RLAB_1 Statystyka opisowa - tabelaryczne i graficzne metody, metody numeryczne

Celem ćwiczenia jest zapoznanie się z podstawowymi funkcjami w środowisku R, oraz tabelarycznymi i graficznym oraz numerycznymi metodami statystyki opisowej

Funkcje: c(), sort(), median(), min(), max(), sort(), sum(), length(), head(), tail(), hist(), morm(), option(), data.frame(), summary(), table(), barplot(), dotchar(), cut(), read.csv(), names(), plot(), library(), rowsums(), cbind(), colsums, rbind() , rownames(), colnames(),Quantile(), var(), sd(), seq(), IQR(), cov(), which(), runif(),

Sprawozdanie nie jest wymagane.

1. Za pomocą R odpowiedz na następujące pytania.

a) suma 137 i 242.

b) Różnica 1206 - 373

c) Iloczyn 547 i 23.

(d) Podziel 8 840 przez 17.

(e) Podnieś 11 do potęgi 3. 11 ^ 3

(f) Znajdź pierwiastek kwadratowy z liczby 64. sqrt(64)

(g) Znajdź pierwiastek sześcienny z 8,000. 8000 ^ (1/3)

2. Wprowadź następujący zestaw danych bezpośrednio do obszaru roboczego języka R i nadaj mu nazwę E1_1:

81, 17, 7, 55, 2, 98, 71, 47, 19, 8, 3, 10, 28, 65, 80.

Sprawdź, czy E1_1 zawiera te elementy i odpowiedz na następujące pytania.

Użyj funkcji c(), aby utworzyć E1_1 obiektu.

>E1_1 <- c(81, 17, 7, 55, 2, 98, 71, 47, 19, 8, 3, 10, 28, 65, 80)

Zbadaj zawartość E1_1.

>E1_1

a) Medianę (posortowanego) zbioru danych wyznacza się jako wartość, która dzieli zbiór danych dokładnie w połowie, pozostawiając taką samą liczbę pozycji danych poniżej, jak powyżej wartości mediany.

Jaka jest mediana E1_1?

Wskazówka: użyj funkcji sort(), aby uszeregować wszystkie wartości danych w E1_1, od najniższej do najwyższej.

```
# (1) Utwórz obiekt E1_1.
>E1_1 <- c(81, 17, 7, 55, 2, 98, 71, 47, 19, 8, 3, 10, 28, 65, 80)

# (2) Użyj funkcji sort(), aby uszeregować dane dotyczące kolejności.
> E1_1 <- sort(E1_1)

# (3) Zbadaj zawartość E1_1. Uwaga: wartość środkowa  (lub mediana) wynosi 28.
>E1_1

# (4) Użyj funkcji median(), aby znaleźć medianę E1_1.
median(E1_1)
```

(b) Korzystając z funkcji max() i min(), znajdź wartości maksymalne i minimalne z E1_1. Ponadto, korzystając z funkcji sum() i mean(), znajdź sumę wszystkich wartości danych, jak również średnią z E1_1.

```
# (1) Użyj funkcji min(), aby znaleźć minimalną wartość w E1_1.
>min(E1_1)

# (2) Użyj funkcji max(), aby znaleźć maksymalną wartość w E1_1.
>max(E1_1)

# (3) Użyj funkcji sum(), aby znaleźć sumę wartości w E1_1.
>sum(E1_1)

# (4) Użyj funkcji mean(), aby znaleźć średnią z E1_1.
>mean(E1_1)
```

c) Policz liczbę wartości danych w E1_1. Chociaż oczywiste jest, że istnieje 15 elementów, funkcja length() może być użyta, gdy chcemy poznać liczbę elementów zawartych w wektorze o nieznanym rozmiarze.

Użyj funkcji length(), aby znaleźć liczbę elementów danych w E1_1.

>length(E1_1)

3. Użyj funkcji sum() i length(), aby obliczyć średnią z E1_1.

(1) Użyj stosunku sum() i length();

>mean <- Sum(E1_1) / length(E1_1)

(2) Zbadaj zawartość średniej.

>mean

Wartość średniej jest taka sama niezależnie od tego, czy wyprowadzimy ją w ten sposób, czy też użyjemy funkcji mean(), aby znaleźć odpowiedź w bardziej bezpośredni sposób.

4. Podstawowy system R obejmuje szereg wbudowanych zestawów danych, które możemy wykorzystać. Aby wyświetlić listę tych bezpłatnych zestawów danych, po prostu wprowadź data() w wierszu polecenia. Na przykład jeden z zestawów danych nosi nazwę LakeHuron.

Aby dowiedzieć się trochę na temat tego zbioru danych, wprowadź

> ? LakeHuron

w wierszu polecenia języka R w konsoli. Kiedy to robimy, otwiera się strona opisująca dane, informująca nas, że zestaw danych LakeHuron składa się z corocznych pomiarów poziomu jeziora Huron, w stopach, 1875 - 1972.

Poniższe pytania dotyczą zbioru danych LakeHuron.

(a) Użyj funkcji head(), aby wyświetlić pierwsze trzy obserwacje jeziora Huron.

>head(LakeHuron, n)

, aby wyświetlić pierwsze n elementów danych. Na przykład:

>head(LakeHuron, 3)

(b) Czy brakuje jakichś danych? Użyj funkcji length(), aby potwierdzić, że istnieje 98 obserwacji (liczba lat od 1875 do 1972).

>leanght(LakeHuron)

(c) Jaki był najniższy poziom (w stopach) jeziora Huron w latach 1875-1972?

>min(LakeHuron)

(d) Jaki był najwyższy poziom wody w jeziorze Huron w tym samym okresie?

>max(LakeHuron)

(e) Jaki jest średni poziom wody w jeziorze Huron w tym okresie?

>mean(LakeHuron)

(f) Co to jest mediana poziomu?

>median(LakeHuron)

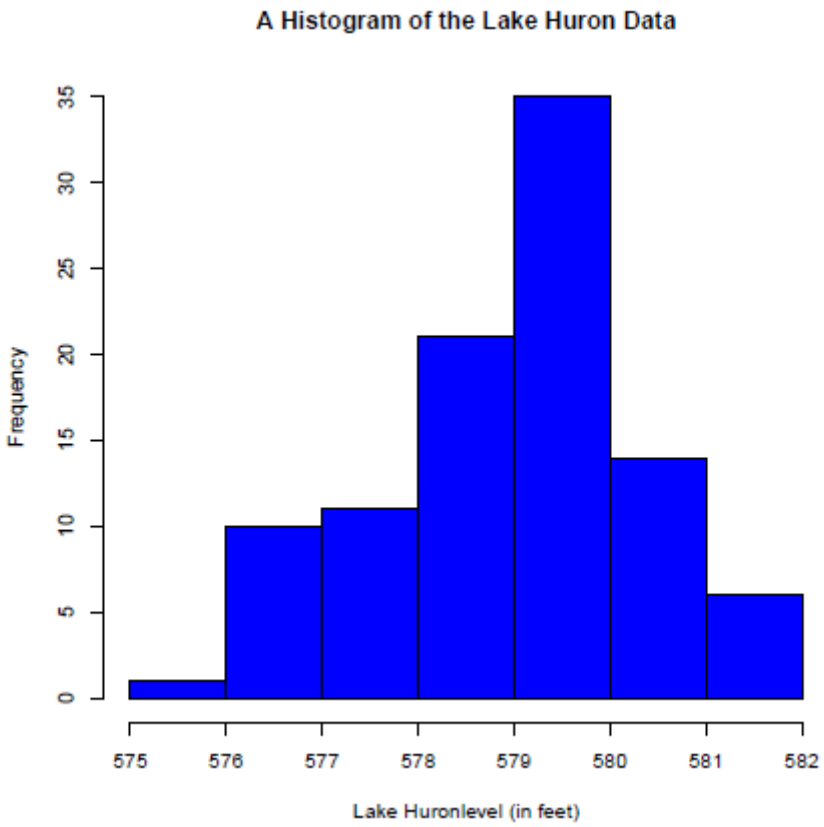
6. Czy jest jakiś sposób, abyśmy mogli użyć języka R, aby uzyskać obraz danych?

Chociaż mamy pewne pojęcie o tym, jak dane są dystrybuowane (najniższy poziom to 576, najwyższy to 582, a średnia to 579), obraz może dostarczyć dodatkowych informacji.

Odpowiedź: Możemy użyć funkcji hist() do stworzenia histogramu danych.

Użyj funkcji hist(), aby podać histogram; Ustaw kolor niebieski.

>hist(LakeHuron, col = blue, xlab = Lake Huronlevel (in feet), ylab = Frequency, main = A Histogram of the Lake Huron Data)



Histogram zapewnia nieco lepszy wgląd w rozkład wartości danych.

W rzeczywistości dane wydają się być rozłożone w miarę normalnie (to znaczy rozkład

jest ukształtowany w sposób zgodny z normalną krzywą w kształcie dzwonu) wokół

średniej 579.

7. Zwracamy uwagę, że prosty histogram zapewnia wizualny wgląd w to, w jaki sposób dane są rozłożone. Skoro właśnie odwołaliśmy się do normalnego rozkładu krzywej dzwonowej, to czy można zobaczyć histogram tego?

Odpowiedź: Możemy użyć funkcji rnorm(n) do wygenerowania zbioru n o rozkładzie normalnym.

wygeneruj dane i przypisz je do obiektu, używamy funkcji hist()

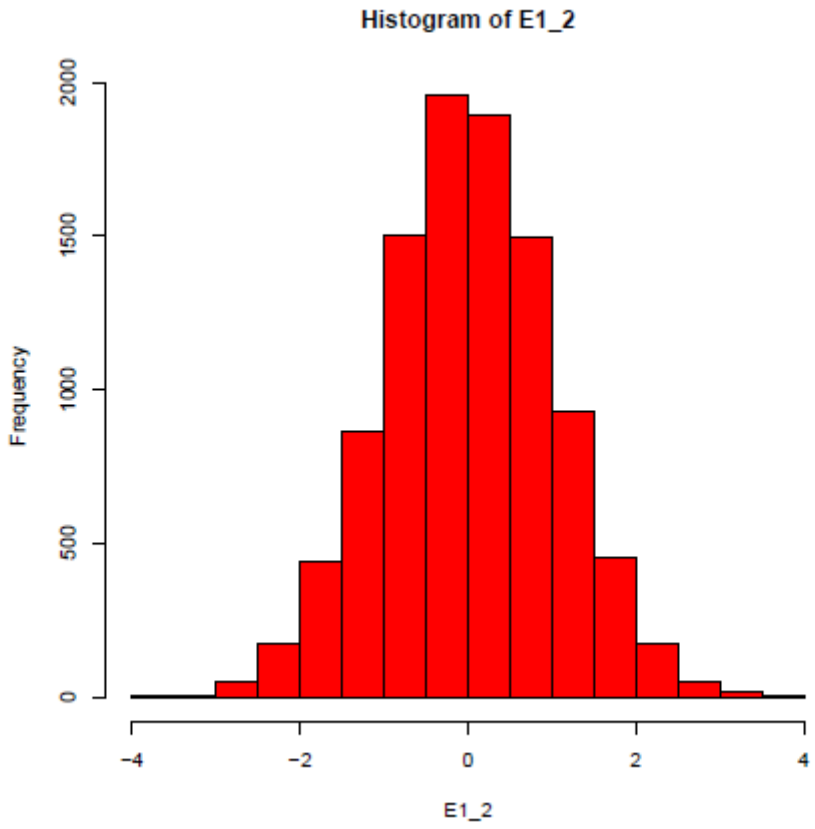
Utwórz histogram.

(1) Użyj funkcji rnorm(10000), aby wygenerować 10 000 normalnie rozłożony wartości danych; Nazwij wynik E1_2.

>E1_2 <- rnorm(10000)

Użyj funkcji hist(), aby utworzyć histogram; Ustaw kolor na czerwony.

>hist(E1_2, col = ‘red’)



8. Czy powinniśmy być bardzo pewni, że dane te są rzeczywiście dobrym odzwierciedleniem rzeczywistego poziomu wody w jeziorze Huron w okresie od 1875 do 1972 roku? Co mogą być jakeś niekontrolowane wpływy na pomiary dokonywane co roku ?

Odpowiedź: Godziny i dni, w których wykonywane są pomiary, będą miały znaczenie. Nie jest to jednak regułą, w której nie ma nic wspólnego z porami roku? Na przykład na wiosnę poziom wody byłby prawdopodobnie wyższy (z powodu spływu z topniejącego śniegu i obfitych opadów wiosennych) niż jesienią (po upalnym okresie parowania). Ponadto, czy pomiary są wykonywane w dokładnie tej samej lokalizacji, przypuszczalnie gdzieś w pobliżu środka jeziora? Możliwe, że nie ma zapisów o tym, gdzie pomiary zostały wykonane, zwłaszcza we

wcześniejszych latach? Chodzi o to, że zawsze musimy być sceptyczni (i zadawać pytania) co do jakości naszych danych zanim będziemy w stanie wyciągnąć z nich rozsądne wnioski.

9. Utwórz ramkę danych składającą się z siedmiu największych narodów świata

zmienne: ludność, PKB i procent ludności miejskiej. (Skorzystaj z tabeli 1.) Nazwa

ramka danych E1 3; nazwij zmienne Naród, Populacja, PKB i Procent Zurbanizowania.

Country	Population	GDP	Urban
Bangladesh	144,000,000	$1,700	28%
Brazil	204,000,000	$10,800	87%
China	1,439,000,000	$7,600	47%
India	1,380,000,000	$3,500	30%
Indonesia	274,000,000	$4,200	44%
Pakistan	221,000,000	$2,500	36%
US	331,000,000	$47,200	82%

Tabela 1: Kraje siedmiu najbardziej zaludnionych krajów świata

(1) co robi poniższa funkcja

>option(scipen = 999)

(2) Utwórz wektor składający się z nazw krajów; przypisz wynik do obiektu o nazwie var1. Uwaga: imiona i nazwiska są zawarte w cudzysłowiu

>var1 <- c(Bangalesh, Brazil, China, India, Indonesia, Pakistan, US)

>var2 <- c(144000000, 204000000, 1439000000, 1380000000, 274000000, 221000000, 331000000)

>var3 <- c(1700, 10800, 7600, 3500, 4200, 2500, 47200)

>var4 <- c(28, 87, 47, 30, 44, 36, 82)

>E1_3 <- data.frame(Nation = var1, Population = var2, GDP = var3, PercentUrban = var4)

Zobacz co utworzyłeś:

>E1_3

10. Odpowiedz na poniższe pytania dotyczące ramki danych E1_3.

a) Znajdź zbiorcze dane statystyczne (średnią, medianę, wartość maksymalną, minimalną, pierwszy i trzeci kwartyl dla zmiennej Populacja.

>summary(E1_3$Population)

b) Znajdź zbiorcze dane statystyczne (średnia, mediana, maksimum, minimum, pierwszy i trzeci kwartyl) dla zmiennej PKB.

>summary(E1_3$PKB)

c) Znajdź zbiorcze dane statystyczne (średnia, mediana, wartość maksymalna,

minimum, pierwszy i trzeci kwartyl) dla zmiennej Percentage Urban.

>summary (E1_3$PercentUrban)

11. Badanie marketingowe przeprowadzone wśród 1095 gospodarstw domowych w celu zbadania postaw wobec następujących marek A, B, C, D, E i F w określonej kategorii produktu ujawnia następująca struktura preferencji marki: 272 preferuje markę A, 212 preferuje markę B, 297 preferuje C, 38 preferuje D, 181 E i 95 F. Utwórz obiekt o nazwie E2 1, który zawiera te informacje, a następnie podaj rozkład częstotliwość preferencji między tymi sześcioma markami.

>E2_1 <- c(rep(A, 272), rep(B, 212), rep(C, 297), rep(D, 38), rep(E, 181), rep(F, 95))
>fd <- table(E2_1)

Sprawdź zawartość:

>Fd

W ten sposób funkcja table() zapewnia rozkład częstotliwości dla sześciu marek.

12. Stwórz względny rozkład częstotliwości preferencji marki. Użyj danych E2_1.

>fd <- table(E2_1)

Utwórz relatywne częstotliwości występowania i przypisz do obiektu rf

>rf <- fd / sum(fd)

>rf

Względny rozkład częstotliwości preferencji marki: A wynosi 0,25, B wynosi 0,19, C wynosi 0,27, D wynosi 0,03, E wynosi 0,17, a F wynosi 0,09.

13. Pokaż wykres słupkowy z częstotliwościami preferencji marki. Ustawianie zakresu pionowego oś od 0 do 300. Określ kolory pasków, od lewej do prawej, jako zielony, niebieski, czerwony, żółty, fioletowy i pomarańczowy. Podaj etykietę zarówno w poziomie, jak i w pionie, a także główny tytuł obrazu. Użyj danych E2 1.

```
>fd <- table(E2_1)

>barplot(fd,
col = c(green, blue, red, yellow, purple, orange),
ylim = c(0, 300),
main = Number of Households Preferring Brand,
xlab = Brands,
ylab = Brand Preference Frequencies)
```

14. Pokaż wykres słupkowy względnych częstości preferencji marki. Ustaw zakres oś pionowa od 0 do 0,30. Oznacz kolory pasków, od lewej do prawej, jako czerwony, niebieski, czerwony, niebieski, czerwony i niebieski. Podaj etykietę zarówno dla osi poziomej, jak i pionowej, a także główny tytuł obrazu. Użyj danych E2_1.

```
>fd <- table(E2_1)

> rf <- fd / sum(fd)

>barplot(rf,
col = c(red, blue, red, blue, red, blue),
ylim = c(0, 0.30),
xlab = Brands,
ylab = Relative Frequencies,
main = Proportion of Households Preferring Brand)
```

15. Pokaż wykres punktowy względnej częstotliwości preferencji marki. Użyj danych E2_1.

```
> fd <- table(E2_1)

> rf <- fd / sum(fd)

> dotchart(sort(rf),
xlab = Relative Frequencies Brand is Preferred,
main = Relative Frequencies by Brand,
pch = 19,
```

col = blue)

Uwaga: konieczne jest posortowanie danych o względnej częstotliwości, jeśli chcemy, aby wykres kropkowy przebiegał w kolejności sekwencyjnej od lewego dolnego rogu do prawego górnego rogu. Odbywa się to przez zagnieżdżenie funkcji sort() jako argumentu w funkcji dotchart(). Jeśli pominiemy funkcję sort() i dołączymy tylko nazwę obiektu (w tym przypadku rf), Punkty na wykresie są domyślnie uporządkowane alfabetycznie.

Uwaga: Ta procedura zawiera komunikat ostrzegawczy, który możemy zignorować, ponieważ

Funkcja dotchart() jest wykonywana pomyślnie i tworzy obraz wykresu kropkowego.

16. Ustaw rozkład częstotliwości dla tych wartości: 24, 29, 34, 29, 37, 26, 30, 34, 30, 11, 12, 14, 18, 38, 17, 13, 16, 12, 33, 35, 35, 29, 28, 26, 25, 34, 11, 16, 19, 11, 13, 36, 12, 12, 12, 26, 36, 16, 26, 22, 15, 29, 38, 34 i 30. Ustaw szerokość klasy na 5.

>E2_2 <- c(24, 29, 34, 29, 37, 26, 30, 34, 30, 11, 12, 14, 18, 38, 17, 13, 16, 12, 33, 35, 35, 29, 28, 26, 25, 34, 11, 16, 19, 11, 13, 36, 12, 12, 12, 26, 36, 16, 26, 22, 15, 29, 38, 34, 30)

Wczytaj dane do obiektu brks:

>brks <- c(10, 14.99, 19.99, 24.99, 29.99, 34.99, 39.99)

Użyj cut() aby przypisać wartości w E2_2 do kategorii zdefiniowanych brks

>categ <- cut(E2_2, brks)

Użyj table() aby uzyskać rozkład częstotliwości występowania wartości danych w categ

> fd <- table(categ)

>fd

Tak więc 11 wartości należy do pierwszej kategorii (od 10 do 15), 7 do drugiej (od

15 do 20), 2 w trzecim, 10 w czwartym, 8 w piątym i 7 w szóstym.

17. Utwórz względny rozkład częstotliwości danych E2_2.

>rf <- fd / sum(fd)

>rf

Tak więc 0,24 obserwacji przypada na pierwszą klasę, 0,16 na drugą, 0,04 na drugą.

trzecia, 0,22 w czwartej, 0,18 w piątej i 0,16 w szóstej klasie.

18. Pokaż histogram częstotliwości dla danych E2_2.

Ustaw zakres oś pozioma z zakresu od 0 do 45, zakres osi pionowej z zakresu od 0 do 12.

Dodaj główny tytuł i etykiety dla osi pionowej i poziomej. Ustaw kolor niebieski jako kolor.

```
hist(E2_2,
breaks = c(9.99, 14.99, 19.99, 24.99, 29.99, 34.99, 39.99),
xlim = c(0, 45),
ylim = c(0, 12),
xlab = x-values,
ylab = Frequencies,
main = Six Categories,
col = blue)
```

19. Duża restauracja typu fast food znajdująca się w centrum Birmingham zbiera próbkę n = 199 czeków klienta sporządzonych w ostatnim dniu roboczym w celu uzyskanie wglądu w rozkład kwoty, jaką wydają ich klienci (£). Użyj zestaw danych check.csv. Jako pierwszy krok do wyznaczenia dobrej szerokości dla kategorii rozkładu częstotliwości, użyj funkcja summary(). Co mówią nam te statystyki?

```
>E2_3 <- read.csv(check.csv)
```

```
> summary(E2_3)
```

Ponieważ minimalne i maksymalne wartości wynoszą odpowiednio 6,72 i 97,74, musimy

podzielić histogram na pięć kategorii z grubsza (97,74-6,72)/5 = 91,02/5 = 20 w przybliżeniu

20. Utwórz rozkład częstotliwości dla E2_3. Wskazówka: najpierw użyj funkcji names(), aby określić, jaka jest nazwa zmiennej.

```
> names(E2_3)
```

Wpisz wartości do brks

```
> brks <- c(0, 20, 40, 60, 80, 100)
```

Użyj cut() aby przypisać wartości w E2_3 do kategorii zdefiniowanych w brks i przypisz do obiektu categ

```
>categ <- cut(E2_3$amount, brks)
```

Użyj funkcji table() by uzyskać rozkład częstotliwości występowania wartości danych w categ
I przypisz wynik do fd

> fd <- table(categ)

>fd

Rozkład częstotliwości jasno wskazuje, że dane nie są dystrybuowane nie w sposób normalny, ale bimodalnie.

21. Pokaż względny rozkład częstotliwości danych E2_3. Utwórz relatywną częstotliwość poprzez podzielenie fd przez całkowitą liczbę danych i przypisz wynik obiektowi rf

>rf <- fd / sum(fd)

>rf

Tak więc 0,09 (zliczeń) należy do pierwszej kategorii, 0,34 do drugiej, 0,15 do

trzeciej, 0,34 w czwartej i 0,08 w piątej.

22. Sporządź histogram danych E2_3 za pomocą 5 kategorii. Dołącz główny tytuł i etykiety dla osi poziomej i pionowej. Określ zakres osi pionowej od 0 do 80 i ustaw kolor fioletowy.

```
> hist(E2_3$amount,
breaks = c(0, 20, 40, 60, 80, 100),
col = purple,
ylim = c(0, 80),
xlab = x values,
ylab = Frequencies,
main = Five Categories of the Amount that Customers Spend)
```

Z histogramu jasno wynika, że nawet jeśli centralna tendencja rozkładu wynosi około 50 (zgodnie ze średnią i medianą), dane są rzeczywiście bimodalne, nie normalne.

23. Pokaż histogram z dziesięcioma klasami. Czy dodatkowa precyzja dziesięciu klas zapewnia jakieś dodatkowe spostrzeżenia podczas próby interpretacji rozkładu danych?

Ponownie dodaj główny tytuł i etykiety dla osi pionowej i poziomej. Sprecyzuj zakres osi pionowej biegnący od 0 do 40 i ustaw kolor zielony jako kolor.

```
> hist(E2_3$amount,
breaks = 10,
col = green,
```

ylim = c(0, 40),
xlab = x values,
ylab = Frequencies,
main = Ten Categories of the Amount that Customers Spend)

Zauważ, że zamiast definiować klasy za pomocą breaks=c(), jak to zrobiliśmy w poprzednim przypadku możemy również skorzystać z breaks=10. Zobacz drugi argument hist() funkcji powyżej. Po bliższym przyjrzeniu się okazuje się, że użycie dziesięciu kategorii, a nie 5 nie daje dokładniejszego wglądu w charakter rozkładu wartości danych. Mimo że tak jest czasami korzystne jest podzielenie danych na więcej (ale węższych) kategorii ponieważ wzorce, które nie były dostrzegalne przy mniejszej liczbie kategorii, mogą zostać ujawnione, gdy dane zostaną podzielone na więcej kategorii.

24. Korzystając z danych parabolic.csv, wyświetl relację między dwiema zmiennymi,

x i y. Która metoda opisowa Twoim zdaniem najlepiej sprawdza się w tym przypadku?

>parabolic <- read.csv(parabolic.csv)

> plot(parabolic$x, parabolic$y,
pch = 21,
col = "blue",
xlab = "Variable x",
ylab = "Variable y",
main = "Two Variables with a Non-Linear Relationship")

Wykres punktowy prawdopodobnie działa najlepiej ze wszystkich, ponieważ zapewnia jasny obraz powiązań między dwiema zmiennymi. W tym przypadku relacja między nimi zmienne x i y nie są liniowe, ale bardziej paraboliczne.

25. Korzystając z danych negative.csv, wyświetl relację między dwiema zmiennymi,

x i y. Jak powinniśmy opisać tę relację?

> negative <- read.csv(negative.csv)

> plot(negative$x, negative$y,
pch = 23,
col = "red",
xlab = "x",
ylab = "y")
Dwie zmienne x i y wydają się być ujemnie (i liniowo) powiązane.

26. Korzystając z danych positive.csv, pokaż relację między dwiema zmiennymi, x

oraz y. Co możemy powiedzieć o tej relacji?

> positive <- read.csv(positive.csv)

```
> plot(positive$x, positive$y,
pch = 25,
col = "purple",
xlab = "x",
ylab = "y")
```

Zmienne x i y wydają się być dodatnio (i liniowo) powiązane.

27. Wyświetl pierwsze siedem obserwacji i pierwsze siedem kolumn (zmiennych) z Cars93

(zbiór danych). W pierwszym kroku zaimportuj dane do E2_4. Wyznacz rozkład częstotliwości

zmiennej Type.

> library(MASS)

> E2_4 <- Cars93

> head(E2_4[1 : 7], 7)

Użyj table() aby uzyskać rozkład częstotliwości Type i przypisz do fd

> fd <- table(E2_4$Type)

Sprawdź co masz

>fd

Rozkład częstotliwości typu pokazuje, że 93 pojazdy są rozmieszczone wedle sześciu typów pojazdów: 22 pojazdy są średniej wielkości, 21 są małe, 16 jest kompaktowych, 14 sportowe, 11 to duże, a 9 to samochody dostawcze.

28. Ustal względny rozkład częstotliwości typu pojazdu dla danych Cars93 i

Przypisz wynik do RFD. Jaki procent stanowią duże samochody? Sprawdź, czy wszystkie

Proporcje (procenty) sumują się do 1.

Rozkład częstotliwości

>rfd <- fd / nrow(E2_4)

>rfd

>sum(rfd)

Udział dużych pojazdów pasażerskich w zbiorze danych Cars93 wynosi prawie 0,12

(0,11828), czyli około 12%. Po zsumowaniu proporcje sumują się do jednego.

29. Zrób wykres słupkowy względnych częstości zmiennej Type w danych Cars93. Określ kolory pasków, od lewej do prawej, jako czerwony, niebieski, żółty, fioletowy, pomarańczowy i zielony. Ustaw zakres osi pionowej tak, aby mieścił się w zakresie od 0 do 0,25.

Dodaj Typy pojazdów" jako etykietę dla osi poziomej, Względne częstotliwości" dla

osi pionowej. Na koniec dodaj Względne częstotliwości typów pojazdów" jako tytuł.

```
>barplot(rfd,
col = c(red, blue, yellow, purple, orange, green),
xlab = Vehicle Types,
ylab = Relative Frequencies,
main = Relative Frequencies of Vehicle Types,
ylim = c(0, 0.25))
```

30. Utwórz wykres punktowy względnych częstości zmiennej Type w danych Cars93. Użyj funkcji sort(), aby uszeregować typy pojazdów od najbardziej reprezentatywnych do najmniej. Ustaw punkty wykresu kropkowego jako niebieskie i dołącz Względne częstotliwości" jako etykietę dla osi poziomej. Dodaj Względne częstotliwości typów pojazdów" jako tytuł.

```
>dotchart(sort(rfd),
main = Relative Frequencies of Vehicle Types,
xlab = Relative Frequencies,
pch = 19,
col = blue)
```

31. Korzystając z danych Cars93, wyznacz rozkład częstotliwości zmiennej Max.Price (maksymalna cena dla każdej z 93 marek i modeli). Ustaw szerokość klasy na 10, określenie najniższego przedziału cenowego na poziomie lub poniżej od 10 000 USD do 20 000 USD do najwyższego przedziału cenowego od 70 000 USD do 80 000 USD Skomentuj rozkład cen pojazdów w 93 samochodach w Cars93.

>brks <- c(0, 10, 20, 30, 40, 50, 60, 70, 80)

Przypisz wartości z E2_4 do przedziałów brks

>categ <- cut(E2_4$Max.Price, brks)

Użyj table() by uzyskać rozkład częstotliwości

>fd <- table(categ)

>fd

Rozkład częstotliwości wskazuje, że tylko 5 pojazdów ma ceny powyżej 40 000 USD;

8 ma ceny 10 000 USD lub mniej. Większość pojazdów, 69 z nich, jest wyceniana w

zakresie od 10 000 dolarów do 30 000 USD.

32. Znajdź względne częstości zmiennej Max.Price danych Cars93. Skomentuj przedziały cenowe i upewnij się, że względne częstotliwości sumują się do jednego.

Uzykaj relatywne częstotliwości

>rfd <- fd / sum(fd)

>rfd

>sum(rfd)

Ze względnych częstotliwości jasno wynika, że prawie 75% maksymalnych cen wszystkie pojazdy osobowe w danych Cars93 mieszczą się od 10 000 USD do 30 000 USD; ponad 17% jest wycenionych powyżej 30 000 USD podczas gdy mniej niż 9% kosztuje mniej niż 10 000 USD Wszystkie Częstości sumują się do jednego.

33. Sporządź histogram częstotliwości zmiennej Max.Price z Cars93. Ustaw kolory pasków histogramu (biegnące od lewej do prawej) jako: czerwony, różowy, niebieski, żółty, fioletowy, pomarańczowy, szary i zielony. Dodaj Maksymalna cena pasażera

Pojazdy (w 000 USD)" jako etykieta dla osi poziomej; dołącz tytuł Częstości

cen." Uwzględnij breaks=8 jako argument funkcji hist().

```
hist(E2_4$Max.Price,
breaks = 8,
xlab = Maximum Price of Passenger Vehicles (in $000),
main = Frequencies of Prices,
col = c(red, pink, blue, yellow, purple, orange,
grey,green))
```

Rozkład częstotliwości przedstawiony na histogramie wydaje się być nieco pochylony (od rozkładu normalnego) w prawo. Pojawiają się dwie wartości odstające, z których jedna znajduje się w 80 000 dolarów (Mercedes Benz 300E) i jeden w 50 400 dolarów.

34. Uporządkuj dane Cars93 w podstawową tabelę krzyżową, która raportuje pojazd obok kraju pochodzenia. Czy w tej konkretnej próbie prawdą jest, że większość dużych pojazdów jest pochodzenia amerykańskiego?

>crosstab <- table(E2_4$Type, E2_4$Origin)

>crosstab

Jak wynika z tabeli krzyżowej, wszystkie duże pojazdy są pochodzenia amerykańskiego.

35. Uporządkuj dane Cars93 w tabeli krzyżowej ze zmiennymi Man.trans.avail (Czy dostępna jest manualna skrzynia biegów?) i Początek zorganizowany wzdłuż dwóch marginesów.

>crosstab <- table(E2_4$Man.trans.avail, E2_4$Origin)

>crosstab

36. Dodaj sumy kolumn i wierszy do tabeli krzyżowej Man.trans.avail i Origin z pliku Cars93. Czy pojazdy amerykańskie są bardziej skłonne (niż pojazdy spoza USA) do oferowana kupującym opcję manualnej skrzyni biegów?

> Totals <- rowSums(crosstab)

>crosstab <- cbind(crosstab, Totals)

>Totals <- colSums(crosstab)

>crosstab <- rbind(crosstab, Totals)

>crosstab

Z tabeli krzyżowej jasno wynika, że znacznie większy odsetek pojazdów oferuje nabywcom korzystających z opcji manualnej skrzyni biegów są pochodzenia spoza USA, 87% (lub 39 z 45) do zaledwie 46% (lub 22 z 48) w przypadku pojazdów pochodzących z USA.

37. Uporządkuj dane Cars93 w tabeli krzyżowej ze zmiennymi Max.Price i Rozmiar silnika. Zmniejsz liczbę kategorii cenowych do czterech|(0,20], (20,40], (40,60], i (60,80]|oraz liczbę kategorii pojemności silnika (w litrach pojemności skokowej) do trzech |(0,2], (2,4] i (4,6).

>brks <- c(0, 2, 4, 6)

>displacement <- cut(E2_4$EngineSize, brks)

>brks <- c(0, 20, 40, 60, 80)

>price <- cut(E2_4$Max.Price, brks)

>crosstab <- table(displacement, price)

>crosstab

38. Dla tabeli krzyżowej zestawu danych Cars93 (zmienne to Max.Price i EngineSize), zmień nazwy wierszy: od 1 do 2 litrów, od 2 do 4 litrów i od 4 do 6 Litrów. Zmień nazwy kolumn: Ekonomiczna, Średnia cena, Wyższa cena i Luksusowy.

> rownames(crosstab) <- c(1 to 2 liters, 2 to 4 liters, 4 to 6 liters)
>colnames(crosstab) <- c(Economy, Mid-Price, Higher-Price, Luxury)
>crosstab

39. Dodaj sumy wierszy i kolumn do tabeli tabelarycznej zestawu danych Cars93

gdzie zmiennymi są Max.Price i EngineSize.

> Totals <- rowSums(crosstab)

>crosstab <- cbind(crosstab, Totals)

>Totals <- colSums(crosstab)

>crosstab <- rbind(crosstab, Totals)

>crosstab

40. Korzystając z danych Cars93, wykonaj wykres punktowy, aby pokazać związek między

Pojemność silnika i maksymalna cena. Dodaj ”Diagram punktowy odnoszący się do pojemności silnika i Cena” jako tytuł; wyznaczają osie jako cenę pojazdu i pojemność silnika

w litrach. Skomentuj relację między tymi dwiema zmiennymi.

>plot(E2_4$Max.Price, E2_4$EngineSize,
main = A Scatter Diagram Relating Engine Size and Price,
pch = 23,
col = purple,
ylab = Engine Displacement in Liters,
xlab = Vehicle Price)

Ogólnie rzecz biorąc, wydaje się, że istnieje pozytywna zależność między wielkością silnika a ceną: Pojemność silnika jest (w przybliżeniu) pozytywnie związana z ceną pojazdu.

41. Skonstruuj wykres punktowy (przy użyciu danych Cars93) dwóch zmiennych: EngineSize (w litrach pojemności użytkowej) w stosunku do koni mechanicznych (maksymalna moc). Dodaj etykietę „Maksymalna moc" i „pojemność silnika" w litrach" do osi wertykalnej i horizontalnej. Dołącz również diagram punktowy odnoszący się do Pojemność silnika i moc jako tytuł; Ustaw kolor niebieski jako kolor kreślonego znaku. Skomentuj związek.

```
>plot(E2_4$EngineSize, E2_4$Horsepower,
main = A Scatter Diagram Relating Engine Size and Horsepower,
xlab = Engine Displacement in Liters,
ylab = Maximum Horsepower,
pch = 19,
col = blue)
```

Zgodnie z oczekiwaniami te dwie zmienne są dodatnio i liniowo powiązane: ogólnie rzecz biorąc, Im większy silnik, tym większa moc.

42. Korzystając z danych Cars93, utwórz wykres punktowy EngineSize (w litrach pojemności względnej) przeciwko MPG.highway (mile autostradowe na galon amerykański). Dodaj nazw etykiet do osie poziome i pionowe, a także tytuł główny. Skomentuj związek.

```
>plot(E2_4$EngineSize, E2_4$MPG.highway,
main = A Scatter Diagram Relating Engine Size and Miles
Per US Gallon,
xlab = Engine Displacement in Liters,
ylab = Highway Miles Per US Gallon,
pch = 24,
col = red)
```

Nic dziwnego, że te dwie zmienne są ujemnie (i nieco liniowo) powiązane:

Ogólnie rzecz biorąc, im większa pojemność silnika, tym niższy przebieg benzyny.

43. Korzystając w dalszym ciągu z danych Cars93, należy utworzyć wykres punktowy przedstawiający Max. Price i RPM (obroty na minutę). Czy te dwie zmienne są w pozytywny czy negatywny sposób? A może wydają się być ze sobą niepowiązane? Dodawać etykiety Prędkości obrotowe na minutę przy maksymalnej mocy i cenie pojazdu do odpowiednio osie pionowe i poziome. Dołącz diagram punktowy odnoszący się do Cena pojazdu i obroty na minutę jako tytuł. Ustaw kolor fioletowy jako znak kreślenia kolor aktu.

```
>plot(E2_4$Max.Price, E2_4$RPM,
```

main = A Scatter Diagram Relating Vehicle Price and
Revs per Minute,
ylab = Revs per Minute at Maximum Horsepower,
xlab = Vehicle Price,
pch = 20,
col = purple)

Ponieważ nie ma powodu, aby podejrzewać, że te dwie konkretne zmienne są ze sobą powiązane, Zarówno pozytywnie, jak i negatywnie, nie jesteśmy zaskoczeni widząc tę chmurę punktów danych.

44. Próbka zawiera następujące wartości danych: 1.50, 1.50, 10.50, 3.40, 10.50, 11.50,

oraz o godz. 2.00. Utwórz wektor o nazwie E3 1. Znajdź średnią.

>E3_1 <- c(1.50, 1.50, 10.50, 3.40, 10.50, 11.50, 2.00)

>mean(E3_1)

45. Znajdź medianę próbki (powyżej) na dwa sposoby: (a) użyj funkcji median(), a następnie (b) użyj funkcji sort(), aby wizualnie zlokalizować środkową wartość.

>median(E3_1)

>sort(E3_1)

46. Utwórz wektor z następującymi elementami: -37.7, -0.3, 0.00, 0.91, e, , 5.1, 2e i 113754, gdzie e jest podstawą logarytmu naturalnego (w przybliżeniu 2,718282...), a stosunek średnicy koła do jego promienia (około 3,141593...). Nazwij obiekt E3 2. Jaka jest mediana i średnia? 78. percentyl? Ile wynosi wariancja i odchylenie standardowe? Zauważ, że R rozumie exp(1) jako e, pi jako Π.

>E3_2 <- c(-37.7, -0.3, 0.00, 0.91, exp(1), pi, 5.1, 2*exp(1), 113754)

>mean(E3_2)

>median(E3_2)

>quantile(E3_2, prob = c(0.78))

>var(E3_2)

>sd(E3_2)

Średnia wynosi 12 637,03; Mediana wynosi 2,718282.. Z uwagi na fakt, że dane wartości w E3_2 są ułożone w porządku rosnącym, medianę można łatwo zidentyfikować jako Środkowa

wartość, e (lub 2,718282...), ponieważ poniżej znajdują się cztery wartości, a cztery wartości. Nie jest to jednak regułą. Co więcej, po prostu sumując wszystkie dziewięć wartości danych i dzieląc przez dziewięć, podaje średnią. 78 percentyl jest podawany jako 5,180775; wariancja i Odchylenie standardowe wynosi odpowiednio 1 437 840 293 i 37 918,86.

47. Weź pod uwagę następujące wartości danych: 10, 20, 30, 40, 50, 60, 70, 80, 90 i 100.

Ile wynoszą 10. i 90. percentyl? Wskazówka: użyj funkcji seq(from=,to=,by=) , aby utworzyć zestaw danych. Nazwij zestaw danych E3 3.

>E3_3 <- seq(from = 10, to = 100, by = 10)
>E3_3
>quantile(E3_3, probs = c(0.1, 0.9))
Percentyl 10. i 90. to odpowiednio 19 i 91. percentyl. Należy pamiętać, że 10. percentyl (19) oznacza wartość, która przekracza co najmniej 10 % pozycji w zbiorze danych; 90. percentyl (91) to wartość, która przekracza co najmniej 90% pozycji. Należy również pamiętać o tym, że możliwe jest zdefiniowanie dowolnych percentyli poprzez ustawienie wartości w probs=c() funkcji quantiles().

48. Jaka jest mediana E3 3? Znajdź środkową wartość wizualnie i za pomocą mediany()

>median(E3_3)

Ten zestaw danych ma parzystą liczbę wartości, wszystkie ułożone rosnąco porządek. W związku z tym medianę ustala się, biorąc średnią wartości w dwie środkowe pozycje: średnia 50 (wartość na 5. pozycji) i 60 (wartość na 6. pozycji) wynosi 55.

49. Modalna to wartość, która występuje z największą częstotliwością w zestawie danych i jest to jedna z miar tendencji centralnej. Rozważmy próbkę z dziewięcioma wartości: 5, 1, 3, 9, 7, 1, 6, 11 i 8. Czy modalna zapewnia miarę centralnej tendencji podobne jak do średnia? Mediana?

>E3_4 <- c(5, 1, 3, 9, 7, 1, 6, 11, 8)

Częstotliwość występowania

>table(E3_4)

>mean(E3_4)

>median(E3_4)

Ponieważ wartość trybu w tym przypadku wynosi 1 (pojawia się dwukrotnie),

daje mniejszy wgląd w centralną tendencję próbki niż średnia (5,667) lub mediana (6).

50. Rozważ inną próbkę z dziewięcioma wartościami: 5, 1, 3, 9, 7, 4, 6, 11 i 8. W jaki sposób modalna oddaje centralną tendencję tej próbki? Ponieważ wszystkie elementy danych pojawiają się tylko raz, nie ma jednej wartości dla modalnej; Dostępnych jest dziewięć modalnych, po jednej dla każdej wartości danych.

51. Znajdź 90. percentyl, 1., 2. i 3. kwartyl, a także minimalną i maksymalną wartości zestawu danych LakeHuron (który jest częścią języka R). Jaka jest średnia? Jaka jest mediana?

>quantile(LakeHuron, prob = c(0.00, 0.25, 0.50, 0.75, 0.90, 1.00))

>mean(LakeHuron)

>median(LakeHuron)

Wartość minimalna (percentyl 0) wynosi 575,960, a wartość maksymalna (percentyl 100. percentyl) wynosi 581,860; 1., 2. i 3. kwartyl to 578,135, 579,120, i 579,875, odpowiednio. Mediana (znana również jako 2. kwartyl lub 50. percentyl) wynosi 579,120. Średnia wynosi 579,0041, a 90. percentyl wynosi 580,646.

52. Znajdź zakres, rozstęp międzykwartylowy, wariancję, odchylenie standardowe i

współczynnik zmienności zbioru danych LakeHuron.

>max(LakeHuron) - min(LakeHuron)

>IQR(LakeHuron)

>var(LakeHuron)

>sd(LakeHuron)

>sd(LakeHuron) / mean(LakeHuron)

Zakres wynosi 5,9 stopy, a rozstęp międzykwartylowy wynosi 1,74 stopy. Co więcej, wariancja a odchylenie standardowe wynoszą odpowiednio 1.737911 i 1.318299 stóp. Wreszcie, współczynnik zmienności wynosi 0,002276838; Oznacza to, że odchylenie standardowe wynosi tylko około 0:228% średniej.

53. Jaki jest przedział i rozstęp między kwartylowy dla następującego zbioru danych: -37,7, -0,3, 0,00, 0,91, e, , 5,1, 2e i 113754? Należy pamiętać, że jest to ten sam zestaw danych, co ten użyty powyżej, gdzie nazwaliśmy go E3 2.

>max(E3_2) - min(E3_2)

>IQR(E3_2)

Zakres wynosi 113791.7; Rozstęp międzykwartylowy wynosi 5,1.

rozstęp międzykwartylowy Udostępnia zakres środkowych 50% danych, podczas gdy zakres obejmuje wszystkie wartości, w tym wartości odstające.

54. Ćwiczenie 11 daje nam możliwość przećwiczenia pisania podstawowego kodu R. Pisanie własnego kodu, oraz wariancja próbki i odchylenie standardowe próbki dane E3 3 (zob. ćwiczenie 4). Sprawdź obie odpowiedzi z odpowiedziami używającymi var() i funkcji sd(). Przypomnijmy, że wyrażenie dla wariancji próbki to

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

>xbar <- mean(E3_3)

> devs <- (E3_3 - xbar)

>sqrd.devs <- (devs) ^ 2

>sum.sqrd.devs <- sum(sqrd.devs)

>variance <- sum.sqrd.devs / (length(E3_3) - 1)

>variance

>standard.deviation <- sqrt(variance)

>standard.deviation

>var(E3_3)

>sd(E3_3)

Wariancja wynosi 916,6667; Odchylenie standardowe wynosi 30,2765.

55. Zbiór danych temps.csv obejmuje wysokie i niskie temperatury (w stopniach Celsjusza) na dzień 1 kwietnia 2021 r. w dziesięciu głównych miastach europejskich; zaimportuj dane do E3 5. Jaka jest kowariancja wysokiej i niskiej temperatury ? Co mówi nam kowariancja?

Odpowiedź: Kowariancja tych dwóch zmiennych wynosi 37,28889 te dwie zmienne są dodatnio powiązane. Nie jest to zaskakujące odkrycie ponieważ, ogólnie rzecz biorąc, miasta, w których temperatury w ciągu dnia są najcieplejsze dni mają najcieplejsze temperatury w nocy.

>temps <- read.csv(temps.csv)

>E3_5 <- temps

>head(E3_5, 3)

>cov(E3_5$Daytemp, E3_5$Nighttemp)

56. Aby zdobyć nieco więcej praktyki w pisaniu kodu R, oblicz kowariancję dwóch zmiennych. Daytemp i Nighttemp z danych E3 5. Przypomnijmy, że kowariancja próbki Pomiędzy dwiema zmiennymi x i y jest:

>devx <- (E3_5$Daytemp - mean(E3_5$Daytemp))

>devy <- (E3_5$Nighttemp - mean(E3_5$Nighttemp))

>crossproduct <- devx * devy

>covariance <- sum(crossproduct) / (length(E3_5$Daytemp) - 1)

>covariance

57. Plik daily_idx_ chg.csv składaj się z procentowej dziennej zmiany (w stosunku do poprzedniego dnia) wartości zamknięcia dla dwóch indeksów giełdowych Dow Jones Industrial Average i S&P500, w ciągu ostatnich 20 dni handlowych. Jaka jest kowariancja ruchów cen dla tych dwóch indeksów? Co kowariancja mówi nam o związku między tymi dwiema zmiennymi? W pierwszym kroku zaimportuj daily_idx_ chg.csv do_ obszaru roboczego języka R, a następnie nazwij ramkę danych E3 6.

>daily_idx_chg <- read.csv(daily_idx_chg.csv)

>E3_6 <- daily_idx_chg

>summary(E3_6)

>cov(E3_6$PCT.DOW.CHG, E3_6$PCT.SP.CHG)

Odpowiedź: Dwie nazwy zmiennych to PCT. DOW. CHG oraz PCT. SP. CHG; Wartości danych

wydają się być wyśrodkowane wokół 0 z wartościami w zakresie od około 1,55 do -1,71

Kowariancja wynosi 0,7693505, co mówi nam tylko, że te dwie zmienne są dodatnio powiązane.

58. Ustandaryzuj dzienne dane idx chg i ponownie oblicz kowariancję. Czy to jest to samo?

>std_indices <- scale(E3_6)

>cov(std_indices)

Odpowiedź: Kowariancja wynosi 0,9573543. Nie, kowariancja nie jest taka sama, nawet

chociaż został on zastosowany do tych samych danych. W rzeczywistości kowariancja na surowych danych nie jest (ogólnie) równa kowariancji dla tych samych danych po standaryzacji.

59. Znajdź korelację tych dwóch zmiennych w daily_ idx chg.csv

>cor(E3_6)

Odpowiedź: Korelacja wynosi 0,9573543, dokładnie taka sama jak kowariancja. Zmienne standaryzowane. Ogólnie rzecz biorąc, korelacja dwóch niestandaryzowanych zmiennych równa się kowariancji tych samych dwóch zmiennych w postaci ustandaryzowanej.

60. Ustandaryzuj daily_ idx chg.csv i ponownie oblicz korelację. Czy to jest to samo?

>cor(std_indices)

Odpowiedź: Korelacja między wartościami znormalizowanymi jest dokładnie taka sama jak

Korelacja między wartościami niestandardowymi: 0,9573543.

61. Zrób wykres punktowy daily_ idx chg.csv za pomocą PCT. DOW. CHG w pozycji poziomej osi, PCT. SP. CHG w pionie. Dodaj główny tytuł i etykiety dla poziomego i osie pionowe. Czy wykres punktowy jest zgodny z sugerowanym dodatnim asocjacją liniową przez współczynnik korelacji?

>plot(E3_6$PCT.DOW.CHG, E3_6$PCT.SP.CHG,

xlab = Percentage Daily Change in the Dow,

ylab = Percentage Daily Change in the S&P500,

pch = 19,

col = purple,

main = A Plot of Daily Percent Changes in the Dow and S&P500)

Odpowiedź: Wykres punktowy jest zgodny ze współczynnikiem korelacji wynoszącym 0,9573543.

Istnieje silnie pozytywna liniowa zależność między tymi dwoma indeksami giełdowymi.

62. Poniżej mamy zależność krzywoliniową, w której punkty mogą być połączone

gładka, paraboliczna krzywą. Zobacz wykres punktowy. Jaka jest najbardziej prawdopodobna korelacja opisujący ten związek? -0,90, -0,50, -0,10, 0,00, +0,10, +0,50 czy

+0.90.

>x <- c(0, -1, -2, -3, -4)

>y <- c(4, 2, 1, 2, 4)

>data <- data.frame(X = x, Y = y)

>plot(data$X, data$Y, pch = 19, xlab = x, ylab = y)

Współczynnik korelacji wynosi 0.

63. Jaki jest najbardziej prawdopodobny współczynnik korelacji opisujący poniższą zależność? Zobacz wykres punktowy. -0,90, -0,50, -0,10, 0,00, +0,10, +0,50 lub +0,90.

x <- c(16, 13, 8, 6, 5)

y <- c(15, 20, 25, 25, 30)

data <- data.frame(X = x, Y = y)

plot(data$X, data$Y, pch = 19, xlab = x, ylab = y)

Odpowiedź: -0,90 to najbliższa wartość, jaką może przyjąć współczynnik korelacji:

Zależność między tymi dwiema zmiennymi jest nie tylko ujemna, ale także liniowa. W

Faktycznie, współczynnik korelacji wynosi -0,9657823.

>cor(data$X, data$Y)

64. Jaki jest najbardziej prawdopodobny współczynnik korelacji opisujący poniższą zależność?

-0,90, -0,50, -0,10, 0,00, +0,10, +0,50 czy +0,90?

```
x <- c(24, 22, 22, 21, 19)

y <- c(27, 24, 23, 21, 19)

data <- data.frame(X = x, Y = y)

plot(data$X, data$Y, pch = 19, xlab = x, ylab = y)
```

Odpowiedź: +0,90 to najbliższa wartość, jaką może przyjąć współczynnik korelacji:

Zależność między tymi dwiema zmiennymi jest nie tylko pozytywna, ale także liniowa.

```
>cor(data$X, data$Y)
```

65. Jaki jest najbardziej prawdopodobny współczynnik korelacji opisujący poniższą zależność? -0,90, -0,50, -0,10, 0,00, +0,10, +0,50 lub +0,90.

```
x <- c(0, -30, -30, -30, -60)

y <- c(-20, 10, -20, -50, -20)

data <- data.frame(X = x, Y = y)

plot(data$X, data$Y, pch = 19, xlab = x, ylab = y)
```

Odpowiedź: W rzeczywistości współczynnik korelacji wynosi 0,00.

```
>cor(data$X, data$Y)
```

66. Reguła empiryczna mówi, że około 68:27% wartości w rozkładzie normalnym przypada

w przedziale od 1 odchylenia standardowego poniżej średniej do 1 odchylenia standardowego powyżej średniej. Zweryfikuj to twierdzenie, poprzez: (a) wygenerowanie n = 1; 000; 000 wartości o rozkładzie normalnym ze średnią 100 i standardowym odchyleniu 15, a następnie (b) zlicz " liczbę wartości danych, które mieszczą się w tym interwałle Jeśli prawda, to w przybliżeniu (0.6827) * (1 000 000) = 682 700 wartości powinno być w przedziale od 85 do 115; w przybliżeniu (0-1587)*(1 000 000) = 158 700 poniżej 85; i w przybliżeniu (01587)*(1 000 000) = 158 700 wartości powyżej 115. Użyj rnorm() do generowania danych.

```
>normal_data <- rnorm(1000000, 100, 15)
```

>a <- length(which(normal_data <= 85))

>a

>b <- length(which(normal_data >= 115))

>b

>c <- (1000000 - (a + b)) / 1000000

>c

Korzystając ze zdolności generowania danych R, możemy stwierdzić, że proporcja

Wartości danych przypadające w przedziale od 1 odchylenia standardowego poniżej średniej do 1 Odchylenie standardowe powyżej średniej wynosi w przybliżeniu 0.6827.

67. Reguła empiryczna mówi nam również, że około 95.45% wartości zawiera się w przedziale od 2 odchyleń standardowych poniżej średniej do 2 odchylenia standardowe powyżej średniej. Jeśli prawda, to w przybliżeniu (0.9545)*(1 000 000) = 954 500 wartości powinno przypadać w przedziale od 70 do 130; w przybliżeniu (0.02275)*(1 000 000) = 22 750 poniżej 70; i w przybliżeniu (0.02775)*(1 000 000) = 22 750 powyżej 130. Użyj normalnych danych z ćwiczenia 23, aby odpowiedzieć na te pytania.

>a <- length(which(normal_data <= 70))

>a

>b <- length(which(normal_data >= 130))

>b

>c <- (1000000 - (a + b)) / 1000000

>c

Proporcja wartości danych przypadających w przedziale od 2 odchyleń standardowych wynosi około 0.9545.

68. Użyj funkcji runif() (patrz poniżej), aby wygenerować n = 1 000 000 wartości, które mają

Jednostajny rozkład biegnący od A = 75 do B = 125. Przypisz wartości danych do wektora. Nazwij to Uniform Data.

>uniform_data <- runif(1000000, 75, 125)

>hist(uniform_data,

breaks = 50,

xlim = c(70, 130),

ylim = c(0, 25000),

col = blue)

Jaka jest proporcja wartości przypadających w przedziale od 90 do 110?

Odpowiedź: Z tego ćwiczenia symulacyjnego widzimy, że proporcja jednolicie rozproszone wartości danych (z zakresu od 75 do 125), które mieszczą się w przedziale od 90

do 110 wynosi (w przybliżeniu) 0.40.

>a <- length(which(uniform_data <= 90))

>a

>b <- length(which(uniform_data >= 110))

>b

>c <- (1000000 - (a + b)) / 1000000

>c

Z histogramu widzimy, że zmienna o jednostajnym rozkładzie przyjmuje kształt

prostokąta, a zatem proporcja wartości danych przypadających w dowolnym przedziale wynosi wprost proporcjonalne do długości tego interwału. W niniejszej sprawie, ponieważ pytanie dotyczy proporcji wartości danych w przedziale szerokości 20 (= 110 - 90) dla

rozkład o szerokości 50 (= 125 - 75), proporcja wartości danych przypadających na

Interwał od 90 do 110 wynosi 20/50 lub 0.40.