PROJEKT NUMER 15

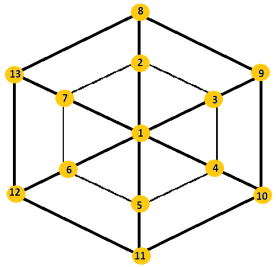
ETAP 1 ORAZ ANALIZA STATYSTYCZNA WYNIKÓW

PODSTAWY TEORII PROCESÓW STOCHASTYCZNYCH

MAGDALENA OLBRYŚ, MMAD, II rok, grupa C

Problem:

Dwa pająki poruszają się losowo między wierzchołkami symetrycznej sieci z 6 promieniami i 2 pętlami:



Jeśli pająk znajduje się w węźle sieci z którego odchodzi m połączeń, to z prawdopodobieństwem p = $\frac{1}{m}$ wybiera jedno z nich w następnym kroku.

(a) Pająki znajdują się na przeciwległych wierzchołkach sieci najbardziej zewnętrznej pętli. Znajdź oczekiwaną liczbę ruchów, aby pająki spotkały się w jednym węźle sieci.

(b) Jeden pająk znajduje się w centrum sieci a drugi w jednym z węzłów sieci na najbardziej zewnętrznej pętli. Znajdź oczekiwaną liczbę ruchów, aby pająki spotkały się w jednym węźle sieci.

1.Opracowanie teorytyczne pojęć potrzebnych do rozwiązania problemu.

1. Prawdopodobieństwo.

Definicja Laplace'a:

Niech dany będzie skończony zbiór Ω wszystkich możliwych zdarzeń elementarnych. Dowolny podzbiór *A* zbioru Ω nazywa się wtedy zdarzeniem.

Prawdopodobieństwem *P(A)* zajścia zdarzenia *A* nazywa się stosunek liczby zdarzeń elementarnych sprzyjających zdarzeniu *A* do liczby wszystkich możliwych zdarzeń elementarnych należących do zbioru Ω. Czyli prawdopodobieństwem zajścia zdarzenia *A* nazywamy liczbę:

$$P(A) = \frac{|A|}{|\,\Omega\,|}$$

gdzie | . | oznacza liczbę elementów danego zbioru.

Jeśli jeden z naszych pająków znajduje się w węźle sieci z którego odchodzi m połączeń, to z prawdopodobieństwem p = $\frac{1}{m}$ wybiera jedno z nich w następnym kroku.

2. Spacer losowy.

Spacer losowy to pojęcie określające ruch losowy: w kolejnych chwilach czasu cząstka przemieszcza się z aktualnego położenia do innego, losowo wybranego.

Załóżmy , że $X_1, X_2, \ldots$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie prawdopodobieństwa takim, że dla każdego $i \in N$:

$$X_i = \begin{cases} 1, & \text{z prawdpopodobieństwem p} \\ -1 & \text{z prawdopodobieństwem } 1-\text{p}. \end{cases}$$

Położenie cząstki po i-tym kroku określamy wzorem:

$$S_n = S_0 + \sum_{i=1}^{n} X_i\,,$$

gdzie S_0 to położenie początkowe.

Nasze pająki poruszają się spacerem losowym.

3. Wartość oczekiwana.

Wartość oczekiwana (wartość średnia, wartość przeciętna) jest wartością spodziewaną w doświadczeniu losowym, czyli w takim gdzie nie możemy z całkowitą pewnością określić wyniku.

Definicja:

Jeżeli *X* jest zmienną losową na przestrzeni probabilistycznej(Ω, F, P) o wartościach w *R*, to wartością oczekiwaną zmiennej *X* nazywa się liczbę:

$$EX \coloneqq \int_{\Omega} XdP$$

o ile ona istnieje tzn. $E|X| = \int_{\Omega} XdP < +\infty$.

W przypadku gdy zmienna losowa *X* ma rozkład dyskretny i przyjmuje tylko skończenie wiele wartości $x_1, x_2, \dots, x_n$ z prawdopodobieństwami wynoszącymi odpowiednio $p_1, p_2, \dots, p_n$ to z powyższej definicji wynika następujący wzór na wartość oczekiwaną *EX*.

$$EX = \sum_{i=1}^{n} x_i * p_i$$

Jeśli zmienna X przyjmuje nieskończenie, ale przeliczalnie wiele wartości to $EX = \sum_{i=1}^{\infty} x_i * p_i$(istnieje ona tylko wtedy gdy szereg ten jest zbieżny bezwzględnie.

W naszym rozkładzie liczymy wartość oczekiwaną dodając do siebie wyniki pojedynczych doświadczeń i dzieląc ją przez ich ilość.

4. Własność Markowa.

Własność markowa to własność procesów stochastycznych polegająca na tym, że przyszłe stany procesu są warunkowo niezależne od stanów przeszłych.

Proces Markowa to proces stochastyczny, który spełnia własność Markowa.

W naszym doświadczeniu kolejna pozycja pająka jest zależna tylko od jego poprzedniej pozycji, nie jest zależna od tych wcześniejszych.

5. Łańcuch Markowa.

Rozważmy ciąg zmiennych losowych $X_0, X_1, X_2, \dots$ odpowiadającym chwilom *0,1,2,…*

Dyskretny łańcuch Markowa *{X_n, n=0,1,2,…}* definiujemy jako dyskretny proces stochastyczny spełniający dla wszystkich liczb naturalnych *n* oraz dla wszystkich stanów X_n warunek:

$$P(\{X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0,\}) = P(\{X_{n+1} = x_{n+1} | X_n = x_n\}).$$

6. Macierz przejścia.

Rozważmy prawdopodobieństwo warunkowe przejścia ze stanu $x_n = i$ do stanu $x_{n+1} = j$.

$$P(\{x_{n+1} = j | x_n = i\}) = P_{ij}(n).$$

Macierzą przejścia nazywamy macierz:

$$P(n) = \begin{bmatrix} P_{00}(n) & P_{01}(n) & P_{02}(n) & \dots \\ P_{10}(n) & P_{11}(n) & P_{12}(n) & \dots \\ P_{20}(n) & P_{21}(n) & P_{22}(n) & \dots \\ \dots & & & \end{bmatrix}$$

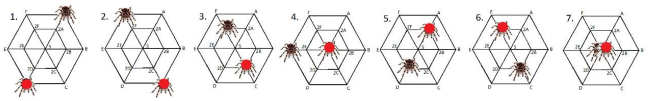
gdzie $0 \le P_{ij}(n) \le 1, \quad \sum_{j=0}^{+\infty} P_{ij}(n) = 1.$

Wyznaczam macierz przejścia dla naszych prawdopodobieństw (13 stanów):

```
P = [[0.0, 1/6, 1/6, 1/6, 1/6, 1/6, 1/6, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
     1/4, 0.0, 1/4, 0.0, 0.0, 0.0, 1/4, 1/4, 0.0, 0.0, 0.0, 0.0, 0.0
     1/4, 1/4, 0.0, 1/4, 0.0, 0.0, 0.0, 0.0, 1/4, 0.0, 0.0, 0.0, 0.0
     1/4, 0.0, 1/4, 0.0, 1/4, 0.0, 0.0, 0.0, 0.0, 1/4, 0.0, 0.0, 0.0
     1/4, 0.0, 0.0, 1/4, 0.0, 1/4, 0.0, 0.0, 0.0, 0.0, 1/4, 0.0, 0.0
     1/4, 0.0, 0.0, 0.0, 1/4, 0.0, 1/4, 0.0, 0.0, 0.0, 0.0, 1/4, 0.0
     1/4, 1/4, 0.0, 0.0, 0.0, 1/4, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1/4
     0.0, 1/3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1/3, 0.0, 0.0, 0.0, 1/3
     0.0, 0.0, 1/3, 0.0, 0.0, 0.0, 0.0, 1/3, 0.0, 1/3, 0.0, 0.0, 0.0
     0.0, 0.0, 0.0, 1/3, 0.0, 0.0, 0.0, 0.0, 1/3, 0.0, 1/3, 0.0, 0.0
     0.0, 0.0, 0.0, 0.0, 1/3, 0.0, 0.0, 0.0, 0.0, 1/3, 0.0, 1/3, 0.0
     0.0, 0.0, 0.0, 0.0, 0.0, 1/3, 0.0, 0.0, 0.0, 0.0, 1/3, 0.0, 1/3
     0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1/3, 1/3, 0.0, 0.0, 0.0, 1/3, 0.0]]
```

2.Plan symulacji pozwalającej rozwiązać problem.

1. Symulację wykonuję w języku Python.
2. Zapisuję wyznaczoną macierz przejścia dla stanów 1-13.
3. Definiuję funkcję los(p) w celu wykonywania przejść pomiędzy stanami.
   - jej argumentem jest wiersz macierzy;
   - losujemy liczbę a z przedziału od 0 do 1;
   - funkcja los(p) poprzez porównanie znajduje miejsce wylosowanej przez nas liczby a w wierszu macierzy i zależnie od jej wartości zwraca liczbę od 1 do 13 (bo tyle właśnie mamy stanów)
4. Definiuje funkcję „pajaczki", której argumentami są kolejno pozycje początkowe obu pająków.
   - jej argumentami są kolejno pozycje początkowe obu pająków;
   - w funkcji dodajemy 2 początkowo puste tablice – trasa 1 pająka, oraz trasa 2 pająka;
   - dodajemy pętle while – działa dopóki pająki nie spotkają się w tym samym miejscu;
   - w pętli losujemy kroki pająków oraz zliczamy ich liczbę;
   - funkcja zwraca liczbę kroków, które pająki musiały wykonać aby się spotkać.



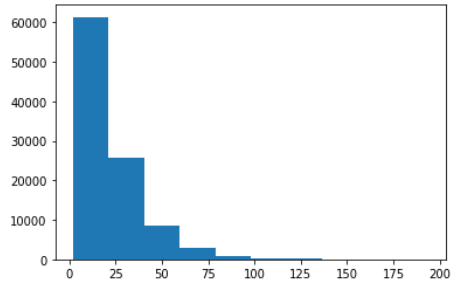
Rys1. Przykładowy „spacer" pająków. Poruszają się losowo do momentu w którym spotkają się (dla podpunktu a).

5. Wywołuję funkcję „pajączki" dla 100000 przypadków i wyniki zapisujemy w liście. Następnie wyliczamy wartość oczekiwaną – dodajemy do siebie te wyniki i dzielimy sumę przez liczbę prób(100000).
6. Zapisuję kolejne liczby kroków, pająków w pliku tekstowym, aby następnie użyć ich do analizy statystycznej w Rstudio.
7. Wykonuję symulację 1000 razy a następnie wyliczam średnią wartość oczekiwaną dzieląc sumę wyników przez liczbę przypadków.
8. Zapisuję listę 1000 wartości oczekiwanych w pliku tekstowym.

Wyniki symulacji:

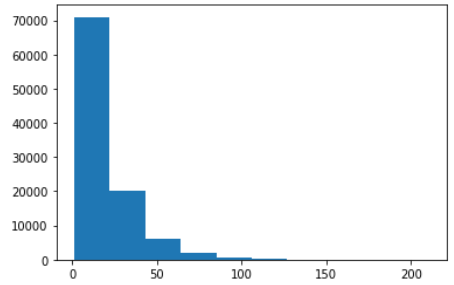
Kod symulacji znajduję się w załączonym pliku (kodpython.py).

a) Średnia oczekiwana liczba ruchów, aby pająki spotkały się w jednym węźle sieci, jeśli pająki znajdują się na przeciwległych wierzchołkach sieci najbardziej zewnętrznej pętli: 22.054.



Histogram przedstawia rozkład liczb kroków, które pająki musiały wykonać aby się spotkać (oś x – liczba kroków, oś y – liczba prób) -dla pojedynczej symulacji.

b) Średnia oczekiwana liczba ruchów, aby pająki spotkały się w jednym węźle sieci, jeśli jeden pająk znajduje się w centrum sieci a drugi w jednym z węzłów sieci na najbardziej zewnętrznej pętli: 17.4761.



Histogram przedstawia rozkład liczb kroków, które pająki musiały wykonać aby się spotkać (oś x – liczba kroków, oś y – liczba prób) – dla pojedynczej symulacji.

3.Metody statystyczne, które umożliwią właściwą analizę problemu.

1. Test statystyczny.

Jest to formuła matematyczna pozwalająca oszacować prawdopodobieństwo spełnienia pewnej hipotezy statystycznej w populacji na podstawie próby losowej z tej populacji. Rozróżniamy testy:

1) Parametryczne, które służą do weryfikacji hipotez parametrycznych, odnoszących się do parametrów rozkładu badanej cechy w populacji generalnej. Najczęściej weryfikują sądy o takich parametrach populacji jak średnia arytmetyczna, wskaźnik struktury i wariancja.

2) Nieparametryczne, które służą do weryfikacji różnorodnych hipotez, dotyczących m.in. zgodności rozkładu cechy w populacji z określonym rozkładem teoretycznym, zgodności rozkładów w dwóch populacjach, a także losowości doboru próby.

Stosując testy statystyczne spróbuję dopasować nasz rozkład do jakiegoś znanego rozkładu.

2. Test Kołmogorowa – Smirnowa dla dwóch prób.

Testu Kołmogorowa-Smirnowa używamy do sprawdzenia, czy dwa jednowymiarowe rozkłady prawdopodobieństwa różnią się od siebie.

Notacje: $X_i, i = 1, \dots, n_x$ mają $rozkład\ taki, że\ F_{x_i} = F_1$

$Y_i, i = 1, \dots, n_Y$ mają $rozkład\ taki, że\ F_{Y_i} = F_1$

$$H_0: F_1 \equiv F_2$$

Statystyka:

$$D_{n_x,n_y} = \sup_{t \in R} |F_1^*(t) - F_2^*(t)|$$

$F_1^*(t)$ – dystrybuanta empiryczna próby X_i

$F_2^*(t)$ – dystrybuanta empiryczna próby Y_i

$$\lambda = \sqrt{n^*} D_{n_x,n_y}$$

$$n^* = \frac{n_x n_y}{n_x + n_y} = \frac{1}{\frac{1}{n_x} + \frac{1}{n_y}}$$

Hipoteza zerowa jest odrzucana na poziomie α jeśli:

$$\lambda > K_\alpha,$$

gdzie K_α dane jest przez $P(K \leq K_\alpha) = 1 - \alpha$.

Test Kołmogorowa – Smirnowa będę wykonywać przy pomocy Rstudio. Najpierw porównam nasz rozkład do rozkładu Poissona.

3. Rozkład Poissona.

Jest to dyskretny rozkład prawdopodobieństwa, wyrażający prawdopodobieństwo szeregu wydarzeń mających miejsce w określonym czasie, gdy te wydarzenia występują ze znaną średnią częstotliwością i w sposób niezależny od czasu jaki upłynął od ostatniego zajścia takiego zdarzenia.

Jeśli oczekiwaną liczbą zdarzeń w tym przedziale jest *λ*, to prawdopodobieństwo, że jest dokładnie *k* wystąpień, gdzie *k* jest nieujemną liczbą całkowitą, *k = 0, 1, 2,* jest równe:

$$f(\lambda, \mathrm{k}) = \frac{\lambda^k * e^{-\lambda}}{k!},$$

gdzie:

e – podstawa logarytmu naturalnego e = 2,71828…

k – liczba wystąpień zdarzenia

λ – dodatnia liczba rzeczywista, równa oczekiwanej liczbie zdarzeń w danym przedziale czasu.

Przy pomocy R studio sprawdzam czy nasz rozkład ma rozkład Poissona:

a) oto wynik testu dla podpunktu a (Rstudio):

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowa and pois_rand
D = 0.38359, p-value < 2.2e-16
alternative hypothesis: two-sided
```

mamy więc podstawy do odrzycenia hipotezy zerowej.

b) a oto testu dla podpunktu b (Rstudio):

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowb and pois_rand
D = 0.41643, p-value < 2.2e-16
alternative hypothesis: two-sided
```

tu również mamy podstawy do odrzucenia hipotezy zerowej.

Nie mamy więc do czynienia z rozkładem Poissona.

Sprawdźmy więc czy jest to ujemny rozkład dwumianowy:

a)

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowa and bin_rand
D = 0.30185, p-value < 2.2e-16
alternative hypothesis: two-sided
```

b)

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowb and bin_rand
D = 0.13049, p-value < 2.2e-16
alternative hypothesis: two-sided
```

mamy podstawy do odrzucenia hipotezy zerowej.

Sprawdźmy czy jest to rozkład wykładniczy:

a)

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowa and exp_rand
D = 0.14619, p-value < 2.2e-16
alternative hypothesis: two-sided
```

b)

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowb and exp_rand
D = 0.09837, p-value < 2.2e-16
alternative hypothesis: two-sided
```

mamy podstawy do odrzucenia hipotezy zerowej.

Sprawdźmy jeszcze rozkład normalny:

a)

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowa and norm_rand
D = 0.14067, p-value < 2.2e-16
alternative hypothesis: two-sided
```


b)

```
        Two-sample Kolmogorov-Smirnov test

data:  liczbykrokowb and norm_rand
D = 0.17268, p-value < 2.2e-16
alternative hypothesis: two-sided
```

w obu podpunktach mamy podstawy do odrzucenia hipotezy zerowej.

Na podstawie analizy statystycznej, nasze dane nie mają żadnego ze znanych rozkładów.

4. Empiryczne przedziały ufności.

Definicja:

Niech cecha X ma rozkład w populacji z nieznanym parametrem θ. Z populacji wybieramy próbę losową ($x_1, x_2,\ldots,x_n$). Przedziałem ufności o współczynniku 1-α nazywamy taki przedział (θ_1, θ_2), który spełnia warunek: $P(\theta_1 < \theta < \theta_2)=1-\alpha$, gdzie θ_1 i θ_2 są funkcjami wyznaczonymi na podstawie próby losowej.

Sprawdzam czy lista 1000 wartości oczekiwanych ma rozkład normalny za pomocą testu Kołmogorowa-Smirnowa w RStudio

a)

```
        Two-sample Kolmogorov-Smirnov test

data:  awartoscioczekiwane and norm_rand
D = 0.02101, p-value = 0.7746
alternative hypothesis: two-sided
```

b)

```
        Two-sample Kolmogorov-Smirnov test

data:  bwartoscioczekiwane and norm_rand
D = 0.02113, p-value = 0.7686
alternative hypothesis: two-sided
```

W obu przypadkach nie mamy podstaw do odrzucenia hipotezy, że jest to rozkład normalny.

Do szacowania przedziału ufności dla średniej, gdy nie znamy rozkładu populacji generalnej, a liczebność próbki n jest duża stosujemy statystykę:

$$Z = \frac{m - \vartheta}{S} * \sqrt{n}$$

gdzie:

ϑ – średnia

m – średnia z próbki

S – odchylenie standardowe z próbki

n – liczebność próbki

Przedziały ufności dla średnich wartości oczekiwanych na poziomie ufności 0.99 (RStudio, plik *wartoscioczekiwane.R*):

a) (22.0495, 22.0586) – wynik 22.054 zawiera się w przedziale ufności.

b) (17.4717, 17.4806) – wynik 17.4761 zawiera się w przedziale ufności.

4. Podsumowanie.

Rozwiązanie problemu przy pomocy symulacji w języku Python:

a) Oczekiwana liczba ruchów, aby pająki spotkały się w jednym węźle sieci, jeśli pająki znajdują się na przeciwległych wierzchołkach sieci najbardziej zewnętrznej pętli: 22.054.

b) Oczekiwana liczba ruchów, aby pająki spotkały się w jednym węźle sieci, jeśli jeden pająk znajduje się w centrum sieci a drugi w jednym z węzłów sieci na najbardziej zewnętrznej pętli: 17.4761.

Ponieważ rozkład liczb kroków, które pająki musiały wykonać aby się spotkać nie przypomina żadnego ze znanych rozkładów, nie możemy wyliczyć wartości oczekiwanych i porównać ich do tych z symulacji oraz obliczyć błędu.

Po powtórzeniu symulacji 1000 razy mogliśmy jednak, dla każdego podpunktu wyliczyć średnią wartość oczekiwaną oraz sprawdzić, czy znajduję się ona w empirycznym przedziale ufności.

Średnie wartości oczekiwane znalazły się w przedziałach ufności, możemy więc przyjąć, że wyniki są wiarygodne.