



Rzeczpospolita
Polska

Politechnika
Warszawska

Unia Europejska
Europejski Fundusz Społeczny



Rachunek Prawdopodobieństwa i Elementy Statystyki Matematycznej

Anna Dembińska

Wydział Matematyki i Nauk Informatycznych

Wykład 5

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”
współfinansowany jest ze środków Unii Europejskiej w ramach
Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach
prowadzonych przez Wydział Matematyki i Nauk Informatycznych”,
realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja –
Rozwój – Współpraca”, współfinansowanego ze środków Unii
Europejskiej w ramach Europejskiego Funduszu Społecznego.

5. ANALIZA DANYCH A WNIOSKOWANIE STATYSTYCZNE

Statystyka obejmuje dwa nurty:

1. analizę danych,
2. wnioskowanie statystyczne.

Celem **analizy danych** jest prezentacja konkretnego zbioru danych, w sposób ukazujący jego własności; w szczególności syntetyczny opis podstawowych jego cech. Otrzymujemy wówczas wnioski, które dotyczą **wyłącznie analizowanego zbioru danych**. Na przykład mamy zebrane informacje na temat studentów studiów 1-go stopnia na kierunku Informatyka na Wydziale MiNI PW, którzy rozpoczęli owe studia w 2017 roku. Dokładniej, mamy listę tych studentów wraz z następującymi danymi:

- data urodzenia,
- płeć,
- czy student obronił pracę inżynierską na MiNI.

Na podstawie tych danych możemy stwierdzić np.

- jaki procent studentów Informatyki, rozpoczynających studia 1-go stopnia na Wydziale MiNI PW w 2017 r., obronił na tym wydziale pracę inżynierską;
- jaki procent owych studentów stanowiły kobiety;
- jaki był średni wiek owych studentów w momencie, gdy rozpoczynali studia na Informatyce na MiNI?.

Otrzymamy wyniki **dokładne i pewne**, ale dotyczyć będą one **jedynie** studentów 1-go stopnia Informatyki na Wydziale MiNI, którzy rozpoczęli te studia w 2017 roku.

Teraz wyobraźmy sobie, że chcemy wiedzieć:

- jaki procent studentów, rozpoczynających studia 1-go stopnia w Polsce na kierunku Informatyka, kończy te studia uzyskaniem dyplomu inżyniera;
- jaki procent studentów, rozpoczynających studia 1-go stopnia w Polsce na kierunku Informatyka, to kobiety;
- ile wynosi średni wiek osób, które rozpoczynają studia 1-go stopnia w Polsce na kierunku Informatyka.

Aby uzyskać dokładną i pewną odpowiedź na powyższe pytania, potrzebowalibyśmy zebrać dane dotyczące wszystkich osób, które rozpoczęły lub rozpoczną w Polsce studia 1-go stopnia na kierunku Informatyka. Jest to zadanie niewykonalne - teraz nie zdobędziemy danych dotyczących przyszłych studentów. Ponadto, nawet gdybyśmy zdecydowali się nasze pytania ograniczyć do studentów dotychczasowych, to zebranie odpowiednich danych byłoby trudne - czasochłonne i kosztowne, a nawet nadal niekoniecznie wykonalne, bo niektóre uczelnie mogą odmówić nam współpracy lub zwlekać z dostarczeniem stosownych danych. Pozostaje wtedy pójść na kompromis - zebrać dane dotyczące tylko wybranych studentów i na ich podstawie wyciągać wnioski o wszystkich studentach. Mamy wtedy do czynienia z **wnioskowaniem statystycznym**. Musimy w nim zwrócić uwagę na dwa aspekty.

1. Bardzo ważny jest odpowiedni wybór studentów do naszego badania - ogólniej - **odpowiedni wybór obserwacji do próby**, tak by dobrze reprezentowały one całą populację.
2. Jeśli tylko jako próby nie weźmiemy całej populacji (a tak we wnioskowaniu statystycznym postępujemy), to **uzyskane wyniki nie będą ani dokładne, ani pewne - pozostaną obarczone błędem**.

Analiza danych

Jak już wspomnieliśmy, celem analizy danych jest opis podstawowych cech konkretnego zbioru danych. Często, aby taki opis uzyskać, musimy najpierw dane, zawarte w zbiorze, uporządkować i uprościć. Porządkowanie danych rozpoczynamy od ustalenia jakiego są one typu. Możemy mieć:

- **dane ilościowe**, czyli dane w postaci liczb; np. czas (w miesiącach) od rozpoczęcia studiów do obronienia dyplomu; wiek (w latach); wysokość miesięcznego stypendium (w PLN);
- **dane jakościowe** opisujące cechę jakościową, jak np. płeć, kolor oczu, zawód itp.

Do opisu danych ilościowych możemy użyć miar liczbowych.

MIARY LICZBOWE DLA DANYCH IŁOŚCIOWYCH

1). Miary położenia:

- miary tendencji centralnej:
 1. średnia (mean): $\bar{x} := \frac{\sum_{i=1}^n x_i}{n}$
`> mean(wektor)` lub `> mean(wektor, na.rm=TRUE)` , gdy są obserwacje brakujące
 2. mediana (median) - wartość środkowa
`> median(wektor)`
 3. moda (dominanta) (mode) - wartość najczęściej pojawiająca się w próbie;
- miary pozycji:
 1. dolny kwartył (lower quartile): Q_1
`> quantile(wektor, 0.25)`
 2. górny kwartył (upper quartile): Q_3
`> quantile(wektor, 0.75)`
 3. decyle, percentyle i kwantyle (deciles, percentiles and quantiles): q_p
`> quantile(wektor, c(0.1, 0.99, 0.85))`
 Powyższa funkcja wyznacza pierwszy decyl, 99-ty percentyl i kwantyl rzędu 0.85.

2). Miary rozproszenia:

1. rozstęp (range): $Max - Min$
`> max(wektor) - min(wektor)`
2. rozstęp międzykwartyłowy (interquartile range): $IQR := Q_3 - Q_1$
`> IQR(wektor)`
3. wariancja (variance): $S^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
`> var(wektor)`
4. odchylenie standardowe (standard deviation): $S := \sqrt{S^2}$
`> sd(wektor)`

3). Miary kształtu:

1. skośność (współczynnik asymetrii) (skewness):

$$A := \frac{n}{(n-1)(n-2)S^3} \sum_{i=1}^n (x_i - \bar{x})^3$$

Jeśli obserwacje są symetrycznie rozłożone względem średniej (która w tej sytuacji równa się medianie), to $A = 0$.

2. kurtoza (współczynnik spłaszczenia) (kurtosis):

$$K := \frac{n(n+1)}{(n-1)(n-2)(n-3)S^4} \sum_{i=1}^n (x_i - \bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Wskazuje czy dane zawierają więcej i bardziej skrajne obserwacje odstające ($K > 0$) czy ich mniej i mniej skrajne ($K < 0$) niż byśmy oczekiwali od danych z rozkładu normalnego.

```
> install.packages("e1071")
> library(e1071)
> skewness(wektor)
> kurtosis(wektor)
```

GRAFICZNA PREZENTACJA DANYCH ILOŚCIOWYCH

1. wykres skrzynkowy (wykres typu *skrzynka z wąsami*) (boxplot)

```
> boxplot(wektor, range=1.5, horizontal=FALSE)
```
2. histogram liczności i histogram częstości (histograms)

```
> hist(wektor, freq=TRUE)      i      > hist(wektor, freq=FALSE)
```
3. jądrowy estymator gęstości (kernel density estimator) - wygładzona wersja histogramu częstości

```
> plot(density(wektor))
```

lub

```
> lines(density(wektor))
```

gdy jądrowy estymator gęstości chcemy nanieść na wcześniej sporządzony wykres, np. na histogram częstości.

GRAFICZNA PREZENTACJA DANYCH JAKOŚCIOWYCH

1. wykres słupkowy (barchart, barplot)

```
> barplot(licznosci, col=c("green", ..., "red"))
```
2. wykres kołowy (piechart)

```
> pie(licznosci, col=c("blue", ..., "yellow"))
```