

Programowanie w języku R. Dane jednowymiarowe

Aleksander Denisiuk
Uniwersytet Warmińsko-Mazurski
Olsztyn, ul. Słoneczna 54
denisjuk@matman.uwm.edu.pl

20 marca 2025

Dane jednowymiarowe

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

Najnowsza wersja tego dokumentu dostępna jest pod adresem

<http://wmii.uwm.edu.pl/~denisjuk/uwm>

Charakterystyki próbki

- ❖ Charakterystyki
- ❖ Średnia
- ❖ Odchylenie
- ❖ boxplot
- ❖ histogram
- ❖ Inne metody
- ❖ Współczynnik wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

Charakterystyki próbki

Charakterystyki

Charakterystyki
próbki

❖ Charakterystyki

❖ Średnia
❖ Odchylenie
❖ boxplot
❖ histogram
❖ Inne metody
❖ Współczynnik
wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

- Dwie podstawowe charakterystyki
- Ogólna tendencja
 - ◆ średnia
 - ◆ mediana
- Rozrzut danych
 - ◆ odchylenie standardowe
 - ◆ kwartyle

Obliczenie średniej wartości

Charakterystyki
próbki

❖ Charakterystyki

❖ Średnia

❖ Odchylenie

❖ boxplot

❖ histogram

❖ Inne metody

❖ Współczynnik
wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

● Średnia i mediana

```
salary <- c(21, 19, 27, 11, 102, 25, 21)
mean(salary); median(salary)
```

● Mediana jest bardziej stabilna

```
a1 <- c(1, 2, 3, 4, 4, 5, 7, 7, 7, 9, 15, 17)
a2 <- c(1, 2, 3, 4, 5, 7, 7, 7, 9, 15, 17)
median(a1)
median(a2)
```

● Kwartyle

```
quantile(salary)
```

Moda

Charakterystyki
próbki

❖ Charakterystyki

❖ Średnia

❖ Odchylenie

❖ boxplot

❖ histogram

❖ Inne metody

❖ Współczynnik
wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

- Najczęściej występująca wartość

```
sex <- c("male", "female", "male", "male",  
         "female", "male", "male")
```

```
t.sex <- table(sex)
```

```
t.sex
```

```
mode <- t.sex[which.max(t.sex)]
```

```
mode
```

- `table()` zlicza ilości

Obliczenie dla całej tablicy

Charakterystyki
próbki

❖ Charakterystyki

❖ Średnia

❖ Odchylenie

❖ boxplot

❖ histogram

❖ Inne metody

❖ Współczynnik
wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

- Dołączyć kolumny ramki do bieżących zmiennych

```
attach(trees)
mean(Girth)
mean(Height)
mean(Volume/Height)
detach(trees)
```

- Funkcja `with()`

```
with(trees, mean(Volume/Height))
```

- Funkcja `lapply()` (dla listy kolumn)

```
lapply(trees, mean)
```

- Pętla `for` nie jest zalecana

Odchylenie

Charakterystyki próbki

❖ Charakterystyki

❖ Średnia

❖ **Odchylenie**

❖ boxplot

❖ histogram

❖ Inne metody

❖ Współczynnik wariacji

Błędne dane

Testy statystyczne

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

● Charakterystyki parametryczne

◆ odchylenie standardowe

```
sd(salary)
```

◆ wariancja

```
var(salary)
```

● IQR

```
IQR(salary)
```


Porównanie średnich

Charakterystyki próbek

❖ Charakterystyki

❖ Średnia

❖ Odchylenie

❖ boxplot

❖ histogram

❖ Inne metody

❖ Współczynnik wariacji

Błędne dane

Testy statystyczne

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

- Dla danych `trees` średnia oraz odchylenie standardowe i mediana oraz IQR są zbliżone:

```
attach(trees)
```

```
mean(Height)
```

```
median(Height)
```

```
sd(Height)
```

```
IQR(Height)
```

```
detach(trees)
```

- ◆ rozkład normalny?
- ◆ Uwaga: zmiana dołączonych kolumn nie ma wpływu na pierwotną ramkę

Wykres *boxplot*

Charakterystyki próbki

❖ Charakterystyki

❖ Średnia

❖ Odchylenie

❖ **boxplot**

❖ histogram

❖ Inne metody

❖ Współczynnik wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

- Czy są dane *mocno odchylające się*?
- Metoda graficzna: wykres `boxplot`
- ◆ dodajmy jeszcze 1000 pracowników

```
new.1000 <- sample(  
  (median(salary) - IQR(salary)) :  
  (median(salary) + IQR(salary)),  
  1000, replace=TRUE)  
salary2 <- c(salary, new.1000)  
boxplot(salary2, log="y")
```

- ◆ `boxplot` jest funkcja wektorowa
`boxplot(trees)`

Wykres histogram

Charakterystyki próbki

- ❖ Charakterystyki
- ❖ Średnia
- ❖ Odchylenie
- ❖ boxplot
- ❖ **histogram**
- ❖ Inne metody
- ❖ Współczynnik wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

- Inna metoda: histogram

```
hist(salary2, breaks=20)
```

- Funkcja `cut()` dzieli cały zakres zmiennej na przedziały i zastępuje dane przez odpowiednie przedziały

```
cut(salary2, 20)  
table(cut(salary2, 20))
```

- Histogram tekstowy:

```
stem(salary, scale=2)
```

- Wykres dystrybuanty

```
plot(density(salary2, adjust=2))  
rug(salary2)
```

◆ `rug()` zaznacza miejsca o większej gęstości

Inne metody

Charakterystyki próbki

- ❖ Charakterystyki
- ❖ Średnia
- ❖ Odchylenie
- ❖ boxplot
- ❖ histogram

❖ Inne metody

- ❖ Współczynnik wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

- Wykres `beeswarm()`

```
library("beeswarm")  
beeswarm(trees)  
boxplot(trees, add=TRUE)
```

- Funkcja `summary()`

```
lapply(list(salary, salary2), summary)
```

- Funkcja `summary()` jest uniwersalna

- ✦ trzęsienia ziemi w Kalifornii

```
summary(attenu)
```

Współczynnik wariancji

Charakterystyki próbek

- ❖ Charakterystyki
- ❖ Średnia
- ❖ Odchylenie
- ❖ boxplot
- ❖ histogram
- ❖ Inne metody
- ❖ Współczynnik wariancji

Błędne dane

Testy statystyczne

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

- Parameter względny: $\frac{\sigma}{m} \times 100\%$
 $100 * \text{apply}(\text{trees}, \text{sd}) / \text{colMeans}(\text{trees})$
 - ◆ funkcja `apply()` — jak `lapply()`, wynik: wektor
 - ◆ `colMeans()`, `rowSums()`, etc

Charakterystyki
próbki

Błędne dane

❖ Pomyłki

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

Błędne dane

Sprawdzanie jakości danych

Charakterystyki
próbki

Błędne dane

❖ Pomyłki

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

- Funkcje `str()` oraz `summary()`

```
err <- read.table("data/errors.txt",  
                  h=TRUE, sep="\t")
```

```
str(err)
```

```
summary(err)
```

- Znajdź trzy pomyłki w danych

Charakterystyki
próbki

Błędne dane

Testy statystyczne

❖ Hipotezy

❖ Błędy

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

Testy statystyczne

Hipotezy statystyczne

Charakterystyki próbek

Błędne dane

Testy statystyczne

❖ Hipotezy

❖ Błędy

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

- Stwierdzenie o populacji generalnej na podstawie próbki
- Dwie hipotezy:
 - ◆ H_0 — hipoteza zerowa
 - brak różnicy
 - ◆ H_1 — hipoteza alternatywna, zaprzeczenie H_0
 - różnica istotna
- Nie można udowodnić H_0 , można
 1. przyjąć H_0
 2. odrzucić H_0 , przyjmując H_1

Błędy statystyczne

Charakterystyki próbek

Błędne dane

Testy statystyczne

❖ Hipotezy

❖ Błędy

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

Próba \ Populacja	Prawidłowa H_0	Prawidłowa H_1
Przyjmujemy H_0	Poprawnie	Błąd II rodzaju
Przyjmujemy H_1	Błąd I rodzaju	Poprawnie

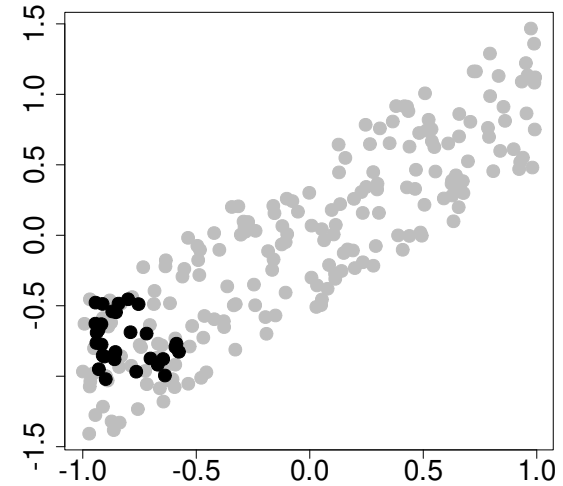
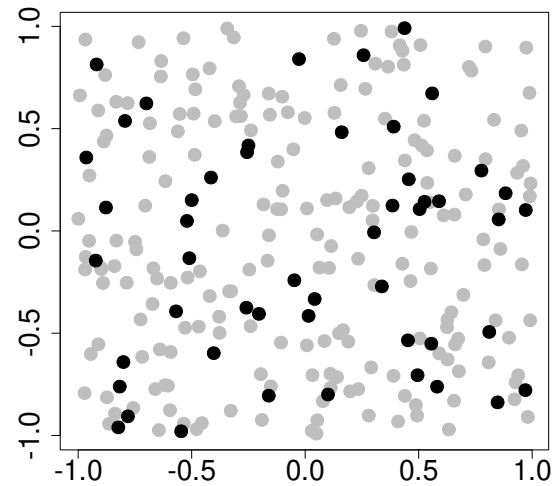
- Prawdopodobieństwo błędu pierwszego rodzaju to jest p -wartość
 - ◆ *poziom ufności testu* to wartość krytyczna, powyżej której trzeba H_0 odrzucić
 - zazwyczaj $\alpha = 0,05$ (0,01)
- *moc testu* to prawdopodobieństwo niepopelnienia błędu drugiego rodzaju: $1 - \beta$.

Błędy statystyczne

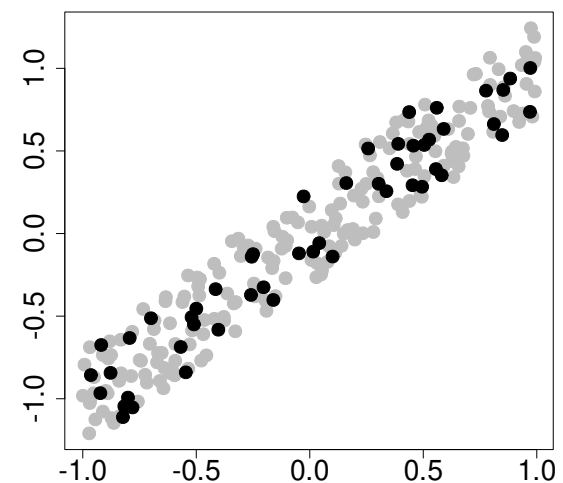
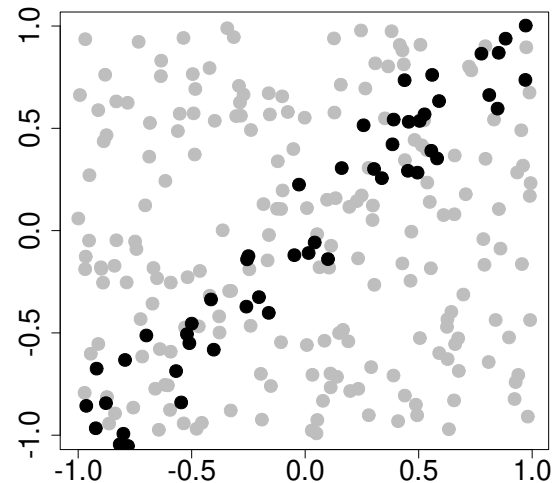
H_0

H_1

H_0



H_1



Charakterystyki próbek

Błędne dane

Testy statystyczne

❖ Hipotezy

❖ Błędy

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

❖ t-test

❖ Wilcoxon

❖ normalność
rozkładu

Własne funkcje

Test proporcji

Jednowymiarowe testy statystyczne

Jednowymiarowy *t*-test Studenta

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

❖ *t*-test

❖ Wilcoxon

❖ normalność
rozkładu

Własne funkcje

Test proporcji

- Hipoteza zerowa H_0 : średnia populacji generalnej zgadza się ze średnią próbki

◆ Hipoteza alternatywna: średnia populacji generalnej nie zgadza się ze średnią próbki

```
t.test(salary, mu=mean(salary))
```

- p — prawdopodobieństwo błędu pierwszego rodzaju (odrzućenie prawidłowej hipotezy zerowej)
- Test parametryczny

Test Wilcoxona

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

❖ t-test

❖ Wilcoxon

❖ normalność
rozkładu

Własne funkcje

Test proporcji

- Test nieparametryczny
- Hipoteza zerowa H_0 : mediana populacji generalnej zgadza się z medianą próbki
 - ◆ Hipoteza alternatywna: mediana populacji generalnej nie zgadza się z medianą próbki

```
wilcox.test(salary, mu=median(salary),  
            conf.int=TRUE)
```

Testy normalności rozkładu

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

❖ t-test

❖ Wilcoxon

❖ normalność
rozkładu

Własne funkcje

Test proporcji

- Test Shapiro-Wilka

```
shapiro.test(salary)
shapiro.test(salary2)
```

- Metoda graficzna

```
qqnorm(salary2, main=" ")
qqline(salary2, col=2)
```

- Test Kolmogorova-Smirnova

```
ks.test(salary2, "pnorm")
```

- ◆ porównuje dwa rozkłady

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

❖ Wekrotyzacja

Test proporcji

Własne funkcje

Wektoryzacja testu Shapiro-Wilka

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

❖ Wektoryzacja

Test proporcji

```
normality1 <- function(data.f) {  
  result <- data.frame(  
    var=names(data.f),  
    p.value = rep(0, ncol(data.f)),  
    normality = rep(FALSE, ncol(data.f))  
  )  
  for (i in 1:ncol(data.f)) {  
    data.sh <- shapiro.test(data.f[, i])$p.value  
    result[i, 2] <- round(data.sh, 5)  
    result[i, 3] <- (data.sh > .05)  
  }  
  return(result)  
}  
normality1(trees)
```

- Może być plik wewnętrzny, przykładowo `normality1.r`
- Załadować poleceniem `source()`

Optymalizacja funkcji

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

❖ Wekrotyzacja

Test proporcji

```
normality2 <- function(data.f, p=.05) {  
  nn <- ncol(data.f)  
  result <- data.frame(  
    var = names(data.f),  
    p.value = numeric(nn),  
    normality = logical(nn))  
  for (i in 1:nn) {  
    data.sh <- shapiro.test(data.f[, i])$p.value  
    result[i, 2:3] <- list(round(data.sh, 5),  
                          data.sh > p)  
  }  
  return(result)  
}  
normality2(trees)
```

Bez pętli *for*

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

❖ Wekrotyzacja

Test proporcji

```
lapply(trees, shapiro.test)
```

- albo tak:

```
lapply(trees,  
       function(.x) {  
         ifelse(shapiro.test(.x)$p.value > .05,  
                "NORMAL", "NOT NORMAL")  
       })
```

- ◆ funkcja *anonimowa*

Nowa funkcja

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

❖ Wekrotyzacja

Test proporcji

```
normality3 <- function(df, p=.05) {  
  lapply(df,  
    function(.x) {  
      ifelse( shapiro.test(.x)$p.value > p,  
              "NORMAL", "NOT NORMAL")  
    }  
  )  
}  
  
normality3(list(salary, salary2))  
normality3(log(trees+1))
```

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

- ❖ Test dwumianowy
- ❖ Test proporcji
- ❖ Dwie próby

Test proporcji

Test dwumianowy

Charakterystyki próbek

Błędne dane

Testy statystyczne

Jednowymiarowe testy statystyczne

Własne funkcje

Test proporcji

❖ Test dwumianowy

❖ Test proporcji

❖ Dwie próby

- Czy częstotliwość wystąpienia zjawiska w próbie istotnie różni się od częstotliwości populacji generalnej
 - ◆ Przykładowo: na WMII 120 studentów spośród 476 nie zaliczyło sesji letniej w pierwszym terminie. Średnio na UWM liczba ta jest 30%. Czy liczba niezaliczających na WMII statystycznie się różni od ogólnej?

```
binom.test(x=120, n=476, p=0.3,  
           alternative="two.sided")
```

- Może być `alternative="greater"` lub `alternative="less"`

Test proporcji

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

❖ Test dwumianowy

❖ Test proporcji

❖ Dwie próby

```
prop.test(x=120, n=476, p=0.3,  
          alternative="two.sided")
```

- Także może być `alternative="greater"` lub `alternative="less"`

Testy dla dwóch prób

Charakterystyki
próbki

Błędne dane

Testy statystyczne

Jednowymiarowe
testy statystyczne

Własne funkcje

Test proporcji

❖ Test dwumianowy

❖ Test proporcji

❖ Dwie próby

- Dla prób niezależnych test proporcji: `prop.test()`
- Dla prób zależnych test McNemara: `mcnemar.test()`