Sprawozdanie 1

Jakub Markowiak
album 255705

14 maja 2021

Spis treści

1 Krótki opis zagadnienia

W tym sprawozdaniu sprawdzimy zastosowanie podstawowych statystyk opisowych oraz spróbujemy na ich podstawie zinterpretować dane `survey` z pakietu `MASS`. Wykonamy tabele liczności i częstości oraz tabele wielodzielcze dla cech jakościowych. Zbadamy również odporność miar położenia i rozproszenia na obserwacje odstające oraz poznamy definicję i zastosowanie współczynnika zmienności.

2 Opis eksperymentów/analiz

Przeprowadzimy następujące analizy i eksperymenty:

1. analiza opisowa danych `survey` z pakietu `MASS` dla cech ilościowych,
2. analiza opisowa danych `survey` z pakietu `MASS` dla cech jakościowych,
3. badanie odporności miar położenia i rozproszenia na obserwacje odstające,
4. zdefiniowanie i zastosowanie współczynnika zmienności w porównaniu dwóch cech.

3 Wyniki

3.1 Analiza opisowa danych survey z pakietu MASS dla cech ilościowych

Rozpoczniemy od wczytania danych `survey` z pakietu `MASS` oraz zbadania liczby obserwacji, liczby cech jakościowych i ilościowych oraz sprawdzenia, czy w danych występują brakujące wartości.

Używając kilku podstawowych funkcji (`ncol, nrow, sapply`) uzyskujemy następujące informacje:

```
## Warning in is.na(df): 'is.na()' zastosowane do nie-listy lub nie-wektora typu
'closure'
```

Liczba cech	Liczba obserwacji	Cechy ilościowe	Cechy jakościowe	Brakujące wartości
12	237	5	7	0

Tabela 1: Wstępne spojrzenie na dane

Następnie wyznaczymy podstawowe statystyki opisowe dla cech `Height` oraz `Age`. Zdefiniujemy pomocniczą funkcję `moda` obliczającą modę z próbki oraz funkcję `my.summary` działającą analogicznie do domyślnego `summary`.

```
moda <- function(v) {
  uniqv <- unique(na.omit(v))
  uniqv[which.max(tabulate(match(na.omit(v), uniqv)))]
}

my.summary <- function(x)
{
  wskazniki <- c(Srednia=mean(x,na.rm=T),
                 Mediana=median(x,na.rm=T),
                 IQR=IQR(x,na.rm=T),
                 Min=min(x,na.rm=T),
                 Maks=max(x,na.rm=T),
                 Odch.stand.=sd(x,na.rm=T),
                 Rozstep=max(x,na.rm=T)-min(x,na.rm=T),
                 Moda=moda(x))
  return(wskazniki)
}
```

Wykorzystując funkcję `my.summary` otrzymujemy podstawowe wskaźniki sumaryczne.

Spoglądając na średnią i odstęp międzykwartylowy widzimy, że zdecydowana większość badanych to młodzi ludzie. Większe zróżnicowanie jest natomiast widoczne we wzroście.

Wyznaczymy teraz przedział typowych wartości dla badanych cech. W tym celu korzystamy ze wzoru

$$\left[\bar{X} - S, \bar{X} + S\right] \tag{1}$$

Tabela 2: Podstawowe wskazniki sumaryczne dla Height i Age

	Srednia	Mediana	IQR	Min	Maks	Odch.stand.	Rozstep	Moda
Height	172.38	171.00	15.0	150.00	200	9.85	50.00	165.0
Age	20.37	18.58	2.5	16.75	73	6.47	56.25	17.5

gdzie $\bar{X}$ - średnia próbkowa, S - odchylenie standardowe. Korzystając z wcześniejszych obliczeń otrzymujemy

	X-S	X+S	Obs. w przedziale	% wszystkich
Age	13.900	26.849	220	92.827
Height	162.533	182.228	143	60.338

Tabela 3: Przedział typowych wartości

Ustalamy teraz zmienną `Sex` jako zmienną grupującą. Policzymy wartości średnie oraz rozrzut cech `Height` oraz `Age` w zależności od przynależności do grupy.

Tabela 4: Średnia, odchylenie standardowe oraz IQR dla Age

Sex	Średnia	Odch. stand.	IQR
Female	20.40753	6.906053	2.47925
Male	20.33196	6.069863	2.37450

Tabela 5: Średnia, odchylenie standardowe oraz IQR dla Height

Sex	Średnia	Odch. stand.	IQR
Female	165.6867	6.151777	7.44
Male	178.8260	8.380252	12.21

Możemy zauważyć, że badane osoby były w podobnym wieku, natomiast znaczne różnice występują przy średnim wzroście – u kobiet jest znacznie mniejszy niż u mężczyzn.

3.2 Analiza opisowa danych survey z pakietu MASS dla cech jakościowych

Wybierzmy teraz dwie cechy jakościowe – `W.Hnd` oraz `Clap`. Poniżej znajduje się tabela liczności i częstości dla `W.Hnd`.

Widzimy, że około 92,37% badanych osób to osoby praworęczne, a tylko 18 spośród wszystkich badanych deklaruje leworęczność. Przygotujemy analogiczną tabelę dla `Clap`.

Możemy z niej odczytać, że około 62,29% badanych deklaruje, że podczas klaskania „przeważa" prawa ręka. Sprawdzimy zatem, czy występują jakieś zależności między cechami `W.Hnd` oraz `Clap`. W tym celu przygotujemy tabelę wielodzielczą dla tych zmiennych.

Widzimy, że 60,85% ankietowanych deklaruje prawą rękę jako tą dominującą i „klaszczącą", natomiast zaledwie 3,82% deklaruje taką kombinację dla lewej ręki. Można wstępnie dostrzec

Tabela 6: Tabela liczności i częstości W.Hnd

W.Hnd	Liczność	Częstość
Left	18	0.0762712
Right	218	0.9237288

Tabela 7: Tabela liczności i częstości Clap

Clap	Liczność	Częstość
Left	39	0.1652542
Neither	50	0.2118644
Right	147	0.6228814

zależność, że osoby praworęczne częściej używają prawej dłoni do klaskania, natomiast osoby leworęczne lewej dłoni.

Analogiczną tabelę przygotujemy dla zmiennych `Sex` oraz `Exer`.

Widzimy, że odsetek osób niećwiczących w obu grupach jest zbliżony i wynosi około 5%, natomiast „systematyczne" ćwiczenia deklaruje o około 7 pkt. procentowych więcej mężczyzn niż kobiet. Nie widać zatem zależności między płcią a wykonywaniem jakichkolwiek ćwiczeń, ale mężczyźni przeważają wśród osób ćwiczących „często", a kobiety wśród osób ćwiczących „trochę".

3.3 Badanie odporności miar położenia i rozproszenia na obserwacje odstające

Zajmiemy się teraz sprawdzeniem, jak odpowiednie miary położenia i rozproszenia zachowują się, gdy w danych występują obserwacje odstające. Najpierw wczytujemy dane do `R`, zapisując je jako wektor `t`. Definiujemy funkcję `sr.ucinana`, obliczającą średnią ucinaną oraz funkcję `my.summary.cut`, która wyświetla w tabeli wybrane wskaźniki rozproszenia i położenia.

```
#Wczytanie danych do R
t <- c(6.5, 5, 6, 4, 7, 7, 5.5, 7.5)
#Definicja funkcji obliczającej średnią ucinaną
sr.ucinana <- function(x,k=1)
{
  x. <- sort(x)
  len.x <- length(x)
  if(k <= len.x/2){
    return(mean(x.[k:(len.x - k)]))
  } else{
    return("Za duża wartość k.")
  }
}
#Definicja funkcji wyświetlającą odpowiednie wskaźniki w tabeli
my.summary.cut <- function(x)
{
  wskazniki <- c(Średnia=mean(x,na.rm=T), Śr.ucinana=sr.ucinana(x),
```

Tabela 8: Tabela wielodzielcza W.Hnd i Clap

W.Hnd	Clap		
	Left	Neither	Right
Left	0.0382979	0.0212766	0.0170213
Right	0.1234043	0.1914894	0.6085106

Tabela 9: Tabela wielodzielcza Sex i Exer

Sex	Exer		
	Freq	None	Some
Female	0.2076271	0.0466102	0.2457627
Male	0.2754237	0.0550847	0.1694915

```
                    IQR=IQR(x,na.rm=T), Mediana=median(x,na.rm=T),
                    Odch.stand.=sd(x,na.rm=T), Wariancja=var(x),
                    Rozstęp=max(x,na.rm=T)-min(x,na.rm=T))
  return(t(wskazniki))
}
```

Używamy teraz wyżej zdefiniowanych funkcji i otrzymujemy następujące wyniki.

Tabela 10: Podstawowe wskaźniki dla t

Średnia	Śr.ucinana	IQR	Mediana	Odch.stand.	Wariancja	Rozstęp
6.0625	5.857143	1.625	6.25	1.178301	1.388393	3.5

Zdefiniujemy wektor t', zamieniając w wektorze t wartość 7.5 na 10.

Wnioskujemy stąd, że najbardziej odporne na odchylenia wskaźniki to `średnia ucinana`, `rozstęp międzykwartylowy` oraz `mediana`. Niską odporność wykazują m.in. `średnia`, `rozstęp` i `wariancja`.

3.4 Zdefiniowanie i zastosowanie współczynnika zmienności w porównaniu dwóch cech

Współczynnik zmienności jest definiowany jako

$$V = \frac{S}{\bar{X}} \cdot 100\%, \qquad (\text{gdy} \bar{X} \neq 0), \tag{2}$$

gdzie S to odchylenie standardowe z próby, a $\bar{X}$ to średnia próbkowa.

Zdefiniujemy funkcję `CV`, która dla wektora $X = (X_1, X_2, \ldots, X_n)$ wylicza współczynnik zmienności.

```
cv <- function(x)
{
  S <- sd(x)
```

Tabela 11: Podstawowe wskaźniki dla t'

Średnia	Śr.ucinana	IQR	Mediana	Odch.stand.	Wariancja	Rozstęp
6.375	5.857143	1.625	6.25	1.787856	3.196429	6

```
  X. <- mean(x)
  if(X. == 0){
    return("Błąd. Średnia próbkowa jest równa zero.")
  } else {
    return(S/X.*100)
  }
}
```

Następnie wprwadzamy do `R` dane przedstawiające wzrost oraz wagę w pewnej grupie uczniów.

```
dane <- data.frame(wzrost = c(151, 160, 162, 155, 154,
                              168, 153, 158, 157, 150, 167),
                   waga = c(61, 69, 73, 65, 64, 78,
                            63, 68, 67, 60, 77))
```

Wykorzystując funkcję `CV` oraz `sd` obliczamy współczynnik zmienności oraz odchylenie standardowe dla cech `waga` oraz `wzrost`.

Tabela 12: CV oraz sd dla wzrostu i wagi

CV.wzrost	sd.wzrost	CV.waga	sd.waga
3.826065	6.034748	8.983467	6.084257

Ponieważ współczynnik zmienności zmiennej `waga` jest wyższy, możemy stwierdzić, że charakteryzuje się ona większą zmiennością niż `wzrost`. Natomiast różnica odchyleń standardowych obu zmiennych jest zbyt mała, aby spoglądając tylko na nią uzyskać taki sam wniosek.

4 Podsumowanie

Poniżej wypunktujemy najważniejsze wnioski, jakie można wyciągnąć z przeprowadzanych analiz:

- wyznaczenie podstawowych wskaźników sumarycznych przy podziale na grupy pozwala nam dostrzec różnice w tych grupach i wysnuć wstępne wnioski,
- wykononanie tabeli wielodzielczej dla dwóch zmiennych jakościowych pozwala nam na porównanie częstości występowania pewnych kombinacji oraz zaobserwowanie ewentualnych zależności między zmiennymi,
- średnia ucinana, rozstęp międzykwartylowy oraz mediana to wskaźniki odporne na obserwacje odstające, natomiast średnia próbkowa, rozstęp oraz wariancja są na nie bardzo podatne,

- współczynnik zmienności pozwala lepiej określić zmienność cechy niż odchylenie standardowe.