Prawdopodobieństwo

1 Pobieranie próbek w R.

Sposoby pobierania próbek i symulacje są bardzo ważne w analizie danych. Dlatego poświęcimy trochę uwagi generowaniu prób losowych przy użyciu programu R oraz symulacją. Symulacja jest potężnym narzędziem do obliczeń wielkości, które są trudne do obliczenia w sposób analityczny. Metody pobierania prób, prawdopodobieństwo i statystyka idą w parze, a zrozumienie sposobów pobierania próbek jest niezbędne dla poprawnej analizy danych o charakterze losowym. Poniżej przedstawiamy kilka podstawowych sposobów pobierania prób w języku R.

1.1 Pobieranie próby bez zwracania

Najczęściej spotykanym w praktyce sposobem pobierania próby jest losowanie bez zwracania. W tam przypadku, po każdym losowaniu element populacji nie może zostać wylosowany ponownie.

```
x1 <- sample (1:6)                    #a random permutation
x2 <- sample (1:6)                    #and another
x3 <- sample (1:6)                    #and another
x1
```

```
[1] 4 1 3 6 5 2
```

```
x2
```

```
[1] 6 4 3 2 5 1
```

```
x3
```

```
[1] 3 5 1 6 2 4
```

Losujemy w ten sposób ze zbioru wszystkich permutacje. Ten sposób losowania jes szeroko stosowane w losowych testach permutacyjnych na przykład, gdy ktoś chce porównać leczenie z grupą kontrolną. W tym przypadku konstruowana jest statystyka testowa (np. różnica średnch w dwóch porównywanych grupach).

1.2 Pobieranie próbek ze zwracaniem

Ten sposób wybierana próby jest wygodny z punktu widzenia teoretycznego gdyż za każdym razem wybieramy elementy do próby z tym samym prawdopodobieństwem. W takim przypadku po każdym losowaniu element wybrany do próby jest może być wylosowany ponownie.

```
x1 <- sample (1:6,10,replace=TRUE)        #sampling with replacement
x2 <- sample (1:6,10,replace=TRUE)        #again
x3 <- sample (1:6,10,replace=TRUE)        #and again
x1
```

```
 [1] 6 1 3 2 2 1 6 3 4 4
```

```
x2
```

```
 [1] 1 5 2 1 1 3 2 5 2 6
```

```
x3
```

```
 [1] 2 3 3 2 1 2 3 4 1 3
```

Po wybraniu próby możemy obliczyć typowe wielkości, które pozwalają scharakteryzować populację z której dokonywaliśmy próbkowania.

```
sum (x1 == 3)
```

```
[1] 2
```

```
max (x1)
```

```
[1] 6
```

1.3 Rozkład zmiennej losowej

Przed przeprowadzeniem eksperymentu nie wiemy dokładnie, jaki będzie jego wynik, ale zwykle znamy wszystkie możliwe wyniki. Na przykład, dla konkretnej osoby z rakiem płuc nie wiemy, czy przeżyje ona kolejne pięć lat. Wiemy jednak, że za pięć lat albo będą martwy, oznaczmy to zdarzenie przez (0), albo będzie żywy, co oznaczymy przez (1). Rozkład takiej zmiennej losowej jest całkowicie scharakteryzowany. Jeśli wiemy, że prawdopodobieństwo bycia martwym jest równe p_0 to prawdopodobieństwo bycia żywym po pięciu latach jest równe $p_1 = 1 - p_0$. Dla zmiennych dyskretnych istnieje tylko skończona lub przeliczalna liczba możliwych wyników, a rozkład jest całkowicie określony, jeśli prawdopodobieństwo każdego możliwego wyniku eksperymentu jest znany. Funkcja rozkładu prawdopodobieństwa zmiennej losowej dyskretnej X jest funkcją, która określa prawdopodobieństwo, tego że X przyjmuje określoną ustaloną wartość. Jeśli $\mathcal{K}$ jest zbiorem wszystkich możliwe wyników, które może przyjmować zmienna losowa X, to funkcja rozkłądu prawdopodobieństwa zmiennej losowej X jest określona jako $p_k = P(X = k), \quad$ dla każdego $k \in \mathcal{K}$. Funkcja rozkładu prawdopodobieństwa p musi spełniać następujące warunki:

1. $p_k \geqslant 0$, dla każdego $k \in \mathcal{K}$.
2. $\sum_{k \in \mathcal{K}} p_k = 1$,

W powyższej suma uwzględniamy wszystkie możliwe wartości wyniku eksperymentu. Wprowadzimy notację, przy użyciu której w miarę formalnie przedstawimy pojęcie prawdopodobieństwa, przy pomocy którego opisujemy i analizujemy zjawiska losowe.

Zilustrujmy zagadnienie funkcji rozkładu prawdopodobieństwa dyskretnej zmiennej losowej na przykładzie zmienneej losowej o rozkłądzie Bernoulliego. Zmienna losowa Bernoulliego jest wynikiem eksperymentu, który może dać w wyniku sukces, oznaczony jako 1 lub niepowodzenie, oznaczony jako 0. Istnieje wiele takich eksperymentów. Przykładami oprócz kalsycznego w rachunku prawdopodobieństwa rzutu monetą dającego w wyniku wyrzucenie orła lub reszki, również zdiagnozowanie u badanej osoby zmaina nowotworowych lub ich brak, przeżycie pięciu lat po zdiagnozowaniu raka płuc lub śmieć w tym okresie, brak powikłań popoperacyjnych po wykonaniu interwencji chirurgicznej lub wystąpienie powikłąń itp.zmarł i nie umarł po pięciu latach, Niech zmienna losowa X opisuje wynikiem rzutu monetą, gdzie $X = 0$ oznacza reszkę, a $X = 1$oznacza orła. Jeśli moneta jest symetryczna, to rozkład tej zmiennej losowej ma postać:

$$p_0 = P(X = 0) = 0,5 \quad \text{oraz } p_1 = P(X = 1) = 0,5.$$

Bardziej zwarty sposób zapisanie tego rozkłądu może mieć postać:

$$p_x = 0,5x0,51 - xdlax = 0,1.$$

Załóżmy teraz, że zmienna losowa X opisuje czy dana osoba z rakiem płuc umrze w przeciągu najbliższych pięć lat czy też przeżyje. Rozkład tej zmiennej losowej ma postać:

$$p_0 = P(X = 0) = \theta \quad \text{oraz } p_1 = P(X = 1) = 1 - \theta,$$

gdzie θ oznacza prawdopodobieństwo śmierci w przeciagu pięciu lat. W bardziej zwarty sposób możemy zapisać te rozkłądu w następującej formie: Zwykle nie znamy wrtości parametru θ i jednym z zagadnień biostatystyki jest oszacowanie wartości tego parametru dla badanej populacji pacjentów na podstawie próby losowej. W programie `R` w prost sposób możemy generować wartości zmiennej losowej orozkłądzie Bernoulliego. Poniżej pokazujemy jak
(1) wygenerować 21 niezależnych próbek z rozkładzu Bernoulliego z prawdopodobieństwem sukcesu 0,5, czyli 21 niezależnych rzutów symetryczną monetą.
(2) wygenerować 21 niezależnych próbek z rozkłądu Bernoulliego z prawdopodobieństwami sukcesu: 0,00, 0,05, 0,10, 0,15,..., 0,95, 1,00.

```
x1<-rbinom(21,1,0.5)
x2<-rbinom(21,1,seq(0,1,length=21))
x1
```

```
 [1] 0 0 1 1 0 1 1 0 0 0 1 1 1 0 1 0 0 0 1 0 0
```

```
x1
[1] 0 0 1 1 0 1 1 0 0 0 1 1 1 0 1 0 0 0 1 0 0
x2
[1] 0 0 0 0 0 0 0 0 1 0 0 1 0 1 1 1 1 1 0 0 1 1
```

Zauważmy, że te dwa wektory są różne i odzwierciedlają różne mechanizmy rzucania monetą. Pierwszy wektor jest bardziej chaotyczny pod względem zmian od 0 do 1, podczas gdy drugi

wektor ma na początku więcej zer i mniej zera pod koniec. Dzieje się tak, ponieważ prawdopodobieństwo sukcesu dla x_2 są znacznie wyższe pod koniec wektora niż na początku. Oczywiście w praktyce nie wiedzielibyśmy nic o prawdopodobieństwie sukcesu lub niepowodzenia, zobaczymy tylko wynik eksperymentu. Pytanie brzmi, czego dane mogłyby zasugerować nam, czego powinniśmy się spodziewać w przyszłości. Aby to stwierdzić musimy ponownie przeprowadzić oba eksperymenty tym razem w następujący sposób:

```
x1<-rbinom(21,1,0.5)
x2<-rbinom(21,1,seq(0,1,length=21))
x1
```

```
 [1] 1 1 0 0 0 1 0 1 1 1 1 1 1 1 0 1 1 0 0 0 1
```

```
x2
```

```
 [1] 0 0 0 0 1 0 0 0 0 1 1 0 0 1 1 1 1 1 1 1 1
```

1.4 Zmienna losowe o rozkładzie Poissona.

Zmienna losowa Poissona jest wynikiem eksperymentu przyjmujacego wartości ze zbioru przeliczalnego

$$\mathcal{K} = \{0, 1, 2, ...\}$$

z prawdopodobieństwami określonymi jako:

$$P(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!} \quad \text{gdzie } k = 0, 1, 2, ...$$

oraz λ jest parametrem okeslajacym wartość oczekiwaną zmiennej losowej X. Istnieje wiele eksperymentów, które wyniki mozna opisać zmienną losową o rozkładzie Poissona, w tym liczby pacjentów przybywających do kliniki danego dnia, liczba patogenów przenoszonych przez wodę w próbce wody lub liczba miejsc crosin-over. W prosty sposób możemy wysymulować próbę z rozkładu Poissona. Załózmy, że chcemy wygenerować zgodnie z rozkładem Poissona niezależnie dwa okresy po 15 dni ze średnią liczbą pacjentów na dzień = 20.

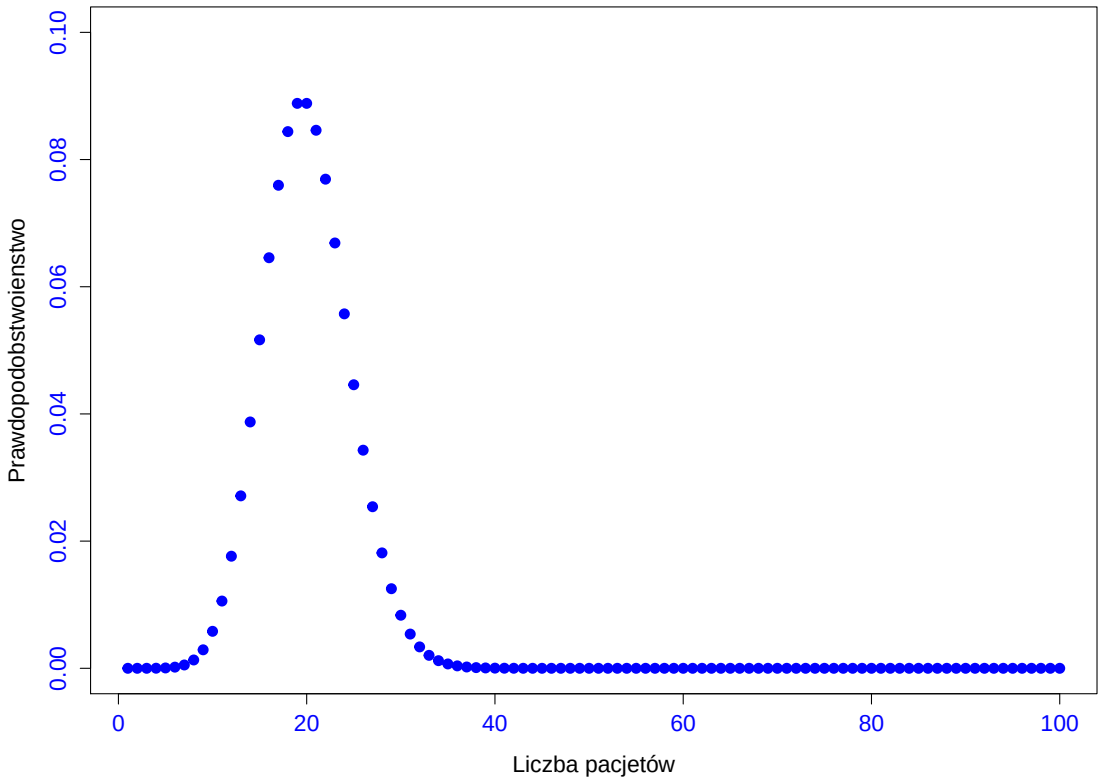
```
rpois(15,20)
```

```
 [1] 18 24 17 15 14 20 11 17 27 17 14 19 17 21 26
```

```
rpois(15,20)
```

```
 [1] 20 17 22 17 18 16 17 19 19 21 20 15 26 19 24
```

Rozważmy zmienną losową X, która ma rozkład $Poissona(\lambda)$. Zwykle jest to oznzczane jako $X \sim Poisson(\lambda)$. Poniższy rysunek pokazuje funkcję rozkładu prawdopodobieństwa zmiennej losowej $Poissona(\lambda)$ gdzie $\lambda = 20$.



Rysunek ten wygenerowaliśmy korzystając z następującego zestawu instrukcji programu R.

```
x=1:100
lambda=20
plot(x,dpois(x,lambda),type="p",pch=19,col="blue",lwd=3,
     xlab="Number of patients",ylab="Probability",cex.lab=1.3,
     cex.axis=1.3,col.axis="blue",ylim=c(0,0.1))
```

Oczywiście w praktyce nie wiemy, że liczba pacjentów w danym dniu jest zgodny z rozkładem *Poissona*(20) lub dowolnym innym rozkładem. Znamy tylko dane, liczbę pacjentów w ciągu wielu dni. Dlatego sensowne jest użycie danych i na ich podstawie spróbować wywnioskować mechanizm, który generuje liczba pacjentów dziennie. Aby zobaczyć, jak to działa, przeprowadzimy symulację liczby pacjentów w ciągu kolejnych 1000 dni (około 3 lat danych) zgodnie z rozkądem *Poissona*(20), a następnie po prostu wykreślmy częstotliwość odwiedzin określonej liczby pacjentów. Na przykład dla liczby odwiedziń 20 obliczymy, ile było dni, w których miało miejsce dokładnie 20 wizyt, a następnie podzielmy tą liczbę przez 1000. W ten sposób możemy zrekonstruować funkcję prawdopodobieństwa na podstawie zaobserwowanych danych.

```
y<-rpois(1000,20)            #simulate 1000 independent Poisson(20)
py=rep(0,100)                #storage vector of Poisson probabilities
for (i in 1:length(py))      #for every possible outcome between 1 and 100
    {py[i]<-mean(y==i)}      #calculate the frequency of observing i subjects
```

Funkcja `table` pozwala nam wyliczć liczbę wystąpienia poszczególnych wartości zmiennej y.

```
tab = table(y)
tab
```

```
y
  8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25
  2   1   3  11  20  37  45  50  55 101  96  81 102  83  66  66  64  34
 26  27  28  29  30  31  32  33  34  36  37
 19  21  19   8   4   4   2   2   2   1   1
```

Podczas gdy funkcja `prop.table` pozwala wyliczyć frakcję poszczególnych wartości w wygenerowanej tablicy.

```
prop.table(tab)
```

```
y
    8     9    10    11    12    13    14    15    16    17    18    19
0.002 0.001 0.003 0.011 0.020 0.037 0.045 0.050 0.055 0.101 0.096 0.081
   20    21    22    23    24    25    26    27    28    29    30    31
0.102 0.083 0.066 0.066 0.064 0.034 0.019 0.021 0.019 0.008 0.004 0.004
   32    33    34    36    37
0.002 0.002 0.002 0.001 0.001
```

Teraz możemy wygenerować wartości funkcji prawdopodobieństwa rozkładu $Poissona(20)$.

```
x=1:100                          #Set the x-axis, where the Poisson pdf is evaluated
lambda=20                        #Set the Poisson success rate
```

Natępnie sporządzić wykres empirycznego rozkładu wraz z naniesionym rozkładem $Poissona(20)$.

Otrzymujemy następujący wykres

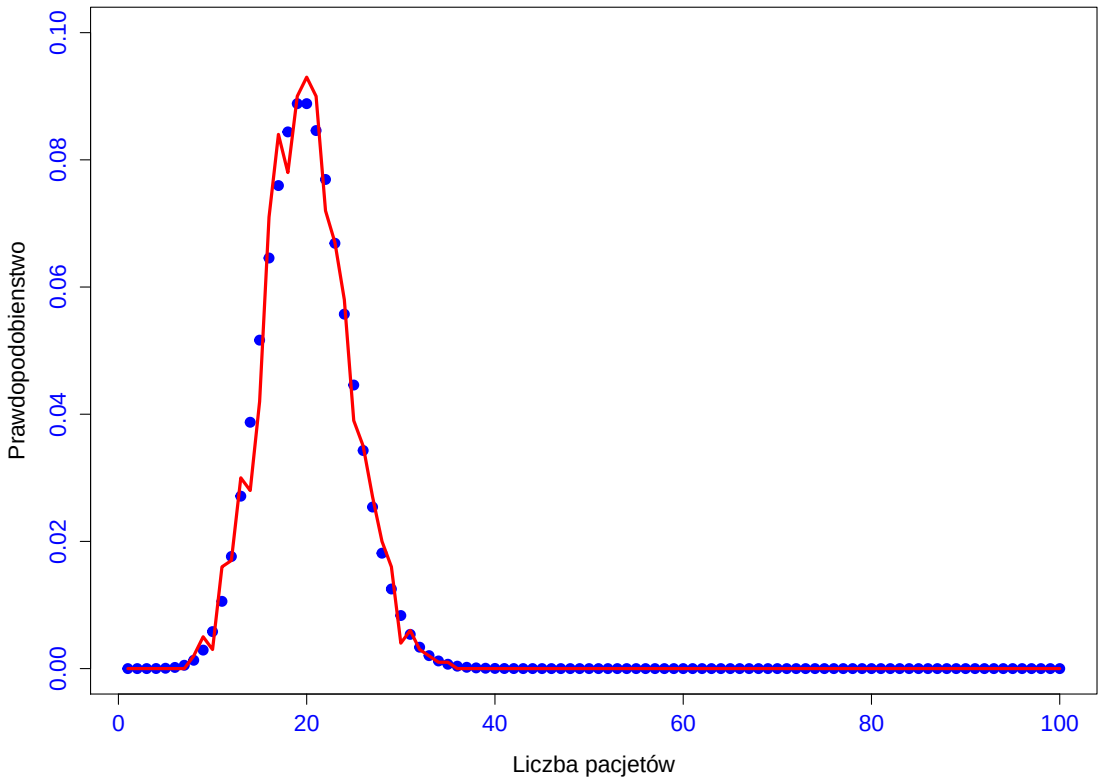
Jeśli, co jest bardziej naturalne, przypuszczamy, że mamy do czynienia z rozkłądem Poissona ale nie znamy wartości parametry λ możemy postąpić w następujący sposób. Najpierw wyestymować, na podstwaie danych, wartość tego parametru, a następnie wykreślić funkcję rozkłądu prawdopodobieństwa Poissona z parametrem λ równą wyestymowanej wartości. W przypadku rozkładu Poissona parametr λ jest równy wartości oczekiwanej więc mozemy wyestymować wartość *lamda* przy pomocy średniej probkowej z wygenerowanej próby.

```
mean(y)
```

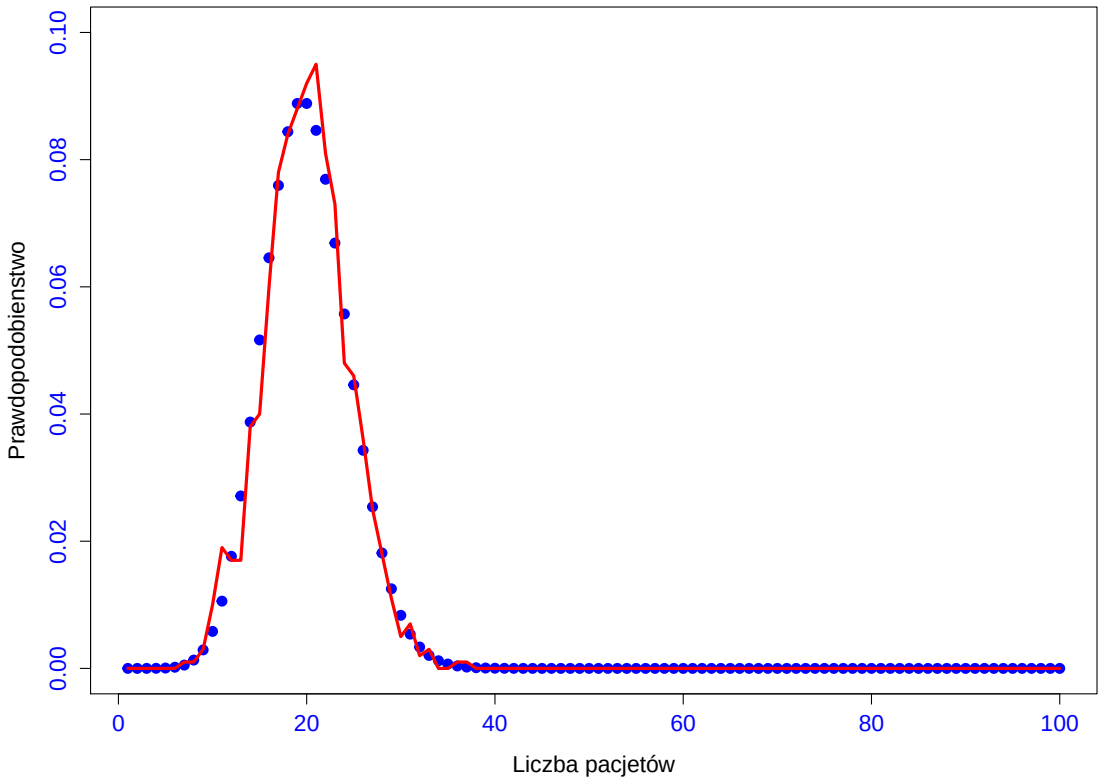
```
[1] 20.013
```

Teraz możemy wygenerować wykres empirycznego rozkładu wraz z rozkładem Poissona z westymowaną watoscią parametru $\lambda = 20,013$.

```r
plot(x,dpois(x,lambda),type="p",pch=19,col="blue",lwd=3,
     xlab="Number of patients",ylab="Probability",cex.lab=1.3,
     cex.axis=1.3,col.axis="blue",ylim=c(0,0.1))
lines(1:100,py,col="red",lwd=3)  #here, lwd controls the thickness
```



Krzysztof Topolski


Prawdopodobienstwo
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
Liczba pacjetów