest dwa razy lżejszy niż pies ważący 10kg, jest poprawne.

Skalę nominalną uważa się za najsłabszą, a ilorazową za najmocniejszą, w tym sensie, że w skali ilorazowej mamy najwięcej możliwości manipulacji wartościami pomiarów.

Innym ważnym podziałem cech jest podział na cechy *ilościowe* (*mierzalne*[13]) oraz *jakościowe* (*niemierzalne, kategoryzacyjne*[14]). Nie ma jednoznacznej definicji tych pojęć. Można zdefiniować zmienną jakościową jako taką, której wartości pomiaru nie są bezpośrednio liczbą (generalnie są definiowane słownie), np. płeć, wykształcenie. A ilościowe jako te, dla których wartość pomiaru jest liczbą. W sensie tej definicji zmienne mierzone w skali nominalnej i porządkowej byłyby jakościowe, a w skali przedziałowej lub ilorazowej ilościowe. W polskiej systematyce zmienną ilościową definiuje się jako taką, którą można zmierzyć i porównać, wiec skala porządkowa „przeszłaby" do cech ilościowych (ale i tak niekiedy dla tej skali jest używana nazwa cecha *quasi-ilościowa*). Potencjalne wartości cech jakościowych nazywamy *poziomami* (*kategoriami*[15]). Natomiast cechy ilościowe można jeszcze podzielić na cechy *dyskretne* (gdy przyjmują skończony lub przeliczalny zbiór wartości, tj. mają rozkład dyskretny) i *ciągłe* (gdy modelujemy je rozkładem ciągłym). Cechy dyskretne są najczęściej efektem zliczania (np. liczba posiadanych samochodów, liczba wypadków) a cechy ciągłe efektem pomiaru własności fizycznych urządzeniami pomiarowymi (np. wzrost, waga, długość).

Ten podrozdział zakończymy dwoma uwagami.

Pierwsza jest taka, że przed przystąpieniem do analiz musimy określić z jakiego typu zmiennych składają się nasze dane, ponieważ to determinuje metody jakie możemy użyć.

Druga, bardziej ogólna, że wyniki metody statystycznej nie powinny zależeć od przeskalowania. Bardziej szczegółowo, rozważmy cechę waga słonia w Afryce. Powiedzmy, że dane mamy w kilogramach i używając ustalonej metody wyszło nam, że na 95% średnia waga słonia jest większa od 2000kg. Dokonujemy przeskalowania na tony i używając tej samej metody wychodzi nam, że na 95% średnia waga słonia jest mniejsza niż 2 tony. Taka metoda byłaby niepoprawna (właściwie bezsensowna).

1.3 Rys historyczny

Statystyka została zapoczątkowana poprzez spisy ludności w państwach starożytnych (Sumeria, Egipt, Chiny, Imperium Rzymskie). Pierwsze użycie słowa statystyka nastąpiło w XVIII wieku i zostało zdefiniowane jako nauka o gromadzeniu i wykorzystaniu danych przez państwo (samo słowo statystyka pochodzi z łacińskiego *status* „państwo"). Ten nurt trwa do dzisiaj i nosi też nazwę *państwoznawstwa*. Jak widać mamy tu do czynienia z badaniami pełnymi i statystyką opisową.

Początki wnioskowania statystycznego można upatrywać w pracach Pettiego i Graunta w drugiej połowie XVII wieku (np. oszacowanie ludności Londynu, określenie szansy przeżycia osób w określonym wieku, oszacowanie dochodu narodowego Anglii). Prawdziwy rozwój statystyki matematycznej nastąpił wraz z rozwojem rachunku prawdopodobieństwa i nastąpił w pierwszych dekadach XX wieku, głównie za sprawą Fishera, Pearsonów i Spława-Neymana.

Dalsze informacje można znaleźć, np. w [12].

1.4 `R`

Zazwyczaj metody statystyczne są obliczeniowo bardzo czasochłonne, więc oczywiście używa się komputera. Istnieje wiele programów statystycznych. Na tym kursie będziemy używać programu `R` (https://cran.r-project.org/). Niemniej proszę nie traktować tego wykładu jako kursu programu `R`. Program ten posłuży nam tylko do implementacji poznanych metod.

[12] ang. *ratio scale.*
[13] ang. *quantitative.*
[14] ang. *qualitative, categorical, factor.*
[15] ang. *levels.*

1.5 Uwagi końcowe

Poniższy wykład ma charakter przeglądowy, nastawiony jest raczej na „stosowanie", a nie „teorię" i jest przeznaczony na 60 godzin zajęć. Z tego powodu znajduje się w nim mało dowodów twierdzeń, choć oczywiście same definicje i twierdzenia się znajdują. Taki cel wykładu został przyjęty świadomie, a motywowane to było na przykład tym, że lepiej jest chyba zdefiniować dziesięć testów z przykładami zastosowań, niż jeden z pełnym dowodem na to jaki rozkład ma funkcja testowa.

Literatura dotycząca statystyki jest olbrzymia. Poniżej prezentujemy literaturę, która była pomocna w pisaniu tego kursu oraz która może służyć dla zainteresowanych do pogłębienia wiedzy z poszczególnych działów.

- Estymacje, testowanie hipotez (teoria): [3, 16, 21];
- Estymacja punktowa: [17];
- Statystyka ogólnie (w formie leksykonu): [18];
- Modele liniowe: [1, 4, 8, 23];
- Statystyka nieparametryczna: [13];
- Wnioskowanie bayesowskie: [11];
- Metody bootstrapowe: [6, 7].
- Metody statystyczne w R: [4, 5, 8, 23].

Rozdział 2

Statystyka opisowa

Statystyka opisowa dostarcza narzędzi do analizy (badania własności, struktury) próby, ale bez uogólnień na całą populację. Niemniej jest kluczowa w pierwszym etapie analizy statystycznej, ponieważ dostarcza nam informacje, które pozwalają nam ocenić „na oko" jakich własności badanej cechy możemy się spodziewać (oraz oczywiście dlatego, że tych informacji używamy potem we wnioskowaniu statystycznym).

Metody statystyki opisowej możemy podzielić na dwa rodzaje: miary (inaczej: statystyki, współczynniki) liczbowe, które charakteryzują próbę, oraz metody graficzne („wizualizacja" danych), to znaczy tworzenie rysunków, na podstawie których jesteśmy w stanie podać własności próby.

2.1 Miary liczbowe

Miary liczbowe możemy podzielić na kilka kategorii (miary położenia, zmienności, skośności i koncentracji) w zależności od tego, jaką własność próby opisują.

2.1.1 Miary położenia

Miary położenia[1] informują nas, gdzie na osi liczbowej „leży" próba. Zaczniemy od *miar położenia centralnego*[2], które wyznaczają „środek" („centrum") próbki. Najczęściej używanymi są średnia arytmetyczna oraz mediana.

Niech $X = (X_1, \ldots, X_n)$ będzie próbą n elementową. Wtedy

$$\bar{X}_n = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

nazywamy *średnią arytmetyczną z próby*[3]. Często mówi się po prostu *średnia*, ale z kontekstu powinno się rozróżnić, czy mówimy o średniej z próby, czy *średniej dla całej populacji*[4][5]. W R tę średnią wyznaczamy następująco

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> mean(X)
[1] 2.055
```

Zdefiniujmy *odchylenie i-ej obserwacji od średniej*[6] $d_i = X_i - \bar{X}$. Wtedy łatwo udowodnić, że

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} d_i = 0.$$

Powyższa własność ma ciekawą interpretację fizyczną. Mówi ona bowiem, że średnia arytmetyczna jest *środkiem ciężkości* (*barycentrum*)[7] figury $\{X_1, \ldots, X_n\}$, zakładając, że każdy punkt X_i ma taką samą masę.

[1] ang. *location measures.*
[2] ang. *measure of central location.*
[3] ang. *(arithmetic) (sample) mean.*
[4] ang. *population mean.*
[5] Ta uwaga dotyczy wszystkich miar liczbowych z próby.
[6] ang. *deviation of i-th observation from the arithmetic mean.*
[7] ang. *center of gravity, balance point.*

Średnia arytmetyczna jest chyba najczęściej stosowaną statystyką we wnioskowaniu statystycznym, niemniej posiada pewną wadę. Jest nią *czułość na wartości odstające*. *Wartości odstające*[8] zostaną formalnie zdefiniowane później, niemniej intuicyjnie są to obserwacje, które są istotnie większe lub mniejsze od pozostałych, które stanowią większość. Czułość na wartości odstające oznacza, że jedna wartość różniąca się od pozostałych może mieć duży wpływ na wartość średniej. Dla przykładu załóżmy, że zapytaliśmy pięciu studentów, ile samochodów by chcieli posiadać i uzyskaliśmy odpowiedzi $1, 0, 2, 2, 1000$. Średnia wynosi $1005/5 = 201$, ale trudno powiedzieć, że „średnio" w tej grupie student chce mieć 201 samochodów, jeśli większość z nich chce co najwyżej dwóch.

Zwróćmy uwagę na inne aspekty użycia średniej arytmetycznej. Będziemy używać tę średnią wyliczoną dla obserwacji **tej samej** cechy (tak została zdefiniowana) mierzonej w skali co najmniej przedziałowej, więc w szczególności, jeśli na przykład obserwacje były w metrach, to średnia też jest w metrach. O ile widać, że liczenie średniej z próby dla cechy mierzonej w skali nominalnej nie ma sensu, to pozostaje pytanie, czy jest sens jej liczenia w skali porządkowej, na przykład w skali ocen w szkole od 1 do 6 [9]. W „życiu codziennym" spotykamy się też z sytuacją liczenia średniej arytmetycznej z obserwacji różnych cech, co z naszego punktu widzenia nie będzie interpretowywalne[10].

Drugą powszechnie stosowaną miarą położenia centralnego jest mediana. Niech $X = (X_1, \ldots, X_n)$ będzie już posortowaną próbą (to znaczy, że zachodzi $X_1 \leqslant \ldots \leqslant X_n$). Wtedy

$$\mathrm{me}(X) = \begin{cases} \frac{x_k + x_{k+1}}{2}, & \text{jeśli } n = 2k \text{ dla pewnego } k \in \mathbb{N} \\ x_k, & \text{jeśli } n = 2k-1 \text{ dla pewnego } k \in \mathbb{N} \end{cases}$$

nazywamy *medianą z próby*[11]. Widzimy więc, że mediana to liczba taka, że połowa obserwacji jest niewiększa od niej, a połowa obserwacji z próby jest od niej niemniejsza. W `R` obliczamy ją następująco

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> median(X)
[1] 1.42
```

Mediana, w przeciwieństwie do średniej arytmetycznej, jest statystyką *odporną*[12] na wartości odstające, w szczególności w poprzednim przykładzie mediana liczby samochodów wynosi 2. Niestety mediana jako funkcja ma słabsze własności matematyczne i jest rzadziej wykorzystywana we wnioskowaniu statystycznym.

Istnieją też modyfikacje średniej arytmetycznej w celu jej „uodpornienia" na wartości odstające. Najpopularniejsze to średnia *obcięta*[13] i *winsorska* lub *winsorowska*[14]. średnia obcięta to średnia z próby po usunięciu z niej ustalonej części najmniejszych i największych obserwacji. W `R` możemy ją policzyć następująco

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> mean(X,trim=0.3)
[1] 1.87
```

(tu usunięto po około 30% obserwacji z obu stron). Średnia ta jest często stosowana w sporcie. Średnia winsorska różni się od obciętej tym, że usunięte obserwacje zastępuje się odpowiednio najmniejszą i największą obserwacją nieusuniętą (na przykład średnia winsorska z próby $1, 3, 4, 5, 6, 9, 10$ to średnia z liczb $4, 4, 4, 5, 6, 6, 6$ przy założeniu, że modyfikujemy po dwie skrajne obserwacje).

W przypadku obserwacji mierzonych w skali nominalnej lub porządkowej najczęściej używaną miarą jest *moda*[15] (*dominanta*), która jest definiowana jako najczęściej występująca obserwacja w próbie. Przykładowo

```
> sort(table(c(2,1,2,2,3,2,3,2,1,5,5,4,3,4)))

1 4 5 3 2
2 2 2 3 5
```

[8] ang. *outliers, extreme observations.*

[9] Co nam mówi informacja, że w klasie V b średnia ocena z historii wynosi 4.12, a w V c wynosi 4.20?

[10] Czy uczeń, który z poszczególnych (różnych) przedmiotów ma oceny $3, 3, 3, 6, 6$ jest „lepszy" od ucznia, który uzyskał odpowiednio oceny $4, 4, 4, 4, 4$, bo średnia ocen jest większa? W tym przykładzie skala pomiarowa jest ta sama, co może „usprawiedliwiać" użycie średniej. Ale na przykład średnia ze wzrostu (w metrach), wagi (w kilogramach) i ostrości wzroku (w dioptriach) byłaby bez sensu.

[11] ang. *sample median.*

[12] ang. *robust statistic.* Oczywiście czasami ta odporność też może być wadą.

[13] ang. *trimmed mean.*

[14] ang. *winsorized mean.*

[15] ang. *mode.*

(tu modą jest 2, która wystąpiła w próbie pięć razy).

Wątek o miarach centralnych zakończymy uwagą, że oczywiście istnieją inne średnie, na przykład ważona, geometryczna, harmoniczna, potęgowa, etc., ale nimi nie będziemy się zajmowali.

Najbardziej ogólną miarą położenia jest kwantyl.

Kwantylem (z próby) rzędu p[16] nazywamy liczbę x_p taką, że co najmniej $p \times 100\%$ obserwacji z próby jest mniejsze lub równe x_p i co najmniej $(1-p) \times 100\%$ obserwacji jest większe lub równe x_p. W R wyznaczamy je następująco[17]

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> quantile(X,p=c(0.1,0.5,0.8))
 10%  50%  80%
0.75 1.42 3.44
```

Najczęściej używanymi kwantylami są:

- *kwartyl dolny*[18] Q_1 oraz *kwartyl górny*[19] Q_3, to jest kwantyle rzędu odpowiednio 0.25 oraz 0.75. Razem z medianą (czyli kwartylem rzędu 0.5) kwartyle dzielą próbę na cztery grupy kwartylowe;
- decyle, to znaczy kwantyle rzędu $0.1, 0.2, \ldots, 0.9$;
- percentyle, to jest kwantyle rzędu $0.01, 0.02, \ldots, 0.99$.

W R funkcja summary dostarcza nam podstawowe statystyki pozycyjne z próby:

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> summary(X)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.300   1.235   1.420   2.055   2.955   4.550
```

2.1.2 Miary zmienności

Kolejną grupą statystyk są *miary zmienności*[20], które badają zmienność (zróżnicowanie) obserwacji (lub inaczej, badają jak bardzo skupione/rozproszone są obserwacje).

Najprostszą miarą zmienności jest *rozstęp*[21]

$$R = \max\{X_1, \ldots, X_n\} - \min\{X_1, \ldots, X_n\}.$$

Wadą tej miary jest to, że zależy tylko od dwóch skrajnych obserwacji, wiec nie informuje nas co się dzieje „w środku" (i w szczególności jest bardzo czuła na wartości odstające).

Inną miarą jest *rozstęp międzykwartylowy*[22]

$$IQR = Q3 - Q1,$$

czyli rozstęp 50% środkowych obserwacji (lub innymi słowy obserwacji z drugiej i trzeciej grupy kwartylowej). W R te statystyki możemy obliczyć następująco:

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> diff(range(X))
[1] 4.25
> IQR(X)
[1] 1.72
```

Zdefiniujemy teraz dwie najczęściej używane miary dyspersji.

Wariancją (z próby)[23] $X_1, \ldots, X_n$ nazywamy

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n},$$

[16] ang. *(sample) quantile corresponding to the probability p, p-quantile.*
[17] Formalnie funkcja quantile domyślnie wyznacza kwantyle używając trochę innej definicji (dostępne jest 9 definicji).
[18] ang. *lower (first) quartile.*
[19] ang. *upper (third) quartile.*
[20] ang. *measures of dispersion (variability, spread).*
[21] ang. *range.*
[22] ang. *interquartile range.*
[23] ang. *(sample) variance.*

a *odchyleniem standardowym (z próby)*[24] jej pierwiastek

$$s = \sqrt{s^2}.$$

Wariancję definiuje się też jako

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

Pierwszej definicji używa się na gruncie statystyki opisowej, a drugiej we wnioskowaniu statystycznym[25]. W R wyznaczamy je następująco (funkcje te wyznaczają wariancję z drugiej definicji):

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> var(X)
[1] 2.55471
> sd(X)
[1] 1.598346
```

Jak widać, wariancja to średni kwadrat odległości obserwacji od średniej. Łatwo też widać, że wariancja i odchylenie standardowe są nieujemne, równe zero tylko, gdy próba jest stała, a im rozproszenie większe, tym te statystyki są większe. Odchylenie standardowe jest chyba używane częściej, ponieważ jest wyrażone w tej samej jednostce co pomiary cechy, a wariancja w kwadracie tej jednostki. Co więcej odchylenie standardowe lepiej zachowuje się przy zamianie skali (na przykład z centymetrów na metry). Wadą wariancji może być to, że obserwacje, które są położone daleko od średniej, mają większy wpływ na wartość tej statystyki, niż obserwacje leżące blisko średniej. W celu uniknięcia tego efektu rozważa się niekiedy *odchylenie przeciętne*[26]

$$d = \frac{\sum_{i=1}^{n}|X_i - \bar{X}|}{n}.$$

Nie będziemy podawać ogólnych własności tych miar, ale można na przykład pokazać, że dla prób o liczności $n \geqslant 4$ zachodzi $0 \leqslant d \leqslant s \leqslant 0.6R$, etc. Z praktycznego punktu widzenia bardziej istotnym problemem jest to, że odchylenia standardowego nie można bezpośrednio użyć do porównania dwóch prób i stwierdzenia, które obserwacje były bardziej różnorodne (w szczególności mogą mieć nawet różne jednostki pomiarowe). Dlatego definiuje się *współczynnik zmienności*[27] wzorem (dla prób ze średnią dodatnią)

$$V = \frac{s}{\bar{X}} \times 100\%,$$

który już dla każdej próby wyrażony jest w procentach i można go użyć do porównywania zmienności dwóch prób. Jeśli $V > 100\%$, to mówimy, że próba ma dużą zmienność. Wadą tego współczynnika jest to, że dla średnich bliskich zero, jest „niestabilny", to znaczy mała zmiana średniej powoduje dużą zmianę wartości tego współczynnika.

2.1.3 Miary skośności

Miary *skośności (asymetrii)*[28] mają za zadanie określenie jak bardzo próba jest skośna (to znaczy niesymetryczna). Przy czym powiemy, że próba jest *symetryczna*, jeśli zbiór $\{X_1, \ldots, X_n\}$ jako figura na osi liczbowej jest symetryczna. Rodzajów niesymetryczności jest wiele, najczęściej wyróżnia się próbę *skośną prawostronnie (w prawo)*[29] (gdy skrajne prawe obserwacje rozciągają się w prawo bardziej niż lewe skrajne w lewo (patrz też rysunek 2.1)) i *skośną lewostonnie (w lewo)*[30] (gdy jest odwrotnie niż w skośności prawostronnej)[31].

Pierwsza miara wykorzystuje fakt, że dla próby symetrycznej średnia arytmetyczna i mediana są równe (a dla próby jednomodalnej prawostronnie skośnej mediana jest mniejsza od średniej). Jest nim *współczynnik skośności Pearsona*[32] dany wzorem

$$As_1 = \frac{3(\bar{X} - \mathrm{me}(X))}{s}.$$

[24] ang. *(sample) standard deviation.*
[25] Powodem jest to, że ta druga statystyka jest estymatorem nieobciążonym wariancji cechy na całej populacji.
[26] ang. *mean (average) absolute deviation.*
[27] ang. *coefficient of variation.*
[28] ang. *coefficient of skewness.*
[29] ang. *skewed to the right, positively skewed.*
[30] ang. *skewed to the left, negatively skewed.*
[31] To nie są matematycznie precyzyjne definicje...
[32] ang. *Pearson's skewness coefficient*

Innym współczynnikiem jest *kwartylowy współczynnik skośności*[33]

$$As_2 = \frac{(Q_3 - \mathrm{me}(X)) - (\mathrm{me}(X) - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2\,\mathrm{me}(X)}{Q_3 - Q_1},$$

który bada jak bardzo kwartyle są niesymetryczne względem mediany (dla próby symetrycznej oczywiście kwartyle są symetryczne względem mediany i wartość tego współczynnika wynosi zero).

Jeszcze innym jest *współczynnik skośności*[34]

$$As_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}.$$

Inaczej mówiąc jest to standaryzowany trzeci moment centralny z próby. W R możemy je policzyć na przykład tak:

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> 3*(mean(X)-median(X))/sd(X)
[1] 1.191857
> (mean((X-mean(X))^3))/(sd(X)^3)
[1] 0.4743385
```

Wszystkie te współczynniki przyjmują wartość zero dla prób symetrycznych i wartości dodatnie (ujemne) dla prób jednomodalnych skośnych w prawo (w lewo).

Zakończymy ten podrozdział uwagą, że współczynniki te są liczbą (bez jednostki), a przy afinicznej zmianie skali ich wartość się nie zmienia (po to istnieją mianowniki w powyższych zbiorach).

2.1.4 Miary koncentracji

Najczęściej stosowaną miarą koncentracji (skupienia) wokół średniej jest *kurtoza*[35] dana wzorem

$$\kappa = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4} - 3.$$

Dla prób pochodzących z rozkładu normalnego wartość kurtozy wynosi około zero (próba o kurtozie bliskiej zero nazywana jest *mezokurtyczną*). Duża kurtoza świadczy o „ogonach" „grubszych" niż w rozkładzie normalnym (takie próby nazywamy *platokurtycznymi*), a ujemna o „cieńszych" (takie próby nazywamy *leptokurtycznymi*). W R mamy

```
> X=c(1.2,1.5,0.3,3.44,4.55,1.34)
> (mean((X-mean(X))^4))/(sd(X)^4)
[1] 1.348537
```

[33] ang. *quartile skewness coefficient*
[34] ang. *skewness.*
[35] ang. *kurtosis*

2.2 Podstawowe metody graficzne

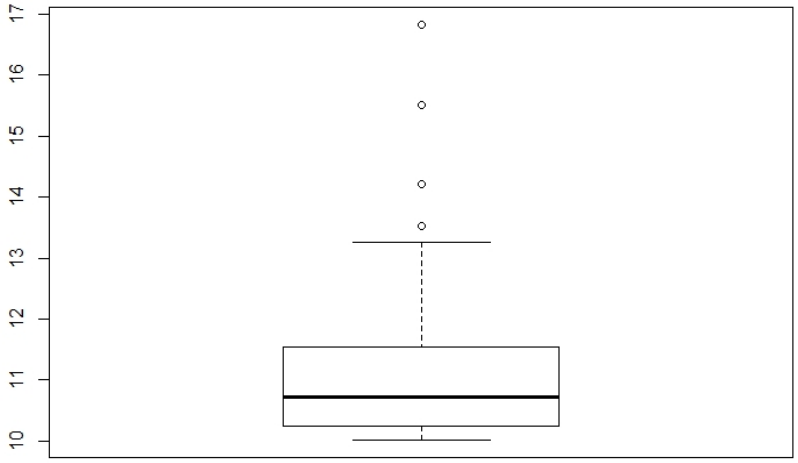
Metod wizualizacji danych jest bardzo dużo. Zdefiniujemy trzy podstawowe rodzaje graficznej prezentacji danych.

2.2.1 Pudełko z wąsami – boxplot

Najpierw zdefiniujemy wartości odstające (to jedna z możliwych definicji).

Definicja 2.1. Niech $\underline{X} = (X_1, \ldots, X_n)$ będzie próbą. Obserwację X_i nazywamy *odstającą*, gdy $X_i > Q_3 + 1.5IQR$ albo $X_i < Q_1 - 1.5IQR$.

Pudełko z wąsami (ang. *boxplot*) to wizualizacja podstawowych statystyk pozycyjnych i ewentualnie występujących obserwacji odstających w danej próbie.



Pogrubiony odcinek wewnątrz prostokąta (pudełka) znajduje się na wysokości mediany z próby. Dolny i górny bok znajdują się odpowiednio na wysokości kwartyla dolnego Q_1 i górnego Q_3. ‚Wąsy' kończą się na wysokości ostatnich wartości nieodstających. Ewentualnie występujące obserwacje odstające są zaznaczane kółeczkami. Z powyższego przykładu możemy na przykład odczytać, że mediana z próby jest odrobinę mniejsza niż 11 oraz, że próba jest trochę skośna w prawo (trzecia grupa kwartylowa jest dłuższa niż druga, górny wąs jest dłuższy od dolnego oraz występują obserwacje odstające w górę), etc. Boxplot dla próby symetrycznej będzie oczywiście symetryczny.

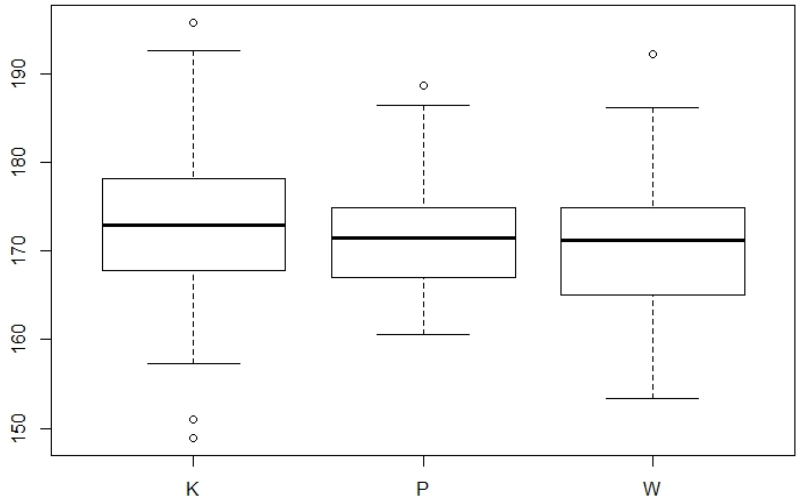
Boxplot może też służyć do wstępnego badania niezależności zmiennej ilościowej i jakościowej. Dla zilustrowania tego zastosowania rozważmy *data frame* dane o trzech zmiennych (dane są fikcyjne): wzrost, waga oraz zamieszkanie – *factor* o trzech poziomach W, K i P (powiedzmy odpowiednio Warszawa, Kraków i Poznań). Kilka pierwszych obserwacji wygląda tak:

```
      wzrost      waga zamieszkanie
1   186.4530  88.24875            P
2   171.3599  69.51805            W
3   164.8907  79.36840            W
4   172.7926  63.67085            K
```

Rezultatem instrukcji

```
> boxplot(wzrost~zamieszkanie,data=dane)
```

będzie następujący rysunek



Boxploty dla obserwacji zmiennej wzrost dla tych trzech poziomów zmiennej zamieszkanie są podobne, więc stwierdzamy, że raczej zmienna zamieszkanie nie ma wpływu na zmienną wzrost, więc ‚na oko' są niezależne.

2.2.2 Histogram

Dla danej próby $\underline{X} = (X_1, \ldots, X_n)$ ustalmy $a_0 \in \mathbb{R}$ oraz $h > 0$ (zwane *szerokością pasma*). Zdefiniujmy *klasy* $K_i = [a_0 + ih, a_0 + (i+1)h)$ oraz *liczności klas* $n_i = \#\{X_s : S_s \in K_i\}$ dla wszystkich $i \in \mathbb{Z}$. Oczywiście $\sum_i n_i = n$ oraz poza skończoną liczbą klas liczności są zerowe. Klasy możemy także definiować jako odcinki o różnych długościach. Ciąg, składający się z par $\{(K_i, n_i)\}_i$, nazywamy *szeregiem rozdzielczym*. Niekiedy obserwacje bezpośrednio mamy w formie szeregu rozdzielczego, gdy na przykład metoda pomiarowa jest niedokładna. Gdy chcemy policzyć jakieś statystyki opisowe z szeregów rozdzielczych, to generalna zasad jest taka, że dla szeregu $\{(K_i, n_i)\}_i$ tworzymy próbę $X_1, \ldots, X_n$ (gdzie $n = \sum_i n_i$), która składa się z n_i środków klasy K_i (dla wszystkich i). Następnie dla tej próby wyznaczamy statystyki.

Histogram jest to graficzne przedstawienie szeregu rozdzielczego (plus ewentualne skalowanie). Precyzyjniej histogram to wykres jednej z poniższych funkcji:

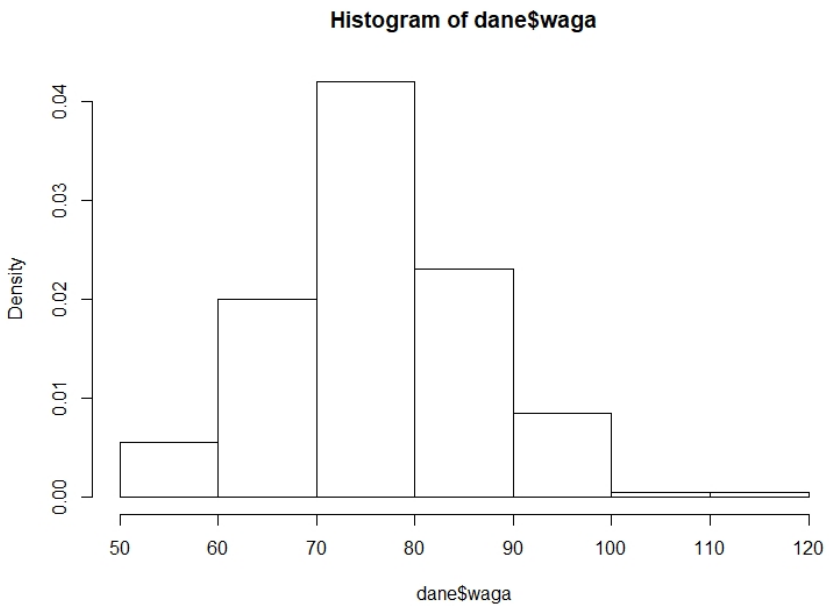
$$h_1(x) = n_i, \text{ gdy } x \in K_i,$$

$$h_2(x) = \frac{n_i}{n}, \text{ gdy } x \in K_i,$$

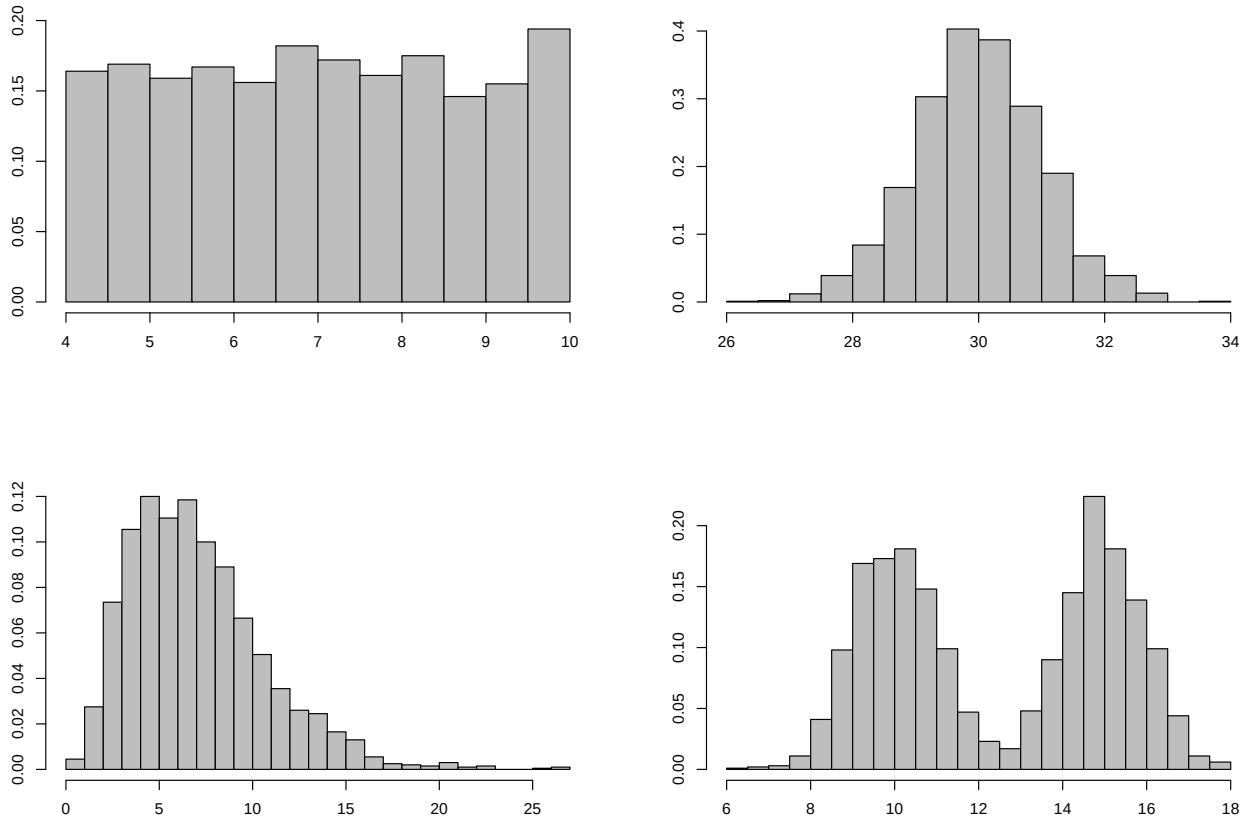
$$h_3(x) = \frac{n_i}{Nh}, \text{ gdy } x \in K_i.$$

Funkcja h_3 jest gęstością, więc można ją potraktować jako *rozkład próby*. Przykład takiego histogramu znajduje się na kolejnym rysunku.

```
> hist(dane$waga,freq=FALSE)
```



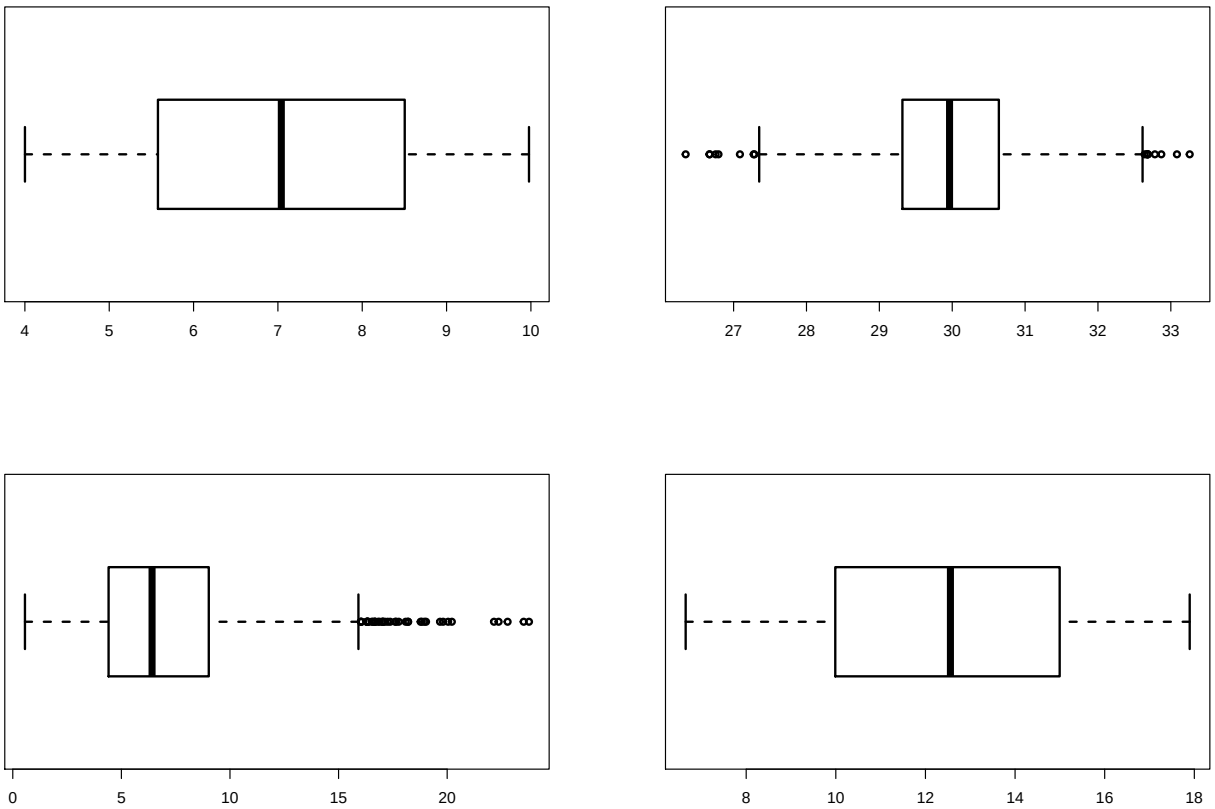
Histogram służy też do opisu rozkładu próby, np. gdy jest symetryczny mówimy, że próba jest symetryczna, jeśli ma dwie ‚górki', że jest dwumodalny, etc. (przykłady na rysunkach 2.1 i 2.2).



Rysunek 2.1: Histogramy odpowiednio dla prób: jednostajnej, symetrycznej jednomodalnej, jednomodalnej skośnej w prawo i dwumodalnej.

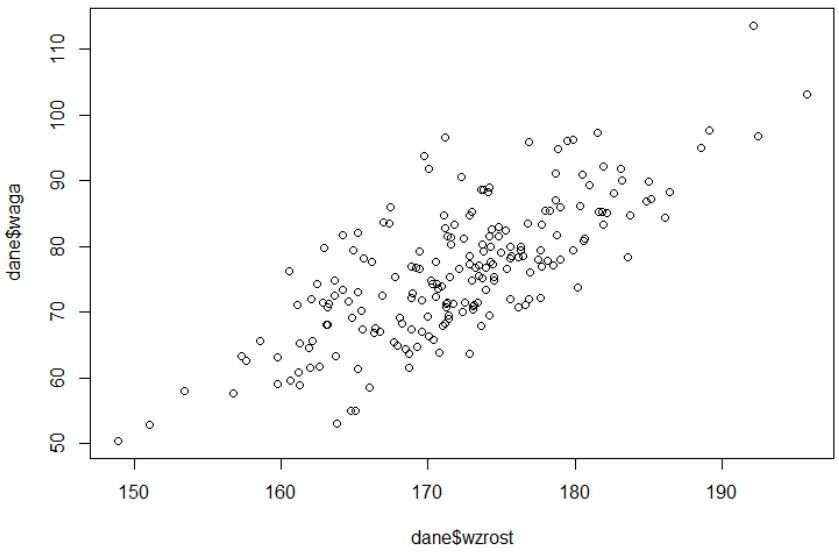
2.2.3 Rysunek rozrzutu. Korelacja liniowa

Rysunek rozrzutu to podstawowe narzędzie do badania zależności między dwoma zmiennymi ilościowymi. Mając próbę dwu zmiennych $(X_1, Y_1), \ldots, (X_n, Y_n)$, rysujemy po prostu punkty o współrzędnych $(X_1, Y_1), \ldots, (X_n, Y_n)$. Dla poprzednich danych dane wygląda to tak:



Rysunek 2.2: Boxploty dla prób z rysunku 2.1.

```
> plot(dane$waga~dane$wzrost)
```



Na tym rysunku widać, że wartości zmiennej waga zmieniają się wraz ze zmianą zmiennej wzrost. Ponadto ta ‚chmurka' punktów przypomina ‚rosnący' pasek, więc ‚na oko' stwierdzamy zależność liniową tych dwóch zmiennych.

Przy okazji zdefiniujemy współczynnik korelacji liniowej Pearsona.

Definicja 2.2. Dla próby $(X_1, Y_1), \ldots, (X_n, Y_n)$ liczbę

$$\rho(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y}$$

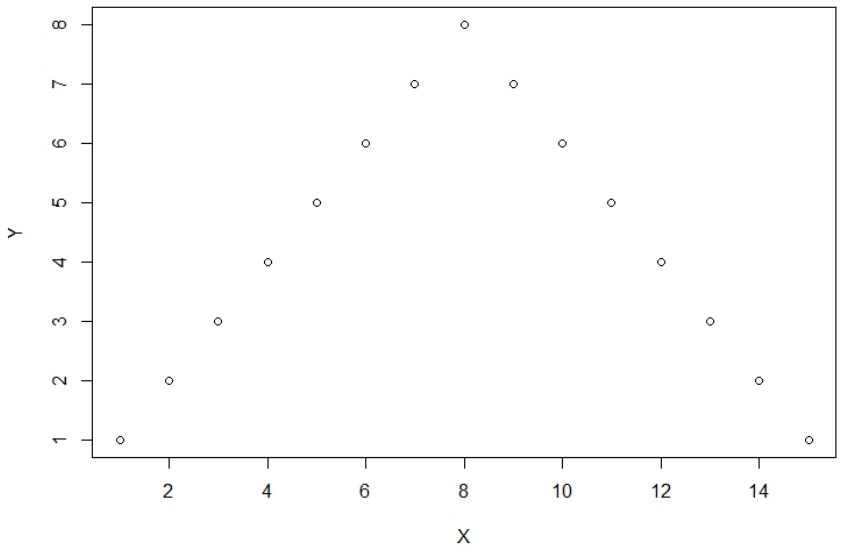
nazywamy *współczynnikiem korelacji liniowej Pearsona*.

Dla poprzednich danych wynosi

```
> cor(dane$waga,dane$wzrost)
[1] 0.7517139
```

Dowodzi się, że $\rho(X,Y) \in [-1,1]$ oraz, że $\rho(X,Y) = 1$ (odpowiednio $\rho(X,Y) = -1$) wtedy i tylko wtedy, gdy istnieją $a > 0$ (odpowiednio $a < 0$) i b takie, że dla wszystkich i mamy $Y_i = aX_i + b$. Wartości tego współczynnika bliskie zera nie świadczą o braku jakiejkolwiek zależności, tylko o braku zależności liniowej. Ilustruje to poniższy przykład.

```
> X=1:15
> Y=c(1,2,3,4,5,6,7,8,7,6,5,4,3,2,1)
> cor(X,Y)
[1] 0
> plot(Y~X)
```



Rozdział 3

Estymacja punktowa

Wracamy do wnioskowania statystycznego. Przypomnijmy punkt wyjścia: badamy populację Ω oraz cechę $X : \Omega \to \mathbb{R}$. Pobieramy próbę osobników $\{\omega_1, \ldots, \omega_n\} \subset \Omega$. Obserwacje cechy X dla tych osobników tworzą próbę $X_1, \ldots, X_n$ ($X_i = X(\omega_i)$). Zakładamy, że rozkłady X_i są takie same jak rozkład cechy X ($P_{X_i} = P_X$). Najbardziej ogólnym celem wnioskowania statystycznego jest znalezienie rozkładu P_X, mając próbę $X_1, \ldots, X_n$. Rozpoczynamy od zdefiniowania rodziny rozkładów, do której potencjalnie należy szukany rozkład P_X.

Definicja 3.1. Rodzinę rozkładów $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ nazywamy *przestrzenią statystyczną*, a Θ *przestrzenią parametrów*.

Innymi słowy, zakładamy, że istnieje ‚prawdziwy' parametr θ_0, tj. taki, że $P_X = P_{\theta_0}$.
W tym miejscu można wyróżnić dwa podstawowe rodzaje wnioskowania statystycznego:
a) *parametryczne*, gdy istnieje $k \in \mathbb{N}$ takie, że $\Theta \subset \mathbb{R}^k$;
b) *nieparametryczne*, gdy dla każdego $k \in \mathbb{N}$ zachodzi $\Theta \not\subset \mathbb{R}^k$.

Definicja 3.2. Próbę $X_1, \ldots, X_n$ nazwiemy *prostą*, gdy zmienne losowe $X_1, \ldots, X_n$ są niezależne.

Przykład 3.3. a) $\mathcal{P} = \{N(\mu, \sigma) \mid \theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+\}$ rodzina składająca się ze wszystkich rozkładów normalnych (wnioskowanie parametryczne);
b) $\mathcal{P} = \{P_f \mid f \in D\}$, gdzie D to zbiór wszystkich gęstości, a P_f rozkład ciągły zadany przez gęstość f (wnioskowanie nieparametryczne).

Wyróżnia się trzy główne metody wnioskowania statystycznego:
1. *Estymację punktową* szukamy takiego $\hat{\theta} \in \Theta$, że $P_X = P_{\hat{\theta}}$;
2. *Estymację przedziałową* gdy $\Theta \subset \mathbb{R}$ szukamy przedziału Θ_1, że

$$P_X \in \{P_\theta \mid \theta \in \Theta_1\};$$

3. *Testowanie hipotez* ustalamy Θ_0 i decydujemy czy

$$P_X \in \{P_\theta \mid \theta \in \Theta_0\}.$$

W tym rozdziale zajmiemy się estymacją punktową (ang. *point estimation*).
Skoro mamy na podstawie próby wskazać jeden parametr $\hat{\theta}$, w naturalny sposób zrobimy to definiując funkcję $(X_1, \ldots, X_n) \to \Theta$. My zdefiniujemy estymator jeszcze nieco ogólniej.

Definicja 3.4. Ustalmy $\mathcal{P} = \{P_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$, $g : \mathbb{R}^d \to \mathbb{R}^k$ oraz próbę $X_1, \ldots, X_n$. Wtedy $T(X_1, \ldots, X_n)$ nazywamy *estymatorem* $g(\theta)$, gdy funkcja $T : \mathbb{R}^n \to \mathbb{R}^k$ jest mierzalna (borelowska).

Często będziemy szukać estymatora θ, tj. dla funkcji $g(\theta) = \theta$. W tej sytuacji często estymator parametru θ oznacza się przez $\hat{\theta}$ lub $\hat{\theta}_n$. Funkcja g jest przydatna, gdy chcemy estymować parametry zmiennej losowej P_X, np. wartości oczekiwanej (np. gdy $X \sim U(a, b)$, $\theta = (a, b)$, to bierzemy $g((a, b)) = (a + b)/2$).

Zauważmy, że estymator $T(X_1, \ldots, X_n)$ traktujemy jako zmienną losową, a jeśli mamy konkretne wartości liczbowe w próbie, to $T(X_1, \ldots, X_n)$ jest elementem Θ, który ma przybliżać prawdziwą wartość parametru θ (ta wartość zwana jest oceną punktową lub estymatem).

3.1 Metoda największej wiarygodności

Nie będziemy na razie odpowiadać na pytanie jaki estymator jest ‚dobry', zaczniemy od metod wyznaczania estymatorów. Jedną z podstawowych metod jest metoda *największej wiarygodności*. Ze względów praktycznych zdefiniujemy ją osobno dla rozkładów dyskretnych i ciągłych.

3.1.1 Dla rozkładów dyskretnych

Załóżmy, że przestrzeń statystyczna $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ składa się z rozkładów dyskretnych ($P_\theta(A) = P_\theta(X \in A)$). Definiujemy wtedy *funkcję wiarygodności*[1] zmiennej θ wzorem

$$L(\theta|X_1, \ldots, X_n) = P_\theta(X_1) \cdots P_\theta(X_n).$$

Gdy próba $(X_1, \ldots, X_n)$ jest prosta, to $L(\theta|X_1, \ldots, X_n)$ jest prawdopodobieństwem tej próby przy założeniu, że $P_{X_i} = P_\theta$ dla każdego i.

Definicja 3.5. Estymatorem największej wiarygodności parametru θ nazywamy

$$ENW(\theta) = \hat{\theta}(X_1, \ldots, X_n) = \operatorname{argmax}_{\theta\in\Theta} L(\theta|X_1, \ldots, X_n).$$

Często rozważa się funkcję *log-wiarygodności* daną wzorem

$$l(\theta|\ldots) = \log(L(\theta|\ldots)).$$

Zachodzi proste twierdzenie (ponieważ logarytm jest funkcją rosnącą)

$$\operatorname{argmax}_{\theta\in\Theta} L(\theta|X_1, \ldots, X_n) = \operatorname{argmax}_{\theta\in\Theta} l(\theta|X_1, \ldots, X_n).$$

Często prościej jest znaleźć maksima funkcji log-wiarygodności niż funkcji wiarygodności. Podamy teraz najprostszy przykład.

Przykład 3.6. (Model ankiety) Najpierw zauważmy, że sformułowania a), b) i c) są równoważne.

a) Wyznaczyć estymator największej wiarygodności parametru p w rozkładzie zero-jedynkowym $(1, 0, p)$ dla próby prostej $X_1, \ldots, X_n$.

b) $X_1, \ldots, X_n$ $i.i.d$[2] $(1, 0, p)$. Wyznacz $ENW(p)$.

c) Badamy cechę X o rozkładzie zero-jedynkowym $(1, 0, p)$ o nieznanej wartości parametru p (tzn. przestrzeń statystyczna składa się z rozkładów zero-jedynkowych). Pobrano próbę prostą $X_1, \ldots, X_n$. Wyznacz estymator największej wiarygodności parametru p.

Rozkład zero-jedynkowy możemy zapisać jako $P_p(x) = p^x(1-p)^{1-x}$ dla $x = 0, 1$. Przestrzeń parametrów to $[0, 1]$. Załóżmy, że w próbie występuje co najmniej jedno zero i jedna jedynka. Wtedy

$$L(p|X_1, \ldots, X_n) = p^{X_1}(1-p)^{1-X_1} \cdots p^{X_n}(1-p)^{1-X_n} = p^{\sum_i X_i}(1-p)^{n-\sum_i X_i},$$

$$l(p|X_1, \ldots, X_n) = \sum_i X_i \log(p) + (n - \sum_i X_i) \log(1-p).$$

Krótkie przeliczenie pokazuje, że maksimum lokalne jest w punkcie $\hat{p} = \frac{\sum_i X_i}{n}$, a ponieważ $L(0) = L(1) = 0$, stwierdzamy, że $ENW(p) = \frac{\sum_i X_i}{n}$.

Przykład 3.7. Załóżmy, że mamy dwie próby proste z różnych rozkładów Poissona: $X_1, \ldots, X_n \sim \mathcal{P}(\lambda)$ oraz $Y_1, \ldots, Y_m \sim \mathcal{P}(3\lambda)$. Wyznaczyć $ENW(\lambda)$.

Rozkład Poissona z parametrem λ dany jest wzorem $P_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!}$ dla $x \in \mathbb{N}$, tak więc funkcja wiarygodności będzie dana formułą

$$L(\lambda|X_1, \ldots, X_n, Y_1, \ldots, Y_m) = e^{-\lambda}\frac{\lambda^{X_1}}{X_1!} \cdots e^{-\lambda}\frac{\lambda^{X_n}}{X_n!} e^{-3\lambda}\frac{(3\lambda)^{Y_1}}{Y_1!} \cdots e^{-3\lambda}\frac{3\lambda^{Y_m}}{Y_m!}.$$

Znowu rozważając funkcję log-wiarygodności i po krótkich przeliczeniach, otrzymujemy wynik

$$ENW(\lambda) = \frac{\sum_i X_i + \sum_j Y_j}{n + 3m}.$$

[1] ang. *likelihood function*

[2] ang. independent, identically distributed.

Ostatni przykład dotyczy *obserwacji obciętych*, tzn. takich, które nie są liczbą, a przedziałem.

Przykład 3.8. Chcemy oszacować średnią liczbę drobnoustrojów na litr wody w basenie. Test, który mamy, jest tylko w stanie stwierdzić, czy w danej próbce są albo nie ma drobnoustrojów. Pobrano losowo 10 próbek po 20ml wody i w 7 z nich stwierdzono drobnoustroje. Ile średnio drobnoustrojów znajduje się w litrze?

Oznaczmy przez X liczbę drobnoustrojów w 20ml wody. Możemy przyjąć, że $X \sim \mathcal{P}(\lambda)$, czyli średnio w 20ml wody jest λ drobnoustrojów. W litrze więc będzie średnio 50λ drobnoustrojów.

Nie mamy dokładnych pomiarów zmiennej losowej X, wiemy tylko, że $X_1, \ldots, X_7 \in [1, +\infty)$ i $X_8 = X_9 = X_{10} = 0$. Funkcja wiarygodności przyjmie postać

$$L(\lambda) = (P_\lambda(X > 0))^7 (P_\lambda(X = 0))^3 = \left(1 - e^{-\lambda}\frac{\lambda^0}{0!}\right)^7 \left(e^{-\lambda}\frac{\lambda^0}{0!}\right)^3$$

$$L(\lambda) = (1 - e^{-\lambda})^7 (e^{-\lambda})^3$$

Podstawiając $t = e^{-\lambda}$, uzyskujemy, że maksimum jest w punkcie $\hat{t} = 3/10$. Ostatecznie więc $50\hat{\lambda} = 50(-\log(3/10)) \approx 60.19864$.

3.1.2 Dla rozkładów ciągłych

Estymator największej wiarygodności dla rozkładów ciągłych definiuje się analogicznie.

Rozważmy przestrzeń statystyczną $\mathcal{P} = \{f_\theta \mid \theta \in \Theta\}$, składającą się z rozkładów ciągłych danych przez gęstości f_θ. Funkcję wiarygodności definiujemy wzorem wzorem

$$L(\theta|X_1, \ldots, X_n) = f_\theta(X_1) \cdots f_\theta(X_n).$$

W sytuacji danych obciętych, gdy $X_i \in A$, w powyższym wzorze składnik $f_\theta(X_i)$ zastępujemy przez $P_\theta(A)$. Definicja estymatora największej wiarygodności jest ta sama jak dla rozkładów dyskretnych. Definicja funkcji log-wiarygodności oraz wspomniana jej własność oczywiście też zachodzi.

Zaczniemy od standardowego przykładu.

Przykład 3.9. Na podstawie próby prostej $X_1, \ldots, X_n$ wyznaczyć $ENW(a)$ w rozkładzie jednostajnym $\mathcal{U}[0, a]$ $(a > 0)$.

Rozkład $\mathcal{U}[0, a]$ jest dany gęstością $f_a(x) = \frac{1}{a}\chi_{[0,a]}(x)$ (χ_A to funkcja charakterystyczna zbioru A). Funkcja wiarygodności przyjmuje więc postać

$$L(a) = \frac{1}{a}\chi_{[0,a]}(X_1) \cdots \frac{1}{a}\chi_{[0,a]}(X_n) = \frac{1}{a^n}\chi_{[0,a]}(X_1) \cdots \chi_{[0,a]}(X_n).$$

Jeśli istnieje i takie, że $X_i < 0$ to $L \equiv 0$ i $ENW(a) = (0, +\infty)$ (zauważmy jednak, że gdy istnieje ujemna obserwacja w próbie, to założenie, że cecha ma rozkład $\mathcal{U}[0, a]$, jest bez sensu).

Załóżmy więc, że wszystkie X_i są dodatnie. Wtedy

$$\chi_{[0,a]}(X_1) \cdots \chi_{[0,a]}(X_n) = 1$$

wtedy i tylko wtedy, gdy $X_i \leqslant a$ dla wszystkich i, czyli gdy
$\max(X_1, \ldots, X_n) \leqslant a$. Ostatecznie więc

$$L(a) = \begin{cases} \frac{1}{a^n}, & \text{jeśli } \max(X_1, \ldots, X_n) \leqslant a \\ 0, & \text{jeśli } \max(X_1, \ldots, X_n) > a \end{cases}$$

Estymatorem największej wiarygodności parametru a jest więc $ENW(a) = \max(X_1, \ldots, X_n)$.

Kolejny przykład będzie dotyczył danych obciętych.

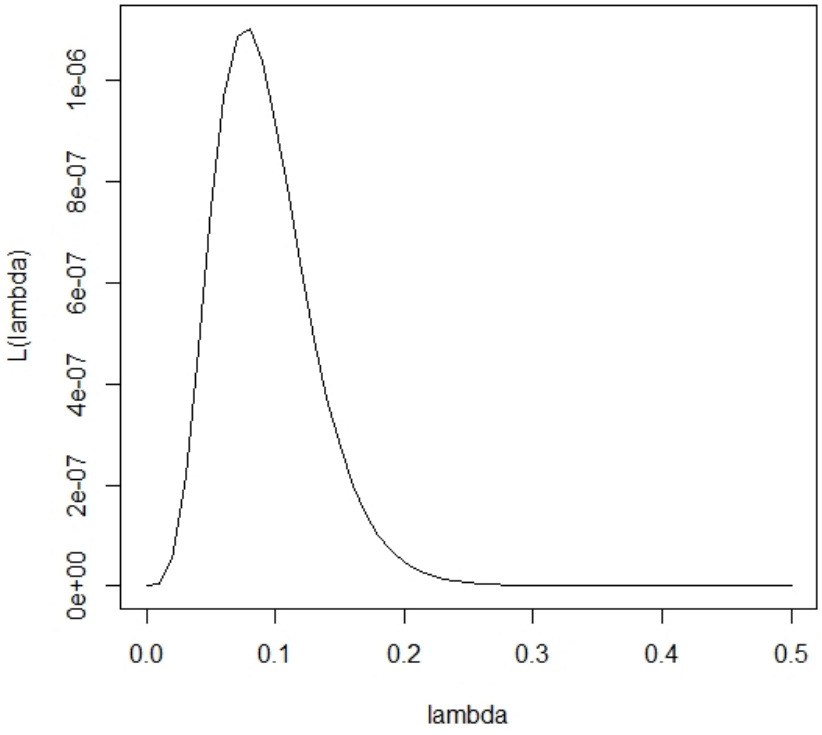
Przykład 3.10. Badamy X czas (w godzinach) bezawaryjnej pracy pewnego typu żarówki w ustalonych warunkach. Wylosowaliśmy 6 sztuk (z założenia z dużej liczby egzemplarzy), włączyliśmy je i poszliśmy do domu. Po powrocie po 8 godzinach okazało się, że dwie z nich są już przepalone. Kolejne żarówki przepalały się odpowiednio w 12, 13 i 15 godzinie pracy. W 18-ej godzinie eksperymentu musieliśmy go przerwać, a ostatnia żarówka jeszcze świeciła. Zakładając, że X ma rozkład wykładniczy $Exp(\lambda)$, chcemy wyznaczyć estymator największej wiarygodności parametru λ.

Gęstością rozkładu $Exp(\lambda)$ jest $f_\lambda(x) = \lambda e^{-\lambda x}\chi_{[0,+\infty]}(x)$, a dystrybuantą $F_\lambda(t) = 1 - e^{-\lambda t}$ dla $t > 0$. Próba jest następująca: $X_1 = X_2 \in [0, 8]$, $X_3 = 12$, $X_4 = 13$, $X_5 = 15$ oraz $X_6 \in [18, +\infty]$. Funkcja wiarygodności jest więc postaci

$$L(\lambda) = (1 - e^{-\lambda 8})^2 \lambda e^{-\lambda 12} \lambda e^{-\lambda 13} \lambda e^{-\lambda 15} e^{-\lambda 18} = (1 - e^{-8\lambda})^2 \lambda^3 e^{-58\lambda}.$$

Analityczne wyznaczenie maksimum tej funkcji może być kłopotliwe, więc zrobimy to numerycznie. Na początek warto narysować wykres.

```
> L=function(lambda) (1-exp(-8*lambda))^2*lambda^3*exp(-58*lambda)
> lambda=seq(0,0.5,by=0.01)
> plot(lambda,L(lambda),type="l")
```



Wyznaczamy teraz estymator numerycznie.

```
> (O=optimize(L,c(0,0.3),maximum=T))
$maximum
[1] 0.07668633

$objective
[1] 1.109826e-06

> 1/O$maximum
[1] 13.04013
```

Otrzymujemy więc, że $ENW(\lambda) \approx 0.0766$ oraz na przykład to, że średni czas bezawaryjnej pracy tej żarówki wynosi około 13.04 godziny. Pozostaje pytanie, czy poza przedziałem $(0, 0.3)$ nie istnieje inne maksimum?(Ćwiczenie)

Przykład 3.11. Mierzyliśmy temperaturę (w stopniach Celsiusza) pewnego obiektu. Wykonaliśmy 7 niezależnych pomiarów i uzyskaliśmy $1.2, 0.9, 0.7, 0.8, 0.77, 2.1, 1.34$ stopni. Niestety po eksperymencie okazało się, że w termometrze zepsuł się wyświetlacz i nie działa znak ‚minus' oraz odchylenie błędu pomiarowego wynosiło aż 1 stopień. Jaką temperaturę miał badany obiekt, jeśli założymy, że temperatura była dodatnia?

Oznaczmy badaną cechę (temperaturę) przez X. Zakładamy, że błędy pomiarowe mają rozkład normalny, więc X ma rozkład normalny $N(m, 1)$, gdzie m oznacza prawdziwą temperaturę badanego obiektu. Jednak próba pochodzi z pomiarów cechy $|X|$. Wyznaczmy na początek rozkład zmiennej $|X|$ (przez Φ i f oznaczmy odpowiednio dystrybuantę i gęstość standardowego rozkładu normalnego $N(0, 1)$, a przez $\Phi_{m,\sigma}$ dystrybuantę rozkładu normalnego $N(m, \sigma)$). Dla $t > 0$ mamy

$$F_{|X|}(t) = P(|X| < t) = P(X \in (-t, t)) = \Phi_{m,1}(t) - \Phi_{m,1}(-t) =$$

$$= \Phi(t - m) - \Phi(t + m).$$

Czyli gęstość zmiennej losowej $|X|$ jest równa

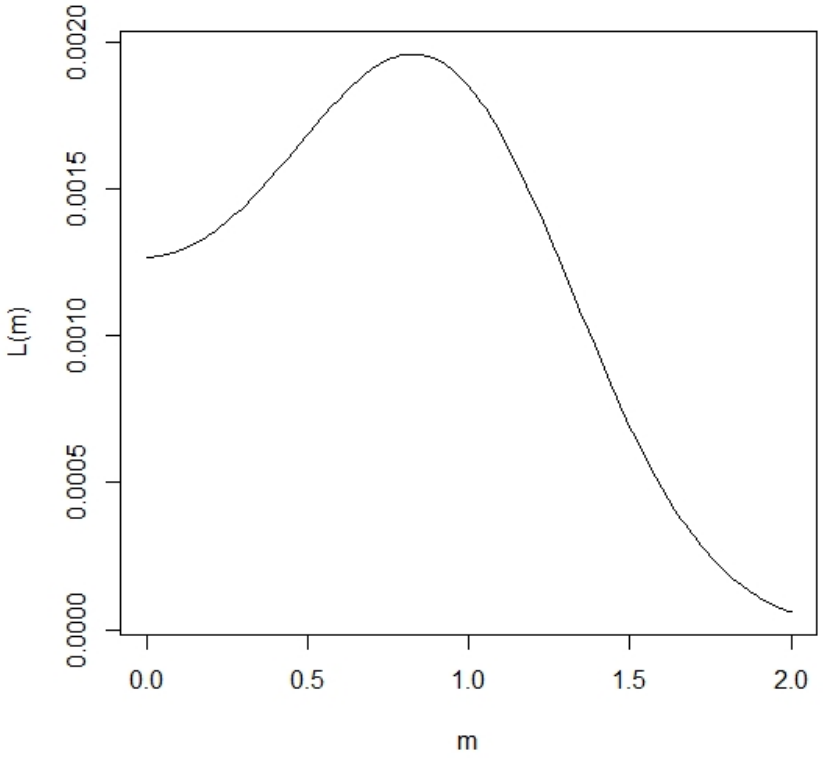
$$g(t) = F'_{|X|}(t) = \Phi'(t - m) + \Phi'(-t - m) = f(t - m) + f(-t - m).$$

Funkcja wiarygodności będzie dana wzorem

$$L(m) = \prod_{i=1}^{6} \Big[\frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - m)^2)}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(-X_i - m)^2)}{2}} \Big].$$

Wyznaczymy maksimum numerycznie.

```
> X=c(1.2,0.9,0.7,0.8,0.77,2.1,1.34)
> L=function(m){
+   n=length(m)
+   w=numeric(n)
+   for (i in 1:n)  w[i]=prod(dnorm(X-m[i])+dnorm(-X-m[i]))
+   return(w)
+ }
> m=seq(0,2,by=0.01)
> plot(m,L(m),type="l")
```




```
> O=optimize(L,c(0,2),maximum=T)
> O[[1]]
[1] 0.8275602
```

Ostatecznie $EMW(m) \approx 0.8275602$. Podobnie jak poprzednio pozostaje pytanie, czy poza przedziałem $(0, 2)$ nie istnieje inne maksimum (Ćwiczenie).

3.2 Metoda momentów

Przedstawimy teraz drugą (a historycznie pierwszą) metodę wyznaczania estymatorów. Na początek przypomnijmy definicje i-tego momentu (momentu rzędu i) zmiennej losowej

$$m_i = E(X^i)$$

(o ile istnieje) oraz i-tego momentu z próby $(X_1, \ldots, X_n)$

$$\hat{m}_i = \frac{1}{n} \sum_{s=1}^{n} X_s^i.$$

W kontekście wnioskowania statystycznego momenty badanej cechy nazywa się momentami *teoretycznymi*, a momenty z próby *empirycznymi*. Metoda momentów szuka parametrów, dla których i-te momenty teoretyczne i empiryczne są równe.

Bardziej precyzyjnie, załóżmy, że nasza przestrzeń statystyczna $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ jest parametryzowana k-wymiarowym parametrem $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k$. Załóżmy dalej, że dla każdego $i = 1, \ldots, k$ istnieje funkcja g_i taka, że moment rzędu i zmiennej losowej X o rozkładzie $P_{(\theta_1,\ldots,\theta_k)}$ jest równy $g_i((\theta_1, \ldots, \theta_k))$ dla wszystkich $(\theta_1, \ldots, \theta_k) \in \Theta$. Niech $X_1, \ldots, X_n$ będzie pobraną próbą. *Estymatorem momentów* nazywamy $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$, który należy do Θ i jest rozwiązaniem układu k równań

$$\begin{cases} \hat{m}_1 = g_1((\theta_1, \ldots, \theta_k)) \\ \vdots \\ \hat{m}_k = g_k((\theta_1, \ldots, \theta_k)) \end{cases}$$

Zaczniemy od najprostszego przykładu.

Przykład 3.12. Wyznaczymy estymator momentów w rozkładzie $\mathcal{U}[0, \alpha]$ $(\alpha > 0)$.

W tym problemie mamy $k = 1$. Wtedy $m_1 = E(X)$, gdy $X \sim \mathcal{U}[0, \alpha]$, czyli $g_1(\alpha) = \alpha/2$. Pierwszy moment z próby to z definicji średnia z próby. Ostatecznie mamy równanie

$$\bar{X} = \frac{\alpha}{2}.$$

Rozwiązaniem jest $\hat{\alpha} = 2\bar{X}$, o ile $\bar{X} > 0$. Zauważmy, że estymator momentów wyszedł inny niż estymator największej wiarygodności.

Rozważmy następny prosty przykład.

Definicja 3.13. Wyznaczmy estymator momentów parametru λ w rozkładzie wykładniczym $Exp(\lambda)$. Pierwszy moment teoretyczny (czyli wartość oczekiwana) wynosi

$$m_1 = \frac{1}{\lambda}.$$

Otrzymujemy równanie

$$\bar{X} = \frac{1}{\lambda}.$$

Estymatorem momentów jest więc

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

(o ile $\bar{X} > 0$). W tym przypadku estymator momentów jest taki sam jak największej wiarygodności.

Następny przykład będzie nieco bardziej złożony.

Przykład 3.14. Wyznaczymy estymator momentów w rozkładzie $\mathcal{U}[a, b]$ $(b > a > 0)$.

Mamy do wyestymowania dwa parametry liczbowe $(k = 2)$. Wyznaczamy pierwszy i drugi moment teoretyczny:

$$m_1 = E(X) = \frac{a+b}{2} =: g_1(a, b);$$

$$m_2 = E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3}(b^2 + ab + a^2) = g_2(a, b).$$

Otrzymujemy ostatecznie układ dwóch równań

$$\begin{cases} \hat{m}_1 = \frac{a+b}{2} \\ \hat{m}_2 = \frac{1}{3}(b^2 + ab + a^2) \end{cases}$$

Po krótkich przekształceniach mamy

$$\begin{cases} a + b = 2\hat{m}_1 \\ ab = 4\hat{m}_1^2 - 3\hat{m}_2 \end{cases}$$

Przez podstawienie otrzymujemy równanie kwadratowe ‚na a'. Jeśli $\hat{m}_2 - \hat{m}_1^2 > 0$ (inaczej wariancja z próby $s_X^2 = \hat{m}_2 - \hat{m}_1^2 > 0$), to mamy dwa rozwiązania, ale tylko jedno spełnia warunek $b > a$

$$\begin{cases} \hat{a} = \hat{m}_1 - \sqrt{3(\hat{m}_2 - \hat{m}_1^2)} = \bar{X} - \sqrt{3} s_X \\ \hat{b} = \hat{m}_1 + \sqrt{3(\hat{m}_2 - \hat{m}_1^2)} = \bar{X} + \sqrt{3} s_X \end{cases}$$

Przed kolejnym przykładem zdefiniujemy rozkład mieszany.

Definicja 3.15. Niech f_1, f_2 będą gęstościami. Wtedy rozkład nazywamy *mieszanym* (rozkładów f_1 i f_2), gdy jego gęstość jest kombinacją wypukłą f_1 i f_2, tzn. $f(x) = \varepsilon f_1(x) + (1 - \varepsilon) f_2(x)$ dla pewnego $\varepsilon \in [0, 1]$ (ozn. (f_1, f_2, ε)).

Przykład 3.16. Załóżmy, że badana cecha X ma rozkład mieszany (f_1, f_2, ε), gdzie f_1 i f_2 są gęstościami odpowiednio rozkładów normalnych $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$. Wyestymować parametr ε, zakładając, że $m_1, \sigma_1, m_2, \sigma_2$ są znane.

Estymator największej wiarygodności parametru ε jest trudny do wyznaczenia, natomiast estymator momentów jest prosty. Mamy bowiem tylko jeden parametr do wyestymowania. Wyznaczmy pierwszy moment teoretyczny:

$$m_1 = \int_{-\infty}^{\infty} x(\varepsilon f_1(x) + (1 - \varepsilon) f_2(x)) dx = \varepsilon m_1 + (1 - \varepsilon) m_2.$$

Otrzymujemy równanie

$$\bar{X} = \varepsilon m_1 + (1 - \varepsilon) m_2.$$

Szukanym estymatorem jest więc

$$\hat{\varepsilon} = \frac{\bar{X} - m_2}{m_1 - m_2}$$

o ile powyższy $\hat{\varepsilon} \in [0, 1]$.

3.3 Własności teoretyczne estymatorów

Rozważmy przestrzeń statystyczną $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ oraz próbę prostą $\underline{X} = (X_1, \ldots, X_n)$ wylosowaną z rozkładu P_θ dla pewnego (nieznanego nam) parametru θ. Chcemy teraz na podstawie próby oszacować wartość tego nieznanego (‚prawdziwego') parametru θ lub ogólniej wartość $g(\theta)$ dla pewnej ustalonej funkcji $g : \mathbb{R}^k \to \mathbb{R}^d$ (na przykład, gdy P_λ to rozkład wykładniczy o parametrze λ, a my chcemy oszacować wartość oczekiwaną rozkładu z którego pochodzi próba, to formalnie estymujemy $g(\lambda) = 1/\lambda$).

Definicja 3.17. Funkcję mierzalną $\hat{\theta}_n : (X_1, \ldots, X_n) \to \mathbb{R}^d$ nazywamy *estymatorem parametru* $g(\theta)$.

Powstaje pytanie jakie estymatory są ‚dobre'. Zdefiniujemy pewne własności, które uznamy za ‚pożądane' oraz kryterium do porównywania estymatorów.

Zgodność estymatorów

Rozważmy ciąg estymatorów $(\hat{\theta}_n)_{n=1}^\infty$ parametru $g(\theta)$. Intuicyjnie, zgodność ciągu estymatorów oznacza, że dla każdego (‚prawdziwego') $\theta \in \Theta$ wraz ze wzrostem liczności próby n estymatory ‚dążą do $g(\theta)$'. Ponieważ estymatory to zmienne losowe, to zbieżność estymatorów można zdefiniować na różne sposoby. Przedstawimy dwa z nich.

Definicja 3.18. Ciąg estymatorów $(\hat{\theta}_n)_{n=1}^\infty$ parametru $g(\theta)$ nazywamy

1. *mocno zgodnym*, gdy dla każdego $\theta \in \Theta$
$$P_\theta(\lim_{n\to\infty} \hat{\theta}_n = g(\theta)) = 1;$$
2. *słabo zgodnym*[3], gdy dla każdego $\theta \in \Theta$ i każdego $\varepsilon > 0$
$$\lim_{n\to\infty} P_\theta(|\hat{\theta}_n - g(\theta)| > \varepsilon)) = 0.$$

Można pokazać, że mocna zgodność implikuje słabą zgodność.

Przykład 3.19. Załóżmy, że próba prosta $(X_1, \ldots, X_n)$ pochodzi z rozkładu ciągłego P_θ danego gęstością f_θ, takiego, że $\theta \in \Theta \subset \mathbb{R}$, a dla każdego θ istnieje wartość średnia $E_\theta(X) = \int_{\mathbb{R}} x f_\theta(x)dx$. Zdefiniujmy $g(\theta) = E_\theta(X)$. Wtedy z prawa wielkich liczb, estymator
$$\hat{\theta}_n(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$$
są mocno zgodnymi estymatorami $g(\theta)$.

Przykład 3.20. Załóżmy, że próba prosta $(X_1, \ldots, X_n)$ pochodzi z rozkładu jednostajnego $\mathcal{U}[0, \alpha]$, dla $\alpha > 0$. Rozważmy estymator $\hat{\alpha}_n = \max(X_1, \ldots, X_n)$ $(n \geqslant 1)$ parametru α (tutaj formalnie $g(\alpha) = \alpha$).

Najpierw znajdziemy rozkład estymatora $\hat{\alpha}_n$ dla ustalonego n i α przez wyznaczenie jego dystrybuanty. Mamy
$$F_{\hat{\alpha}_n}(t) = P(\hat{\alpha}_n < t) = P(\max(X_1, \ldots, X_n) < t) = P(X_1 < t, \ldots, X_n < t)$$
$$= P(X_1 < t) \cdots P(X_n < t) = \left(\frac{1}{\alpha}t\right)^n,$$
dla $t \in [0, \alpha]$ i $F_{\hat{\alpha}_n}(t) = 1$ dla $t > \alpha$.

Ustalmy teraz $\varepsilon > 0$. Jeśli $\varepsilon \geqslant \alpha$ to od razu dostajemy, że
$$P(|\hat{\alpha}_n - \alpha| \leqslant \varepsilon) = 1.$$
W przypadku, gdy $\varepsilon < \alpha$, mamy
$$P(|\hat{\alpha}_n - \alpha| \leqslant \varepsilon) = P(\alpha - \varepsilon \leqslant \hat{\alpha}_n \leqslant \alpha + \varepsilon) = F_{\hat{\alpha}_n}(\alpha + \varepsilon) - F_{\hat{\alpha}_n}(\alpha - \varepsilon)$$
$$= 1 - \left(\frac{1}{\alpha}(\alpha - \varepsilon)\right)^n = 1 - \left(1 - \frac{\varepsilon}{\alpha}\right)^n.$$
Ostatecznie otrzymujemy, że
$$\lim_{n\to\infty} P(|\hat{\alpha}_n - \alpha| > \varepsilon) = 1 - \lim_{n\to\infty} P(|\hat{\alpha}_n - \alpha| \leqslant \varepsilon) = 1.$$
Estymator $\hat{\alpha}_n$ jest więc estymatorem słabo zgodnym parametru α w rozkładzie $\mathcal{U}[0, \alpha]$.

[3] ang. *weakly consistent*.

Błąd średniokwadratowy, nieobciążoność estymatora

Przedstawimy kryterium liczbowe oceny dokładności estymacji parametru rzeczywistego $g(\theta)$.

Definicja 3.21. *Błędem średniokwadratowym*[4] estymatora $\hat{\theta}_n$ parametru $g(\theta)$ nazywamy

$$MSE_\theta(\hat{\theta}_n) = E_\theta((\hat{\theta}_n - g(\theta))^2).$$

Twierdzenie 3.22. *Błąd średniokwadratowy jest równy*

$$MSE_\theta(\hat{\theta}_n) = D^2_\theta(\hat{\theta}_n) + b^2(\hat{\theta}_n, g(\theta)),$$

gdzie $b(\hat{\theta}_n, g(\theta)) = E_\theta(\hat{\theta}_n) - g(\theta)$ *nazywamy* obciążeniem[5] *estymatora* $\hat{\theta}_n$.

Dowód.

$$\begin{aligned}
MSE_\theta(\hat{\theta}_n) &= E_\theta((\hat{\theta}_n - g(\theta))^2) = E_\theta\left((\hat{\theta}_n - E_\theta(\hat{\theta}_n) + E_\theta(\hat{\theta}_n) - g(\theta))^2\right) \\
&= E_\theta\left((\hat{\theta}_n - E_\theta(\hat{\theta}_n))^2 + 2(\hat{\theta}_n - E_\theta(\hat{\theta}_n))(g(\theta) - E_\theta(\hat{\theta}_n)) + (g(\theta) - E_\theta(\hat{\theta}_n))^2\right) \\
&= E_\theta\left((\hat{\theta}_n - E_\theta(\hat{\theta}_n))^2\right) + (g(\theta) - E_\theta(\hat{\theta}_n))E_\theta(\hat{\theta}_n - E_\theta(\hat{\theta}_n)) + E_\theta((g(\theta) - E_\theta(\hat{\theta}_n))^2) \\
&= D^2_\theta(\hat{\theta}_n) + (g(\theta) - E_\theta(\hat{\theta}_n))^2 = D^2(\hat{\theta}_n) + (E_\theta(\hat{\theta}_n) - g(\theta))^2 \\
&= D^2_\theta(\hat{\theta}_n) + b^2(\hat{\theta}_n, g(\theta)).
\end{aligned}$$

Błąd średniokwadratowy estymatora jest więc sumą jego wariancji i kwadratu obciążenia. Łatwo widać też, że $MSE_\theta(\hat{\theta}_n) \geqslant 0$. □

Definicja 3.23. Estymator T jest lepszy od estymatora W, gdy istnieje θ taka, że $MSE_\theta(T) < MSE_\theta(W)$ oraz $MSE_\theta(T) \leqslant MSE_\theta(W)$ dla każdego θ.

Definicja 3.24. Estymator $\hat{\Theta}$ nazywamy dopuszczalnym, gdy nie istnieje estymator lepszy od niego. Nazywamy go zaś niedopuszczalnym, gdy nie jest dopuszczalny.

Przykład 3.25. Załóżmy, że $\Theta = \mathbb{R}$. Zdefiniujmy dwa trywialne[6] estymatory parametru θ: $T(X_1, \ldots, X_n) = \theta_1$ i $W(X_1, \ldots, X_n) = \theta_2$ dla pewnych $\theta_1 < \theta_2$. Wtedy:

1. $MSE_{\theta_1}(T) = MSE_{\theta_2}(W) = 0$ (oraz $D^2_\theta(T) = 0$), ale

   $$MSE_\theta(T) = (\theta_1 - \theta)^2 \neq (\theta_2 - \theta)^2 = MSE_\theta(W),$$

   dla $\theta \in \mathbb{R} \setminus \{(\theta_1 + \theta_2)/2\}$. Czyli ani T nie jest lepszy od W, ani W nie jest lepszy od T.

2. Jeśli $\hat{\theta}$ to dowolny estymator taki, że $MSE_{\theta_1}(\hat{\theta}) > 0$, to nie jest on lepszy od T.

Powyższe przykłady pokazują, że nie ma sensu szukać estymatora, który jest najlepszy dla wszystkich $\theta \in \Theta$. Dlatego szuka się takich estymatorów na pewnych podzbiorach (‚klasach') estymatorów (ale uwaga, estymator dopuszczający w pewnej klasie nie musi być dopuszczalny w klasie szerszej). Najczęściej tą klasą są estymatory *nieobciążone*, których definicja jest następująca.

Definicja 3.26. Estymatory $(\hat{\theta}_n)_{n=1}^\infty$ parametru $g(\theta)$ nazywamy

1. *nieobciążonymi*, gdy $E_\theta(\hat{\theta}_n) = g(\theta)$ (tj. obciążenie jest równe zero) dla każdego n i θ;

2. *asymptotycznie nieobciążonymi*, gdy $\lim_{n\to\infty} E_\theta(\hat{\theta}_n) = g(\theta)$ dla każdego θ.

Przykład 3.27. Załóżmy, że próba $(X_1, \ldots, X_n)$ pochodzi z rozkładu P_θ, takiego, że istnieje jego średnia μ. Wtedy

$$E_\theta(\bar{X}) = \frac{1}{n}\sum_{i=1}^n E_\theta(X_i) = \frac{1}{n} n\mu = \mu.$$

Czyli średnia z próby jest estymatorem nieobciążonym wartości średniej.

[4] ang. *mean square error*.
[5] ang. *bias*.
[6] (by nie rzec ‚patologiczne')

W kilku punktach przedstawmy podstawowe własności estymatorów nieobciążonych.

1. Estymator nieobciążony nie musi istnieć, na przykład nie istnieje estymator nieobciążony parametru $g(p) = p^2$ w rozkładzie $(1, 0, p)$ dla prób jednoelementowych, albo parametru $g(p) = 1/p$ w rozkładzie $b(n, p)$.

2. Jeśli T i W to estymatory nieobciążone parametru $g(\theta)$, to dla każdego $\alpha \in [0, 1]$ estymator $Z_\alpha = \alpha T + (1-\alpha)W$ też jest nieobciążony (czyli jeśli istnieją już dwa różne estymatory nieobciążone pewnego parametru, to istnieje ich już nieskończenie wiele).

3. Jeśli $h(x) = ax+b$, a T jest estymatorem nieobciążonym parametru $g(\theta)$, to $h(T)$ jest estymatorem nieobciążonym parametru $h(g(\theta))$. Ale uwaga, własność ta nie musi zachodzić dla dowolnej funkcji h, na przykład wariancja z próby $s^2 = (1/(n-1)) \sum_{i=1}^n (X_i - \bar{X})^2$ jest estymatorem nieobciążonym wariancji rozkładu (jeśli ta istnieje), ale $\sqrt{s^2}$ nie jest estymatorem nieobciążonym odchylenia standardowego rozkładu.

4. Jeśli T jest estymatorem nieobciążonym parametru $g(\theta)$, to wprost z definicji

$$MSE_\theta(T) = D^2_\theta(T).$$

Stąd na przykład wniosek dla dwóch estymatorów nieobciążonych T i W, że jeśli $D^2_\theta(T) < D^2_\theta(W)$ dla wszystkich θ, to T jest lepszy od W.

To uzasadnia definicję estymatora o najmniejszej wariancji.

Definicja 3.28. Oznaczmy przez U zbiór estymatorów nieobciążonych parametru $g(\theta)$ i załóżmy, że jest niepusty. Wtedy estymator nieobciążony T parametru $g(\theta)$ nazywamy *estymatorem nieobciążonym o minimalnej wariancji*[7] (ozn.
$ENMW(g(\theta))$, gdy $E_\theta(T^2) < \infty$ oraz dla każdego $W \in U$ i każdego $\theta \in \Theta$

$$D^2_\theta(T) \leqslant D^2_\theta(W).$$

Przykład 3.29. Rozważmy dwa estymatory wartości oczekiwanej μ rozkładu P_θ o średniej μ i skończonej wariancji σ^2

$$T(X_1, \ldots, X_n) = \bar{X}_n$$

oraz

$$W(X_1, \ldots, X_n) = \bar{X}_{n-1} = \frac{1}{n-1}(X_1 + \cdots + X_{n-1}).$$

Wtedy $E(T) = E(W) = \mu$, więc oba są nieobciążonymi estymatorami μ. Policzmy wariancję estymatora T

$$D^2(T) = D^2\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}(D^2(X_1) + \cdots + D^2(X_n))$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Analogicznie

$$D^2(W) = \frac{\sigma^2}{n-1}.$$

Mamy więc dla każdego θ

$$D^2(T) = \frac{\sigma^2}{n} < \frac{\sigma^2}{n-1} = D^2(W),$$

więc w szczególności $MSE_\theta(T) < MSE_\theta(W)$ i estymator T jest lepszy od estymatora W (więc estymator W jest niedopuszczalny). Wnioskujemy też, że estymator W nie jest $ENMW(\mu)$.

Definicja 3.30. Dla próby prostej $\underline{X} = (X_1, \ldots, X_n)$ pochodzącej z rozkładu P_θ definiujemy n-tą *informację Fishera*

$$I_n(\theta) = E_\theta\left(\left[\frac{\partial}{\partial\theta} \log P_\theta(\underline{X})\right]^2\right),$$

gdzie $P_\theta(\underline{X}) = P_\theta(X_1) \cdots P_\theta(X_n)$.

[7] ang. *minimum-variance unbiased estimator (MVUE).*

Twierdzenie 3.31. *(Nierówność Rao–Craméra). Przy pewnych założeniach, jeśli $\hat{\theta}_n$ jest estymatorem nieobciążonym parametru θ, to*

$$D^2_\theta(\hat{\theta}_n) \geqslant \frac{1}{nI_1(\theta)}.$$

Przykład 3.32. Niech próba prosta $\underline{X} = (X_1, \ldots, X_n)$ pochodzi z rozkładu $N(\mu, \sigma)$ dla znanego σ. Wtedy mamy

$$P_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

oraz

$$\frac{\partial}{\partial \mu} \log P_\mu(X_1) = \frac{\partial}{\partial \mu} \left(-\log(\sqrt{2\pi}\sigma) - \frac{(X_1-\mu)^2}{2\sigma^2} \right) = \frac{X_1 - \mu}{\sigma^2}.$$

Skoro $X_1 \sim N(\mu, \sigma)$, to pierwsza informacja Fishera wynosi

$$I_1(\mu) = E_\mu \left(\left[\frac{X_1 - \mu}{\sigma^2} \right]^2 \right) = \frac{1}{\sigma^4} D^2_\mu(X_1) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Z nierówności Rao–Craméra wynika więc, że wariancja dowolnego estymatora nieobciążonego μ jest niemniejsza niż $1/(nI_1(\mu)) = \sigma^2/n$. Z drugiej strony $D^2_\mu(\bar{X}_n) = \sigma^2/n$, więc średnia z próby jest $ENMW(\mu)$.

Rozdział 4

Przedziały ufności

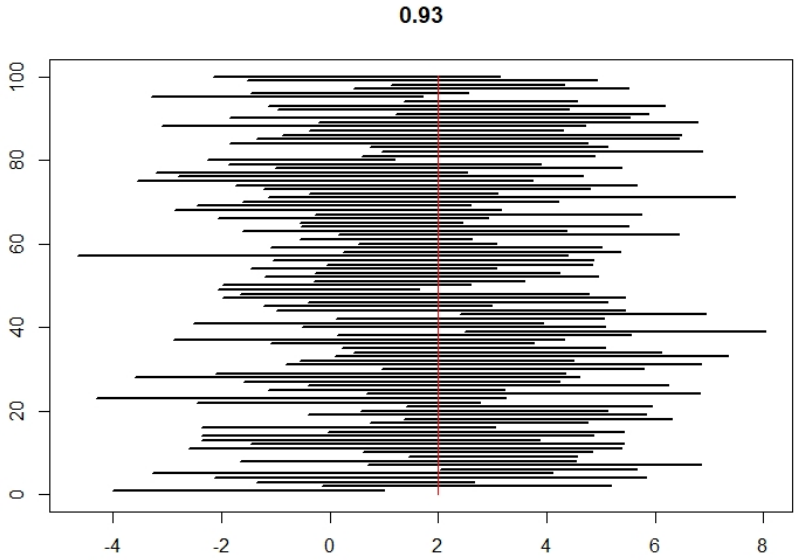
W tym rozdziale przedstawimy kolejną metodę wnioskowania statystycznego: estymację przedziałową. Metoda ta została wprowadzona przez Jerzego Neymana w roku 1937 ([20]).

Definicja 4.1. Niech $X_1, \ldots, X_n$ będzie próbą z rozkładu z parametrem $\theta \in \Theta \subset \mathbb{R}$. *Przedziałem ufności* [1] parametru θ, na poziomie ufności [2] $1-\alpha$, nazywamy przedział
$(\hat{\theta_1}(X_1, \ldots, X_n), \hat{\theta_2}(X_1, \ldots, X_n))$ z własnością:

$$P_\theta(\hat{\theta_1}(X_1, \ldots, X_n) < \theta < \hat{\theta_2}(X_1, \ldots, X_n)) \geqslant 1-\alpha$$

dla wszystkich θ.

Metoda ta dostarcza nam więcej wniosków niż estymacja punktowa. Długość tego przedziału daje nam dodatkową informację o dokładności estymacji (błędzie statystycznym). Parametr α jest zaś równy prawdopodobieństwu (ryzyku) błędu, tj. sytuacji kiedy nasz przedział może nie zawierać prawdziwego parametru. Wizualizuje to następująca symulacja: zakładamy, że cecha ma rozkład $N(2,4)$ i 100 razy losujemy próbę prostą 10-cio elementową. Dla każdej próby wyznaczamy 95% przedział ufności według wzoru (4.2). Na rysunku mamy zaznaczone te 100 przedziałów ufności. Widzimy, że 93 z nich były dobre, tj. zawierały średnią cechy 2. Najlepiej więc by było, gdyby $1-\alpha = 0$. Wtedy niestety długość przedziału obejmuje całe Θ, czyli mamy wniosek pewny, ale zupełnie niedokładny. Gdy natomiast $1-\alpha$ maleje, to dokładność rośnie, ale ryzyko też. Najczęściej więc bierze się $1-\alpha = 0.95$, co jest kompromisem między dokładnością, a ryzykiem błędu.



[1] ang. *confidence interval (CI)*
[2] ang. *confidence level*

4.1 Rozkład t-Studenta

Omówimy teraz rozkład t-Studenta, który będzie potrzebny w dalszej części wykładu.

Definicja 4.2. Rozkład dany gęstością

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}.$$

nazywamy *rozkładem t-Studenta z* $k \in \mathbb{N}_+$ stopniami swobody (ozn. $X \sim t(k)$).

Jak widać, gęstość jest funkcją parzystą, a dla dużych k jest bliska funkcji $Ce^{-x^2/2}$. W konsekwencji dla dużych k rozkład $t(k)$ można aproksymować standardowym rozkładem normalnym. Przez $t(p,k)$ będziemy oznaczać kwantyl rzędu p rozkładu $t(k)$ (tj. $t(p,k) = F^{-1}_{t(k)}(p)$). Z parzystości gęstości dostajemy, że $t(1-p,k) = -t(p,k)$ dla każdego p i k. Dla dużych k kwantyle tego rozkładu możemy przybliżać kwantylami standardowego rozkładu normalnego $t(p,k) \approx u(p)$. Inne podstawowe własności tego rozkładu są zawarte w poniższych twierdzeniach.

Twierdzenie 4.3. *Jeśli* $X \sim t(k)$*, to*

1. $EX = 0$ *dla* $k > 1$ *i* EX *nie istnieje dla* $k = 1$.
2. $D^2(X) = \frac{k}{k-2}$ *dla* $k > 2$, $D^2(X) = \infty$ *dla* $k = 2$ *oraz wariancja nie jest zdefiniowana dla* $k = 1$.

Twierdzenie 4.4. *Niech* $X_1, \ldots, X_n$ *będzie ciągiem niezależnych zmiennych losowych o rozkładzie* $N(\mu,\sigma)$*. Wtedy zmienna losowa*

$$t = \frac{\bar{X}_n - \mu}{s_n}\sqrt{n},$$

gdzie $\bar{X}_n = (\sum X_i)/n$ *i* $s_n = \sqrt{\frac{1}{n-1}\sum (X_i - \bar{X})^2}$*, ma rozkład t-Studenta o* $n-1$ *stopniach swobody.*

4.2 Przedziały ufności dla średniej

Przedział ufności dla μ w rozkładzie $N(\mu,\sigma)$ przy znanym σ

Wiadomo, że jeśli $X_1, \ldots, X_n$ jest ciągiem niezależnych zmiennych losowych o rozkładzie $N(\mu,\sigma)$, to zmienna losowa

$$u = \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n},$$

ma standardowy rozkład normalny $N(0,1)$. Ustalmy $1-\alpha$. Wtedy istnieją u_1 i u_2 takie, że

$$P(u_1 < u < u_2) = 1 - \alpha.$$

Możemy wziąć $u_1 = u(\alpha/2)$ oraz $u_2 = u(1-\alpha/2)$. Wtedy

$$P(u(\alpha/2) < \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} < u(1-\alpha/2) = 1-\alpha,$$

a po krótkich przekształceniach, wykorzystując własność $u(1-p) = -u(p)$, ostatecznie otrzymujemy

$$P\left(\bar{X}_n - u(1-\alpha/2)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + u(1-\alpha/2)\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha. \tag{4.1}$$

Gdy w powyższym rozumowaniu weźmiemy inne u_1, u_2, to otrzymamy inne przedziały ufności. Przedział dany wzorem (4.1) nazywamy *dwustronnym*. O tym przedziale możemy powiedzieć, że ma następujące własności.

1. Gdy ustalimy α i n, to wraz ze wzrostem σ długość przedziału rośnie (czyli, gdy modelujemy cechę o dużej zmienności, to wnioski są mniej dokładne).
2. Gdy ustalimy α i σ, to wraz ze wzrostem n długość przedziału maleje (czyli im większa próba, tym wnioski są dokładniejsze).
3. Gdy ustalimy n i σ to gdy $1-\alpha$ dąży do 1 to długość przedziału dąży do nieskończoności (czyli im pewniejszy wniosek tym mniej jest dokładny).

Przedział ufności dla μ w rozkładzie $N(\mu, \sigma)$

W sytuacji, gdy nie znamy σ, postępujemy podobnie, z tą różnicą, że skorzystamy z twierdzenia 4.4. Ustalmy $1-\alpha$ oraz rozważmy zmienną losową t z tego twierdzenia. Wtedy

$$t = \frac{\bar{X}_n - \mu}{s_n}\sqrt{n} \sim t(n-1).$$

Wtedy istnieją t_1, t_2 takie, że

$$P(t_1 < t < t_2) = 1 - \alpha.$$

Biorąc $t_1 = t(\alpha/2, n-1)$ i $t_2 = t(1-\alpha/2, n-1)$, po przekształceniach otrzymujemy

$$P\left(\bar{X}_n - t(1-\alpha/2, n-1)\frac{s}{\sqrt{n}} < \mu < \bar{X}_n + t(1-\alpha/2, n-1)\frac{s}{\sqrt{n}}\right) = 1-\alpha. \tag{4.2}$$

Przykład 4.5. Mamy dużą partię worków cementu, które powinny ważyć średnio po 20 kg. Wylosowaliśmy kilka, zważyliśmy je i otrzymaliśmy pomiary $19.5, 19.6, 20.2, 20.1, 19.9$. Wyznaczymy 95% przedział ufności dla średniej wagi worka. Możemy tu założyć, że waga worka pochodzi z rozkładu normalnego $N(\mu, \sigma)$. Przedział ufności wyznaczamy w R:

```
> przedzial.ufnosci=function(X,alfa=0.05){
+
+   n=length(X)
+   m=mean(X)
+   l=qt(1-alfa/2,n-1)*sd(X)/sqrt(n)
+   return(c(m-l,m+l))
+ }
>
> X=c(19.5,19.6,20.2,20.1,19.9)
> przedzial.ufnosci(X)
[1] 19.48134 20.23866
```

Jak widać powyższy przedział zawiera 20, więc możemy wnioskować, że na poziomie ufności 0.95 średnia waga worka wynosi(w granicach błędu statystycznego) 20 kg.

Przedział ufności dla średniej dowolnego rozkładu dla dużej liczności próby

Załóżmy, że $X_1, \ldots, X_n$ próbą prostą z populacji o średniej μ i wariancji $0 < \sigma^2 < \infty$. Wtedy z Centralnego Twierdzenia Granicznego wynika, że

$$u = \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \approx N(0,1)$$

dla dużych n. Estymujemy σ przez s_n i otrzymujemy, że

$$u = \frac{\bar{X}_n - \mu}{s_n}\sqrt{n} \approx N(0,1)$$

dla dużych n. Po podobnych przekształceniach jak wcześniej otrzymujemy przedział ufności dla μ na poziomie ufności $1-\alpha$

$$\mu \in \left(\bar{X}_n - u(1-\alpha/2)\frac{s_n}{\sqrt{n}}, \bar{X}_n + u(1-\alpha/2)\frac{s_n}{\sqrt{n}}\right). \tag{4.3}$$

Przykład 4.6. Przepytano 600 losowo wybranych kierowców w Krakowie, ile wydają miesięcznie złotych na paliwo. Z odpowiedzi uzyskano średnią 438 zł z odchyleniem standardowym 187 zł. Ile średnio miesięcznie wydaje kierowca z Krakowa na paliwo? (Przyjmijmy $1-\alpha = 0.9$).

Nie znamy rozkładu wydatków na paliwo, ale próba jest duża. Zastosujemy więc wzór (4.3).

```
> 438-qnorm(0.95)*187/sqrt(600)
[1] 425.4428
> 438+qnorm(0.95)*187/sqrt(600)
[1] 450.5572
```

Otrzymaliśmy przedział $(425.44, 450.55)$.

Przedział ufności dla frakcji dla dużej liczności próby

Ostatni model możemy zastosować dla rozkładu zero-jedynkowego $(1, 0, p)$. Niech $\hat{p} = k/n$ oznacza frakcję z próby. Wtedy wzór (4.3) przyjmie postać

$$p \in \left(\hat{p} - u(1-\alpha/2)\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + u(1-\alpha/2)\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right). \tag{4.4}$$

Przedział ufności dla frakcji

Można pokazać, że przedział ufności dla frakcji (dla dowolnej liczności próby n) jest dany wzorem

$$p \in (B(\alpha/2, k, n-k+1), B(1-\alpha/2, k+1, n-k)), \tag{4.5}$$

gdzie wykorzystujemy kwantyle rozkładu Beta (patrz definicja 4.13).

Przykład 4.7. W sondażu na 978 losowo wybranych respondentów 430 powiedziało, że popiera kandydata AA. Jaki jest 95% przedział ufności dla prawdziwego poparcia tego kandydata?

W tym przypadku mamy $n = 978$, $k = 430$ oraz $1 - \alpha = 0.95$. Możemy zastosować oba wzory (4.4) oraz (4.5):

```
> przedzial.ufnosci=function(k,n,alfa=0.05){
+   p=k/n
+   l=qnorm(1-alfa/2)*sqrt(p*(1-p))/sqrt(n)
+   p1=qbeta(alfa/2,k,n-k+1)
+   p2=qbeta(1-alfa/2,k+1,n-k)
+
+   print(paste("Przedzial przyblizony: (",p-l, ",",p+l,")"))
+   print(paste("Przedzial dokładny: (",p1,",",p2,")"))
+ }
> przedzial.ufnosci(430,978)
[1] "Przedzial przyblizony: ( 0.408565358502969 , 0.470780244769014 )"
[1] "Przedzial dokładny: ( 0.408275907334687 , 0.47143394041474 )"
```

Jak widać na poziomie ufności 0.95 poparcie mieści się w przedziale $(40.8\%, 47.1\%)$, choć wynikiem ankiety będzie $430/978 \times 100\% = 43.96\%$.

Wybór liczności próby do estymacji p.

Zauważmy, że przedział ufności (4.4) jest postaci $(\hat{p} - B, \hat{p} + B)$. Nazwijmy B *błędem estymacji*. Możemy teraz postawić następujący problem: dla ustalonych $1 - \alpha$ oraz B wyznaczyć (przed pobraniem próby) najmniejsze takie n, aby $p \in (\hat{p} - B, \hat{p} + B)$.

Wiemy, że

$$B = u(1 - \frac{\alpha}{2})\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}.$$

Po krótkich przekształceniach uzyskujemy

$$n = \frac{(u(1-\frac{\alpha}{2}))^2 \hat{p}(1-\hat{p})}{B^2}.$$

Niestety prawa strona, a konkretnie $\hat{p}$ zależy od już wykonanej ankiety (więc też od n). Na szczęście wiemy, że $\hat{p} \in [0, 1]$, więc

$$\hat{p}(1-\hat{p}) \leqslant \frac{1}{4}.$$

Ostatecznie uzyskujemy

$$n \geqslant \frac{(u(1-\frac{\alpha}{2}))^2}{4B^2}.$$

Przykład 4.8. Dla $1 - \alpha = 0.95$ i $B = 0.03$ uzyskujemy

```
> qnorm(0.975)^2/(4*0.03^2)
[1] 1067.072
```

Czyli $n = 1068$.

4.3 Rozkład χ^2 i Beta

Definicja 4.9. Rozkład χ^2 (,*chi-kwadrat*') z k stopniami swobody (ozn. $X \sim \chi^2(k)$)jest rozkładem o gęstości danej wzorem

$$f(x) = \begin{cases} \frac{x^{\frac{k}{2}-1}e^{-x/2}}{2^{k/2}\Gamma(\frac{k}{2})}, & \text{jeśli } x \geqslant 0 \\ 0, & \text{jeśli } x < 0 \end{cases}$$

Przez $\chi^2(p,k)$ będziemy oznaczać kwantyl rzędu p rozkładu $\chi^2(k)$. Najważniejsze własności tego rozkładu są zawarte w poniższych twierdzeniach.

Twierdzenie 4.10. *Niech $X_1, \ldots, X_n$ będzie ciągiem niezależnych zmiennych losowych o standardowym rozkładzie normalnym oraz niech $Z = \sum_{i=1}^{k} X_i^2$. Wtedy $Z \sim \chi^2(k)$. W szczególności, $X_1^2 \sim \chi^2(1)$.*

Twierdzenie 4.11. *Niech zmienne losowe $X \sim \chi^2(k)$ i $Y \sim \chi^2(l)$ są niezależne. Wtedy*

1. *$EX = k$ oraz $D^2(X) = 2k$.*
2. *$X + Y \sim \chi^2(k+l)$.*
3. *$\chi^2(k) \approx N(k, \sqrt{2k})$ dla dużych k.*
4. *$\chi^2(2) = Exp(1/2)$.*
5. *Jeśli $Z \sim \mathcal{U}[0,1]$, to $-2\log(Z) \sim \chi^2(2)$.*

Twierdzenie 4.12. *Niech $Z \sim N(0,1)$ oraz $V \sim \chi^2(k)$ są niezależne. Wtedy*

$$T = \frac{Z}{\sqrt{\frac{V}{k}}} \sim t(k).$$

Zdefiniujemy teraz rozkład Beta.

Definicja 4.13. Rozkład Beta o parametrach α, β (ozn. $\mathcal{B}(\alpha,\beta)$) jest dany gęstością

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\chi_{[0,1]}(x).$$

Przez $B(p,\alpha,\beta)$ oznaczać będziemy kwantyl rzędu p rozkładu $\mathcal{B}(\alpha,\beta)$.

4.4 Przedział ufności dla wariancji

Przedział ufności dla wariancji w rozkładzie normalnym.

Twierdzenie 4.14. *Niech $X_1, \ldots, X_n$ będzie ciągiem niezależnych zmiennych losowych o rozkładzie normalnym ze średnią μ i wariancją σ^2. Wtedy*

$$\chi^2 = \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1).$$

Wykorzystując powyższe twierdzenie, analogicznie do poprzednich przedziałów, możemy uzyskać $1-\alpha$ (dwustronny) przedział ufności dla σ^2

$$\sigma^2 \in \left(\frac{(n-1)s_n^2}{\chi^2(1-\alpha/2, n-1)}, \frac{(n-1)s_n^2}{\chi^2(\alpha/2, n-1)} \right).$$

Przykład 4.15. Producent urządzenia do napełniania butelek jednolitrowych zapewnia, że odchylenie standardowe napełnień wynosi 1ml. W losowej próbie 10 napełnień, odchylenie standardowe z próby wyniosło 1.4ml. Czy zakładając normalność napełnień, możemy na poziomie ufności $1-\alpha = 0.95$ powiedzieć, że producent ma rację?

Wyznaczamy przedział ufności dla wariancji

```
> (10-1)*1.4^2/qchisq(1-0.05/2,10-1)
[1] 0.9273099
> (10-1)*1.4^2/qchisq(0.05/2,10-1)
[1] 6.532391
```

oraz dla odchylenia standardowego

```
> sqrt(0.9273099)
[1] 0.9629693
> sqrt(6.532391)
[1] 2.555854
```

Ostatecznie otrzymujemy

$$\sigma \in (0.962, 2.555),$$

więc nie możemy zaprzeczyć stwierdzeniu producenta.

Rozdział 5

Testowanie hipotez

Testowanie hipotez to trzecia podstawowa technika wnioskowania statystycznego.

5.1 Ogólna teoria

Załóżmy, że badamy cechę $X : \Omega \to \mathbb{R}$. Niech $X_1, \ldots, X_n$ będzie próbą. *Testowanie hipotez* to procedura, którą składa się z następujących kroków.

1. Sformułowanie hipotez H_0 and H_1.

Na początku stawiamy dwie hipotezy o badanej zmiennej: *hipotezę zerową* H_0[1] oraz *hipotezę alternatywną* H_1[2]. Celem testu jest ustalenie, czy informacja zawarta w próbie wystarczająco świadczy przeciwko hipotezie zerowej na rzecz hipotezy alternatywnej. Gdy próba wystarczająco nie przeczy hipotezie zerowej, to mówimy, że nie ma podstaw do odrzucenia hipotezy zerowej. Testowanie hipotez polega więc na podziale przestrzeni prób na dwie części: próby, dla których odrzucimy hipotezę zerową i mówimy, że hipoteza alternatywna jest prawdziwa, oraz próby, które nie przeczą hipotezie zerowej.

2. Statystyka testowa. Ten podział będziemy realizować za pomocą mierzalnej funkcji

$$T : (X_1, \ldots, X_n) \to \mathbb{R},$$

którą nazywamy *statystyką testową*[3]. Następnie musimy wyznaczyć rozkład T przy założeniu prawdziwości hipotezy zerowej H_0 (będziemy używać oznaczenia

$$T \overset{H_0}{\sim} R,$$

gdy statystyka testowa ma rozkład R przy prawdziwości H_0).

3. Ustalenie α i wyznaczenie zbioru krytycznego.

Ustalamy $\alpha \in (0, 1)$ (zazwyczaj jest równy 0.05) nazywany *poziomem istotności*[4]. Następnie wyznaczamy *zbiór krytyczny* (*odrzuceń*, ang. *rejection region*) K o własności

$$P(T \in K | H_0) = \alpha \tag{5.1}$$

próbując zminimalizować

$$P(T \notin K | H_1) = \beta. \tag{5.2}$$

Parametr α jest prawdopodobieństwem *błędu I rodzaju*, β jest prawdopodobieństwem *błędu II rodzaju*, zaś $1 - \beta$ jest nazywane *mocą* (ang. *power*) testu.

4. Decyzja. Nasz wniosek opiera się na wartości statystyki testowej dla naszej próby:

– jeśli $T(X_1, \ldots, X_n) \in K$, to odrzucamy hipotezę zerową H_0 na rzecz H_1;

– jeśli $T(X_1, \ldots, X_n) \notin K$, to nie mamy podstaw do odrzucenia H_0 (próba nie przeczy hipotezie H_0);

Przed przykładem sformułujemy jeszcze kilka uwag.

– Hipotezy H_0 i H_1 stawiamy **przed** analizą próby.

[1] ang. *null hypothesis*
[2] ang. *alternative hypothesis*
[3] ang. *test statistic*
[4] ang. *significance level*

– Często hipoteza alternatywna ma postać, która nie pozwala wyznaczyć β. Skoro więc kontrolujemy prawdopodobieństwo błędu I rodzaju (i jest małe), a nie wiemy ile dokładnie wynosi β, to odrzucenie hipotezy zerowej jest wynikiem bardziej znaczącym niż jej nieodrzucenie.

– Jeśli hipoteza zerowa jest prawdziwa to i tak prawdopodobieństwo, że ją odrzucimy wynosi α. Załóżmy, że pewną prawdziwą hipotezę zerową bada niezależnie od innych 100 zespołów na poziomie istotności $\alpha = 0.05$. Wtedy około 5 zespołów odrzuci tę hipotezę w teście i opublikują ten wynik, a ok. 95 zespołów nie jej nie odrzuci i raczej nie opublikuje tych wyników. Wtedy powstanie nieprawdziwy obraz w literaturze. Nosi to nazwę *obciążenia publikacyjnego*[5].

Rozważmy konkretny test.

Test na μ w rozkładzie $N(\mu,\sigma)$ przy znanym σ.

Załóżmy, że $X_1,\dots,X_n$ jest próbą prostą pochodzącą z rozkładu normalnego $N(\mu,\sigma)$ ze znanym σ. Stawiamy hipotezę zerową o parametrze μ, mając trzy możliwości postawienia hipotezy alternatywnej

$$H_0:\ \mu=\mu_0$$

a) $H_1:\mu\neq\mu_0$ *alternatywa dwustronna*[6];

b) $H_1:\mu>\mu_0$ *alternatywa prawostronna*[7];

c) $H_1:\mu<\mu_0$ *alternatywa lewostronna*[8];

Przed testowaniem musimy się na jedną z tych alternatyw zdecydować. Statystyka testowa z zdefiniowana poniższym wzorem ma przy prawdziwości hipotezy zerowej standardowy rozkład normalny

$$z=\frac{\bar X_n-\mu_0}{\sigma}\sqrt{n}\ \overset{H_0}{\sim}\ N(0,1).$$

Teraz musimy wyznaczyć zbiory krytyczne. Oczywiście istnieje niekończenie wiele zbiorów spełniających definicję (5.1), ale, próbując minimalizować prawdopodobieństwo błędu II rodzaju (5.2), otrzymujemy dla poszczególnych hipotez alternatywnych zbiory krytyczne postaci

a) $K=(-\infty,-u(1-\frac{\alpha}{2}))\cup(u(1-\frac{\alpha}{2}),\infty)$;

b) $K=(u(1-\alpha),\infty)$;

c) $K=(-\infty,-u(1-\alpha))$.

Powyższe zbiory można uzasadnić na dwa sposoby. Dla ustalenia uwagi rozważmy alternatywę prawostronną. Skoro średnia z próby $\bar X_n$ jest estymatorem parametru μ to za przyjęciem alternatywy $\mu>\mu_0$ wnioskujemy, gdy $\bar X_n$ jest dużo większa niż μ_0, a to jest równoważne z przyjmowaniem przez z dużych wartości. Więc zbiór krytyczny powinien być postaci $(c,+\infty)$.

Możemy też rozważyć hipotezę alternatywną postaci $H_1:\mu=\mu_1$, gdzie $\mu_1>\mu_0$. Przy takiej alternatywie jesteśmy już w stanie wyznaczyć rozkład statystyki testowej z, i co za tym idzie, policzyć β. Ustalmy więc α i zbiór krytyczny $K_1=(u(1-\alpha),\infty)$. Wtedy prawdopodobieństwo błędu II rodzaju wynosi

$$\begin{aligned}
\beta_1 &= P(z\notin K_1\mid\mu=\mu_1)=P(z<u(1-\alpha)\mid\mu=\mu_1)\\
&=P\left(\frac{\bar X_n-\mu_0}{\sigma}\sqrt{n}<u(1-\alpha)\mid\mu=\mu_1\right)\\
&=P\left(\frac{\bar X_n-\mu_1+\mu_1-\mu_0}{\sigma}\sqrt{n}<u(1-\alpha)\mid\mu=\mu_1\right)\\
&=P\left(\frac{\bar X_n-\mu_1}{\sigma}\sqrt{n}+\frac{\mu_1-\mu_0}{\sigma}\sqrt{n}<u(1-\alpha)\mid\mu=\mu_1\right)\\
&=P\left(\frac{\bar X_n-\mu_1}{\sigma}\sqrt{n}<u(1-\alpha)-\frac{\mu_1-\mu_0}{\sigma}\sqrt{n}\mid\mu=\mu_1\right)\\
&=\Phi\left(u(1-\alpha)-\frac{\mu_1-\mu_0}{\sigma}\sqrt{n}\right).
\end{aligned}$$

[5] ang. *publication bias*
[6] ang. *two-sided alternative*
[7] ang. *right-sided alternative*
[8] ang. *leftt-sided alternative*

Dla zbioru krytycznego postaci $K_2 = (-\infty, -u(1-\alpha))$ prawdopodobieństwo błędu II rodzaju wynosi

$$\begin{aligned}\beta_2 &= P(z \notin K_2 \mid \mu = \mu_1) = P(z > -u(1-\alpha) \mid \mu = \mu_1)\\ &= 1 - P\left(\frac{\bar{X}_n - \mu_1}{\sigma}\sqrt{n} < -u(1-\alpha) - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n} \mid \mu = \mu_1\right)\\ &= 1 - \Phi\left(-u(1-\alpha) - \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right)\\ &= \Phi\left(u(1-\alpha) + \frac{\mu_1 - \mu_0}{\sigma}\sqrt{n}\right) > \beta_1.\end{aligned}$$

Możemy teraz rozumować, że skoro zbiór krytyczny K_1 jest lepszy od zbioru K_2 dla każdej alternatywy postaci $\mu = \mu_1, \mu_1 > \mu_0$, to jest też lepszy dla alternatywy $\mu > \mu_0$, która jest sumą wszystkich alternatyw postaci $\mu = \mu_1, \mu_1 > \mu_0$[9].

Zauważmy, że dla ustalonej postaci zbioru krytycznego, jeśli odrzucamy hipotezę zerową na poziomie istotności α, to odrzucimy hipotezę zerową na każdym poziomie istotności $\alpha' > \alpha$. To uzasadnia poniższą definicję.

Definicja 5.1. Liczbę

$$pv = \inf\{\alpha \mid \text{ odrzucamy hipotezę } H_0 \text{ na poziomie istotności } \alpha\}$$

nazywamy *pvalue*.

Wprost z tej definicji wynika, że jeśli dla danego testu otrzymamy pvalue pv, to jeśli testujemy na poziomie istotności $\alpha < pv$, to hipotezy zerowej nie odrzucimy, a jeśli na poziomie istotności $\alpha > pv$, to hipotezę zerową odrzucimy.

Można też udowodnić, że jeśli statystyka testowa ma rozkład ciągły, to pvalue ma rozkład jednostajny $\mathcal{U}[0,1]$.

Dla przykładu wyznaczmy pvalue w powyższym teście dla alternatywy prawostronnej. Wtedy pvalue to poziom istotności dla którego wartość statystyki testowej znajdzie się na brzegu zbioru krytycznego. Mamy więc

$$u(1-pv) = z(X_1, \ldots, X_n).$$

Obkładając dystrybuantą rozkładu $N(0,1)$, uzyskujemy

$$pv = 1 - \Phi(z(X_1, \ldots, X_n)).$$

Zauważmy, że to pvalue także możemy zapisać jako

$$pv = P(z > z(X_1, \ldots, X_n) \mid H_0),$$

stąd interpretacja, że pvalue to prawdopodobieństwo pobrania próby bardziej przeczącej hipotezie H_0 niż analizowana próba.

W `R` do implementacji tego testu można użyć funkcji `z.test{BSDA}`.

[9] Oczywiście nie pokazaliśmy, że zbiór krytyczny $(c, +\infty)$ ma najmniejsze β pośród wszystkich zbiorów krytycznych, tylko, że jest lepszy od zbioru postaci $(-\infty, c)$.

Rozdział 6

Przegląd podstawowych testów parametrycznych

W tym rozdziale zaprezentujemy podstawowe testy.

6.1 Podstawowe testy parametryczne

6.1.1 Testy o jednym parametrze

6.1.1.1 Test na μ w rozkładzie $N(\mu, \sigma)$ (test t o średniej w rozkładzie normalnym).

Załóżmy, że $X_1, \ldots, X_n$ jest próbą prostą pochodzącą z rozkładu normalnego $N(\mu, \sigma)$. Tak jak w poprzednim teście są możliwe trzy alternatywy o tych samych nazwach. Ponieważ test ten jest bardzo podobny do poprzedniego wypiszemy go ‚skrótowo':

$$H_0: \ \mu = \mu_0 \qquad H_1 : a)\ \mu \neq \mu_0 \quad b)\ \mu > \mu_0 \quad c)\ \mu < \mu_0.$$

$$t = \frac{\bar{X}_n - \mu_0}{s_n}\sqrt{n} \overset{H_0}{\sim} t(n-1).$$

a) $K = (-\infty, -t(1-\frac{\alpha}{2}, n-1)) \cup (t(1-\frac{\alpha}{2}, n-1), \infty)$;

b) $K = (t(1-\alpha, n-1), \infty)$;

c) $K = (-\infty, -t(1-\alpha, n-1))$.

W `R` do implementacji tego testu używamy funkcji `t.test`.

Przykład 6.1. Mamy dużą partię worków cementu, które powinny według producenta ważyć średnio po 20kg. Podejrzewamy, że producent zawyża średnią wagę worka. Wylosowaliśmy kilka, zważyliśmy je i otrzymaliśmy pomiary $19.5, 19.6, 20.2, 20.1, 19.9, 19.6, 19.4$. Możemy tu założyć, że waga worka pochodzi z rozkładu normalnego $N(\mu, \sigma)$. Stawiamy hipotezę zerową $H_0 : \mu = 20$ i alternatywną $H_1 : \mu < 20$ (ponieważ chcemy pokazać, że producent zawyża średnią wagę worka, to jest prawdziwa średnia waga worka jest mniejsza niż 20). Testowanie w `R` wygląda tak

```
> W=c(19.5,19.6,20.2,20.1,19.9,19.6,19.4)
> t.test(W,mu=20,alternative = "less")

One Sample t-test

data:  W
t = -2.0717, df = 6, p-value = 0.04184
alternative hypothesis: true mean is less than 20
95 percent confidence interval:
     -Inf 19.98493
sample estimates:
mean of x
 19.75714
```

Uzyskujemy między innymi informację, że wartość statystyki testowej to $t = -2.0717$, etc. Najistotniejszą informacją jest jednak to, że pvalue jest równe 0.04184. Na poziomie istotności 0.05 odrzucimy wiec hipotezę zerową na rzecz alternatywnej, czyli uznamy, że producent zawyżył wagę worka.

Oczywiście wynik testu zależy od przyjętej alternatywy, na przykład

```
> t.test(W,mu=20,alternative="two.sided")

One Sample t-test

data:  W
t = -2.0717, df = 6, p-value = 0.08369
alternative hypothesis: true mean is not equal to 20
95 percent confidence interval:
 19.47031 20.04398
sample estimates:
mean of x
 19.75714
```

czy

```
> t.test(W,mu=20,alternative="greater")

One Sample t-test

data:  W
t = -2.0717, df = 6, p-value = 0.9582
alternative hypothesis: true mean is greater than 20
95 percent confidence interval:
 19.52936      Inf
sample estimates:
mean of x
 19.75714
```

6.1.1.2 Test na wariancję w rozkładzie normalnym

Załóżmy, że $X_1, \ldots, X_n$ jest próbą prostą pochodzącą z rozkładu normalnego $N(\mu, \sigma)$. Stawiamy hipotezę zerową o wariancji σ^2. Tradycyjnie mamy trzy możliwe hipotezy alternatywne.

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : a)\ \sigma^2 \neq \sigma_0^2 \quad b)\ \sigma^2 > \sigma_0^2 \quad c)\ \sigma^2 < \sigma_0^2.$$

$$\chi^2 = \frac{(n-1)s_n^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n-1).$$

a) $K = (0, \chi^2(\frac{\alpha}{2}, n-1)) \cup (\chi^2(1-\frac{\alpha}{2}, n-1), \infty)$;

b) $K = (\chi^2(1-\alpha, n-1), \infty)$;

c) $K = (0, \chi^2(\alpha, n-1))$.

6.1.1.3 Test na wariancję w dowolnym rozkładzie

Bez założenia normalności, ale dla dużej próby, możemy skonstruować test ze statystyką testową

$$u = \frac{s_n^2 - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{n}{2}} \underset{n\to\infty}{\overset{H_0}{\sim}} N(0,1),$$

która asymptotycznie ma standardowy rozkład normalny.

6.1.1.4 Testy na frakcję

Załóżmy, że $X_1, \ldots, X_n$ jest próbą prostą pochodzącą z rozkładu zero-jedynkowego $(1, 0, p)$. Niech k to liczba jedynek w próbie, a $\hat{p} = k/n$. Stawiamy hipotezę zerową o frakcji p. Tradycyjnie mamy trzy możliwe hipotezy alternatywne.

$$H_0 : \ p = p_0 \qquad H_1 : a)\ p \neq p_0 \quad b)\ p > p_0 \quad c)\ p < p_0.$$

Mamy dwie wersje tego testu.

I. n małe (a właściwie dowolne).

$$T = k \overset{H_0}{\sim} b(n, p_0).$$

Ponieważ rozkład statystyki testowej jest dyskretny, to żeby nie bawić się kwantylami, prościej jest od razu wyznaczyć pvalue. I tak na przykład dla alternatywy b) rozważamy zbiór krytyczny prawostronny, więc $pv = P(T \geqslant k \mid H_0)$, itd. W `R` implementujemy ten test funkcją `binom.test` (patrz przykład 6.2).

II. n duże ($n > 100, np_0 > 50$).

$$u = \frac{k - np_0}{\sqrt{np_0(1-p_0)}} \overset{H_0}{\underset{n\to\infty}{\sim}} N(0,1).$$

Zbiory krytyczne są podobne jak poprzednio, to znaczy

a) $K = (-\infty, -u(1-\frac{\alpha}{2})) \cup (u(1-\frac{\alpha}{2}), \infty)$;

b) $K = (u(1-\alpha), \infty)$;

c) $K = (-\infty, -u(1-\alpha))$.

Przykład 6.2. W pewnej gminie poparcie dla partii ABC w ostatnich wyborach wynosiło 24.5%. Teraz przeprowadzono ankietę i na 134 losowo wybrane osoby 45 popierają tę partię. Czy poparcie się zmieniło?

Stawiamy hipotezę zerową $H_0 : p = 0.245$ i alternatywną $H_1 : p \neq 0.245$.

```
> binom.test(45, 134, p = 0.245)

Exact binomial test

data:  45 and 134
number of successes = 45, number of trials = 134, p-value = 0.02033
alternative hypothesis: true probability of success is not equal to 0.245
95 percent confidence interval:
 0.2565963 0.4224826
sample estimates:
probability of success
             0.3358209
```

Ponieważ nie jest powiedziane inaczej rozważamy standardowy poziom istotności $\alpha = 0.05$. Ponieważ $pv = 0.02033$, to hipotezę zerową odrzucamy i stwierdzamy, że od ostatnich wyborów poparcie tej partii się zmieniło.

6.1.2 Testy t na średnią w dwóch populacjach

6.1.2.1 Testy t dla dwóch prób niezależnych

Załóżmy, że niezależne od siebie próby proste $X_{1,1}, \ldots, X_{1,n_1}$ oraz $X_{2,1}, \ldots, X_{2,n_2}$ pochodzą odpowiednio z rozkładów $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$. Niech $\bar{X}_1$, $\bar{X}_2$, S_1^2 oraz S_2^2 oznaczają odpowiednio średnie i wariancje z próby dla próby pierwszej i drugiej. Stawiamy hipotezę zerową o różnicy średnich oraz jak zawsze trzy możliwe alternatywy

$$H_0 : \ \mu_1 - \mu_2 = \mu_0$$
$$H_1 : a)\ \mu_1 - \mu_2 \neq \mu_0 \quad b)\ \mu_1 - \mu_2 > \mu_0 \quad c)\mu_1 - \mu_2 < \mu_0.$$

Mamy dwa testy: przy założeniu, że wariancje są równe albo, że są dowolne.

I. Gdy $\sigma_1 = \sigma_2$.

Używamy statystyki testowej

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{S_{X_1X_2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t(n_1 + n_2 - 2),$$

gdzie

$$S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

W R mamy `t.test(`X_1`,`X_2`,mu=`μ_0`,var.equal=TRUE,paired=F)`.

II. Dla dowolnych σ_1, σ_2.

Używamy statystyki testowej

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{S_{\bar{X}_1 - \bar{X}_2}} \overset{H_0}{\sim} t(df),$$

gdzie

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

oraz

$$df = \left\lfloor \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \right\rfloor.$$

W R mamy `t.test(`X_1`,`X_2`,mu=`μ_0`,var.equal=FALSE,paired=F)`.

W obu testach odpowiednio dla alternatyw a), b) i c) stosujemy zbiór krytyczny dwustronny, prawostronny i lewostronny.

Przykład 6.3. Badano wzrost 18-latków w Krakowie i Warszawie. Pobrano losowe próby i uzyskano dane 176, 180, 170, 172, 173, 175 dla Krakowa i 176, 178, 179, 180, 170, 172, 174, 175 dla Warszawy. Czy średni wzrost 18-latków w Krakowie i Warszawie się różni?

Rozważamy hipotezę alternatywną dwustronną.

```
> K=c(176,180,170,172,173,175)
> W=c(176,178,179,180,170,172,174,175)
> t.test(K,W,var.equal = FALSE,paired = FALSE)

Welch Two Sample t-test

data:  K and W
t = -0.61969, df = 10.855, p-value = 0.5483
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.317168  2.983835
sample estimates:
mean of x mean of y
 174.3333  175.5000
```

Jak widać nie ma podstaw do odrzucenia hipotezy o równości średnich w tych dwóch miastach.

6.1.2.2 Test t dla dwóch prób zależnych

Załóżmy, że mamy próbę prostą zmiennej losowej dwuwymiarowej $(X_1, Y_1), \ldots, (X_n, Y_n)$ (to znaczy badamy cechy X i Y na tych samych osobnikach, są *sparowane*). Zakładamy, że zmienna losowa $D = X - Y$ ma rozkład normalny. Stawiamy hipotezę zerową $H_0 : E(X) - E(Y) = \mu_0$ (i znowu mamy trzy możliwe hipotezy alternatywne). Test polega na użyciu zwykłego testu t na próbie

$$D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n.$$

W R używamy funkcji `t.test(`X_1`,`X_2`,mu=`μ_0`,var.equal=FALSE,paired=TRUE)`.

Przykład 6.4. Badano skuteczność pewnego leku na obniżenie ciśnienia. Wybrano losowo wybraną grupę pacjentów i dwukrotnie zmierzono im ciśnienie, przed podaniem leku i godzinę po. Otrzymano pomiary: przed 134, 145, 143, 154, 149 i po 129, 147, 140, 150, 147. Czy lek działa?

```
> Przed=c(134,145,143,154,149)
> Po=c(129,147,140,150,147)
> t.test(Po,Przed,alternative="less",paired=TRUE)

Paired t-test

data:  Po and Przed
t = -1.9863, df = 4, p-value = 0.05898
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1759203
sample estimates:
mean of the differences
                  -2.4
```

Na poziomie istotności 0.05 stwierdzamy, że nie mamy podstaw do stwierdzenia, że lek (średnio) obniża ciśnienie.

6.1.3 Rozkład F

W tym podrozdziale zdefiniujemy kolejny ważny w statystyce rozkład, rozkład F (zwany też *Fishera, Snedecora–Fishera, F-Snedecora*).

Definicja 6.5. *Rozkład F z d_1 i d_2 stopniami swobody* (ozn. $X \sim F(d_1, d_2)$)) to rozkład zadany gęstością

$$f(x) = \frac{1}{B(\frac{d_1}{2}, \frac{d_2}{2})} \Big(\frac{d_1}{d_2}\Big)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \Big(1 + \frac{d_1}{d_2} x\Big)^{-\frac{d_1+d_2}{2}} I_{\mathbb{R}_+}(x).$$

Przez $F(\alpha, d_1, d_2)$ będziemy oznaczać kwantyl rzędu α rozkładu $F(d_1, d_2)$.

Najważniejsze własności tego rozkładu zawarte są w poniższych twierdzeniach.

Twierdzenie 6.6. *Jeśli $X \sim F(d_1, d_2)$ to*

1. $EX = \frac{d_2}{d_2-2}$ *dla* $d_2 > 2$;
2. *dla* $d_2 > 4$

$$D^2(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2-2)^2(d_2-4)}.$$

Twierdzenie 6.7. *Niech $X \sim \chi^2(d_1)$ i $Y \sim \chi^2(d_2)$ będą niezależnymi zmiennymi losowymi. Wtedy*

$$F = \frac{X/d_1}{Y/d_2} \sim F(d_1, d_2).$$

Twierdzenie 6.8. *Zachodzą też następujące własności.*

1. *Jeśli $X \sim F(d_1, d_2)$, to $X^{-1} \sim F(d_2, d_1)$;*
2. *Jeśli $Y \sim t(k)$, to $Y^2 \sim F(1, k)$ oraz $Y^{-2} \sim F(k, 1)$.*

6.1.4 Testy na porównywanie wariancji w dwóch populacjach o rozkładach normalnych

Załóżmy, że niezależne od siebie próby proste $X_{1,1}, \ldots, X_{1,n_1}$ oraz $X_{2,1}, \ldots, X_{2,n_2}$ pochodzą odpowiednio z rozkładów $N(\mu_1, \sigma_1)$ oraz $N(\mu_2, \sigma_2)$. Niech S_1^2 oraz S_2^2 oznaczają wariancje z próby odpowiednio dla próby pierwszej i drugiej. Stawiamy hipotezę zerową o ilorazie wariancji oraz mamy trzy możliwe alternatywy

$$H_0 : \frac{\sigma_1^2}{\sigma_2} = 1;$$

$$H_1 : a)\ \frac{\sigma_1^2}{\sigma_2} \neq 1 \quad b)\ \frac{\sigma_1^2}{\sigma_2} > 1 \quad c)\ \frac{\sigma_1^2}{\sigma_2} < 1.$$

Stosujemy statystykę testową

$$F = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F(n_1 - 1, n_2 - 1).$$

Zbiór krytyczny jest dla powyższych hipotez alternatywnych odpowiednio dwustronny, prawostronny i lewostronny. W `R` używamy instrukcji `var.test`.

6.1.5 Testy na porównywanie frakcji w dwóch populacjach dla dużych prób

Załóżmy, że niezależne od siebie próby proste $X_{1,1}, \ldots, X_{1,n_1}$ oraz $X_{2,1}, \ldots, X_{2,n_2}$ pochodzą odpowiednio z rozkładów $(1, 0, p_1)$ oraz $(1, 0, p_2)$. Załóżmy, że liczności prób n_1 i n_2 są duże. Niech $\hat{p_1} = k_1/n_1$ oraz $\hat{p_2} = k_2/n_2$ oznaczają frakcje z próby odpowiednio dla próby pierwszej i drugiej. Przedstawimy dwa testy.

Test równości frakcji

Stawiamy hipotezę zerową o równości parametrów p_1 i p_2 oraz mamy trzy możliwe alternatywy

$$H_0 : \ p_1 = p_2$$
$$H_1 : a)\ p_1 \neq p_2 \quad b)\ p_1 > p_2 \quad c)\ p_1 < p_2.$$

Używamy statystyki testowej

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} \underset{n_1, n_2 \to \infty}{\overset{H_0}{\sim}} N(0,1),$$

gdzie

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2}.$$

Zbiór krytyczny jest dla powyższych hipotez alternatywnych odpowiednio dwustronny, prawostronny i lewostronny.

Przykład 6.9. W Krakowie na 300 losowo wybranych ludzi 60-ciu popiera partię ABC, a w Warszawie 70-ciu na 400. Czy poparcie dla tej partii w Krakowie i Warszawie się różni?

Stosujemy hipotezę alternatywną dwustronną:

```
> row.frak.test=function(k1,n1,k2,n2){
+
+   print("Test dwustronny na rownosc frakcji")
+   print("p.value=")
+   p1=k1/n1
+   p2=k2/n2
+   p=(k1+k2)/(n1+n2)
+
+   z=(p1-p2)/sqrt(p*(1-p)*(1/n1+1/n2))
+
+   return(2*(1-pnorm(abs(z))))
+ }
>
> row.frak.test(60,300,70,400)
[1] "Test dwustronny na rownosc frakcji"
[1] "p.value="
[1] 0.3999416
```

Pvalue 0.3999416 świadczy, że nie odrzucimy hipotezy o równości poparcia w tych miastach.

Test o różnicy frakcji

$$H_0 : \ p_1 - p_2 = D;$$
$$H_1 : \ p_1 - p_2 > D.$$

Używamy statystyki testowej

$$z = \frac{(\hat{p_1} - \hat{p_2}) - D}{\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}} \underset{n_1, n_2 \to \infty}{\overset{H_0}{\sim}} N(0,1),$$

z prawostronnym zbiorem krytycznym.

Przykład 6.10. Wykonano dwie niezależne ankiety: przed i po kampanii reklamowej pewnej marki piwa. Przed kampanią na 1100 losowo wybranych piwoszy 100-stu zadeklarowało chęć zakupu piwa tej marki. Po kampanii 150-ciu na 1200-tu. Czy kampania reklamowa spowodowała wzrost udziału w rynku o co najmniej $D = 0.01$?

```
> row.frak.D.test=function(k1,n1,k2,n2,D){
+
+   print("Test na roznice frakcji")
+   p1=k1/n1
+   p2=k2/n2
+   print(paste("p1=",p1))
+   print(paste("p2=",p2))
+
+   z=(p1-p2-D)/sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
+   print("p.value=")
+   return(1-pnorm(z))
+
+ }
>
> row.frak.D.test(150,1200,100,1100,D=0.01)
[1] "Test na roznice frakcji"
[1] "p1= 0.125"
[1] "p2= 0.0909090909090909"
[1] "p.value="
[1] 0.03086311
```

Na poziomie istotności $\alpha = 0.05$ stwierdzamy, że kampania reklamowa spowodowała wzrost udziału w rynku o co najmniej 0.01.

Rozdział 7

Przegląd podstawowych testów nieparametrycznych

W tym rozdziale przedstawimy podstawowe testy nieparametryczne.

7.1 Testy zgodności i niezależności

W testach zgodności testujemy hipotezę, że badana cecha ma ustalony rozkład, np. $H_0 : X \sim N(0,1)$, $H_0 : X \sim \mathcal{P}(2)$, etc. Ogólnie $H_0 : F_X = F_0$, to znaczy, że dystrybuanta badanej cechy jest równa F_0.

7.1.1 Test zgodności χ^2 (Pearsona)

W tym teście nie mamy żadnych dodatkowych założeń o badanej zmiennej (co jest wielką zaletą), niemniej dla zmiennych o rozkładzie ciągłym test Kołmogorowa (patrz podrozdział 7.1.4) jest lepszy.

Testujemy hipotezę zerową i alternatywną

$$H_0 : F_X = F_0 \quad H_1 : \neg H_0.$$

Próbę prostą $(X_1, \ldots, X_n)$ przekształcamy w szereg rozdzielczy $\{(K_i, O_i)\}_{i=1,\ldots k}$ (gdy badana zmienna ma rozkład dyskretny to klasy w tym szeregu są zazwyczaj definiowane przez poziomy tego rozkładu). O_i nazywamy *częstościami empirycznymi (zaobserwowanymi)*[1] Zdefiniujmy

$$p_i = P(X \in K_i | H_0)$$

nazywane *prawdopodobieństwami teoretycznymi*. Definiujemy *częstości spodziewane*[2] wzorem

$$E_i = np_i.$$

Statystyka testowa χ^2 mierzy różnicę między częstościami spodziewanymi i zaobserwowanymi i ma asymptotycznie rozkład $\chi^2(k-1)$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \underset{n\to\infty}{\overset{H_0}{\sim}} \chi^2(k-1).$$

Zbiór krytyczny jest prawostronny, jest więc równy $K = (\chi^2(1-\alpha, k-1), \infty)$. Zazwyczaj aproksymacji tej używamy, gdy $E_i \geqslant 5$ dla każdego $i = 1, \ldots, k$.

Jak widać wynik testu zależy od próby (a tak naprawdę od liczności klas $O_1, \ldots, O_k$) oraz prawdopodobieństw teoretycznych $p = (p_1, \ldots, p_k)$. W `R` używamy funkcji

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Zwracam uwagę na domyślną wartość argumentu p.

[1] ang. *observed frequencies*
[2] ang. *expected frequencies*

Przykład 7.1. Rzucamy monetą. Otrzymaliśmy 25 orłów i 35 reszek. Czy moneta jest symetryczna?

W tym przypadku mamy dwie klasy o licznościach (częstościach) 25 i 35. Testujemy hipotezę, że $p_1 = p_2 = 1/2$.

```
> chisq.test(c(25,35))

Chi-squared test for given probabilities

data:  c(25, 35)
X-squared = 1.6667, df = 1, p-value = 0.1967
```

Na poziomie istotności 0.05 nie odrzucamy hipotezy o symetryczności tej monety.

7.1.2 Test niezależności χ^2

Niech X, Y będą dowolnymi zmiennymi. Chcemy testować hipotezę o niezależności tych zmiennych. Wtedy oczywiści musimy obserwować wartości tych zmiennych na tych samych osobnikach, więc nasza próba jest postaci $(X_1, Y_1), \ldots, (X_n, Y_n)$. Testujemy hipotezę zerową

$$H_0 : \ X \text{ i } Y \text{ są niezależne}$$

z hipotezą alternatywną, że nie są niezależne.

Zarówno próbę cechy X i Y dzielimy na klasy, a następnie wyznaczamy liczności we wszystkich kombinacjach tych klas. Formalnie

$$O_{ij} = \#\{(X_s, Y_s) \ : \ X_s \in K_i, Y_s \in L_j\}.$$

Macierz $O = [O_{ij}]_{i=1,\ldots,k;j=1,\ldots,l}$ nazywamy *tablicą wielodzielczą* (*tablicą kontyngencji*[3]).

Oznaczmy przez $r_i = \sum_{j=1}^{l} O_{ij}$ i $c_j = \sum_{i=1}^{k} O_{ij}$ odpowiednio sumę i-tego wiersza i j-ej kolumny macierzy O. Wtedy

$$\begin{aligned} p_{ij} &:= P(X \in K_i \wedge Y \in L_j | H_0) = P(X \in K_i) P(Y \in L_j) \\ &\approx \frac{r_i}{n} \frac{c_j}{n} = \frac{r_i c_j}{n^2}. \end{aligned}$$

Wyznaczamy teraz spodziewaną liczność w (i, j)-ej komórce

$$E_{ij} = p_{ij} n = \frac{r_i c_j}{n}.$$

Statystyka testowa jest postaci

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \underset{n\to\infty}{\overset{H_0}{\sim}} \chi^2\big((k-1)(l-1)\big).$$

Zbiór krytyczny jest prawostronny $K = (\chi^2(1-\alpha, (k-1)(l-1)), \infty)$ (zazwyczaj aproksymacji tej używamy, gdy $E_{ij} \geqslant 5$ dla wszystkich i, j).

Jak widać wynik testu zależy tylko od tablicy wielodzielczej O (a sama ta procedura nazywa się też analizą *tablicy wielodzielczej*). W R mamy `chisq.test(O)`.

Przykład 7.2. (Dane fikcyjne). Badamy dwie cechy: miejsce urodzenia (z poziomami 'NW', 'NE', 'SW' i 'SE') oraz grupę krwi. Uzyskano dane

```
> O=matrix(c(20,30,35,25,40,60,50,
+            30,30,70,40,30,50,50,70,50),
+          ncol=4,
+          dimnames=list(c("NW","NE","SW","SE"),
+                        c("O","A","B","AB")))
> O
    O  A  B AB
NW 20 40 30 50
NE 30 60 70 50
SW 35 50 40 70
SE 25 30 30 50
```

[3] ang. *contingency table*

Czy rozkład grupy krwi zależy od miejsca urodzenia?

```
> chisq.test(O)

Pearson's Chi-squared test

data:  O
X-squared = 18.599, df = 9, p-value = 0.02883
```

Na poziomie istotności 0.05 hipotezę zerową o niezależności odrzucamy. Stwierdzamy więc, że grupa krwi zależy od miejsca urodzenia.

7.1.3 Test Fischera

Załóżmy, że badamy niezależność dwóch cech X i Y, a tablica wielodzielcza jest wymiaru 2×2. Wtedy zamiast testu niezależności χ^2, gdzie znamy tylko asymptotyczny rozkład statystyki testowej, możemy użyć *dokładnego testu Fishera*[4]. Jest to szczególnie istotne dla małej liczby obserwacji. Zobaczmy poniższy przykład.

Przykład 7.3. Załóżmy, że tablica wielodzielcza jest następująca

```
> M=matrix(c(4,24,6,7),ncol=2)
> M
     [,1] [,2]
[1,]    4    6
[2,]   24    7
```

Użyjmy zarówno testu Fishera jak i testu χ^2.

```
> fisher.test(M)

Fisher's Exact Test for Count Data

data:  M
p-value = 0.04851
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.03196256 1.13355348
sample estimates:
odds ratio
 0.2039797

> chisq.test(M)

Pearson's Chi-squared test with Yates' continuity correction

data:  M
X-squared = 3.3138, df = 1, p-value = 0.0687

Komunikat ostrzegawczy:
W poleceniu 'chisq.test(M)': Aproksymacja chi-kwadrat może być niepoprawna
```

Zauważmy, że stosując test dokładny, hipotezę zerową odrzucamy, w przeciwieństwie do testu χ^2 (na poziomie istotności 0.05).

[4]ang. *Fisher's Exact Test*

7.1.4 Test zgodności Kołmogorowa

Testujemy hipotezę zerową i alternatywną

$$H_0 : F_X = F_0 \quad H_1 : \neg H_0,$$

zakładamy jednak, że dystrybuanta F_0 jest **ciągła**.

Statystyka testowa jest równa

$$D_n = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F_0(t)|,$$

gdzie $\hat{F}_n$ jest dystrybuantą empiryczną z próby. Kołmogorow pokazał (zob. [15]), że rozkład statystyki D_n nie zależy od F_0 oraz podał jej asymptotyczny rozkład. Precyzyjniej pokazał, że $\sqrt{n}D_n \to D$ (przy $n \to \infty$), gdzie rozkład D dany jest dystrybuantą (dla $t > 0$)

$$F_D(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{2k^2t^2}.$$

W `R` używamy funkcji `ks.test`.

Przykład 7.4. Przykład pokarzemy na danych symulowanych.

```
> X=runif(25,0,1)
> ks.test(X,"punif",min=0,max=1)$p.value
[1] 0.450178
```

Testowaliśmy hipotezę zerową prawdziwą i jak widać test Kołmogorowa zadziałał prawidłowo.

7.2 Test znaków

Test znaków jest prostym testem (o medianach), co więcej wymaga co najwyżej porządkowej skali pomiarowej.

7.2.1 Test znaków dla mediany

Załóżmy, że badamy cechę X oraz mamy próbę prostą $X_1, \ldots, X_n$. Niech md oznacza medianę cechy X. Testujemy hipotezę zerową, mając trzy możliwe alternatywy

$$H_0 : md = \mu_0 \qquad H_1 : a)\ md \neq \mu_0 \quad b)\ md > \mu_0 \quad c)\ md < \mu_0.$$

Niech $n_0 = \#\{X_i \ : \ X_i \neq \mu_0\}$. Statystyka testowa jest równa

$$S = \#\{X_i \ : \ X_i > \mu_0\} \overset{H_0}{\sim} b(n_0, 1/2).$$

Do poszczególnych hipotez alternatywnych używamy odpowiednio zbioru krytycznego dwustronnego, prawostronnego i lewostronnego. W `R` używamy funkcji `SIGN.test{BSDA}` (patrz przykład 7.5).

7.2.2 Test znaków dla dwóch prób zależnych

Załóżmy, że mamy próbę prostą zmiennej losowej dwuwymiarowej $(X_1, Y_1), \ldots, (X_n, Y_n)$ (to znaczy badamy cechy X i Y na tych samych osobnikach). Niech md_X oznacza medianę cechy X, a md_Y cechy Y. Stawiamy hipotezę zerową

$$H_0 : md_X - md_Y = \mu_0$$

(i znowu mamy trzy możliwe hipotezy alternatywne). Zauważmy, że gdy $\mu_0 = 0$ to równoważnie możemy sformułować hipotezę zerową jako

$$H_0 : p := P(X > Y) = 1/2.$$

Test polega na użyciu testu znaków na próbie

$$D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n.$$

W `R` używamy funkcji `SIGN.test(`X_1`,`X_2`,md=`μ_0`,alternative=...)`.

Przykład 7.5. Producent batonów bada czy projekt nowego opakowania podoba się bardziej niż aktualne. Wylosowano kilku klientów i każdy miał ocenić te dwa opakowania w skali od 1 do 5. Uzyskano dane: dla starego opakowania $1, 2, 3, 2, 4, 5, 5, 5, 2, 3, 2, 3, 3, 3$ oraz dla nowego $2, 4, 3, 2, 2, 4, 3, 4, 3, 4, 1, 1, 4, 4$.

```
> Stare=c(1,2,3,2,4,5,5,5,2,3,2,3,3,3)
> Nowe=c(2,4,3,2,2,4,3,4,3,4,1,1,4,4)
```

a) Czy mediana oceny starego opakowania wynosi 2?

```
> SIGN.test(Stare,md=2)

One-sample Sign-Test

data:  Stare
s = 9, p-value = 0.02148
alternative hypothesis: true median is not equal to 2
95 percent confidence interval:
 2.000000 4.165934
sample estimates:
median of x
         3
```

Na poziomie istotności 0.05 stwierdzamy, że mediana oceny starego opakowania nie jest równa 2.

b) Czy projekt nowego opakowania podoba się (średnio) bardziej?

```
> SIGN.test(Stare,Nowe,alternative="less")

Dependent-samples Sign-Test

data:  Stare and Nowe
S = 6, p-value = 0.6128
alternative hypothesis: true median difference is less than 0
95 percent confidence interval:
 -Inf    1
sample estimates:
median of x-y
           0
```

Na poziomie istotności 0.05 nie możemy stwierdzić, że nowy projekt jest lepszy.

7.3 Testy serii i rangowe

W tym podrozdziale przedstawimy podstawowe testy serii oraz testy rangowe. Zanim do nich przejdziemy zdefiniujemy pojęcie serii oraz rangi.

7.3.0.1 Serie

Rozważmy ciąg a składający się z n zer oraz m jedynek (formalnie $a \in \{0,1\}^{n+m}$). *Serią*seria[5] nazywamy maksymalny podciąg a składający się z kolejnych, takich samych wyrazów. Na przykład ciąg $a = (1, 0, 0, 1, 1, 1, 0, 0)$ składa się z 4 serii: (1), $(0, 0)$, $(1, 1, 1)$ oraz $(0, 0)$. Zachodzi następujące twierdzenie.

Twierdzenie 7.6. *Rozważmy zbiór wszystkich ciągów $\Omega = \{0,1\}^{n+m}$ z prawdopodobieństwem klasycznym. Niech R oznacza zmienną losową, która ciągowi z Ω przyporządkowuje liczbę serii w tym ciągu. Wtedy*

[5] ang. *run*

a) rozkład R jest następujący

$$P(R = 2k) = \frac{2\binom{n-1}{k-1}\binom{m-1}{k-1}}{\binom{n+m}{n}}$$

oraz

$$P(R = 2k+1) = \frac{\binom{n-1}{k-1}\binom{m-1}{k} + \binom{n-1}{k}\binom{m-1}{k-1}}{\binom{n+m}{n}}.$$

b)

$$E(R) = \frac{2nm}{n+m} + 1;$$

c)

$$D(R) = \sqrt{\frac{2nm(2nm-n-m)}{(n+m)^2(n+m-1)}}.$$

Dla dużych wartości n i m stosuje się aproksymację $R \approx N(E(R), D(R))$.

7.3.0.2 Rangi

Załóżmy, że ciąg $\underline{X} = (X_1, \ldots, X_n)$ składa się z różnych wartości (przy założeniu, że badamy cechę o rozkładzie ciągłym, to założenie jest teoretycznie spełnione). Wtedy *rangą* i-ej obserwacji w ciągu $\underline{X}$ nazywamy miejsce X_i ciągu, który powstał przez uporządkowanie $\underline{X}$ od najmniejszego do największego i oznaczamy przez $r_{\underline{X}}(X_i)$ lub $r_{X_1,\ldots,X_n}(X_i)$. W szczególności element najmniejszy w ciągu $\underline{X}$ będzie miał rangę 1, a największy rangę n. Suma wszystkich rang wyniesie zaś $1 + \cdots + n = n(n+1)/2$. Zamianę obserwacji na rangi nazywamy *rangowaniem*. Zauważmy, że do rangowania wystarczy nam skala porządkowa, a niekiedy obserwacja może być bezpośrednio rangą (na przykład gdy badaną cechą jest miejsce w zawodach, uzyskana lokata w rankingu, etc.).

Przykładowo wektor rang obserwacji z ciągu $\underline{X} = (8, 7, 3, 10)$ wyniesie $(3, 2, 1, 4)$.

W praktyce obserwacje czasami się powtarzają i wtedy ciąg sortujemy, a ranga obserwacji to średnia arytmetyczna rang wszystkich takich samych obserwacji, na przykład w ciągu $(10, 10, 7, 4, 4, 4)$ rangi poszczególnych obserwacji wynoszą $(5.5, 5.5, 4, 2, 2, 2)$. To tak zwane rangi *wiązane*[6]. Zazwyczaj testy oparte na rangach są odporne na niewielkie ilości rang wiązanych. Przy większej ich ilości testy się lekko modyfikuje (zob. [13]).

Definicja 7.7. Niech X i Y będą zmiennymi losowymi, a F i G odpowiednio ich dystrybuantami. Wtedy mówimy, że X jest *stochastycznie mniejsza* od Y (lub X jest *mniejsza* od Y w porządku stochastycznym zwykłym) (ozn. $X \leqslant^S Y$ lub ogólnie dla rozkładów $F \leqslant^S G$), gdy dla wszystkich $t \in \mathbb{R}$ zachodzi $F(t) \geqslant G(t)$.

7.3.1 Test losowości serii (Walda–Wolfowitza)

Ogólnie testami serii nazywamy testy, w których używa się pojęcia serii i wykorzystuje się własności z rozkładu z twierdzenia 7.6. Przedstawimy *test losowości serii* (ang. *Single-Sample Runs Test for Randomness, Wald–Wolfowitz Runs Test*).

Niech $X_1, \ldots, X_n$ będzie próbą. Testujemy hipotezę zerową i alternatywną

$$H_0: \text{ próba } X_1, \ldots, X_n \text{ jest losowa;}$$

$$H_1: \text{ nie jest losowa.}$$

Jeśli próba składa się bezpośrednio z dwóch symboli (na przykład w ankiecie odpowiedzi ‚tak' i ‚nie'), to statystyką testową jest liczba serii w tej próbie. Gdy obserwacje dotyczą cechy ilościowej, to najpierw wyznaczamy ciąg zero-jedynkowy $a = (a_1, \ldots, a_n)$ ze wzoru $a_i = 0$, gdy $X_i < md$, albo $a_i = 1$, gdy $X_i \geqslant md$ (gdzie md to mediana z próby $X_1, \ldots, X_n$). Wtedy statystyką testową jest liczba serii w ciągu a. Hipotezę o losowości odrzucamy, gdy liczba serii jest za mała (czyli mamy do czynienia z dodatnią korelacją w próbie) lub za duża (czyli mamy do czynienia z ujemną korelacją w próbie). Zbiór krytyczny jest więc dwustronny, a do wyznaczania zbioru krytycznego używamy kwantyli rozkładu zmiennej R z twierdzenia 7.6.

Możemy użyć też wersji jednostronnych tego testu, na przykład przyjąć hipotezę alternatywną, że próba jest ujemnie skorelowana i przyjąć zbiór krytyczny prawostronny.

Uwaga. Oczywiście liczba serii w ciągu zależy od kolejności, więc testujemy losowość próby w kolejności jej pobierania.

W `R` używamy funkcji `runs.test{lawstat}` (patrz przykład 7.8).

[6] ang. *tied ranks*

7.3.2 Testy rangowe

Testy rangowe, jak wskazuje nazwa, opierają się na rangach.

7.3.2.1 Test losowości Bartelsa

Jest to rangowa wersja testu losowości von Neumanna (zob. [2]).

Testujemy hipotezę zerową

$$H_0: \text{ próba } X_1, \ldots, X_n \text{ jest losowa;}$$

$$H_1: \text{ nie jest losowa.}$$

Niech $r_i = r_{X_1,\ldots,X_n}(X_i)$ oznacza rangę i-ej obserwacji w ciągu $X_1, \ldots, X_n$. Statystyka testowa jest dana wzorem

$$RVN = \frac{\sum_{i=1}^{n-1}(r_i - r_{i+1})^2}{\sum_{i=1}^{n}(r_i - (n+1)/2)^2}.$$

Statystyka ta ma asymptotycznie rozkład $N(2, \sigma)$, gdzie

$$\sigma = \sqrt{\frac{4(n-2)(5n^2-2n-9)}{5n(n+1)(n-1)^2}}.$$

Statystyka RVN przyjmuje małe wartości dla prób z trendem (dodatnio skorelowane), a duże dla prób ‚oscylujących' (ujemnie skorelowanych). W `R` używamy funkcji `bartels.test{lawstat}`.

Przykład 7.8. Zobaczmy, czy generator generuje liczby statystycznie losowe.

```
> X=runif(100)
> runs.test(X)

Runs Test - Two sided

data:  X
Standardized Runs Statistic = 0.40204, p-value = 0.6877

> bartels.test(X)

Bartels Test - Two sided

data:  X
Standardized Bartels Statistic = -0.1925, RVN Ratio = 1.9615,
p-value = 0.8474
```

Jak widać w obu testach nie odrzucamy hipotezy o losowości wygenerowanej próby.

7.3.2.2 Test sumy rang Wilcoxona (U-Manna–Whitneya)

Ten test traktuje się jako odpowiednik testu t dla dwóch prób niezależnych, ale bez założenia normalności. Formalnie jest to test *jednorodności*, to znaczy testujemy hipotezę, że dwie próby pochodzą z tego samego rozkładu.

Niech $\underline{X} = (X_1, \ldots, X_m)$ oraz $\underline{Y} = (Y_1, \ldots, Y_n)$ będą losowymi, niezależnymi próbami pochodzącymi z dwóch populacji o rozkładach danych przez odpowiednio dystrybuanty **ciągłe** F i G. Testujemy hipotezę zerową

$$H_0: \ F = G.$$

Możliwe są trzy hipotezy alternatywne

a) $F \leqslant^S G$ i $F \neq G$ (w `R` alternative='less');

b) $F \geqslant^S G$ i $F \neq G$ (w `R` 'greater');

c) ($F \leqslant^S G$ lub $F \geqslant^S G$) i $F \neq G$ (w `R` alternatywa domyślna);

Alternatywę c) możemy też bardziej ogólnie traktować jako $F \neq G$.

Oznaczmy przez

$$S_i = r_{X_1,\ldots,X_n,Y_1,\ldots,Y_m}(Y_i).$$

Niech

$$W = \sum_{i=1}^{n} S_i.$$

Statystyką testową jest

$$U' = W - \frac{n(n+1)}{2} - mn.$$

Przy prawdziwości hipotezy zerowej W ma rozkład, którego kwantyle uzyskujemy z funkcji `qwilcox`. Asymptotycznie U' ma rozkład normalny $N(E(U'), D(U'))$ dla

$$E(U') = \frac{mn}{2}$$

oraz

$$D(U') = \sqrt{\frac{mn(m+n+1)}{12}}.$$

Można korzystać z tego przybliżenia, gdy $n, m \geqslant 4$ oraz $n + m \geqslant 20$.

Duże wartości statystyki U' świadczą o dużej wartości W, a to oznacza, że rangi obserwacji cechy Y są większe niż cechy X, więc obserwacje cechy Y są większe niż cechy X. Oznacza to, że dla alternatywy a) zbiór krytyczny będzie prawostronny, dla b) lewostronny, a dla c) dwustronny. W `R` zaimplementujemy go funkcją `wilcox.test(X,Y,alternative=..`

Test ten można uogólnić na *model z przesunięciem*[7]. Testujemy w nim hipotezę zerową

$$H_0 : \ \forall t \in \mathbb{R} \quad F(t - \Delta) = G(t),$$

dla ustalonego Δ nazywanego *location shift*. Hipoteza ta jest równoważna hipotezie, że zmienna $X + \Delta$ ma taki sam rozkład jak zmienna Y (oznaczamy to $X + \Delta \stackrel{d}{=} Y$). Wtedy w szczególności $EY - EX = \Delta$. Hipotezy alternatywne formułujemy analogicznie jak wcześniej. Oczywiście dla $\Delta = 0$ otrzymujemy poprzednią wersję testu. W `R` implementujemy ją za pomocą funkcji

`wilcox.test(X,Y,mu=Δ,alternative=...)`.

Przykład 7.9. (Dane fikcyjne). Wylosowano grupę czwarto- i piątoklasistów i zmierzono ile minut dziennie oglądają telewizję. Uzyskano dane

```
> k4=c(25,28,43,34,24,88,36,39)
> k5=c(33,56,45,12,98,38)
```

Czy te dwie grupy się różnią?

Testujemy hipotezę, że rozkłady czasu oglądania telewizji w tych dwóch grupach są takie same, z alternatywą, że nie są.

```
> wilcox.test(k4,k5)

Wilcoxon rank sum test

data:  k4 and k5
W = 18, p-value = 0.4908
alternative hypothesis: true location shift is not equal to 0
```

Na poziomie istotności 0.05 nie ma podstaw do odrzucenia hipotezy o równości tych rozkładów.

[7] ang. *location-shift model*

7.3.2.3 Test znakowanych rang Wilcoxona dla prób zależnych

Załóżmy, że mamy próbę prostą zmiennej losowej dwuwymiarowej $(X_1, Y_1), \ldots, (X_n, Y_n)$ (to znaczy badamy cechy X i Y na tych samych osobnikach). Niech md_X oznacza medianę cechy X, a md_Y cechy Y. Niech $D_i = X_i - Y_i$ dla $i = 1, \ldots, n$ oraz niech F_D oznacza **ciągłą** dystrybuantę D.

Testujemy hipotezę zerową

$$H_0 : \ D \stackrel{d}{=} -D \text{ (równoważnie } F_D(t) = 1 - F_D(-t) \text{ lub } F_D = F_{-D}).$$

Zauważmy, że z hipotezy zerowej wynika w szczególności, że mediana zmiennej D wynosi zero. Możliwe są trzy hipotezy alternatywne

a) $F_D \leqslant^S F_{-D}$ i $F_D \neq F_{-D}$ (w R alternative='less');

b) $F_D \geqslant^S F_{-D}$ i $F_D \neq F_{-D}$ (w R 'greater');

c) ($F_D \leqslant^S F_{-D}$ lub $F_D \geqslant^S F_{-D}$) i $F_D \neq F_{-D}$ (w R alternatywa domyślna);

Statystyką testową jest

$$W = \sum_{i:D_i>0} r_{|D_1|,\ldots,|D_n|}(|D_i|).$$

Dla małych n używamy rozkładu Wilcoxona, dla $n > 16$ rozkład W przybliżamy rozkładem normalnym $N(E(W), D(W))$, gdzie

$$E(W) = \frac{n(n+1)}{4}$$

oraz

$$D(W) = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

W R zaimplementujemy go funkcją

```
wilcox.test(X,Y,paired=TRUE, alternative=...)
```

lub

```
wilcox.test(X-Y, alternative=...).
```

Możemy też testować hipotezę, że zmienna $D = X - \Delta - Y$ jest symetryczna wokół zera. W R implementujemy ją za pomocą funkcji

```
wilcox.test(X,Y,mu=Δ,paired=TRUE, alternative=...)
```

lub

```
wilcox.test(X-Y,mu=Δ, alternative=...)
```

Możemy też zastosować ten test dla zmiennej $D = X - \Delta$, używając

```
wilcox.test(X,mu=Δ, alternative=...)
```

(testujemy wtedy, że zmienna X ma rozkład symetryczny względem Δ, w szczególności, że mediana X jest równa Δ).

Przykład 7.10. Grupa losowo wybranych konsumentów oceniała równocześnie dwa gatunki masła, powiedzmy A i B, w skali od 0 do 100. Uzyskano dane

```
> A=c(34,67,44,50,60,70,45,51,67,62)
> B=c(37,70,50,55,54,75,55,69,65,80)
```

Czy masło B smakuje lepiej niż masło A?

```
> wilcox.test(A,B,paired=T,alternative="less")

Wilcoxon signed rank test with continuity correction

data:  A and B
V = 7.5, p-value = 0.02314
alternative hypothesis: true location shift is less than 0
```

Na poziomie istotności 0.05 stwierdzamy, że masło B smakuje lepiej.

7.4 Inne testy rangowe i testy jednorodności

Testy jednorodności[8] służą do testowania hipotezy, że dwie (lub więcej) prób mają ten sam rozkład.

7.4.1 Test Kołmogorowa–Smirnowa

Test Kołmogorowa–Smirnowa jest rozwinięciem testu zgodności Kołmogorowa.

Załóżmy, że niezależne od siebie próby proste $X_1, \ldots, X_n$ oraz $Y_1, \ldots, Y_m$ pochodzą z rozkładów ciągłych danych odpowiednio przez dystrybuanty F_X i F_Y.

Testujemy hipotezę zerową i alternatywną

$$H_0 : F_X = F_Y \quad H_1 : \neg H_0.$$

Statystyka testowa jest równa

$$D_{n,m} = \sup_{t \in \mathbb{R}} |\hat{F}_{X,n}(t) - \hat{F}_{Y,m}(t)|,$$

gdzie $\hat{F}_{X,n}$ i $\hat{F}_{Y,m}$ są odpowiednio dystrybuantą empiryczną z pierwszej i drugiej próby. Duże wartości statystyki testowej świadczą przeciwko hipotezie zerowej. Smirnow pokazał (zob. [22]), że dystrybuanta statystyki $\sqrt{\frac{nm}{n+m}} D_{n,m}$ przy prawdziwości hipotezy zerowej jest zbieżna (gdy $n \to \infty$, $m \to \infty$ oraz $\sqrt{\frac{nm}{n+m}} \to \infty$) do dystrybuanty danej wzorem (dla $t > 0$)

$$F_D(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{2k^2 t^2}.$$

(Jest to ta sama dystrybuanta, która występuje w teście Kołmogorowa).

W `R` używamy funkcji `ks.test`.

Przykład 7.11. (Dane fikcyjne). W dwóch jeziorach, A i B, losowo wyłowiono po kilka leszczy i zmierzono ich długość. Dla jeziora A uzyskano pomiary $30.4, 34.5, 29.8, 40.7, 12.5, 31.0$, a dla jeziora B $11.1, 14.7, 23.8, 32.7, 25.5, 36.7$. Czy rozkłady długości leszczy w tych jeziorach są takie same?

```
> A=c(30.4,34.5,29.8,40.7,12.5,31.0)
> B=c(11.1,14.7,23.8,32.7,25.5,36.7)
> ks.test(A,B)

Two-sample Kolmogorov-Smirnov test

data:  A and B
D = 0.5, p-value = 0.474
alternative hypothesis: two-sided
```

Otrzymano pvalue równe 0.474, więc nie ma podstaw (na poziomie istotności 0.05) do odrzucenia hipotezy o równości badanych rozkładów.

Dla większej liczby prób możemy użyć na przykład testu Kruskala–Wallisa, który zostanie przedstawiony później.

7.4.2 Test jednorodności χ^2 dla rozkładów dyskretnych

Test ten służy do testowania hipotezy, że k ($k \geqslant 2$) niezależnych prób ma taki sam rozkład dyskretny. Bardziej precyzyjnie, załóżmy, że $X_{i,j}$ to j-ta obserwacja z i-ej próby (dla $i = 1, \ldots, k$ oraz $j = 1, \ldots, n_i$) oraz że te próby pochodzą z rozkładów dyskretnych przyjmujących wartości $w_1, \ldots, w_s$.

Testujemy hipotezę zerową, że próby pochodzą z tego samego rozkładu z hipotezą alternatywną $H_1 : \neg H_0$.

Definiujemy tablicę częstości $N = [N_{i,j}]_{i=1,\ldots,k;j=1,\ldots,s}$, gdzie

$$N_{i,j} = \#\{X_{i,t} \ : \ X_{i,t} = w_j\}.$$

Następnie procedura testowa jest identyczna jak w teście niezależności χ^2.

[8] ang. *tests for homogeneity.*

Przykład 7.12. (Dane fikcyjne). W trzech miastach A, B i C badano jak często uczniowie liceum czytają książki w skali nigdy, rzadko, czasami i często. Pobrano próby losowe i uzyskano następującą tablicę częstości N

```
   nigdy rzadko czasami czesto
A    13     10       4      3
B    34     30       8      9
C    22      8       9     15
```

Używamy teraz funkcji `chisq.test`.

```
> chisq.test(N)

Pearson's Chi-squared test

data:  N
X-squared = 13.351, df = 6, p-value = 0.0378
```

Na poziomie istotności 0.05 odrzucamy hipotezę zerową. Stwierdzamy więc, że częstotliwość czytania książek przez licealistów nie jest taka sama w tych trzech miastach.

Test jednorodności χ^2 jest (zazwyczaj) szczególnym przypadkiem testu niezależności χ^2. Mianowicie, jeśli badamy cechę X na k podpopulacjach wyznaczonych przez poziomy zmiennej jakościowej Y, to niezależność X i Y to to samo, co równość rozkładów cechy X na poziomach cechy Y (w powyższym przykładzie cechą X byłaby częstotliwość czytania, a Y miastem o poziomach A, B i C).

7.4.3 Test Kruskala–Wallisa

Test Kruskala–Wallisa jest testem rangowym jednorodności dla k grup ($k \geqslant 2$) dla prób niezależnych (i można go uznać za uogólnienie testu Wilcoxona).

Niech

$$X_{11}, \ldots, X_{1n_1}$$

$$\vdots$$

$$X_{k1}, \ldots, X_{kn_k}$$

będą k niezależnymi próbami prostymi pochodzącymi z rozkładów danych odpowiednio przez dystrybuanty ciągłe $F_1, \ldots, F_k$.

Testujemy hipotezę zerową

$$H_0 : F_1 = \ldots = F_k$$

i alternatywną

$$H_1 : \exists\, i, j \quad : \quad F_i \leqslant^S F_j \text{ lub } F_i \geqslant^S F_j \text{ (w szczególności } F_i \neq F_j),$$

gdzie nierówności stochastyczne są takie, jak w definicji7.7.

Niech $n = n_1 + \cdots + n_k$, a $r_{ij} = r_{X_{11},\ldots,X_{kn_k}}(X_{ij})$ oznacza rangę obserwacji X_{ij} w próbie powstałej z połączenia k powyższych prób. Niech $R_i = \sum_{j=1}^{n_i} r_{ij}$ oznacza sumę rang obserwacji z i-ej próby. Wtedy statystyka testowa jest dana wzorem

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1).$$

Dla małych wartości n_i rozkład statystyki H wyznacza się metodami kombinatorycznymi. Dla większych korzysta się z twierdzenia, że asymptotycznie rozkład H jest równy rozkładowi $\chi^2(k-1)$ (przybliżenie jest już dobre dla $n_i > 5$ i $k \geqslant 4$).

Zbiór krytyczny jest prawostronny. Intuicyjnie można to uzasadnić następująco. Gdy hipoteza zerowa jest prawdziwa, to obserwacje z tych k grup powinny się równomiernie przemieszać, więc (dla uproszczenia załóżmy, że grupy są równoliczne) suma rang w każdej grupie powinna być podobna. Wiedząc, że $R_1 + \cdots + R_k = n(n+1)/2$, gdy w jednej grupie suma rang wzrasta, to wartość funkcji H też wzrasta. Duże wartości H świadczą więc o nierównomierności sumy rang w grupach.

Przykład 7.13. (Dane fikcyjne). Na pewnej uczelni wylosowano po kilku studentów z trzech wydziałów A, B, C i poproszono ich aby dokładnie liczyli jaką kwotę wydadzą na żywność w nadchodzącym miesiącu. Po miesiącu uzyskano dane

```
> A=c(623,714,699,807,430,766)
> B=c(777,789,804,680,823,902,759)
> C=c(550,655,780,781,644)
```

Testujemy, czy na tych wydziałach rozkład wydatków na żywność jest taki sam.

```
> kruskal.test(list(A,B,C))

Kruskal-Wallis rank sum test

data:  list(A, B, C)
Kruskal-Wallis chi-squared = 4.5383, df = 2, p-value = 0.1034
```

Jak widać na standardowym poziomie istotności nie ma podstaw, aby odrzucić hipotezę o równości tych rozkładów na tych trzech wydziałach.

7.4.4 Test McNemary

Test McNemary służy do testowania hipotezy o równości rozkładów dwóch zmiennych o rozkładach dwupunktowych dla prób zależnych (powiązanych, sparowanych).

Załóżmy, że mamy próbę prostą $(X_1, Y_1), \ldots, (X_n, Y_n)$ zmiennej losowej dwuwymiarowej (X, Y), takiej, że X i Y mają rozkłady dwupunktowe (dychotomiczne, binarne). Testujemy hipotezę, że zmienne X i Y mają te same rozkłady, naprzeciwko hipotezy, że nie mają tych samych rozkładów.

Dla ustalenia uwagi, załóżmy, że zmienne X i Y przyjmują wartości 0 i 1. Zdefiniujmy $B = \#\{(X_i, Y_i) : X_i = 0, Y_i = 1\}$ oraz $C = \#\{(X_i, Y_i) : X_i = 1, Y_i = 0\}$. Statystykę testową definiujemy wzorem

$$\chi^2 = \frac{(B-C)^2}{B+C}$$

(lub z tzw. korektą ciągłości $\chi^2 = (|B-C|-1)^2/(B+C)$). Przy prawdziwości hipotezy zerowej statystyka testowa ma asymptotycznie rozkład $\chi^2(1)$ (zob. [19]), a zbiór krytyczny jest prawostronny (intuicyjnie wynika to z faktu, że hipoteza zerowa jest równoważna hipotezie, że $P(X = 1, Y = 0) =: p_{10} = p_{01} := P(X = 0, Y = 1)$).

Przykład 7.14. (Dane fikcyjne). Badano zachorowalność na pewną chorobę w wieku 10 lat (cecha X) i 12 lat (cecha Y). Losowo wybrane 120 dzieci pytano więc dwukrotnie (w wieku 10 i 12 lat) czy w ostatnim półroczu były chore. Uzyskano dane

```
> Dane=matrix(c(40,20,30,30),ncol=2,
+             dimnames=list(c("chore jako 10 latek",
+                             "zdrowe jako 10 latek"),
+                           c("chore jako 12 latek",
+                             "zdrowe jako 12 latek")))
> Dane
                    chore jako 12 latek zdrowe jako 12 latek
chore jako 10 latek                  40                   30
zdrowe jako 10 latek                 20                   30
```

Testujemy hipotezę o równości rozkładów cech X i Y.

```
> mcnemar.test(Dane)

McNemar's Chi-squared test with continuity correction

data:  Dane
McNemar's chi-squared = 1.62, df = 1, p-value = 0.2031
```

Na poziomie istotności 0.05 nie ma podstaw do odrzucenia hipotezy o równości tych rozkładów.

Uogólnieniem tego testu na przypadek większej liczby prób (tj. na badanie równości rozkładów współrzędnych k wymiarowej zmiennej $(X_1, \ldots, X_k)$, przy założeniu dwupunktowości rozkładów) jest test Q Cochrana (zob. [13]).

7.4.5 Test Friedmana

Test Friedmana jest testem rangowym jednorodności rozkładów k zmiennych dla prób zależnych przy założeniu ciągłości rozkładów.

Załóżmy, że mamy n elementową próbę prostą

$$\begin{matrix} X_{11}, & \ldots, & X_{1k} \\ \vdots & & \\ X_{n1}, & \ldots, & X_{nk} \end{matrix}$$

zmiennej losowej k-wymiarowej $(X_1, \ldots, X_k)$, takiej, że zmienne $X_1, \ldots, X_k$ mają rozkłady dany odpowiednio przez dystrybuanty ciągłe $F_1, \ldots, F_k$.

Testujemy hipotezę zerową

$$H_0 : F_1 = \ldots = F_k$$

i alternatywną

$$H_1 : \exists\, i, j \quad : \quad F_i \leqslant^S F_j \text{ lub } F_i \geqslant^S F_j \text{ (w szczególności } F_i \neq F_j).$$

Niech $r_{ij} = r_{X_{i1},\ldots,X_{ik}}(X_{ij})$, a $R_i = \sum_{j=1}^{n} r_{ji}$ (tzn. przy powyższej formie danych rangujemy „w poziomie", a sumujemy rangi „w pionie"). Statystyka testowa jest dana wzorem

$$T = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1),$$

która asymptotycznie ma rozkład $\chi^2(k-1)$ (zob. [9]). Zbiór krytyczny jest prawostronny.

Przykład 7.15. (Dane fikcyjne). Chcemy ocenić, czy troje skrzypiec (A, B i C) są podobne. Na każdych gra 10-ciu skrzypków (z zawiązanymi oczami, w losowej kolejności, etc). Każdy ocenia skrzypce w skali od 1 do 10. Uzyskano dane:

```
> Skrzypce
     A   B   C
1   5.4 7.4 2.1
2   4.5 5.0 2.0
3   3.7 5.8 5.3
4   6.7 6.9 4.5
5   8.9 2.5 2.8
6   5.6 7.8 2.6
7   6.0 4.5 4.1
8   7.7 8.6 9.8
9   8.8 8.9 7.8
10 9.0 8.5 6.9
```

Stosujemy test Friedmana.

```
> friedman.test(Skrzypce)

Friedman rank sum test

data:  Skrzypce
Friedman chi-squared = 6.2, df = 2, p-value = 0.04505
```

Na poziomie istotności 0.05 odrzucamy hipotezę zerową i stwierdzamy, że rozkłady ocen tych trzech skrzypiec się różnią.

Znormalizowana statystyka Friedmana nosi nazwę współczynnika zgodności W Kenndala, który służy do oceny zgodności ocen eksperckich (zob. [14]).

7.4.6 Test korelacji rang Spearmana

Załóżmy, że mamy próbę prostą $(X_1, Y_1), \ldots, (X_n, Y_n)$ zmiennej losowej dwuwymiarowej (X, Y). Niech $R_i = r_{X_1,\ldots,X_n}(X_i)$ oraz $S_i = r_{Y_1,\ldots,Y_n}(Y_i)$. Definiujemy wtedy *współczynnik korelacji Spearmana* ρ_S jako współczynnik korelacji liniowej rang $R_1, \ldots, R_n$ oraz $S_1, \ldots, S_n$ lub równoważnie

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n(n^2-1)}.$$

Wprost z definicji wynika na przykład, że gdy ten współczynnik ma wartość 1 to związek między zmiennymi jest rosnący, etc. Zaletą tego współczynnika jest też to, że można go używać dla zmiennych mierzonych w skali porządkowej.

Testujemy hipotezę zerową, że zmienne X i Y są nieskorelowane, z trzema możliwymi hipotezami alternatywnymi (korelacja niezerowa albo dodatnia albo ujemna). Dla małych liczności próby można wyznaczyć dokładny rozkład ρ_S. Dla $10 \leqslant n \leqslant 200$ statystyka

$$T = \rho_S \sqrt{\frac{n-2}{1-\rho_S^2}}$$

ma w przybliżeniu rozkład $t(n-2)$, natomiast asymptotycznie $\rho_S\sqrt{n-1}$ ma rozkład $N(0,1)$.

Przykład 7.16. W pewnym kursie uczestniczyło 15-stu kursantów. Przeprowadzono testy na początku oraz na końcu kursu i poszczególni kursanci zajęli na nich następujące miejsca

```
> WST=1:15
> KON=c(7,14,10,3,5,13,1,12,8,9,15,4,11,6,2)
```

(czyli na przykład jeden kursant był najlepszy w teście wstępnym, a siódmy na końcowym). Czy oceny wstępne i końcowe są skorelowane?

```
> cor(WST,KON,method="spearman")
[1] -0.1607143
> cor.test(WST,KON,method="spearman")

Spearman's rank correlation rho

data:  WST and KON
S = 650, p-value = 0.5667
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1607143
```

Jak widać współczynnik korelacji Spearmana dla tych prób wynosi -0.1607143, a pvalue z testu 0.5667, więc hipotezy zerowej nie odrzucamy.

7.5 Paradoks Simpsona

Paradoksem Simpsona nazywamy sytuację (w rachunku prawdopodobieństwa i statystyce), gdy analiza na pewnej populacji daje inne (sprzeczne) wyniki, niż ta analiza przeprowadzona na pewnych podzbiorach tej populacji. Zilustrujemy to dwoma standardowymi przykładami.

Przykład 7.17. (Dane fikcyjne). W tablicy `Katar` znajdują się liczby wyleczonych i niewyleczonych pacjentów leczonych lekiem A albo B na lekki albo mocny katar.

```
> Katar=array(c(600, 90, 20, 300,
+                 300, 10, 80, 600),
+          dim = c(2, 2, 2),
+           dimnames = list(
```

```
+                       Lek = c("A", "B"),
+                       Rodzaj_kataru = c("lekki", "mocny"),
+                       Pacjenci = c("wyleczeni", "niewyleczeni")))
>
> Katar
, , Pacjenci = wyleczeni

   Rodzaj_kataru
Lek lekki mocny
  A   600    20
  B    90   300

, , Pacjenci = niewyleczeni

   Rodzaj_kataru
Lek lekki mocny
  A   300    80
  B    10   600
```

Wyliczmy skuteczność (to jest frakcję wyleczonych pacjentów) tych leków dla pacjentów z lekkim i mocnym katarem.

```
> Skutecznosc=apply(Katar,c(1,2),function(x) x[1]/sum(x))
> Skutecznosc
   Rodzaj_kataru
Lek     lekki     mocny
  A 0.6666667 0.2000000
  B 0.9000000 0.3333333
```

Jak widzimy lek *B* jest lepszy w leczeniu zarówno lekkiego, jak i mocnego kataru. Jeśli połączymy pacjentów (nie weźmiemy pod uwagę jaki rodzaj kataru mieli), to uzyskamy wynik przeciwny, lek *A* jest skuteczniejszy.

```
> Skutecznosc_bez_rozroznienia_rodzaju_kataru=
+     apply(apply(Katar,c(1,3),sum),1,function(x) x[1]/sum(x))
>
> Skutecznosc_bez_rozroznienia_rodzaju_kataru
   A    B
0.62 0.39
```

Ten paradoks był możliwy dlatego, że liczba leczonych pacjentów w tych czterech grupach nie była porównywalna, czego oczywiście staramy się unikać w zaplanowanych badaniach.

```
> apply(Katar,c(1,2),sum)
   Rodzaj_kataru
Lek lekki mocny
  A   900   100
  B   100   900
```

Jako ćwiczenie można do tych danych użyć testu na porównywanie frakcji.

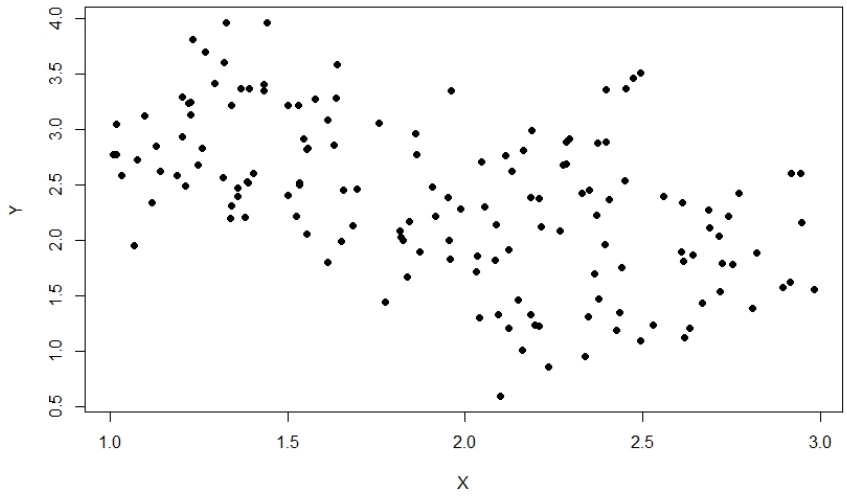
Kolejny przykład dotyczy cech ilościowych.

Przykład 7.18. (Dane fikcyjne). W danych `Dane`

```
           X         Y R
1   1.097196 3.1257216 1
2   1.684475 2.1353762 1
3   1.017309 2.7681811 1
```

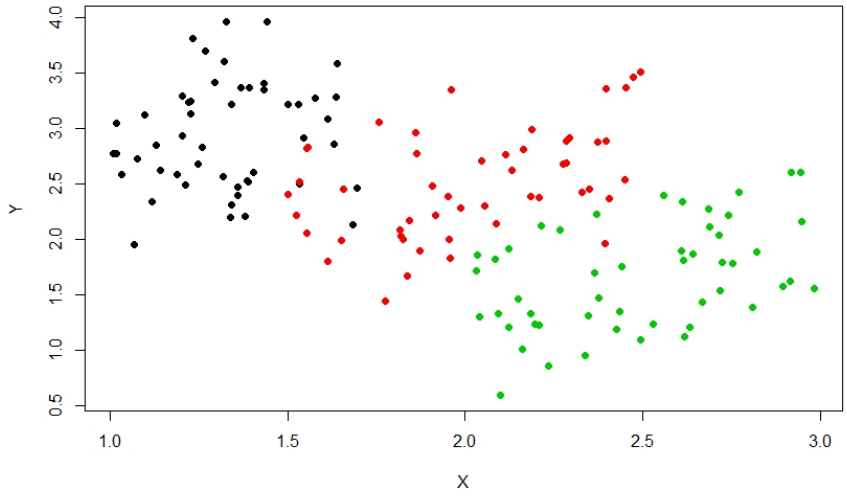
zawarto pomiary następujących zmiennych dla pewnego gatunku chrząszcza: X długość, Y waga oraz R rodzaj lasu (1– liściasty, 2– mieszany, 3– iglasty) w jakim występuje. Gdy nie uwzględnimy zmiennej R uzyskamy następujący rysunek rozrzutu oraz współczynnik korelacji liniowej.

```
> cor(Dane$X,Dane$Y)
[1] -0.5050072
```



Uzyskujemy wniosek, że im ten chrząszcz jest dłuższy, tym jest (średnio) lżejszy. Gdy uwzględnimy miejsce występowania, uzyskamy wnioski przeciwne. W każdym rodzaju lasu im ten chrząszcz jest dłuższy, tym jest (średnio) cięższy.

```
> for (i in 1:3) print(cor(Dane$X[Dane$R==i],Dane$Y[Dane$R==i]))
[1] 0.1237896
[1] 0.4575853
[1] 0.4477652
```



Przykłady te obrazują nam, że uwzględniając bądź nie jakąś zmienną, możemy uzyskać jakościowo różne wnioski. W modelowaniu (szczególnie złożonych układów) istnieje więc niebezpieczeństwo, że uzyskaliśmy błędne wnioski, ponieważ nie uwzględniliśmy jakiejś istotnej zmiennej (o istnieniu której możemy nawet nie mieć pojęcia).

7.6 Współczynnik τ Kendalla

W tym podrozdziale przedstawimy nieparametryczny współczynnik korelacji τ-Kendalla używany do badania niezależności zmiennych losowych.

Definicja 7.19. Rozważmy wektor losowy (X, Y). Wtedy *współczynnikiem korelacji τ-Kendalla* (między zmiennymi X i Y) nazywamy

$$\tau = 2P((Y_2 - Y_1)(X_2 - X_1) > 0) - 1,$$

gdzie X_1, X_2 to niezależne zmienne losowe o takim samym rozkładzie jak X, a Y_1, Y_2 to niezależne zmienne losowe o takim samym rozkładzie jak Y.

Zauważmy, że gdy X i Y są niezależne, to

$$\begin{aligned}\tau &= 2P((Y_2 - Y_1)(X_2 - X_1) > 0) - 1\\ &= 2(P(Y_2 - Y_1 > 0, X_2 - X_1 > 0) + P(Y_2 - Y_1 < 0, X_2 - X_1 < 0)) - 1\\ &= 2(P(Y_2 - Y_1 > 0)P(X_2 - X_1 > 0) + P(Y_2 - Y_1 < 0)P(X_2 - X_1 < 0)) - 1\\ &= 2(1/2 \cdot 1/2 + 1/2 \cdot 1/2) - 1 = 0.\end{aligned}$$

Implikacja w drugą stronę nie zachodzi.

Definicja 7.20. Dla próby $(X_1, Y_1), \ldots, (X_n, Y_n)$ *statystyką Kendalla z próby* nazywamy

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sign}((Y_j - Y_i)(X_j - X_i)),$$

a

$$\hat{\tau} = \frac{K}{n(n-1)/2}$$

nazywamy *współczynnikiem korelacji Kendalla z próby*.

Testujemy hipotezę zerową (przy założeniu, że rozkład wektora (X, Y) jest ciągły[9])

$$H_0 : X \text{ i } Y \text{ są niezależne}$$

(w szczególności $\tau = 0$). Możliwe są trzy alternatywy

$$H_1 : a)\ \tau \neq 0 \quad b)\ \tau > 0 \quad c)\ \tau < 0.$$

Używamy statystyki testowej K. Dla powyższych alternatyw zbiór krytyczny jest odpowiednio dwustronny, prawostronny i lewostronny. Dla małych n trzeba wyznaczać rozkład K dla każdego n osobno. Dla dużych n wykorzystujemy aproksymację rozkładem normalnym wykorzystując fakt, że

$$E(K) = 0$$

oraz

$$D(K) = \sqrt{\frac{n(n-1)(2n+5)}{18}}.$$

Przykład 7.21. Rozważmy problem z przykładu 7.10. Wyznaczmy współczynnik korelacji Kendalla dla ocen masła, a następnie przetestujmy hipotezę o niezależności ocen.

```
> cor(A,B,method = "kendall")
[1] 0.6590909
> cor.test(A,B,method="kendall")

Kendall's rank correlation tau

data:  A and B
z = 2.6146, p-value = 0.008933
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.6590909
```

Jak widać na poziomie 0.05 oceny są zależne.

[9] Przy tym założeniu dla każdego i, j $(Y_j - Y_i)(X_j - X_i)$ jest niezerowe.

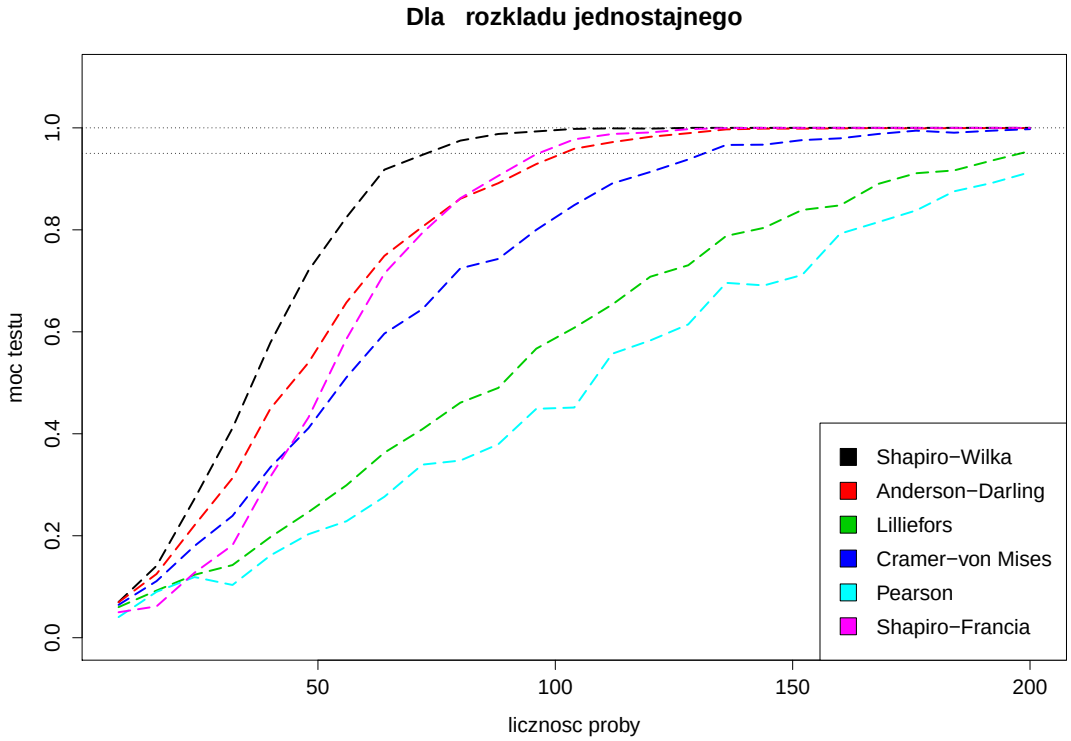
Rozdział 8

Testy normalności

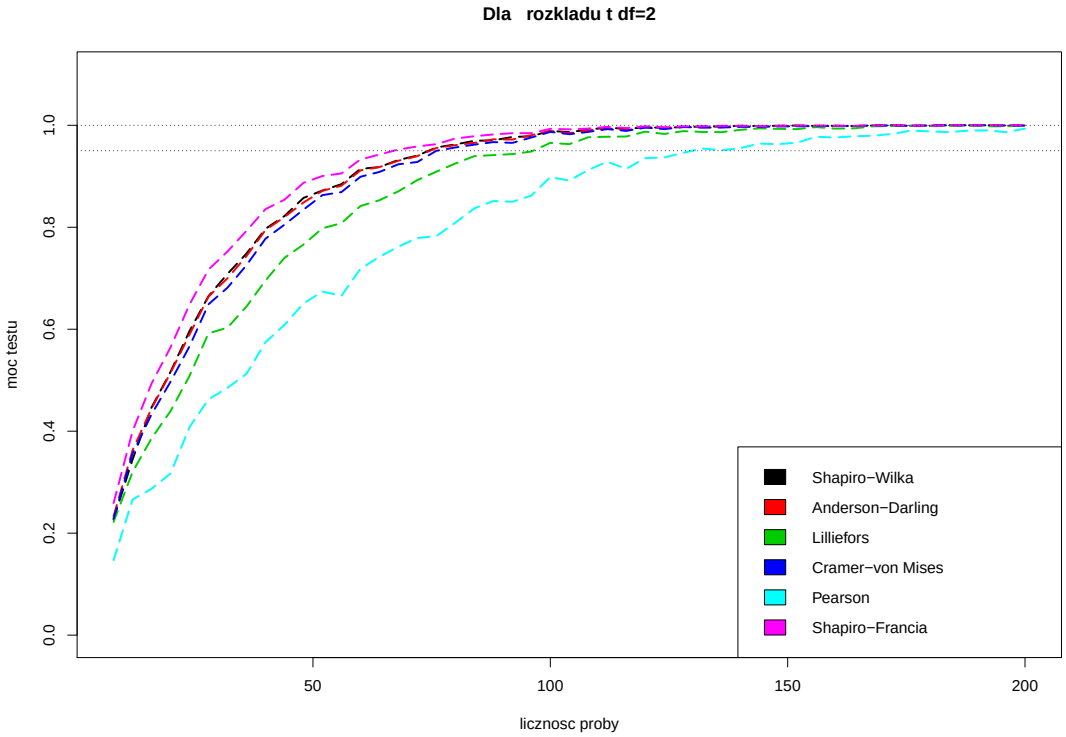
Testy normalności testują hipotezę, że próba pochodzi z rozkładu normalnego. Wiele metod wymaga normalności badanej zmiennej, więc powstało (i powstaje) wiele takich testów. Wśród nich są następujące testy (w `R` znajdują się w pakiecie `nortest`):

1. Shapiro–Wilka (`shapiro.test(X)`);
2. Andersona–Darlinga (`ad.test(X)`);
3. χ^2-Pearsona (`pearson.test(X)`);
4. Craméra–von Misesa (`cvm.test(X)`);
5. Lillieforsa (`lillie.test(X)`);
6. Shapiro–Francii (`sf.test(X)`);

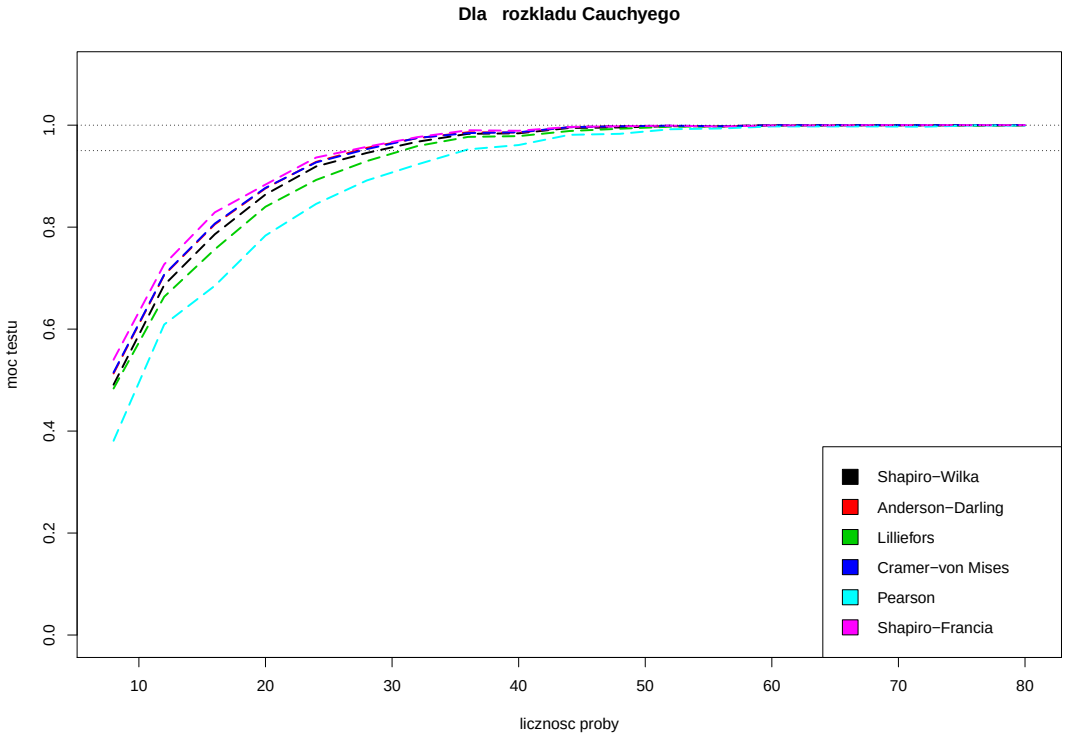
Oczywiście powstaje pytanie, który z nich stosować. Niestety nie da się wyznaczyć mocy tych testów, ponieważ hipoteza alternatywna jest zbyt złożona. Moc wyznacza się więc symulacyjnie: dla ustalonej liczności próby losujemy wielokrotnie próbę z rozkładu, który nie jest normalny, i sprawdzamy jaka jest frakcja odrzuceń hipotezy zerowej na ustalonym poziomie istotności (u nas to będzie 0.05). Rozkład normalny można scharakteryzować tym, że to rozkład jednomodalny, symetryczny o ‚cienkich' ogonach. Do symulacji mocy testu bierze się więc próby z rozkładów symetrycznych, ale o ‚grubych' ogonach lub jeszcze ‚cieńszych', rozkładów skośnych (o różnych ogonach), czy rozkłady dwu- lub trzymodalnych.



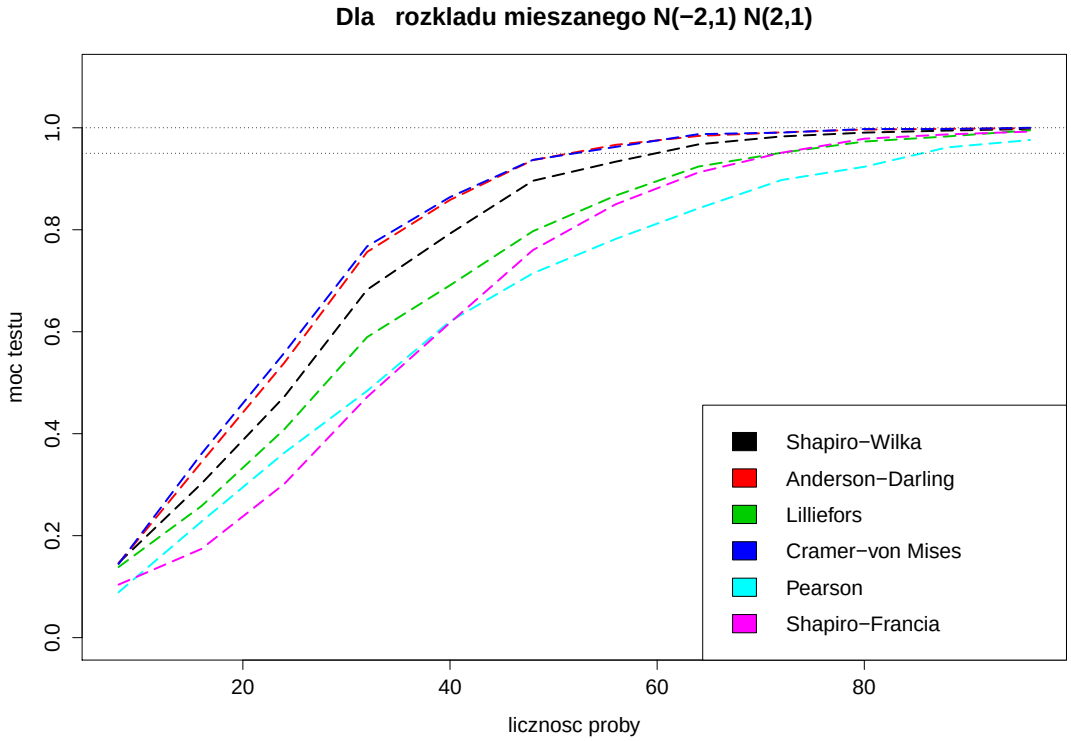
Jak widać dla rozkładu jednostajnego najlepszym (tj. najmocniejszym) testem jest test Shapiro–Wilka, a potem Andersona–Darlinga oraz Shapiro–Francii. Moc około 0.95 test Shapiro–Wilka uzyskuje dla liczności prób około 70. Pozostałe testy potrzebują ponad 100, a najgorsze nawet i 200. Zwraca uwagę fakt, że dla małych liczności prób moc jest mała (nawet 0.1).



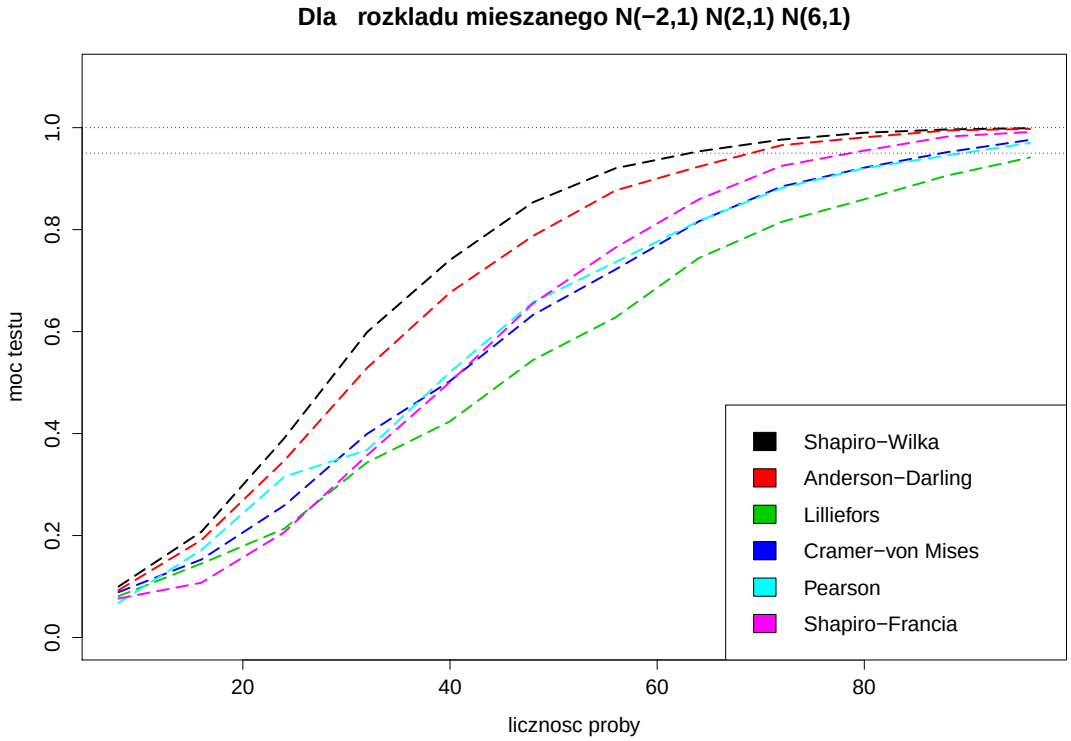
Dla rozkładu t-Studenta o 2 stopniach swobody wszystkie testy oprócz testu χ^2-Pearsona zachowują się podobnie i osiągają moc 0.95 dla liczności około 70.



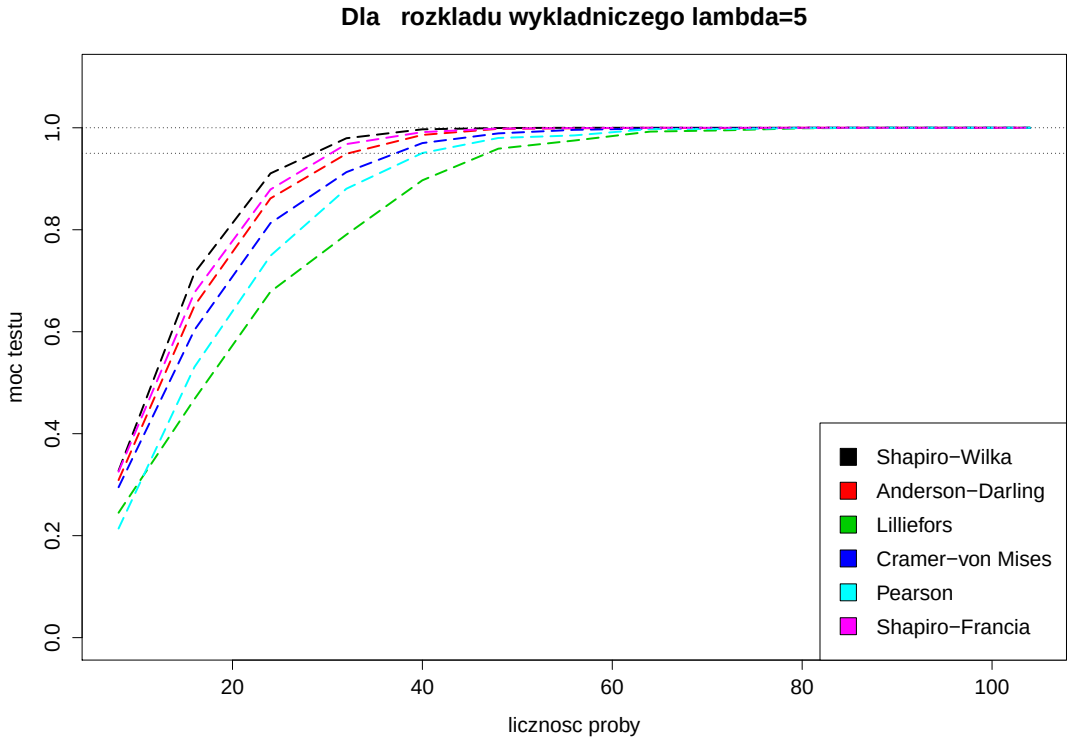
Dla rozkładu Cauchego, a więc rozkładu o bardzo grubych ogonach, wszystkie testy dla liczności próby około 35 osiągają moc 0.95.

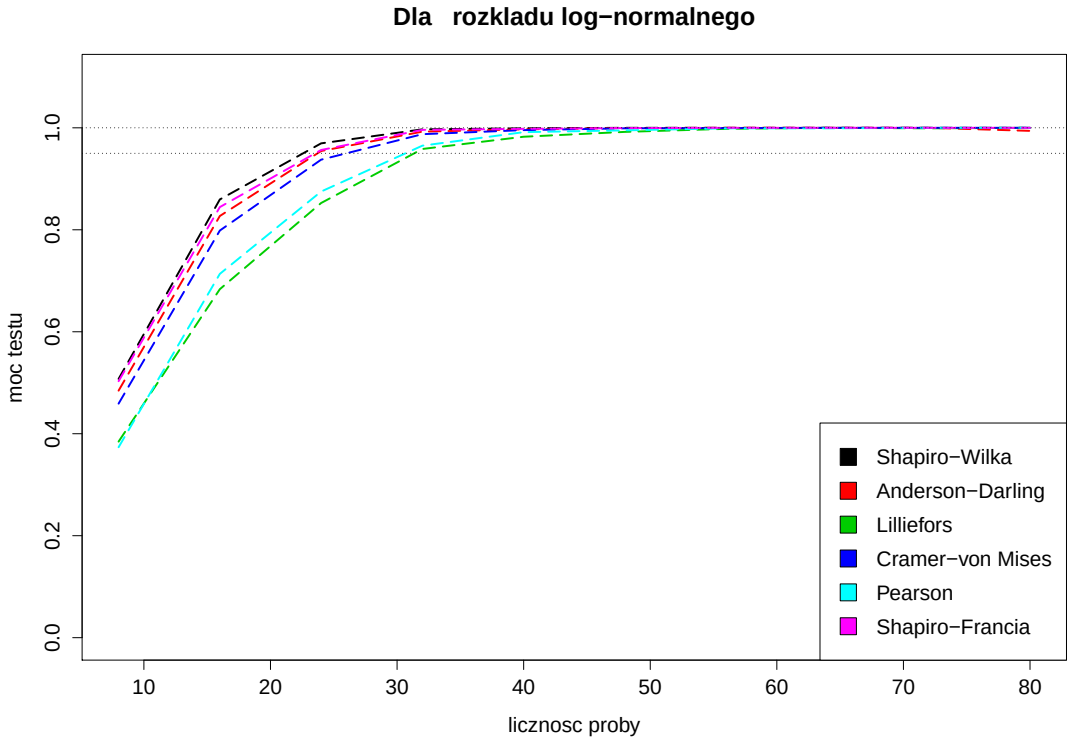


Dla rozkładu symetrycznego, dwumodalnego o ogonach podobnych do rozkładu normalnego najmocniejsze są testy Craméra–von Misesa i Andersona–Darlinga, słabszy jest test Shapiro–Wilka.

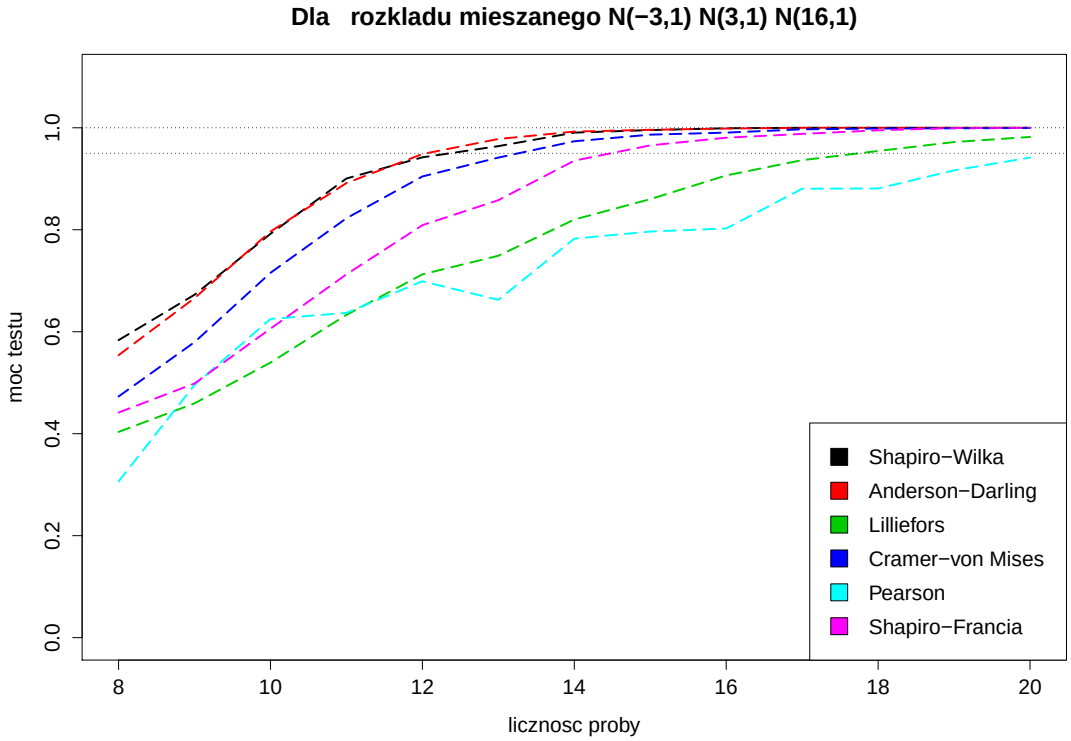


Dla rozkładu niesymetrycznego trzymodalnego najmocniejszy jest test Shapiro–Wilka, a trochę słabszy jest test Andersona–Darlinga.

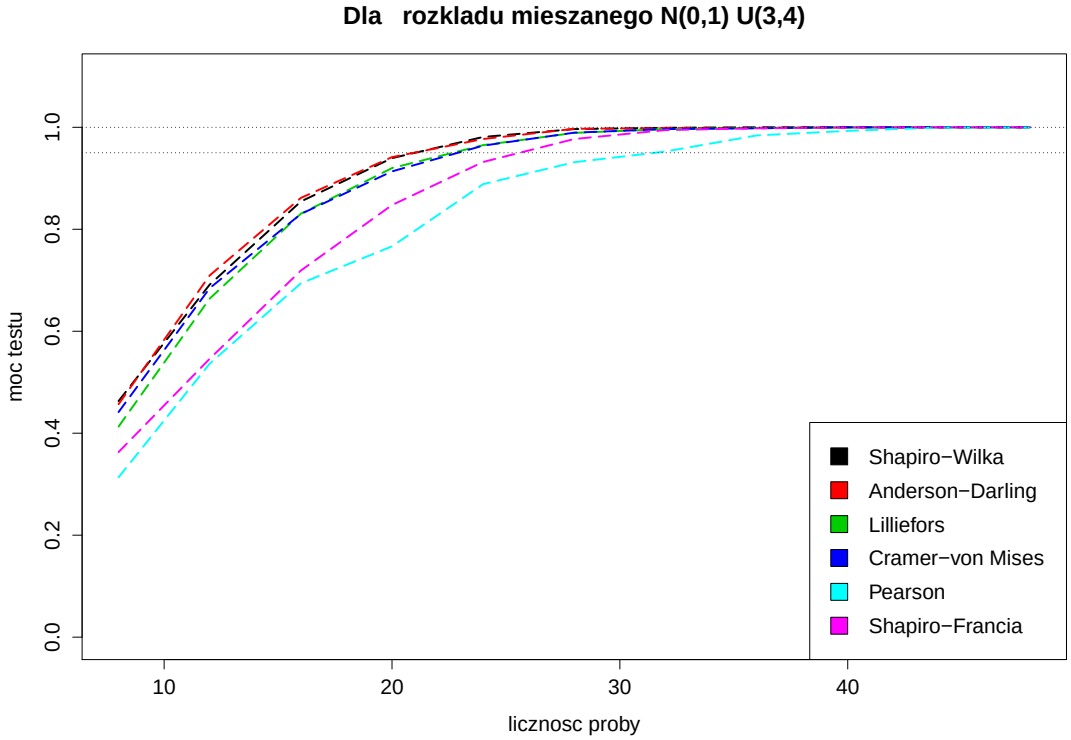




Dla rozkładów skośnych (wykładniczego i log-normalnego) wszystkie testy działają dobrze, najmocniejszym zaś wydaje się test Shapiro–Wilka.



Jak widać, gdy rozkład bardzo odbiega od rozkładu normalnego, już dla liczności próby 12 możemy osiągnąć moc w okolicach 0.95.



Na koniec symulacja dla ‚egzotycznego' rozkładu (mieszanego normalnego i jednostajnego).

Podsumowanie. Jak widać najtrudniej od rozkładu normalnego jest odróżnić rozkład jednostajny. Przy rozkładach skośnych lub o ‚grubych' ogonach testy radzą sobie dużo lepiej. Żaden test nie jest jednoznacznie najlepszy, niemniej widać, że testy Shapiro–Wilka i Andersona–Darlinga można uznać ogólnie za najlepsze.

Rozdział 9

Wybrane zagadnienia

9.1 Estymacja punktowa nieparametryczna

Z wnioskowaniem statystycznym nieparametrycznym mamy do czynienia, gdy przestrzeń parametrów jest nieskończenie wymiarowa. Rozważymy dwie sytuacje, gdy badamy zmienną o rozkładzie ciągłym danym gęstością f, więc mamy do czynienia z przestrzenią statystyczną

$$\mathcal{P} = \{P_f \ : \ f \in D\}$$

składającą się ze wszystkich rozkładów ciągłych P_f zadanych przez gęstość $f \in D$, lub jeszcze ogólniej przestrzeń statystyczna składa się ze wszystkich rozkładów zadanych dystrybuantą $F \in \mathcal{F}$

$$\mathcal{P} = \{P_F \ : \ F \in \mathcal{F}\}.$$

Szukamy więc estymatora gęstości $\hat{f}$ lub w drugiej sytuacji estymatora dystrybuanty $\hat{F}_n$. W przypadku estymacji gęstości kryterium oceny takiego estymatora jest *scałkowany błąd średniokwadratowy*[1].

Definicja 9.1. *Scałkowanym błędem średniokwadratowym estymatora $\hat{f}_n$ gęstości f* nazywamy

$$R(\hat{f}_n) = E \int_{\mathbb{R}} \left(\hat{f}_n(x) - f(x)\right)^2 dx.$$

Wprost z definicji wynika, że błąd ten przyjmuje wartości nieujemne, a im jest mniejszy, tym estymator $\hat{f}_n$ lepiej szacuje f (w skrajnej sytuacji, wartość $R(\hat{f}_n) = 0$ oznacza, że dla prawie wszystkich prób wartość estymatora jest równa prawie wszędzie szukanej gęstości f).

Przedstawimy dwie podstawowe metody estymacji gęstości: histogram oraz estymację jądrową[2].

Estymacja dystrybuanty

Niech $(X_1, \ldots, X_n)$ będzie próbą prostą wylosowaną z rozkładu P_F dla pewnej dystrybuanty $F \in \mathcal{F}$. Wyznaczymy teraz estymator największej wiarygodności dystrybuanty. Niech

$$L(F; (X_1, \ldots, X_n) = \prod_{i=1}^{n} P_F(X_i).$$

Kiefer i Wolfowitz w 1956 roku udowodnili, że funkcję L maksymalizuje *dystrybuanta empiryczna*[3]

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i). \tag{9.1}$$

Dowód polega na zauważeniu, że jeśli dystrybuanta F jest ciągła w którymś z punktów X_i, to $L(F) = 0$. Niech więc $p_i = P_F(X_i)$ i problem sprowadzamy do sytuacji skończenie wymiarowej: dla jakich dodatnich $p_1, \ldots, p_n$ takich, że $p_1 + \cdots + p_n = 1$ iloczyn $p_1 \cdots p_n$ jest największy? Rozwiązaniem jest $p_1 = \cdots = p_n = 1/n$, więc dystrybuanta jest dana wzorem (9.1).

[1] ang. *MISE – mean integrated squared error.*
[2] ang. *KDE – kernel density estimation.*
[3] ang. *empirical cumulative distribution function.*

Estymacja gęstości

Histogram

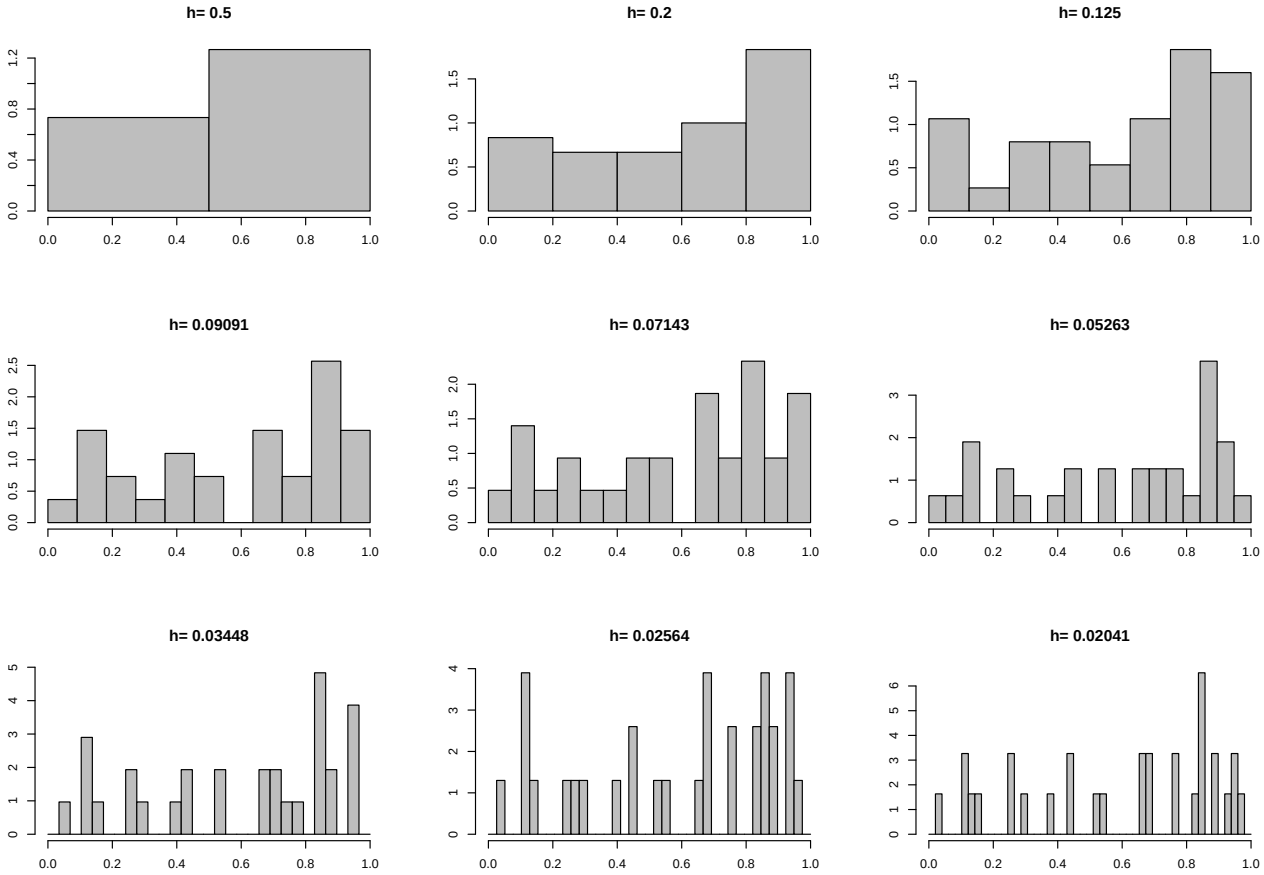
Definicja 9.2. Ustalmy próbę $(X_1, \ldots, X_n)$, $x_0 \in \mathbb{R}$ oraz $h > 0$ (zwane *szerokością klasy*, *szerokością pasma*). Niech

$$I_m = [x_0 + mh, x_0 + (m+1)h)$$

dla $m \in \mathbb{Z}$ oznacza *klasy*. Wtedy *histogramem* nazywamy funkcję $\hat{f}_n : \mathbb{R} \to \mathbb{R}$ daną wzorem

$$\hat{f}_n(x) = \frac{\#\{X_i \ : \ X_i \in I_m \ni x\}}{nh}.$$

Rysunek 9.1 pokazuje nam jak histogram zależy od parametru h. Jak widać histogram istotnie zależy od wartości tego parametru. Powstaje więc pytanie, jak dobrać wartość h w praktyce (aby był jak ‚najlepszy'). Teoretycznie odpowiada na to twierdzenie 9.3.



Rysunek 9.1: Histogramy dla tej samej próby dla różnych wartości szerokości klasy h.

Twierdzenie 9.3. *(Por. [10]). Rozważmy przestrzeń statystyczną*

$$\mathcal{P} = \left\{ P_f \ : \ f \in D, \ \int f^2 < \infty, \int (f')^2 < \infty \right\}.$$

Wtedy optymalną (w sensie scałkowanego błędu średniokwadratowego) szerokością pasma jest

$$h_0 = \frac{c}{\sqrt[3]{n}}, \quad gdzie \ \ c = \sqrt[3]{\frac{6}{\int (f')^2}}.$$

Twierdzenie to jest w praktyce nieprzydatne, ponieważ optymalne pasmo zależy od szukanej gęstości f. Gdy dodatkowo założymy, że szukana gęstość pochodzi z rozkładu normalnego $N(\mu, \sigma)$, to $c = 2\sqrt[3]{3}\sqrt[6]{\pi}\sigma \approx 3.486\sigma$, co dalej zależy od szukanej gęstości, więc w praktyce stosujemy wzór $h_0 = 3.486 s_X$ (gdzie s_X jest odchyleniem standardowym z próby)[4]. Istnieją też inne metody wyznaczania pasma, np. $h = 2IQR(X)/(n^{-1/3})$ (metoda Freedmana–Diaconisa, [10]).

[4] *Scott's normal rule.*

Prostota (obliczeniowa) jest zaletą histogramu, ale matematycznie jest wadą, ponieważ jako funkcja kawałkami stała jest nieciągła (a w praktyce często estymujemy gęstości klasy C^k).

W `R` do wyznaczenia histogramu można użyć funkcji `hist`.

Estymator jądrowy

Do zdefiniowania estymatora jądrowego potrzebujemy zdefiniować *jądro*[5].

Definicja 9.4. Funkcję $K : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ nazywamy *jądrem*, gdy spełnione są dwa warunki

- $\int_{-\infty}^{\infty} K(x)dx = 1$,
- $\forall x \in \mathbb{R}\ K(-x) = K(x)$.

Innymi słowy jądrem nazywamy parzystą gęstość. Najczęściej używanymi jądrami są:

- jądro gaussowskie K_g (gęstość standardowego rozkładu normalnego);
- jądro optymalne (Epanecznikowa)

$$K_e(t) = \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right)\chi_{[-\sqrt{5},\sqrt{5}]}(t);$$

- jądro prostokątne (gęstość rozkładu jednostajnego $\mathcal{U}[-1,1]$);
- jądro trójkątne

$$K(t) = (1 - |t|)\chi_{[-1,1]}(t).$$

Zdefiniujmy teraz estymator jądrowy.

Definicja 9.5. Ustalmy próbę $(X_1, \ldots, X_n)$, jądro K oraz $h > 0$ (zwane szerokością pasma lub parametrem wygładzającym[6]). Wtedy *estymatorem jądrowym* nazywamy funkcję $\hat{f}_n : \mathbb{R} \to \mathbb{R}$ daną wzorem

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

Rysunki 9.2, 9.3 oraz 9.4 ilustrują nam ideę jaka stoi za tą definicją oraz obrazuje wpływ parametru h na wartości estymatora jądrowego.

Powstaje problem, jakie jądro i szerokość pasma wybrać w praktyce. Poniższe twierdzenie odpowiada teoretycznie na to pytanie.

Twierdzenie 9.6. *Rozważmy przestrzeń statystyczną*

$$\mathcal{P} = \left\{P_f \ : \ f \in D,\ \int (f'')^2 < \infty\right\}.$$

Wtedy asymptotycznie (tj. dla $n \to \infty$) optymalnym jądrem (w sensie scałkowanego błędu średniokwadratowego) w klasie całkowalnych w kwadracie jąder jest jądro Epanecznikowa. Natomiast dla jądra K (całkowalnego w kwadracie) asymptotycznie optymalną szerokością pasma jest

$$h = \frac{c}{\sqrt[5]{n}},$$

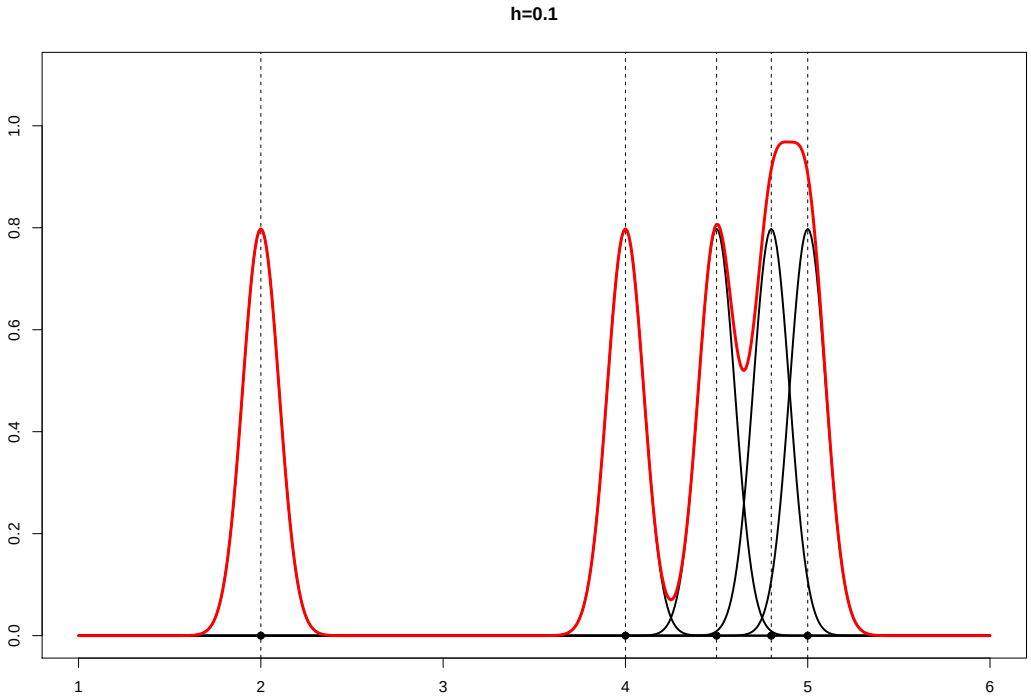
gdzie

$$c = \left(\int t^2 K(t)dt\right)^{-2/5}\left(\int K^2(t)dt\right)^{1/5}\left(\int (f''(t))^2 dt\right)^{-1/5}.$$

Ponownie to twierdzenie nie można bezpośrednio stosować w praktyce, ponieważ oszacowanie optymalnego pasma zależy od szukanej gęstości f. Gdy założymy, że f ma rozkład normalny $N(\mu, \sigma)$, to dla jądra gaussowskiego otrzymamy $h \approx 1.06\sigma n^{-1/5}$, a dla jądra Epanecznikowa $h \approx 1.05\sigma n^{-1/5}$. W praktyce okazuje się, że jądro gaussowskie i Epanecznikowa dają podobne rezultaty (zob. [10]).

[5] ang. *kernel.*

[6] ang. *bandwith, smoothing parameter.*



Rysunek 9.2: Estymator jądrowy dla próby $2, 4, 4.5, 4.8, 5$ dla jądra gaussowskiego i $h = 0.1$ (czerwony). Na czarno wykresy funkcji, których przeskalowana suma tworzy estymator jądrowy.

9.2 Metody komputerowe

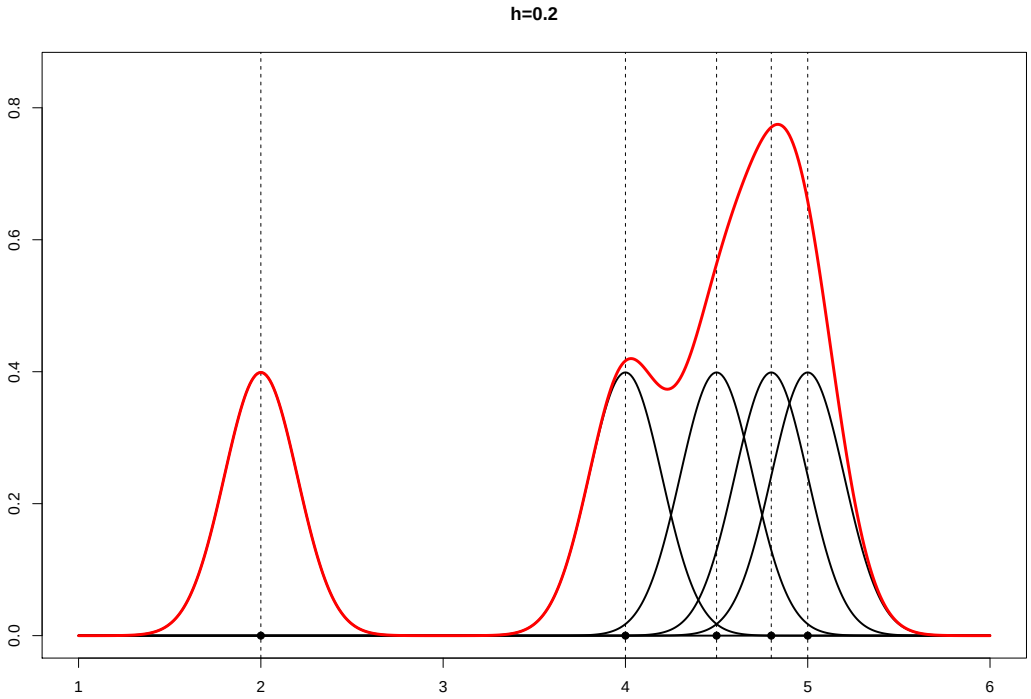
Przedstawimy kilka podstawowych metod komputerowych, które pozwalają rozwiązywać problemy, które są trudne dla klasycznego wnioskowania statystycznego, na przykład wyznaczania rozkładów funkcji testowych, czy wyznaczania przedziałów ufności dla małych prób lub bez szczegółowych założeń, czy mocy testów dla ustalonych hipotez alternatywnych. W kolejnych podrozdziałach omówimy symulacyjne wyznaczanie rozkładów, testy permutacyjne oraz metodę bootstrap.

Symulacyjne wyznaczanie rozkładów

Symulacyjne wyznaczenie rozkładu funkcji testowej może nam się przydać, gdy nie znamy jego dokładnego rozkładu, na przykład gdy znamy tylko rozkład asymptotyczny (bądź nie znamy go wcale). Zilustrujemy to standardowymi przykładami.

Przykład 9.7. Wyznaczymy symulacyjne rozkład statystyki testowej w teście niezależności χ^2 dla tablicy 2×2 przy zadanych rozkładach brzegowych i porównamy go do rozkładu $\chi^2(1)$. Zwróćmy uwagę, że dokładny rozkład statystyki testowej w tym teście jest zawsze dyskretny, więc trudno podejrzewać, że będzie on dobrze aproksymowany (dla małych liczności prób) przez rozkład ciągły $\chi^2(1)$.

```
WarStat=function(n=40,p=c(1,2,1.5),B=10000){
 k=length(p)
 S=numeric(B)
 for (i in 1:B){
 X=sample(1:k,n,prob=p,replace=TRUE)
 Y=sample(1:k,n,prob=p,replace=TRUE)
 X=as.factor(X)
 Y=as.factor(Y)
 names(X)=1:k
 names(Y)=1:k
 S[i]=chisq.test(table(X,Y))$statistic
 }
 return(S)
}

>W=WarStat(n=10,p=c(0.3,0.7),B=2000)
```



Rysunek 9.3: Estymator jądrowy dla próby $2, 4, 4.5, 4.8, 5$ dla jądra gaussowskiego i $h = 0.2$ (czerwony). Na czarno wykresy funkcji, których przeskalowana suma tworzy estymator jądrowy.

```
>plot(ecdf(W),xlab="",ylab="",main="")
>lines(sort(W),pchisq(sort(W),df=1),col="red")
```

Dla liczności próby $n = 10$ rezultat symulacji znajduje się na rysunku 9.5, a dla liczności próby $n = 1000$ na rysunku 9.6.

Jak widać dla małej liczności próby rozkład asymptotyczny $\chi^2(1)$ różni się od rozkładu wyznaczonego symulacyjnie.

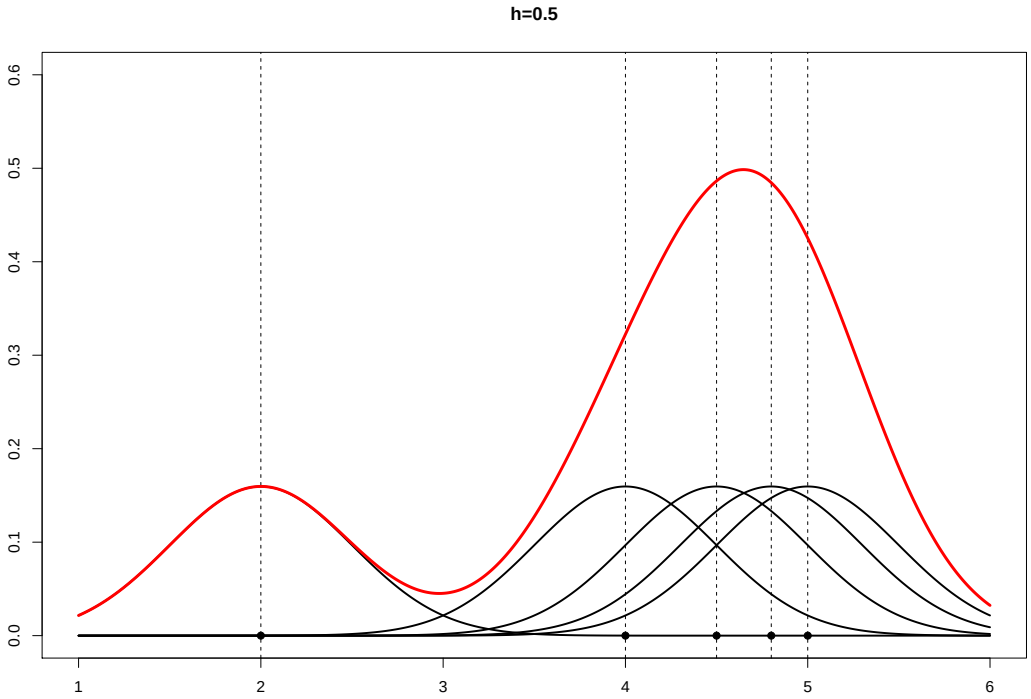
W kolejnym przykładzie wyznaczymy pvalue w teście zgodności χ^2, wykorzystując rozkład statystyki testowej wyznaczony symulacyjnie.

Przykład 9.8. Najpierw zaimplementujmy test zgodności χ^2 dla wektora wartości zaobserwowanych O i prawdopodobieństw teoretycznych p.

```
chisq.sym.test=function(O,p,B=1000){
  k=length(O)
  n=sum(O)
  S=numeric(B)
  for (i in 1:B){
    X=as.factor(sample(1:k,n,prob=p,replace=TRUE))
    levels(X)=1:k
    S[i]=chisq.test(table(X))$statistic
  }
  return(sum(S>=chisq.test(O)$statistic)/B) #zbiór krytyczny prawostronny
}
```

Zaimplementujmy teraz ten test i dla porównania ten sam test funkcją `chisq.test` z opcją `simulate.p.value = TRUE` dla mało licznej próby

```
> chisq.sym.test(O=c(3,8),p=c(0.5,0.5),B=2000)
[1] 0.227
> chisq.test(c(3,8),simulate.p.value = TRUE)

Chi-squared test for given probabilities with simulated p-value
(based on 2000 replicates)
```



Rysunek 9.4: Estymator jądrowy dla próby $2, 4, 4.5, 4.8, 5$ dla jądra gaussowskiego i $h = 0.5$ (czerwony). Na czarno wykresy funkcji, których przeskalowana suma tworzy estymator jądrowy.

```
data:  c(3, 8)
X-squared = 2.2727, df = NA, p-value = 0.2309
```

oraz dla dużej (już z wykorzystaniem rozkładu asymptotycznego)

```
> chisq.sym.test(O=c(80,60),p=c(0.5,0.5),B=2000)
[1] 0.109
> chisq.test(c(80,60))

Chi-squared test for given probabilities

data:  c(80, 60)
X-squared = 2.8571, df = 1, p-value = 0.09097
```

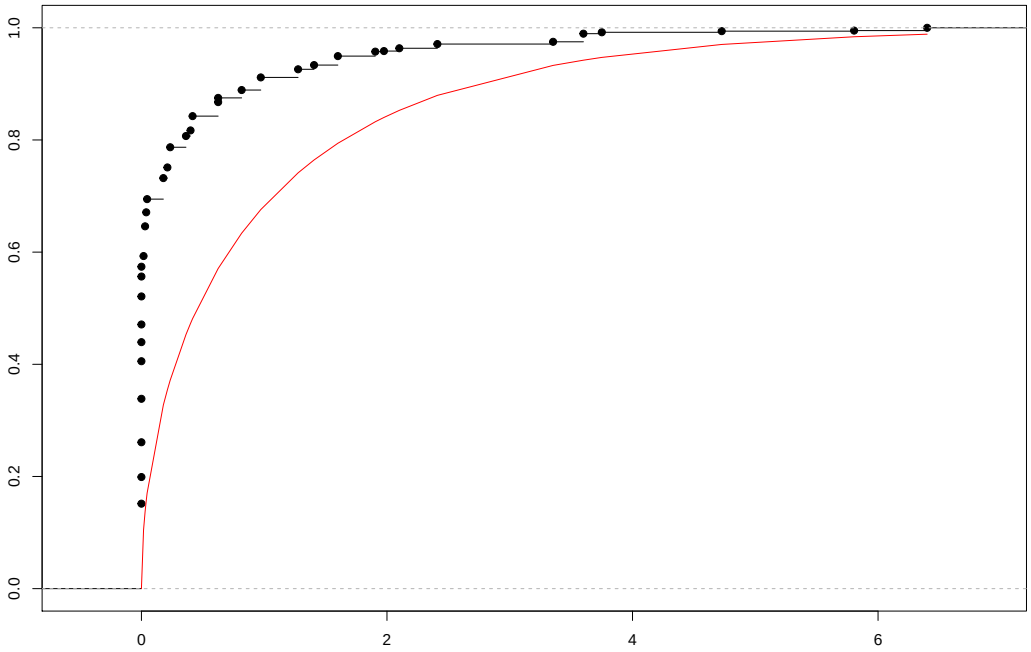
9.2.1 Testy permutacyjne

Ustalmy test statystyczny (i w szczególności formę danych). *Test permutacyjny* polega na wielokrotnej (odpowiedniej) permutacji danych i wyznaczeniu wartości statystyki testowej dla każdej z nich. Pvalue wyznaczamy jako frakcje tych permutacji, dla których statystyka testowa ‚bardziej' przeczyła hipotezie zerowej niż statystyka testowa dla danych oryginalnych (w szczególności dla zbioru krytycznego prawostronnego pvalue definiujemy jako frakcje permutacji, dla których statystyka testowa miała wartość większą lub równą statystyce testowej dla danych oryginalnych). Powstaje pytanie, ile permutacji trzeba wykonać. Najczęściej przyjmuje się, że dla poziomu istotności α należy wykonać co najmniej $50/\alpha$ permutacji.

Wielką zaletą testów permutacyjnych jest właściwie brak dodatkowych założeń, a wadą może być długi czas wykonania.

Podamy przykład testu, który porównuje dwie próby.

Przykład 9.9. Rozważmy test o statystyce testowej `stat`, który testuje hipotezę o dwóch próbach X i Y oraz ma zbiór krytyczny prawostronny (tzn. im większa wartość statystyki testowej tym ‚gorzej' dla hipotezy zerowej). Załóżmy, że próba X jest n elementowa, a Y m elementowa. Permutacja danych polega na tym, że próby X i Y łączymy w jedną próbę $n+m$ elementową, permutujemy ją i pierwsze n obserwacji traktujemy jako próbę $'X'$, a pozostałe jako próbę $'Y'$. W `R` będzie to wyglądać następująco.



Rysunek 9.5: Do przykładu 9.7: dla $n = 10$, na czarno dystrybuanta empiryczna rozkładu statystyki testowej, a na czerwono dystrybuanta rozkładu $\chi^2(1)$.

```
TestPermut=function(X,Y,stat,B=999){

  n=length(X)
  m=length(Y)
  Z=c(X,Y)
  S=stat(X,Y)
  Stat=numeric(B)
  for (i in 1:B){
    probka=Z[sample(m+n)]
    Stat[i]=stat(probka[1:n],probka[n+(1:m)])
  }
  print(1-ecdf(Stat)(S))          #pvalue z ecdf
  return((1+sum(Stat>=S))/(B+1)) #pvalue wprost z symulacji
}
```

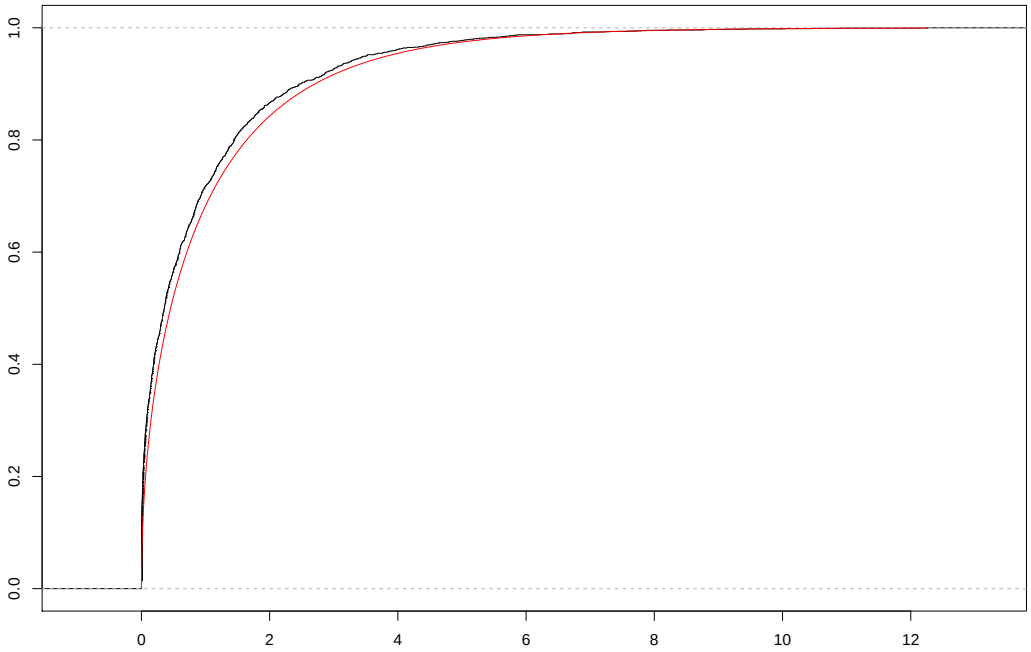
Zaimplementujmy test t o równości średnich i odpowiadający mu test permutacyjny.

```
> x=rnorm(40,0,1)
> y=rnorm(50,0,5)
> TestPermut(x,y,stat=function(x,y) t.test(x,y)$statistic,B=1001)
[1] 0.2647353
[1] 0.2654691
> t.test(x,y,alternative = "greater")$p.value
[1] 0.2569116
```

Jak widać wartości pvalue są podobne. Podkreślmy, że w teście permutacyjnym nie trzeba zakładać normalności cech, ani nic o liczności prób, więc możemy go implementować na przykład tak

```
> v1=rexp(10,0.5)
> v2=rpois(20,1)
> TestPermut(v1,v2,stat=function(x,y) t.test(x,y)$statistic,B=1001)
[1] 0.001998002
[1] 0.002994012
```

Test ten możemy zastosować też do ‚własnych' statystyk testowych (poniżej dla testu równości median), o rozkładach których ciężko cokolwiek powiedzieć.



Rysunek 9.6: Do Przykładu 9.7: dla $n = 1000$, na czarno dystrybuanta empiryczna rozkładu statystyki testowej, a na czerwono dystrybuanta rozkładu $\chi^2(1)$.

```
> TestPermut(x,y,B=2500,stat=function(x,y) abs(median(x)-median(y)))
[1] 0.2103159
```

Możemy też symulacyjnie wyznaczyć moc tego testu (na poziomie istotności 0.05) dla prób z rozkładu normalnego i rozkładu t, których mediny różnią się o 1. W tym przypadku to może już chwilę potrwać, bo wykonujemy w sumie milion permutacji.

```
> alpha=0.05
> B=1000
> S=numeric(B)
> fun=function(x,y) abs(median(x)-median(y))
> for (i in 1:B){
+   x=rnorm(30,0,1)
+   y=rt(20,df=2)+1
+   S[i]=TestPermut(x,y,B=1000,stat=fun)
+ }
> print(sum(S<alpha)/B)
[1] 0.773
```

9.2.2 Bootstrap

Rozważmy próbę $\underline{X} = (X_1, \ldots, X_n)$ (pobraną z pewnej populacji) oraz estymator T pewnego parametru θ. W sytuacji, gdy na przykład n jest małe nie jesteśmy w stanie określić z jakiego rozkładu pochodzi ta próba i w konsekwencji nie możemy wyznaczyć rozkładu estymatora T oraz skonstruować dokładnego przedziału ufności dla θ. Metoda bootstrapu pozwala rozwiązać ten problem.

Metoda ta polega na wygenerowaniu (poniżej zdefiniujemy trzy sposoby) replikacji oryginalnej próby $\underline{X}_1^\star, \ldots, \underline{X}_R^\star$ (z których każda składa się z n obserwacji). Następnie wyznaczamy wartość statystyki T na tych replikacjach: $t_1^\star = T(\underline{X}_1^\star), \ldots, t_R^\star = T(\underline{X}_R^\star)$. Metoda bootstrap opiera się na zasadzie, że $t^\star = (t_1^\star, \ldots, t_R^\star)$ pochodzi z rozkładu podobnego do rozkładu T na całej populacji. Możemy więc wnioskować o własnościach T na podstawie replikacji oraz próby $t^\star$. W szczególności możemy wnioskować na podstawie histogramu dla $t^\star$, zdefiniować *błąd standardowy* estymatora jako

$$SE_{t^\star} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(t_i^\star - \bar{t^\star})^2}.$$

Przedział ufności dla parametru θ na poziomie ufności $1-\alpha$ można zdefiniować na przykład następująco

- *percentylowy przedział ufności*
$$(q_{\frac{\alpha}{2}}(t^\star), q_{1-\frac{\alpha}{2}}(t^\star)),$$
gdzie $q_p(t^\star)$ to kwantyl z próby rzędu p;
- *normalny przedział ufności* (gdy $t^\star$ ma rozkład zbliżony do normalnego)
$$\left(\hat{t} - u(1-\frac{\alpha}{2})SE_{t^\star}, \hat{t} + u(1-\frac{\alpha}{2})SE_{t^\star}\right),$$
gdzie $\hat{t} = T(\underline{X})$.

Najważniejszymi sposobami generowania replikacji są

- *bootstrap nieparametryczny*

  Elementy replikacji $\underline{X}_i^\star$ pochodzą z losowania z rozkładu danego przez dystrybuantę empiryczną $\hat{F}_n$ próby $\underline{X}$, innymi słowy replikacja składa się z n elementów wylosowanych ze zwracaniem z $\underline{X}$. Metoda ta nie wymaga żadnych dodatkowych założeń.

- *bootstrap parametryczny*

  Zakładamy, że próba $\underline{X}$ pochodzi z ustalonej rodziny rozkładów, estymujemy jej parametry, a następnie z tego rozkładu losujemy elementy replikacji.

- *bootstrap wygładzający*

  Ta metoda zostanie zdefiniowana poniżej w osobnym paragrafie.

Przykład 9.10. Porównamy percentylowy przedział ufności z klasycznym przedziałem ufności dla średniej w rozkładzie normalnym.

```
> X=rnorm(50)
> l=qt(0.975,df=49)*sd(X)/sqrt(length(X))
> c(mean(X)-l,mean(X)+l)   #klasyczny przedzial ufnosci dla sredniej
[1] -0.3968296  0.2030779
>
> R=999    # Liczba probek bootstrapowych
> but=numeric(R)
> for (i in 1:R)  but[i]=mean(sample(X,replace=T))
>
> quantile(but,c(0.025,0.975)) #bootstrapowy przedzial nieparametryczny dla sredniej
      2.5%      97.5%
-0.4053695  0.1861921
```

Jak widać przedziały te są podobne.

9.2.2.1 Bootstrap wygładzający

W sytuacji, gdy próba $\underline{X}$ jest niewielka lub obserwacje się powtarzają, to możliwych wartości statystyki na replikacjach nieparametrycznych jest też niewiele, więc przedziały ufności będą trywialne. Zobaczmy to na przykładzie mediany.

```
> X=c(1,1,1,2,2,3,4)
> R=999    # Liczba probek bootstrapowych
> but=numeric(R)
> for (i in 1:R)  but[i]=median(sample(X,replace=T))
> quantile(but,c(0.025,0.975)) #bootstrapowy przedzial nieparametryczny dla mediany
 2.5% 97.5%
    1     3
```

Aby temu zapobiec, do każdego elementu replikacji dodaje się ‚lekki' szum, najczęściej liczbę wylosowaną z rozkładu normalnego $N(0,\sigma)$. Nosi to nazwę *bootstrapu wygładzającego*.

```
> n=length(X)
> but_wygl=numeric(R)
> for (i in 1:R)  but_wygl[i]=median(sample(X,replace=T)+rnorm(n,0,1))
> quantile(but_wygl,c(0.025,0.975))
     2.5%     97.5%
0.6337888 3.2867821
```

9.2.2.2 Testowanie hipotez i bootstrap

Metodę bootstrap można też użyć do wyznaczanie rozkładu statystyki testowej. Dla przykładu rozważmy klasyczny test t dla średniej z rozkładu normalnego. W sytuacji, gdy próba pochodzi z tego rozkładu, możemy użyć dokładnego rozkładu statystyki testowej. Gdy nie wiemy, czy to założenie jest spełnione, możemy użyć metody bootstrap.

```
boot.t.test=function(X,mu=0,R=999,alfa=0.05){

T=function(X,m0=0) (mean(X)-m0)*sqrt(length(X))/sd(X)

Tbut=numeric(R)
mbut=numeric(R)
for (i in 1:R){
  Xb=sample(X,replace=TRUE)
  mbut[i]=mean(Xb)
  Tbut[i]=T(Xb,m0=mean(X))
}

k=ecdf(Tbut)(T(X,m0=mu))
p_value=1-2*abs(k-0.5)
print(paste("p.value=",p_value))

print(paste(1-alfa,"-przedzial ufnosci dla sredniej: (",
            quantile(mbut,alfa/2),",",quantile(mbut,1-alfa/2),")",sep=""))
}
```

Najpierw porównajmy ten test z dokładnym testem t, gdy próba pochodzi z rozkładu normalnego.

```
> boot.t.test(X,5)
[1] "p.value= 0.566566566566567"
[1] "0.95-przedzial ufnosci dla sredniej: (4.05187484385895,5.4934240931606)"
> t.test(X,mu=5)

One Sample t-test

data:  X
t = -0.524, df = 19, p-value = 0.6063
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.036303 5.577780
sample estimates:
mean of x
 4.807041
```

Jak widać testy te dały podobne wyniki. Ale oczywiście test ‚bootstrapowy' możemy użyć dla dowolnej próby.

```
> X=rexp(20,rate=0.5)
> boot.t.test(X,3)
[1] "p.value= 0.122122122122122"
[1] "0.95-przedzial ufnosci dla sredniej: (1.00872123219612,2.62442261544203)"
> boot.t.test(X,2)
[1] "p.value= 0.592592592592593"
[1] "0.95-przedzial ufnosci dla sredniej: (0.98193164894082,2.60071261374941)"
```

Rozdział 10

Wstęp do metod bayesowskich

10.1 Rozkład Beta

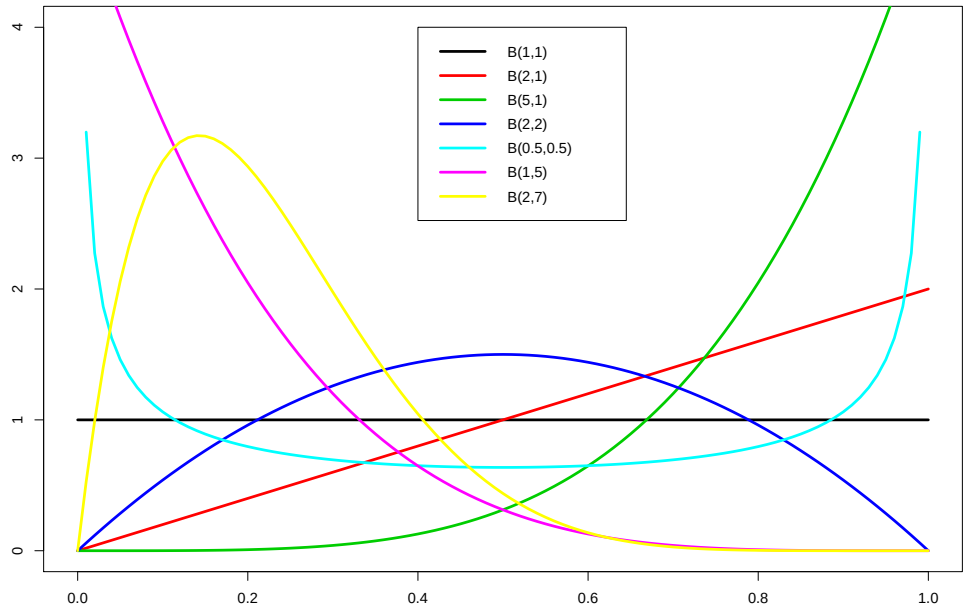
Zdefiniujemy rozkład Beta oraz podamy jego podstawowe własności.

Definicja 10.1. *Rozkład Beta o parametrach* α, β (ozn. $\mathcal{B}(\alpha, \beta)$) jest dany gęstością

$$f_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\chi_{[0,1]}(x).$$

Przez $B(p, \alpha, \beta)$ oznaczać będziemy kwantyl rzędu p rozkładu $\mathcal{B}(\alpha, \beta)$.

Rozkład $\mathcal{B}(\alpha, \beta)$ służy głównie do modelowania parametrów z odcinka jednostkowego (najczęściej frakcję). Jego zaletą jest to, że dla różnych parametrów α, β dostajemy rozkłady o różnych własnościach (symetryczne, skośne, „U"-kształtne, etc). W szczególności rozkład $\mathcal{B}(1,1)$ to rozkład jednostajny $\mathcal{U}(0,1)$, $\mathcal{B}(2,1)$ to rozkład trójkątny, etc. Inne przykłady przedstawia rysunek 10.1.



Rysunek 10.1: Gęstości rozkładu $\mathcal{B}(\alpha, \beta)$ dla różnych parametrów.

Podstawowe własności rozkładu Beta przedstawiają poniższe twierdzenia.

Twierdzenie 10.2. *Niech* $X \sim \mathcal{B}(\alpha, \beta)$. *Wtedy*

1. $E(X) = \frac{\alpha}{\alpha+\beta}$;
2. $D^2(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$;
3. $\operatorname{argmax}(f_{\alpha,\beta}) = \frac{\alpha-1}{\alpha+\beta-2}$, *dla* $\alpha, \beta > 1$.

Twierdzenie 10.3. *Rozkład Beta ma następujące własności.*

1. *Jeśli* $X \sim \mathcal{B}(\alpha, \beta)$, *to* $1 - X \sim \mathcal{B}(\beta, \alpha)$.
2. *Jeśli* $X \sim \mathcal{B}(n/2, m/2)$, *to* $\frac{mX}{n(1-X)} \sim F(n, m)$.
3. *Jeśli* $X \sim \mathcal{B}(\alpha, 1)$, *to* $-\log(X) \sim Exp(\alpha)$.
4. *Jeśli zmienne losowe* $X \sim \chi^2(\alpha)$ *i* $Y \sim \chi^2(\beta)$ *są niezależne, to* $\frac{X}{X+Y} \sim \mathcal{B}(\alpha/2, \beta/2)$.
5. *Jeśli* $X \sim \mathcal{U}(0, 1)$ *i* $\alpha > 0$, *to* $X^{1/\alpha} \sim \mathcal{B}(\alpha, 1)$.

10.2 Wnioskowanie bayesowskie

Wnioskowanie bayesowskie to wnioskowanie statystyczne wykorzystujące wzór Bayesa i operujące pojęciem *prawdopodobieństwa subiektywnego*. Nie będziemy tu wchodzić w dyskusję (chyba już nawet nie do końca matematyczną, tylko filozoficzną, metodologiczną), która interpretacja prawdopodobieństwa, subiektywna, czy obiektywna (‚częstotliwościowa') jest właściwa[1], niemniej zwróćmy uwagę, że jak dotąd interpretowaliśmy metody statystyczne ‚częstotliwościowo' (np. poziom ufności przedziału ufności to była częstość zdarzenia, że przedział ufności zawiera badany parametr, przy dużej liczbie powtórzeń). Jednak jak zobaczymy, wnioskowanie bayesowskie możemy traktować jako uogólnienie poznanych dotąd metod. Przejdźmy do szczegółów.

Podobnie jak wcześniej zakładamy, że badana zmienna (cecha) ma rozkład P_θ pochodzący z ustalonej rodziny rozkładów $\{P_\theta : \theta \in \Theta\}$. We wnioskowaniu bayesowskim dodajemy dodatkową strukturę na przestrzeń parametrów Θ, tj. rozważamy rozkład prawdopodobieństwa na Θ nazywany rozkładem *a priori*, który wyraża nasze opinie **przed** pobraniem próby (subiektywne lub wynikające z naszej wiedzy o badanej zmiennej) o tym, jakie jest prawdopodobieństwo, że parametr θ jest tym prawdziwym (tj. takim, że $P_X = P_\theta$). Na przykład przeprowadzamy ankietę w celu określenia frakcji p poparcia partii A. Przed badaniem stwierdzamy, że to poparcie na pewno nie jest większe niż 0.6 (bo w ostatnich 30-tu latach żadna partia nie miała takiego poparcia lub z jakiegokolwiek innego powodu), a pozostałe frakcje są równo prawdopodobne. Jako rozkład *a priori* weźmiemy więc rozkład jednostajny na odcinku $[0, 0.6]$. Jeśli nie mamy żadnych opinii o parametrze θ, możemy zawsze rozważyć *nieinformacyjny* rozkład *a priori* (np. jednostajny na Θ, jeśli taki ma sens).

Oznaczmy teraz przez $\underline{X} = (X_1, \ldots, X_n)$ pobraną próbę prostą. Wtedy wykorzystując wzór Bayesa otrzymujemy wzór na rozkład *a posteriori*

$$P(\theta|\underline{X}) = \frac{P(\underline{X}|\theta)P(\theta)}{P(\underline{X})}, \tag{10.1}$$

to jest rozkład parametru θ pod warunkiem wylosowania próby $\underline{X}$. W liczniku mamy wiarygodność próby $P(\underline{X}|\theta)$ oraz rozkład *a priori* $P(\theta)$, a w mianowniku prawdopodobieństwo wylosowania próby $\underline{X}$ (przy rozkładzie *a priori*). Ponieważ mianownik to liczba (która normalizuje licznik) można skrótowo powiedzieć, że rozkład *a posteriori* jest proporcjonalny do iloczynu wiarygodności próby i rozkładu *a priori* ($P(\theta|\underline{X}) \propto P(\underline{X}|\theta)P(\theta)$). Co więcej w praktyce często trudno wyliczyć mianownik, więc podaje się tylko licznik (jak zobaczymy nie zawsze to przeszkadza w dalszej analizie).

Do wnioskowania używamy teraz rozkładu *a posteriori*.

Bayesowskim estymatorem punktowym[2] nazywamy te parametry θ, które są najbardziej prawdopodobne *a posteriori*. Bardziej precyzyjnie, gdy rozkład *a posteriori* jest dyskretny (oznaczmy go przez π_1), to

$$\hat{\theta}_B = \operatorname{argmax}_{\theta \in \Theta}(\pi_1(\theta)),$$

a gdy ciągły zadany przez gęstość f_1, to

$$\hat{\theta}_B = \operatorname{argmax}_{\theta \in \Theta}(f_1(\theta)).$$

Odpowiednikiem ‚częstotliwościowych' przedziałów ufności[3] są we wnioskowaniu bayesowskim *przedziały wiarygodności*[4]. Można je zdefiniować na różne sposoby, przedstawimy dwie możliwości.

Intuicyjnie $a\%$ *przedział HDI*[5] zawiera $a\%$ parametrów o największych prawdopodobieństwach. Bardziej formalnie

[1] Standardowym przykładem różnicy między tymi pojęciami może być pytanie jakie jest prawdopodobieństwo, że Polska wygra najbliższe Mistrzostwa Europy w piłce nożnej? Zwolennik podejścia obiektywnego powie, że nie da się tego określić obiektywnie, bo te mistrzostwa są jedyne (nie da się ich powtórzyć w identycznych warunkach), natomiast subiektywnie można odpowiedzieć np. 0.1 (na podstawie swojej intuicji, opinii o drużynach, interpretacji danych historycznych, etc.).

[2] ang. *MAP estimator (a maximum* a posteriori *probability estimator).*

[3] ang. *Confidence Intervals*

[4] ang. *Credible Intervals*

[5] ang. *Highest Density Interval*

(np. dla przypadku, gdy rozkład *a posteriori* jest zadany gęstością f_1) to zbiór

$$\{\theta \ : \ f_1(\theta) > w \ : \ \int_{\{\theta: f_1(\theta)>w\}} f_1(\theta)d\theta = a/100\}.$$

Natomiast $(1-a)\%$ *przedział ETI*[6] jest równy $(q(a/2), q(1-a/2))$, gdzie $q(p)$ jest kwantylem rzędu p rozkładu *a posteriori*.

Wprost z definicji wynika, że punktowy estymator bayesowski zawsze należy do przedziału HDI, natomiast nie musi należeć do przedziału ETI. Przedział ETI z definicji jest przedziałem, a przedział HDI nie musi być przedziałem (być spójnym) na przykład dla odpowiedniego rozkładu dwumodalnego.

Zakończymy ten wstęp dwoma uwagami. Pierwszą rozpoczniemy od następującego twierdzenia.

Twierdzenie 10.4. *Niech $\underline{X}, \underline{Y}$ są próbami niezależnymi. Niech $R(\pi, \underline{X})$ oznacza rozkład* a posteriori *dla rozkładu* a priori *π i próby $\underline{X}$. Wtedy*

$$R(\pi_0, (\underline{X}, \underline{Y})) = R(R(\pi_0, \underline{X}), \underline{Y}).$$

Twierdzenie to jest przydatne w praktyce, gdy wyznaczenie rozkładu *a posteriori* dla próby n-elementowej jest zbyt trudne rachunkowo. Wtedy można równoważnie n razy przeprowadzić analizę na próbie jednoelementowej. Często więc problemy są rozważane dla próby $\underline{X} = (X_1)$.

Druga uwaga dotyczy ogólnego wzoru (10.1). Jak często bywa w rachunku prawdopodobieństwa wzory ogólne operujące dowolnym rozkładem są mało praktyczne. Zazwyczaj rozważa się rozkłady dyskretne albo ciągłe i podamy teraz cztery szczególne przypadki tego wzoru.

10.2.1 Rozkłady próby i *a priori* dyskretne

Niech rozkład *a priori* będzie dyskretny i oznaczmy go przez π_0. Załóżmy także, że rozkład cechy X też jest dyskretny i oznaczmy go przez P_θ (czyli $P_\theta(x) = P_\theta(X = x)$). Wtedy rozkład *a posteriori* π_1 dany jest wzorem

$$\pi_1(\theta) = \frac{P_\theta(\underline{X})\pi_0(\theta)}{\sum_{\theta\in\Theta} P_\theta(\underline{X})\pi_0(\theta)},$$

gdzie $P_\theta(\underline{X}) = P_\theta(X_1)\cdots P_\theta(X_n)$. Widzimy, że gdy $\pi_0(\theta) = 0$, to i $\pi_1(\theta) = 0$. Stąd wniosek, że rozkład *a posteriori* też jest dyskretny.

Przykład 10.5. (Ankieta). Badamy zmienną X o rozkładzie zerojedynkowym z parametrem $p \in [0,1]$ (skrótowo $X|p \sim (1,0,p)$). Rozważmy dwupunktowy rozkład *a priori* $\pi_0(0.1) = \pi_0(0.2) = 0.5$. Załóżmy, że o próbie prostej $\underline{X} = (X_1, \ldots, X_n)$. Wiemy tylko ile jedynek w niej wystąpiło. Wtedy można równoważnie zapisać ją jako $\underline{X} = (n, k)$, gdzie k to liczba jedynek, i $P_p((n,k)) = \binom{n}{k}p^k(1-p)^{n-k}$. Dla prostoty załóżmy, że $\underline{X} = (3,2)$.

Wyznaczamy teraz rozkład *a posteriori*:

$$\pi_1(0.1) = \frac{\binom{3}{2}0.1^2 0.9^1\ 0.5}{\binom{3}{2}0.1^2 0.9^1\ 0.5 + \binom{3}{2}0.2^2 0.8^1\ 0.5} = \ldots = \frac{9}{41}.$$

oraz analogicznie

$$\pi_1(0.2) = \frac{\binom{3}{2}0.2^2 0.8^1\ 0.5}{\binom{3}{2}0.1^2 0.9^1\ 0.5 + \binom{3}{2}0.2^2 0.8^1\ 0.5} = \ldots = \frac{32}{41}.$$

Bayesowską estymacją parametru p będzie więc $\hat{p}_B = 0.2$.

10.2.2 Rozkład próby dyskretny, a *a priori* ciągły

Niech rozkład *a priori* będzie ciągły, a jego gęstość oznaczmy przez f_0. Załóżmy także, że rozkład cechy X jest dyskretny i oznaczmy go przez P_θ. Wtedy rozkład *a posteriori* jest ciągły, którego gęstość f_1 dana jest wzorem

$$f_1(\theta) = \frac{P_\theta(\underline{X})f_0(\theta)}{\int_{\theta\in\Theta} P_\theta(\underline{X})f_0(\theta)d\theta}.$$

[6] ang. *Equal-Tailed Interval*

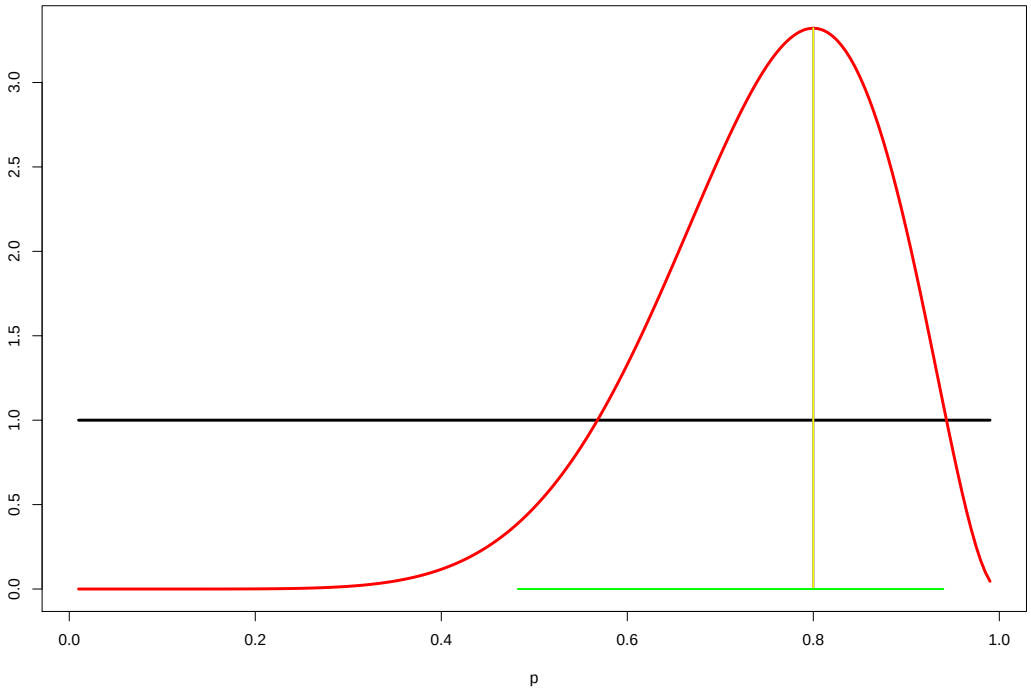
Przykład 10.6. (Ankieta). Tak jak w przykładzie (10.5) załóżmy, że $\underline{X}|p \sim b(n,p)$, natomiast rozkład *a priori* jest rozkładem $\mathcal{B}(\alpha,\beta)$. Wtedy

$$f_1(p) = \frac{\binom{n}{k}p^k(1-p)^{n-k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\chi_{[0,1]}(p)}{\int_0^1 \binom{n}{k}p^k(1-p)^{n-k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}dp}$$
$$= \frac{p^{\alpha+k-1}(1-p)^{\beta+n-k-1}\chi_{[0,1]}(p)}{\int_0^1 p^k(1-p)^{n-k}p^{\alpha-1}(1-p)^{\beta-1}dp}.$$

Nie musimy wyznaczać dokładnej wartości mianownika, ponieważ widać, że f_1 jest gęstością rozkładu $\mathcal{B}(\alpha+k,\beta+n-k)$ (w takim przypadku mówimy, że rozkład Beta jest rozkładem *a priori sprzężonym* dla rozkładu dwumianowego). W szczególności dla nieinformacyjnego rozkładu *a priori* $\mathcal{B}(1,1)$ uzyskujemy rozkład *a posteriori* $\mathcal{B}(1+k,1+n-k)$. Zgodnie z twierdzeniem 10.2 uzyskujemy

$$\hat{p}_B = \operatorname{argmax}(f_{1+k,1+n-k}) = \frac{1+k-1}{1+k+1+n-k-2} = \frac{k}{n}.$$

Jak widzimy w tym przypadku estymator bayesowski jest równy estymatorowi największej wiarygodności. Ilustruje to rysunek 10.2.



Rysunek 10.2: Gęstość rozkładu *a priori* $\mathcal{B}(1,1)$ (czarny), gęstość rozkładu *a posteriori* dla próby $n = 10$, $k = 8$, $\mathcal{B}(9,3)$ (czerwony), wartość estymatora bayesowskiego i największej wiarygodności (żółty), 95% przedział ETI (zielony).

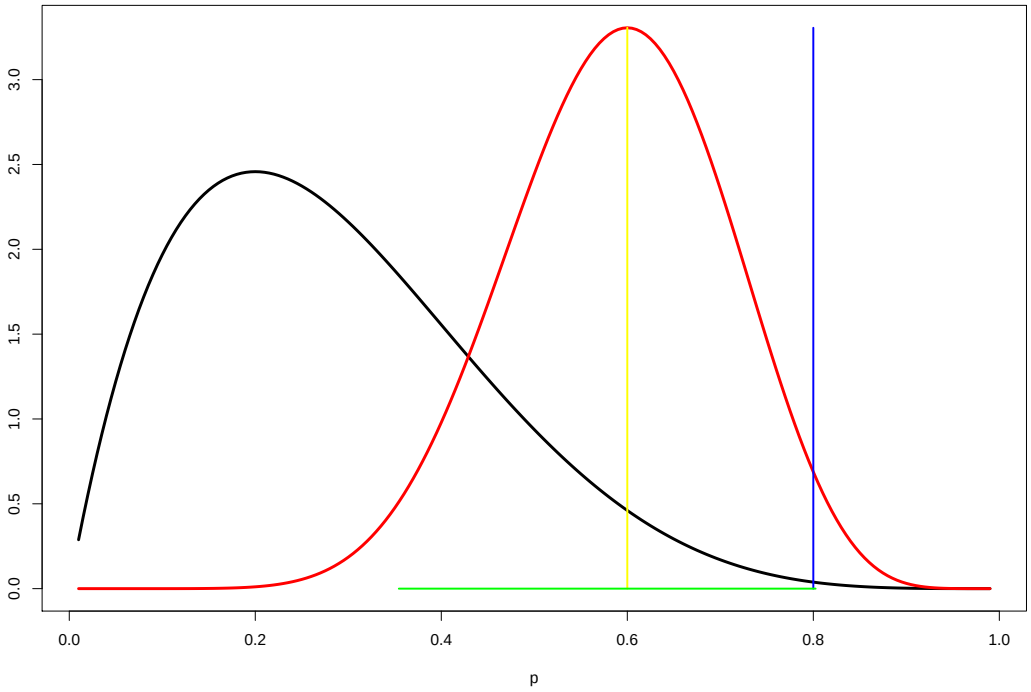
Dla innego rozkładu *a priori* ta równość oczywiście nie musi zachodzić, co można zobaczyć na Rysunku 10.3.

10.2.3 Rozkłady próby i *a priori* ciągłe

W sytuacji, gdy zarówno rozkład próby jak i rozkład *a priori* jest ciągły, rozkład *a posteriori* też jest ciągły, a jego gęstość dana jest wzorem

$$f_1(\theta) = \frac{f_\theta(\underline{X})f_0(\theta)}{\int_{\theta\in\Theta} f_\theta(\underline{X})f_0(\theta)d\theta},$$

gdzie f_0 jest gęstością rozkładu *a priori*.



Rysunek 10.3: Gęstość rozkładu *a priori* $\mathcal{B}(2,5)$ (czarny), gęstość rozkładu *a posteriori* dla próby $n = 10$, $k = 8$, $\mathcal{B}(10,7)$ (czerwony), wartość estymatora bayesowskiego (żółty), wartość estymatora największej wiarygodności (niebieski), 95% przedział ETI (zielony).

10.2.4 Rozkład próby ciągły, a *a priori* dyskretny

W sytuacji, gdy rozkład próby jest ciągły, a rozkład *a priori* jest dyskretny, rozkład *a posteriori* jest dyskretny, a jego rozkład dany jest wzorem

$$\pi_1(\theta) = \frac{f_\theta(\underline{X})\pi_0(\theta)}{\sum_{\theta\in\Theta} f_\theta(\underline{X})\pi_0(\theta)}.$$

Rozdział 11

Wielowymiarowy rozkład normalny

W tym rozdziale zdefiniujemy wielowymiarowy rozkład normalny i podamy jego podstawowe własności. Wcześniej musimy jednak przypomnieć kilka podstawowych pojęć algebry liniowej.

11.1 Macierze dodatnio i nieujemnie określone

Rozważmy przestrzeń wektorową $\mathbb{R}^n$. Niech $x, y \in \mathbb{R}^n$. Przez $< x, y >= x^T y$ oznaczmy standardowy iloczyn skalarny, przez $M(n,m)$ macierze $n \times m$, a przez $M(r)$ macierze kwadratowe $r \times r$. Powiemy, że macierz A jest *symetryczna*, gdy $A = A^T$.

Definicja 11.1. Macierz symetryczną A nazywamy

- *nieujemnie określoną* (ozn. $A \geqslant 0$), gdy $x^T Ax \geqslant 0$ dla każdego $x \in \mathbb{R}^n$;
- *dodatnio określoną* (ozn. $A > 0$), gdy $A \geqslant 0$ oraz $x^T Ax = 0$ wtedy i tylko wtedy, gdy $x = 0$.

Stwierdzenie 11.2. *Niech $X \in M(n,m)$. Niech $A = X^T X$. Wtedy*

- *$A \geqslant 0$;*
- *$A > 0$ wtedy i tylko wtedy, gdy X jest monomorfizmem.*

Dowód. Ustalmy $x \in \mathbb{R}^n$. Wtedy

$$x^T Ax = x^T X^T Xx = (Xx)^T Xx =< Xx, Xx >= \|Xx\|^2 \geqslant 0.$$

Druga teza wynika z własności, że X jest monomorfizmem wtedy i tylko wtedy, gdy $Xx = 0$ wtedy i tylko wtedy, gdy $x = 0$ oraz $< x, x >= 0$ wtedy i tylko wtedy, gdy $x = 0$. □

Macierze symetryczne mają też prostą, rzeczywistą postać macierzy Jordana.

Twierdzenie 11.3. *Niech $A = A^T \in M(r)$. Wtedy wszystkie wartości własne A są rzeczywiste oraz istnieje macierz ortonormalna $P \in M(r)$ (to znaczy taka, że $P^T P = PP^T = I$) taka, że $A = PJP^T$, gdzie*

$$J = \begin{bmatrix} z_1 & & 0 \\ & \ddots & \\ 0 & & z_r \end{bmatrix},$$

gdzie $z_1, \ldots, z_r \in \sigma(A)$.

Wykorzystując powyższe twierdzenie uzyskujemy następujące własności macierzy nieujemnie i dodatnio określonych.

Wniosek 11.4. *Niech A będzie macierzą symetryczną. Wtedy*

1. *$A \geqslant 0$ wtedy i tylko wtedy, gdy $\lambda \geqslant 0$ dla każdej wartości własnej $\lambda \in \sigma(A)$;*
2. *$A > 0$ wtedy i tylko wtedy, gdy $\lambda > 0$ dla każdej wartości własnej $\lambda \in \sigma(A)$;*
3. *Jeśli $A \geqslant 0$, to $A > 0$ wtedy i tylko wtedy, gdy A jest nieosobliwa;*
4. *Jeśli $A \geqslant 0$, to istnieje dokładnie jedna macierz B taka, że $B \geqslant 0$ i $B^2 = A$ (wtedy definiujemy $A^{1/2} := B$);*
5. *$A > 0$ wtedy i tylko wtedy, gdy $A^{1/2} > 0$.*

11.2 Wektory losowe

Definicja 11.5. Niech Y będzie n-wymiarowym wektorem losowym (to znaczy $Y : \Omega \to \mathbb{R}^n$ jest mierzalna i $Y(\omega) = (Y_1(\omega), \ldots, Y_n(\omega))^T$). Wtedy *wartością oczekiwaną* Y nazywamy

$$E(Y) = \begin{bmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{bmatrix} = \mu \in \mathbb{R}^n,$$

natomiast *macierzą kowariancji (wariancji)* nazywamy macierz $\Sigma(Y) = cov(Y) \in M(n)$ daną wzorem

$$\Sigma(Y) = [E((Y_i - \mu_i)(Y_j - \mu_j))]_{i,j=1,\ldots,n}.$$

Wprost z definicji wynika, że macierz kowariancji jest zawsze symetryczna.

W trzech poniższych twierdzeniach sformułowane są podstawowe własności wartości oczekiwanej i macierzy kowariancji dowolnego wektora losowego.

Twierdzenie 11.6. *Niech $Y : \Omega \to \mathbb{R}^n$, $A \in M(m,n)$, $b \in \mathbb{R}^m$, $\mu = E(Y)$, $\Sigma = \Sigma(Y)$ i $Z = AY + b$. Wtedy*

1. $E(Z) = A\mu + b$;
2. $\Sigma(Z) = A\Sigma A^T$;
3. $E(\|Y\|^2) = \|\mu\|^2 + tr(\Sigma)$.

Dowód. Dowód punktów (1) i (2) wynika z liniowości operatora E. Dla punktu (3) mamy

$$E(\|Y\|^2) = E\left(\sum_{i=1}^n Y_i^2\right) = \sum_{i=1}^n E(Y_i^2) = \sum_{i=1}^n D^2(Y_i) + \sum_{i=1}^n E(Y_i)^2 = tr(\Sigma) + \|\mu\|^2.$$

□

Twierdzenie 11.7. *Niech $Y : \Omega \to \mathbb{R}^n$ będzie wektorem losowym. Wtedy $\Sigma = cov(Y) \geqslant 0$.*

Dowód. Ustalmy $a \in \mathbb{R}^n$ i zdefiniujmy zmienną losową

$$Z = a_1 Y_1 + \ldots + a_n Y_n = a^T Y.$$

Z twierdzenia 11.6 mamy

$$a^T \Sigma a = cov(Z) = D^2(Z) \geqslant 0.$$

□

Twierdzenie 11.8. *Załóżmy, że wektor losowy $Y : \Omega \to \mathbb{R}^n$ ma rozkład ciągły dany gęstością f. Wtedy $\Sigma = cov(Y) > 0$.*

Dowód. Załóżmy do dowodu nie wprost, że istnieje $0 \neq a \in \mathbb{R}^n$ takie, że $a^T \Sigma a = 0$. Zdefiniujmy $Z = a^T Y$. Wtedy $D^2(Z) = a^T \Sigma a = 0$, więc istnieje c takie, że $P(Z = c) = 1$. Ale z drugiej strony mamy

$$1 = P(Z = c) = P(a^T Y = c) = \int_{\{y \in \mathbb{R}^n : a^T y = c\}} f(x) dx_1 \ldots dx_n = 0.$$

Ostatnia równość wynika z faktu, że całkujemy po zbiorze będącym podzbiorem przestrzeni co najwyżej $n-1$ wymiarowej, która ma n-wymiarową miarę Lebesgue'a równą zero. Otrzymaliśmy sprzeczność. □

11.3 Funkcja generująca momenty

Definicja 11.9. Niech $Y : \Omega \to \mathbb{R}^n$ będzie wektorem losowym. Definiujemy wtedy *funkcję generującą momenty* $M_Y : \mathbb{R}^n \to \mathbb{R}$ wzorem

$$M_Y(t) = E\left(e^{t^T Y}\right) = E\left(e^{t_1 Y_1 + \cdots t_n Y_n}\right).$$

Oczywiście, gdy wektory losowe Y i Z mają ten sam rozkład, to mają te same funkcje tworzące momenty. O implikacji w drugą stronę mówi poniższe twierdzenie.

Twierdzenie 11.10. *Niech Y i Z będą n-wymiarowymi wektorami losowymi. Jeśli istnieje niepusty zbiór otwarty U w $\mathbb{R}^n$ taki, że $M_Y(t) = M_Z(t)$ dla każdego $t \in U$ (w szczególności te funkcje są określone na U), to rozkłady Y i Z są takie same.*

Potrzebne nam będzie także takie twierdzenie.

Twierdzenie 11.11. *Niech $Y : \Omega \to \mathbb{R}^n$, $A \in M(m,n)$, $b \in \mathbb{R}^m$ i $Z = AY + b$. Wtedy*

1. $M_Z(t) = e^{b^T t} M_Y(A^T t)$;
2. *Oznaczmy $Y = (Y_1, Y_2)^T$, gdzie $Y_1 \in \mathbb{R}^d$, $Y_2 \in \mathbb{R}^{n-d}$. Jeśli wektory losowe Y_1 i Y_2 są niezależne to*

$$M_Y\left(\begin{bmatrix} t_1 \\ t_2 \end{bmatrix}\right) = M_Y\left(\begin{bmatrix} t_1 \\ 0 \end{bmatrix}\right) M_Y\left(\begin{bmatrix} 0 \\ t_2 \end{bmatrix}\right) = M_{Y_1}(t_1) M_{Y_2}(t_2).$$

Przejdźmy teraz do definicji wielowymiarowego rozkładu normalnego.

11.4 Wielowymiarowy rozkład normalny

Najpierw w trzech krokach przedstawimy ideę, która wyjaśni nam (na końcu) definicję wielowymiarowego rozkładu normalnego.

1. Dla jednowymiarowej zmiennej losowej $Z \sim N(0,1)$ funkcja generująca momenty wynosi $M_Z(t) = e^{\frac{1}{2}t^2}$.
2. Załóżmy teraz, że $Z_1, \ldots, Z_n$ to ciąg niezależnych zmiennych losowych o rozkładach $N(0,1)$. Utwórzmy wektor losowy $Z = (Z_1, \ldots, Z_n)^T$. Wtedy z twierdzenia 11.11 mamy

$$M_Z(t) = \prod_{i=1}^{n} M_{Z_i}(t_i) = \prod_{i=1}^{n} e^{\frac{1}{2}t_i^2} = e^{\frac{1}{2}(t_1^2 + \cdots + t_n^2)} = e^{\frac{1}{2}t^T t} = e^{\frac{1}{2}\|t\|^2}.$$

3. Weźmy Z z wcześniejszego punku oraz ustalmy $A \in M(m,n)$, $\mu \in \mathbb{R}^m$ oraz niech $Y = AZ + \mu$. Wtedy ponownie wykorzystując twierdzenie 11.11, otrzymujemy

$$\begin{aligned} M_Y(t) &= e^{\mu^T t} M_Z(A^T t) = e^{\mu^T t} e^{\frac{1}{2}\|A^T t\|^2} = e^{\mu^T t} e^{\frac{1}{2}<A^T t, A^T t>} \\ &= e^{\mu^T t} e^{\frac{1}{2}(A^T t)^T A^T t} = e^{\mu^T t} e^{\frac{1}{2}t^T A A^T t} = e^{\mu^T t + \frac{1}{2}t^T A A^T t}. \end{aligned}$$

Przyjmując oznaczenie $\Sigma = AA^T$, otrzymujemy ostatecznie

$$M_Y(t) = e^{\mu^T t + \frac{1}{2}t^T \Sigma t}.$$

Postawmy teraz definicje.

Definicja 11.12. Niech $\mu \in \mathbb{R}^n$, $\Sigma \in M(n)$ oraz $\Sigma \geqslant 0$. Wtedy powiemy, że wektor losowy $Y : \Omega \to \mathbb{R}^n$ ma *n-wymiarowy rozkład normalny o parametrach μ i Σ* (ozn. $Y \sim N_n(\mu, \Sigma)$), gdy

$$M_Y(t) = e^{\mu^T t + \frac{1}{2}t^T \Sigma t}.$$

Tak więc mówimy, że Y ma n-wymiarowy rozkład normalny, gdy ma rozkład taki, jak wektor losowy, który jest afinicznym przekształceniem wektora Z z punktu (2).

Zauważmy też, że jeśli $Y \sim N_1(0, [0])$, to $M_Y(t) = 1$, wiec $P(Y = 0) = 1$.

Twierdzenie 11.13. *Niech wektor losowy Y ma rozkład normalny $N_n(\mu, \Sigma)$.*

1. *Wtedy $E(Y) = \mu$ oraz $cov(Y) = \Sigma$;*
2. *Jeśli $A \in M(m,n)$, $b \in \mathbb{R}^m$, a $W = AY + b$, to $W \sim N_m(A\mu + b, A\Sigma A^T)$;*
3. *$a^T Y$ ma rozkład normalny dla każdego $a \in \mathbb{R}^n$ (implikacja odwrotna też jest prawdziwa);*
4. *Jeśli $Y = (Y_1, Y_2)^T$, $\mu = (\mu_1, \mu 2)^T$, Y_i jest wektorem n_i-wymiarowym, $n_1 + n_2 = n$ oraz*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

(gdzie $\Sigma_{11} \in M(n_1, n_1)$, etc), to

   (a) *$Y_i \sim N_{n_i}(\mu_i, \Sigma_{ii})$ dla $i = 1, 2$;*
   (b) *Y_1 i Y_2 są niezależne wtedy i tylko wtedy, gdy macierze Σ_{12} i Σ_{21} składają się z samych zer.*

Twierdzenie 11.14. *Niech wektor losowy Y ma rozkład normalny $N_n(\mu, \Sigma)$. Wtedy*

1. *Y ma rozkład ciągły wtedy i tylko wtedy, gdy $\Sigma > 0$. W takiej sytuacji gęstość $f_Y : \mathbb{R}^n \to \mathbb{R}$ dana jest wzorem*

$$f_Y(y) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{det(\Sigma)}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)}; \tag{11.1}$$

2. *Σ jest nieujemnie określona, ale nie dodatnio określona wtedy i tylko wtedy, gdy istnieje podprzestrzeń afiniczna $V \subset \mathbb{R}^n$ wymiaru co najwyżej $n-1$, taka, że $P(Y \in V) = 1$.*

Dla $n = 1$ i $\Sigma = [\sigma^2]$ ($\sigma > 0$) wzór (11.1) sprowadza się do gęstości rozkładu normalnego $N(\mu, \sigma)$, więc w oznaczeniach $N_1(\mu, \sigma^2) = N(\mu, \sigma)$.

Definicja 11.15. Rozkład normalny $N_n(\mu, \sigma^2 I)$, gdzie $\sigma > 0$, a $I \in M(n)$ jest macierzą identycznościową, nazywamy *sferycznym rozkładem normalnym*.

Twierdzenie 11.16. *Sferyczny rozkład normalny ma następujące własności.*

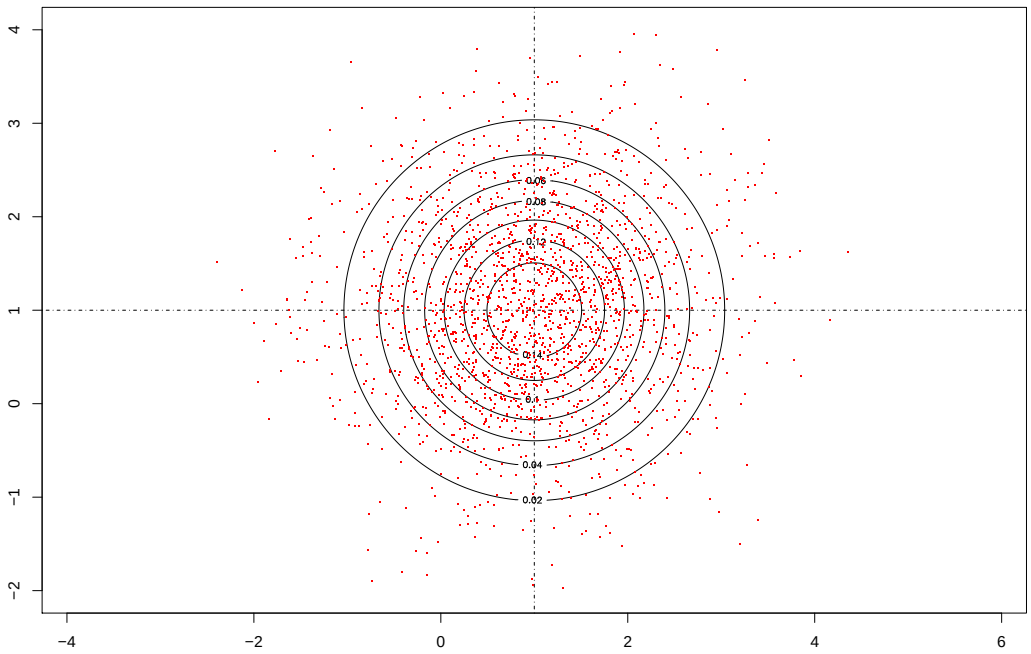
1. *Jeśli $Y \sim N_n(\mu, \sigma^2 I)$, to dla macierzy ortonormalnej A ($A^T = AA^T = I$) wektor losowy $W = A(Y - \mu) + \mu$ ma ten sam rozkład sferyczny $N_n(\mu, \sigma^2 I)$;*
2. *Jeśli $Y \sim N_n(\mu, \sigma^2 I)$, to jej gęstość dana jest wzorem*

$$f_Y(y) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\|y-\mu\|^2}{2\sigma^2}}; \tag{11.2}$$

3. *$Y = (Y_1, \ldots, Y_n)^T \sim N_n(\mu, \sigma^2 I)$ wtedy i tylko wtedy, gdy $Y_1, \ldots, Y_n$ są niezależne oraz $Y_i \sim N(\mu_i, \sigma)$ dla każdego $i = 1, \ldots, n$.*

Jak widać ze wzoru (11.2) poziomice gęstości sferycznego rozkładu normalnego są sferami (stąd nazwa). Z powyższego twierdzenia wynika też, że zmienna losowa o rozkładzie sferycznym $N_n(\mu, \sigma^2 I)$ po obrocie wokół μ ma ten sam rozkład.

Na rysunkach 11.1, 11.2, 11.3 oraz 11.4 przedstawiono poziomice gęstości przykładowych dwuwymiarowych rozkładów normalnych razem z punktami wylosowanymi z tych rozkładów.



Rysunek 11.1: 2000 punktów wylosowanych z rozkładu $N_2(\mu, \Sigma)$ i poziomice jego gęstości dla $\mu = (1,1)^T$ oraz $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

11.5 Projekcje ortogonalne

Rozpoczniemy od kilku twierdzeń z algebry liniowej.

Uwaga 11.17. *Dla dowolnej macierzy $A \in M(n,p)$ oraz wektorów $x \in \mathbb{R}^p$ i $y \in \mathbb{R}^n$ zachodzi*

$$\langle Ax, y\rangle = (Ax)^T y = x^T A^T y = \langle x, A^T y\rangle.$$

Definicja 11.18. Dwa wektory $x, y \in \mathbb{R}^n$ są prostopadłe (inaczej ortogonalne, ozn. $x \perp y$), gdy $\langle x, y\rangle = 0$.

Twierdzenie 11.19. *Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Wtedy istnieje dokładnie jedna podprzestrzeń liniowa $V^\perp \subset \mathbb{R}^n$ taka, że $V \oplus V^\perp = \mathbb{R}^n$ oraz $x \perp y$ dla wszystkich $x \in V$ oraz $y \in V^\perp$.*

Uwaga 11.20. *Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Wtedy*
$V^\perp = \{y \in \mathbb{R}^n \ : \ \forall x \in V \ \ x \perp y\}$ *oraz* $\dim V + \dim V^\perp = n$.

Niech $X \in M(n,p)$. Możemy traktować X także jako odwzorowanie liniowe $\mathbb{R}^p \ni x \to Xx \in \mathbb{R}^n$. Wtedy $\dim \ker X + \dim \operatorname{im} X = p$.

Oznaczmy przez $e_1, \ldots, e_p$ standardową bazę $\mathbb{R}^p$. Wtedy $X_{\bullet j} = Xe_j$ będzie j-tą kolumną macierzy X. Przez $r(X) = \dim \operatorname{im} X$ oznaczmy rząd macierzy X.

Twierdzenie 11.21. *Niech $X \in M(n,p)$. Rozważmy podprzestrzeń liniową $V = \operatorname{im} X \subset \mathbb{R}^n$. Wtedy*

1. *$X_{\bullet 1}, \ldots, X_{\bullet p}$ generuje podprzestrzeń V;*
2. *$X_{\bullet 1}, \ldots, X_{\bullet p}$ jest bazą V wtedy i tylko wtedy, gdy $r(X) = p$ (tzn. gdy X jest monomorfizmem).*

Twierdzenie 11.22. *Niech $X \in M(n,p)$ i $V = \operatorname{im} X$. Wtedy $V^\perp = \ker X^T$.*

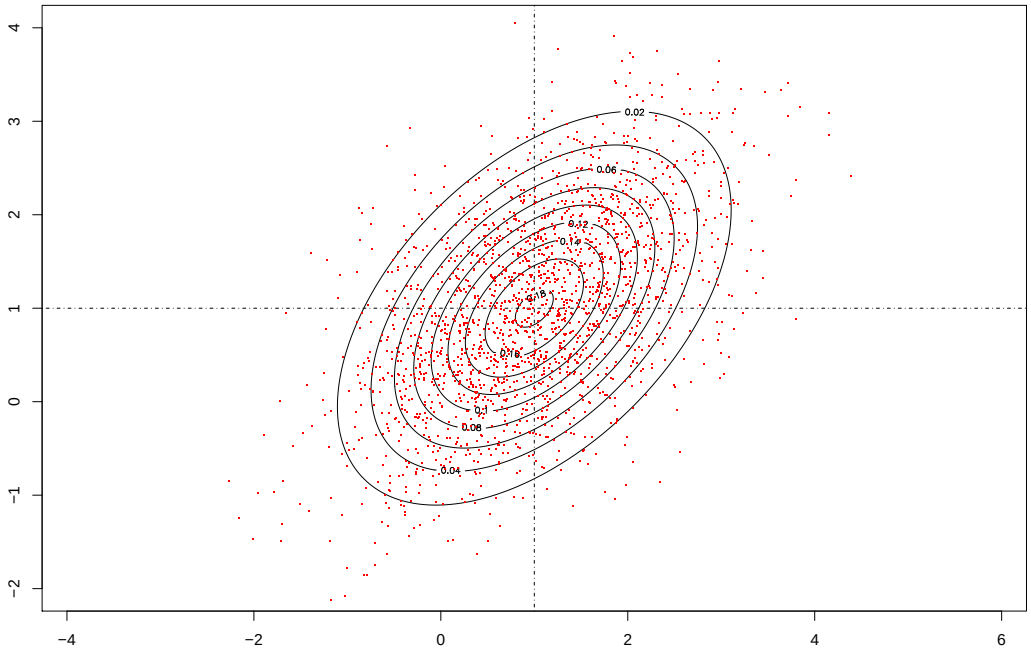
Dowód. Na początek zauważmy, że

$$\dim V^\perp = n - \dim V = n - r(X) = n - r(X^T) = \dim \ker X^T.$$

Wystarczy więc udowodnić, że podprzestrzeń $\ker X^T \subset V^\perp$. Ustalmy więc dowolne $y \in \ker X^T$ oraz $z \in V$. Wtedy $z = Xx$ dla pewnego $x \in \mathbb{R}^p$. Korzystając z Uwagi 11.17 uzyskujemy

$$\langle y, z\rangle = \langle y, Xx\rangle = \langle X^T y, x\rangle = \langle 0, x\rangle = 0,$$

czyli $z \in V^\perp$. □



Rysunek 11.2: 2000 punktów wylosowanych z rozkładu $N_2(\mu, \Sigma)$ i poziomice jego gęstości dla $\mu = (1,1)^T$ oraz $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

Twierdzenie 11.23. *Niech $X \in M(n,p)$. Wtedy* $\operatorname{im}(X^TX) = \operatorname{im}(X^T)$.

Dowód. Niech $V = \operatorname{im} X \subset \mathbb{R}^n$.

Wprost z definicji obrazu wiemy, że $\operatorname{im}(X^TX) \subset \operatorname{im}(X^T)$.

Dla dowodu przeciwnej inkluzji ustalmy dowolne $y \in \operatorname{im}(X^T)$. Wtedy $y = X^Tx$ dla pewnego $x \in \mathbb{R}^n$. Korzystając z faktu, że $V \oplus V^\perp = \mathbb{R}^n$, możemy zapisać, że $x = x_V + x_{V^\perp}$ dla pewnych (jednoznacznie wyznaczonych) $x_V \in V$ i $x_{V^\perp} \in V^\perp$. Wtedy $x_V = Xz$ dla pewnego $z \in \mathbb{R}^p$. Ostatecznie, używając twierdzenia 11.22, otrzymujemy

$$y = X^Tx = X^T(x_V + x_{V^\perp}) = X^Tx_V + X^Tx_{V^\perp} = X^Tx_V = X^TXz \in \operatorname{im}(X^TX).$$

□

Jako wnioski otrzymujemy następujące własności.

Wniosek 11.24. *Niech $X \in M(n,p)$. Wtedy*

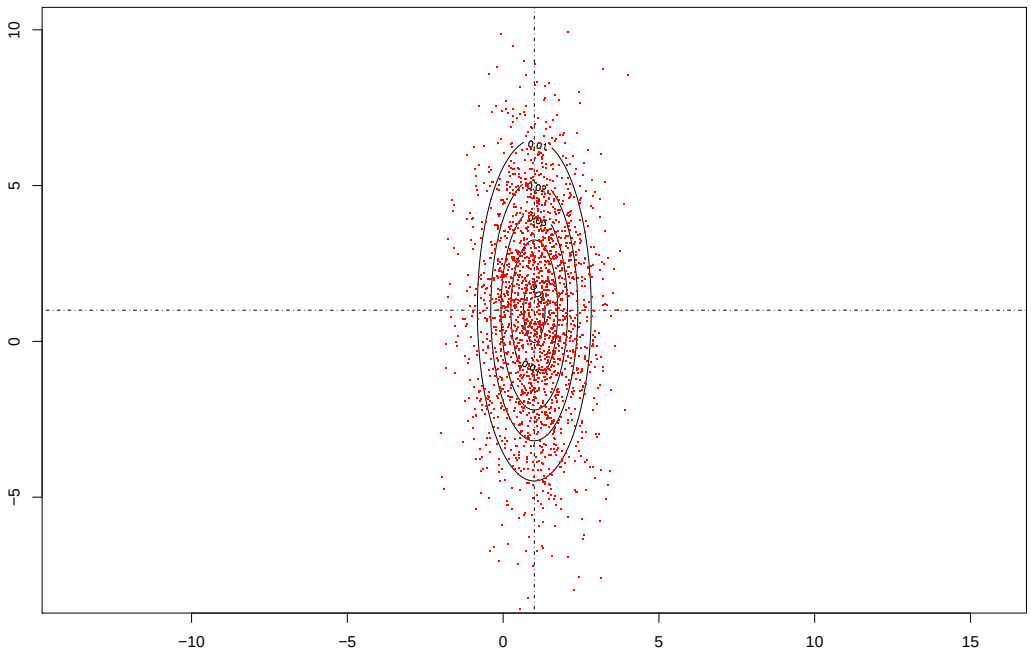
1. $r(X^TX) = r(X^T) = r(X)$;
2. $r(X) = p$ *wtedy i tylko wtedy, gdy X^TX jest nieosobliwa.*

Zdefiniujemy teraz rzutowanie ortogonalne na podprzestrzeń liniową oraz w serii twierdzeń podamy jej własności.

Definicja 11.25. Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Skoro $V \oplus V^\perp = \mathbb{R}^n$, to dla każdego $x \in \mathbb{R}^n$ istnieje dokładnie jeden $x_V \in V$ oraz dokładnie jeden $x_{V^\perp} \in V^\perp$ taki, że $x = x_V + x_{V^\perp}$. Definiujemy odwzorowanie $P_V : \mathbb{R}^n \to \mathbb{R}^n$ wzorem $P_V(x) = x_V$. Odwzorowanie P_V nazywamy *rzutowaniem ortogonalnym* (*projekcją ortogonalną*) na V.

Uwaga 11.26. *Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Łatwo widać, że*

1. P_V *jest odwzorowaniem liniowym;*
2. $x = P_Vx + P_{V^\perp}x$ *dla każdego $x \in \mathbb{R}^n$ (lub równoważnie, że $I = P_V + P_{V^\perp}$);*
3. $P_Vv = v$ *dla każdego $v \in V$;*
4. $P_Vw = 0$ *dla każdego $w \in V^\perp$.*



Rysunek 11.3: 2000 punktów wylosowanych z rozkładu $N_2(\mu, \Sigma)$ i poziomice jego gęstości dla $\mu = (1,1)^T$ oraz $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}$.

Twierdzenie 11.27. *(Pitagorasa) Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Wtedy dla każdego $y \in \mathbb{R}^n$*

$$\|y\|^2 = \|P_V y\|^2 + \|P_{V^\perp} y\|^2.$$

Uwaga 11.28. *Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Ustalmy dowolne $y \in \mathbb{R}^n$ oraz $u \in V$. Wtedy*

$$\|y-u\|^2 = \|P_V(y-u)\|^2 + \|P_{V^\perp}(y-u)\|^2$$
$$= \|P_V y - P_V u\|^2 + \|P_{V^\perp} y - P_{V^\perp} u\|^2 = \|P_V y - u\|^2 + \|P_{V^\perp} y\|^2.$$

Definicja 11.29. Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Wtedy dla $y \in \mathbb{R}^n$ definiujemy

$$\operatorname{dist}(y, V) = \inf\{\|y-u\| \ : \ u \in V\}.$$

Z uwagi 11.28 wynika następujące twierdzenie.

Twierdzenie 11.30. *Niech $V \subset \mathbb{R}^n$ będzie podprzestrzenią liniową. Wtedy dla każdego $y \in \mathbb{R}^n$*

$$\operatorname{dist}(y, V) = \|P_{V^\perp} y\| = \|y - P_V y\|.$$

Poniższe twierdzenie podaje nam wzór na rzutowanie ortogonalne.

Twierdzenie 11.31. *Niech $X \in M(n,p)$ i $V = \operatorname{im} X$. Jeśli $r(X) = p$, to dla każdego $y \in \mathbb{R}^n$*

$$P_V y = X(X^T X)^{-1} X^T y.$$

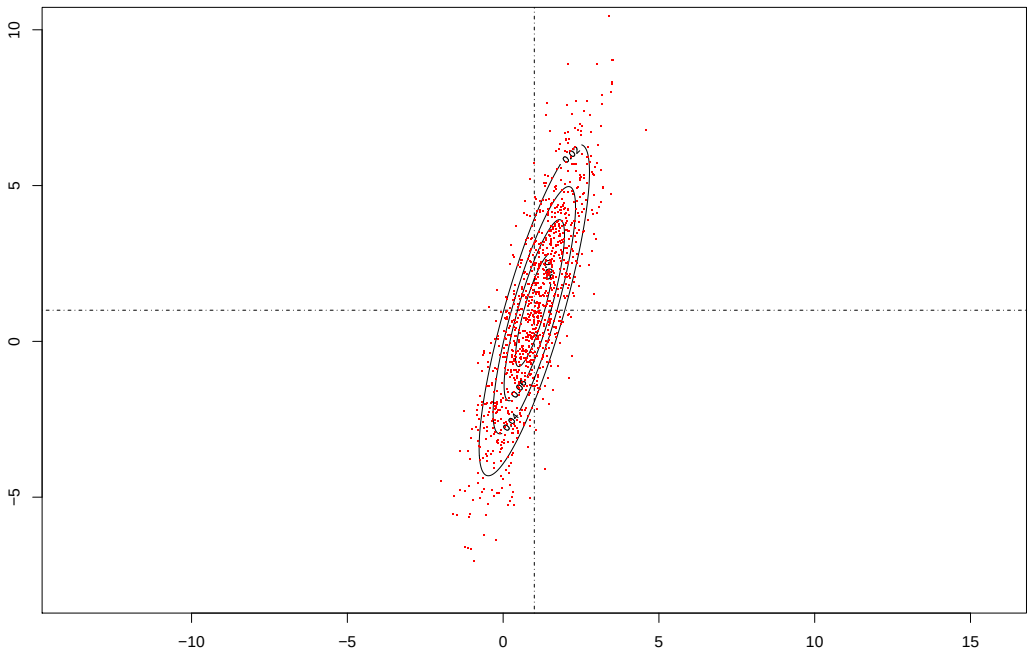
Dowód. Ustalmy $y \in \mathbb{R}^n$. Wtedy wprost z definicji obrazu widzimy, że $X(X^T X)^{-1} X^T y \in V$. Możemy też przedstawić y jako

$$y = X(X^T X)^{-1} X^T y + (y - X(X^T X)^{-1} X^T y).$$

Wystarczy więc udowodnić, że $z = y - X(X^T X)^{-1} X^T y \in V^\perp$. A skoro $V^\perp = \ker(X^T)$, to z kolei wystarczy pokazać, że $X^T z = 0$. Teraz

$$X^T z = X^T (y - X(X^T X)^{-1} X^T y) = X^T y - X^T X (X^T X)^{-1} X^T y$$
$$= X^T y - X^T y = 0.$$

□



Rysunek 11.4: 1000 punktów wylosowanych z rozkładu $N_2(\mu, \Sigma)$ i poziomice jego gęstości dla $\mu = (1,1)^T$ oraz $\Sigma = \begin{bmatrix} 1 & 2.5 \\ 2.5 & 9 \end{bmatrix}$.

Wniosek 11.32. *Niech $X \in M(n,p)$ i $V = \operatorname{im} X$. Jeśli $X_{\bullet 1}, \ldots, X_{\bullet p}$ jest bazą ortonormalną V, to $P_V y = XX^T y$.*

Wniosek 11.33. *Niech $X \in M(n,p)$ i $V = \operatorname{im} X$. Wtedy*

1. *Jeśli $X_{\bullet 1}, \ldots, X_{\bullet p}$ jest bazą ortonormalną V, to $P_V y = XX^T y$;*
2. *$P_{V^\perp} y = (I - X(X^T X)^{-1} X^T) y$;*
3. *$P_V = P_V P_V$ oraz $P_V = P_V^T$ (twierdzenie odwrotne też jest prawdziwe, tzn. jeśli $P \in M(n)$ i $PP = P$ oraz $P = P^T$, to P jest projekcją ortogonalną na pewną podprzestrzeń liniową).*

Twierdzenie 11.34. *Niech $Y \sim N_n(\mu, \sigma^2 I)$, a $V, W \subset \mathbb{R}^n$ będą podprzestrzeniami liniowymi. Wtedy*

1. *$P_V Y \sim N_n(P_V \mu, \sigma^2 P_V)$;*
2. *Jeśli $V \perp W$, to $P_V Y$ i $P_W Y$ są niezależne.*

Dowód.

1. Ze wzoru na rozkład afinicznego przekształcenia wektora losowego o rozkładzie normalnym i z wniosku 11.33 mamy
$$P_V Y \sim N_n(P_V \mu, P_V \sigma^2 I P_V^T) = N_n(P_V \mu, \sigma^2 P_V P_V) = N_n(P_V \mu, \sigma^2 P_V).$$
2. Ćwiczenie.

□

Rozdział 12

Modele liniowe

12.1 Teoria ogólna

Definicja 12.1. (Y, V) nazywamy *modelem liniowym*, gdy $Y \sim N_n(\mu, \sigma^2 I)$, $\mu \in V$, $\sigma > 0$, a $V \subset \mathbb{R}^n$ jest podprzestrzeniami liniową.

Innymi słowy modelem liniowym nazywamy wektor losowy, którego rozkład jest elementem rodziny $\{N_n(\mu, \sigma^2 I) \mid \mu \in V, \sigma > 0\}$ dla ustalonej podprzestrzeni liniowej V. Dalej będziemy rozważać modele liniowe z próbą 1-elementową, na podstawie której będziemy chcieli na przykład wyestymować parametry μ i σ.

Przykłady zaczniemy od najprostszej sytuacji.

Przykład 12.2. Niech $Y_1, \ldots, Y_n$ będzie próbą prostą z wylosowaną z rozkładu $N(m, \sigma)$. Wtedy możemy zdefiniować wektor losowy $Y = (Y_1, \ldots, Y_n)^T$, który będzie miał rozkład $N_n(\mu, \sigma^2 I)$ dla $\mu \in V$, gdzie

$$V = \{\mu \in \mathbb{R}^n \ : \ \mu_1 = \cdots = \mu_n\}$$

(w szczególności $\dim V = 1$). Widzimy więc, że model liniowy jest uogólnieniem standardowego problemu rozważanego wcześniej.

Przykład 12.3. (Analiza wariancji, ANOVA jednoczynnikowa). Załóżmy, że mamy próby proste z k grup wylosowanych z rozkładów normalnych o tej samej wariancji

$$\begin{array}{ccc} Y_{11}, \ldots, Y_{1n_1} & \sim & N(\mu_1, \sigma) \\ \vdots & \vdots & \vdots \\ Y_{k1}, \ldots, Y_{1n_k} & \sim & N(\mu_k, \sigma) \end{array}$$

(Y_{ij} oznacza j-tą obserwację z i-ej grupy). Niech $n = n_1 + \cdots + n_k$. Wtedy

$$Y = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{kn_k} \end{bmatrix} \sim N_n(\mu, \sigma^2 I), \quad \mu \in V,$$

gdzie

$$V = \{\mu \in \mathbb{R}^n \ : \ \mu = (\underbrace{\mu_1, \ldots, \mu_1}_{n_1}, \ldots, \underbrace{\mu_k, \ldots, \mu_k}_{n_k})^T \ : \ \mu_1, \ldots \mu_k \in \mathbb{R}\}$$

(w szczególności $\dim V = k$). Zastąpiliśmy więc n-elementową próbę z rozkładu jednowymiarowego, próbą jednoelementową z rozkładu n-wymiarowego.

Przykład 12.4. (Regresja liniowa) Niech $X = [X_{ij}] \in M(n, p)$ składa się z obserwacji zmiennych objaśnianych (precyzyjniej będziemy rozważać X, w której pierwsza kolumna składa się z samych jedynek), $\varepsilon \sim N_n(0, \sigma^2 I)$. Wtedy $Y = X\beta + \varepsilon \sim N_n(\mu, \sigma^2 I)$ dla $\mu \in V$, gdzie $V = \operatorname{im} X$.

Uwaga 12.5. *Niech $X \in M(n, p)$ będzie monomorfizmem ($r(X) = p$) oraz niech $V = \operatorname{im} X$. Wtedy dla każdego $\mu \in V$ istnieje dokładnie jedno $\beta \in \mathbb{R}^p$ takie, że $\mu = X\beta$. Skoro tak, to $X^T\mu = X^T X\beta$ i w konsekwencji $\beta = (X^T X)^{-1} X^T \mu$.*

Mówimy wtedy, że rozważamy model liniowy we współrzędnych.

12.1.1 Estymatory w modelu liniowym

Rozważmy model liniowy (Y, V) w $\mathbb{R}^n$ taki, że $\dim V = p$, i próbę jednoelementową $\underline{Y} \in \mathbb{R}^n$ (później często próbę też będziemy oznaczać Y i z kontekstu będzie wynikać, czy Y oznacza zmienną losową, czy próbę (obserwacje)). Możemy rozważyć funkcję wiarygodności $L : V \times \mathbb{R}_+ \to \mathbb{R}$

$$L(\mu, \sigma|\underline{Y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\|\underline{Y}-\mu\|^2/\sigma^2}$$

oraz log-wiarygodności

$$l(\mu, \sigma|\underline{Y}) = \log(L(\mu, \sigma|\underline{Y})) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{\|\underline{Y}-\mu\|^2}{\sigma^2}.$$

Widać, że dla każdego ustalonego σ, funkcja przyjmuje wartość największą dla $\hat{\mu} \in V$ takiego, że $\|\underline{Y} - \mu\|^2$ jest najmniejsze. Stąd otrzymujemy, że $\hat{\mu} = P_V \underline{Y}$. Następnie można pokazać, że maksimum funkcji l jest w $\widehat{\sigma^2} = \|P_{V^\perp}\underline{Y}\|^2/n$ (jednak ten estymator jest obciążony).

Twierdzenie 12.6 ([1])**.** *Niech (Y, V) będzie modelem liniowym w $\mathbb{R}^n$ z $\dim V = p$. Wtedy estymatory*

$$\hat{\mu} = P_V Y, \quad \widehat{\sigma^2} = \frac{\|P_{V^\perp} Y\|^2}{n-p}$$

mają następujące własności

1. *$\hat{\mu}$ i $\widehat{\sigma^2}$ są niezależne;*
2. *są nieobciążone, tzn. $E(\hat{\mu}) = \mu$ oraz $E(\widehat{\sigma^2}) = \sigma^2$;*
3. *$\widehat{\sigma^2}(n-p)/\sigma^2 \sim \chi^2(n-p)$.*

Przykład 12.7. Rozważmy sytuację z przykładu 12.2. Wtedy $V = \operatorname{im} X$ dla

$$X = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

oraz

$$\hat{\mu} = P_V Y = X(X^T X)^{-1} X^T Y = X[n]^{-1} X^T Y = \frac{1}{n}[1]_{n\times n} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix},$$

gdzie $\bar{Y} = (\sum_{i=1}^n Y_i)/n$. W konsekwencji otrzymujemy $\hat{m} = \bar{Y}$, czyli estymator największej wiarygodności parametru m w rozkładzie $N(m, \sigma)$.

Rozdział 13

Regresja liniowa

13.1 Uwagi ogólne o modelowaniu statystycznym

Załóżmy, że na pewnej populacji badamy kilka zmiennych (cech). Ustalamy, którą z nich chcemy opisać za pomocą pozostałych. My będziemy ją oznaczać jako Y i nazywać będziemy zmienną *objaśnianą (zależną, regresantem)*[1]. Pozostałe oznaczane przez $X_1, \ldots, X_{p-1}$ nazywamy zmiennymi *objaśniającymi (niezależnymi, regresyjnymi)*[2]. Można wyróżnić dwa podstawowe cele modelowania: chcemy wyjaśnić zależności między badanymi zmiennymi lub/i chcemy przewidywać (predykować) wartości zmiennej objaśnianej dla ustalonych wartości zmiennych objaśniających.

Tworzenie modelu (parametrycznego) możemy podzielić na trzy etapy:

1. Tworzymy postać modelu: definiujemy jaki ma rozkład zmienna Y przy ustalonych wartościach zmiennych objaśniających. Rozkład ten zależy od pewnej liczby parametrów.
2. Estymujemy parametry modelu (używając obserwacji)[3].
3. Wykonujemy diagnostykę wyestymowanego modelu, to znaczy weryfikujemy założenia modelu, jego dopasowanie, czy da się go uprościć, etc.

Z zasady *brzytwy Ockhama*, że nieistotne czynniki w modelu usuwamy, wynikają następujące postulaty:

1. Model powinien mieć możliwie jak najmniej założeń i parametrów.
2. Model liniowy jest lepszy od nieliniowego.
3. Zmienna jest umieszczana w modelu, jeśli jej usunięcie istotnie pogarsza model.

Parafrazując Einsteina, możemy powiedzieć, że model powinien być tak prosty, jak to możliwe, ale nie prostszy.

Podrozdział ten zakończymy uwagą, że model zawsze jest uproszczonym opisem danego zjawiska, więc zawsze powinniśmy patrzeć na niego krytycznie.

13.2 Postać modelu regresji liniowej

Postać modelu regresji liniowej można zapisać równoważnie na kilka sposobów. Zakładamy na razie, że wszystkie zmienne $Y, X_1, \ldots, X_{p-1}$ są ilościowe.

Niech n oznacza liczbę osobników z populacji w próbie prostej. Wtedy możemy zadać postać modelu następująco

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \text{ dla } i = 1, \ldots, n,$$

gdzie Y_i to obserwacja cechy objaśnianej dla i-ego osobnika z próby, $X_{i,j}$ to obserwacja j-ej cechy objaśniającej dla i-ego osobnika z próby, $\varepsilon_1, \ldots, \varepsilon_n$ to liczby wylosowane niezależnie z rozkładu $N(0, \sigma)$, a $\beta_0, \ldots, \beta_{p-1}, \sigma$ to parametry modelu (wspólne dla wszystkich osobników w populacji). Rozważając ogólnie, możemy zapisać, że

$$Y_i \sim N(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1}, \sigma), \text{ dla } i = 1, \ldots, n,$$

[1] ang. *response variable.*
[2] ang. *explanatory variables, predictors.*
[3] Mówimy też, że dopasowujemy model, kalibrujemy model.

albo

$$\begin{cases} Y_i|(X_1,\ldots,X_{p-1}) \sim N(\mu,\sigma) \\ \mu = E(Y_i|(X_1,\ldots,X_{p-1})) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} \end{cases},$$

dla $i = 1,\ldots,n$.

Zamiast opisywać każdą obserwację z osobna, można utworzyć z nich wektory i całość zapisać macierzowo. Zdefiniujmy

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},\ X = \begin{bmatrix} 1 & X_{1,1} & \ldots & X_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \ldots & X_{n,p-1} \end{bmatrix},$$

Wtedy możemy zapiać ten model jako

$$Y = X\beta + \varepsilon,\ \text{gdzie } \varepsilon \sim N_n(0,\sigma^2 I),$$

lub

$$Y \sim N_n(X\beta,\sigma^2 I),$$

albo

$$\begin{cases} Y|X \sim N_n(\mu,\sigma^2 I) \\ \mu = E(Y|X) = X\beta \end{cases}.$$

Jak widać model składa się z dwóch części: $X\beta$ nazywamy *składnikiem systematycznym*[4], a ε nazywamy *składnikiem losowym (błędem losowym)*[5]. Wektor β nazywamy *parametrami strukturalnymi*, β_0 *wyrazem wolnym*[6], σ *odchyleniem błędu losowego*[7].

13.3 Estymacja parametrów

Z postaci macierzowej wynika natychmiast, że mamy do czynienia z modelem liniowym we współrzędnych, więc mamy już gotowe wzory na estymatory i łatwo znajdziemy ich własności.

Jeśli $r(X) = p$, $V = \operatorname{im} X$, to estymatorem parametrów β i σ^2 są odpowiednio

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

i

$$\widehat{\sigma^2} = \frac{\|P_{V^\perp} Y\|^2}{n-p}.$$

Powyższy estymator $\hat{\beta}$ jest też *estymatorem najmniejszych kwadratów*[8] [9], to znaczy minimalizuje *funkcję*[10] $RSS : \mathbb{R}^p \to \mathbb{R}$ daną wzorem

$$RSS(\beta) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1}))^2.$$

Można to łatwo zauważyć, ponieważ

$$\begin{aligned} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1}))^2 &= (Y - X\beta)^T (Y - X\beta) \\ &= \|Y - X\beta\|^2 = \|P_{V^\perp} Y\|^2. \end{aligned}$$

Wprowadźmy jeszcze kilka oznaczeń i definicji.

[4] ang. *systematic component.*
[5] ang. *random component.*
[6] ang. *intercept.*
[7] ang. *residual standard error.*
[8] ang. *LSE – least squares estimator or OLS estimator – ordinary least squares.*
[9] Historycznie patrząc, estymator najmniejszych kwadratów był zdefiniowany wcześniej, wg Encyklopedii PWN, Legendre 1805r, Gauss później twierdził, że on używał od 1794r.
[10] ang. *residual sum of squares*

Zdefiniujmy

$$H = P_V = X(X^T X)^{-1} X^T \quad (\textit{macierz daszkowa}^{11});$$
$$\hat{Y} = P_V Y = HY = X\hat{\beta} \quad (\textit{wartości dopasowane}^{12});$$
$$\hat{\varepsilon} = Y - \hat{Y} \quad (\textit{reszty, błędy obserwowalne}^{13});$$
$$SSE = \hat{\varepsilon}^T \hat{\varepsilon} \quad (\textit{błąd kwadratowy (suma kwadratów reszt, dewiancja)}^{14}).$$

Zauważmy, że wtedy

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta} = P_{V^\perp} Y,$$
$$SSE = \hat{\varepsilon}^T \hat{\varepsilon} = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = \|P_{V^\perp} Y\|^2 = RSS(\hat{\beta})$$

oraz

$$\widehat{\sigma^2} = \frac{SSE}{n-p}.$$

Mianownik w ostatnim wzorze, to znaczy $n-p$ nazywamy *stopniem swobody*[15] modelu. Zauważmy też, że SSE może służyć jako miara dopasowania modelu, to znaczy $SSE = 0$ wtedy i tylko wtedy gdy wektor reszt jest zerowy (oznacza to idealne dopasowanie), a im SSE jest większe tym dopasowanie jest gorsze.

Poniższe twierdzenie podaje nam rozkład estymatora $\hat{\beta}$.

Twierdzenie 13.1. *Niech $Y \sim N_n(X\beta, \sigma^2 I)$, $X \in M(n,p)$ oraz $r(X) = p$. Wtedy $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$.*

Z powyższego twierdzenia wynika, że odchylenie standardowe estymatora $i-1$ parametru strukturalnego wynosi $sd(\hat{\beta}_{i-1}) = \sqrt{\sigma^2 (X^T X)^{-1}_{ii}}$ (dla $i = 1, \ldots, p$). Estymując σ^2, otrzymujemy definicję *błędu standardowego*[16] estymatora $\hat{\beta}_{i-1}$

$$SE(\hat{\beta}_{i-1}) = \hat{\sigma} \sqrt{(X^T X)^{-1}_{ii}}.$$

13.4 Przedziały ufności dla β oraz dla predykacji

13.4.1 Przedział ufności dla β

Dowodzi się (zob. [1]), że $(\hat{\beta}_i - \beta_i)/SE(\hat{\beta}_i) \sim t(n-p)$. Wykorzystując ten fakt, konstruujemy przedział ufności na poziomie ufności $1-\alpha$ (stosujemy zapis $(x \pm a) = (x-a, x+a)$)

$$\beta_{i-1} \in \left(\hat{\beta}_{i-1} \pm \hat{\sigma} \sqrt{(X^T X)^{-1}_{ii}}\, t(1 - \frac{\alpha}{2}, n-p) \right).$$

13.4.2 Przedziały ufności dla predykacji

Załóżmy, że x_0 to wektor obserwacji zmiennych objaśniających dla pewnego osobnika z badanej populacji. Wtedy, zgodnie z postacią naszego modelu, predykacja jest zmienną losową $y_0 = x_0^T \beta + \varepsilon_0 \sim N(x_0^T \beta, \sigma)$. Predykację punktową definiujemy jako

$$\hat{y}_0 = x_0^T \hat{\beta}.$$

Zwróćmy jednak uwagę, że tak naprawdę $\hat{y}_0$ to estymacja wartości średniej rozkładu zmiennej objaśnianej dla osobników o wartościach zmiennych objaśniających x_0. Błąd estymacji ma dwa źródła: błąd estymatora $\hat{\beta}$ i σ. Formalnie

$$var(\hat{y}_0) = var(x_0^T \hat{\beta} + \varepsilon) = var(x_0^T \hat{\beta}) + var(\varepsilon) = x_0^T\, var(\hat{\beta}) x_0 + \sigma^2$$
$$= x_0^T (X^T X)^{-1} \sigma^2 x_0 + \sigma^2 = \sigma^2 \left(1 + x_0^T (X^T X)^{-1} x_0 \right).$$

Ostatecznie uzyskujemy przedział ufności dla wartości średniej predykacji

$$(\hat{y}_0 \pm t(1 - \frac{\alpha}{2}, n-p) \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0})$$

oraz dla predykacji

$$(\hat{y}_0 \pm t(1 - \frac{\alpha}{2}, n-p) \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}).$$

[11] ang. *hat matrix.*
[12] ang. *fitted values.*
[13] ang. *residuals, observed residuals.*
[14] ang. *SSE-error sum of squares; RSS-residual sum of squares; deviance.*
[15] ang. *degree of freedom.*
[16] ang. *standard error.*

13.5 Testowanie hipotez

Dwa modele liniowe (Y, V) i (Y, W) są *zagnieżdżone*[17], jeśli $W \subset V$. Dla modelu regresji liniowej modele zagnieżdżone możemy oznaczyć jako Ω i ω, gdzie $\omega \subset \Omega \subset \{1, X_1, \ldots, X_{p-1}\}$, które oznaczają zbiór zmiennych objaśniających (innymi słowy każda zmienna objaśniająca w modelu ω jest zmienną objaśniającą w modelu Ω). Oznaczmy przez SSE_Ω i SSE_ω błędy kwadratowe tych modeli. Łatwo widać, że wtedy $SSE_\Omega \leqslant SSE_\omega$. Zauważmy też, że te błędy kwadratowe możemy traktować jako zmienne losowe (ponieważ zależą od próby). Wtedy zachodzi twierdzenie.

Twierdzenie 13.2 ([1]). *Przy powyższych oznaczeniach i założeniach zmienna losowa F dana wzorem*

$$F = \frac{\frac{SSE_\Omega - SSE_\omega}{p-q}}{\frac{SSE_\Omega}{n-p}}$$

ma rozkład $F(p-q, n-p)$, gdzie p i q oznacza odpowiednio liczbę parametrów strukturalnych w modelu Ω i ω, a n liczbę obserwacji.

Twierdzenie to może posłużyć do testowania hipotezy zerowej, że modele Ω i ω są tak samo dobrze dopasowane, a równoważnie, że zmienne objaśniające ze zbioru $\Omega \setminus \omega$ są nieistotne statystycznie **w modelu Ω**, czyli parametry strukturalne odpowiadające zmiennym objaśniającym ze zbioru $\Omega \setminus \omega$ są równe zero. Statystyką testową jest powyższa funkcja F, a zbiór krytyczny jest prawostronny $K = (F(1-\alpha, p-q, n-p), +\infty)$.

Uwaga 13.3. *Zauważmy, że badamy istotność zmiennej objaśniającej w ustalonym modelu Ω. Może się zdarzyć, że ta sama zmienna jest istotna w modelu Ω, a nieistotna w modelu Ω' ($\Omega \neq \Omega'$)!*

Najczęściej używa się dwóch szczególnych wersji tego testu.

1. (Test F) $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$.

   Porównujemy wtedy model $Y = X\beta + \varepsilon$ z modelem *pustym* $Y = \beta_0 + \varepsilon$ (to znaczy tylko z wyrazem wolnym).

2. (Test t) Dla ustalonego i testujemy $H_0 : \beta_i = 0$.

   Porównujemy wtedy model $Y = X\beta + \varepsilon$ z tym modelem ale bez i-ej zmiennej objaśniającej. Okazuje się też (zob. [1]), że numerycznie („rachunkowo") takim samym testem jest zastosowanie statystyki testowej $\hat{\beta}_i / SE(\hat{\beta}_i) \sim t(n-p)$.

13.6 Implementacja w R

Rozważmy dane `leafburn` z pakietu `faraway`. Występuje tam 30 obserwacji czterech zmiennych (dla różnych próbek z liści tytoniu): `burntime`-czas spalania w sekundach, procentowa (wagowo) zawartość azotu (`nitrogen`), chloru (`chlorine`) i potasu `potassium`.

Implementujemy model postaci

$$burntime = \beta_0 + \beta_{nitrogen} nitrogen + \beta_{chlorine} chlorine + \beta_{potassium} potassium + \varepsilon.$$

```
> model<-lm(burntime~nitrogen+chlorine+potassium, data=leafburn)
> summary(model)

Call:
lm(formula = burntime ~ nitrogen + chlorine + potassium, data = leafburn)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2442  -3.3475  -0.6669   2.8171  16.7331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.967      8.345   3.471 0.001824 **
nitrogen      -9.089      2.077  -4.376 0.000174 ***
chlorine      -6.743      2.180  -3.092 0.004698 **
```

[17] ang. *nested*.

```
potassium      3.193      1.213   2.632 0.014096 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.373 on 26 degrees of freedom
Multiple R-squared:  0.582,Adjusted R-squared:  0.5338
F-statistic: 12.07 on 3 and 26 DF,  p-value: 3.891e-05
```

Powyżej widzimy między innymi:

1. Podstawowe statystyki reszt (jest to ważne w diagnostyce modelu).
2. Dla czterech parametrów strukturalnych w kolejnych kolumnach mamy: wartość wyestymowana, błąd standardowy, wartość statystyki testowej z testu t oraz *pvalue* z tego testu.
3. Oszacowanie $\hat{\sigma}$ (6.373) oraz stopnie swobody modelu.
4. W ostatniej linii wynik testu F, w szczególności *pvalue* równe $3.891e-05$.

Wyznaczmy teraz 95% przedziały ufności dla parametrów strukturalnych.

```
> confint(model)
                  2.5 %    97.5 %
(Intercept)  11.8143773 46.119244
nitrogen    -13.3586897 -4.819667
chlorine    -11.2244733 -2.260751
potassium     0.6991953  5.686891
```

Przedziały ufności dla predykacji dla trzech przykładowych osobników są wyznaczane za pomocą funkcji `predict`.

```
> dane_do_predykacji=data.frame(nitrogen=c(1,2,3),
+                               chlorine=c(1,2,3),
+                               potassium=c(1,2,3))
> predict(model,newdata = dane_do_predykacji,
+         interval = "confidence") # CI dla E(burntime)
        fit        lwr       upr
1 16.328064   2.467249 30.188878
2  3.689316  -8.414133 15.792766
3 -8.949431 -21.492152  3.593291
>
> predict(model,newdata = dane_do_predykacji,
+         interval = "prediction") # dla predykacji
        fit       lwr       upr
1 16.328064  -2.74320 35.399327
2  3.689316 -14.14560 21.524237
3 -8.949431 -27.08533  9.186467
```

Niepokojące jest to, że niektóre predykacje są ujemne (a przecież czas spalania jest liczbą dodatnią). O czym to może świadczyć?

Użyjmy jeszcze funkcji `anova` do implementacji testu.

```
> model_02=lm(burntime~potassium, data=leafburn)
> anova(model_02,model)  #H_0: \beta_{nitrogen}=\beta_{chlorine}=0
Analysis of Variance Table

Model 1: burntime ~ potassium
Model 2: burntime ~ nitrogen + chlorine + potassium
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     28 2480.2
2     26 1055.9  2    1424.3 17.536 1.509e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Widzimy, że na standardowym poziomie istotności „Model 2” (z kodu powyżej) jest lepiej dopasowany, niż „Model 1”.

13.7 Współczynnik R^2

Twierdzenie 13.4. *Dla modelu regresji liniowej zachodzi*

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{SST}=\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{SSR}+\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{SSE},$$

gdzie SST to całkowita suma kwadratów[18]*, a SSR* regresyjna suma kwadratów[19]

Definicja 13.5. *Współczynnik dopasowania*[20] R^2 definiujemy wzorem

$$R^2=\frac{SSR}{SST}.$$

Twierdzenie 13.6. *Współczynnik dopasowania R^2 ma następujące własności.*

1. $R^2 \in [0,1]$*;*
2. $R^2 = 1 - SSE/SST$*;*
3. *Dla regresji prostej (tzn. $Y=\beta_0+\beta_1X+\varepsilon$) $R^2=\rho(Y,X)^2$, gdzie $\rho(Y,X)$ to współczynnik korelacji liniowej Pearsona.*
4. $|R^2|=1$ *wtedy i tylko wtedy, gdy $Y, X_1, \ldots, X_{p-1}$ są liniowo zależne.*

Uwaga 13.7. (Interpretacja). *Dla modelu pustego (tzn. $Y=\beta_0+\varepsilon$) $SSR=0$, więc $SST=SSE$. Z tego powodu dla dowolnego już modelu mamy*

$$R^2=1-\frac{SSE}{SST}=1-\frac{SSE}{SSE(model\ \ pusty)}.$$

Wynika stąd, że możemy wykorzystać współczynnik R^2 do porównania dwóch modeli, gdy oba mają wyrazy wolne β_0.

Nie ma jednoznacznych kryteriów, które mówią kiedy współczynnik R^2 jest dobry. Oczywiście im bliższy wartości 1 tym lepiej.

13.8 Diagnostyka modelu

Diagnostyka modelu polega na weryfikacji poprawności założeń modelu. Można ją podzielić na trzy grupy.

1. Założenia o błędach losowych:
   a) stałość (jednorodność) wariancji[21];
   b) niezależność;
   c) normalność.
2. Liniowość składnika systematycznego ('$Y=X\beta$').
3. Analiza obserwacji odstających i/lub wpływowych.

Metody wykorzystywane w diagnostyce modelu modelu możemy podzielić na *metody graficzne* (definiujemy różnego rodzaju wykresy/rysunki i sprawdzamy założenia ‚na oko') lub *analityczne* (testowanie hipotez, etc.).

[18] ang. *total sum of squares.*
[19] *wariancja wyjaśniona przez model*, ang. *regression sum of squares, explained variation.*
[20] Współczynnik determinacji, ang. *goodness of fit, coefficient of determination, percentage of variance explained.*
[21] homoskedastyczność, ang. *homoscedascity.*

13.8.1 Sprawdzanie założeń o błędach

Zauważmy na początek, że reszty modelu nie są realizacjami błędów losowych ε

$$\hat{\varepsilon} = Y - \hat{Y} = (I - H)Y = (I - H)(X\beta + \varepsilon) = (I - H)X\beta + (I - H)\varepsilon = (I - H)\varepsilon,$$

czyli

$$\hat{\varepsilon}_i = \varepsilon_i - \sum_{j=1}^{n} H_{ij}\varepsilon_j$$

dla $i = 1, \ldots, n$. Stąd widać, że nawet jeśli ε ma równe wariancje, to już $\hat{\varepsilon}$ niekoniecznie. A z drugiej strony, jeśli ε nie ma rozkładu normalnego, to $\hat{\varepsilon}$ już może mieć rozkład zbliżony do normalnego. W końcu zaś, jeśli błędy losowe są niezależne, to reszty już nie są niezależne (o ile H nie jest diagonalna). Niemniej nie mamy innej możliwości jak tylko użyć reszt to weryfikacji założeń o błędach.

13.8.1.1 Homoskedastyczność

Metody graficzne

Nie da się badać równości wariancji, analizując same reszty. Musimy badać reszty w stosunku do innych własności. Najczęściej porównujemy reszty do wartości dopasowanych, bądź zmiennych objaśniających. Formalnie rysujemy zbiory punktów

1. $\{(\hat{Y}_i, \hat{\varepsilon}_i),\ i = 1, \ldots, n\}$,
2. $\{(X_{i,j}, \hat{\varepsilon}_i),\ i = 1, \ldots, n\}$,

i oceniamy, czy rozrzut reszt wzdłuż danej własności jest podobny, a średnia wartość reszt wynosi zero.

Testy

Wszystkie poniższe testy testują jednorodność wariancji reszt (zob. [4]).

Test Breuscha–Pagana. W tym teście sprawdza się, czy reszty nie zależą od zmiennych objaśniających. Implementujemy model regresji liniowej $(\varepsilon)^2 = Xb + e$ i testujemy hipotezę, że wszystkie współczynniki b są równe zero. W `R` używamy funkcji `bptest{lmtest}`.

Test Goldfelda–Quandta polega na podziale reszt na dwie grupy i przetestowaniu równości wariancji w tych dwóch grupach. Podział odbywa się za pomocą zmiennej `order.by` i kwantylu rzędu `point`, to znaczy, że do pierwszej grupy bierzemy te obserwacje, dla których wartość cechy `order.by` jest mniejsza niż kwantyl rzędu `point`. W `R` używamy funkcji `gqtest{lmtest}(formula,order.by,point)`.

Test Harrisona–McCabe'a. Tak samo jak w poprzednim teście dzielimy obserwacje na dwie grupy, a funkcją testową jest stosunek kwadratów reszt w pierwszej grupie do sumy kwadratów reszt wszystkich obserwacji. Dodatkowo możemy narysować wykres wartości funkcji testowej w zależności od `point`. W `R` używamy funkcji `hmctest{lmtest}(formula,order.by,point,plot=TRUE)`.

13.8.1.2 Niezależność

Sprawdzając niezależność reszt, możemy badać autokorelacje rzędu k. Możemy to zrobić, oceniając zbiór punktów $\{(\hat{\varepsilon}_i, \hat{\varepsilon}_{i-k}), i = k + 1, \ldots, n\}$ lub stosując **test Durbina–Watsona** (dla rzędu 1) lub **test Breuscha–Godfreya** dla rzędów do `order`. W `R` używamy odpowiednio funkcji `dwtest{lmtest}` i `bgtest{lmtest}`.

13.8.1.3 Normalność

Do sprawdzania normalności używamy testów normalności, na przykład testu Shapiro–Wilka lub Andersona–Darlinga. Dla większej liczby obserwacji zaleca się analizę *wykresu kwantylowego*[22].

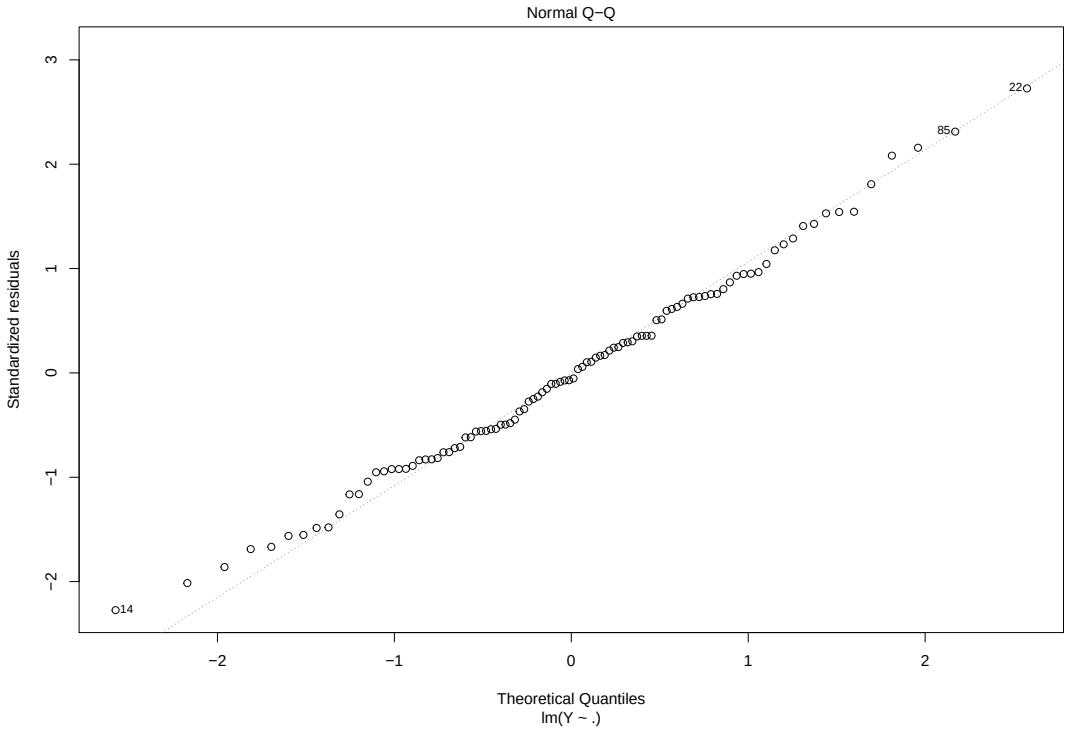
[22] ang. *QQ plot.*

13.8.1.4 Implementacja w R

Zobaczmy jak wygląda diagnostyka dla modelu dopasowanego do symulowanych danych spełniających wszystkie założenia.

```
> X1=runif(100)
> X2=runif(100)
> X3=runif(100)
> Y=1+2*X1+3*X2+4*X3+rnorm(100,0,0.5)
> Dane=data.frame(Y,X1,X2,X3)
>
> model=lm(Y~.,Dane)
> plot(model,2)
> shapiro.test(residuals(model))

Shapiro-Wilk normality test

data:  residuals(model)
W = 0.99312, p-value = 0.895

> library(nortest)
> ad.test(residuals(model))

Anderson-Darling normality test

data:  residuals(model)
A = 0.19359, p-value = 0.8914

> plot(model,1)
> library(lmtest)
> bptest(Y~.,data=Dane)

studentized Breusch-Pagan test

data:  Y ~ .
BP = 6.2447, df = 3, p-value = 0.1003

> gqtest(Y~.,data=Dane)

Goldfeld-Quandt test

data:  Y ~ .
GQ = 1.1262, df1 = 46, df2 = 46, p-value = 0.3443
alternative hypothesis: variance increases from segment 1 to 2

> hmctest(Y~.,data=Dane,plot=TRUE)

Harrison-McCabe test

data:  Y ~ .
HMC = 0.46887, p-value = 0.354

> plot(residuals(model)[-1]~residuals(model)[-100])
>
> dwtest(model)

Durbin-Watson test

data:  model
DW = 2.1414, p-value = 0.7718
alternative hypothesis: true autocorrelation is greater than 0

> bgtest(Y~.,data=Dane,order=2)
```

```
Breusch-Godfrey test for serial correlation of order up to 2

data:  Y ~ .
LM test = 0.72774, df = 2, p-value = 0.695
```



Rysunek 13.1: Wykres kwartylowy dla reszt.

Jak widać wszystkie testy ‚zadziałały', to znaczy nie odrzuciły prawdziwych hipotez zerowych.

13.8.2 Liniowość modelu

Przez liniowość modelu rozumiemy, że zmienna objaśniana zależy liniowo od zmiennych objaśniających. Wydaje się, że najlepiej badałoby się to analizując rysunki rozrzutu Y vs. zmienne objaśniające X_k. Niestety tak nie jest, co widać na przykładzie rysunku 13.5, gdzie umieszczono takie rysunki dla danych liniowych bez błędu losowego. Jak widać jedna zmienna może „zaburzyć" rysunki dla innych zmiennych. Dlatego raczej analizuje się rysunki rozrzutu reszt vs. wartości dopasowane lub zmienne objaśniające. W sytuacji, gdy „na oko" reszty nie są losowo dodatnie i ujemne, może to świadczyć o nieliniowości modelu.

13.8.2.1 Testy

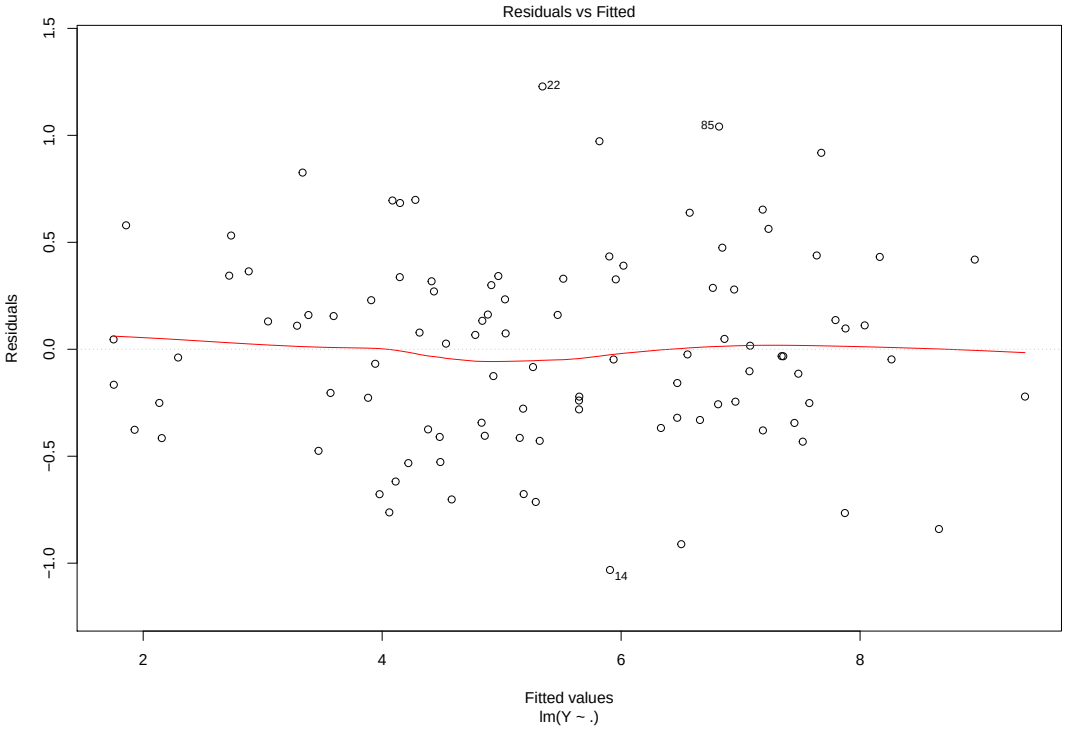
Zainteresowanych dokładniejszą analizą poniższych testów odsyłamy do [4].

Test Harveya–Colliera bazuje na tzw. resztach rekurencyjnych[23], których średnia nie jest równa zero, gdy model nie jest liniowy. Test weryfikuje więc hipotezę, że średnia wartość reszt rekurencyjnych wynosi zero. W `R` używamy funkcji `hmctest{lmtest}(formula,order.by)`.

Test Rainbow. Zauważmy, że jeśli model liniowy jest prawdziwy, to model wyestymowany na podzbiorze obserwacji powinien być podobny do modelu dopasowanego na wszystkich obserwacjach (patrz rysunek 13.6). Test Rainbow porównuje model dopasowany na pełnych danych do modelu dopasowanego na obserwacjach, dla których zmienna `order.by` jest mniejsza niż swoja mediana. W `R` używamy funkcji `raintest{lmtest}(formula,order.by)`.

Test RESET[24]. Test ten porównuje badany model z tym modelem, ale do którego dodano (domyślnie drugie i trzecie) potęgi zmiennych objaśniających. W `R` używamy funkcji `resettest{lmtest}(formula,type= ''regressor'')`.

[23] ang. *recursive residuals*.
[24] ang. *regression specification error*.



Rysunek 13.2: Wykres reszt względem wartości dopasowanych.

13.8.3 Obserwacje wpływowe i odstające

Ogólnie mówiąc (definicje pojawią się za chwilę) obserwacje wpływowe[25] to takie, które mają duży wpływ na estymacje modelu, a obserwacje odstające[26] to takie, które mają dużą resztę. Gdy model liniowy jest prawdziwy wszystkie obserwacje powinny mieć mniej więcej taką samą „wpływowość" i odstawanie. Dlatego identyfikacja obserwacji wpływowych i odstających jest przydatna w diagnostyce.

Definicja 13.8. *Dźwignią*[27] i-ej obserwacji nazywamy i-ty wyraz na przekątnej macierzy H, tj. $h_i = H_{ii}$.

Wiemy, że

$$\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (I - H)Y,$$

czyli

$$var(\hat{\varepsilon}) = (I - H)\sigma^2,$$

i w szczególności

$$var(\hat{\varepsilon}_i) = (1 - H_{ii})\sigma^2 = (1 - h_i)\sigma^2. \tag{13.1}$$

Wynika stąd, że jeśli dźwignia i-ej obserwacji jest duża, to wariancja reszty tej obserwacji jest mała, więc obserwacja ta potencjalnie „przyciąga" płaszczyznę regresji. Zauważmy też, że

$$\hat{Y}_i = h_i Y_i + \sum_{j \neq i} H_{ij} Y_j,$$

więc dźwignia i-ej obserwacji mówi nam jaki wpływ ma Y_i na wyznaczenie $\hat{Y}_i$. Można też udowodnić, że $h_i > 1/n$ oraz

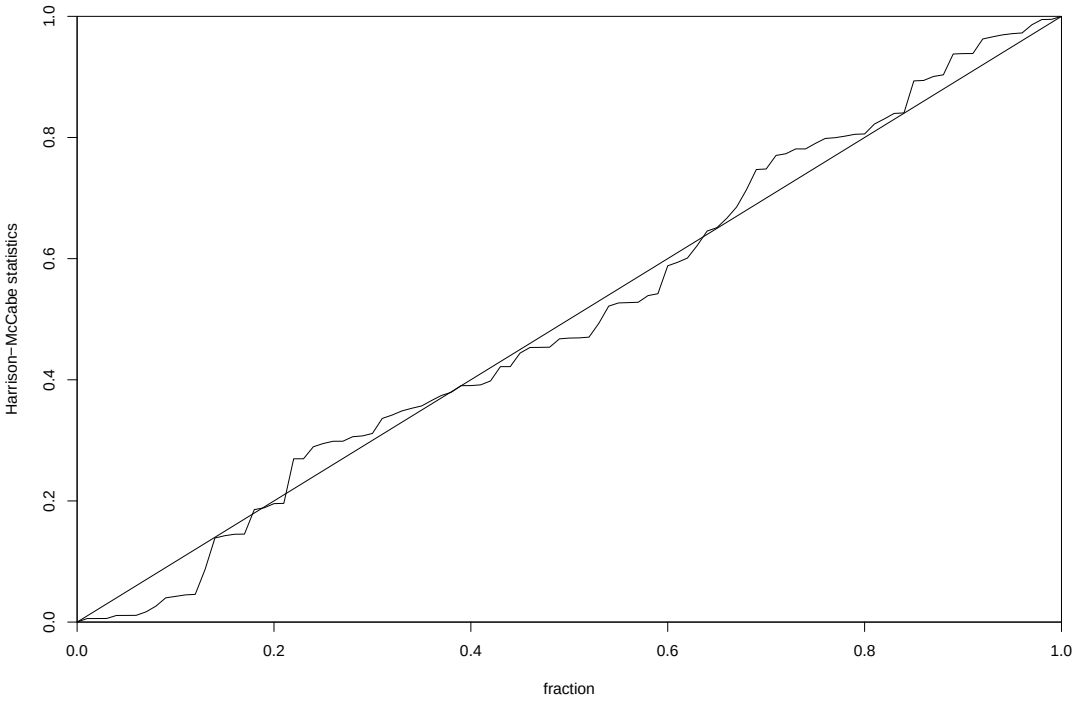
$$\sum_{i=1}^{n} h_i = \operatorname{tr} H = p.$$

Wynika z tego, że średnia dźwignia równa się p/n.

Przyjmuje się, że dźwignia jest duża, jeśli jest większa od $2p/n$ (a dla małych prób większa niż $3p/n$).

Skoro zachodzi (13.1), to możemy standaryzować reszty.

[25] ang. *influential*.
[26] ang. *outliers*.
[27] ang. *leverage*.



Rysunek 13.3: Wykres z testu Harrisona–McCabe'a.

Definicja 13.9. *Resztą standaryzowaną*[28] i-ej obserwacji nazywamy

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}.$$

Pewną wadą dźwigni jest to, że zależy ona tylko od obserwacji cech objaśniających, a nie zależy od obserwacji zmiennej objaśnianej. W `R` używamy odpowiednio funkcji `hatvalues(model)` i `rstandard(model)`.

Inną ideą jak zbadać wpływ i-ej obserwacji na estymowany model jest porównanie modelu dopasowanego na pełnej próbie z modelem dopasowanym na próbie bez i-tej obserwacji[29]. Oznaczmy przez $\hat{\beta}_{(i)}$, $\hat{\sigma}_{(i)}$ wartości estymatorów dla próby bez i-ej obserwacji oraz niech $\hat{Y}_{(i)} = X\hat{\beta}_{(i)} = (\hat{Y}_{(i),1}, \ldots, \hat{Y}_{(i),n})^T$. W `R` aby uzyskać wartości $\hat{\sigma}_{(i)}$ dla wszystkich i używamy funkcji `influence(model)$sigma`.

Możemy teraz badać różnicę $Y_i - \hat{Y}_{(i),i}$. Dlaczego jest to przydatne ilustruje rysunek 13.7.

Definicja 13.10. *Resztą studentyzowaną*[30] i-ej obserwacji nazywamy

$$t_i = \frac{Y_i - \hat{Y}_{(i),i}}{sd(Y_i - \hat{Y}_{(i),i})}.$$

Można udowodnić (zob. [4]), że reszty studentyzowane posiadają następujące własności.

Twierdzenie 13.11. *Przy spełnionych założeniach modelu liniowego t_i ma rozkład $t(n-p-1)$.*

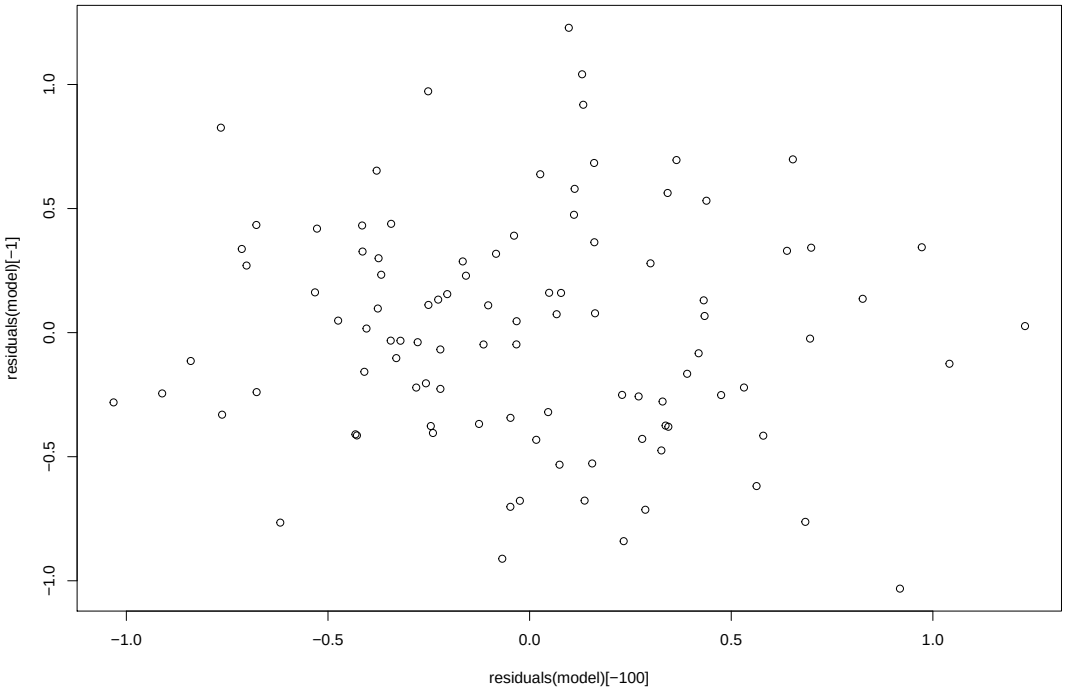
Twierdzenie 13.12. *Reszta studentyzowana jest równa*

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}} = r_i \left(\frac{n-p-1}{n-p-r_i^2}\right)^{1/2}.$$

Pierwsze twierdzenie pozwala nam zdefiniować obserwację odstającą na poziomie istotności α: i-ta obserwacja jest odstająca, jeśli[31]

$$|t_i| > t\left(1 - \frac{\alpha}{2n}, n-p-1\right).$$

[28] ang. *internally studentized residual.*
[29] ang. *one leave out.*
[30] ang. *externally studentized residual, jackknife residual.*
[31] Uwzględniono w tym wzorze tzw. *korekcję Bonferroniego*, zob. [4].



Rysunek 13.4: Wykres do zbadania autokorelacji reszt rzędu 1.

Drugie pozwala wyznaczyć wartości reszt studentyzowanych bez implementowania n różnych regresji.
W `R` używamy funkcji `rstudent(model)`.
Inną metodą badania wpływu i-tej obserwacji na model jest badanie różnicy $\hat{\beta} - \hat{\beta}_{(i)}$. Zauważmy, że zachodzi

$$X(\hat{\beta} - \hat{\beta}_{(i)}) = \hat{Y} - \hat{Y}_{(i)}. \tag{13.2}$$

Definicja 13.13. *Odległością Cooka*[32] i-ej obserwacji nazywamy

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^T (\hat{Y} - \hat{Y}_{(i)})}{p\hat{\sigma}^2}.$$

Wykorzystując (13.2) otrzymujemy, że

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^T (\hat{Y} - \hat{Y}_{(i)})}{p\hat{\sigma}^2} = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X(\hat{\beta} - \hat{\beta}_{(i)})}{p\hat{\sigma}^2}.$$

Można też udowodnić następujące twierdzenie.

Twierdzenie 13.14 ([4])**.** *Odległość Cooka jest równa*

$$D_i = \frac{1}{p} r_i^2 \frac{h_i}{1 - h_i}.$$

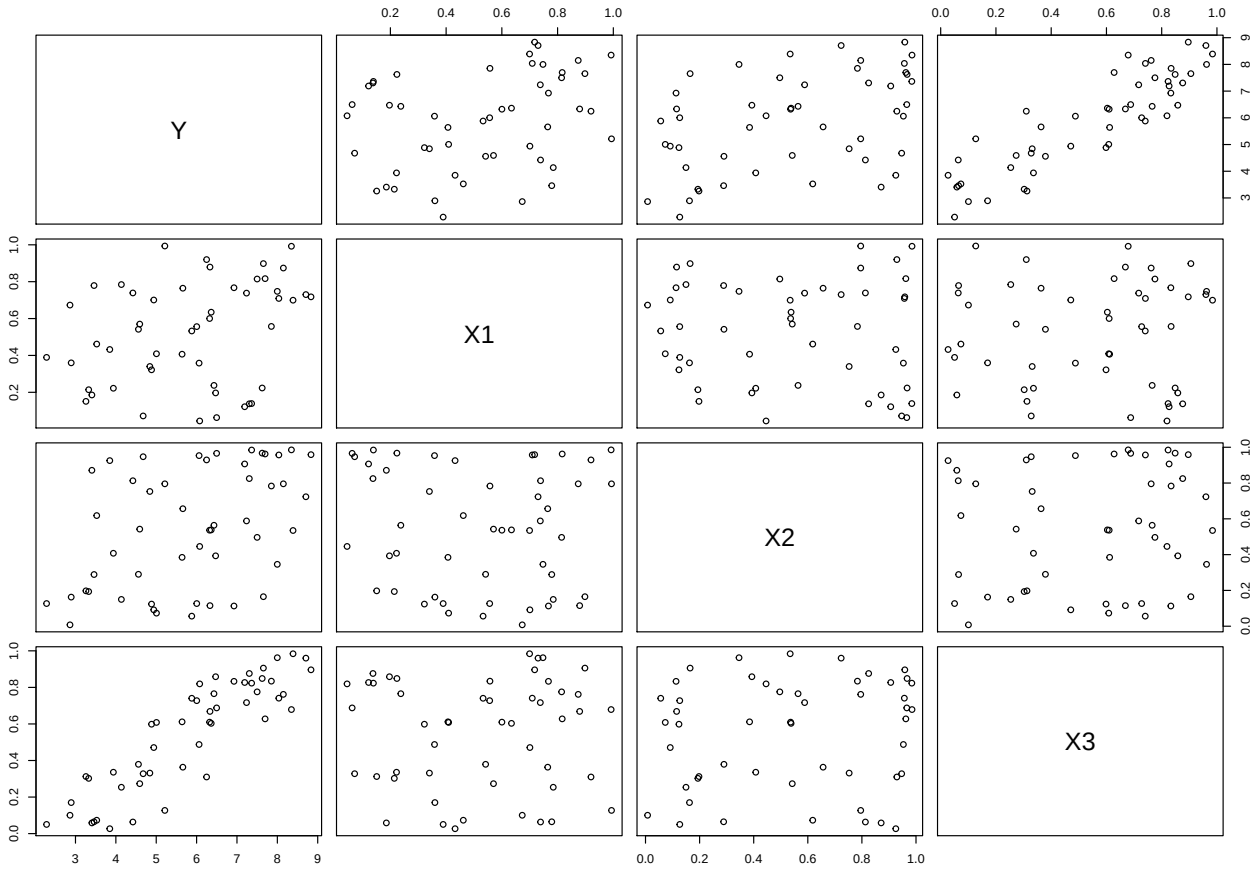
Jak widać odległość Cooka jest kombinacją reszty standaryzowanej i dźwigni. Przyjmuje się (zob. [4]), że obserwacja o odległości Cooka większej niż 0.5, jest „podejrzana", a większej od 1 „bardzo podejrzana". W `R` używamy funkcji `cooks.distance(model)`. Do uzyskania różnych rysunków diagnostycznych możemy użyć instrukcji `plot(model)`. Rysunki te dla przykładowego modelu znajdują się na rysunku 13.8.

Zakończymy kilkoma uwagami o obserwacjach odstających.

1. Jeśli znajdziemy obserwację odstającą, to sprawdzamy, czy nie znalazła się w danych z powodu błędu, badamy fizyczny kontekst (oczywiście nie zawsze jest to możliwe).

2. Dwie lub więcej obserwacji odstających obok siebie mogą spowodować, że się ich nie wykryje. Niemniej „skupisko" wartości odstających raczej świadczy o innej strukturze danych niż liniowa.

[32] ang. *Cook's distance.*



Rysunek 13.5: Rysunki rozrzutu dla danych liniowych.

3. Obserwacja odstająca w jednym modelu nie musi nią być w innym.

4. Usunięcie wartości odstających może poprawić diagnostykę, ale oczywiście zmienia wartości estymatorów i ma wpływ na wnioski z tego modelu. Dlatego powinno się zawsze raportować, jakie obserwacje zostały usunięte i dlaczego.

13.9 Transformacje zmiennych

Jeśli diagnostyka modelu wychodzi źle, inną możliwością oprócz usuwania obserwacji odstających i zmniejszeniu liczby zmiennych objaśniających (o czym dokładniej będzie za chwile) jest transformacja zmiennych: objaśnianej i/lub objaśniających.

Podstawową transformacją jest transformacja wielomianowa. Dokładamy kolejne potęgi zmiennych objaśniających i ich interakcje, na przykład

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_1 X_2 + \beta_5 X_2^2 + \varepsilon.$$

W `R` byłoby tak:

```
lm(Y~polym(X1,X2,degree=2,raw=TRUE),Dane).
```

Model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_2 + \varepsilon$ implementujemy następująco:

```
lm(Y~X1+X1:X2+X2,Dane)
```

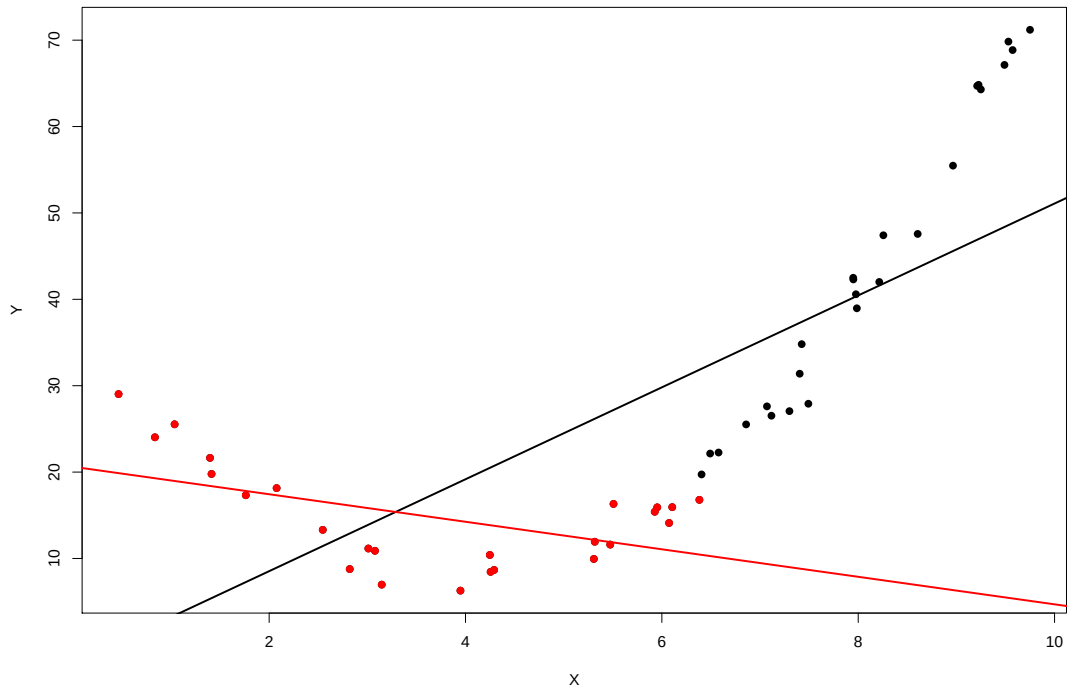
lub

```
lm(Y~(X1+X2)^2,Dane).
```

Do transformacji używa się często logarytmu lub pierwiastka. Oczywiście możemy też transformować zmienną objaśnianą. Na przykład model

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_2 + \beta_4 \sqrt{X_1} + \varepsilon$$

zaimplementujemy tak



Rysunek 13.6: Model dopasowany do wszystkich obserwacji na czarno, do obserwacji w kolorze czerwonym na czerwono.

`lm(log(Y)~X1+X1:X2+X2+sqrt(X1),Dane).`

Oczywiście można stosować jakiekolwiek inne transformacje (wykładnicze, trygonometryczne, etc.), na przykład model

$$\log\left(\frac{1}{1+Y^2}\right) = \beta_0 + \beta_1 X_1 + \beta_2 \frac{1}{X_1} + \beta_3 X_2 + \beta_4 e^{X_2} + \beta_5 \sin(X_1 X_2) + \varepsilon$$

zaimplementujemy w poniższy sposób

`lm(log(1/(1+Y^2))~X1+I(1/X1)+X2+exp(X2)+sin(I(X1*X2)),Dane).`

13.9.1 Metoda Boxa–Coxa

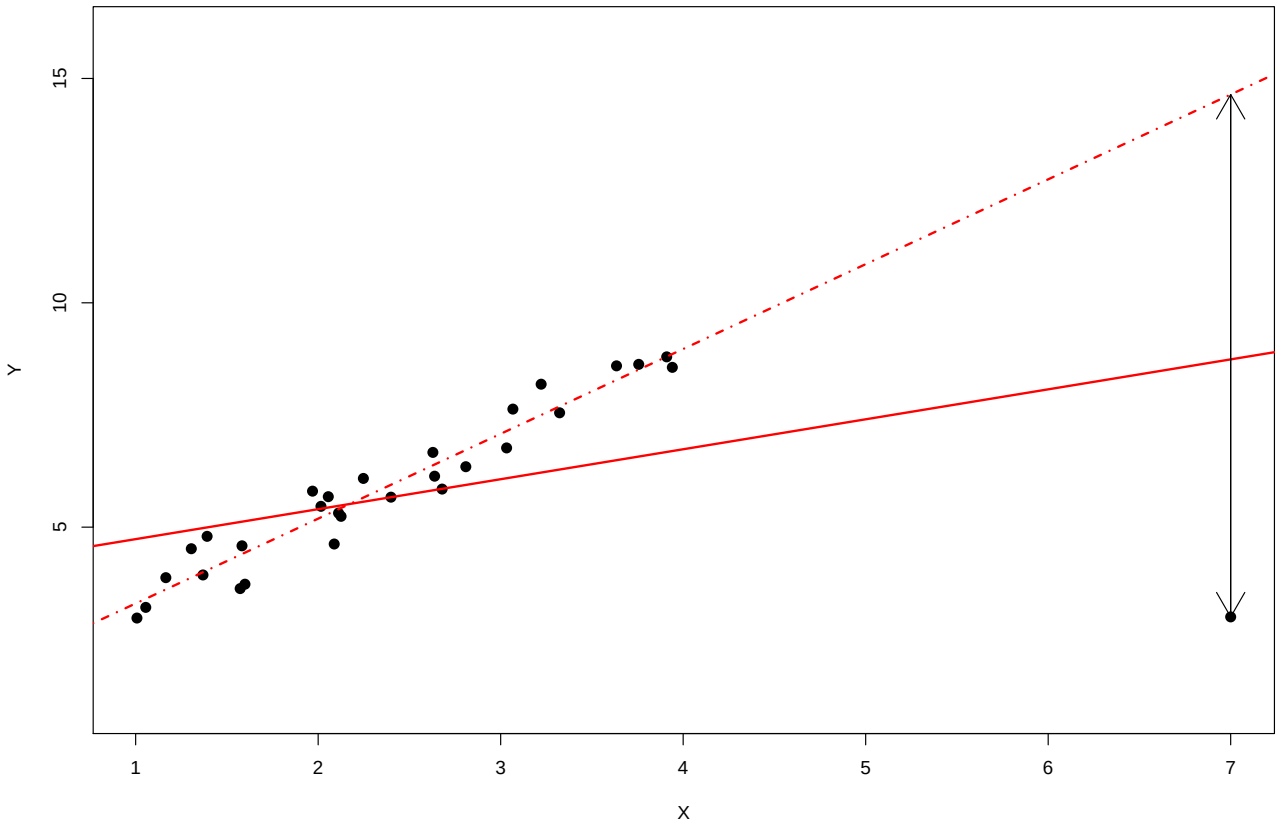
Nie ma ogólnych reguł jak dobierać transformacje. Do wyboru transformacji zmiennej objaśnianej pomocna może być następująca metoda *Boxa–Coxa*. Rozważamy w niej rodzinę transformacji g_λ daną wzorem

$$g_\lambda(Y) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \text{dla } \lambda \neq 0 \\ \log(Y), & \text{dla } \lambda = 0 \end{cases},$$

przy założeniu, że $\min_i\{Y_i\} > 0$. Następnie szukamy argumentu $\hat{\lambda}$ takiego, że w nim jest maksimum funkcji L, która parametrowi λ przyporządkowuje wartość funkcji log-wiarygodności dla dopasowanego modelu postaci $g_\lambda(Y) = X\beta + \varepsilon$. W programie `R` metoda ta jest zaimplementowana w funkcji `boxcox{MASS}(model,plotit=T)`, a przykładowy rezultat działania tej funkcji przedstawia rysunek 13.9. Jeśli $\min_i\{Y_i\} < 0$ to można „przesunąć" obserwacje o a, tak aby $\min_i\{Y_i + a\} > 0$, ale zaleca się, aby a nie było zbyt duże. Po drugie, jeśli na przykład $\hat{\lambda} = 0.45$, to wybieramy bardziej naturalną wartość 0.5. Po trzecie, jeśli $\max_i\{Y_i\}/\min\{Y_i\}$ jest małe, a obserwacje są daleko od 0 to metoda ta wiele nie daje.

Na podobnej zasadzie działa metoda zaimplementowana w funkcji `logtrans{MASS}(model)`, która rozważa rodzinę transformacji zmiennej objaśnianej h_α daną wzorem

$$h_\alpha(Y) = \log(Y + \alpha).$$



Rysunek 13.7: Dwustronną strzałką zaznaczona jest różnica $Y_i - \hat{Y}_{(i),i}$ dla skrajnej prawej obserwacji.

13.10 Wybór (selekcja) zmiennych do modelu

Załóżmy, że mamy zmienne objaśniające $X_1, \ldots, X_{p-1}$. Są to zarówno bezpośrednio obserwowane zmienne na osobnikach, jak ich potencjalne transformacje, interakcje[33], etc. Możemy więc utworzyć 2^{p-1} różnych modeli. Wśród nich chcemy wybrać „najlepszy". Oczywiście nie da się tego zrobić jednoznacznie, chociażby dlatego, że model dobry do predykacji niekoniecznie może dobrze opisywać prawdziwe związki miedzy zmiennymi i na odwrót. Po drugie, aby wybrać model „najlepszy" potrzebujemy umieć porównywać ze sobą dwa dowolne modele. W tym celu musimy zdefiniować odpowiednie kryteria.

13.10.1 Kryteria

Kryterium to funkcja, która modelowi przyporządkowuje liczbę rzeczywistą, a porównanie dwóch modeli względem tego kryterium to porównywanie tych liczb. Zgodnie z zasadą, że model jest „dobry", gdy jest dobrze dopasowany i ma możliwie małą liczbę parametrów, przyjmuje się postulat, że kryterium powinno „nagradzać" model za dobre dopasowanie (w naszym przypadku za małe SSE) i „karać" za dużą liczbę parametrów. Bardziej precyzyjnie, gdy porównujemy dwa modele o tej samej liczbie parametrów, to kryterium powinno wskazać jako lepszy model o mniejszej wartości SSE, a jeśli porównujemy dwa modele o takiej samej wartości SSE to powinien „wygrać" model o mniejszej liczbie parametrów.

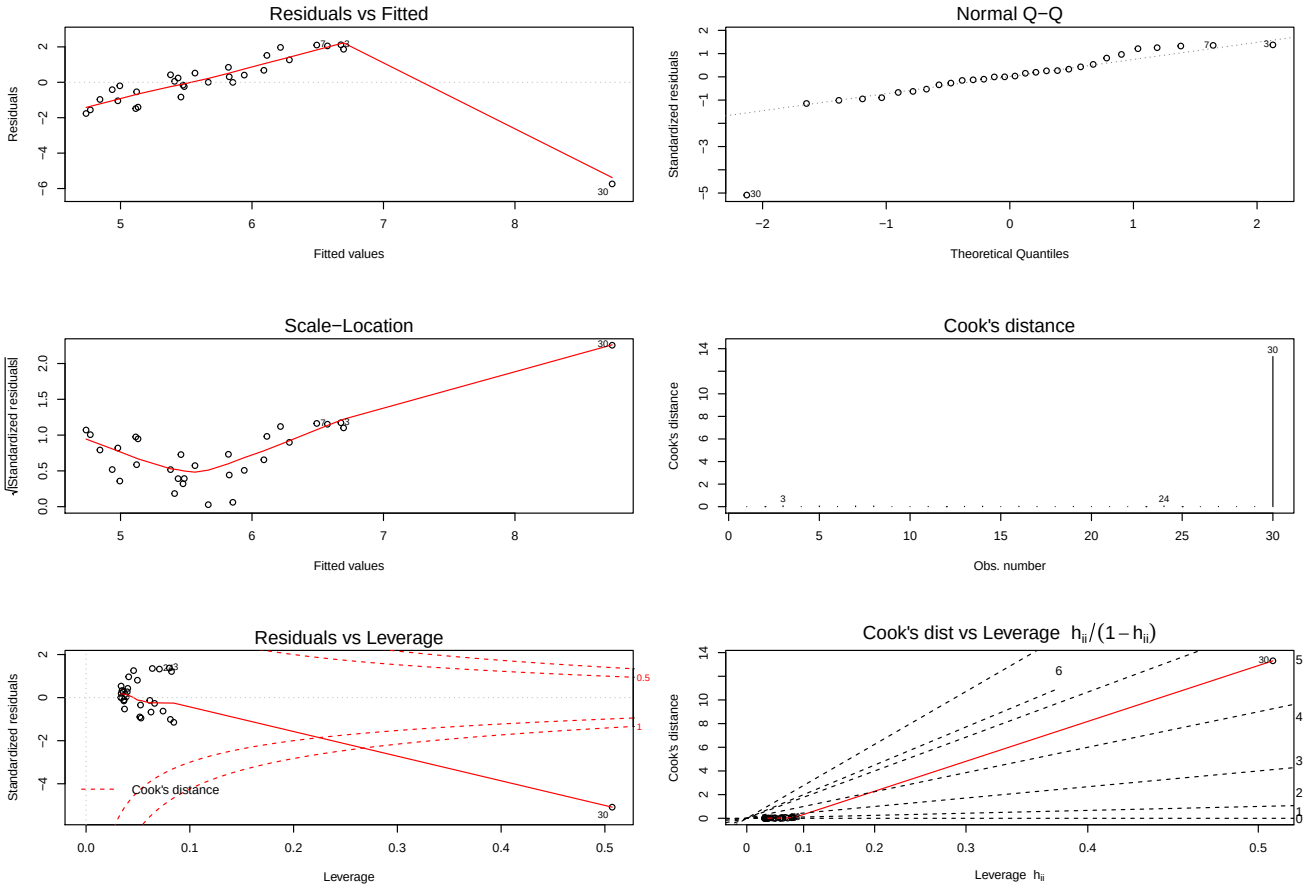
Najczęściej wykorzystywane kryteria (zob. [4]) to:

1. *Zmodyfikowany współczynnik* R^2_{adj}[34] dany wzorem

$$R^2_{adj} = 1 - \frac{SSE_M/(n-p)}{SST_M/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2).$$

   Jak widać, to kryterium im jest większe tym lepiej. Zauważmy też, że sam współczynnik R^2 nie „karze" za liczbę parametrów.

[33] Dołączenie do zmiennych objaśniających ich transformacji nosi nazwę *kodowania* zmiennych.
[34] ang. *Adjust* R^2.



Rysunek 13.8: Przykładowe rysunki diagnostyczne.

2. *Kryterium GIC*[35] zdefiniowane jako

$$GIC(M) = -2\log(L(M)) + k|M|,$$

gdzie $L(M)$ to wartość funkcji wiarygodności modelu M, a $|M|$ to liczba parametrów modelu M. To kryterium jest im jest mniejsze tym lepiej. W programie `R` do wyznaczenia wartości tego kryterium używamy funkcji `AIC(model,k)`. Używa się najczęściej dwóch szczególnych przypadków tego kryterium:

a) *Kryterium Akaika AIC*[36] dla $k = 2$;

b) *Kryterium Schwartza BIC*[37] dla $k = log(n)$.

Jak widać kryterium *BIC* „karze" model za liczbę parametrów bardziej niż *AIC*. W praktyce, jeśli chcemy uzyskać model do predykacji używamy kryterium *AIC*, a gdy interesuje nas model, który dobrze opisuje związki między zmiennymi, używamy kryterium *BIC*.

3. *Statystyka C_p Mallowsa* dana wzorem

$$C_p(M) = \frac{SSE_M}{\hat{\sigma}^2} + 2p - n,$$

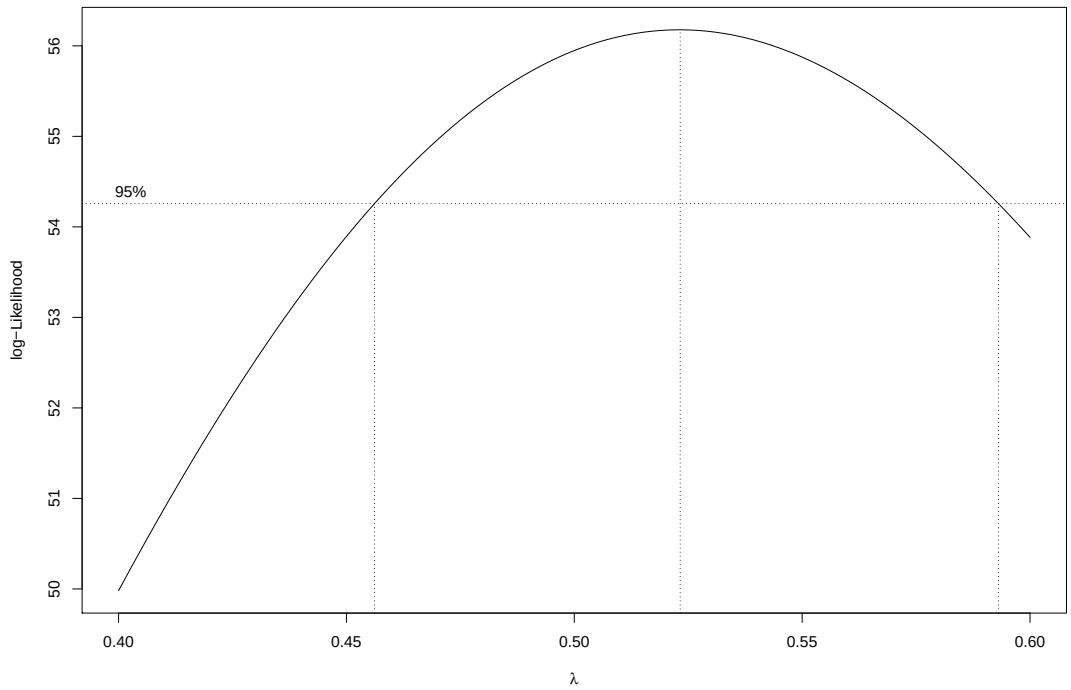
gdzie p to liczba parametrów modelu M, a $\hat{\sigma}^2$ to oszacowanie wariancji błędu losowego dla modelu pełnego. To kryterium im mniejsze tym model jest lepszy.

13.10.2 Metoda krokowa

Nie zawsze jest możliwe (na przykład czasowo) sprawdzenie wszystkich modeli i znalezienie optymalnego modelu w sensie ustalonego kryterium. Wtedy można zastosować metody krokowe dla ustalonego kryterium:

1. *Metoda krokowa w przód*[38]. Zaczynamy od modelu pustego, a w każdym kolejnym kroku dodajemy zmienną,

[35] ang. *Generalized Information Criterion.*
[36] ang. *Akaike Information Criterion.*
[37] ang. *Bayesian Information Criterion.*
[38] ang. *forward selection.*



Rysunek 13.9: Przykładowy rezultat działania funkcji `boxcox(model,lambda=seq(0.4,0.6,by=0.02),plotit=T)`.

która najbardziej polepsza model.

2. *Metoda krokowa w tył*[39]. Zaczynamy od modelu pełnego, a w każdym kolejnym kroku usuwamy zmienną, której brak najbardziej polepsza model.

3. *Metoda krokowa*[40]. Zaczynamy od modelu pełnego, a w każdym kolejnym kroku usuwamy zmienną, której brak najbardziej polepsza model i sprawdzamy, czy dodanie jakiejś zmiennej usuniętej w poprzednich krokach nie poprawi modelu.

Oczywiście metoda krokowa może nie wskazać nam rzeczywistego optimum danego kryterium.

W programie `R` do implementacji metody krokowej dla kryterium *GIC* używamy funkcji `step(model,k,direction, ...)`, gdzie `direction` może być równe `'both','backward'` lub `'forward'`.

Uwaga 13.15. *W metodzie krokowej w każdym kroku porównujemy modele zagnieżdżone, więc możemy także użyć testu t jako kryterium czy usunięcie/dodanie danej zmiennej poprawia model.*

Uwaga 13.16. *Jeśli wśród zmiennych objaśniających istnieje naturalna hierarchia to usuwamy zmienne zgodnie z tą hierarchią. Na przykład gdy występują wielomiany zmiennych, to najpierw usuwamy wyrazy wyższych rzędów.*

13.10.3 Uwagi końcowe

Zakończymy ten rozdział kilkoma uwagami.

1. Modele optymalizujące różne kryteria mogą być różne.

2. Diagnostyka modelu optymalizującego ustalone kryterium może być zła.

3. Dlatego selekcja zmiennych powinna być równoległa z diagnostyką, wykrywaniem obserwacji odstających, etc.

[39] ang. *backward selection.*
[40] ang. *stepwise selection.*

13.11 Analiza kowariancji – ANCOVA

Analizą kowariancji (ANCOVA[41]) nazywamy model regresji liniowej, w którym występuje pośród zmiennych objaśniających co najmniej jedna zmienna ilościowa i co najmniej jedna zmienna jakościowa.

Załóżmy, że zmienna jakościowa V ma k poziomów $l_1, \ldots, l_k$. Wtedy kodujemy tą zmienną następująco: jeden z tych poziomów określamy jako *referencyjny*[42] (w R domyślnie jest to pierwszy poziom w kolejności alfabetycznej), a dla każdego pozostałego poziomu tworzymy zmienną charakterystyczną $\mathbb{1}_{\{V=l_i\}}$ nazywaną zmienną *pustą*[43] i te zmienne umieszczamy w modelu.

Prześledźmy to na przykładzie prostych danych (fikcyjnych)

```
  Waga Wzrost E
1 67.1  172.0 P
2 78.3  176.0 P
3 83.1  185.2 P
4 70.1  178.0 S
5 66.8  180.0 S
6 83.3  185.0 S
7 70.0  167.0 W
8 88.8  173.0 W
9 97.5  186.0 W
```

gdzie zmienne `Waga` i `Wzrost` są ilościowe, a zmienna `E` („wykształcenie") jest jakościowa o trzech poziomach: „P", „S" oraz „W". Niech `Waga` będzie zmienną objaśnianą. Wtedy poziomy zmiennej E dzielą obserwacje na trzy grupy i możemy rozważyć dwa modele: *addytywny*, w którym parametr przy zmiennej `Wzrost` jest taki sam w każdej grupie, a tylko wyrazy wolne są różne albo model *z interakcją*, w którym dopasowujemy model regresji do każdej grupy osobno.

Formalnie model addytywny będzie następujący

$$Waga = \beta_0 + \beta_1 Wzrost + \beta_2 \mathbb{1}_{\{E='S'\}} + \beta_3 \mathbb{1}_{\{E='W'\}} + \varepsilon. \tag{13.3}$$

W R implementujemy go następująco

```
> model_2=lm(Waga~Wzrost+E,Dane)
> summary(model_2)

Call:
lm(formula = Waga ~ Wzrost + E, data = Dane)

Residuals:
      1       2       3       4       5       6       7       8       9
-1.4741  4.4288 -2.9547  0.6729 -5.2757  4.6029 -4.3976  6.4567 -2.0591

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -159.203     54.826  -2.904  0.03364 *
Wzrost         1.324      0.308   4.300  0.00772 **
ES            -7.093      4.514  -1.571  0.17695
EW            12.445      4.463   2.789  0.03850 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.39 on 5 degrees of freedom
Multiple R-squared:  0.8422,Adjusted R-squared:  0.7476
F-statistic: 8.898 on 3 and 5 DF,  p-value: 0.01897
```

Czyli dopasowany model jest następujący:

$$Waga = -159.203 + 1.324 \times Wzrost + \varepsilon$$

dla osób na poziomie referencyjnym „P";

[41] ang. *Analysis of covariance.*
[42] ang. *reference level, baseline level.*
[43] ang. *dummy variable.*

$$Waga = (-159.203 + (-7.093)) + 1.324 \times Wzrost + \varepsilon$$

dla osób na poziomie referencyjnym „S";

$$Waga = (-159.203 + 12.445) + 1.324 \times Wzrost + \varepsilon$$

dla osób na poziomie referencyjnym „W".

W **tym** modelu współczynnik dla $\mathbb{1}_{\{E='W'\}}$ jest istotny statystycznie, więc poziom „W" różni się istotnie od poziomu referencyjnego. Ogólnie diagnostykę wykonujemy analogicznie jak poprzednio. Dla tego modelu macierz $\boldsymbol{X}$ wygląda następująco

```
> model.matrix(model_2)
  (Intercept) Wzrost ES EW
1           1  172.0  0  0
2           1  176.0  0  0
3           1  185.2  0  0
4           1  178.0  1  0
5           1  180.0  1  0
6           1  185.0  1  0
7           1  167.0  0  1
8           1  173.0  0  1
9           1  186.0  0  1
```

Możemy jeszcze porównać ten model z modelem bez zmiennej E:

```
> model_1=lm(Waga~Wzrost,Dane)
> anova(model_2,model_1)
Analysis of Variance Table

Model 1: Waga ~ Wzrost + E
Model 2: Waga ~ Wzrost
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      5 145.26
2      7 656.59 -2   -511.33 8.8005 0.02302 *
```

Jak widać model z addytywnie dodaną zmienną $\boldsymbol{E}$ jest lepiej dopasowany ($pvalue = 0.02302$).

Model z interakcją będzie zaś następujący

$$Waga = \beta_0 + \beta_1 Wzrost + \beta_2 \mathbb{1}_{\{E='S'\}} + \beta_3 \mathbb{1}_{\{E='W'\}} + \beta_4 Wzrost \times \mathbb{1}_{\{E='S'\}} + \beta_5 Wzrost \times \mathbb{1}_{\{E='W'\}} + \varepsilon.$$

W R implementujemy go następująco

```
 > model_3=lm(Waga~Wzrost*E,Dane)
> summary(model_3)

Call:
lm(formula = Waga ~ Wzrost * E, data = Dane)

Residuals:
     1      2      3      4      5      6      7      8      9
-2.806  4.026 -1.220  3.173 -4.442  1.269 -4.414  6.452 -2.037

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -117.9123   118.5863  -0.994    0.393
Wzrost         1.0920     0.6669   1.637    0.200
ES          -199.2300   255.7803  -0.779    0.493
EW           -28.4908   143.9315  -0.198    0.856
Wzrost:ES      1.0657     1.4185   0.751    0.507
Wzrost:EW      0.2303     0.8129   0.283    0.795

Residual standard error: 6.384 on 3 degrees of freedom
Multiple R-squared:  0.8672,Adjusted R-squared:  0.6459
F-statistic: 3.919 on 5 and 3 DF,  p-value: 0.1451
```

Czyli dopasowany model jest następujący:

$$Waga = -117.9123 + 1.0920 \times Wzrost + \varepsilon$$

dla osób na poziomie referencyjnym „P";

$$Waga = (-117.9123 + (-199.2300)) + (1.0920 + 1.0657) \times Wzrost + \varepsilon$$

dla osób na poziomie referencyjnym „S";

$$Waga = (-117.9123 + (-28.4908)) + (1.0920 + 0.2303) \times Wzrost + \varepsilon$$

dla osób na poziomie referencyjnym „W".

Macierz X tego modelu jest równa

```
> model.matrix(model_3)
  (Intercept) Wzrost ES EW Wzrost:ES Wzrost:EW
1           1  172.0  0  0         0         0
2           1  176.0  0  0         0         0
3           1  185.2  0  0         0         0
4           1  178.0  1  0       178         0
5           1  180.0  1  0       180         0
6           1  185.0  1  0       185         0
7           1  167.0  0  1         0       167
8           1  173.0  0  1         0       173
9           1  186.0  0  1         0       186
```

Na koniec porównajmy model z interakcją z modelem addytywnym:

```
> anova(model_3,model_2)
Analysis of Variance Table

Model 1: Waga ~ Wzrost * E
Model 2: Waga ~ Wzrost + E
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      3 122.25
2      5 145.26 -2   -23.005 0.2823 0.7721
```

Jak widać dołożenie interakcji nie poprawiło istotnie dopasowania ($pvalue = 0.7721$).

Rozdział 14

Analiza wariancji – ANOVA

14.1 ANOVA jednoczynnikowa

Analizą wariancji jednoczynnikową[1] nazywamy model liniowy, w którym występuje tylko jedna zmienna („czynnik") objaśniająca i jest ona jakościowa.

Załóżmy, że zmienna objaśniająca X ma k poziomów. Wtedy nasza próba jest podzielona na k grup. Mamy więc próby proste zmiennej objaśnianej, wylosowanych z rozkładów normalnych o tej samej wariancji, zgrupowane w k grup:

$$\begin{array}{ccc} Y_{11}, \ldots, Y_{1n_1} & \sim & N(\mu_1, \sigma) \\ \vdots & \vdots & \vdots \\ Y_{k1}, \ldots, Y_{1n_k} & \sim & N(\mu_k, \sigma) \end{array}$$

(Y_{ij} oznacza j-tą obserwację z i-ej grupy). Niech $n = n_1 + \cdots + n_k$. Wtedy

$$Y = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{kn_k} \end{bmatrix} \sim N_n(\mu, \sigma^2 I),$$

gdzie

$$\mu = (\underbrace{\mu_1, \ldots, \mu_1}_{n_1}, \ldots, \underbrace{\mu_k, \ldots, \mu_k}_{n_k})^T.$$

Bardziej skrótowo, możemy zapisać, że $Y_{ij} \sim N(\mu_i, \sigma)$ dla $i = 1, \ldots, k$. Równoważnie, że

$$Y_{ij} \sim N(\mu + \alpha_i, \sigma)$$

dla $i = 1, \ldots, k$ i $\sum_i \alpha_i = 0$ (gdzie μ to średnia globalna zmiennej Y, a α_i to wpływ i-ego poziomu czynnika X na średnią) albo

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

dla $\varepsilon_{ij} \sim N(0, \sigma)$, albo (jak w `R`), że

$$Y_{1j} \sim N(\mu, \sigma) \text{ i } Y_{ij} \sim N(\mu + \alpha_i, \sigma)$$

dla $i = 2, \ldots, k$ (wtedy μ to średnia zmiennej Y dla poziomu referencyjnego).

Analiza wariancji sprowadza się do porównania tego modelu z modelem pustym (testem F), więc do przetestowania hipotezy o równości średnich w grupach

$$H_0 : \ \mu_1 = \cdots = \mu_k$$

z hipotezą alternatywną

$$H_1 : \ \exists\, i, j \ : \ \mu_i \neq \mu_j.$$

Dla przykładu rozważmy (fikcyjne dane):

[1] ang. *One-way analysis of variance.*

```
> Dane
  Waga X
1 67.1 P
2 72.3 P
3 83.1 P
4 70.1 S
5 66.8 S
6 73.3 S
7 80.0 W
8 88.8 W
9 97.5 W
```

Wtedy implementujemy analizę wariancji następująco

```
> anova(lm(Waga~X,data=Dane))
Analysis of Variance Table

Response: Waga
          Df Sum Sq Mean Sq F value  Pr(>F)
X          2 579.66 289.830  5.6556 0.04164 *
Residuals  6 307.48  51.247
```

lub

```
> model_anova<-aov(Waga~X,data=Dane)
> summary(model_anova)
            Df Sum Sq Mean Sq F value Pr(>F)
X            2  579.7  289.83   5.656 0.0416 *
Residuals    6  307.5   51.25
```

Jak widać $pvalue = 0.0416$, więc na poziomie 0.05 odrzucamy hipotezę zerową o równości średnich w tych grupach.

14.1.1 Analiza post hoc

Gdy odrzucamy hipotezę zerową wykonujemy tzw. analizę *post hoc* w celu zlokalizowania grup, których średnie się różnią. Najczęściej używa się następujących testów post hoc (zob. [4], są one dostępne w pakiecie `agricolae`):

1. *Test HSD*[2] *Tuckeya* dla grup równolicznych:

   ```
   > print(HSD.test(model_anova, "X")$groups)
         Waga groups
   W 88.76667      a
   P 74.16667     ab
   S 70.06667      b
   ```

   Grupy, których część wspólna oznaczeń jest pusta, różnią się średnią. W tym przypadku średnie grup W i S się różnią.

2. *Test Studenta–Newmana–Keulsa* dla grup równolicznych:

   ```
   > print(SNK.test(model_anova, "X")$groups)
         Waga groups
   W 88.76667      a
   P 74.16667      b
   S 70.06667      b
   ```

   Ten test wskazał, że średnią różnią się grupy W i P oraz W i S.

3. *Test LSD*[3] *Fishera* (tu nie ma założenia o równoliczności prób). Test ten polega na porównaniu wszystkich par testem t z korekcją pvalue ze względu na liczbę wykonanych testów.

[2] ang. *Honestly Significant Differences.*
[3] ang. *Least Significant Difference.*

```
> print(LSD.test(model_anova, "X",p.adj="holm")$group)
      Waga groups
W 88.76667      a
P 74.16667      a
S 70.06667      a
```

W tym przypadku test nie wykrył par różniących się średnią.

14.1.2 Sprawdzenie jednorodności wariancji i normalności reszt

Normalność reszt sprawdzamy standardowo:

```
> shapiro.test(residuals(lm(Waga~X,data=Dane)))

Shapiro-Wilk normality test

data:  residuals(lm(Waga ~ X, data = Dane))
W = 0.94121, p-value = 0.5947
```

Do badania jednorodności wariancji używamy innych testów niż poprzednio, a mianowicie testów jednorodności wariancji w k grupach. Mogą to być następujące testy (zob. [4], testy 2-4 są dostępne w pakiecie `lawstat`):

1. *Test Bartletta* (przy założeniu normalności prób w grupach):

   ```
   > bartlett.test(Waga~X,data=Dane)

   Bartlett test of homogeneity of variances

   data:  Waga by X
   Bartlett's K-squared = 1.4982, df = 2, p-value = 0.4728
   ```

2. *Test Browna–Forsytha*:

   ```
   > levene.test(Dane$Waga,Dane$X)

   Modified robust Brown-Forsythe Levene-type test based on the absolute
   deviations from the median

   data:  Dane$Waga
   Test Statistic = 0.61119, p-value = 0.5733
   ```

3. *Test Levene'a*:

   ```
   > levene.test(Dane$Waga,Dane$X,location="mean")

   Classical Levene's test based on the absolute deviations from the mean

   data:  Dane$Waga
   Test Statistic = 0.98592, p-value = 0.4264
   ```

4. *Test Flignera–Killeena*:

   ```
   > fligner.test(Dane$Waga~Dane$X)

   Fligner-Killeen test of homogeneity of variances

   data:  Dane$Waga by Dane$X
   Fligner-Killeen:med chi-squared = 1.3257, df = 2, p-value = 0.5154
   ```

14.2 ANOVA dwuczynnikowa

Analizą wariancji dwuczynnikową[4] nazywamy model liniowy, w którym występują tylko dwie jakościowe zmienne („czynniki") objaśniające, powiedzmy A i B.

Załóżmy, że zmienna objaśniająca A ma k poziomów, a B ma l poziomów. Wtedy nasza próba jest podzielona na $k \times l$ grup. Mamy więc próby proste zmiennej objaśnianej, wylosowanych z rozkładów normalnych o tej samej wariancji, zgrupowane w $k \times l$ grup. Oznaczmy przez Y_{ijm} m-tą obserwację zmiennej Y w grupie wyznaczonej przez i-ty poziom zmiennej A i j-ty poziom zmiennej B. Wtedy możemy zapisać, że

$$Y_{ijm} \sim N(\mu_{ij}, \sigma)$$

dla $i = 1, \ldots, k$, $j = 1, \ldots, l$ i $m = 1, \ldots, n_{ij}$, bądź równoważnie, że

$$Y_{ijm} = \mu_{ij} + \varepsilon_{ijm}$$

dla niezależnych $\varepsilon_{111}, \ldots, \varepsilon_{kln_{kl}}$ o rozkładach $N(0, \sigma)$.

Możemy rozważyć następujące modele (w nawiasach formuła, która implementuje te modele w `R`):

1. (model pusty) $Y_{ijm} = \mu + \varepsilon_{ijm}$ ($Y \sim 1$);
2. $Y_{ijm} = \mu + \alpha_i + \varepsilon_{ijm}$, $\alpha_1 = 0$ ($Y \sim A$);
3. $Y_{ijm} = \mu + \beta_j + \varepsilon_{ijm}$, $\beta_1 = 0$ ($Y \sim B$);
4. (model addytywny)
$$Y_{ijm} = \mu + \alpha_i + \beta_j + \varepsilon_{ijm},$$
$\alpha_1 = \beta_1 = 0$ ($Y \sim A + B$);
5. (model z interakcją, model pełny)
$$Y_{ijm} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijm},$$
$\alpha_1 = \beta_1 = 0$, $\gamma_{1j} = 0$ i $\gamma_{i1} = 0$ dla wszystkich i i j ($Y \sim A \star B$ lub $Y \sim A + B + A : B$).

Możemy testować istotność czynnika A (czyli zerowość parametrów α_i), czynnika B (czyli zerowość parametrów β_j) oraz istotność interakcji (czyli zerowość parametrów γ_{ij}). Istnieje wiele różnych możliwości wykonania takich testów, na przykład aby sprawdzić istotność czynnika A możemy porównać model $Y \sim A$ z modelem pustym albo model $Y \sim B$ z modelem addytywnym. Dlatego rozważa się dwa typy testów:

1. *typu I* (*sekwencyjne* [5]): porównujemy modele w następującej kolejności:
   - model pusty vs. $Y \sim A$;
   - $Y \sim A$ vs. $Y \sim A + B$;
   - $Y \sim A + B$ vs. $Y \sim A \star B$.

   W `R` implementujemy te testy za pomocą `anova(lm(Y~A*B))` lub `summary(aov(Y~A*B))`. Testy te **zależą** od jakiej zmiennej ‚zaczynamy' (od A czy od B), to znaczy, dwa pierwsze testy wywołane przez `summary(aov(Y~A*B))` i `summary(aov(Y~B*A))` będą inne (i zazwyczaj dadzą inne wyniki, gdy zmienne A i B będą zależne).
2. *typu III* (*brzegowe*[6]): istotność danego czynnika badamy porównując do model pełny z modelem pełnym bez tego czynnika:
   - $Y \sim B + A : B$ vs. $Y \sim A \star B$;
   - $Y \sim A + A : B$ vs. $Y \sim A \star B$;
   - $Y \sim A + B$ vs. $Y \sim A \star B$.

[4] ang. *two-way analysis of variance.*
[5] ang. *sequential.*
[6] ang. *marginal.*

W R implementujemy te testy za pomocą `summary(lm(Y~A*B))`. Ten typ testów nie zależy już od kolejności, niemniej preferuje się testowanie sekwencyjne, ponieważ w testowaniu brzegowym występują dosyć nienaturalne modele $Y \sim B + A : B$ i $Y \sim A + A : B$.

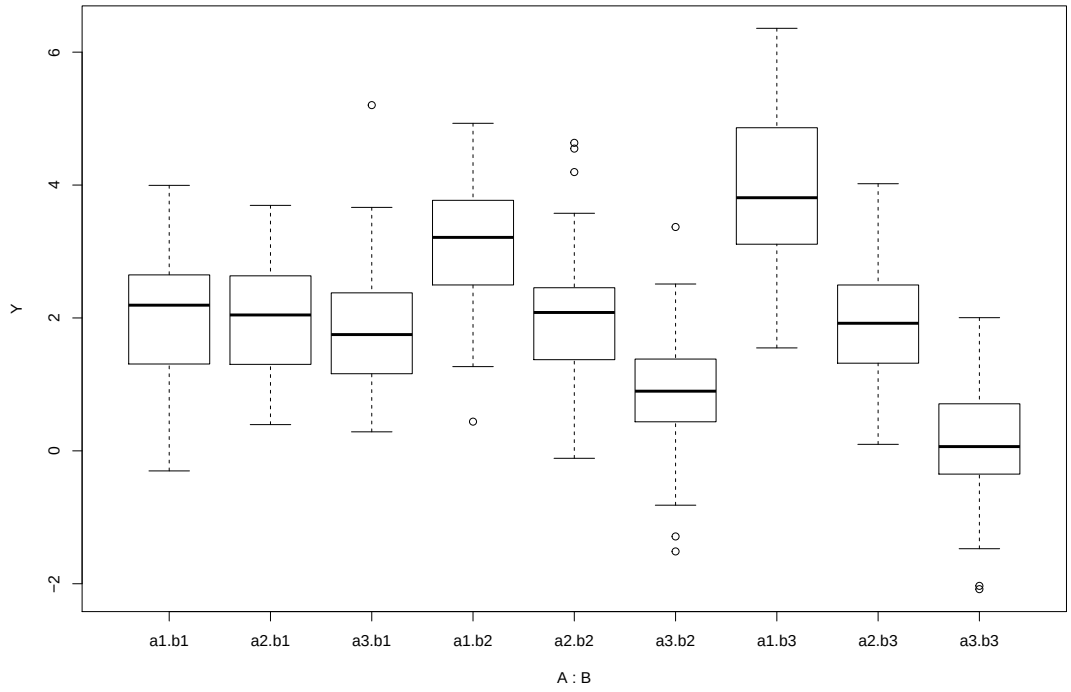
Dla przykładu rozważmy teraz (fikcyjne) dane `Dane`:

```
> head(Dane)
          Y  A  B
1 1.5409642 a1 b1
2 1.8392981 a1 b1
3 2.2129297 a1 b1
4 0.1921069 a1 b1
5 2.5333257 a1 b1
6 2.2020255 a1 b1
> attach(Dane)
```

Na początek można wyznaczyć boxploty dla zmiennej Y w każdej grupie.

```
>boxplot(Y~A*B)
```

Efektem jest Rysunek 14.1. Widzimy ‚na oko', że średnie w grupach raczej nie są równe.



Rysunek 14.1: Boxploty dla zmiennej Y w każdej grupie dla danych `Dane`.

Implementujemy analizę wariancji z testami sekwencyjnymi.

```
> model1=aov(Y~A*B)
> summary(model1)
             Df Sum Sq Mean Sq F value Pr(>F)
A             2  322.3  161.17 172.695 <2e-16 ***
B             2    0.1    0.03   0.029  0.971
A:B           4  167.5   41.88  44.870 <2e-16 ***
Residuals   441  411.6    0.93
```

Jak widać interakcja i czynnik A są istotne (pvalue mniejsze niż $2e-16$). Czynnik B nie jest istotny, ale zostawiamy go, zgodnie z zasadą, że pierwszą usuwalibyśmy interakcję.

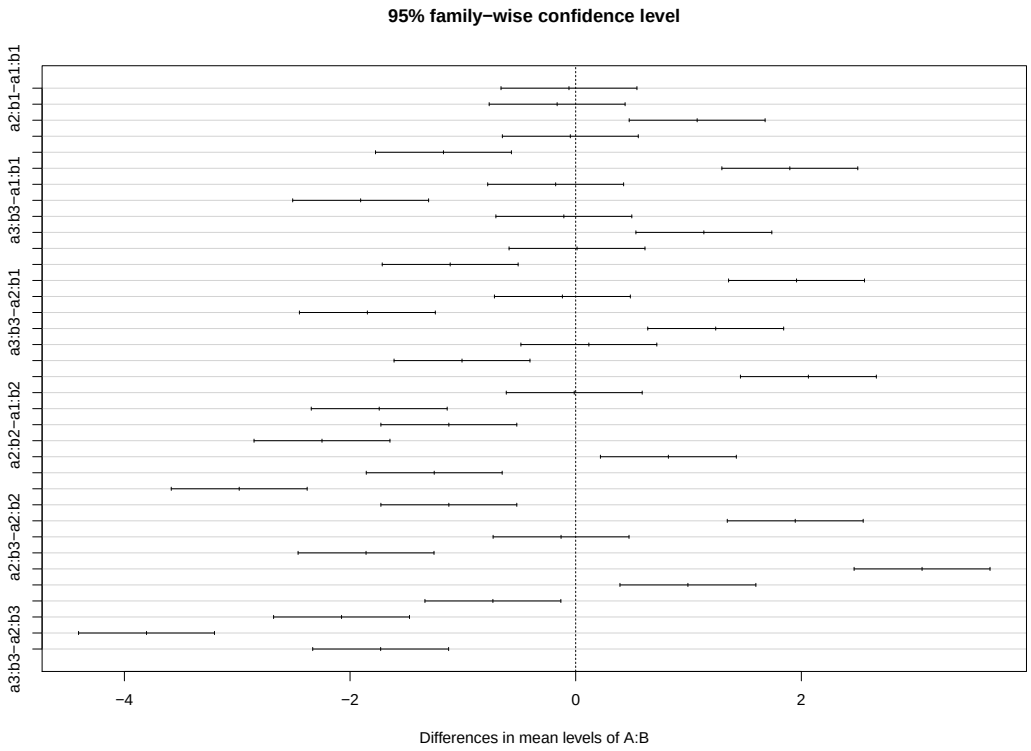
Następnie przeprowadzamy analizę post hoc.

```
> TukeyHSD(aov(Y~A*B))
  Tukey multiple comparisons of means
```

```
    95% family-wise confidence level

Fit: aov(formula = Y ~ A * B)

$A
            diff       lwr        upr p adj
a2-a1 -1.0864012 -1.348730 -0.8240728     0
a3-a1 -2.0723213 -2.334650 -1.8099929     0
a3-a2 -0.9859201 -1.248249 -0.7235917     0

$B
             diff        lwr       upr     p adj
b2-b1  0.02690879 -0.2354196 0.2892372 0.9684337
b3-b1  0.01284402 -0.2494844 0.2751724 0.9927178
b3-b2 -0.01406476 -0.2763931 0.2482636 0.9912742

$`A:B`
                     diff        lwr        upr     p adj
a2:b1-a1:b1 -0.05950649 -0.6617988  0.5427858 0.9999976
a3:b1-a1:b1 -0.16445684 -0.7667492  0.4378355 0.9951474
a1:b2-a1:b1  1.07655013  0.4742578  1.6788425 0.0000016
a2:b2-a1:b1 -0.04773962 -0.6500320  0.5545527 0.9999996
```

Jak widać wszystkie parametry α_i różnią się między sobą, natomiast wszystkie parametry β_j są statystycznie równe. Następnie są różnice wszystkich par grup (w naszym przypadku jest ich 36, powyżej nie są wszystkie wypisane). Można to też ‚zwizualizować' (rysunek 14.2):

```
> plot(TukeyHSD(aov(Y~A*B)))
```



Rysunek 14.2: Efekt porównania wszystkich par.

To samo dla czynników A i B otrzymujemy inną funkcją:

```
> print(HSD.test(model1, "A"))
          Y groups
a1 3.0465315      a
a2 1.9601303      b
a3 0.9742102      c
```

```
> print(HSD.test(model1, "B"))
          Y groups
b2 2.007282      a
b3 1.993217      a
b1 1.980373      a
```

14.2.1 ANOVA hierarchiczna

Z analizą wariancji *hierarchiczną* mamy do czynienia, gdy poziomy jednego faktora są *zagnieżdżone*[7] w poziomach drugiego (a nie *przecięte*[8] jak w poprzednim przykładzie). Zobaczmy to na przykładzie (fikcyjnych) danych `Dane2`, gdzie badamy ośmioklasistów

```
> head(Dane2)
  Wzrost Miasto Szkola
1 175.0    Wro    SP2
2 159.5     Kr    SP2
3 170.7    War    SP2
```

gdzie `Wzrost` to wzrost, a `Szkola` i `Miasto` to odpowiednio szkoła, do której ośmioklasista uczęszcza i miasto, gdzie ta szkoła się znajduje. Poniżej widzimy, że we wszystkich trzech miastach mamy szkoły o tym samym numerze 1, 2 i 3, a *SP1* w Krakowie to co innego niż *SP1* w Warszawie.

```
> by(Dane2$Szkola,Dane2$Miasto,table)
Dane2$Miasto: Kr

SP1 SP2 SP3
 10  12   9
------------------------------------------------------------
Dane2$Miasto: War

SP1 SP2 SP3
 14  13   6
------------------------------------------------------------
Dane2$Miasto: Wro

SP1 SP2 SP3
 14   8  14
```

Dlatego rozważamy tutaj model

$$Y_{ijm} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijm},$$

gdzie α_i to w naszym przypadku efekt i-ego miasta ($i \geqslant 2$), a β_{ij} to efekt j-tej szkoły w i-tym mieście, a Y_{ijm} to wzrost m-ego ośmioklasisty w j-ej szkole w i-tym mieście. W `R` implementujemy ten model następująco (z testami sekwencyjnymi)

```
> mh=aov(Wzrost~Miasto/Szkola,Dane2)
> summary(mh)
              Df Sum Sq Mean Sq F value Pr(>F)
Miasto         2    130   64.98   0.734  0.483
Miasto:Szkola  6    289   48.22   0.545  0.773
Residuals     91   8054   88.51
```

Jak widać zarówno efekt szkoły jak i miasta jest nieistotny statystycznie.

14.2.2 Kontrasty

Załóżmy, że rozważamy ANOVĘ jednoczynnikową, w której czynnik A ma k poziomów. Niech $\mu_1, \ldots, \mu_k$ oznaczają średnie wartości cechy objaśnianej w grupach wyznaczonych przez poziomy czynnika A. Wtedy *kontrastem* nazywamy kombinacje liniową tych średnich, to znaczy $c^T\mu$, dla pewnego $c \in \mathbb{R}^k$, gdzie $\mu = (\mu_1, \ldots, \mu_k)^T$.

[7] ang. *nested*.
[8] ang. *crossed*.

Analiza kontrastu polega na testowaniu hipotezy zerowej

$$H_0 : c^T \mu = 0.$$

Dla przykładu rozważmy ponownie dane `Dane`, w których czynnik A ma trzy poziomy. Ustalmy kontrast $c1 = (0, 1, -1)^T$. Wtedy hipoteza zerowa przyjmie postać $H_0 : \mu_2 = \mu_3$. W `R` implementujemy to następująco

```
> c1=c(0,1,-1)
> summary(lm(Y~A,contrasts = list(A=cbind(c1))))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.99362    0.06411  31.097  < 2e-16 ***
Ac1          0.49296    0.07852   6.278 8.09e-10 ***
```

Hipotezę zerową odrzucamy ($pvalue = 8.09e - 10$) i mówimy, że kontrast $c1$ jest istotny (niezerowy), czyli średnie w drugiej i trzeciej grupie nie są równe.

Możemy testować istotność kilku kontrastów równocześnie, na przykład

```
> c1=c(0,1,-1)
> c2=c(2,-1,-1)
> kontrast=cbind(c1,c2)
> summary(lm(Y~A,contrasts = list(A=kontrast)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.99362    0.05366  37.155  < 2e-16 ***
Ac1          0.49296    0.06572   7.501 3.44e-13 ***
Ac2          0.52645    0.03794  13.875  < 2e-16 ***
```

Teraz oba kontrasty są istotne (przy czym dla kontrastu $c2$ mamy hipotezę zerową $H_0 : \mu_1 = (\mu_2 + \mu_3)/2$).

W `R` mamy zaimplementowane najczęściej używane kontrasty, na przykład `contr.helmert`, `contr.sdif`, czy `contr.poly`. Przykładowo

```
> contr.poly(3)
                .L         .Q
[1,] -7.071068e-01  0.4082483
[2,] -7.850462e-17 -0.8164966
[3,]  7.071068e-01  0.4082483
> summary(lm(Y~A,contrasts = list(A=contr.poly(3))))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.99362    0.05366  37.155   <2e-16 ***
A.L         -1.46535    0.09294 -15.767   <2e-16 ***
A.Q          0.04102    0.09294   0.441    0.659
```

Bibliografia

[1] S. F. Arnold, *The theory of linear models and multivariate analysis,* Wiley, 1981.

[2] R. Bartels, *The rank version of von Neumann's ratio test for randomness*, Journal of the American Statistical Association, 77 (377): 40–46, 1982.

[3] J. Bartoszewicz, *Wykłady ze statystyki matematycznej,* PWN, 1989.

[4] P. Biecek, *Analiza danych z programem R,* Wydawnictwo Naukowe PWN, 2012.

[5] P. Biecek, *Przewodnik po pakiecie R,* Oficyna Wydawnicza GiS, 2017.

[6] A. C. Davison, D. V. Hinkley, *Bootstrap Methods and their Application,* Cambridge University Press, 1997.

[7] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *An Introduction to the Bootstrap,* Chapman and Hall, 1994.

[8] J. Faraway, *Linear Models with R,* Chapman and Hall, 2014.

[9] M. Friedman, *A comparison of alternative tests of significance for the problem of m rankings*, The Annals of Mathematical Statistics, 11 (1): 86–92, (1940).

[10] L. Gajek, M. Kałuszka, *Wnioskowanie statystyczne. Modele i metody*, WNT, 2000.

[11] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin, *Bayesian Data Analysis,* Chapman and Hall, 2013, `http://www.stat.columbia.edu/ gelman/book/BDA3.pdf`

[12] T, Górecki, *Podstawy statystyki z przykładami w R*, Wydawnictwo BTC, 2011.

[13] M. Hollander, D. Wolfe, E. Chicken, *Nonparametric Statistical Methods,* Wiley, 2014.

[14] M.G. Kendall, B. Babington Smith, *The Problem of m Rankings*, The Annals of Mathematical Statistics, 10 (3): 275–287, (1939).

[15] A. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione*, G. Ist. Ital. Attuari, 4: 83—91, (1933).

[16] M. Krzyśko, *Statystyka matematyczna,* Wydawnictwo naukowe UAM, 2004.

[17] E. L. Lehman, *Teoria estymacji punktowej,* Wydawnictwo Naukowe PWN, 1991.

[18] R. Magiera, *Modele i metody statystyki matematycznej,* Oficyna Wydawnicza GiS, 2018.

[19] Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika, 12 (2): 153–157, (1947).

[20] J. Neyman, *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, (1937).

[21] S. D. Silvey, *Wnioskowanie statystyczne,* PWN, 1978.

[22] N. Smirnov, *Table for estimating the goodness of fit of empirical distributions*, Annals of Mathematical Statistics, 19 (2): 279–281, 1948.

[23] M. Walesiak, E. Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R,* Wydawnictwo Naukowe PWN, 2009.

Indeks

A
analiza
kowariancji ANCOVA, 109
post hoc, 113
wariancji ANOVA
dwuczynnikowa, 115
hierarchiczna, 118
jednoczynnikowa, 112

B
badanie
częściowe, niepełne, 3
pełne, 3
błąd
estymacji, 32
I rodzaju, 35
II rodzaju, 35
kwadratowy, 94
standardowy, 94
standardowy estymatora, 74
średniokwadratowy, 26
bootstrap
nieparametryczny, 75
parametryczny, 75
wygładzający, 75
boxplot, 12

C
cecha, 3
ciągła, 5
dyskretna, 5
ilościowa, mierzalna, 5
jakościowa, niemierzalna, 5
poziom, kategoria, 5
quasi-ilościowa, 5
cechy
sparowane, 41
częstość
empiryczna, 45
spodziewana, 45

D
dominanta, 8
dystrybuanta
empiryczna, 67
dźwignia, 101

E
estymacja
przedziałowa, 17
punktowa, 17
estymator, 17
bayesowski punktowy, 78
jądrowy, 69
momentów, 23
najmniejszych kwadratów, 93
największej wiarygodności, 18
nieobciążony, 26
asymptotycznie, 26
o minimalnej wariancji, 27
parametru, 25
zgodny
mocno, 25
słabo, 25

F
funkcja
generująca momenty, 84
log-wiarygodności, 18
RSS, 93
wiarygodności, 18

H
hipoteza
alternatywna, 35
dwustronna, 36
lewostronna, 36
prawostronna, 36
testowanie, 17, 35
zerowa, 35
histogram, 13, 68

I
informacja
Fishera, 27

J
jądro
estymatora, 69
gaussowskie, 69
optymalne, Epanecznikowa, 69
prostokątne, 69
trójkątne, 69

K
klasa, 13
liczność, 13
szerokość, 68
kontrast, 118
kryterium
Akaika AIC, 107
GIC, 107
Schwartza BIC, 107
kurtoza, 11
kwantyl
z próby, 9
kwartyl
dolny, 9
górny, 9

L
liczność
próby, 3

M
macierz
daszkowa, 94
kowariancji (wariancji), 83
symetryczna, 82
dodatnio określona, 82
nieujemnie określona, 82
mediana
z próby, 8
metoda
analityczna, 97
Boxa–Coxa, 105
graficzna, 97
krokowa, 108
krokowa w przód, 107
krokowa w tył, 108
największej wiarygodności, 18
miara
położenia, 7
centralnego, 7
skośności, 10
zmienności, 9
moda, 8
model
liniowy, 90
parametry strukturalne, 93
pusty, 95
składnik losowy, 93
składnik systematyczny, 93
zagnieżdżony, 95
moment
empiryczny, 23
teoretyczny, 23

O
obciążenie, 26
publikacyjne, 36
obserwacja
obciętych, 19
odstająca, 12, 101
wpływowa, 101
odchylenie
od średniej, 7
przeciętne, 10
standardowe, 10
odległość Cooka, 103
osobnik, 3

P
paradoks Simpsona, 58
pasmo
szerokość, 13, 68
populacja, 3
poziom
istotności, 35
prawdopodobieństwo
subiektywne, 78
teoretyczne, 45
próba, 3
leptokurtyczna, 11
liczność, 3
mezokurtyczna, 11
platokurtyczna, 11
prosta, 4, 17
reprezentatywna, 4
skośna, 10
symetryczna, 10
przedział
ufności, 29
dwustronny, 30
normalny, 75
percentylowy, 75
wiarygodności, 78
ETI, 79
HDI, 78
przestrzeń
parametrów, 17
statystyczna, 17

R
ranga, 50
wiązana, 50
rangowanie, 4, 50
reszta
standaryzowana, 102
studentyzowana, 102
reszty obserwowalne, 94
rozkład
a posteriori, 78
a priori, 78
Beta, 77
χ^2, 33

F, 42
mieszany, 24
normalny
sferyczny, 85
wielowymiarowy, 84
próby, 13
t-Studenta, 30
rozstęp, 9
międzykwartylowy, 9
rzutowanie ortogonalne, 87

S
scałkowany błąd średniokwadratowy MISE, 67
skala
ilorazowa, 5
nominalna, 4
pomiarowa, 4
porządkowa, 4
przedziałowa, 4
statystyka
C_p Mallowsa, 107
czułość na wartości odstające, 8
Kendalla z próby, 61
matematyczna, 3
odporna, 8
opisowa, 3
testowa, 35
stochastyczna
nierówność $\leqslant^S$, $\leqslant^S$, 50
stopień
swobody, 94
suma kwadratów
całkowita, SST, 97
regresyjna, SSR, 97
szereg
rozdzielczy, 13

Ś
średnia
arytmetyczna z próby, 7
obcięta, 8
populacji, 7
winsorska, 8
środek
ciężkości, 7

T
tablica
wielodzielcza, kontyngencji, 46
test
Bartletta, 114
Browna–Forsytha, 114
brzegowy, 115
Fishera
dokładny, 47
Flignera–Killeena, 114
Harveya–Colliera, 100
HSD, 113
jednorodności, 51, 54
Kołmogorowa–Smirnowa, 54
Levene'a, 114
losowości serii Walda–Wolfowitza, 50
LSD Fishera, 113
moc, 35
permutacyjny, 72
Rainbow, 100
RESET, 100
sekwencyjny, 115
Studenta–Newmana–Keulsa, 113

W
wariancja
z próby, 9
wartości
odstające, 8
wartości dopasowane, 94
wartość oczekiwana, 83
wnioskowanie
nieparametryczne, 4, 17
parametryczne, 4, 17
statystyczne, 3
współczynnik
dopasowania, 97
korelacji
Kendalla z próby, 61
liniowej Pearsona, 16
Spearmana, 58
τ-Kendalla, 61
R^2 zmodyfikowany, 106
skośności, 11
kwartylowy, 11
Pearsona, 10
zmienności, 10
wykres
kwantylowy, *QQ plot*, 98
pudełko z wąsami, *boxplot*, 12

Z
zbiór
krytyczny, odrzuceń, 35
zmienna, 3
objaśniająca, niezależna, 92
objaśniana, zależna, 92