1 Prawdopodobieństwo

Przypomnimy krótko podstawowe pojecia i fakty dotyczace teorii prawdopodobiębństwa. Na wstępie ograniczymy się do najprostrzych faktów. W kolejnych wykładach dotyczących już statystyki postaramy się stopniowo przypominać lub wprowadzić niezbene pojecia i fakty dotyczące rachunku prawdopodobieństwa.

1.1 Notacja

Wprowadzimy notację, przy użyciu której w miarę formalnie przedstawimy pojęcie prawdopodobieństwa, przy pomocy którego opisujemy i analizujemy zjawiska losowe.

- **Przestrzeń prób lub inaczej przestrzeń zdarzeń elementarnych** to zbiór możliwych wyników eksperymentu którego wynik jest losowy.

  Przykłady. *Wyniki rzutu rzut kostką do gry* $\Omega = \{1, 3, 4, 5, 6\}$. *Wyniki rzutu monetą* $\Omega = \{orze, reszka\}$. *Wzrost losowo wybranej osoby* $\Omega = \mathbb{R}_+$.

- **Zdarzenie losowe** to zbiór A będący podzbiór przestrzeni zdarzeń elementarnych $A \subseteq \Omega$.

  Przykłady. *W wynik rzutu kością otrzymaliśmy parzystą liczbę oczek,* $A = \{2, 4, 6\}$.

- **Zdarzenie elementarne** jest szczególnym wynikiem eksperymentu.

  Przykład. Wyrzucenie kostką 4 oczek.

- Symbolem zbioru pustego, $\emptyset$, oznaczamy zdarzenie niemożliwe.

  Przykład. Wyrzucenie sześcienną kostką do gry 7 oczek.

Przestrzeń próbki może być bardzo złożona lub bardzo prosta. Rozważmy przykład rzutu kostką. Jeśli kostka zostanie rzucona raz, dowolna liczba oczek od 1 do 6 może być wynikiem rzutu. Gdy eksperyment polega na rzucie kostką dwa razy, wynikiem jest już jedna z możliwych par: $(1, 1)$, $(1, 2)$, ... , $(6, 6)$ co daje łącznie 36 możliwych wyników. Tutaj liczymy $(1, 2)$ oraz $(2, 1)$ jako dwa różne wyniki. Jeśli eksperyment polega na rzucie kostką n razy, wyniki są wszystkimi możliwymi wektorami rozmairu n, $\mathbf{v} = (v_1, v_2, ..., v_n,)$, których każda ze współrzednych jest jedną z liczbach od 1 do 6. Daje to przestrzen zdarzeń elementarnych 6^n elemetową. Aby uzmysłwić sobie, jak duża to liczba wystarczy zauważ, że liczba ta dla $n = 130$ jest większa niż 10^{82}, co mozna porównać z szacunkowaną liczba atomów w obserwowanym Wszechświecie. Dlatego nie najmniejszych szans, aby rzucajac w ten sposób kostką do gry uzyskać wszystkie możliwe wyniki. Możemy jednak przewidzieć średnią liczbę oczek z n rzutów lub przewidzieć, ile wyników będzie większych niż 5. Biostatystyka zajmuje się pozyskiwaniem przydatnych, informacji ze skomlikowanych zdarzeń losowych, które mogą być wynikiem nawet najprostszych eksperymentów.

1.2 Rachunek prawdopodobieństwa

Miara prawdopodobieństwa, P, jest funkcją o wartościach rzeczywistych zdefiniowaną w zbiorze możliwych zdarzeniń, spełniająca następujace warunki:

1. Dla każdego zdarzenia $E \subseteq \Omega, \quad 0 \leqslant P(E) \leqslant 1$.
2. $P(\Omega) = 1$.
3. Jeśli E_i, dla $i = 1, 2, ...$ są zdarzeniami wzajemnie się wykluczającymi, czyli $E_i \cap E_j = \emptyset$ dla każdego $i \neq j$, to
$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

Własność (3) nazywa się to własnością przeliczalnej addytywności. W szczególności własność (3) implikuje skończoną addytywność

$$P\left(\bigcup_{i=1}^{N} E_i\right) = \sum_{i=1}^{N} P(E_i).$$

W oparciu o reguły teorii mnogości dotyczące działania na zbiorach jest stosunkowo łatwo pokazać następujace własności funkcji prawdopodobieństwa:

- $P(\emptyset) = 0$,
- $P(E) = 1 - P(\Omega \setminus E)$,
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,
- Jeśli $A \subseteq B, \quad$ to $\quad P(A) \leqslant P(B)$,
- $P\left(\bigcup_{i=1} E_i\right) \leqslant \sum_{i=1}^{n} P(E_i)$,
- $P\left(\bigcup_{i=1}^{n} E_i\right) \geqslant \max_i P(E_i)$.

Przykład 1. Według badań wynika, że że około 3% populacji Europy cierpi na bezdech senny natomiast 10% populacji Europy ma zespół niespokojnych nóg (RLS). Przeprowadzone badania dowodzą, że 58% dorosłych w Europie cierpi na bezsenność. Czy to oznacza, że 71% ludzi będzie miało co najmniej jedną z tych trzech zaburzeń snu?

Odpowiedź brzmi „nie”, ponieważ wydarzenia nie wykluczają się wzajemnie. To znaczy, jeśli osoba ma bezdech senny to nie oznacza, że nie może cierpieć na bezsenność lub doświadczać RLS. Zdefiniuj następujace zdarzenia

- $A_1 = \{\text{Osoba ma bezdech senny}\}$,
- $A_2 = \{\text{Osoba ma RLS}\}$,
- $A_3 = \{\text{Osoba cierpi na bezsenność}\}$.

Zdarzenie $\{$*Osoba ma co najmniej jeden z trzech problemów ze snem*$\}$ można zapisać je formalnie jako $A_1 \cup A_2 \cup A_3$. Jesteśmy zainteresowani wyznaczeniem prawdopodobieństwa tego zdarzenia $P(A_1 \cup A_2 \cup A_3)$, a więc prawdopodobieństwa, że losowo wybrana osoba z populacji będzie miała co najmniej jeden z trzech problemów ze snem. Wiemy już, że

$$P(A \cup B) = P(A) + P(B) - P(A \setminus B).$$

Oznaczając przez $A = A_1 \cup A_2$ i $B = A_3$ otrzymujemy

$$P(A_1 \cup A_2 \cup A_3) = P(A_1 \cup A_2) + P(A_3) - P(\{A1 \cup A_2\} \setminus A_3).$$

Ponieważ $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
oraz

$$P(\{A_1 \cup A_2\} \cap A_3) = P(\{A_1 \cap A_3\} \cup \{A_1 \cap A_3\}),$$

wynika z tego

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(\{A_1 \cap A_3\} \cup \{A_2 \cap A_3\}).$$

Łacząc wszystko razem otrzymujemy

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$$

Zatem

$$P(A_1 \cup A_2 \cup A_3) = 0,71 - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3).$$

Powyższy wzór sugeruje, jakie dodatkowe informacje są niezbędne aby właściwie obliczyć prawdopodobieńco najmnstwoiej jednego z trzech problemów ze snem.

Przykład 2. *Problem urodzinowy.* Jakie jest prawdopodobieństwo, że co najmniej dwie osoby siedzące na sali wykładowej mają urodziny tego samego dnia roku, choc być może nie w tym samym roku? Załóżmy, że dzień urodzin przypada w losowym dniu roku z takim samym prawdopodobieństwem dla każdego dnia roku. Zacznijmy od obliczenia tego prawdopodobieństwa dla pokoju z $n = 2$ osobami. Niech A oznacza zdarzenie "co najmniej dwie osoby na sali wykładowej mają urodziny w tym samym dniu roku" wtedy zdarzenie $A^c = \Omega \setminus A$ oznacza "każda osoba na sali ma urodziny w innym dniu roku" oraz

$$P(A) = 1 - P(A^c).$$

Obliczmy na ile sposobów możemy rozłożyć dni urodzin w taki sposób aby nie było urodzin tego samego dnia roku. Dla dwóch osób mamy w sumie 365×365 możliwych par dni roku, w których dwie osoby mogą mieć urodziny. Liczba par urodziny, które nie są takie same to 365×364, ponieważ pierwsza osoba może mieć urodziny w dowolnym dniu roku (365), podczas gdy druga osoba może mieć urodziny w dowolnym dniu roku za wyjatkiem dnia w którym urodziny ma pierwsza osoba (364). Zatem, gdy $n = 2$ to

$$P(A) = 1 - \frac{365}{\times}\frac{364}{365}.$$

Korzystając z podobnego rozumowania dla n = 3, mamy

$$P(A) = 1 - \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365}.$$

i dla dowolnego n mamy

$$P(A) = 1 - \frac{365}{365} \times \frac{364}{365} \times \ ... \ \times \frac{365 - n + 1}{365}.$$

Przyjeliśmy założenia, które pozwoliły nan w prosty sposób obliczyć prawdopodobieństwo zdarzenia A. Pierwsze z przyjętych założeń to niezależność dni urodzin każdej z osób na sali. To oczywiście nie byłoby prawdą, jeśli osoby nie zostały wybrane niezależnie od siebie. Na przykład, jeśli obie osoby były by wybrane tak aby ich dni urodzin wypadały w tym samym miesiącu. Drugim założeniem było to, że każdy dzień ma jednakowe prawdopodobieństwo. Obliczenia byłyby trudniejsze, ale można je wykonać. Jeśli jest więcej niż 365 osób, prawdopodobieństwo wynosi 1, tutaj ignorujemy lata przestępne. Dlatego obliczamy prawdopodobieństwo tylko dla pokoje z mniej niż $n < 365$ osób.

```
n=1:364                          #Vector of number of people in the room
pn=n                             #Vector that will contain the probabilitie
for (i in 1:364) {
  pn[i]<- 1-prod(365:(365-i+1))/365^i
}  #Prob. of >= 2 people with same B-day
```

Możemy obliczyć prawdopodobieństwa dla liczby osób na sali wykładowej równej 23 i 57.

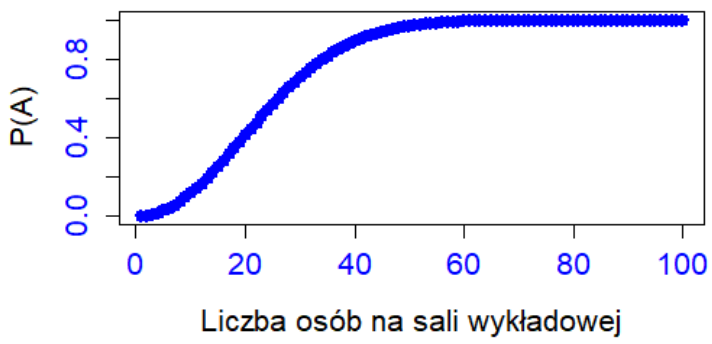
```
round(pn [c (23,57)], digital = 3)
```

```
[1] 0,507 0,990
```

Zwróćmy uwagę, że dla grupy 23 osób istnieje 50% szansa, że co najmniej dwie osoby będą miały urodziny tego samego dnia roku, podczas gdy dla grupy 57 osób jest już 99% szans. Te wyniki mogą być nieco zaskakujące. Możemy narysować wykres funkcji, która podaje prawdopodobieństwo, że co najmniej dwie osóby będą urodziny w tym samym dniu, w zależności od liczby osób na sali.

```
plot(n[1:100],pn[1:100],type="p",pch=19,col="blue",lwd=3,
     xlab="Number of people in the room",ylab="P(at least 2 same-day B-days)",
     cex.lab=1.3,cex.axis=1.3,col.axis="blue")
```

Na rysunku przedstawiamy jedynie prawdopodobieństwo tego zdarzenia dla $n < 100$ gdyż dla $n > 100$ prawdopodobieństwo to jest bardzo bliskie wartości 1.



Prawdopodobieństwo można oszacować za pomocą symulacji metodą Monte Carlo. Wykorzystamy w tym celu następujący kod w R.

```
set.seed(7654098)
n=23                    #number of people in the room
n_sim=10000             #number of Monte Carlo simulated rooms with n people
are_2_birthdays<-rep(NA,n_sim)  #vector of indicators for >=2 same B-day
for (i in 1:n_sim)
  {#begin simulations
  #calculate the number of unique birthdays in a group of n people
  n_birthdays<-length(unique(sample(1:365,n,replace = TRUE)))
  are_2_birthdays[i]<-(n_birthdays<n)
  }#end simulations
mean(are_2_birthdays)
```

```
[1] 0,4994
```

Wynik symulacji Monte Carlo nie zapewnia idealnego obliczenia prawdopodobieństwo, ale podana wartość jest bardzo blisko prawdziwej. Można to jeszcze poprawić poprzez zwiększenie liczby symulacji do 100 000 lub więcej. Zapewnia to przykład tego, jak potężne mogą być symulacje; w rzeczywistości można łatwo dostosować kod uwzględniający nierówne prawdopodobieństwo urodzin dla różnych osób dni w roku lub bardziej skomplikowane reguły decyzyjne. Załóżmy na przykład, że jesteśmy zainteresowani uzyskaniem prawdopodobieństwa, że w pokoju z n ludźmi dokładnie dwie dziewczynki urodziy się tego samego dnia, a jeden chłopiec urodził się tego samego dnia co te dwiema dziewczynami, ale co najmniej trzech chłopców rodzi się tego samego dnia, który jest różny niż dzień urodzin tych dziewczynek. Oczywiście to bardzo sztuczny przykład, ale stanowi przykład pytania, na które może z łatwością odpowiedzi przy pomocy symulowane, ale dużo trudej odpowiedzieć przy użyciu jawnych obliczeń.

Krzysztof Topolski