

Statystyka opisowa

Katedra Statystyki

Contents

Wprowadzenie	5
1 Zbiorowość, jednostka, cecha	7
1.1 Zbiory danych	11
2 Wizualizacja danych	13
2.1 Rodzaje rozkładów	13
2.2 Szeregi statystyczne	17
2.3 Dystrybuanta	17
2.4 Histogramy	19
3 Analiza struktury	27
3.1 Miary klasyczne	28
3.2 Miary pozycyjne	34
3.3 Szereg jednostkowy i przedziałowy	38
3.4 Podsumowanie miar	41
3.5 Przedziały ufności	46
3.6 Testy statystyczne	51
4 Korelacje	57
4.1 Cechy jakościowe	57
4.2 Cechy ciągłe	61
5 Regresja	65
5.1 Regresja prosta	65
5.2 Trend liniowy	75
6 Sezonowość	79
6.1 Trend liniowy	79
6.2 Model addytywny	82
6.3 Model multiplikatywny	85
6.4 Ocena jakości	87
6.5 Błąd prognozy	87

7	Analiza szeregu dynamicznego	89
7.1	Przyrosty absolutne	89
7.2	Przyrosty względne	90
7.3	Indeksy	91
7.4	Średnie tempo zmian	93
8	Indeksy indywidualne i zespołowe	95
8.1	Indeksy indywidualne	95
8.2	Indeksy zespołowe (agregatowe)	96
8.3	Przykład	98

Wprowadzenie

Niniejszy skrypt stanowi pomoc dydaktyczną do przedmiotów Statystyka oraz Statystyka opisowa. Celem przedmiotu jest poznanie metod statystycznych, ich zastosowań oraz interpretacja otrzymanych wyników.

Chapter 1

Zbiorowość, jednostka, cecha

Dane, dane, dane! - wołał niecierpliwie - Nie mogę lepić cegieł nie mając gliny.

– Przygoda w Copper Beeches, Artur Conan Doyle

Statystyk do pracy potrzebuje danych. Te mogą być pozyskiwane na wiele różnych sposobów. Najpopularniejszym sposobem zbierania danych są badania ankietowe. Mogą być realizowane praktycznie przez każdego - osoby fizyczne (np. studentów), firmy, ośrodki badania opinii publicznej, administrację rządową, itd. W Polsce podmiotem odpowiedzialnym za przeprowadzanie badań społecznych jest Główny Urząd Statystyczny. Dzięki prowadzonym badaniom ankietowym dostarcza informacji na temat praktycznie każdej dziedziny życia obywatela Polski. Przykładowym badaniem prowadzonym przez GUS jest Badanie Aktywności Ekonomicznej Ludności. Zebranie informacji na temat profilu społeczno-demograficznego rynku pracy kosztuje rocznie ponad 17 milionów złotych (przygotowanie ankiet, zebranie danych, opracowanie wyników, druk publikacji). Próba badania to około 100 tys. osób czyli niecałe 3 promile całej populacji Polski.

Innym źródłem danych są rejestry administracyjne, które gromadzą mniej danych, ale na temat całej zbiorowości - np. PESEL.

Poza tym ogromne ilości danych gromadzą firmy prywatne - banki, telekomy, towarzystwa ubezpieczeniowe, portale społecznościowe. Przykładowo dane Facebooka to 300 petabajtów z dziennym przyrostem rzędu 600 TB (dane z roku 2014).

W odniesieniu do każdego zbioru danych można zdefiniować zbiorowość statystyczną, jednostkę statystyczną oraz cechy statystyczne - stałe i zmienne.

Zbiorowość statystyczna - zbiór elementów objętych badaniem statystycznym, posiadających co najmniej jedną cechę stałą (wspólną) oraz co najmniej jedną cechę zmienną (różniące te elementy między sobą). Przykładowo, ankietowano mieszkańców Poznania (cecha stała ponieważ wszyscy respondenci byli mieszkańcami Poznania) pod kątem sympatii do prezydenta (cecha zmienna ponieważ mieszkańcy udzielali różnych odpowiedzi). Zbiorowością będą oczywiście mieszkańcy Poznania.

Jednostka statystyczna - pojedynczy element zbiorowości. W wyżej przedstawionym przykładzie jednostką będzie mieszkaniec Poznania.

Z kolei cechy zmienne dzielą się na:

- czasowe, np. rok urodzenia
- przestrzenne, np. miejsce zamieszkania
- rzeczowe
 - niemierzalne (jakościowe)
 - * dychotomia, np. płeć
 - * politomia, np. wykształcenie
 - mierzalne (ilościowe)
 - * ciągłe, np. wzrost, waga, wiek, średnia ocen
 - * skokowe, np. liczba rodzeństwa, liczba pokoi

Wyżej wprowadzone definicje najlepiej zobrazuje przykład ankiety.

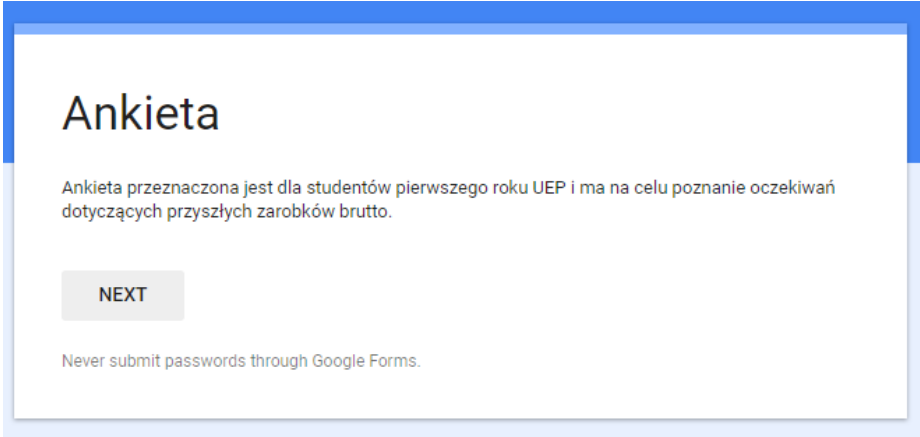


Figure 1.1: Wprowadzenie do ankiety

Na podstawie zdania wprowadzającego do ankiety możemy zdefiniować zbiorowość oraz jednostkę. Zbiorowością są studenci pierwszego roku UEP, co jest ich cechą stałą (wspólną). Jednostką jest z kolei student pierwszego roku UEP - pojedynczy ankietowany.

Cechy zmienne można z kolei przypisać do kolejnych pytań w ankiecie:

Ankieta

Dane metryczkowe

Płeć

☐ Kobieta

☐ Mężczyzna

Miejsce urodzenia

Your answer

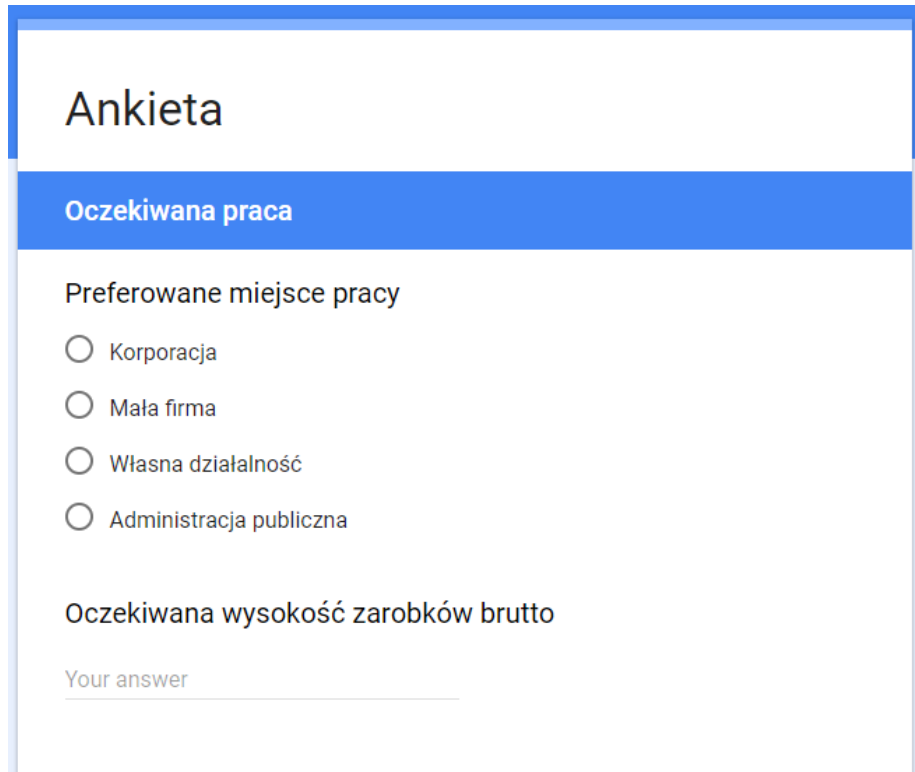
Data urodzenia

Date

dd.mm.rrrr

Figure 1.2: Pytania metryczkowe

Płeć jest przykładem cechy rzeczowej, jakościowej i dychotomicznej. Data urodzenia to cecha czasowa, a miejsce urodzenia jest cechą przestrzenną.



Ankieta

Oczekiwana praca

Preferowane miejsce pracy

☐ Korporacja

☐ Mała firma

☐ Własna działalność

☐ Administracja publiczna

Oczekiwana wysokość zarobków brutto

Your answer

Figure 1.3: Kolejne pytania

Preferowane miejsce pracy jest cechą rzeczową, jakościową i politomiczną, natomiast wysokość oczekiwanych zarobków brutto to cecha rzeczowa, ilościowa, ciągła.

Zadania

Zdefiniuj jednostkę oraz 3 przykładowe cechy wraz z rodzajem dla podanych zbiorowości:

1. użytkownicy Facebooka
2. gospodarstwa domowe
3. przedsiębiorstwa budowlane
4. klienci banku
5. drzewa w parku Sołackim

1.1 Zbiory danych

Podczas zajęć będziemy korzystać z dwóch zbiorów danych:

- dane sprzedażowe w sklepach Rossmann w Niemczech w 2014 roku - plik
- dane dotyczące sprzedaży piwa 0,5 l w 2011 roku - plik

Rossmann

Zmienne

- sklep_id - identyfikator sklepu
- dzien_tyg - liczba oznaczająca dzień tygodnia
- data - data w formacie rok-miesiąc-dzień
- sprzedaz - dzienny obrót (w euro)
- liczba_klientow - liczba klientów
- czy_otwarty - czy sklep był otwarty danego dnia
- czy_promocja - czy danego dnia rozpoczęła się promocja
- czy_swieto - czy sklep był zamknięty w dane święto
- czy_swieto_szkolne - czy na sklep miały wpływ dni wolne od szkoły
- sklep_typ - typ sklepu
- sklep_asort - dostępny asortyment
- sklep_konkurencja - odległość w metrach do najbliższego konkurencyjnego sklepu

Piwa

Zmienne

- miesiac - miesiąc zakupu
- poj - pojemność piwa
- wojewodztwo - województwo
- sklepy - rodzaj sklepu
- browar - pochodzenie piwa
- cena - jednostkowa cena piwa
- sztuki - liczba zakupionych sztuk piwa

Zadania

Z wykorzystaniem narzędzia tabel przestawnych w Excelu (wstążka “Wstawianie” -> Tabele przestawne) określ rodzaj cechy dla każdej zmiennej.

Chapter 2

Wizualizacja danych

Bardzo istotną częścią statystyki jest wizualizacja wyników. Poprawne korzystanie z wykresów wymaga poznania kilku, czasami nieoczywistych, zasad:

- efekt 3D na wykresach zaburza percepcję i utrudnia porównywanie danych,
- wykresy liniowe służą przede wszystkim do prezentacji zmian w czasie,
- ...

Zasoby internetowe są pełne przykładów i wzorców tworzenia wykresów:

- Graficzna prezentacja danych statystycznych - Wykresy, mapy, GIS
- Zbiór esejów o wizualizacji danych
- Flowing Data
- D3

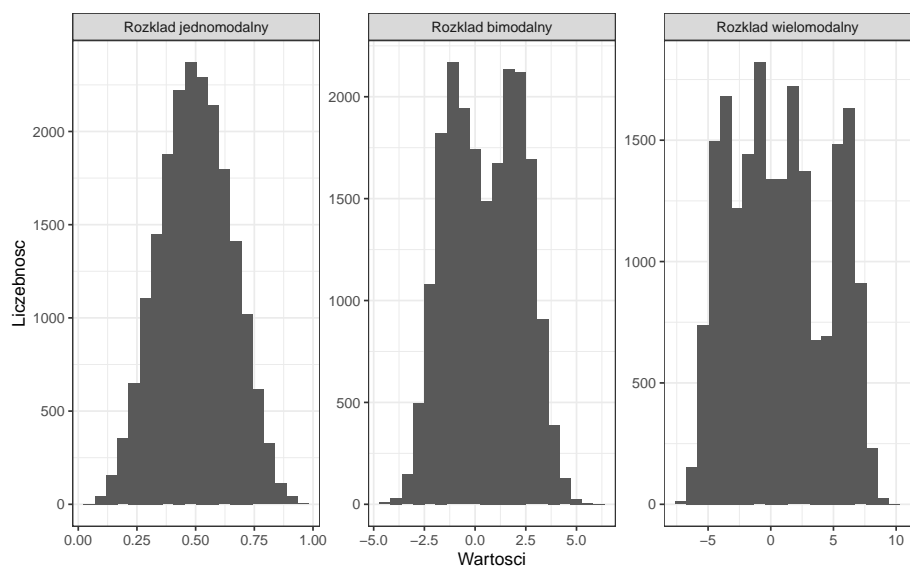
W analizie statystycznej bardzo ważne jest pojęcie rozkładu cechy.

Rozkładem empirycznym cechy nazywamy przyporządkowanie kolejnym wartościom zmiennej (x_i) odpowiadającym im liczebności (n_i). Rozkład odzwierciedla strukturę badanej zbiorowości z punktu widzenia określonej cechy.

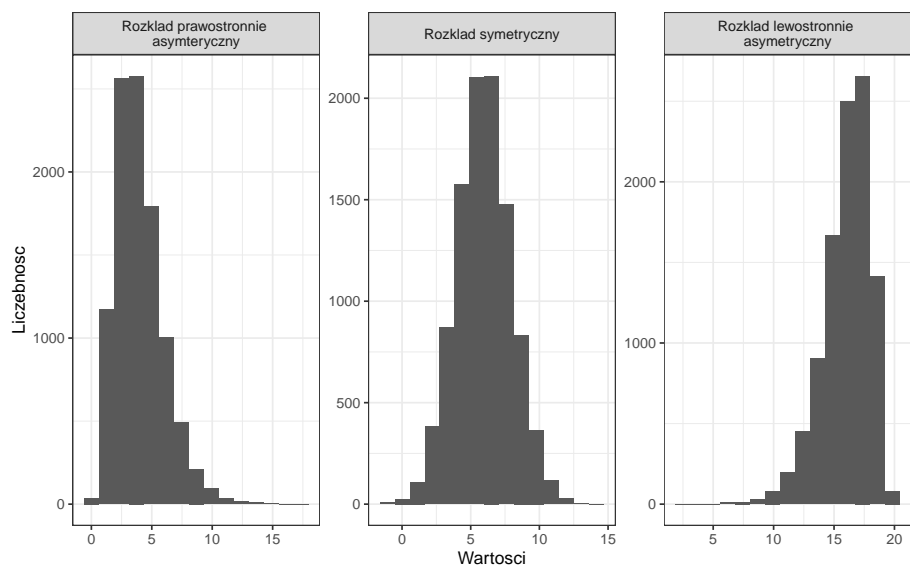
Najdogodniejszym sposobem graficznej prezentacji rozkładu jest histogram przedstawiający częstość poszczególnych kategorii. Histogram można utworzyć na podstawie tabeli przestawnej. W sytuacji kiedy nie wszystkie wartości są reprezentowane histogram może wyglądać dziwnie, dlatego stosuje się grupowanie wartości. W Excelu jest to możliwe z wykorzystaniem funkcji CZĘSTOŚĆ. Funkcja ta oblicza rozkład częstości występowania wartości w zakresie wartości (działa po zaznaczeniu odpowiedniego zakresu z~kombinacją CTRL+SHIFT+ENTER). Prawe przedziały są domknięte.

2.1 Rodzaje rozkładów

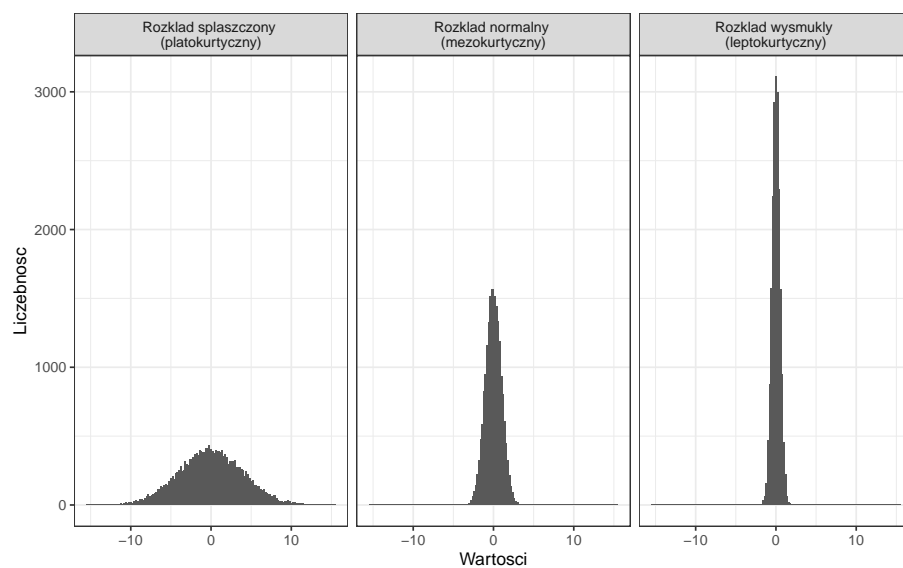
Ze względu na liczbę punktów ekstremalnych:



Ze względu na rodzaj zmienności:

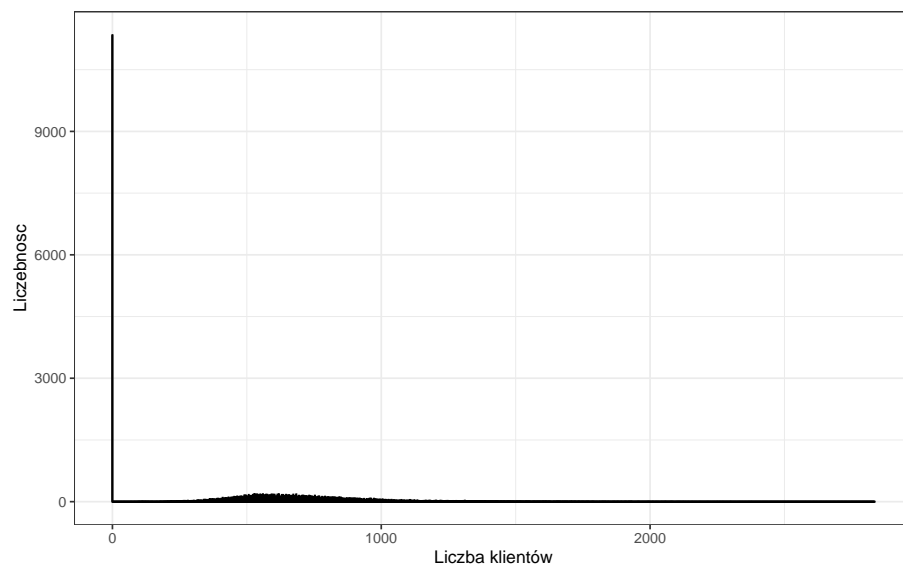


Ze względu na skupienie wokół średniej:

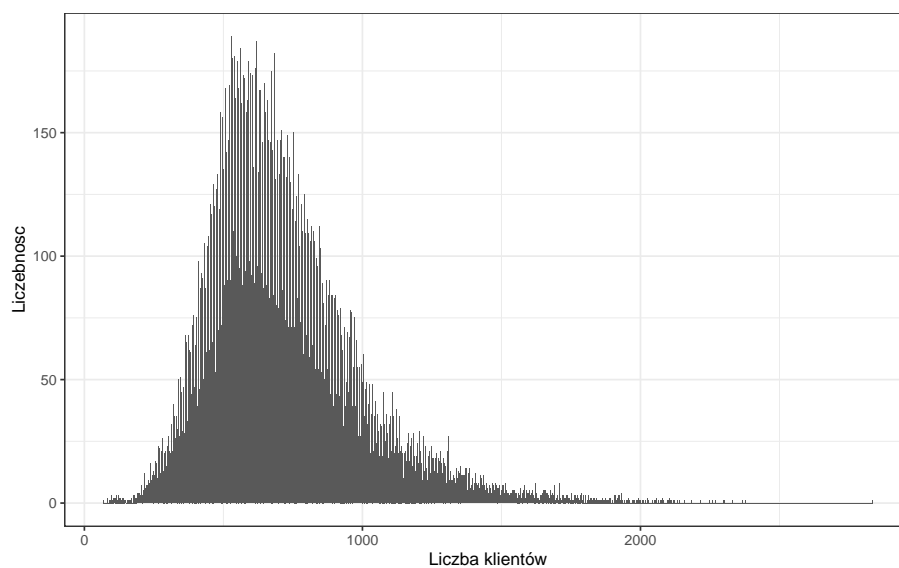


W ten sposób możemy opisywać histogramy, natomiast w dalszej części zajęć dowiemy się jakie miary definiują te cechy.

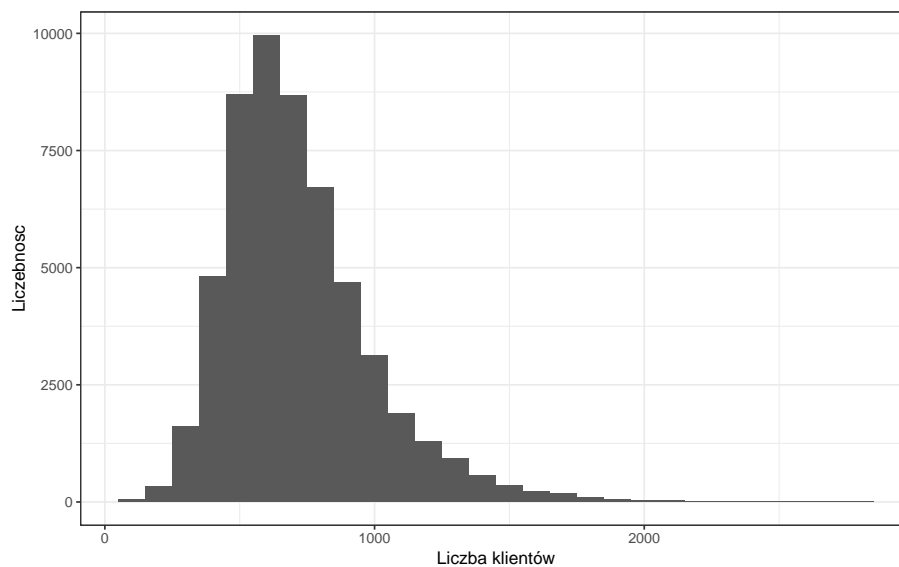
Przejdźmy do naszego zbioru danych i przeanalizujmy liczbę klientów.



Okazuje się, że występuje bardzo dużo wartości równych 0, wynikających z obserwacji dni, w których sklep był zamknięty. Musimy wyeliminować zera z naszych danych.



Obecnie rozkład liczby klientów jest dużo bardziej czytelny. Niemniej możemy zauważyć wiele wartości, które występują częściej od pozostałych. Zgrupujmy słupki w przedziałach o rozpiętości 100 klientów.



Najliczniejszą kategorią stanowią dni, w których liczba klientów pochodziła z przedziału 500-600 osób. Można także zaobserwować, że rozkład charakteryzuje się asymetrią prawostronną.

2.2 Szeregi statystyczne

Przeprowadzając powyższe grupowanie utworzyliśmy kilka rodzajów szeregów statystycznych.

Szereg statystyczny jest to ciąg wielkości statystycznych usystematyzowanych według określonego ściśle kryterium. Powstaje on w wyniku grupowania bądź porządkowania. Stanowi podstawę dla prowadzenia numerycznej analizy statystycznej.

- **szereg prosty** to wykaz wszystkich wariantów badanej cechy np. liczba klientów dla każdego sklepu danego dnia
- **szereg rozdzielczy jednostkowy (punktowy)** wykaz wariantów cechy i liczebności poszczególnego wariantu np. szereg utworzony z wykorzystaniem tabeli przestawnej
- **szereg rozdzielczy przedziałowy zamknięty o równych przedziałach klasowych** wykaz zgrupowanych wariantów cechy i liczebności poszczególnych wariantów np. szereg utworzony z wykorzystaniem funkcji CZĘSTOŚĆ
- **szereg rozdzielczy przedziałowy zamknięty o nierównych przedziałach klasowych**
- **szereg rozdzielczy przedziałowy otwarty** - kiedy w pierwszej lub/i ostatniej grupie znajduje się przedział otwarty (zwrot poniżej/powyżej)

2.3 Dystrybuanta

Kolejnym zagadnieniem związanym z rozkładem cechy jest **dystrybuanta**.

Dystrybuanta empiryczna to funkcja ukazująca skumulowany rozkład cechy w n -elementowej zbiorowości. Funkcję $F(x)$ definiuje się jako skumulowane prawdopodobieństwo wystąpienia - tj. sumę prawdopodobieństw od danego przedziału klasowego w rozkładzie empirycznym badanej cechy. Wyraża się wzorem:

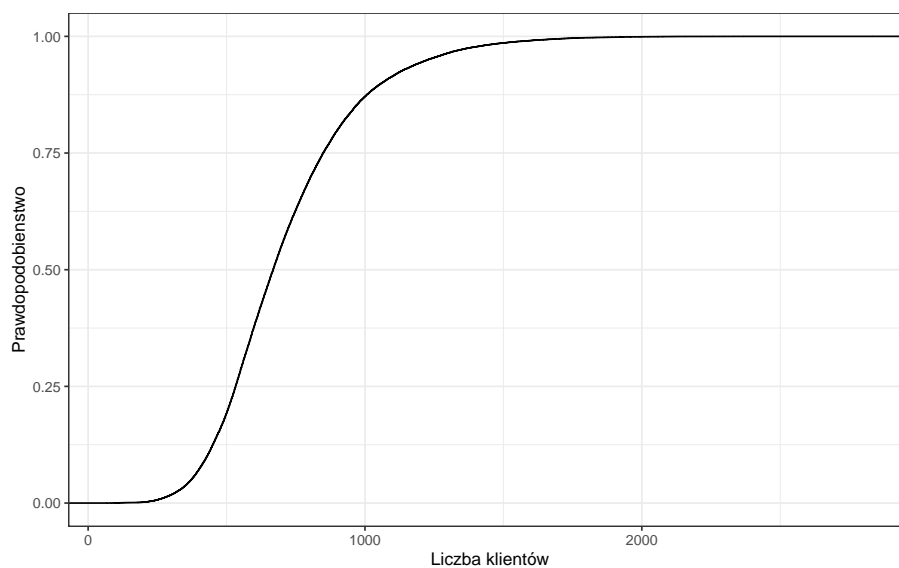
$$F(x) = \sum_{x_i < x} p_i,$$

gdzie: p_i — prawdopodobieństwo wystąpienia wariantu.

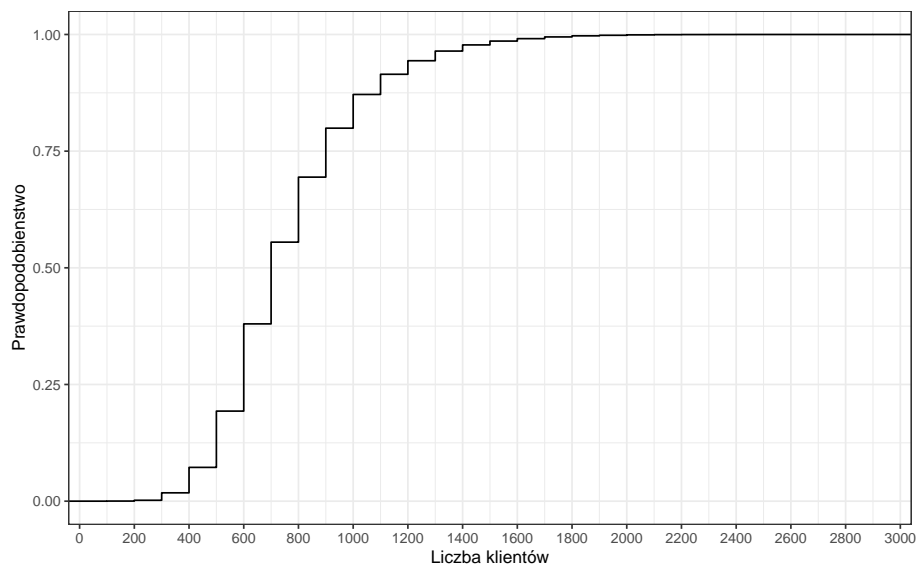
Dystrybuanta empiryczna jest funkcją:

- niemalejącą,
- lewostronnie ciągłą.

Dystrybuanta ciągła liczby klientów



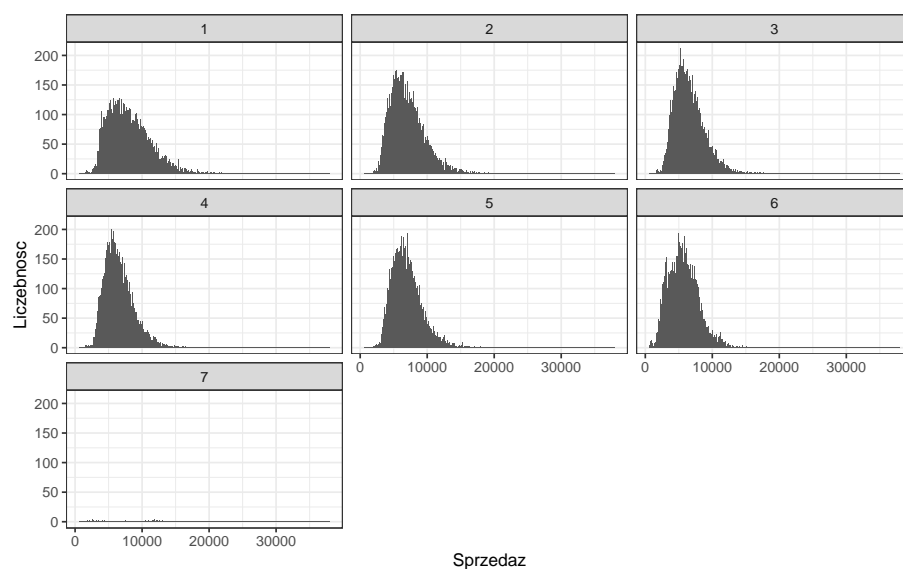
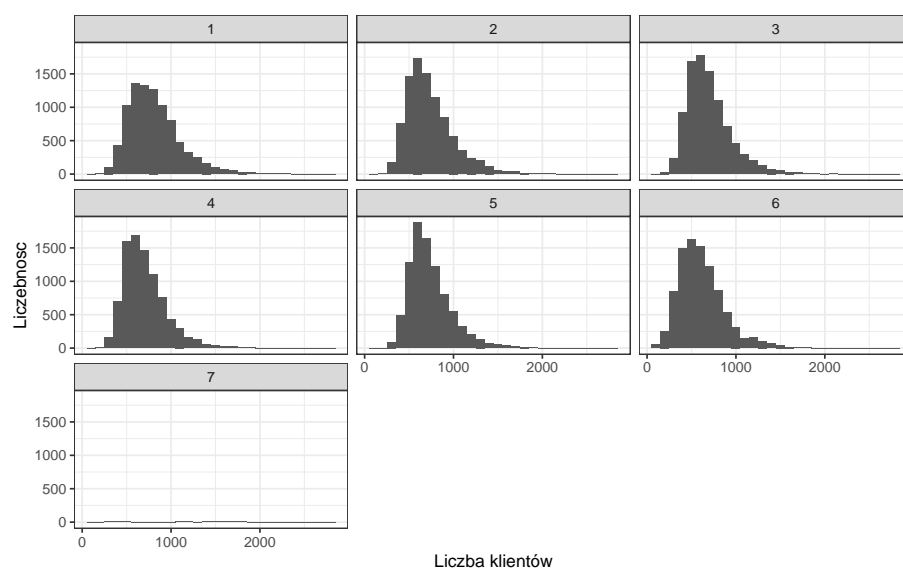
Dystrybuanta skokowa liczby klientów



Przykładowo prawdopodobieństwo, że wystąpi dzień, w którym sklep obsłuży do 700 klientów wynosi 55%.

Zadania

Z wykorzystaniem histogramu lub innych poznanych podczas zajęć metod określ w jaki dzień tygodnia sklepy Rossmann odwiedza najwięcej klientów.



2.4 Histogramy

Jak utworzyć histogram?

Aby utworzyć histogram musimy określić szerokość przedziału. Istnieje wiele metod tworzenia tych przedziałów, najczęściej występujące przedstawiam poniżej.

Poniżej przedstawiam tabelę z rozkładem zmiennej, którą badamy

minimum	q1	median	mean	q3	maximum
71.90225	93.71676	100.0921	100.1613	106.646	132.4104

2.4.1 Metoda arbitralnie ustalonej szerokosci przedziałów

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil,$$

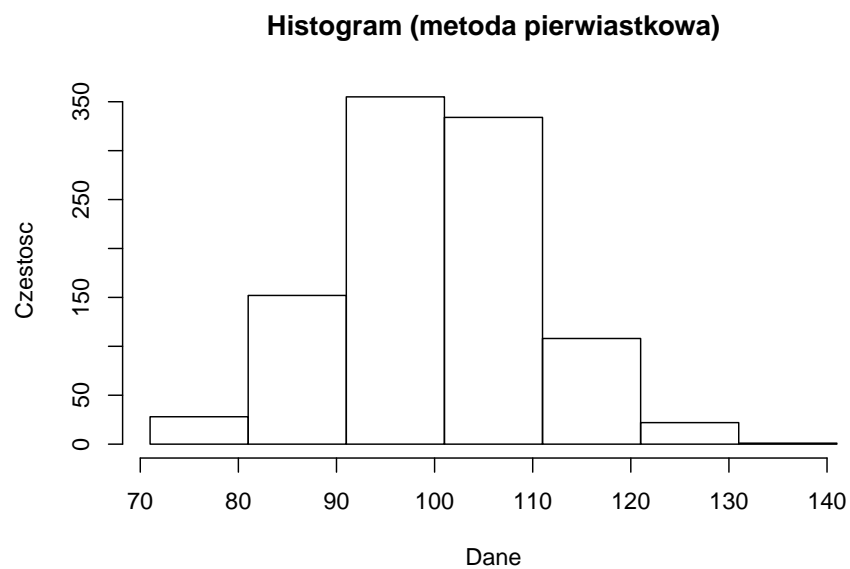
gdzie x oznacza badaną cechę ciągłą, h oznacza arbitralnie ustawioną szerokość przedziałów, a k liczba przedziałów.



2.4.2 Metoda pierwiastkowa

$$k = \sqrt{n}$$

gdzie n oznacza liczbę obserwacji.

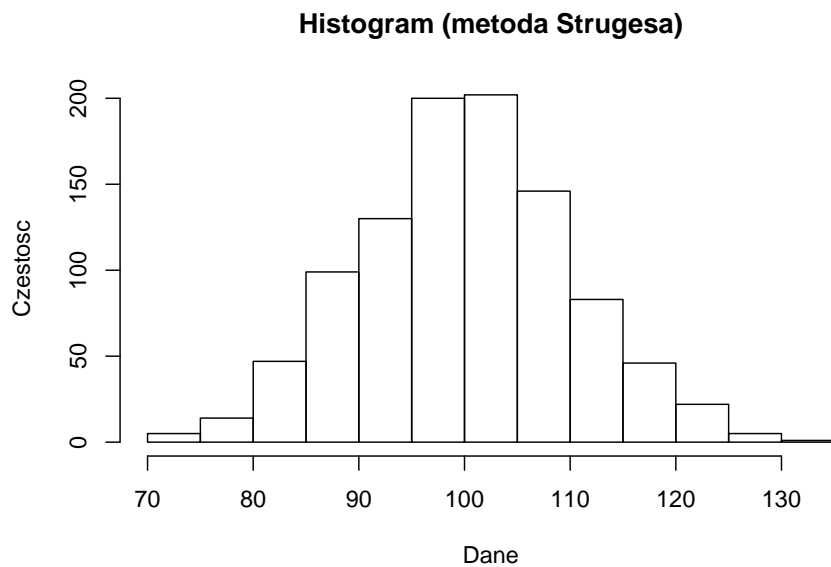


2.4.3 Metoda Struges'a

Sturges, H. A. (1926). The choice of a class interval. Journal of the american statistical association, 21(153), 65-66. [LINK](#)

$$k = \lceil \log_2 n \rceil + 1$$

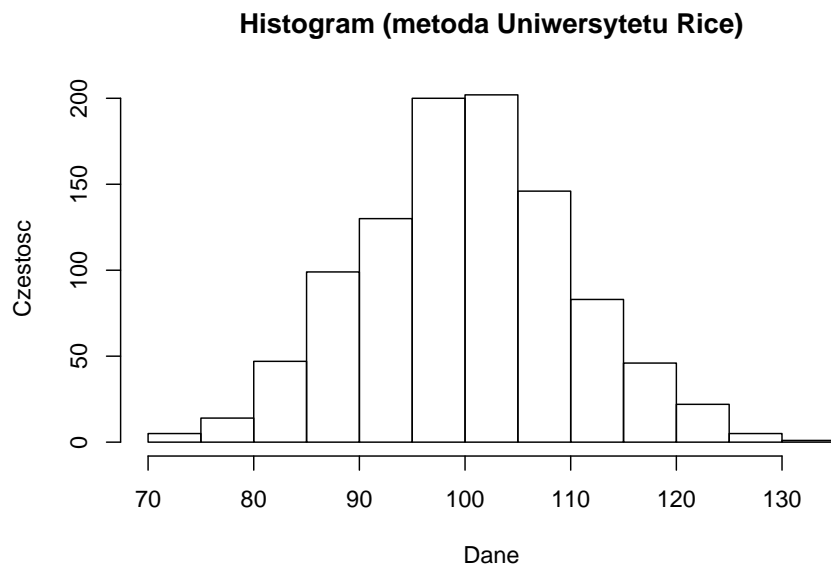
gdzie $\lceil \bullet \rceil$ oznacza sufit lub cecha g6rna liczby rzeczywistej.



UWAGA: Tej metody nie należy stosować dla małych prób (np. $n < 30$) oraz asymetrycznych rozkładów (tj. innych niż rozkład normalny). Zachęcam do zapoznania się z dyskusją na temat tej metody: <http://www.robjhyndman.com/papers/sturges.pdf>

2.4.4 Metoda Uniwersytetu Rice

$$k = \lceil 2n^{1/3} \rceil$$

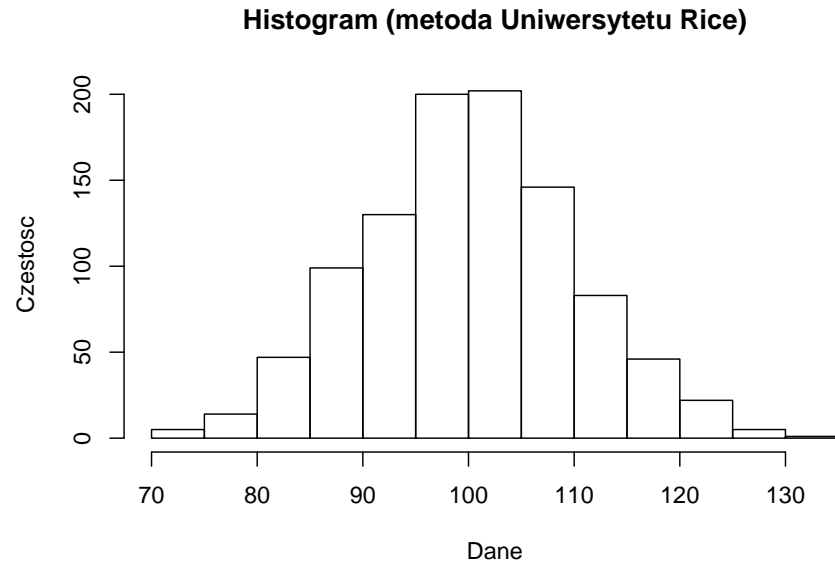


2.4.5 Metoda Doane'a

Doane, D. P. (1976). Aesthetic frequency classifications. The American Statistician, 30(4), 181-183. [LINK](#)

$$k = 1 + \log_2(n) + \log_2 \left(1 + \frac{|g_1|}{\sigma_{g_1}} \right)$$

gdzie g_1 oznacza trzeci moment (asymetrię), a $\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$.

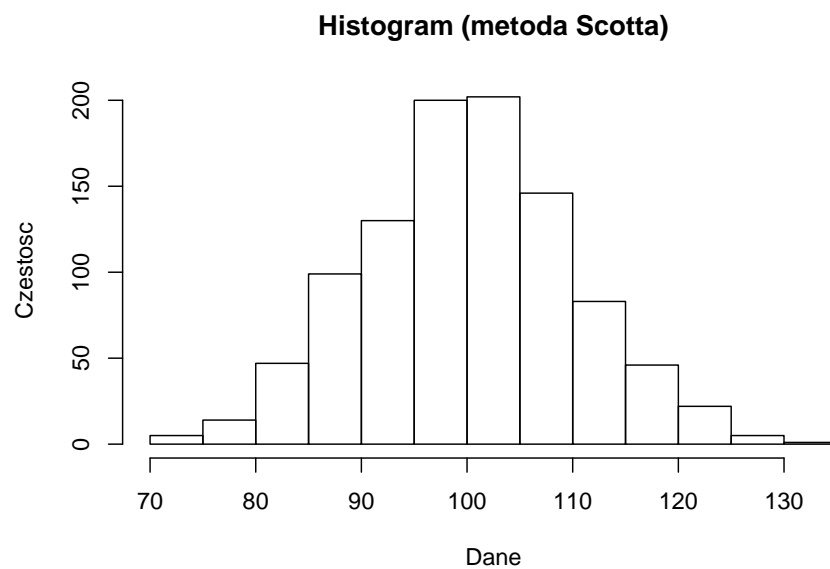


2.4.6 Metoda Scott'a

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 605-610. [LINK](#)

$$h = \frac{3.5 s(x)}{n^{1/3}}$$

gdzie $s(x)$ to odchylenie standardowe (np. z próby).

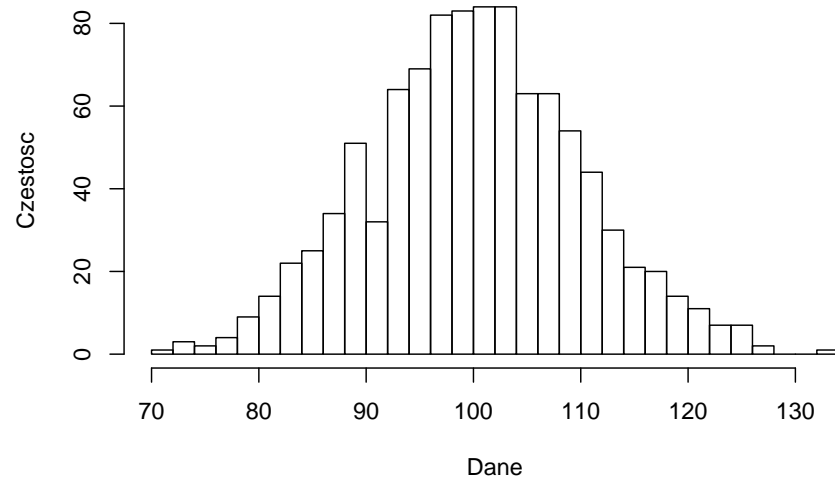


2.4.7 Metoda Freedmana–Diaconis’a

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4), 453-476. [LINK](#)

$$h = 2 \frac{R(x)}{n^{1/3}}$$

gdzie $R(x) = Q3 - Q1$, a $Q1$ oznacza kwartył pierwszy, a $Q3$ kwartył trzeci.

Histogram (metoda Freedmana–Diaconisa)

Chapter 3

Analiza struktury

Kompleksowa analiza struktury oznacza wyczerpujący opis cech zbiorowości statystycznej. Do charakterystyk najczęściej wykorzystywanych przy opisie struktury zbiorowości należą:

- miary przeciętne - służące do określania tej wartości zmiennej opisanej przez rozkład, wokół której skupiają się pozostałe wartości zmiennej,
- miary rozproszenia (dyspersji) - służące do badania stopnia zróżnicowania wartości zmiennej,
- miary asymetrii - służące do badania asymetrii rozkładu,
- miary koncentracji - służące do analizy stopnia skupienia poszczególnych jednostek wokół średniej.

Analiza struktury bazuje na dwóch typach miar:

- miary klasyczne - obliczane na podstawie wszystkich obserwacji,
- miary pozycyjne - wartość miary wskazuje dana jednostka.

Celem analizy struktury jest dostarczenie kilku liczb, które w łatwy sposób pozwolą na opis i porównania badanych cech.

Dominanta czyli najczęściej występująca wartość. Inaczej moda, modalna, tryb (w Excelu - kalka językowa z angielskiego słowa *mode*. Wartość dominanty można ustalić jedynie dla rozkładów jednomodalnych.

W Excelu jest funkcja:

- WYST.NAJCZĘŚCIEJ,

jednak dla rozkładów wielomodalnych zwróci ona pierwszą modalną.

3.1 Miary klasyczne

Najpopularniejszym przedstawicielem miar klasycznych jest **średnia arytmetyczna**. Wyrażona jest wzorem:

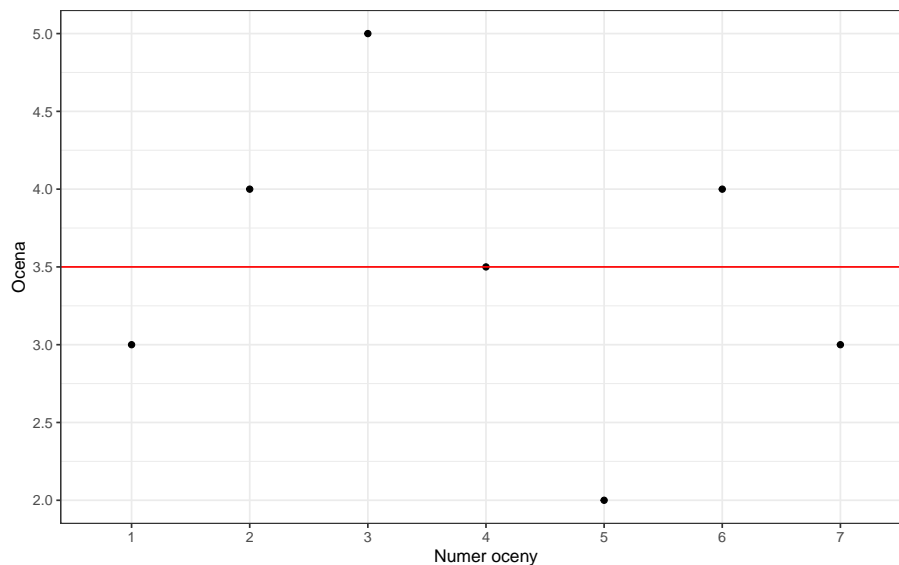
$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N},$$

gdzie:

- \bar{x} - symbol średniej arytmetycznej,
- x_i - wariant cechy mierzalnej,
- N - liczebność badanej zbiorowości.

Co sprawia, że średnia jest tak powszechną i uniwersalną miarą? Jest to liczba, która ma najwięcej wspólnego z każdą wartością cechy w zbiorowości. Innymi słowy, odległość wartości cechy od średniej jest najmniejsza z możliwych.

Przykładowo, dane są oceny jednego ze studentów: 3, 4, 5, 3+, 2, 4, 3



Na powyższym wykresie punkty oznaczają kolejne oceny, natomiast średnia została zaznaczona kolorem czerwonym - wynosi ona 3,5.

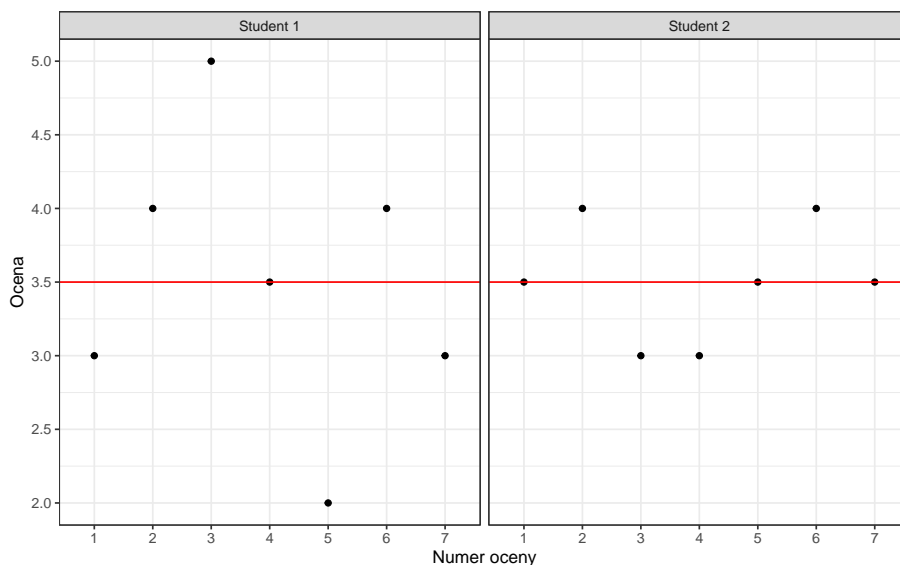
Jako miarę odległości poszczególnych ocen od średniej możemy przyjąć wartość bezwzględną różnicy danej oceny i średniej. W tej sytuacji pierwsza ocena różni się od średniej o 0,5, druga ocena także, natomiast trzecia o 1,5, itd. Po zsumowaniu tych wartości otrzymujemy sumę odchyłeń równą 5. Jest to najmniejsza wartość jaką jesteśmy w stanie otrzymać. Jeżeli stwierdzimy, że w

naszym mniemaniu wartość 3,55 jest lepszą miarą przeciętną to suma odchyleń będzie już większa i wyniesie 5.05.

W Excelu istnieje funkcja:

- ŚREDNIA.

Średnia stanowi także dobrą miarę jeśli chcemy porównać jakieś grupy. Co jednak zrobić w sytuacji, kiedy przykładowo dwaj studenci mają identyczne średnie ocen? Czy to oznacza, że ich oceny są także takie same? Taka sytuacja może się zdarzyć, ale występuje dosyć rzadko. Poniżej zostały przedstawione oceny dwóch studentów, którzy mają identyczną średnią.



To co możemy zauważyć gołym okiem to fakt, że oceny studenta nr 2 są bliżej średniej. Miarą zróżnicowania cechy jest **wariancja** dana formułą:

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

gdzie:

- s^2 - symbol wariancji,
- \bar{x} - średnia arytmetyczna w zbiorowości,
- x_i - wariant cechy mierzałnej,
- N - liczebność badanej zbiorowości.

Jeżeli przeanalizujemy wzór na wariancję jest on bardzo logiczny. W pierwszym kroku liczymy odchylenia wartości cechy od średniej. Następnie otrzymane

wartości podnosimy do kwadratu w celu uniknięcia wartości ujemnych, a następnie wszystko uśredniamy. Możemy zatem powiedzieć, że wariancja jest średnią kwadratów odchyleń wartości od średniej.

Wariancja ocen pierwszego studenta wynosi 0.79, natomiast drugiego 0.14. Na podstawie tej miary jesteśmy w stanie stwierdzić, że większe zróżnicowanie ocen występuje u pierwszego studenta. Nie możemy jednak powiedzieć jak bardzo się różnią ponieważ wariancji nie jesteśmy w stanie zinterpretować. Wynika to z faktu, że wynik wariancji jest podawany w jednostkach do kwadratu, co zwykle jest pozbawione sensu.

W Excelu dysponujemy dwiema funkcjami do wyliczenia wariancji:

- WARIANCJA.POP (we wzorze znajduje się $\frac{1}{N}$),
- WARIANCJA.PRÓBKI (we wzorze znajduje się $\frac{1}{N-1}$).

W zależności od tego czy mamy informację o populacji czy tylko próbie powinniśmy stosować odpowiednią formułę. Podczas zajęć przyjmujemy, że dysponujemy całą populacją i będziemy stosować odpowiednie funkcje.

Pierwiastek z wariancji czyli **odchylenie standardowe** umożliwia liczbowe określenie zróżnicowania. Informuje o ile jednostki zbiorowości różnią się średnio od średniej. W interpretacji odchylenia standardowego musimy pamiętać o pojawiającym się dwa razy słowie *średnia*. Pierwsze dotyczy średniej zastosowanej we wzorze na wariancję, a drugie określa policzoną wcześniej średnią arytmetyczną.

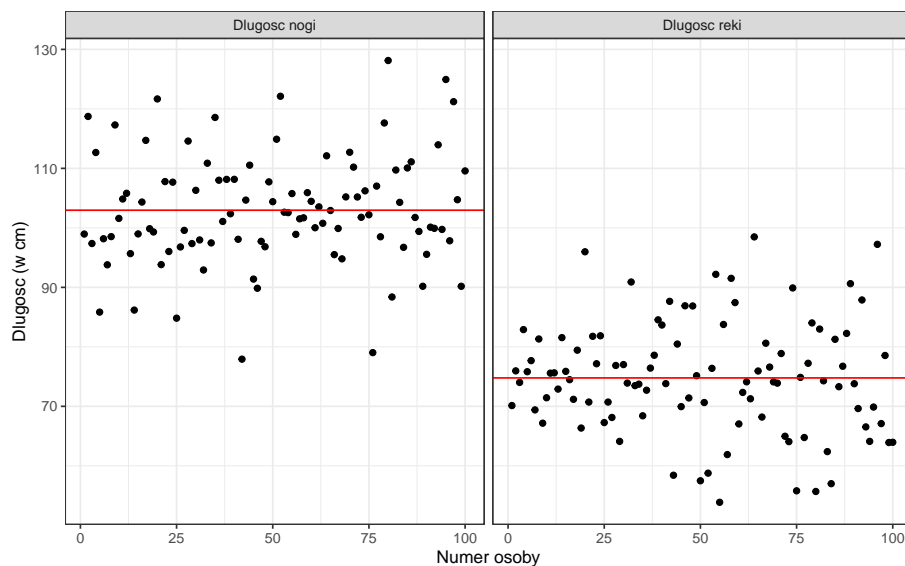
O pierwszym studencie powiemy, że jego oceny różnią się średnio od średniej o 0.89 oceny, natomiast oceny drugiego studenta odchylają się średnio od średniej o 0.37 oceny.

Podobnie jak w przypadku wariancji w Excelu znajdują się dwie funkcje do wyznaczania odchylenia standardowego:

- ODCH.STAND.POPUL,
- ODCH.STANDARD.PRÓBKI.

Jeśli średnie są takie same to do oceny zróżnicowania wystarczy odchylenie standardowe. Sytuacja się jednak komplikuje w przypadku występowania różnic pomiędzy średnimi. Jak zatem porównać zróżnicowanie cech, które mają różne średnie i odchylenia standardowe?

Przeprowadzono eksperyment, w którym 100 osobom zmierzono długość ręki i nogi.



Średnia długość nogi wynosiła 102.97 cm, a odchylenie standardowe 9.24 cm. Z kolei długość ręki charakteryzowała się wartością 74.78 cm z odchyleniem standardowym rzędu 9.33 cm. Ocena zróżnicowania cech o różnych średnich jest możliwe z wykorzystaniem **klasycznego współczynnika zmienności**:

$$V_s = \frac{s}{\bar{x}} \cdot 100,$$

gdzie:

- s - odchylenie standardowe,
- \bar{x} - średnia arytmetyczna.

Współczynnik zmienności wyrażony jest w procentach i można przyjąć kilka umownych progów:

- 0%-20% - cecha mało zróżnicowana,
- 21%-40% - cecha umiarkowanie zróżnicowana,
- 41%-60% - cecha silnie zróżnicowana,
- powyżej 60% - cecha bardzo silnie zróżnicowana.

Oczywiście wszystko zależy od tego jaką cechę analizujemy i jakie jest jej typowe zróżnicowanie.

Obliczając wartość współczynnika zmienności dla długości nogi otrzymamy 8.97%, natomiast dla długości ręki 12.48%. Na tej podstawie możemy stwierdzić, że długość ręki charakteryzuje się większym zróżnicowaniem.

Klasyczny współczynnik zmienności nie ma oprogramowanej odpowiedniej funkcji w Excelu. Można natomiast w prosty sposób tę wartość obliczyć.

Odchylenie standardowe oraz średnią zestawiamy ze sobą także podczas wyznaczania **typowego obszaru zmienności**:

$$\bar{x} - s < x_{typ} < \bar{x} + s$$

Zgodnie z definicją w tym przedziale mieści się około 2/3 wszystkich jednostek analizowanej cechy.

Typowy obszar zmienności dla długości nogi to przedział od 93.73 cm do 112.21 cm i w rzeczywistości zawiera 74% obserwacji.

Patrz też: Reguła trzech sigm.

Do kompletnego opisu struktury brakuje tylko miar określających asymetrię oraz skupienie wokół średniej. **Klasyczny współczynnik asymetrii** nazywany także trzecim momentem centralnym albo skośnością jest wyrażony wzorem:

$$\alpha_3 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3}{s^3},$$

gdzie:

- α_3 - symbol klasycznego współczynnika asymetrii,
- s - odchylenie standardowe w zbiorowości,
- \bar{x} - średnia arytmetyczna w zbiorowości,
- x_i - wariant cechy mierzalnej,
- N - liczebność badanej zbiorowości.

Pozwala określić czy rozkład cechy jest:

- symetryczny - rozkład jest symetryczny, $\alpha_3 = 0$,
- lewostronnie asymetryczny - wydłużone lewe ramię rozkładu, $\alpha_3 < 0$,
- prawostronnie asymetryczny - wydłużone prawe ramię rozkładu, $\alpha_3 > 0$.

Skośność dla długości nogi wynosi 0.1134617, co oznacza, że rozkład długości nóg cechuje się lekką prawostronną asymetrią.

W Excelu znajduje się funkcja o nazwie:

- SKOŚNOŚĆ.

Skupienie wokół średniej definiuje **klasyczny współczynnik koncentracji**, inaczej czwarty moment centralny lub kurtosa:

$$\alpha_4 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4}{s^4},$$

gdzie:

- α_4 - symbol klasycznego współczynnika koncentracji,
- s - odchylenie standardowe w zbiorowości,
- \bar{x} - średnia arytmetyczna w zbiorowości,
- x_i - wariant cechy mierzalnej,
- N - liczebność badanej zbiorowości.

Pozwala określić czy rozkład cechy jest:

- normalny - $\alpha_4 = 3$,
- spłaszczony - wartości nie są mocno skoncentrowane wokół średniej, $\alpha_4 < 3$,
- wysmukły - wartości są mocno skoncentrowane wokół średniej, $\alpha_4 > 3$.

Niektóre programy zamiast kurtozy wyznaczają tzw. eksces:

$$Ex = \alpha_4 - 3$$

Wówczas wartość tej miary interpretujemy przyjmując za punkt odniesienia wartość 0.

Kurtoza dla długości nogi wynosi 6.4025412, co oznacza, że rozkład długości nóg jest wysmukły.

W Excelu znajduje się funkcja o nazwie:

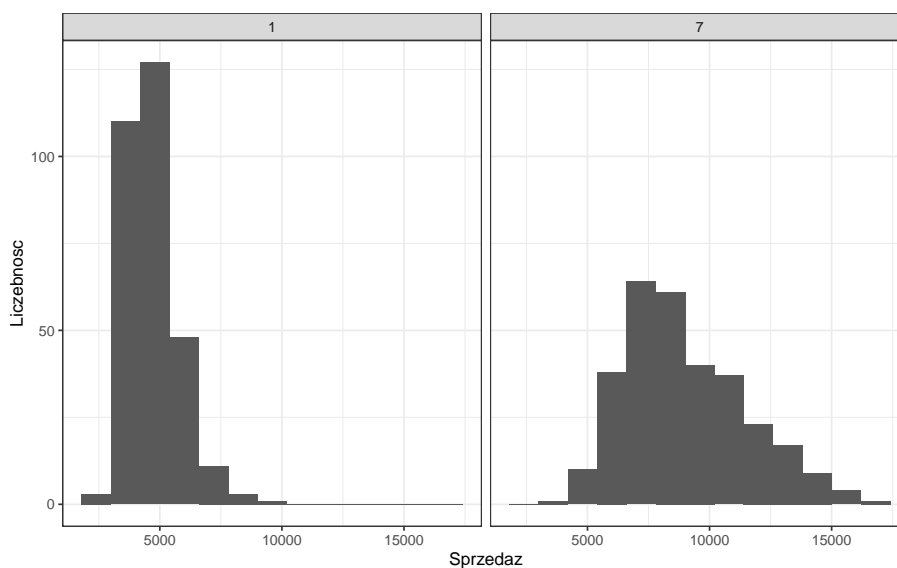
- KURTOZA.

W rzeczywistości wynikiem działania tej funkcji jest eksces. W interpretacji zatem wynik odnosimy do wartości 0.

Do wyznaczenia powyższych miar można także wykorzystać dodatek programu Excel: Analiza danych znajdujący się po prawej stronie we wstążce DANE. Jeśli nie widzimy tego dodatku to klikamy *Przycisk pakietu Office* w lewym górnym rogu ekranu, następnie *Opcje*. W nowym oknie przechodzimy do *Dodatki* i na dole okna przycisk *Przejdź*. Zaznaczamy *Analysis ToolPak* i wybieramy *OK*.

Przykład

Wykorzystując zbiór danych na temat sklepów Rossmann przeprowadzimy kompleksową analizę porównawczą struktury sprzedaży w dwóch wybranych sklepach. Pierwszy ze sklepów (id=1) posiada asortyment podstawowy i jest typu *c*, natomiast drugi (id=7) posiada asortyment rozszerzony i jest typu *a*. W pierwszym kroku zobaczymy jak wygląda rozkład analizowanej cechy po wyeliminowaniu dni, w którym sklep był zamknięty.



Już na pierwszy rzut oka widać różnice w rozkładzie sprzedaży dla poszczególnych sklepów. Pierwszy z rozkładów jest bardziej wysmukły, natomiast w drugim przypadku obserwujemy wyższe wartości sprzedaży. Z wykorzystaniem miar klasycznych dokonamy analizy sprzedaży.

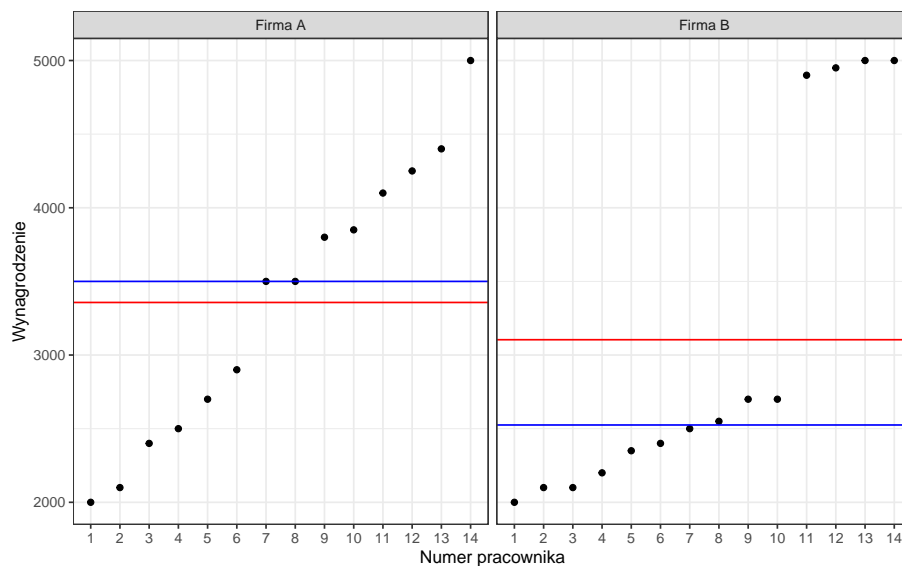
	Sklep_1	Sklep_7
n	303.00	305.00
x_sr	4730.72	8975.03
s	1057.28	2487.50
v_s	22.35	27.72
alpha_3	1.00	0.58
alpha_4	4.53	2.83

3.2 Miary pozycyjne

Podstawowe miary pozycyjne nie są obliczane z wykorzystaniem wszystkich obserwacji, jak ma to miejsce w przypadku miar klasycznych, tylko szukamy obserwacji która wskazuje wartość wybranej miary pozycyjnej. Najpopularniejszą z miar pozycyjnych jest **mediana** (kwartył 2, wartość środkowa, Q_2), która wyznacza wartość dla której 50% jednostek zbiorowości ma wartości cechy niższe bądź równe medianie, a 50% równe bądź wyższe od mediany.

Medianę wyznacza się poprzez posortowanie wartości cechy rosnąco i wybór wartości środkowej (jeśli N jest nieparzyste) lub średniej z wartości środkowych (jeśli N jest parzyste).

Zaletą mediany jest mniejsza wrażliwość na obserwacje odstające. Rozważmy przypadek wynagrodzeń w pewnych przedsiębiorstwach:



W firmie A wynagrodzenia pracowników nie są zróżnicowane, ale nie występują pomiędzy nimi zbyt duże różnice. Średnia pensja (kolor czerwony) wynosi 3357 zł, natomiast mediana (kolor niebieski) odpowiada wynagrodzeniom 7 i 8 pracownika - 3500 zł. Można powiedzieć, że obie wartości dobrze odzwierciedlają realne zarobki pracowników. Z kolei w firmie B nierówności dochodowe są znacznie większe, możliwe że zestawiono wynagrodzenia pracowników szeregowych oraz kadry zarządzającej. Średnia wynosząca 3104 zł nie oddaje prawdziwych zarobków ani pierwszej ani drugiej grupy. Natomiast wartość mediany wynosząca 2525 zł jest bardziej odporna na wartości odstające. Mediana wynagrodzenia w firmie B oznacza, że 50% pracowników otrzymuje pensję w wysokości 2525 zł lub mniej, natomiast drugie 50% zatrudnionych uzyskuje wynagrodzenie w wysokości 2525 zł lub więcej.

W Excelu możemy skorzystać z funkcji:

- MEDIANA(wartości cechy),
- KWARTYL.PRZEDZ.ZAMK(wartości cechy, 2).

Mediana podzieliła nam jednostki zbiorowości na dwie połowy. Jeśli podzielimy pierwszą połowę ponownie na pół otrzymamy wartość **kwartyla pierwszego (dolnego)**, który informuje, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q_1 , a 75% równe bądź wyższe od tego kwartyla. Z kolei po podzieleniu drugiej połowy obserwacji uzyskujemy wartość **kwartyla trzeciego (górnego)**, który informuje, że 75% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi trzeciemu Q_3 , a 25% równe bądź wyższe od tego kwartyla.

Do wyznaczenia wartości kwartyli w Excelu korzystamy z funkcji:

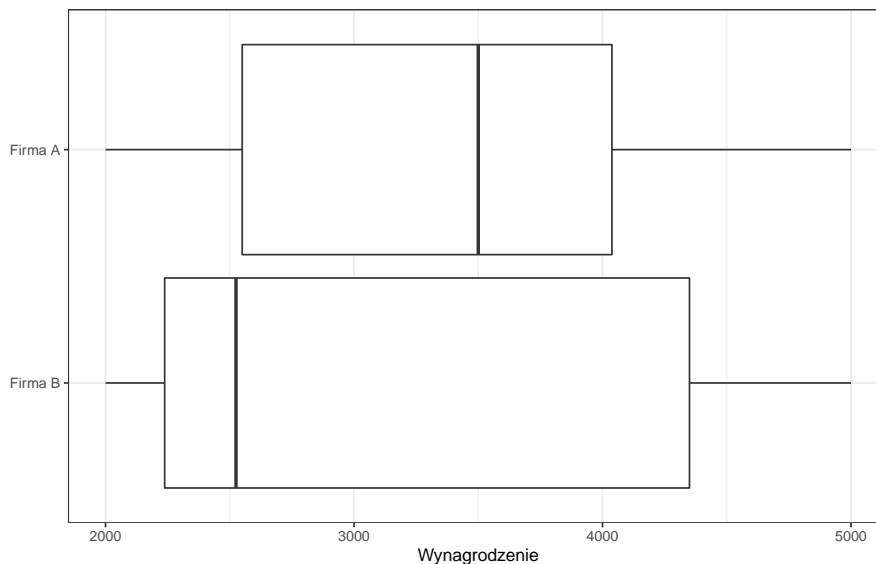
- KWARTYL.PRZEDZ.ZAMK(wartości cechy, numer kwartyła),

gdzie numer kwartyła to:

- 0 - minimum,
- 1 - kwartył dolny,
- 2 - mediana,
- 3 - kwartył górny,
- 4 - maksimum.

W firmie A kwartył dolny wynagrodzeń wyniósł 2550 zł, co oznacza, że 25% pracowników uzyskuje pensję równą bądź niższą niż 2550 zł, a 75% równą bądź wyższą niż 2550. Z kolei 75% pracowników otrzymuje wynagrodzenie mniejsze lub równe 4038 zł, a 25% większe bądź równe 4038 zł. W firmie B kwartył pierwszy jest równy 2238 zł, a trzeci 4350 zł.

Wartości kwartyli można przedstawić na wykresie pudełkowym (ang. boxplot):



W miarach pozycyjnych opartych na kwartyłach zróżnicowanie wartości od mediany mierzy **odchylenie ćwiartkowe**:

$$Q = \frac{(Q_3 - Q_1)}{2}$$

gdzie:

- Q - symbol odchylenia ćwiartkowego,
- Q_1 - kwartył pierwszy,
- Q_3 - kwartył trzeci.

Mierzy ono przeciętne odchylenie wartości cechy zbiorowości od mediany u 50% środkowych jednostek - między kwartylem dolnym i górnym. Przykładowo w firmie A przeciętne odchylenie wynagrodzenia od mediany wynosi 744 zł.

Zestawienie odchylenia ćwiartkowego oraz mediany pozwala na obliczenie **pozycyjnego współczynnika zmienności**:

$$V_Q = \frac{Q}{Q_2} \cdot 100$$

gdzie:

- V_Q - symbol pozycyjnego współczynnika zmienności,
- Q - odchylenie ćwiartkowe,
- Q_2 - mediana.

Podobnie jak w przypadku klasycznego współczynnika zmienności korzystamy z umownych progów dotyczących zróżnicowania. W firmie A pozycyjny współczynnik zmienności był równy 21% co oznacza, że wynagrodzenia w tej firmie cechowały się umiarkowanym zróżnicowaniem, natomiast w firmie B było to 42% czyli silne zróżnicowanie wynagrodzeń.

Ostatnią miarą opartą na kwartylach jest **pozycyjny współczynnik asymetrii**, który określa kierunek i siłę asymetrii jednostek znajdujących się między pierwszym i trzecim kwartylem:

$$A_Q = \frac{(Q_1 + Q_3 - 2 \cdot Q_2)}{(2 \cdot Q)}$$

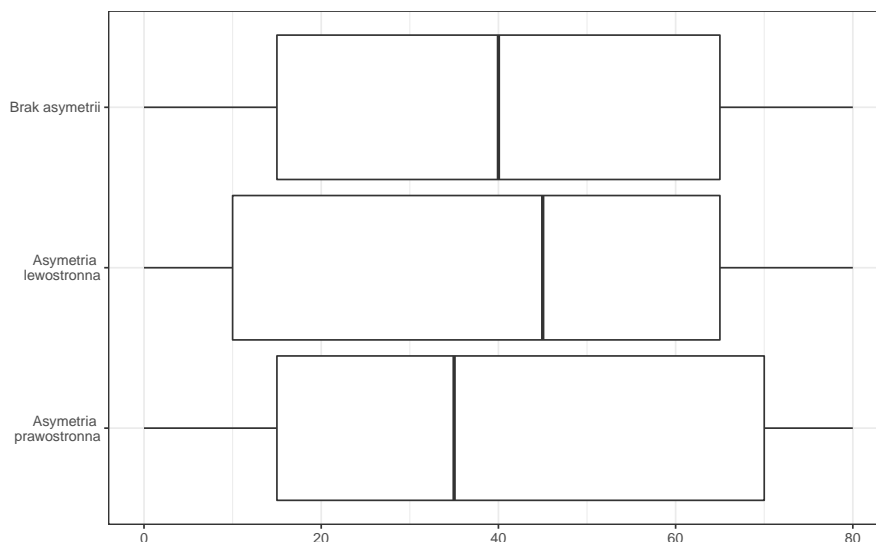
gdzie:

- A_Q — symbol pozycyjnego współczynnika asymetrii,
- Q_1 — kwartył pierwszy,
- Q_3 — kwartył trzeci,
- Q_2 — mediana,
- Q — odchylenie ćwiartkowe.

Interpretacja pozycyjnego współczynnika asymetrii przebiega identycznie jak w przypadku klasycznego współczynnika asymetrii:

- symetryczny - mediana pomiędzy wartościami kwartyli dolnego i górnego, $A_Q = 0$,
- lewostronnie asymetryczny - mediana bliżej wartości kwartyla górnego, $A_Q < 0$,
- prawostronnie asymetryczny - mediana bliżej wartości kwartyla dolnego, $A_Q > 0$.

Tę informację możemy także odczytać z wykresu pudełkowego, określając umiejscowienie mediany względem pozostałych kwartyli:



W firmie A pozycyjny współczynnik asymetrii był równy -0.28 , co pociąga za sobą informację o asymetrii lewostronnej, natomiast w firmie B występowała asymetria prawostronna (0.73).

Przykład

Wyznamy miary pozycyjne dla dwóch sklepów Rossmann analizowanych wcześniej:

	Sklep_1	Sklep_7
n	303.00	305.00
q1	3908.00	7129.00
q2	4607.00	8592.00
q3	5286.00	10681.00
q	689.00	1776.00
v_q	14.96	20.67
aq	-0.01	0.18

Zadania

Przeprowadzić kompleksową analizę struktury sprzedaży/liczby klientów dla poszczególnych dni tygodnia.

3.3 Szereg jednostkowy i przedziałowy

Nie zawsze dysponujemy danymi zebranymi w szeregu prostym. W opracowaniach statystycznych dane publikowane są w postaci szeregów jednostkowych oraz przedziałowych. W tej części opracowania skupimy się na analizie struktury takich danych.

3.3.1 Szereg jednostkowy

W przypadku szeregu jednostkowego możliwe jest odtworzenie szeregu prostego bądź zastosowanie wzorów, w których odpowiednio przeważymy obserwacje. Odpowiednimi wagami będą liczebności.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i n_i$$

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \cdot n_i$$

$$\alpha_3 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3 \cdot n_i}{s^3}$$

$$\alpha_4 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4 \cdot n_i}{s^4}$$

gdzie:

- \bar{x} - średnia arytmetyczna w zbiorowości,
- s - odchylenie standardowe w zbiorowości,
- α_3 - symbol klasycznego współczynnika asymetrii,
- α_4 - symbol klasycznego współczynnika koncentracji,
- x_i - wariant cechy mierzalnej,
- n_i - liczba obserwacji dla wariantu,
- N - liczebność badanej zbiorowości.

3.3.2 Szereg przedziałowy

W przypadku szeregu przedziałowego przeprowadzanie analizy struktury nie jest już takie oczywiste. Nie mamy jednoznacznie określonego wariantu cechy. W związku z tym wyznaczamy środek przedziału klasowego i tą wartość traktujemy jako wariant cechy. Z takiego podejścia do sprawy wynikają dwie istotne kwestie:

- poniższe wzory możemy zastosować wyłącznie do analizy szeregów rozdzielczych przedziałowych zamkniętych o równych przedziałach klasowych,
- sposób utworzenia szeregu rozdzielczego będzie miał wpływ na precyzję wyników.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x'_i n_i$$

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x'_i - \bar{x})^2 \cdot n_i$$

$$\alpha_3 = \frac{\frac{1}{N} \sum_{i=1}^N (x'_i - \bar{x})^3 \cdot n_i}{s^3}$$

$$\alpha_4 = \frac{\frac{1}{N} \sum_{i=1}^N (x'_i - \bar{x})^4 \cdot n_i}{s^4}$$

gdzie:

- \bar{x} - średnia arytmetyczna w zbiorowości,
- s - odchylenie standardowe w zbiorowości,
- α_3 - symbol klasycznego współczynnika asymetrii,
- α_4 - symbol klasycznego współczynnika koncentracji,
- x'_i - środek przedziału klasowego dla wariantu cechy,
- n_i - liczba obserwacji dla wariantu,
- N - liczebność badanej zbiorowości.

Przy założeniu, że dominanta znajduje się w najliczniejszym przedziale możemy zastosować poniższy wzór:

$$D = x_D + \frac{n_D - n_{D-1}}{2n_D - n_{D-1} - n_{D+1}} \cdot c_D$$

gdzie:

- D - symbol dominanty,
- x_D - początek przedziału, w którym znajduje się dominanta,
- n_D - liczebność najliczniejszego przedziału,
- n_{D-1} - liczebność przedziału wcześniejszego niż najliczniejszy,
- n_{D+1} - liczebność przedziału późniejszego niż najliczniejszy,
- c_D - rozpiętość najliczniejszego przedziału.

Dla szeregu rozdzielczego możemy także wyznaczyć wartości kwartyli stosując wzory interpolacyjne:

$$Q_1 = xQ_1 + \frac{\frac{N}{4} - cumQ_1^{-1}}{nQ_1} \cdot cQ_1$$

gdzie:

- Q_1 - oznaczenie kwartyła pierwszego,
- N - liczebność badanej zbiorowości,
- xQ_1 - początek przedziału, w którym znajduje się kwartył pierwszy,
- $cumQ_1^{-1}$ - skumulowana liczebność z przedziału wcześniejszego niż ten, który zawiera kwartył pierwszy,
- nQ_1 - liczebność przedziału zawierającego kwartył pierwszy,
- cQ_1 - rozpiętość przedziału zawierającego kwartył pierwszy.

$$Q_2 = xQ_2 + \frac{\frac{N}{2} - cumQ_2^{-1}}{nQ_2} \cdot cQ_2$$

gdzie:

- Q_2 - oznaczenie mediany,
- N - liczebność badanej zbiorowości,
- xQ_2 - początek przedziału, w którym znajduje się mediana,
- $cumQ_2^{-1}$ - skumulowana liczebność z przedziału wcześniejszego niż ten, który zawiera medianę,
- nQ_2 - liczebność przedziału zawierającego medianę,
- cQ_2 - rozpiętość przedziału zawierającego medianę.

$$Q_3 = xQ_3 + \frac{\frac{3N}{4} - cumQ_3^{-1}}{nQ_3} \cdot cQ_3$$

gdzie:

- Q_3 - oznaczenie kwartyła trzeciego,
- N - liczebność badanej zbiorowości,
- xQ_3 - początek przedziału, w którym znajduje się kwartył trzeci,
- $cumQ_3^{-1}$ - skumulowana liczebność z przedziału wcześniejszego niż ten, który zawiera kwartył trzeci,
- nQ_3 - liczebność przedziału zawierającego kwartył trzeci,
- cQ_3 - rozpiętość przedziału zawierającego kwartył trzeci.

Po wyznaczeniu wartości kwartyli pozostałe miary liczymy w tradycyjny sposób.

3.4 Podsumowanie miar

3.4.1 Schemat

3.4.2 Miary klasyczne

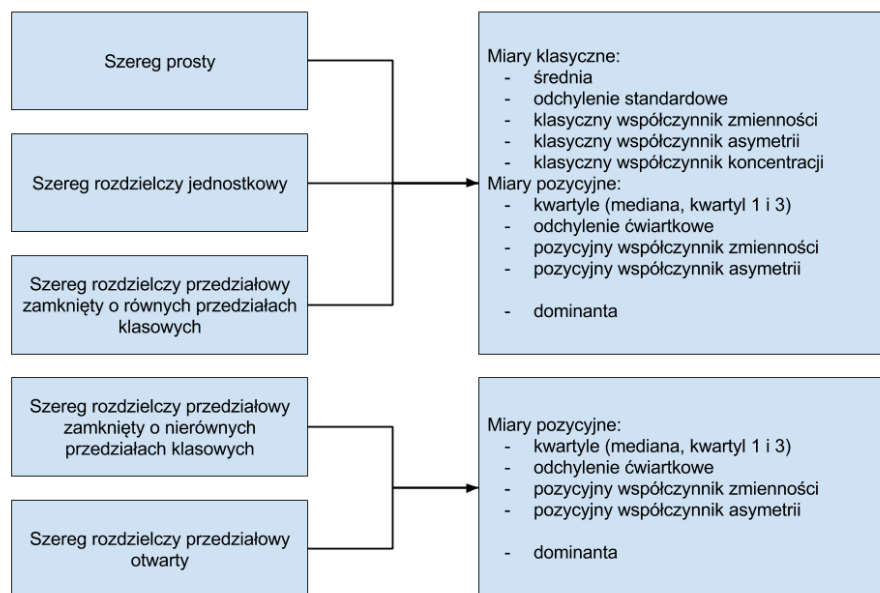


Figure 3.1: Analiza struktury w zależności od typu szeregu

Miara	Oznaczenie	Wzór	Interpretacja i wykorzystanie	Funkcja w Excelu
Średnia arytmetyczna	\bar{x}	$\frac{\sum_{i=1}^N x_i}{N}$	Wartość przeciętna	ŚREDNIA(x)
Średnia harmoniczna	\bar{x}_h	$\frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$	Wartość przeciętna	ŚREDNIA.GEOMETRYCZNA(x)
Średnia geometryczna	\bar{x}_g	$\sqrt[n]{x_1 x_2 \dots x_N}$	Wartość przeciętna	ŚREDNIA.HARMONICZNA(x)

Miara	Oznaczenie	Wzór	Interpretacja i wykorzystanie	Funkcja w Excelu
Odchylenie przeciętne (średnie)	d	$\frac{1}{N} \sum x_i - \bar{x} $	O ile wszystkie jednostki badanej zbiorowości różnią się średnio ze względu na wartość zmiennej od średniej arytmetycznej tej zmiennej	ODCH.ŚREDNIE(x)
Odchylenie kwadratowe	d^2	$\sum_{i=1}^N (x_i - \bar{x})^2$	Kwadrat odchylenia przeciętnego	ODCH.KWADRATOWE(x)
Odchylenie standardowe (dla populacji)	s	$\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$	O ile przeciętnie odchylają się wartości od średniej	ODCH.STAND.POPUL(x)
Odchylenie standardowe (dla próby)	s	$\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}}$	O ile przeciętnie odchylają się wartości od średniej	ODCH.STANDARD.PRÓBK(x)
Wariancja (dla populacji)	s^2	$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}$	Informuje o zróżnicowaniu populacji	WARIANCJA.POP(x)
Wariancja (dla próby)	s^2	$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}$	Informuje o zróżnicowaniu próby	WARIANCJA.PRÓBK(x)

Miara	Oznaczenie	Wzór	Interpretacja i wykorzystanie	Funkcja w Excelu
Rozstęp	R	$\max(x) - \min(x)$	Empiryczny obszar zmienności, wartość maksymalna cechy x minus wartość minimalna tej cechy	–
Typowy obszar zmienności	–	$\bar{x} - s < x_{typ} < \bar{x} + s$	Informuje o relatywnym zróżnicowaniu populacji (próby). Zwykle wykorzystujemy do porównań dwóch lub więcej grup.	–
Współczynnik zmienności	V_x	$\frac{s}{\bar{x}}$	Informuje o relatywnym zróżnicowaniu populacji (próby). Zwykle wykorzystujemy do porównań dwóch lub więcej grup. Wyrażamy w procentach.	–

Miara	Oznaczenie	Wzór	Interpretacja i wykorzystanie	Funkcja w Excelu
Współczynnik asymetrii	α_3	$\frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3}{s^3}$	Pozwala zidentyfikować czy rozkład jest symetryczny lub asymetryczny	SKOŚNOŚĆ(x)
Współczynnik koncentracji	α_4	$\frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4}{s^4}$	Pozwala zidentyfikować czy rozkład jest wysmukły czy spłaszczony	KURTOZA(x)
Eksces	Ex	$\alpha_4 - 3$	Pozwala zidentyfikować czy rozkład jest wysmukły czy spłaszczony (porównujemy do 0)	—

3.4.3 Miary pozycyjne

Miara	Oznaczenie	Wzór	Interpretacja i wykorzystanie	Funkcja w Excelu
Kwartył 1	Q_1		Dzieli populację na dwie części w stosunku 25 / 75	KWARTYL.PRZEDZ.ZAMK(x, 1)
Kwartył 2, Mediana	Q_2, Me		Dzieli populację na dwie części w stosunku 50 / 50	KWARTYL.PRZEDZ.ZAMK(, 2) lub MEDIANA(x)

Miara	Oznaczenie	Wzór	Interpretacja i wykorzystanie	Funkcja w Excelu
Kwartyl 3	Q_3		Dzieli populację na dwie części w stosunku 75 / 25	KWARTYL.PRZEDZ.ZAMK(x,3)
Odchylenie ćwiartkowe	Q	$Q = \frac{Q_3 - Q_1}{2}$	Mierzy ono przeciętne odchylenie wartości cechy zbiorowości od mediany	–
Pozycyjny współczynnik zmienności	V_Q	$\frac{Q}{Me}$	Mierzy przeciętne zróżnicowanie cechy	–
Pozycyjny współczynnik asymetrii	A_Q	$\frac{Q_1 + Q_3 - 2Me}{2Q}$	Mierzy (a)symetrię rozkładu	–

3.5 Przedziały ufności

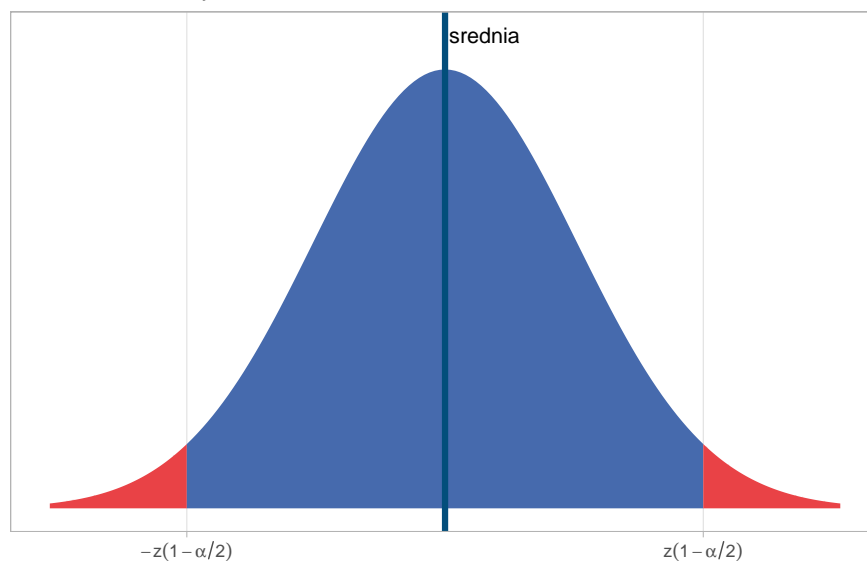
Dotychczas analizowane dane dotyczyły populacji, zatem obliczone wartości statystyk można uznać za precyzyjne i nieobciążone błędem. Natomiast większość prowadzonych badań ogranicza się do analizy jedynie fragmentu populacji. Wówczas, oprócz obliczenia interesującej nas miary ważne jest także podanie możliwego błędu. Popularną praktyką jest obliczanie **przedziałów ufności**, które prezentują zakres, w którym z określonym prawdopodobieństwem znajduje się prawdziwa wartość parametru. Zwykle bierze się pod uwagę następujące prawdopodobieństwa: 90%, 95% i 99%, niemniej można wybrać dowolną wartość z przedziału 0-100%. We wzorach operuje się pojęciem **poziomu istotności** oznaczanym przez α .

Skupimy się na wyznaczaniu następujących przedziałów ufności:

- dla średniej w populacji normalnym ze znanym odchyleniem standardowym,
- dla średniej w populacji normalnym z nieznanym odchyleniem standardowym dla małej próby,
- dla średniej w populacji normalnym z nieznanym odchyleniem standardowym dla dużej próby,
- dla odsetka (proporcji, frakcji).

W każdym przypadku będziemy musieli wyznaczyć kwantyl rozkładu, który odpowiada przyjętemu poziomowi prawdopodobieństwa i tym samym obszar, który pokryje wyznaczony przedział ufności. Dla rozkładu normalnego ta sytuacja jest przedstawiona na wykresie:

Rozkład normalny



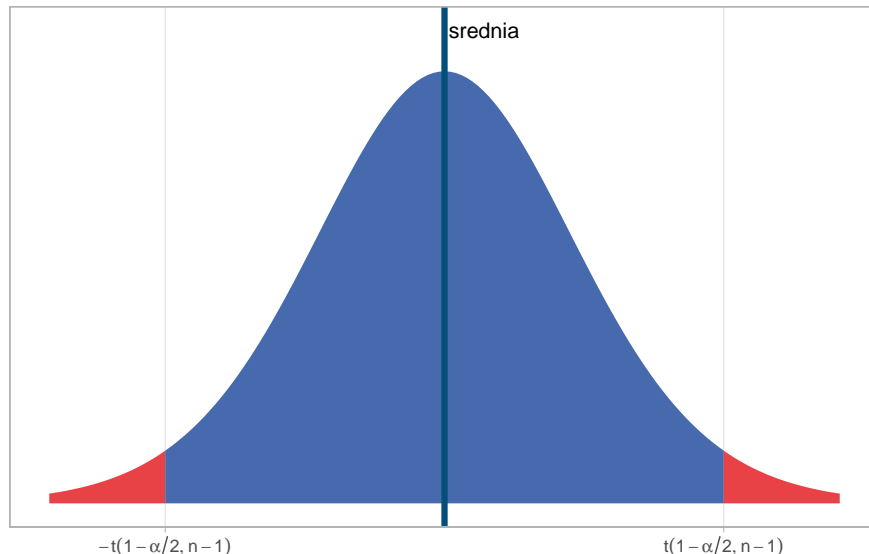
W tabeli przedstawiono relację przyjętego prawdopodobieństwa, poziomu istotności oraz wartości kwantyli rozkładu normalnego.

Prawdopodobieństwo	Poziom istotności	Kwantyl rozkł. norm.
99% (0,99)	0,01	2,58
95% (0,95)	0,05	1,96
90% (0,90)	0,10	1,64

W Excelu te wartości można wyznaczyć z wykorzystaniem funkcji: ROZKŁ.NORMALNY.S.ODWR($1-\alpha/2$).

Te wartości dla najpopularniejszych poziomów prawdopodobieństwa zawsze będą takie same. Natomiast w przypadku małych prób ($n < 30$) należy skorzystać z rozkładu t-Studenta. Kształt tego rozkładu jest zbliżony do normalnego, natomiast wartości kwantyli zależą od dwóch parametrów - przyjętego poziomu istotności oraz liczebności próby pomniejszonej o 1 (liczba stopni swobody).

Rozkład t-Studenta z 16 stopniami swobody



Do wyznaczenia wartości kwantyla z rozkładu t-Studenta wykorzystuje się funkcje: ROZKŁ.T.ODWR($1-\alpha/2; n-1$) lub ROZKŁ.T.ODWR.DS($\alpha; n-1$)

3.5.1 Średnia w populacji normalnej ze znanym odchyleniem standardowym

W sytuacji, w której znane jest odchylenie standardowe w populacji np. z wcześniejszych badań można wykorzystać następujący wzór:

$$P \left\{ \bar{X} - z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}} < m < \bar{X} + z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

gdzie:

- m - prawdziwa wartość średniej w populacji,
- \bar{X} - estymator średniej,
- $z_{(1-\alpha/2)}$ - kwantyl rozkładu normalnego obliczony dla poziomu istotności α ,
- σ - znane odchylenie standardowe,
- n - liczebność próby.

Przykład

W zakładzie produkcyjnym postanowiono zbadać staż pracy pracowników. W tym celu z populacji pracowników wylosowano próbę o liczebności 196 osób, na podstawie której obliczono, że $\bar{x} = 6,9$ lat. Dotychczasowe doświadczenie

wskazuje, że rozkład stażu pracowników jest opisany rozkładem normalnym z odchyleniem standardowym 2,8 lat. Dla $\alpha = 0,05$ zbudować przedział ufności.

Kwantyl rozkładu normalnego można obliczyć z wykorzystaniem formuły ROZKŁ.NORMALNY.S.ODWR. W tym przypadku zostanie użyta formuła ROZKŁ.NORMALNY.S.ODWR(1-0,05/2), co skutkuje otrzymaniem wartości 1,96. Po podstawieniu do wzoru:

$$6,9 - 1,96 \frac{2,8}{\sqrt{196}} < m < 6,9 + 1,96 \frac{2,8}{\sqrt{196}}$$

$$6,508 < m < 7,292$$

Przedział od 6,5 do 7,3 lat z prawdopodobieństwem 95% pokrywa prawdziwą wartość stażu pracy wszystkich pracowników.

3.5.2 Średnia w populacji normalnej z nieznanym odchyleniem standardowym - mała próba

Jeśli odchylenie standardowe w populacji nie jest znane to można wykorzystać wzór, w którym używa się wartości odchylenia standardowego obliczonego na podstawie próbie. Trzeba jednak rozróżnić małą oraz dużą próbę. Przyjętym w statystyce progiem jest $n > 30$, kiedy uznaje się próbę za dużą. Podstawową różnicą jest wykorzystanie we wzorze przedziału ufności dla małej próby kwantyla rozkładu t-Studenta, a dla dużej kwantyla rozkładu normalnego.

$$P \left\{ \bar{X} - t_{(1-\alpha/2, n-1)} \frac{s}{\sqrt{n-1}} < m < \bar{X} + t_{(1-\alpha/2, n-1)} \frac{s}{\sqrt{n-1}} \right\} = 1 - \alpha$$

gdzie:

- m - prawdziwa wartość średniej w populacji,
- \bar{X} - estymator średniej,
- $t_{(1-\alpha/2, n-1)}$ - kwantyl rozkładu t-Studenta dla poziomu istotności α z $n-1$ stopniami swobody,
- s - odchylenie standardowe z próby,
- n - liczebność próby.

Przykład

Postanowiono oszacować średni czas potrzebny do wykonania detalu. Z populacji robotników wylosowano próbę 17 osób i dokonano pomiaru czasu wykonywania detalu. Okazało się, że średni czas wykonania detalu wynosił 15 minut, a odchylenie standardowe 2 minuty. Rozkład czasu wykonania tego detalu ma rozkład normalny i przyjęto poziom istotności $\alpha = 0,05$.

Wartość kwantyla rozkładu t-Studenta obliczamy za pomocą formuły $\text{ROZKŁ.T.ODWR}(1-0,05/2;16)$ lub $\text{ROZKŁ.T.ODWR.DS}(0,05;16)$, co daje wartość 2,12.

$$15 - 2,12 \frac{2}{\sqrt{16}} < m < 15 + 2,12 \frac{2}{\sqrt{16}}$$

$$13,94 < m < 16,06$$

Średni czas wykonania detalu z 95% prawdopodobieństwem jest nie mniejszy niż 13,94 i nie większy niż 16,06 minuty.

3.5.3 Średnia w populacji normalnej z nieznanym odchyleniem standardowym - duża próba

W przypadku dużej próby przedział ufności można przybliżyć z wykorzystaniem kwantyla rozkładu normalnego.

$$P \left\{ \bar{X} - z_{(1-\alpha/2)} \frac{s}{\sqrt{n}} < m < \bar{X} + z_{(1-\alpha/2)} \frac{s}{\sqrt{n}} \right\} \approx 1 - \alpha$$

gdzie:

- m - prawdziwa wartość średniej w populacji,
- \bar{X} - estymator średniej,
- $z_{(1-\alpha/2)}$ - kwantyl rozkładu normalnego obliczony dla poziomu istotności α ,
- s - odchylenie standardowe z próby,
- n - liczebność próby.

3.5.4 Proporcja

Na podobnej zasadzie można także wyznaczyć przedział ufności dla odsetka.

$$P \left\{ \frac{m}{n} - z_{(1-\alpha/2)} \sqrt{\frac{\frac{m}{n} (1 - \frac{m}{n})}{n}} < p < \frac{m}{n} + z_{(1-\alpha/2)} \sqrt{\frac{\frac{m}{n} (1 - \frac{m}{n})}{n}} \right\} \approx 1 - \alpha$$

gdzie:

- p - wartość proporcji w populacji,
- $z_{(1-\alpha/2)}$ - kwantyl rozkładu normalnego obliczony dla poziomu istotności α ,
- m - liczba jednostek posiadających daną cechę,
- n - liczebność

Przykład

Spośród 10 tysięcy pracowników wylosowano próbę liczącą 200 osób i przeprowadzono badanie dotyczące opuszczenia zakładu pracy. Okazało się, że 20 z 200 respondentów zamierza, z różnych względów, opuścić zakład pracy. Dla poziomu ufności 90% wyznaczyć przedział ufności dla wskaźnika pracowników planujących opuścić bieżące miejsce pracy.

$$\frac{20}{200} - 1,64\sqrt{\frac{\frac{20}{200}\left(1 - \frac{20}{200}\right)}{200}} < p < \frac{20}{200} + 1,64\sqrt{\frac{\frac{20}{200}\left(1 - \frac{20}{200}\right)}{200}}$$

$$0,065 < p < 0,135$$

$$6,5\% < p < 13,5\%$$

Z prawdopodobieństwem 90% możemy stwierdzić, że pracowników planujących opuścić zakład pracy jest nie mniej niż 6,5% i nie więcej niż 13,5%.

3.6 Testy statystyczne

Test statystyczny to procedura pozwalająca oszacować prawdopodobieństwo spełnienia pewnej hipotezy statystycznej w populacji na podstawie danych pochodzących z próby losowej. Hipoteza statystyczna to układ dwóch hipotez: zerowej i alternatywnej.

- hipoteza zerowa H_0 - zakładamy, że pomiędzy estymatorem i parametrem lub rozkładem empirycznym i teoretycznym nie ma różnic (H_0 zawsze zawiera znak równości):

$$H_0 : m = m_0$$

- hipoteza alternatywna H_1 - dopuszcza istnienie różnic pomiędzy estymatorem a parametrem. Może przyjmować trzy warianty:

$$H_1 : m \neq m_0$$

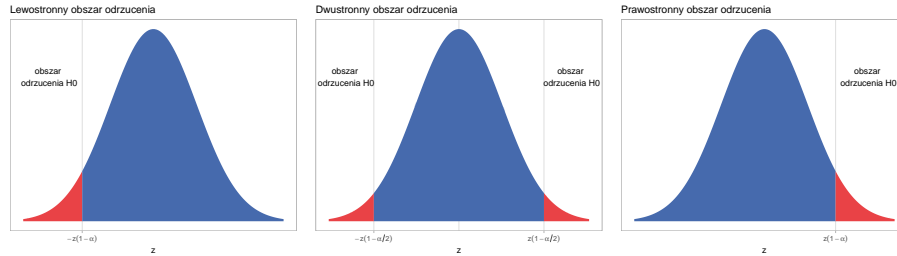
$$H_1 : m > m_0$$

$$H_1 : m < m_0$$

gdzie m_0 - hipotetyczna wartość średniej w populacji generalnej.

Weryfikacja testu statystycznego polega na obliczeniu wartości statystyki testowej oraz kwantyla odpowiedniego rozkładu i sprawdzenie czy wartość statystyki testowej wpada do przedziału odrzucenia. Jeśli wartość statystyki testowej

jest większa od wartości kwantyla rozkładu (np. $|t| \geq t_\alpha$) to są podstawy do odrzucenia hipotezy zerowej, natomiast w przypadku kiedy $|t| < t_\alpha$ to nie ma podstaw do odrzucenia hipotezy zerowej.



Obecnie wszystkie programy do analiz statystycznych zwracają także wartość p czyli najmniejszy poziom istotności przy którym nie ma podstaw do odrzucenia hipotezy zerowej. Przyjmuje się, że jeśli $p < \alpha$ to są podstawy do odrzucenia hipotezy zerowej, natomiast w przypadku $p \geq \alpha$ nie ma podstaw do odrzucenia H_0 . Jest to jednak bardzo uproszczona definicja i warto zgłębić temat np. korzystając z Wikipedii.

3.6.1 Test t dla jednej średniej

Sprawdzamy czy średnia w populacji jest równa określonej wartości. Testowany jest jeden z poniższych układów hipotez:

- $H_0 : m = m_0; H_1 : m \neq m_0$
- $H_0 : m = m_0; H_1 : m > m_0$
- $H_0 : m = m_0; H_1 : m < m_0$

Statystyka testowa jest następująca:

$$t = \frac{(\bar{x} - m_0)}{s} \sqrt{n}$$

i ma rozkład t -Studenta o $n - 1$ stopniach swobody.

Jeśli wartość statystyki testowej jest większa od wartości kwantyla rozkładu ($|t| \geq t_\alpha$) to są podstawy do odrzucenia hipotezy zerowej, natomiast w przypadku kiedy $|t| < t_\alpha$ nie ma podstaw do odrzucenia hipotezy zerowej.

Hipotezę zerową w teście t dla jednej średniej można także zweryfikować na podstawie przedziału ufności. Nie będzie podstaw do odrzucenia hipotezy zerowej jeśli weryfikowana wartość m_0 będzie znajdować się w wyznaczonym przedziale ufności. Jeśli będzie poza nim to są podstawy do odrzucenia H_0 .

Przykład

Z 24 gospodarstw zebrano dane na temat plonów żyta (w tonach na hektar):
30, 31, 27, 35, 31, 32, 36, 25, 31, 32, 28, 29, 24, 30, 25, 31, 25, 29, 25, 27, 22,

29, 32, 29. Czy prawdziwa jest hipoteza, że w gospodarstwach w województwie uzyskuje się średnio 30 ton z hektara żyta? Przyjmij poziom istotności 0,05.

Układ hipotez jest następujący:

- $H_0 : m = 30$
- $H_1 : m \neq 30$

W pierwszej kolejności obliczamy wartość średnią, odchylenie standardowe i na tej podstawie wartość statystyki testowej:

$$t = \frac{\bar{x} - m_0}{s} \sqrt{n} = \frac{29 - 30}{3,5} \sqrt{24} = -1,47$$

Natomiast wartość kwantyla rozkładu t-Studenta z 23 stopniami swobody dla poziomu istotności $\alpha = 0,05$ wynosi 2,07 ($\text{ROZKŁ. T. ODWR}(1-0,05/2; 24-1)$). W takim razie obszar odrzucenia hipotezy zerowej to przedział od $-\infty$ do -2,07 i od 2,07 do $+\infty$. Wartość statystyki testowej nie znajduje się w tym przedziale, zatem nie ma podstaw do odrzucenia hipotezy zerowej - średnia w populacji nie różni się od 30 ton z hektara.

Z kolei przedział ufności dla średniej z plonów żyta to 27,46 do 30,46 ton i można zauważyć, że weryfikowana wartość m_0 znajduje się w tym przedziale.

Natomiast w sytuacji, w której m_0 zostałyby przyjęte na poziomie 32 ton to wówczas statystyka testowa wynosi -4,29 i wpada do obszaru odrzucenia w rozkładzie t-Studenta. Ponadto można zauważyć, że ta wartość znajduje się poza wyznaczonym przedziałem ufności.

Zadania

1. Na grupie 25 kobiet przeprowadzono badanie dotyczące zarobków i zebrano następujące wyniki: 4330, 3063, 3012, 3486, 3415, 3097, 2451, 3418, 2970, 4050, 2828, 4076, 3011, 3575, 3939, 3089, 3733, 3347, 2719, 3238, 4372, 3272, 2909, 3368, 3598. Przyjmując poziom istotności równy 0.1 zweryfikuj hipotezę, że średnie zarobki kobiet w populacji nie różnią się istotnie od 3500 zł.
2. W 17 szkołach podstawowych przeprowadzono wśród uczniów testy kompetencji. Otrzymano następujące wyniki punktowe: 84, 82, 96, 78, 73, 83, 82, 77, 83, 80, 78, 85, 80, 71, 66, 79, 91. Zweryfikuj hipotezę, że średni wynik testu kompetencji w szkołach podstawowych wynosi 80 punktów przyjmując poziom istotności 0,01.

3.6.2 Test istotności proporcji

Testowany jest jeden z poniższych układów hipotez:

- $H_0 : p = p_0; H_1 : p \neq p_0$
- $H_0 : p = p_0; H_1 : p > p_0$

- $H_0 : p = p_0; H_1 : p < p_0$

Statystyka testowa:

$$z = \frac{\left(\frac{m}{n} - p_0\right)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Statystyka testowa ma rozkład normalny standaryzowany.

Przykład

W pewnym mieście przeprowadzono badanie aktywności zawodowej. Przebadano 980 osób, spośród których 674 zadeklarowały się jako osoby pracujące. Na poziomie istotności 0,1 zweryfikuj hipotezę, że odsetek pracujących w tym mieście jest równy 72%.

Układ hipotez jest następujący:

- $H_0 : p = 0,72$
- $H_1 : p \neq 0,72$

W pierwszej kolejności obliczamy wartość średnią, odchylenie standardowe i na tej podstawie wartość statystyki testowej:

$$z = \frac{\left(\frac{m}{n} - p_0\right)}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0,69 - 0,72)}{\sqrt{\frac{0,72(1-0,72)}{980}}} = -2,25$$

Natomiast wartość kwantyla rozkładu normalnego dla poziomu istotności $\alpha = 0,1$ wynosi 1,64 (ROZKŁ.NORMALNY.S.ODWR(1-0,1/2)). W takim razie obszar odrzucenia hipotezy zerowej to przedział od $-\infty$ do -1,64 i od 1,64 do $+\infty$. Wartość statystyki testowej znajduje się w tym przedziale, zatem są podstawy do odrzucenia hipotezy zerowej - odsetek pracujących w całym mieście istotnie różni się od 72%.

Zadania

1. Wysunięto przypuszczenie, że palacze papierosów stanowią 40% populacji. W celu sprawdzenia tej hipotezy wylosowano 500 osób. Okazało się, że wśród nich było 230 palaczy. Zweryfikować postawioną hipotezę na poziomie istotności 0,05.
2. W pewnym powiecie na 119 przedsiębiorstw z sekcji PKD C w badaniu DG 1 wzięło udział 14 przedsiębiorstw. Na poziomie istotności 0,05 zweryfikuj hipotezę, że odsetek przedsiębiorstw biorących udział w badaniu wynosi 10%.

3.6.3 Test t dla dwóch średnich niezależnych

Test ma za zadanie sprawdzić czy średnie w dwóch grupach różnią się od siebie w sposób istotny statystycznie. Możemy rozpatrywać następujące układy hipotez:

- $H_0 : m_1 = m_2; H_1 : m_1 \neq m_2$
- $H_0 : m_1 = m_2; H_1 : m_1 > m_2$
- $H_0 : m_1 = m_2; H_1 : m_1 < m_2$

A statystyka testowa ma następującą postać:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Statystyka testowa ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody.

Przykład

Na grupie 50 mężczyzn i kobiet przeprowadzono badanie dotyczące zarobków i zebrano następujące wyniki:

- Kobiety: 4330, 3063, 3012, 3486, 3415, 3097, 2451, 3418, 2970, 4050, 2828, 4076, 3011, 3575, 3939, 3089, 3733, 3347, 2719, 3238, 4372, 3272, 2909, 3368, 3598
- Mężczyźni: 4182, 4258, 3840, 4266, 3494, 2862, 3611, 3594, 2874, 4025, 3486, 3710, 3165, 4019, 4556, 3449, 3755, 4579, 4174, 4565, 3798, 3739, 3596, 4374, 3286

Czy średnie zarobki różnią się w grupach płci? Przyjmij poziom istotności równy 0,1.

- H_0 : średnie w grupach płci są takie same
- H_0 : średnie w grupach płci różnią się

W Excelu najprościej przeprowadzić ten test wykorzystując funkcję `T.TEST` zakładając rozkład dwustronny oraz równość wariancji. Funkcja w rezultacie zwraca wartość p, która wynosi 0,003034 i jest mniejsza od przyjętego poziomu istotności $\alpha = 0,1$, co oznacza, że są podstawy do odrzucenia hipotezy zerowej. Średnie w grupach płci różnią się między sobą.

3.6.4 Test istotności dwóch proporcji

Testowany jest jeden z poniższych układów hipotez:

- $H_0 : p_1 = p_2; H_1 : p_1 \neq p_2$
- $H_0 : p_1 = p_2; H_1 : p_1 > p_2$
- $H_0 : p_1 = p_2; H_1 : p_1 < p_2$

Statystyka testowa:

$$Z = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{\bar{p}\bar{q}}{n}}}$$

gdzie:

- $\bar{p} = \frac{m_1+m_2}{n_1+n_2}$
- $\bar{q} = 1 - \bar{p}$
- $n = \frac{n_1 \cdot n_2}{n_1+n_2}$

Statystyka testowa ma rozkład normalny standaryzowany.

Chapter 4

Korelacje

Korelacja [łac.], mat. wzajemne powiązanie, współzależność zjawisk lub obiektów; w teorii prawdopodobieństwa i statystyce — współzależność liniowa zmiennych losowych (jej liczbową miarą jest współczynnik korelacji) [źródło: słownik PWN].

4.1 Cechy jakościowe

Celem analizy współzależności jest określenie siły związku pomiędzy dwiema cechami jakościowymi. Sprawdźmy czy istnieje zależność pomiędzy wynikiem z egzaminu a płcią?

Podstawą takiej analizy jest tablica kontyngencji albo tablica krzyżowa. W przypadku obserwacji statystycznej dotyczącej dużej ilości zmiennych, operowanie wartościami szczegółowymi jest uciążliwe. W celu stwierdzenia istnienia lub braku związku korelacyjnego konstruuje się tablicę korelacyjną. Na skrzyżowaniu kolumn z wierszami wpisuje się liczebności jednostek zbiorowości statystycznej, u których zaobserwowano jednocześnie występowanie określonej wartości x_i i y_i . Ogólna postać tablicy krzyżowej jest następująca:

cecha X / cecha Y	y_1	y_2	\cdots	y_j	\cdots	y_r	\sum_j
x_1	n_{11}	n_{12}	\cdots	n_{1j}	\cdots	n_{1r}	$n_{1.}$
x_2	n_{21}	n_{22}	\cdots	n_{2j}	\cdots	n_{2r}	$n_{2.}$
\cdots	\cdots	\cdots	\cdots	\cdots	\cdots	\cdots	\cdots
x_i	n_{i1}	n_{i2}	\cdots	n_{ij}	\cdots	n_{ir}	$n_{i.}$
\cdots	\cdots	\cdots	\cdots	\cdots	\cdots	\cdots	\cdots
x_k	n_{k1}	n_{k2}	\cdots	n_{kj}	\cdots	n_{kr}	$n_{k.}$
\sum_i	$n_{.1}$	$n_{.2}$	\cdots	$n_{.j}$	\cdots	$n_{.r}$	n

Jak wynika z tablicy zmienna losowa X przyjmuje k wariantów ($i=1,2,\dots,k$), zaś zmienna losowa Y przyjmuje r wariantów ($j=1,2,\dots,r$).

Symbol $n_{.j}$ oznacza liczbę jednostek, które mają wariant y_j zmiennej Y , natomiast symbol $n_{i.}$ - liczbę jednostek, które mają wariant x_i zmiennej X . Symbole n_{ij} oznaczają liczbę jednostek, które posiadają jednocześnie wariant x_i cechy X i warianty y_j cechy Y . Symbol n oznacza liczebność próby, przy czym:

$$n = \sum_{i=1}^k n_{i.} = \sum_{j=1}^r n_{.j} = \sum_{i=1}^k \sum_{j=1}^r n_{ij}$$

W analizowanym przykładzie pozyskaliśmy informację od 500 osób na temat wyniku egzaminu oraz płci. Tablica krzyżowa tych danych wygląda następująco:

Płeć / Wynik	Nie zdany	Zdany	Suma
Mężczyzna	100	70	170
Kobieta	130	200	330
Suma	230	270	500

Do odpowiedzi na pytanie czy istnieje zależność pomiędzy tymi cechami wykorzystamy statystykę chi-kwadrat (χ^2). Nazwę tej statystyki czytamy tak samo jak piszemy.

W pierwszym kroku musimy obliczyć **oczekiwane (teoretyczne) częstości** dla każdej komórki czyli wartości jakie musiałyby występować, żeby zależności nie było. Wzór na liczebności teoretyczne jest następujący:

$$\hat{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

Przykładowo, liczebność teoretyczna dla mężczyzn, którzy nie zdali egzaminu to iloczyn liczby wszystkich mężczyzn i liczby wszystkich, którzy nie zdali egzaminu podzielony przez wszystkie obserwacje:

$$\hat{n}_{11} = \frac{170 \cdot 230}{500} = 78,2$$

Liczebności oczekiwane po wstawieniu do tabeli:

Płeć / Wynik	Nie zdany	Zdany	Suma
Mężczyzna	78,2	91,8	170
Kobieta	151,8	178,2	330
Suma	230	270	500

Częstości teoretyczne nie muszą być wartościami całkowitymi, ale suma w wierszu, kolumnie i dla całej tablicy krzyżowej musi pozostać taka sama.

W kolejnym kroku wyznaczamy **standardowe współczynniki różnicy** pomiędzy częstościami oczekiwanymi a zaobserwowanymi. Suma tych współczynników da nam wartość **statystyki** χ^2 .

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

gdzie:

- r - liczba wariantów cechy Y,
- k - liczba wariantów cechy X,
- n_{ij} - liczebności empiryczne dla i-tego wariantu cechy X i j-tego wariantu cechy Y,
- \hat{n}_{ij} - liczebności teoretyczne dla i-tego wariantu cechy X i j-tego wariantu cechy Y.

$$\chi^2 = \frac{(100 - 78,2)^2}{78,2} + \frac{(70 - 91,8)^2}{91,8} + \frac{(130 - 151,8)^2}{151,8} + \frac{(200 - 178,2)^2}{178,2} = 17$$

W przypadku, gdy dysponujemy tablicą o wymiarach 2x2 możemy skorzystać z prostszego sposobu wyznaczenia statystyki χ^2 korzystając ze wzoru. Jeśli tablica kontyngencji jest w postaci:

cecha X / cecha Y	y_1	y_2	$n_{i.}$
x_1	a	b	a+b
x_2	c	d	c+d
$n_{.j}$	a+c	b+d	n

a liczebności w komórkach większe niż 5 to wzór na chi-kwadrat jest następujący:

$$\chi^2 = \frac{n \cdot (a \cdot d - b \cdot c)^2}{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}$$

W przypadku występowania częstości mniejszych od 5 musimy zastosować wzór uwzględniający poprawkę Yatesa:

$$\chi^2 = \frac{n \cdot (|a \cdot d - b \cdot c| - 0,5 \cdot n)^2}{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}$$

W analizowanym przypadku zależności płci i wyniku z egzaminu wartości w komórkach a, b, c i d są większe od 5, więc można wykorzystać pierwszą formułę:

$$\chi^2 = \frac{500 \cdot (100 \cdot 200 - 70 \cdot 130)^2}{170 \cdot 330 \cdot 230 \cdot 270} = \frac{59405000000}{3483810000} = 17$$

Sama wartość statystyki chi-kwadrat nie informuje o sile zależności pomiędzy analizowanymi zmiennymi. W celu określenia siły zależności musimy wyznaczyć jedną z dostępnych miar korelacji: **współczynnik V-Cramera** lub **współczynnik zbieżności T-Czuprowa**:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, r-1)}}$$

$$T = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(k-1)(r-1)}}}$$

gdzie:

- χ^2 - wyznaczona wartość statystyki χ^2 ,
- n - liczba wszystkich obserwacji,
- k - liczba kolumn tabeli kontyngencji bez sumy (liczba wariantów pierwszej cechy),
- r - liczba wierszy tabeli kontyngencji bez sumy (liczba wariantów drugiej cechy).

Możemy przyjąć pewne umowne progi dotyczące interpretacji tych miar:

- od 0,00 do 0,29 - słaby związek pomiędzy zmiennymi,
- od 0,30 do 0,49 - umiarkowany związek pomiędzy zmiennymi,
- od 0,50 do 1,00 - silny związek pomiędzy zmiennymi.

W naszym przykładzie współczynnik V-Cramera jest równy współczynnikowi T-Czuprowa i wynosi $V = T = 0,18$, co oznacza, że pomiędzy płcią a wynikiem z egzaminu występuje słaba zależność.

Miary korelacji cech jakościowych w Excelu

Niestety, aktualnie w MS Excel nie ma funkcji, które umożliwiają obliczenie statystyki χ^2 . Istnieje natomiast funkcja **CHI.TEST**, która liczy statystykę χ^2 ale jej nie zwraca. Informuje jedynie o istotności statystycznej, która wykracza poza zakres materiału realizowanego w ramach tych zajęć.

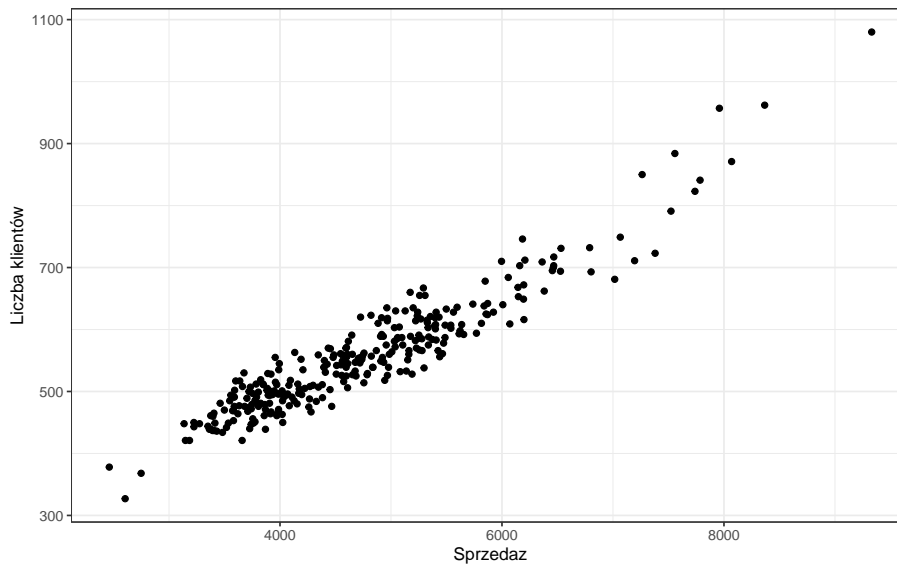
Zadanie

Oceń siłę zależności pomiędzy faktem wejścia promocji, a liczbą klientów (poniżej/powyżej średniej) w sklepie numer 7.

4.2 Cechy ciągłe

W odróżnieniu do cech jakościowych, w przypadku cech ciągłych oprócz siły zależności określamy także kierunek tej zależności pomiędzy dwoma zmiennymi. Silnie skorelowane ze sobą zmienne zachowują się “jak gdyby równocześnie się poruszały”.

Sprawdźmy czy liczba klientów jest skorelowana ze sprzedażą w sklepie nr 1. Pierwszym krokiem w analizie korelacji jest stworzenie wykresu rozrzutu:



Na tej podstawie możemy już stwierdzić, że zależność jest dodatnia - wzrost wartości jednej cechy pociąga za sobą wzrost wartości drugiej cechy:

- wzrost temperatury, większa sprzedaż lodów;
- wzrost wynagrodzenia, zwiększenie wydatków;
- mniej czasu spędzonego na działaniach marketingowych, mniej klientów.

Więcej przykładów

Z korelacją ujemną mielibyśmy do czynienia, gdy wartości jednej cechy by rosły, a drugiej malały. Przykłady ujemnej korelacji:

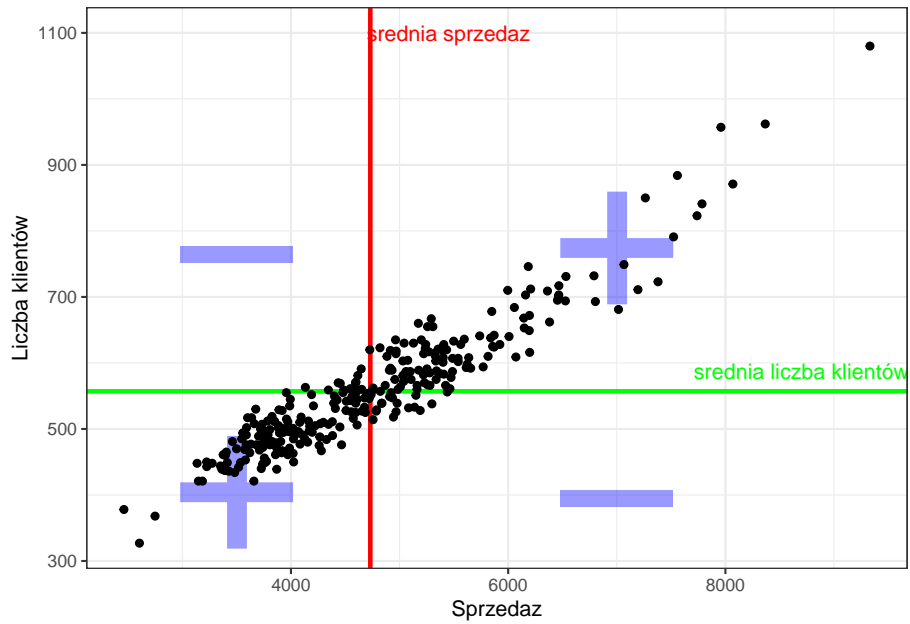
- liczba nieobecności na zajęciach jest zwykle związana z niższymi ocenami;
- większa prędkość pociągu, krótszy czas dotarcia do stacji końcowej;
- spadek temperatury, wzrost sprzedaży grzejników.

Więcej przykładów

Wartością liczbową, która określa kierunek korelacji jest **kowariancja**. Wyznaczenie kowariancji polega na policzeniu różnic wartości obu cech od średniej, a następnie ich przemnożeniu i uśrednieniu, zgodnie ze wzorem:

$$\text{cov}(x, y) = \text{cov}(y, x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Znak kowariancji determinuje kierunek zależności, który jest ustalany na podstawie iloczynu różnic pomiędzy wartościami średnich a analizowanymi cechami. Sumowane są wartości dodatnie i ujemne, co pokazuje poniższy wykres:



Jeśli kowariancja będzie:

- $\text{cov}(x, y) = 0$ — brak zależności,
- $\text{cov}(x, y) < 0$ — ujemna zależność,
- $\text{cov}(x, y) > 0$ — dodatnia zależność.

W przypadku sklepu nr 1 kowariancja wynosi 94843, co oczywiście pociąga za sobą dodatnią zależność. Na podstawie kowariancji nie możemy natomiast wyznaczyć siły zależności ponieważ jest wyznaczona w dziwnych jednostkach - osobo-euro. Poza tym może przyjąć wartości z całego zakresu liczb rzeczywistych: $(-\infty; +\infty)$.

Standaryzując kowariancję z wykorzystaniem odchylenia standardowego każdej cechy otrzymamy **współczynnik korelacji liniowej Pearsona**:

$$r_{xy} = r_{yx} = \frac{\text{cov}(X, Y)}{S_x \cdot S_y}$$

lub

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}}$$

Współczynnik ten jest wielkością unormowaną, przyjmuje wartości z przedziału $r \in (-1; 1)$.

Jeśli:

- $r_{xy} = 1$ — korelacja dodatnia doskonała,
- $0 < r_{xy} < 1$ — korelacja dodatnia niedoskonała (słaba/umiarkowana/silna)
- $r_{xy} = 0$ — brak zależności,
- $-1 < r_{xy} < 0$ — korelacja ujemna niedoskonała (słaba/umiarkowana/silna)
- $r_{xy} = -1$ — korelacja ujemna doskonała.

W sklepie nr 1 współczynnik korelacji liniowej Pearona wynosi 0.94, co oznacza, że pomiędzy wartością sprzedaży a liczbą klientów występuje silna dodatnia korelacja liniowa.

W przypadku, gdy w zbiorze analizowanych cech znajdują się wartości odstające, które zaburzają liniowość relacji, współczynnik korelacji liniowej może nie spełniać swojej funkcji. Wówczas należy skorzystać ze **współczynnika korelacji rang Spearmana**, który jest współczynnikiem korelacji liniowej Pearsona, ale obliczanym na **rangach**.

Rangowanie polega na posortowaniu wartości jednej cechy rosnąco - przypisanie kolejnych wartości od 1 do n (jak w sporcie), na następnie powtórzenie operacji dla drugiej cechy. Jeśli jakaś wartość będzie się powtarzać (*ex aequo*), wówczas wyznaczamy wartość tzn. rangi wiązanej - średniej arytmetycznej z rang tej wartości.

W Excelu można wyznaczyć wyłącznie współczynnik korelacji liniowej Pearsona korzystając z funkcji:

- PEARSON(tablica1, tablica2).

Korelacja nie oznacza przyczynowości - informuje jedynie o współwystępowaniu cech. Do analizy przyczynowo-skutkowej służą metody regresji.

Kilka dodatkowych wyjaśnień

Pozorne korelacje

Gra - zgadnij współczynnik korelacji

Zadanie

Przeanalizuj sprzedaż oraz liczbę klientów w sklepie nr 7 w maju. Utwórz wykres rozrzutu oraz oblicz i zinterpretuj współczynnik korelacji liniowej Pearsona korzystając z funkcji Excela.

Chapter 5

Regresja

Metoda regresji wykorzystywana jest do funkcyjnego odwzorowania zależności pomiędzy badanymi zmiennymi. Jej celem jest poszukanie określonej klasy funkcji, która w możliwie najlepszy sposób charakteryzowałaby zależność pomiędzy zmiennymi. Funkcję tą określa się mianem funkcji regresji. Budując model chcemy osiągnąć określone cele poznawcze.

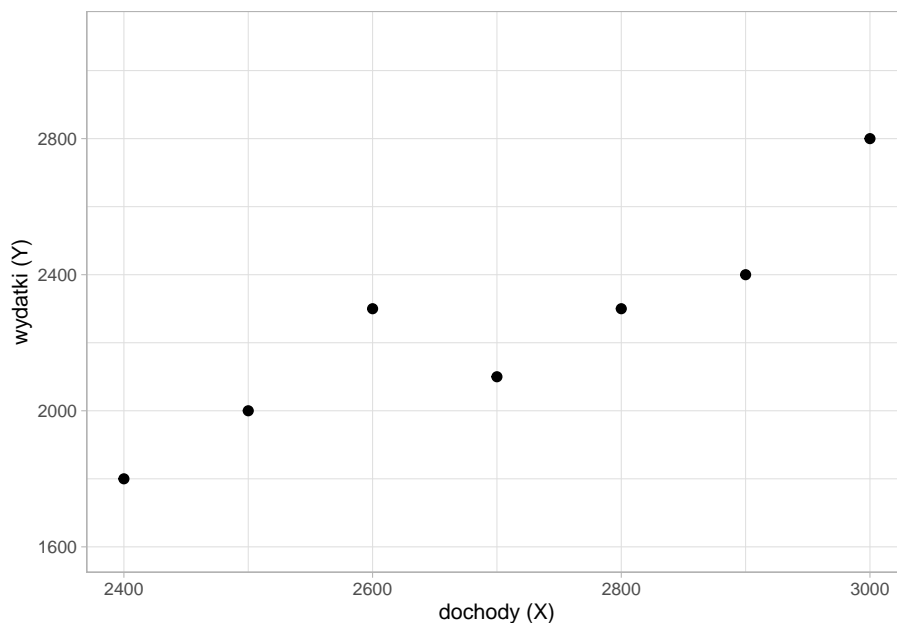
5.1 Regresja prosta

Celem regresji jest zbudowanie na podstawie dostępnych informacji modelu opisującego rzeczywistość. Taki model pełni funkcję poznawczą - dostarcza wiedzy na temat zjawiska, a także umożliwia prognozowanie (predykcję) nieznannej wartości analizowanej cechy.

Weźmy pod uwagę prosty przykład dochodów i wydatków:

wydatki	dochody
2300	2600
1800	2400
2400	2900
2300	2800
2800	3000
2000	2500
2100	2700

Podobnie jak w analizie korelacji punktem wyjścia w regresji prostej jest utworzenie wykresu rozrzutu.



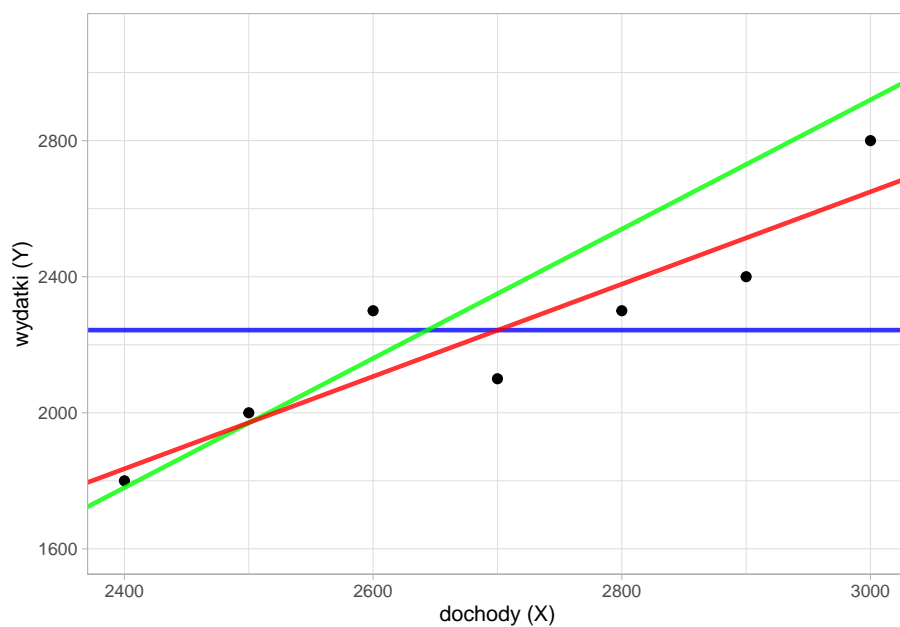
Ażeby móc zastosować model regresji musi występować związek korelacyjny pomiędzy zmiennymi oraz musi to być związek liniowy. Następnie na podstawie kryteriów merytorycznych określamy zmienną objaśnianą (y) oraz zmienną objaśniającą (x).

Zależność wydatków od dochodów wydaje się oczywista - za y przyjmujemy wydatki, a x to będą dochody. Dobrą praktyką jest umieszczanie zmiennej objaśnianej na osi OY, a zmiennej objaśniającej na osi OX. Interesuje nas tworzenie modelu upraszczającego rzeczywistość do poziomu wzoru na prostą, której ogólna postać jest następująca:

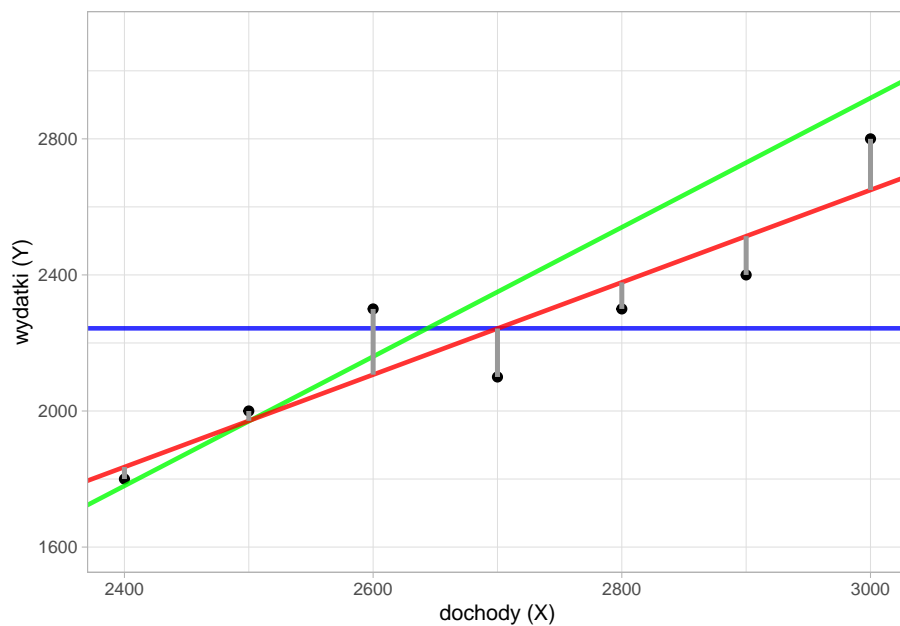
$$y_i = a_1 \cdot x_i + a_0$$

W przypadku tylko dwóch punktów wyznaczenie współczynników a_1 i a_0 nie stanowiłoby żadnego problemu. Natomiast dla podanego przykładu trzeba posłużyć się Klasyczną Metodą Najmniejszych Kwadratów (KMNK), w której minimalizujemy odległość punktów od dopasowywanej prostej.

Spróbujmy teraz dopasować kilka prostych - mogą one przebiegać na wiele różnych sposobów.



W następnym kroku obliczamy różnice pomiędzy istniejącymi punktami, a odpowiadającym im wartościami na prostej:



Oznaczając y_i jako rzeczywistą wartość wydatków i \hat{y}_i jako wartość leżącą

na prostej zależy nam na minimalizowaniu wyrażenia $\sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min$.

Różnica $y_i - \hat{y}_i$ jest nazywana resztą (ang. residual). Wyznaczając te wartości dla analizowanych przez nas prostych otrzymamy następujące wyniki:

name	suma_kwadratow_reszt
czerwona	101430
zielona	264300
niebieska	22462143

Jak możemy zauważyć najmniejsza wartość sumy kwadratów reszt obserwowana jest dla linii w kolorze czerwonym. Interesuje nas teraz wzór tej prostej. Przyjmując wcześniejsze oznaczenia ogólna postać prostej regresji jest następująca:

$$\hat{y}_i = a_1 x_i + a_0$$

gdzie y z daszkiem (\hat{y}) oznacza wartość teoretyczną, leżącą na wyznaczonej prostej.

Wobec tego wartości empiryczne/rzeczywiste (y) będą opisane formułą:

$$y_i = a_1 x_i + a_0 + u_i$$

w której u_i oznacza składnik resztowy wyliczany jako $u_i = y_i - \hat{y}_i$.

Model zależności wydatków od dochodu ma następującą postać:

$$\hat{y}_i = 1,357x_i - 1421,429$$

Po podstawieniu pierwszej wartości dochodu - 2400 zł do tego wzoru otrzymamy teoretyczną/modelową wartość wydatków:

$$\hat{y}_1 = 1,357 \cdot 2400 - 1421,429 = 1835,371$$

Ta wartość leży na czerwonej prostej i różni się od rzeczywistej wartości wydatków uzyskanych przez tę osobę, która wynosi 1800 zł. Różnica pomiędzy wartością rzeczywistą a modelową nazywana jest resztą i wynosi w tym przypadku:

$$u_1 = 1800 - 1835,371 = -35,371$$

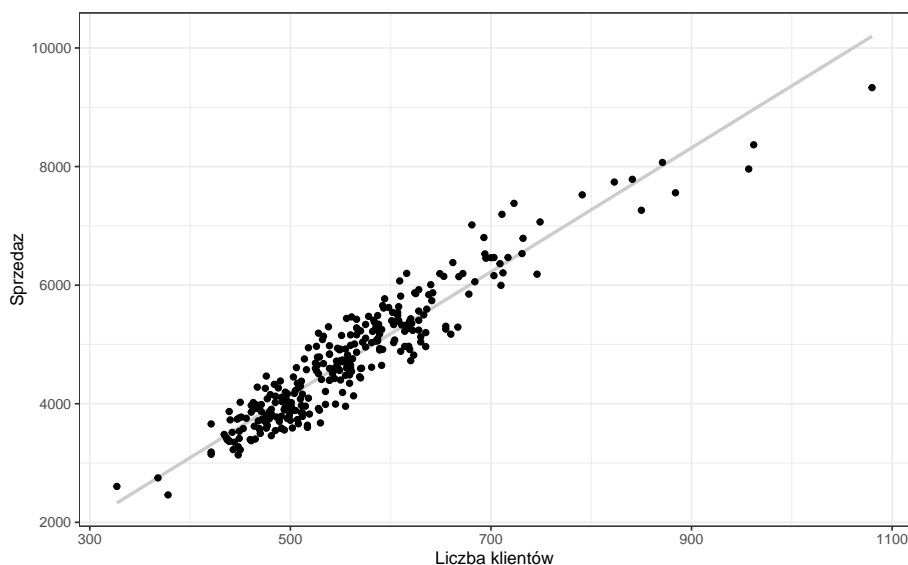
Można zatem powiedzieć, że stworzony model nie różni się zbyt od rzeczywistości w przypadku tej obserwacji. Na podstawie dwóch powyższych wartości możliwe jest wyznaczenie wartości rzeczywistej:

$$y_1 = 1,357 \cdot 2400 - 1421,429 - 35,371 = 1800$$

Wiedząc już jaka intuicja przyświeca analizie regresji przejdziemy do analizy wybranego sklepu Rossmann i na tej podstawie wyznaczymy parametry modelu, a także je zinterpretujemy. Stworzony model zostanie też wykorzystany do predykcji.

Na podstawie wartości sprzedaży oraz liczby klientów w danym sklepie Rossmann i chcielibyśmy wyznaczyć możliwy poziom sprzedaży przy danej liczbie klientów np. 1000 klientów.

W analizowanym przez nas przypadku sklepu Rossmann zmienną objaśnianą będzie poziom sprzedaży (y), który będziemy wyjaśniać liczbą klientów (x). Naszym celem jest znalezienie wzoru prostej, która będzie przebiegać możliwie najbliżej wszystkich punktów wykresu. Musimy wyznaczyć współczynnik kierunkowy tej prostej (a_1) oraz punkt przecięcia z osią OY (a_0).



Wartości tych współczynników możemy policzyć z wykorzystaniem następujących wzorów:

$$a_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

lub znając wartość współczynnika korelacji liniowej Pearsona:

$$a_1 = r \frac{S_y}{S_x}$$

z kolei wartość wyrazu wolnego można uzyskać ze wzoru:

$$a_0 = \bar{y} - a_1 \bar{x}$$

gdzie:

- r - współczynnik korelacji liniowej Pearsona pomiędzy cechą x i y ,
- S_y - odchylenie standardowe dla cechy y ,
- S_x - odchylenie standardowe dla cechy x ,
- \bar{y} - średnia dla cechy y ,
- \bar{x} - średnia dla cechy x .

Na tej podstawie ustalamy, że interesująca nas prosta ma następujący wzór:

$$\hat{y}_i = 10,45x_i - 1091,22$$

Współczynnik kierunkowy (a_1) informuje o ile przeciętnie zmieni się wartość zmiennej objaśnianej (y), gdy wartość zmiennej objaśniającej (x) wzrośnie o jednostkę. W naszym przypadku wzrost liczby klientów o 1 osobę spowoduje średni wzrost sprzedaży o 10,45 euro.

Z kolei wyraz wolny (a_0) to wartość zmiennej objaśnianej (y), w sytuacji w której wartość zmiennej objaśniającej (x) będzie równa 0. Należy zachować szczególną ostrożność przy interpretacji tego współczynnika, ponieważ często jest on pozbawiony sensu. W analizowanym przykładzie współczynnik a_0 informuje, że przy zerowej liczbie klientów sprzedaż w sklepie nr 1 wyniesie -1091,22 euro.

Kolejnym elementem analizy regresji jest ocena dopasowania modelu. W tym celu posługujemy się kilkoma miarami.

Pierwszą miarą, która opisuje dopasowanie funkcji regresji do danych empirycznych jest **odchylenie standardowe składnika resztowego**, które jest pierwiastkiem z sumy kwadratów reszt podzielonej przez liczbę obserwacji pomniejszoną o 2. To pomniejszenie mianownika wynika z faktu, że w modelu mamy dwa współczynniki a_1 i a_0 , które w ten sposób uwzględniamy. Formalnie można to zapisać w następujący sposób:

$$S_u = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

lub

$$S_u = \sqrt{\frac{\sum_{i=1}^n u_i^2}{n - 2}}$$

Miara ta określa, o ile, przeciętnie biorąc (+/-), wartości empiryczne zmiennej objaśnianej odchylają się od wartości teoretycznych tej zmiennej, obliczonej na podstawie funkcji regresji. W analizowanym przypadku możemy stwierdzić, że znane wartości sprzedaży odchylają się od wartości teoretycznych średnio o +/- 351,57 euro.

Odchylenie standardowe składnika resztowego jest także miarą błędu prognozy. Przykładowo, chcemy sprawdzić jak będzie kształtować się sprzedaż przy liczbie klientów równej 1000 osób. Po podstawieniu tej wartości do funkcji regresji otrzymamy:

$$y_{1000} = 10,45 \cdot 1000 - 1091,22 = 9358,78$$

Na tej podstawie stwierdzamy, że przy 1000 klientów prognozowana sprzedaż wyniosłaby 9358,78 euro +/- 351,57 euro.

Kolejna miara to **współczynnik zmienności resztowej**, który otrzymujemy poprzez podzielenie odchylenia standardowego składnika resztowego przez średni poziom cechy:

$$V_u = \frac{S_u}{\bar{y}} \cdot 100\%$$

Współczynnik ten wskazuje, jaki procent średniego poziomu zmiennej objaśnianej stanowią wahania losowe, których miarą jest S_u . Parametr V_u jest więc miernikiem relatywnej wielkości błędu losowego. Niektórzy autorzy postulują, że błąd ten można umownie uznać za dopuszczalny, jeśli $V_u < 15\%$. Należy się jednak wystrzegać przed „dogmatycznym” podejściem do oceny modeli regresji i jedynie słusznych progów.

W naszym przypadku ten współczynnik będzie równy $V_u = \frac{351,57}{4730,72} \cdot 100\% = 7\%$ co oznacza, że 7% średniego poziomu sprzedaży stanowią wahania losowe.

Równie ważną miarą dopasowania funkcji regresji do danych empirycznych jest **współczynnik determinacji** lub bardziej potocznie **współczynnik r kwadrat** — od symbolu, którym jest oznaczany. Współczynnik ten obliczany jest na podstawie reszt z modelu oraz odchylen wartości empirycznych od średniej:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

lub jako kwadrat współczynnika korelacji liniowej Pearsona:

$$R^2 = r_{xy}^2$$

Określa, jaki procent wariancji zmiennej objaśnianej został wyjaśniony przez funkcję regresji. R^2 przyjmuje wartości z przedziału $< 0; 1 >$ ($< 0\%; 100\% >$), przy czym model regresji tym lepiej opisuje zachowanie się badanej zmiennej objaśnianej, im R^2 jest bliższy jedności (bliższy 100%)

Analizowany przez nas model regresji jest bardzo dobry: $R^2 = 0,89$, co oznacza, że oszacowany model regresji wyjaśnia 89% zmienności sprzedaży.

Przeciwnieństwem współczynnika determinacji R^2 jest **współczynnik zbieżności (indeterminacji)**. Tę miarę można wyznaczyć korzystając ze wzoru:

$$\varphi^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$$

bądź odejmując od wartości 1 wartość współczynnika R^2 :

$$\varphi^2 = 1 - R^2$$

Współczynnik zbieżności φ^2 określa, jaka część wariancji badanej zmiennej objaśnianej nie została wyjaśniona przez funkcję regresji. Oczywiście jest więc, że korzystna sytuacja występuje wówczas, gdy φ^2 jest bliższy zera.

W przyjętym przez nas modelu regresji $\phi^2 = 11\%$, co oznacza, że 11% zmienności sprzedaży nie została wyjaśniona przez funkcję regresji. Można także powiedzieć, że 11% zmienności sprzedaży stanowią czynniki losowe nie wyjaśniane przez funkcję regresji.

Ostatnim elementem analizy jest ocena jakości parametrów funkcji regresji a_1 i a_0 . Równanie regresji wyznaczyliśmy na podstawie dostępnych danych, ale nie znamy równania tej prostej w populacji. W związku z czym mogliśmy się trochę pomylić przy obliczaniu współczynników a_1 i a_0 . W celu oceny skali tych błędów wyznacza się **błędy średnie szacunku ocen parametrów funkcji regresji** według wzorów:

$$S_{a_1} = \frac{S_u}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

oraz

$$S_{a_0} = \sqrt{\frac{S_u^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}$$

Błędy te wskazują, o ile, przeciętnie biorąc (+/-), odchylają się oceny parametrów modelu regresji od ich wartości prawdziwych. Jest oczywiście pożądane, żeby te błędy były możliwie jak najmniejsze. W związku z powyższym przyjmuje się, że ilorazy:

$$V_{a_1} = \frac{S_{a_1}}{a_1}$$

$$V_{a_0} = \frac{S_{a_0}}{a_0}$$

nie powinny przekraczać wartości 0,5 (50%) w wartości bezwzględnej.

Jest to szczególnie istotne w przypadku parametru współczynnika kierunkowego a_1 , natomiast dla wyrazu wolnego a_0 ta własność nie musi być spełniona.

W analizowanym przez nas modelu wartość parametru a_1 odchyła się od jego wartości prawdziwej o +/- 0,21 co stanowi 2% wartości tego parametru. Z kolei wartość parametru a_0 odchyła się od jego wartości prawdziwej o +/- 119,81 co stanowi 11% wartości tego parametru.

Regresja prosta w Excelu

• Sposób nr 1

Parametry funkcji regresji można także wyznaczyć korzystając z wbudowanej funkcji programu Excel — REGLINP. Składnia jest następująca:

- REGLINP(wektor_y; wektor_x; stała; statystyka)

gdzie:

- wektor_y — zestaw wartości zmiennej objaśnianej (y),
- wektor_x — zestaw wartości zmiennej objaśniającej (y),
- stała — jeśli podamy wartość 1 to wyraz wolny jest obliczany normalnie, jeśli podamy 0 to zostanie oszacowany model bez wyrazu wolnego,
- statystyka — jeśli argument ma wartość 1 to funkcja REGLINP zwraca dodatkowe statystyki regresji, natomiast jeśli ma wartość 0 to funkcja zwraca tylko wartości współczynnika kierunkowego oraz wyrazu wolnego.

Po napisaniu funkcji i uwzględnieniu wszystkich argumentów naciskamy ENTER — powinna pojawić się jedna wartość. Następnie należy zaznaczyć obszar 2 kolumny na 5 wiersze uwzględniając w lewej górnej komórce otrzymaną wcześniej wartość. W kolejnym kroku przechodzimy do PASKU FORMUŁY programu Excel i korzystamy z tajemnej formuły CTRL+SHIFT+ENTER.

W rezultacie otrzymujemy tabelę o wymiarach 2x5, która zawiera następujące elementy:

Współczynnik kierunkowy (a_1)	Wyraz wolny (a_0)
-----------------------------------	-----------------------

Średni błąd szacunku parametru (S_{a_1})	Średni błąd szacunku parametru (S_{a_0})
Współczynnik determinacji (R^2)	Odchylenie standardowe składnika resztowego (S_u)
Statystyka F (F)	Liczba stopni swobody ($n - 2$)
Regresyjna suma kwadratów ($\sum (\hat{y} - \bar{y})^2$)	Suma kwadratów reszt ($\sum (y - \hat{y})^2$)

• Sposób nr 2

Zaznaczamy punkty na wykresie rozrzutu i klikamy prawym przyciskiem myszy. Wybieramy **Dodaj linię trendu**, a następnie zaznaczamy opcje Wyświetl równanie na wykresie oraz Wyświetl wartości R-kwadrat na wykresie.

• Sposób nr 3

Do wyznaczenia parametrów regresji można także wykorzystać graficzne środowisko analizy danych. W tych celu wybieramy zakładkę DANE i po prawej stronie ANALIZA DANYCH. W menu zaznaczamy REGRESJA i klikamy OK. W opcjach wejścia zaznaczamy:

- Zakres wejściowy Y — zestaw wartości zmiennej objaśnianej (y),
- Zakres wejściowy X — zestaw wartości zmiennej objaśniającej (x),
- Tytuły — jeśli zostały zaznaczone kolumny wraz z nagłówkami.

W opcjach wyjścia określamy miejsce wyświetlenia wyniku: bieżący arkusz/nowy arkusz/nowy skoroszyt.

W rezultacie otrzymujemy następujący wynik:

PODSUMOWANIE - WYJŚCIE

Statystyki regresji	
Wielokrotność R	r
R kwadrat	R^2
Dopasowany R kwadrat	
Błąd standardowy	S_u
Obserwacje	n

ANALIZA WARIANCJI

	df	SS	MS	F	Istotność F
Regresja	1	$\sum (\hat{y} - \bar{y})^2$		F	
Resztkowy	$n - 2$	$\sum (y - \hat{y})^2$			
Razem	$n - 1$				

	Współczynniki	Błąd standardowy	t Stat	Wartość-p
Przecięcie	a_0	S_{a_0}		
zmienna x	a_1	S_{a_1}		

Zależności:

- Jeżeli ze wzoru na odchylenie standardowe składnika resztowego usuniemy pierwiastek to otrzymamy wariancję składnika resztowego, którą należy najpierw spierwiastkować, aby móc przeprowadzić interpretację.
- Jeśli bardzo chcemy policzyć wartość odchylenia standardowe składnika resztowego na podstawie wartości surowych to wartość licznika możemy odczytać z funkcji REGLINP — 5 wiersz, 2 kolumna. Wówczas wystarczy podzielić tę wartość przez 4 wiersz drugiej kolumny i spierwiastkować, aby otrzymać wartość S_u . Podobnie postępujemy, jeśli korzystamy z narzędzia REGRESJA.

Zadania

Ilu klientów powinno przyjść do sklepu nr 7, żeby możliwe było osiągnięcie sprzedaży na poziomie 20000 euro? Zapisz uzyskany model, zinterpretuj parametry regresji oraz oceń jakość dopasowania.

5.2 Trend liniowy

Oprócz określania nieznanymi wartości cechy, regresja jest także wykorzystywana do prognozowania w czasie. Przykładowo mając dane dotyczące miesięcznej sprzedaży w roku 2014 spróbujemy określić możliwą sprzedaż w wybranym miesiącu 2015 roku.

Sposób postępowania jest bardzo podobny do regresji prostej z tym, że zamiast wartości cechy x mamy kolejne numery okresów $t = 1, 2, 3, \dots, n$. Wówczas równanie trendu ma następującą postać:

$$\hat{y}_i = a_1 t_i + a_0$$

Po oszacowaniu parametrów a_1 i a_0 nieco inaczej je zinterpretujemy. Wartość parametru a_1 informuje o średniej zmianie cechy y z okresu na okres, z kolei a_0 to wartość wynikająca z modelu dla okresu poprzedzającego analizę.

Spróbujemy określić możliwą sprzedaż w styczniu 2015 roku dla sklepu nr 7. W tym zagregowaliśmy dane do postaci miesięcznej. Model regresji ma następującą postać:

$$\hat{y}_i = 4850t_i + 196592$$

Wynika z niego, że z miesiąca na miesiąc sprzedaż rosła średnio o 4850 euro. Natomiast możliwa sprzedaż w grudniu 2013 roku wynosiła 196592 euro.

Ocena jakości modelu przebiega analogicznie jak w przypadku regresji prostej. Inaczej wyznacza się błąd prognozy, co wynika z faktu, że im bardziej “oddalimy” się od okresu na podstawie którego oszacowaliśmy parametry trendu, tym błąd prognozy będzie większy.

Przy obliczaniu błędu prognozy korzystamy ze wzoru:

$$D(y_T^P) = S_u \sqrt{1 + \frac{1}{n} + \frac{(T - \bar{t})^2}{\sum_{t=1}^n (t_i - \bar{t})^2}}$$

gdzie:

- S_u — odchylenie standardowe składnika resztowego,
- n — liczba znanych okresów,
- \bar{t} — średnia z numerów okresów,
- T — numer okresu, na który stawiana jest prognoza.

Odchylenie standardowe składnika resztowego wynosiło 20031 euro, co oznacza, że znane wartości miesięcznej sprzedaży odchylają się od wartości wynikających z trendu średnio o ± 20031 euro.

Wyznaczając błąd prognozy musimy uwzględnić dodatkowy składnik uwzględniający czas. W związku z tym prognozowana sprzedaż w styczniu 2015 roku wyniesie:

$$\hat{y}_{13} = 4850 \cdot 13 + 196592 = 259639$$

a błąd prognozy:

$$D(y_{13}^P) = 20031 \sqrt{1 + \frac{1}{12} + \frac{(13 - 6,5)^2}{143}} = 23521$$

Czyli prognozowana miesięczna sprzedaż w styczniu 2015 roku wyniesie 259 639 euro $\pm 23\,521$ euro.

Natomiast dla lutego wartość błędu prognozy będzie już większa:

$$\hat{y}_{14} = 4850 \cdot 14 + 196592 = 264489$$

$$D(y_{14}^P) = 20031 \sqrt{1 + \frac{1}{12} + \frac{(14 - 6,5)^2}{143}} = 24341$$

Zadania

Ile wynosi prognozowana miesięczna sprzedaż (oraz błąd prognozy) w sklepie nr 5 w kwietniu 2015 roku.

Zadanie egzaminacyjne

Postanowiono zbadać zależność pomiędzy dzienną liczbą klientów (w tys. osób) a dziennym przychodem (w tys. zł) w pewnej sieci sklepów. W tym celu wybrano 16 sklepów (po jednym z każdego województwa), w których badano te wielkości. Analiza wykazała, że średnia liczba klientów wynosiła 1,5 tys. osób, a dziennego przychodu 40 tys. zł. Współczynnik zmienności dla liczby klientów był równy 13%, a dla przychodu 17%. Suma kwadratów reszt wynosiła 70, a współczynnik korelacji liniowej Pearsona 0,89.

- Wyznacz parametry funkcji regresji i zapisz jej postać.
- Oceń jakość otrzymanej funkcji.
- Ile wynosi prognozowany dzienny przychód (oraz błąd prognozy) dla sklepu obsługującego dziennie 800 osób?

Rozwiązanie

Dane:

- $\bar{x} = 1,5$
- $\bar{y} = 40$
- $V_x = 0,13$
- $V_y = 0,17$
- $\sum (y - \hat{y})^2 = 70$
- $n = 16$
- $r = 0,89$

Na podstawie powyższych danych wyznaczamy $S_x = V_x \cdot \bar{x} = 0,13 \cdot 1,5 = 0,195$

oraz $S_y = V_y \cdot \bar{y} = 0,17 \cdot 40 = 6,8$. Następnie $S_u = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{70}{16-2}} = 2,24$.

Na podstawie odpowiednich wzorów wyznaczamy parametry regresji: $a_1 = r \cdot \frac{S_y}{S_x} = 0,89 \cdot \frac{6,8}{0,195} = 31,04$ oraz $a_0 = \bar{y} - a_1 \cdot \bar{x} = 40 - 31,04 \cdot 1,5 = -6,55$. W związku z tym funkcja regresji ma następującą postać:

$$\hat{y} = 31,04 \cdot x - 6,55$$

Jakość modelu oceniamy na podstawie wartości S_u oraz $R^2 = r^2 = 0,89^2 = 0,79$.

Dzienny przychód dla sklepu obsługującego dziennie 800 osób będzie wynosił $\hat{y}(0,8) = 31,04 \cdot 0,8 - 6,55 = 18,27$ tys. zł $+/- 2,24$ tys. zł.

Chapter 6

Sezonowość

Jednym z rodzajów szeregu statystycznego jest szereg czasowy, który można zdefiniować jako ciąg obserwacji jakiegoś zjawiska w kolejnych jednostkach czasu (latach, kwartałach, miesiącach). Rozważane zjawisko może podlegać pewnym prawidłowościom, których wykrycie i opis jest celem analizy szeregów czasowych. Najczęściej rozważa się cztery czynniki wpływające na rozwój zjawiska w czasie:

- trend (T_t) — długookresowe, systematyczne zmiany, jakim podlega dane zjawisko,
- wahania sezonowe (S_t) — regularne odchylenia od tendencji rozwojowej (trendu) związane np. z porami roku (warunkami klimatycznymi),
- wahania cykliczne (C_t) — związane z cyklem koniunkturalnym,
- wahania przypadkowe (I_t) — nieregularne zmiany.

Analiza danych, które mogą charakteryzować się sezonowością rozpoczyna się od wizualizacji oraz estymacji parametrów modelu liniowego. W tym celu posłużymy się dwoma przykładami. Pierwszy będzie dotyczył zużycia energii elektrycznej, a drugi przewozów ładunków w Polsce - plik.

W obu przypadkach dysponujemy danymi kwartalnymi za lata 2003–2005. Na pierwszy rzut oka możemy wskazać pewne prawidłowości: zużycie energii jest widocznie wyższe w drugich i czwartych kwartałach analizowanych lat. Z kolei przewozy ładunków wzrastają od kwartału pierwszego do trzeciego (w którym osiągają maksimum w danym roku), by następnie spaść.

Celem analizy będzie ilościowe określenie wielkości zmian sezonowych, tak aby było możliwe prognozowanie z uwzględnieniem tych czynników.

6.1 Trend liniowy

Pierwszym krokiem w analizie szeregu czasowego jest estymacja parametrów trendu liniowego.

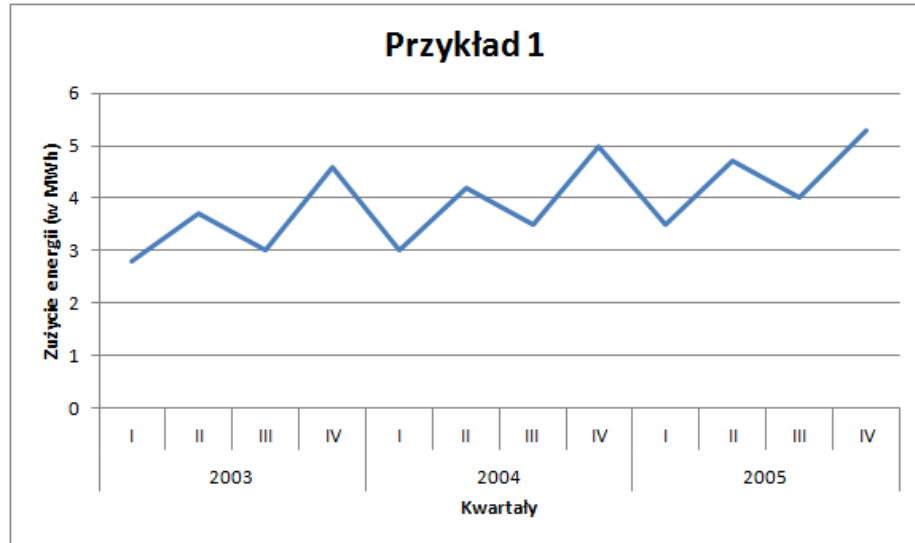


Figure 6.1: Zużycie energii - dane oryginalne

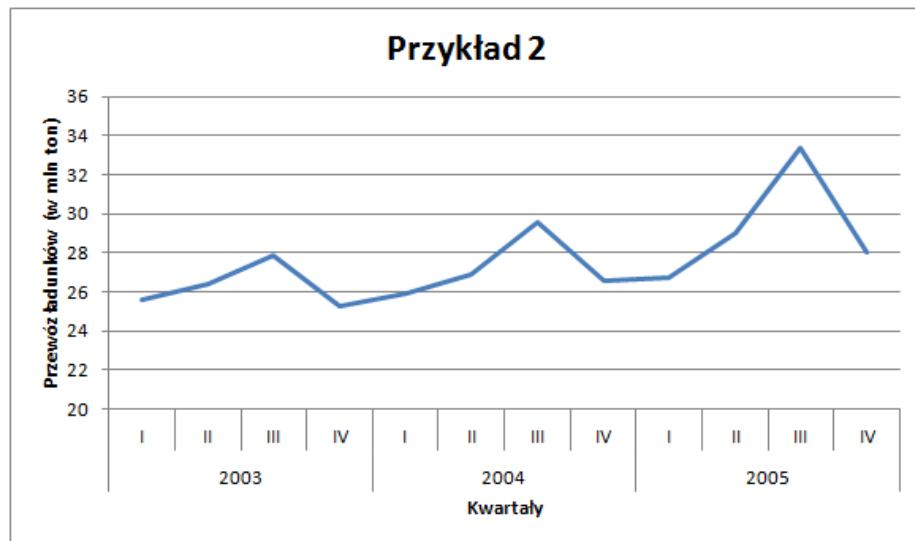


Figure 6.2: Przewóz ładunków - dane oryginalne

Dla przykładu pierwszego dotyczącego zużycia energii funkcja regresji przyjmuje następującą postać:

$$\hat{y}_t = 0,15 \cdot t + 2,99$$

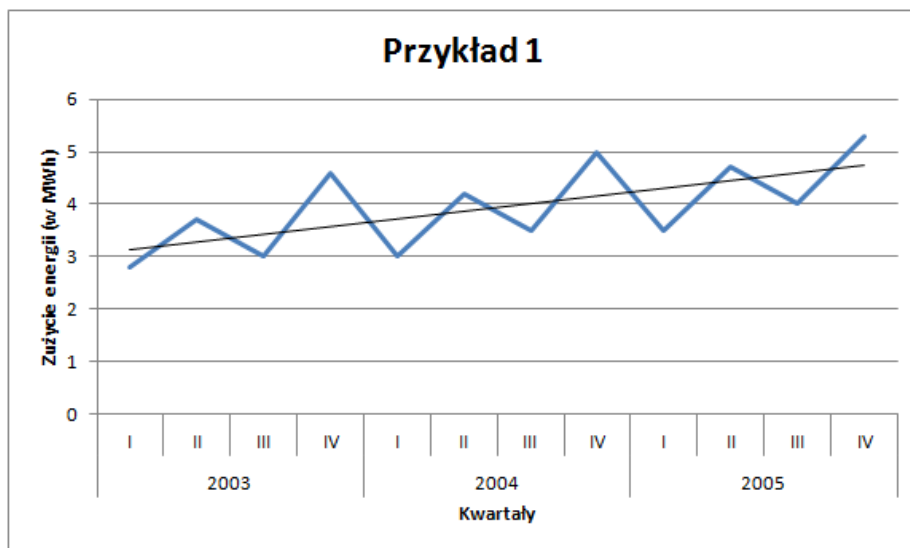
w której współczynnik kierunkowy informuje o tym, że z kwartału na kwartał zużycie energii rosło przeciętnie o 0,15 MWh. Z kolei wyraz wolny równy 2,99 oznacza, że w okresie $t = 0$ czyli w IV kwartale 2002 roku, teoretyczne zużycie energii wynosiło 2,99 MWh.

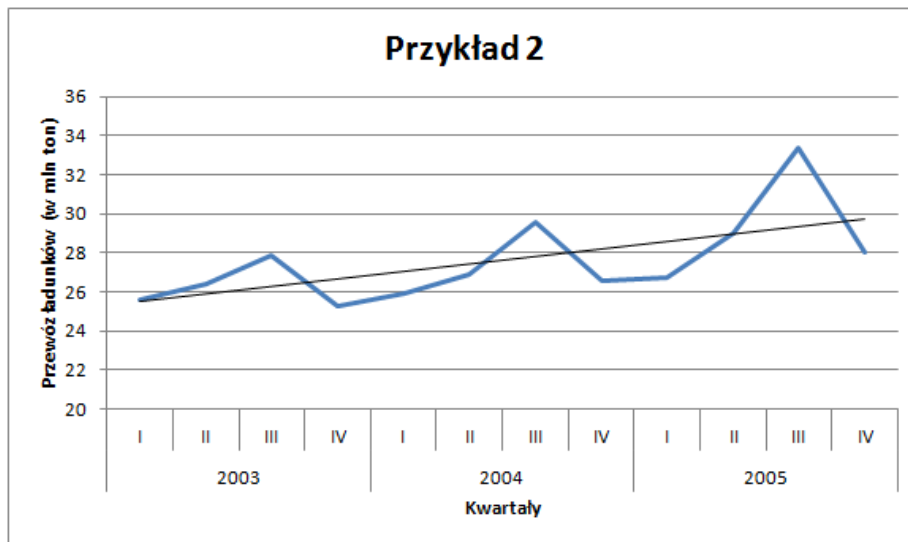
W drugim z analizowanych przykładów — przewozów ładunków — model wyglądał następująco:

$$\hat{y}_t = 0,38 \cdot t + 25,13$$

co oznacza, że z kwartału na kwartał przewóz ładunków wzrastał średnio o 0,38 mln ton, natomiast w IV kwartale 2002 roku modelowa wartość przewozów ładunków wynosiła 25,13 mln ton.

Na podstawie wyznaczonych funkcji regresji można obliczyć wartości teoretyczne (\hat{y}_t) zużycia energii oraz przewozów ładunków i pod postacią prostej przedstawić na wykresie.





Otrzymane wartości wynikające z funkcji trendu (\hat{y}_t) mają charakter liniowy i prawdę rzecząc słabo dopasowują się do danych empirycznych. Współczynnik R^2 w przykładzie pierwszym wynosi 41%, a w przykładzie drugim tylko 37%. Ponadto, jeśli chcielibyśmy prognozować na kolejne okresy to według funkcji trendu wartości zużycia energii dla kwartałów pierwszych byłyby przeszacowane, a dla kwartałów czwartych niedoszacowane. Stąd zachodzi potrzeba uwzględnienia w modelu występowania sezonowości, którą obserwujemy w danych.

Pierwszym krokiem jest identyfikacja rodzaju tej sezonowości. Może ona mieć charakter addytywny — wtedy wahania sezonowe są stałe w poszczególnych okresach (por. przykład 1) lub multiplikatywny, kiedy czynniki sezonowe są proporcjonalne do funkcji trendu (por. przykład 2). W zależności od zidentyfikowanego charakteru należy obliczyć wskaźniki sezonowości. W pierwszej kolejności rozważymy model addytywny.

6.2 Model addytywny

Analizę modelu addytywnego należy rozpocząć od wyznaczenia różnic pomiędzy wartościami empirycznymi (y) a modelowymi (\hat{y}) dla poszczególnych okresów zgodnie ze wzorem:

$$S_t^i = y_t - \hat{y}_t$$

Następnie dla każdego z analizowanych podokresów (półroczy, kwartałów, miesięcy) oblicza się surowe wskaźniki sezonowości uśredniając wyznaczone wcześniej różnice:

$$S_i = \frac{\sum_{t=1}^m S_t^i}{p}$$

gdzie:

- m — liczba podokresów (półroczy, kwartałów, miesięcy),
- p — liczba analizowanych lat.

W analizowanym przez nas przykładzie musimy wyznaczyć surowe wskaźniki sezonowości dla każdego kwartału. Ponadto jeśli spełniona będzie zależność $\sum_{i=1}^m S_i = 0$ to oznacza, że wskaźniki sezonowości są wolne od wahań przypadkowych. W praktyce jednak rzadko zdarza się taka sytuacja. W takim przypadku należy jeszcze wyznaczyć współczynnik korygujący zgodnie z wzorem:

$$k = \frac{\sum_{i=1}^m S_i}{m}$$

a następnie skorygować surowe wskaźniki sezonowości według formuły

$$So_i = S_i - k$$

otrzymując tzw. oczyszczone wskaźniki sezonowości, które informują o średnich odchyleniach od funkcji trendu w poszczególnych podokresach. Dla tych wskaźników zachodzi zależność: $\sum_{i=1}^m So_i = 0$. W przykładzie 1 oczyszczone wskaźniki sezonowości dla poszczególnych kwartałów są równe:

Wskaźnik	Wartość	Interpretacja
So_1	-0,62	w pierwszych kwartałach lat 2003–2005 zużycie energii było mniejsze średnio o 0,62 MWh niż wynika to z funkcji trendu
So_2	0,33	w drugich kwartałach lat 2003–2005 zużycie energii było większe średnio o 0,33 MWh niż wynika to z funkcji trendu
So_3	-0,51	w trzecich kwartałach lat 2003–2005 zużycie energii było mniejsze średnio o 0,51 MWh niż wynika to z funkcji trendu
So_4	0,81	w czwartych kwartałach lat 2003–2005 zużycie energii było większe średnio o 0,81 MWh niż wynika to z funkcji trendu
Suma	0,00	wskaźniki sezonowości są wolne od wahań przypadkowych

Kolejnym etapem analizy jest wyznaczenie zmodyfikowanych wartości teoretycznych uwzględniających sezonowość. Te wartości oznaczane jako \hat{y}^* uzyskujemy dodając do wartości teoretycznych (\hat{y}) odpowiednie dla poszczególnych

podokresów oczyszczone wskaźniki sezonowości So_i . Formalny zapis jest następujący:

$$\hat{y}^* = \hat{y} + So_i$$

Wartości \hat{y}^* przedstawione na wykresie już znacznie lepiej pasują do posiadanych danych empirycznych:

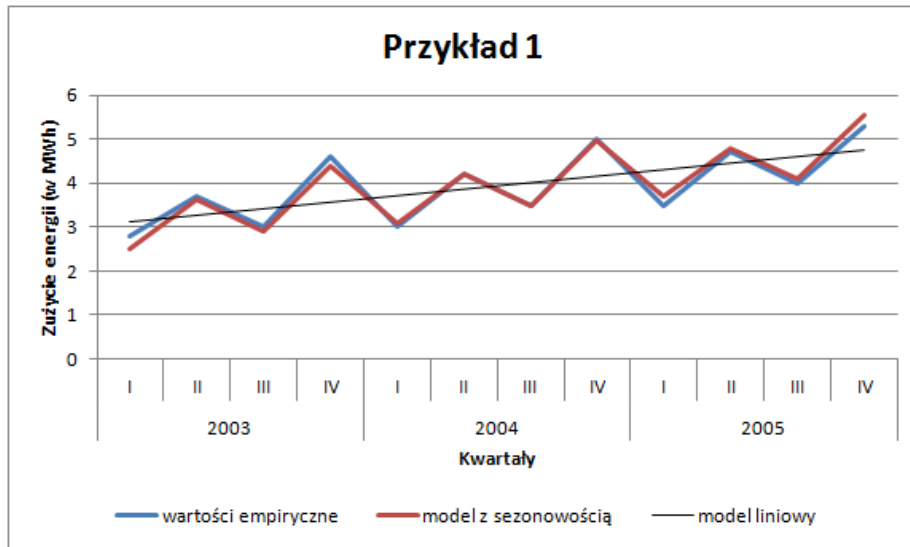


Figure 6.3: Zużycie energii - trend z sezonowością

Na podstawie tak zmodyfikowanego modelu można prognozować przyszłe wartości z dużo większą precyzją. Prognozowanie w modelu addytywnym polega na podstawieniu numeru okresu dla którego się prognozuje do funkcji trendu, a następnie dodanie odpowiedniego wskaźnika sezonowości:

$$\hat{y}_T^P = \hat{y} + So_i = a_1 \cdot T + a_0 + So_i$$

Interesuje nas prognozowane zużycie energii w IV kwartale 2008 roku. Ten okres przyjmuje wartość $t = 24$, natomiast wskaźnik sezonowości dla czwartego kwartału jest równy 0,81 MWh. Powyższe wartości podstawiamy do wzoru:

$$\hat{y}_{24}^P = 0,15 \cdot 24 + 2,99 + 0,81 = 7,4$$

co oznacza, że prognozowane zużycie energii w IV kwartale 2008 roku wyniesie 7,4 MWh.

6.3 Model multiplikatywny

W modelu multiplikatywnym zamiast różnic pomiędzy wartościami teoretycznymi a modelowymi oblicza się ich iloraz zgodnie ze wzorem:

$$S_t^i = \frac{y_t}{\hat{y}_t}$$

Następnie dla każdego z analizowanych podokresów (półroczy, kwartałów, miesięcy) oblicza się surowe wskaźniki sezonowości uśredniając wyznaczone wcześniej ilorazy:

$$S_i = \frac{\sum_{t=1}^m S_t^i}{p}$$

gdzie:

- m — liczba podokresów (półroczy, kwartałów, miesięcy),
- p — liczba analizowanych lat.

W analizowanym przez nas przykładzie musimy wyznaczyć surowe wskaźniki sezonowości dla każdego kwartału. W przypadku sezonowości multiplikatywnej zależność oznaczająca, że wskaźniki sezonowości są wolne od wahań przypadkowych jest wyrażona następująco: $\sum_{i=1}^m S_i = m$. W praktyce jednak rzadko zdarza się taka sytuacja. W takim przypadku należy jeszcze wyznaczyć współczynnik korygujący zgodnie z wzorem:

$$k = \frac{\sum_{i=1}^m S_i}{m}$$

a następnie skorygować surowe wskaźniki sezonowości według formuły

$$So_i = S_i/k$$

otrzymując tzw. oczyszczone wskaźniki sezonowości, które informują o średnich odchyleniach od funkcji trendu w poszczególnych podokresach. Dla tych wskaźników zachodzi zależność: $\sum_{i=1}^m So_i = m$. W przykładzie 2 oczyszczone wskaźniki sezonowości możemy zapisać w postaci procentowej i dla poszczególnych kwartałów są równe:

Wskaźnik	Wartość	Interpretacja
So_1	96,5%	w pierwszych kwartałach lat 2003–2005 rzeczywiste przewozy były średnio o 3,5% niższe niż wynika to z funkcji trendu

Wskaźnik	Wartość	Interpretacja
So_2	100,1%	w drugich kwartałach lat 2003–2005 rzeczywiste przewozy były średnio o 0,1% wyższe niż wynika to z funkcji trendu
So_3	108,9%	w trzecich kwartałach lat 2003–2005 rzeczywiste przewozy były średnio o 8,9% wyższe niż wynika to z funkcji trendu
So_4	94,5%	w czwartych kwartałach lat 2003–2005 rzeczywiste przewozy były średnio o 5,5% niższe niż wynika to z funkcji trendu
Suma	400,00%	wskaźniki sezonowości są wolne od wahań przypadkowych

Kolejnym etapem analizy jest wyznaczenie zmodyfikowanych wartości teoretycznych uwzględniających sezonowość. Te wartości oznaczane jako \hat{y}^* uzyskujemy mnożąc wartości teoretyczne (\hat{y}) odpowiednie dla poszczególnych podokresów przez oczyszczone wskaźniki sezonowości So_i . Formalny zapis jest następujący:

$$\hat{y}^* = \hat{y} \cdot So_i$$

Wartości \hat{y}^* przedstawione na wykresie już znacznie lepiej pasują do posiadanych danych empirycznych:

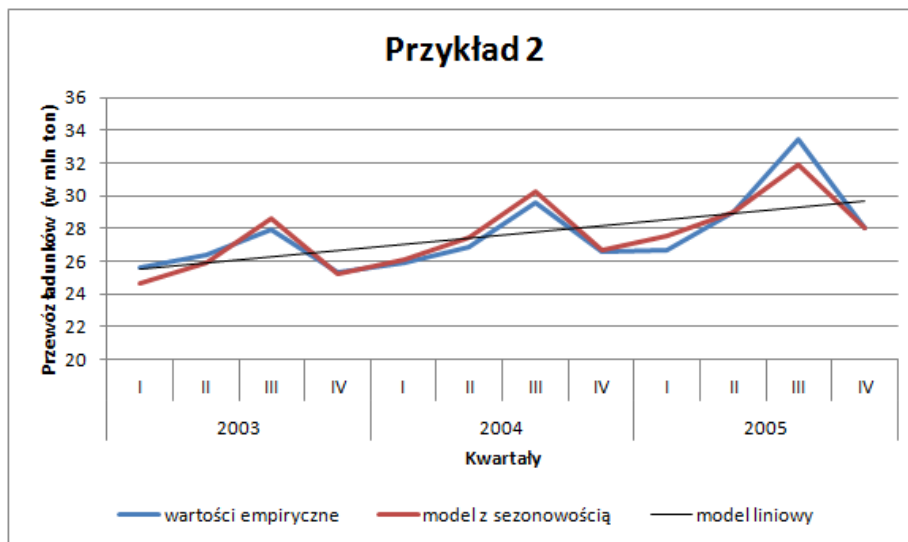


Figure 6.4: Przewóz ładunków - trend z sezonowością

Na podstawie tak zmodyfikowanego modelu można prognozować przyszłe wartości z dużo większą precyzją. Prognozowanie w modelu multiplikatywnym polega na podstawieniu numeru okresu dla którego się prognozuje do funkcji trendu, a następnie przemnożenie przez odpowiedni wskaźnik sezonowości:

$$\hat{y}_T^P = \hat{y} \cdot So_i = (a_1 \cdot T + a_0) \cdot So_i$$

Interesuje nas prognozowane zużycie energii w III kwartale 2006 roku. Ten okres przyjmuje wartość $t = 15$, natomiast wskaźnik sezonowości dla kwartału trzeciego jest równy 108,9%. Powyższe wartości podstawiamy do wzoru:

$$\hat{y}_{15}^P = (0,38 \cdot 15 + 25,13) \cdot 108,9\% = 33,6$$

co oznacza, że prognozowane przewozy ładunków w III kwartale 2006 roku wyniosą 33,6 mln ton.

6.4 Ocena jakości

Ostatnim elementem analizy sezonowości jest ocena jakości otrzymanego modelu. W takim przypadku nie wyznaczamy współczynnika R^2 ponieważ z definicji dotyczy on wyłącznie zależności liniowej. Główną miarą jakości będzie odchylenie standardowe składnika resztowego z uwzględnieniem sezonowości:

$$S_u^* = \sqrt{\frac{\sum_{t=1}^n (y_t - \hat{y}_t^*)^2}{n - 2}}$$

Licznik odchylenia standardowego zawiera sumę kwadratów odchyleń wartości empirycznych (y_t) od wartości modelowych z sezonowością (\hat{y}_t^*). Nie ma już tutaj znaczenia czy model był addytywny czy multiplikatywny.

W przykładzie pierwszym S_u^* wynosiło 0,16 MWh, co oznacza, że rzeczywiste zużycie energii różni się od zużycia teoretycznego wyznaczonego na podstawie szeregu czasowego średnio o +/- 0,16 MWh. Z kolei w przykładzie drugim S_u^* wynosiło 0,74 mln ton, a co za tym idzie rzeczywiste przewozy różnią się od przewozów teoretycznych uzyskanych w oparciu o model szeregu czasowego średnio o +/- 0,74 mln ton.

6.5 Błąd prognozy

Wyliczona wartość S_u^* niezbędna jest przy wyznaczaniu błędu prognozy zgodnie ze wzorem:

$$D(y_T^P) = S_u^* \sqrt{1 + \frac{1}{n} + \frac{(T - \bar{t})^2}{\sum_{t=1}^n (t - \bar{t})^2}}$$

w którym uwzględniamy możliwość wzrostu tego błędu wraz z oddalaniem się od zakresu danych, które posiadamy.

Dla analizowanych przykładów otrzymano następujące błędy prognozy:

- przykład 1 — zużycie energii

Przy prognozie dla IV kwartału 2008 roku

$$D(y_{24}^P) = 0,29$$

co oznacza, że prognozowane zużycie energii w IV kwartale 2008 roku wyniesie 7,4 +/- 0,29 MWh. - przykład 2 — przewóz ładunków

Przy prognozie dla III kwartału 2006 roku

$$D(y_{24}^P) = 0,93$$

co oznacza, że prognozowane przewozy w III kwartale 2006 roku wyniosą 33,6 +/- 0,93 mln ton.

Na podstawie otrzymanych prognoz oraz ich błędów można wyznaczyć przedziały, w których spodziewamy się wartości rzeczywistej.

Chapter 7

Analiza szeregu dynamicznego

Badanie szeregu dynamicznego umożliwia dokładne określenie zmian jakie zachodzą w kolejnych okresach. Za przykład posłuży sprzedaż piwa w województwie wielkopolskim w kolejnych okresach.

miesiac	liczba_sztuk
1	2113
2	1991
3	2084
4	2775
5	2594
6	2654
7	2771
8	2983

7.1 Przyrosty absolutne

Odejmowanie dwóch wielkości liczbowych daje w wyniku dodatni lub ujemny **przyrost absolutny (bezwzględny)**. Przyrosty mogą być obliczane w stosunku do jednego okresu (momentu) lub też okresu (momentu) stale zmieniającego się.

Jeśli poszczególne wyrazy szeregu dynamicznego oznaczymy przez:

$$y_1, y_2, y_3, \dots, y_n$$

to ciąg przyrostów absolutnych o podstawie stałej y_1 (**jednopodstawowych**) przedstawia się następująco:

$$y_2 - y_1, y_3 - y_1, \dots, y_{n-1} - y_1, y_n - y_1$$

Natomiast ciąg przyrostów absolutnych o podstawie zmiennej (**łańcuchowych**) ma postać:

$$y_2 - y_1, y_3 - y_2, \dots, y_{n-1} - y_{n-2}, y_n - y_{n-1}$$

Przyrosty absolutne informują o tym, o ile jednostek wzrósł (znak plus) lub zmalał (znak minus) poziom badanego zjawiska w okresie (momencie) badanym w porównaniu z okresem (momentem) przyjętym za podstawę. Przyrosty absolutne są wyrażone w tych samych jednostkach miary co badane zjawisko.

Przyrosty absolutne o podstawie w maju oraz przyrosty absolutne łańcuchowe przedstawione są w tabeli.

miesiac	liczba_sztuk	Paj	Pal
1	2113	-481	NA
2	1991	-603	-122
3	2084	-510	93
4	2775	181	691
5	2594	0	-181
6	2654	60	60
7	2771	177	117
8	2983	389	212

Na podstawie przyrostów absolutnych jednopodstawowych stwierdzamy, że sprzedaż piwa w lutym była o 603 sztuki mniejsza w porównaniu do maja. Z kolei przyrosty absolutne łańcuchowe informują o tym, że lipcu liczba sprzedanych piw była większa o 117 sztuk w porównaniu do czerwca.

7.2 Przyrosty względne

Iloraz przyrostów absolutnych zjawiska i jego poziomu w okresie (momencie) przyjętym za podstawę porównań nazywamy **przyrostem względnym**.

Ciąg wartości przyrostów względnych o stałej podstawie y_1 (**jednopodstawowych**) jest następujący:

$$\frac{y_2 - y_1}{y_1}, \frac{y_3 - y_1}{y_1}, \dots, \frac{y_{n-1} - y_1}{y_1}, \frac{y_n - y_1}{y_1}$$

Ciąg wartości przyrostów względnych o podstawie zmiennej (**łańcuchowych**) ma postać:

$$\frac{y_2 - y_1}{y_1}, \frac{y_3 - y_2}{y_2}, \dots, \frac{y_{n-1} - y_{n-2}}{y_{n-2}}, \frac{y_n - y_{n-1}}{y_{n-1}}$$

Przyrosty względne wyrażane są w procentach. Informują o ile wyższy lub niższy jest poziom badanego zjawiska w danym okresie w stosunku do okresu przyjętego za podstawę (przyrosty względne jednopodstawowe) lub w stosunku do okresu bezpośrednio poprzedzającego (przyrosty względne łańcuchowe). Przyrosty względne mogą być wartościami dodatnimi, ujemnymi lub równymi zero. Określane są niekiedy mianem wskaźników tempa przyrostu.

miesiac	liczba_sztuk	Pwj	Pwl
1	2113	-18.54	NA
2	1991	-23.25	-5.77
3	2084	-19.66	4.67
4	2775	6.98	33.16
5	2594	0.00	-6.52
6	2654	2.31	2.31
7	2771	6.82	4.41
8	2983	15.00	7.65

Na podstawie przyrostów względnych jednopodstawowych stwierdzamy, że sprzedaż piwa w lutym była o 22,25% mniejsza w porównaniu do maja. Z kolei przyrosty względne łańcuchowe informują o tym, że lipcu liczba sprzedanych piw była większa o 4,41% w porównaniu do czerwca.

7.3 Indeksy

Indeksem (wskaźnikiem dynamiki) nazywamy każdą liczbę względną powstałą przez podzielenie wielkości danego zjawiska w okresie badanym (sprawozdawczym) przez wielkość tego zjawiska w okresie podstawowym (bazowym). Jeżeli poziom zjawiska w okresie (momencie) badanym oznaczmy symbolem y_1 , a w okresie podstawowym symbolem y_0 , to ogólny wzór na indeks przyjmie postać: $i = \frac{y_1}{y_0}$.

Indeks jest wielkością niemianowaną i może być wyrażony w ułamku lub w procentach. Jeżeli indeks przyjmie wartość z przedziału $0 \leq i < 1$, świadczy to o spadku poziomu zjawiska w okresie badanym w stosunku do okresu podstawowego. Większa od 1 (lub od 100%) wartość indeksu informuje o wzroście poziomu zjawiska w okresie badanym w porównaniu z okresem podstawowym. Wreszcie indeks równy 1 oznacza, że poziomy zjawiska w okresach badanym i podstawowym są takie same.

Ciąg indeksów o stałej podstawie y_1 (**jednopodstawowych**) można zapisać:

$$\frac{y_1}{y_1}, \frac{y_2}{y_1}, \dots, \frac{y_{n-1}}{y_1}, \frac{y_n}{y_1}$$

Ciąg indeksów o podstawie zmiennej (**łańcuchowych**) ma postać:

$$\frac{y_2}{y_1}, \frac{y_3}{y_2}, \dots, \frac{y_{n-1}}{y_{n-2}}, \frac{y_n}{y_{n-1}}$$

Między przyrostami względnymi a indeksami istnieje ścisły związek. Indeksy jednopodstawowe można otrzymać z przyrostów względnych o podstawie stałej poprzez dodanie 100 (lub jedności, jeśli posługujemy się ułamkami, a nie wielkościami procentowymi). W analogiczny sposób można dokonać przejścia z indeksów łańcuchowych na przyrosty względne łańcuchowe. Oczywiście można dokonać również operacji odwrotnej, tzn. zamienić indeksy na przyrosty względne.

miesiac	liczba_sztuk	Ij	II
1	2113	81.46	NA
2	1991	76.75	94.23
3	2084	80.34	104.67
4	2775	106.98	133.16
5	2594	100.00	93.48
6	2654	102.31	102.31
7	2771	106.82	104.41
8	2983	115.00	107.65

W praktyce badań statystycznych częściej wykorzystuje się indeksy niż przyrosty względne. Na indeksach wygodniej jest bowiem dokonywać określonych przekształceń i działań algebraicznych. Działania, jakie mogą być dokonywane na indeksach, sprowadzają się w zasadzie do zamiany indeksów jednopodstawowych na łańcuchowe i odwrotnie oraz do zmiany podstawy w szeregu indeksów o podstawie stałej.

Zamiany indeksów jednopodstawowych na łańcuchowe dokonuje się poprzez dzielenie indeksów jednopodstawowych przez siebie:

$$\frac{y_i}{y_1} / \frac{y_{i-1}}{y_1} = \frac{y_i}{y_{i-1}}$$

W praktyce indeksy łańcuchowe na podstawie indeksów jednopodstawowych liczymy tak samo jak zwykle indeksy łańcuchowe, tylko zamiast wartości bezwzględnych podstawiamy wartości indeksów jednopodstawowych.

Zamiany indeksów łańcuchowych na jednopodstawowe dokonuje się według następujących zasad:

1. indeks jednopodstawowy w okresie przyjętym za podstawę wynosi 100%;
2. indeks jednopodstawowy w okresie następującym bezpośrednio po okresie przyjętym za podstawę jest taki sam jak indeks łańcuchowy;
3. dalsze indeksy jednopodstawowe po okresie przyjętym za podstawę otrzymujemy, mnożąc wyznaczony indeks jednopodstawowy z okresu wcześniejszego przez indeks łańcuchowy dla analizowanego okresu;

4. indeksy jednopodstawowe przed okresem podstawowym są ilorazem indeksu jednopodstawowego z okresu następnego oraz indeksu łańcuchowego z okresu następnego.

Zmiany podstawy w indeksach jednopodstawowych dokonuje się poprzez dzielenie poszczególnych indeksów przy danej podstawie przez indeks jednopodstawowy tego okresu, który przyjmuje się za nową podstawę. Innymi słowy, indeks jednopodstawowy o nowej podstawie liczymy w taki sam sposób jak byśmy liczyli indeks jednopodstawowy na podstawie wartości bezwzględnych, ale zamiast nich podstawiamy wartości indeksów jednopodstawowych.

7.4 Średnie tempo zmian

Indeksy jednopodstawowe i łańcuchowe pozwalają na ocenę zmian badanego zjawiska między dwoma wyróżnionymi okresami (momentami). Często jednak zachodzi konieczność oceny zmian danego zjawiska w całym okresie objętym obserwacją. Do tego celu wykorzystuje się średnią geometryczną.

Średnia geometryczna jest pierwiastkiem n -tego stopnia z iloczynu n zmiennych. Średnie tempo zmian zjawisk ujętych w formie szeregów czasowych oblicza się najczęściej na podstawie indeksów łańcuchowych. Ponieważ z n wielkości absolutnych można utworzyć $n - 1$ indeksów łańcuchowych, przez to wzór na średnią geometryczną indeksów łańcuchowych przyjmuje postać:

$$\bar{y}_g = \sqrt[n-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \frac{y_4}{y_3} \cdot \dots \cdot \frac{y_{n-1}}{y_{n-2}} \cdot \frac{y_n}{y_{n-1}}} = \sqrt[n-1]{\prod_{i=2}^n \frac{y_i}{y_{i-1}}} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Średnie tempo zmian wyrażone w procentach określa, jaki jest przeciętny okresowy przyrost procentowy analizowanego zjawiska w badanym przedziale czasowym.

Średnia geometryczna indeksów łańcuchowych sprzedanych piw wynosi 105.05%, co oznacza, że liczba sprzedanych piw w województwie wielkopolskim rosła z miesiąca na miesiąc przeciętnie o 5.05%.

Obliczone średnie tempo zmian można wykorzystać w celach prognostycznych. Zakładając, że w kolejnych okresach badane zjawisko będzie rozwijać się w dotychczasowym tempie:

$$y_T = y_n \cdot \bar{y}_g^{(T-n)}$$

gdzie:

- T — numer okresu prognozowanego,
- n — numer ostatniego okresu.

Przeprowadzenie szacunku z wykorzystaniem powyższej metody wymaga jednak, by dotychczasowy rozwój tej zmiennej był jednokierunkowy i nie podlegał zbyt dużej zmienności. Stosując tę regułę nie można ustalić błędu prognozy.

Zakładając, że we wrześniu popyt na piwo nie spadnie, prognozowana sprzedaż wyniesie 3134 sztuk.

Znając wartość cechy w ostatnim okresie możemy, z wykorzystaniem indeksów łańcuchowych, odtworzyć wartości cechy dla okresów wcześniejszych dzieląc wartość cechy przez indeks łańcuchowy.

Zadania

1. Na podstawie danych dotyczących liczby sprzedanych sztuk piwa w województwie mazowieckim:
 - oblicz średnie tempo zmian i je zinterpretuj
 - o ile procent zmieniła się sprzedaż piwa w marcu w porównaniu do czerwca?
 - jaka jest prognozowana sprzedaż piwa w październiku?
2. Wyznacz indeksy łańcuchowe dla liczby sprzedanych sztuk piwa w sklepach sieci detalicznej. Wyłącznie na ich podstawie określ:
 - o ile procent zmieniła się sprzedaż piwa w sierpniu w porównaniu do czerwca?
 - sprzedaż piwa w listopadzie wiedząc, że w sierpniu sprzedano 4146 sztuk
 - ile wynosiła sprzedaż piwa w lutym?
 - w których miesiącach sprzedaż piwa była mniejsza niż w miesiącu poprzedzającym?

Chapter 8

Indeksy indywidualne i zespołowe

Indeksy indywidualne i zespołowe umożliwiają ocenę zmian w dwóch różnych okresach uwzględniając stałą ilość bądź wartość dla danego koszyka produktów. Najbardziej znanym przykładem indeksu zespołowego jest wskaźnik cen towarów i usług konsumpcyjnych czyli miara inflacji/deflacji.

8.1 Indeksy indywidualne

Indeksem indywidualnym nazywamy stosunek poziomów tego samego pojedynczego zjawiska z dwóch różnych okresów (momentów). W statystyce społeczno-ekonomicznej rozpatruje się zwykle trzy rodzaje indywidualnych wskaźników dynamiki, a mianowicie: indeksy cen, ilości i wartości.

Indeks indywidualny cen wyraża relację poziomu cen określonego dobra w okresie badanym i w okresie podstawowym, co można zapisać następująco:

$$i_p = \frac{p_1}{p_0}$$

gdzie:

- i_p — indywidualny indeks cen,
- p_1 — cena jednostki wyrobu w okresie badanym,
- p_0 — cena jednostki wyrobu w okresie podstawowym.

Indeks indywidualny ilości oblicza się jako stosunek ilości określonego wyrobu wytworzonego w okresie badanym i w okresie podstawowym:

$$i_q = \frac{q_1}{q_0}$$

gdzie:

- i_q — indywidualny indeks ilości,
- q_1 — ilość wyrobu wyprodukowanego w okresie badanym,
- q_0 — ilość wyrobu wyprodukowanego w okresie podstawowym.

Iloczyn ilości wyrobu wytworzonego w okresie badanym i ceny tego wyrobu z okresu badanego daje w wyniku wartość wyrobu w okresie badanym. Podobnie wylicza się wartość wyrobu w okresie podstawowym. W związku z tym **indywidualny indeks wartości** to iloraz wartości wyrobu wytworzonego w okresie badanym i w okresie podstawowym:

$$i_w = \frac{q_1 p_1}{q_0 p_0} = \frac{w_1}{w_0}$$

gdzie:

- i_w — indywidualny indeks wartości,
- w_1 — wartość wyrobu w okresie badanym,
- w_0 — wartość wyrobu w okresie podstawowym.

Indywidualne indeksy cen, ilości i wartości informują o zmianie (wzroście lub spadku) tych wielkości w okresie badanym w porównaniu z okresem przyjętym za podstawę porównań. Między indywidualnymi indeksami cen, ilości i wartości obliczonymi dla tego samego produktu zachodzi następujący związek:

$$i_w = i_p i_q$$

Relacja określona powyższym wzorem nosi nazwę **równości indeksowej dla indeksów indywidualnych**. W przypadku gdy nie dysponujemy informacjami wyjściowymi, równość indeksowa umożliwia — przy znajomości dowolnej pary spośród trzech indeksów — obliczenie trzeciego indeksu.

8.2 Indeksy zespołowe (agregatowe)

W praktyce badań statystycznych niejednokrotnie zachodzi potrzeba obliczenia indeksów dotyczących nie indywidualnych jednostek, ale całego zespołu (agregatu, zbioru) jednostek. Do badania dynamiki całego zespołu zjawisk — zwykle niejednorodnych i bezpośrednio niesumowalnych — stosowane są indeksy zespołowe (agregatowe). Konstrukcja indeksów zespołowych opiera się na wykorzystaniu określonych współczynników przeliczeniowych, odgrywających rolę wag. Rolę wag najczęściej spełniają ceny lub ilości.

Agregatowy indeks wartości określonego zespołu artykułów jest ilorazem sum wartości badanych dóbr w okresie badanym i w okresie podstawowym:

$$I_w = \frac{\sum q_1 p_1}{\sum q_0 p_0}$$

gdzie:

- I_w — agregatowy indeks wartości badanego zespołu artykułów,
- $\sum q_1 p_1$ — suma wartości badanego zespołu w okresie badanym,
- $\sum q_0 p_0$ — suma wartości badanego zespołu w okresie podstawowym.

Agregatowy indeks wartości wyraża zmiany, jakie nastąpiły w okresie badanym w porównaniu z okresem podstawowym zarówno w ilościach określonego zespołu artykułów, jak i w ich cenach. W celu obliczenia siły i kierunku zmian wyłącznie ilości lub wyłącznie cen wyrobów wschodzących w skład agregatu buduje się agregatowe indeksy ilości i agregatowe indeksy cen. Do uzyskania agregatowego indeksu ilości unieruchamiane (ustalane na stałym poziomie) są ceny, natomiast do uzyskania agregatowego indeksu cen unieruchamiane są ilości. Każdy z tych indeksów może być wyliczony w oparciu o dwie formuły — Laspeyresa i Paaschego. W indeksie Laspeyresa unieruchamia się ilość lub cenę na poziomie okresu podstawowego, natomiast we wzorze Paaschego stała jest ilość bądź cena na poziomie okresu badanego.

Agregatowy indeks ilości Laspeyresa ma postać:

$$I_q^L = \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

Agregatowy indeks ilości Paaschego oblicza się następująco:

$$I_q^P = \frac{\sum q_1 p_1}{\sum q_0 p_1}$$

Agregatowe indeksy ilości informują o tym, o ile — przeciętnie biorąc — wzrosła lub zmalała ilość określonego zbioru artykułów w okresie badanym w porównaniu z odpowiednim okresem podstawowym.

Agregatowy indeks cen Laspeyresa ma postać:

$$I_p^L = \frac{\sum q_0 p_1}{\sum q_0 p_0}$$

Agregatowy indeks cen Paaschego oblicza się następująco:

$$I_p^P = \frac{\sum q_1 p_1}{\sum q_1 p_0}$$

Agregatowe indeksy cen odpowiadają na pytanie, jak zmieniły się — przeciętnie biorąc — ceny danego zbioru artykułów w okresie badanym w porównaniu z okresem podstawowym, przy unieruchomieniu ilości w obu okresach, zgodnie z przyjętą formułą.

W przypadku niezbyt odległych okresów porównawczych (tzn. okresu podstawowego i badanego) obliczane są też agregatowe indeksy cen i ilości według formuły Fishera.

Agregatowy indeks ilości Fishera wyrażony jest formułą:

$$I_q^F = \sqrt{I_q^L \cdot I_q^P}$$

Agregatowy indeks cen Fishera oblicza się następująco:

$$I_p^F = \sqrt{I_p^L \cdot I_p^P}$$

Agregatowe indeksy cen i ilości Fishera informują o tym, o ile — przeciętnie biorąc — wzrosła lub zmalała ilość lub cena określonego zbioru artykułów w badanym okresie.

Pomiędzy agregatowymi indeksami wartości, cen i ilości zachodzą następujące związki:

$$I_w = I_p^L \cdot I_q^P = I_p^P \cdot I_q^L = I_p^F \cdot I_q^F$$

Związek określony powyższą relacją nosi nazwę **równości indeksowej dla indeksów agregatowych (zespołowych)**. Jeśli dysponujemy informacjami o poziomach dwóch spośród trzech omawianych indeksów agregatowych, to możemy obliczyć wielkość trzeciego indeksu.

8.3 Przykład

W tabeli zebrano informacje na temat sprzedaży piwa (w sztukach) oraz średniej ceny (w zł) z trzech browarów w dwóch kolejnych miesiącach.

	styczeń		luty	
	p0	q0	p1	q1
Browar				
Carlsberg	2,30	5009	2,28	4437
Kompania Piwowarska	2,69	5806	2,63	5882
Zywiec	2,51	7934	2,45	7613

- Jak zmieniła się wartość sprzedanych piw w porównywanych okresach?
- Jaki wpływ na zmianę wartości miała dynamika cen, a jaka dynamika ilości

sprzedawanych piw?

W kolejnej tablicy przedstawione są wyniki obliczeń pomocniczych:

Piwo	q0p0	q1p1	q0p1	q1p0
Carlsberg	11526,95	10133,76	11440,17	10210,64
Kompania Piwowarska	15590,44	15456,79	15257,08	15794,52
Zywiec	19909,22	18671,07	19458,33	19103,72
Razem	47026,62	44261,62	46155,57	45108,88

Podstawiając odpowiednie wartości do wzoru na agregatowy indeks wartości, otrzymujemy:

$$I_w = \frac{44261,62}{47026,62} \cdot 100\% = 94,12\%$$

Otrzymany wynik oznacza, że łączna wartość piw ze wszystkich trzech browarów w lutym jest o 5,88% niższa od wartości ze stycznia. Spadek wartości o 5,88% spowodowany jest zmianami cen i ilości produkowanych wyrobów.

Agregatowe indeksy ilości Laspeyresa i Paaschego są równe:

$$I_q^L = \frac{45108,88}{47026,62} \cdot 100\% = 95,92\%$$

$$I_q^P = \frac{44261,62}{46155,57} \cdot 100\% = 95,90\%$$

Agregatowy indeks ilości Laspeyresa informuje o tym, że sprzedaż piw z trzech browarów łącznie w lutym w porównaniu ze styczniem spadła o 4,08%, przy założeniu, że ceny w lutym były takie same jak w styczniu.

Agregatowy indeks ilości obliczony według formuły Paaschego wskazuje, że sprzedaż piw z trzech browarów łącznie w lutym spadła — w porównaniu ze styczniem — o 4,10% przy stałych cenach z lutego.

Agregatowe indeksy cen Laspeyresa i Paaschego wynoszą:

$$I_p^L = \frac{46155,57}{47026,62} \cdot 100\% = 98,15\%$$

$$I_p^P = \frac{44261,62}{45108,88} \cdot 100\% = 98,12\%$$

Agregatowy indeks cen Laspeyresa oznacza, że — przeciętnie biorąc — cena piwa składających się na badany agregat spadła w lutym w porównaniu do stycznia

o 1,85%, przy zachowaniu umownego założenia, że w lutym sprzedano te same ilości każdego piwa co w styczniu.

Agregatowy indeks cen Paaschego informuje o tym, że ceny badanych piw w lutym spadły — średnio biorąc — w porównaniu ze styczniem o 1,88%, przy założeniu, że w styczniu roku sprzedano te same ilości piwa co w lutym.

Agregatowe indeksy cen i ilości przy zastosowaniu formuły Fishera są równe:

$$I_q^F = \sqrt{0,9592 \cdot 0,9590} = 0,9591 \cdot 100\% = 95,91\%$$

$$I_p^F = \sqrt{0,9815 \cdot 0,9812} = 0,9813 \cdot 100\% = 98,13\%$$

Agregatowy indeks ilości Fishera oznacza, że ilość sprzedanych piw spadła w lutym w porównaniu do stycznia średnio o 4,09%.

Agregatowy indeks cen Fishera informuje o tym, że ceny sprzedanych piw spadły w lutym w porównaniu do stycznia średnio o 1,87%.

Zadania

1. Przeprowadź wszechstronną analizę dynamiki cen i ilości w sklepach typu dyskont, hypermarket, sieci detalicznej, spożywczych oraz supermarketach w miesiącach maj i lipiec.
2. Wartość produkcji dwóch wyrobów trwałego użytku w Polsce w latach 2004-2005 przedstawia następująca tabela. Wiedząc, że łączna wartość sprzedaży tych dóbr w 2004 wynosiła 262479 mln zł, scharakteryzuj jej dynamikę, uwzględniając zmiany cen oraz ilości.

Nazwa wyrobu	Wartość sprzedaży w 2005 (mln zł.)	Zmiany ilości
Radio	38410	Spadek o 35%
Telewizor	423800	Wzrost o 49%
Suma	462210	x

Opracowano na podstawie: Sobczyk Mieczysław, 2002, *Statystyka*, Wydawnictwo Naukowe PWN.