Wprowadzenie do rachunku prawdopodobieństwa

1. Lista wszystkiego, co może się wydarzyć (**stany świata**)
	1. Każdy stan musi być wyczerpującym opisem wszystkiego, co jest ważne w danym problemie decyzyjnym
2. **Zdarzenie** to jakaś kolekcja (zbiór) stanów świata
	1. Zdarzenie może się wydarzyć bądź nie (oznaczać będziemy kolejno jako lub )
	2. Zdarzenia specjalne:
		1. Zdarzenie pewne
		2. Zdarzenie niemożliwe
		3. Zdarzenia jednostkowe
3. Zdefiniowane są **działania**:
	1. Przecięcie: (zachodzi jednocześnie zdarzenie i )
	2. Suma: (zachodzi zdarzenie lub[[1]](#footnote-1) zdarzenie )
	3. Komplement: (nie zachodzi zdarzenie )

**Przykład:** Populacja 1000 pracowników pewnej branży przemysłu chemicznego:

Stany świata: , gdzie

* CR miał kontakt z chemikaliami i zachorował na raka
* NCR nie miał kontaktu z chemikaliami i zachorował na raka
* CNR miał kontakt z chemikaliami i nie zachorował na raka
* NCNR nie miał kontaktu z chemikaliami i nie zachorował na raka

Zdarzenia:

: zdarzenie pewne, dana osoba należy do badanej populacji

: zdarzenie puste lub niemożliwe

Działania:

1. **Prawdopodobieństwo** to nieujemna liczba przyporządkowana zdarzeniom, która mierzy ich prawdopodobność zdarzenia się
	1. Suma prawdopodobieństw:
	2. Suma prawdopodobieństw zdarzeń rozłącznych
	3. **Prawdopodobieństwo warunkowe**:
	4. Dwa zdarzenia są **niezależne**, jeśli
	5. **Prawdopodobieństwo całkowite**:

Stąd, jeśli oznaczymy , to

Czyli jeśli , to

A jeśli , to

* 1. **Twierdzenie Bayesa**:

**Przykład**

Niech prawdopodobieństwa określone będą poprzez liczebność poszczególnych kategorii:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *R* | *NR* |  |
| *C* | 220 | 135 | 355 |
| *NC* | 48 | 597 | 645 |
|  | 268 | 732 | 1000 |

A zatem prawdopodobieństwa odpowiednich stanów świata oraz zdarzeń wynoszą:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *R* | *NR* |  |
| *C* | 0,220 | 0,135 | 0,355 |
| *NC* | 0,048 | 0,597 | 0,645 |
|  | 0,268 | 0,732 | 1,000 |

Możemy zweryfikować część reguł rachunku prawdopodobieństwa:

* Prawdopodobieństwo sumy zdarzeń:

Prawdopodobieństwo tego, że pracownik miał kontakt z chemikaliami lub zachorował na raka jest równe 0,403

* Prawdopodobieństwo warunkowe:

Prawdopodobieństwo tego, że pracownik zachorował na raka pod warunkiem, że miał kontakt z chemikaliami jest dużo wyższe niż bezwarunkowe prawdopodobieństwo zachorowania na raka, które z kolei jest dużo wyższe niż prawdopodobieństwo zachorowania na raka pod warunkiem, że pracownik nie miał kontaktu z chemikaliami. Wniosek jest taki, że zachorowanie na raka i kontakt z chemikaliami nie są niezależne: te dwa zdarzenia są pozytywnie ze sobą skorelowane[[2]](#footnote-2).

* Prawdopodobieństwo całkowite:
* Prawdopodobieństwo warunkowe:

Prawdopodobieństwo tego, że pracownik miał kontakt z chemikaliami pod warunkiem tego, że jest chory na raka wynosi 0,821

1. **Zmienna losowa** to funkcja przyporządkowująca stanom świata wartości liczbowe.
	1. Zmienne losowe X są charakteryzowane za pomocą **rozkładów prawdopodobieństwa**, czyli listę wartości zmiennych losowych wraz z przyporządkowanymi tym wartościom prawdopodobieństwom
	2. Mając dwie zmienne losowe możemy określić ich **łączny rozkład prawdopodobieństwa**: listę par wartości dwóch rozkładów wraz z przyporządkowanymi im prawdopodobieństwom
	3. Mając łączny rozkład prawdopodobieństwa dwóch zmiennych losowych X i Y, osobne rozkłady zmiennej X i Y nazywamy **rozkładami krańcowymi**
	4. Zmienne losowe X i Y są **niezależne**, jeśli wiedza o jednej zmiennej nie pomaga nam w przewidzeniu drugiej zmiennej:
		1. Niezależność zmiennych losowych odpowiada niezależności zdarzeń: jakiekolwiek zdarzenie zdefiniowane w ramach jednej zmiennej losowej jest niezależne od jakiegokolwiek zdarzenia zdefiniowanego w ramach drugiej zmiennej losowej
		2. Zmienne losowe są niezależne, jeśli ich łączny rozkład jest iloczynem ich rozkładów krańcowych
		3. Szczególny przypadek to zmienne IID: to są **zmienne z identycznym i niezależnym rozkładem**, np. n rzutów kostką do gry
2. **Wartość oczekiwana zmiennej losowej** to średnia ważona jej wartości (wagi są równe prawdopodobieństwom tych wartości)
	1. Jeśli wraz z odpowiadającym rozkładem prawdopodobieństwa , to
3. **Wariancja zmiennej losowej** to wartość oczekiwana odchyleń kwadratowych od średniej
	1. Jeśli wraz z odpowiadającym rozkładem prawdopodobieństwa , to

Odchylenie standardowe to pierwiastek wariancji

**Przykład**:

Departament księgowy firmy przeprowadził rachunek opłacalności pracy pracowników. Wnioski z raportu są następujące: ponieważ największa grupa pracowników (597 osób) to NCNR to wobec tego względne zyski z pracy różnych grup pracowników zostały ocenione względem tej największej grupy. Okazało się, że praca osoby mającej kontakt z chemikaliami generuje dla firmy zysk o 1000 złotych większy niż praca osoby nie mającej kontaktu z chemikaliami. Jednak w sytuacji, gdy którykolwiek pracownik zachoruje na raka, firma ponosi dodatkowe koszty związane z utratą pracownika w wysokości 1000 złotych. Względny zysk firmy z pracy jednego pracownika przedstawia zmienna losowa X:

|  |  |  |
| --- | --- | --- |
|  | Wartość zmiennej X | Prawdopodobieństwo |
|  | -1000 | 0,048 |
|  | 0 | 0,817 |
|  | +1000 | 0,135 |

Wartość oczekiwana zmiennej X wynosi:

Wartość oczekiwana w wysokości 87 informuje nas, że w obecnej sytuacji firma średnio zarabia na pracy przeciętnego pracownika o 87 złotych więcej niż w sytuacji gdyby wszyscy pracownicy należeli do największej grupy pracowników NCNR (jest to hipotetyczna sytuacja, w której wszyscy pracownicy pracują bez kontaktu z chemikaliami i jednocześnie żaden z pracowników nie zachoruje na raka – hipotetyczna, bo przecież nawet ci, którzy nie mają kontaktu z chemikaliami mogą zachorować na raka). Wniosek, jaki firma wyprowadziła z przeprowadzonej analizy jest taki, że aby podnieść zyski firma powinna skierować więcej pracowników do pracy z chemikaliami.
Odchylenie standardowe zmiennej X wynosi:

Odchylenie standardowe informuje nas o „typowym” odchyleniu od wartości średniej, którego możemy się spodziewać w tej sytuacji.

Załóżmy, że NFZ otrzymuje od wszystkich pracowników składkę w wysokości 200 złotych. Jeśli któryś z pracowników choruje na raka, to NFZ musi zwrócić koszty leczenia w wysokości średnio 1000 złotych. Bilans pieniężny NFZ dla grupy pracowników rozważanej firmy jest opisany za pomocą zmiennej losowej Y:

|  |  |  |
| --- | --- | --- |
|  | Wartość zmiennej Y | Prawdopodobieństwo |
|  | -800 | 0,268 |
|  | +200 | 0,732 |

Wartość oczekiwana zmiennej Y wynosi:

A zatem w przypadku bilansu ubezpieczenia pracowników tej firmy NFZ odnotowuje średnio na każdym z nich 68 złotych straty.
Odchylenie standardowe wynosi:

Łączny rozkład prawdopodobieństwa tych dwóch zmiennych to:

|  |  |  |  |
| --- | --- | --- | --- |
|  | -800 | 200 | Rozkład X |
| -1000 | 0,048 | 0,000 | 0,048 |
| 0 | 0,220 | 0,597 | 0,817 |
| 1000 | 0,000 | 0,135 | 0,135 |
| Rozkład Y | 0,268 | 0,732 | 1,000 |

Zmienne X i Y nie są niezależne, ponieważ prawdopodobieństwa rozkładu łącznego nie są iloczynem rozkładów krańcowych.

Zauważmy, że ponieważ rozkład X i Y nie są niezależne, to jeśli znamy wartość jednej zmiennej to zmieniają się prawdopodobieństwa poszczególnych wartości drugiej zmiennej.

Np.

Jeśli zmienne X i Y byłyby niezależne to ich rozkład powinien wyglądać następująco:

|  |  |  |  |
| --- | --- | --- | --- |
|  | -800 | 200 | Rozkład X |
| -1000 | 0,013 | 0,035 | 0,048 |
| 0 | 0,219 | 0,598 | 0,817 |
| 1000 | 0,036 | 0,099 | 0,135 |
| Rozkład Y | 0,268 | 0,732 | 1,000 |

Załóżmy teraz, że Państwu udało się zmusić firmę, aby wszyscy jej pracownicy nie mieli kontaktu z chemikaliami. Wówczas tabela liczebności wyglądałaby następująco:

|  |  |  |
| --- | --- | --- |
| **Pierwotna sytuacja** |  | **Nowa sytuacja 1** |
|  | *R* | *NR* |  |  |  | *R* | *NR* |  |
| *C* | 220 | 135 | 355 |  | *C* | 0 | 0 | 0 |
| *NC* | 48 | 597 | 645 |  | *NC* | 74 | 926 | 1000 |
|  | 268 | 732 | 1000 |  |  | 74 | 926 |  |

Liczebności w nowej sytuacji otrzymujemy przy założeniu, że ponieważ nikt nie ma kontaktu z chemikaliami, to prawdopodobieństwo zachorowania na raka dla wszystkich jest równe prawdopodobieństwu warunkowemu (obliczone na podstawie sytuacji pierwotnej) zachorowania na raka pod warunkiem braku kontaktu z chemikaliami .

W nowej sytuacji zmienia się rozkład prawdopodobieństwa zamiennej losowej X i Y. Nazwijmy nowe zmienne losowe X’ i Y’:

|  |  |  |
| --- | --- | --- |
|  | Wartość zmiennej X’ | Prawdopodobieństwo |
|  | -1000 | 0,074 |
|  | 0 | 0,926 |
|  | +1000 | 0,000 |

|  |  |  |
| --- | --- | --- |
|  | Wartość zmiennej Y’ | Prawdopodobieństwo |
|  | -800 | 0,074 |
|  | +200 | 0,926 |

Nowe wartości oczekiwane i wariancje zmiennych to:

A teraz załóżmy, że to firmie udało się „przenieść” wszystkich pracowników do pracy w kontakcie z chemikaliami.

|  |  |  |
| --- | --- | --- |
| **Pierwotna sytuacja** |  | **Nowa sytuacja 2** |
|  | *R* | *NR* |  |  |  | *R* | *NR* |  |
| *C* | 220 | 135 | 355 |  | *C* | 620 | 380 | 1000 |
| *NC* | 48 | 597 | 645 |  | *NC* | 0 | 0 | 0 |
|  | 268 | 732 | 1000 |  |  | 620 | 380 | 1000 |

Liczebności w nowej sytuacji otrzymujemy przy założeniu, że ponieważ wszyscy mają kontakt z chemikaliami, to prawdopodobieństwo zachorowania na raka dla wszystkich jest równe prawdopodobieństwu warunkowemu (obliczone na podstawie sytuacji pierwotnej) zachorowania na raka pod warunkiem kontaktu z chemikaliami .

W nowej sytuacji zmienia się rozkład prawdopodobieństwa zamiennej losowej X i Y. Nazwijmy nowe zmienne losowe X’’ i Y’’:

|  |  |  |
| --- | --- | --- |
|  | Wartość zmiennej X’’ | Prawdopodobieństwo |
|  | -1000 | 0,000 |
|  | 0 | 0,620 |
|  | +1000 | 0,380 |

|  |  |  |
| --- | --- | --- |
|  | Wartość zmiennej Y’’ | Prawdopodobieństwo |
|  | -800 | 0,620 |
|  | +200 | 0,380 |

Nowe wartości oczekiwane i wariancje zmiennych to:

Podsumowanie:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Obecna sytuacja | Bez chemikaliów | Tylko chemikalia |
| wartość oczekiwana względnego zysku firmy  | 87 | -74 | 380 |
| wartość oczekiwana bilansu NFZ  | -68 | 126 | -420 |

1. **Rozkład normalny** to rodzina rozkładów sparametryzowanych za pomocą średniej i odchylenia standardowego
	1. Pełni ważną rolę w statystyce z powodu centralnego twierdzenia granicznego
2. **Prawo wielkich liczb** mówi, że jeśli obserwujesz długi ciąg zmiennych o identycznych i niezależnym rozkładach, to ich średnia będzie coraz bliższa ich wspólnej wartości oczekiwanej
	1. Np. jeśli rzucasz kostką do gry wiele razy, to średnia częstość szóstek będzie coraz bliższa 1/6

|  |  |
| --- | --- |
| **Liczba rzutów** | **Częstość szóstek** |
| 10 | 0,6 |
| 100 | 0,15 |
| 1000 | 0,158 |
| 10000 | 0,1652 |
|  | 0,16666667 |



1. **Centralne twierdzenie graniczne** mówi, że jeśli spojrzeć na średnią n identycznych i niezależnych zmiennych losowych (parę mniej istotnych warunków musi być spełnione), to ta średnia ma rozkład coraz bardziej zbliżony do rozkładu normalnego
	1. Dzięki prawu wielkich liczb wiemy, że średnia zbliżona jest do wartości oczekiwanej, a dzięki centralnemu twierdzeniu wiemy jaki jest rozkład tej średniej i możemy określić błąd jaki możemy zrobić
	2. Np. rozkład średniej liczby oczek w n rzutach kostką szybko dąży do rozkładu normalnego:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 rzut | 2 rzuty | 3 rzuty | 4 rzuty |
|  |  |  |  |

1. Słowo „lub” użyte jest tutaj w znaczeniu łącznym, tj. może zachodzić jedno zdarzenie, drugie zdarzenie oraz oba naraz. [↑](#footnote-ref-1)
2. Korelacja daje podstawy do stwierdzenia, że dwa zjawiska idą ze sobą w parze, ale nie daje podstaw do wskazania związku przyczynowo-skutkowego pomiędzy tymi zjawiskami. Do wskazania takiego związku potrzebne są dodatkowe informacje: logika i reguły zwykłego następstwa zdarzeń, informacje fachowe na temat natury tych zjawisk, wyobraźnia. [↑](#footnote-ref-2)