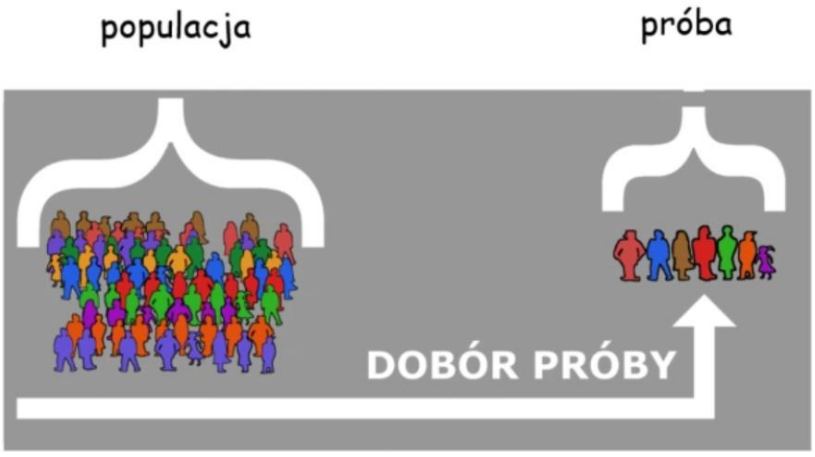
STATYSTYKA
vs.
PRAWDOPODOBIEŃSTWO

Statystyka i prawdopodobieństwo



www.agh.edu.pl

Statystyka i prawdopodobieństwo

www.agh.edu.pl

Statystyka i prawdopodobieństwo

PRAWDOPODOBIEŃSTWO

Jeśli **wiemy, że istnieje** 30 czerwonych skarpet, 20 niebieskich i **50 czarnych**, możemy użyć **prawdopodobieństwa**, aby **odpowiedzieć na pytania** dotyczące **składu losowej próbki** tych skarpet.

Jeśli natomiast **nie mamy wiedzy o rodzajach skarpetek** w szufladzie, to wkraczamy w sferę **statystyki**. **Statystyki** pomagają nam **wnioskować** o **właściwościach populacji** na podstawie **losowej próby**.

STATYSTYKA

www.agh.edu.pl

Statystyka i prawdopodobieństwo

PRAWDOPODOBIEŃSTWO

Pytania tego typu to:

- ✓ „**Jakie jest prawdopodobieństwo**, że wyciągniemy z szuflady dwie niebieskie i dwie czerwone?"
- ✓ „**Jakie jest prawdopodobieństwo**, że wyciągniemy 3 skarpetki i będziemy mieć pasującą parę?"
- ✓ „**Jakie jest prawdopodobieństwo**, że wylosujemy pięć skarpetek **ze zwracaniem** i wszystkie są czarne?"

Pytania o charakterze statystycznym to:

- ✓ Losowe pobranie 10 skarpetek z szuflady, dało: 1 niebieską, 4 czerwone i 5 **czarnych**. **Jaki jest całkowity udział czarnych, niebieskich i czerwonych skarpet w szufladzie?**
- ✓ Próbujemy losowo 10 skarpetek z szuflady, zapisujemy **liczbę czarnych**, a następnie **wkładamy je z powrotem** do szuflady. Ten proces jest wykonywany **pięć razy**. **Średnia liczba** skarpetek czarnych w każdej z tych prób wynosi 7. **Jaka jest liczba czarnych skarpetek w szufladzie?**

STATYSTYKA

Statystyka i prawdopodobieństwo

Prawdopodobieństwo i statystyka mają ze sobą **wiele wspólnego**.

Statystyki są **zbudowane na podstawie** prawdopodobieństwa.

Chociaż zazwyczaj **nie mamy pełnych informacji o populacji**, możemy **użyć twierdzeń i wyników dotyczących prawdopodobieństwa**, aby uzyskać wyniki statystyczne. Te wyniki informują nas o populacji.

www.agh.edu.pl

AGH

ZDARZENIA LOSOWE

AGH

Zdarzenia losowe

- **Doświadczenie** (zdarzenie) nazywamy **losowym**, jeżeli pomimo precyzyjnych warunków, w których jest ono realizowane – nie jesteśmy w stanie przewidzieć jego wyników.
- W każdym doświadczeniu losowym można wyróżnić najprostsze, nierozkładalne zdarzenia (wyniki doświadczenia), które nazywamy **zdarzeniami elementarnymi.**
- Zbiór wszystkich zdarzeń elementarnych związanych z dowolnym doświadczeniem nazywamy **przestrzenią zdarzeń elementarnych, Ω.** Każdy podzbiór przestrzeni zdarzeń elementarnych nazywamy **zdarzeniem losowym, ω.**

Doświadczenie - zdarzenia – definiowanie przestrzeni zdarzeń– tworzenie modelu

Doświadczenie: **egzamin**

Opis dowolnego zdarzenia losowego, jakie może mieć miejsce w danym doświadczeniu:

Doświadczenie - zdarzenia – definiowanie przestrzeni zdarzeń– tworzenie modelu

Doświadczenie: **egzamin**

Zdarzenie: **ocena z egzamin**

1 4 3 5- 4- 6

Opis zbioru zdarzeń elementarnych:

$$\Omega = \{2, 3, 3.5, 4, 4.5, 5\}; \qquad \#\,\Omega = 6$$

Opis dowolnego zdarzenia losowego, jakie może mieć miejsce w danym doświadczeniu:

- ✓ A: egzamin oblany: $A = \{2\}$
- ✓ B: egzamin zdany = uzyskanie oceny co najmniej 3: $B = \{3, 3.5, 4, 4.5, 5\}$
- ✓ C: wynik egzaminu satysfakcjonujący = uzyskanie oceny co najmniej 4 $C = \{4, 4.5, 5\}$

www.agh.edu.pl

Zdarzenia, przestrzeń zdarzeń – formalizacja opisu

- Niech ω_i oznacza jeden z możliwych wyników prowadzonego doświadczenia (eksperymentu)
- $\omega_i \in \Omega$, ω_i jest elementem zbioru
- $\Omega = \{\omega_1, \omega_2 \ldots \ldots \omega_n\}, \quad \#\Omega = n$
- Zbiór zdarzeń elementarnych, zawiera wszystkie możliwe wyniki danego doświadczenia (eksperymentu)
- Ω może być zbiorem skończonym albo zbiorem nieskończonym, to zależy od doświadczenia i liczby możliwych wyników

Prawdopodobieństwo - Aksjomatyczna definicja prawdopodobieństwa

Zakładamy, że: A jest zdarzeniem losowym: tzn. $A \subset \Omega$

Prawdopodobieństwo P jest funkcją :

$$P: A \rightarrow P(A)$$

spełniającą następujące aksjomaty:

1. $P(A) \in [0,1]$
2. $P(\Omega) = 1 \qquad P(\varnothing)=0$
3. $P(A \cup B) = P(A)+P(B)$ jeśli $A \cap B = \varnothing$

albo

3.' $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Definicje prawdopodobieństwa (rachunkowe)

Ω jest zbiorem wszystkich zdarzeń elementarnych,
(rozłącznych i jednakowo możliwych)

A ⊂ Ω, A jest zdarzeniem losowym

Prawdopodobieństwo zajścia zdarzenia A → $P(A) = \frac{n(A)}{n(\Omega)}$ (Liczba zdarzeń sprzyjających zajściu A / Liczba możliwych zdarzeń)

Klasyczna definicja - wzór Laplace'a

$$P(A) = \frac{\overline{\overline{A}}}{\overline{\overline{\Omega}}} = \frac{liczba\ zdarzeń\ elementarnych\ sprzyjajacych\ zdarzeniu\ A}{liczba\ wszystkich\ mozliwych\ zdarzen\ elementarnych}$$

Geometryczna

$$P(A) = \frac{\mu A}{\mu \Omega} = \frac{miara\ geometryczna\ zbioru\ A}{miara\ geometryczna\ zbioru\ \Omega}$$

Statystyczna

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} = \frac{liczba\ zaobserwowanych\ zdarzenA}{liczba\ przeprowadzonych\ obserwacji}$$

Probabilistyczne modele danych
Zmienne losowe

Zmienna losowa jest to funkcja rzeczywista X,

określona na zbiorze zdarzeń elementarnych Ω

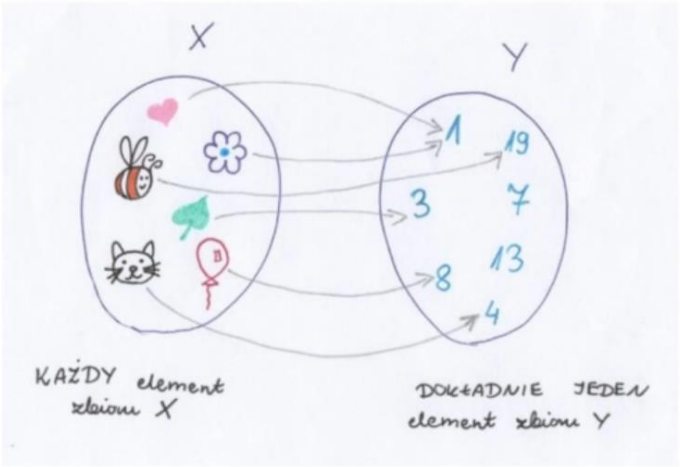
$X: \Omega \to W$

Zmienne losowe zwykle oznacza się dużymi literami z końca alfabetu : X, Y, Z.

Wartości zmiennych losowych zwykle oznacza się małymi literami z końca alfabetu: x,y,z.

Probabilistyczne modele danych
Zmienne losowe

Definiowanie zmiennej losowej to **przypisanie wartości** (liczbowych) zdarzeniom elementarnym.



https://www.statystyczny.pl/zmienna-losowa/

www.agh.edu.pl

Probabilistyczne modele danych
Zmienne losowe dyskretne i ciągłe

ZMIENNA LOSOWA DYSKRETNA

Rozkład prawdopodobieństwa

ZMIENNA LOSOWA CIĄGŁA

Funkcja gęstości

Dystrybuanta

Charakterystyki liczbowe zmiennych losowych

- ✓ Wartość oczekiwana EX
- ✓ Wariancja D^2X
- ✓ Odchylenie standardowe DX

Probabilistyczne modele danych
Zmienne losowe

Z partii wyrobów zawierającej wyroby dobre i wyroby wadliwe losuję jeden wyrób, wtedy

$$\Omega = \{\omega_d, \omega_w\}$$

gdzie

ω_d - oznacza wylosowanie wyrobu dobrego

ω_w - oznacza wylosowanie wyrobu wadliwego

Określam zmienną losową X w następujący sposób:

$$X(\omega_d) = 1 \qquad X(\omega_w) = 0$$

Probabilistyczne modele danych
Rozkład prawdopodobieństwa zmiennej losowej dyskretnej

Jeżeli w przedstawionym przykładzie, dotyczącym kontroli jakości wyrobów, 90% wyrobów było dobrych, natomiast 10% stanowiły wybraki, to możemy mówić o prawdopodobieństwie zdarzeń:

$$P(\{\omega : X(\omega) = 0\}) = 0{,}1$$
$$P(\{\omega : X(\omega) = 1\}) = 0{,}9$$

Zdefiniowaliśmy rozkład prawdopodobieństwa

Tablicowy zapis rozkładu prawdopodobieństwa zmiennej losowej X		
x_i	0	1
p_i	0,1	0,9

Rozkład prawdopodobieństwa zmiennej losowej X jest zbiorem par {x, p}, gdzie x jest wartością zmiennej X, p- prawdopodobieństwem wystąpienia wartości x.

Probabilistyczne modele danych

Dystrybuanta zmiennej losowej dyskretnej

Dystrybuantą, $F_X(x_0)$, zmiennej losowej X jest prawdopodobieństwo zdarzenia, polegającego na tym, że zmienna ta przyjmie wartości mniejsze od x_0.

$$F_X(x_0) = P(X < x_0)$$

Dystrybuantę zmiennej losowej X oznaczamy zwykle jako F_X

$$F_X(x_0) = P_X((-\infty, x_0)) = P(X < x_0)$$

Dystrybuanta jest funkcją:

- określoną na zbierze liczb rzeczywistych;
- o wartościach z przedziału [0-1];
- niemalejącą
- prawostronnie ciągłą

www.agh.edu.pl

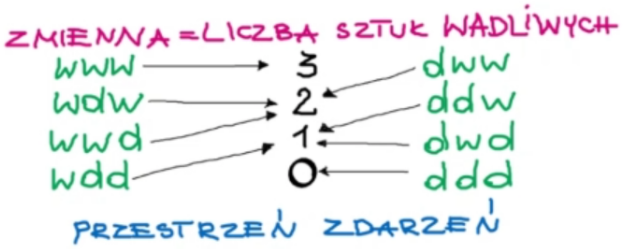
Probabilistyczne modele danych

Wyznaczanie rozkładu zmiennej losowej dyskretnej

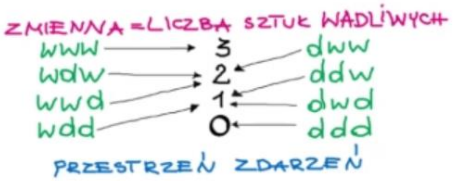
Z partii wyrobów losujemy 3 sztuki

Na rysunku przedstawiono:

- ✓ Przestrzeń możliwych zdarzeń
- ✓ Sposób określenia zmiennej losowej



Probabilistyczne modele danych
Wyznaczanie rozkładu zmiennej losowej dyskretnej

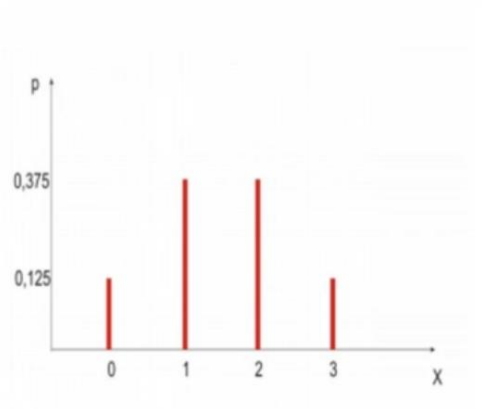


Rozkład prawdopodobieństwa zmiennej losowej

$$p_1 = P(X = 0) = \frac{1}{8}$$
$$p_2 = P(X = 1) = \frac{3}{8}$$
$$p_3 = P(X = 2) = \frac{3}{8}$$
$$p_4 = P(X = 3) = \frac{1}{8}$$

Probabilistyczne modele danych
Wyznaczanie rozkładu zmiennej losowej dyskretnej

$$p_1 = P(X = 0) = \frac{1}{8}$$
$$p_2 = P(X = 1) = \frac{3}{8}$$
$$p_3 = P(X = 2) = \frac{3}{8}$$
$$p_4 = P(X = 3) = \frac{1}{8}$$



Rozkład prawdopodobieństwa zmiennej losowej				
x_i	0	1	2	3
p_i	1/8	3/8	3/8	1/8

www.agh.edu.pl

Probabilistyczne modele danych

Dystrybuanta rozkładu zmiennej losowej dyskretnej

Rozkład prawdopodobieństwa zmiennej losowej				
x_i	0	1	2	3
p_i	1/8	3/8	3/8	1/8

Jakie będzie prawdopodobieństwo wylosowania mniej niż 2 wadliwych sztuk?

$$F_X(x_0) = P_X((-\infty, x_0)) = P(X < x_0)$$

$$F(2) = P(X < 2) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$$

AGH

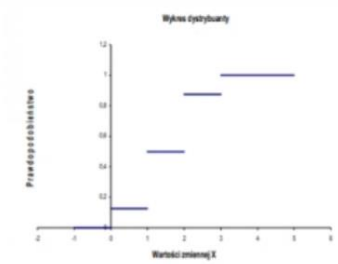
Probabilistyczne modele danych

Wyznaczanie dystrybuanty zmiennej losowej dyskretnej

Rozkład prawdopodobieństwa zmiennej losowej				
x_i	0	1	2	3
p_i	1/8	3/8	3/8	1/8
$F(x)$	0	1/8	1/2	7/8

Dystrybuanta

$$F_X(x_0) = P_X((-\infty, x_0)) = P(X < x_0)$$



$$F_X(0) = P_X((-\infty, 0)) = P(X < 0) = 0$$
$$F_X(1) = P_X((-\infty, 1)) = P(X < 1) = P(X = 0) = 1/8$$
$$F_X(2) = P_X((-\infty, 2)) = P(X < 2) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8}$$
$$F_X(3) = P_X((-\infty, 3)) = P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8$$
$$F_X(4) = P_X((-\infty, 4)) = P(X < 4) = 1$$

www.agh.edu.pl

Probabilistyczne modele danych
Parametry rozkładu zmiennej losowej dyskretnej

Znamy rozkład prawdopodobieństwa wylosowania sztuki wadliwej, ale nie wiemy, jakiego wyniku tak naprawdę możemy się spodziewać.

Odpowiedź na to pytanie da nam **wartość oczekiwana**, która jest jednym z **parametrów rozkładu prawdopodobieństwa**, który mówi o tym, jakiej **przeciętnej wartości zmiennej losowej** należy się spodziewać w długiej serii jej realizacji.

AGH

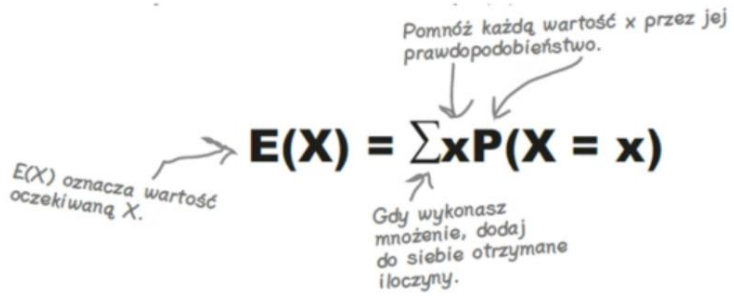
Probabilistyczne modele danych
Parametry rozkładu zmiennej losowej dyskretnej

Wartość oczekiwana $E(X)$ zmiennej losowej X przypomina nieco średnią arytmetyczną, tyle że jest wyznaczana na podstawie rozkładu prawdopodobieństwa.

$$E(X) = \sum_{i=0}^{n} x_i p_i$$

Wartość oczekiwaną liczymy mnożąc wartość każdej realizacji przez jej prawdopodobieństwo, a następnie sumujemy otrzymane iloczyny.

Probabilistyczne modele danych
Parametry rozkładu zmiennej losowej dyskretnej



x_i	0	1	2	3	$\sum$
p_i	1/8	3/8	3/8	1/8	1
$x_i \cdot p_i$	0	3/8	6/8	3/8	12/8=1,5

$$E(X) = 1{,}5$$

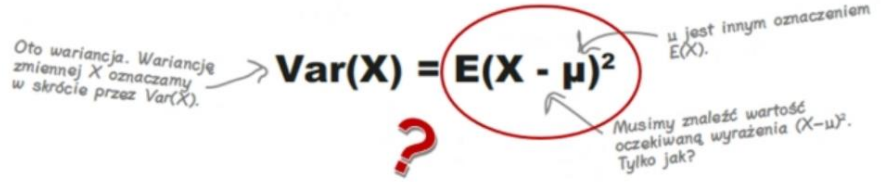
Probabilistyczne modele danych
Parametry rozkładu zmiennej losowej dyskretnej

Wartość oczekiwana $E(X)$ pozwala oszacować typową, przeciętną realizację zmiennej losowej, ale nie mówi nic o tym, jak bardzo jej wartości mogą się zmieniać.

WARIANCJA/ODCHYLENIE STANDARDOWE

www.agh.edu.pl

Probabilistyczne modele danych
Parametry rozkładu zmiennej losowej dyskretnej



$$Var(X) = D^2(X) = E[X - E(X)]^2 = E(X^2) - E^2(X) = \sum_{i=1}^{n} (xi - E(X)^2$$

$$D(X) = \sqrt{E(X^2) - E^2(X)}$$

Probabilistyczne modele danych
Parametry rozkładu zmiennej losowej dyskretnej

$$D^2(X) = E(X^2) - E^2(X)$$

x_i	0	1	2	3	
p_i	1/8	3/8	3/8	1/8	1
$x_i \cdot p_i$	0	3/8	6/8	3/8	12/8=1,5
$x_i^2 p_i$	0	3/8	12/8	9/8	3

$E(X) = 1{,}5$

$E(X^2) = 3$

$$D^2(X) = 3 - (1{,}5)^2 = 3 - 2{,}25 = 0{,}75$$

$$D(X) = \sqrt{0{,}75} = 0{,}87$$

Probabilistyczne modele danych
Rozkład zmiennej losowej ciągłej

Zmienną losową, która teoretycznie może przyjąć **każdą wartości z określonego przedziału** nazywamy **zmienną losową ciągłą**.

Zmiennymi losowymi ciągłymi są na przykład:
- ✓ długość łodygi lnu, $X \in [0, 120]$ (cm),
- ✓ ciężar jaja przepiórki, $X \in [0, 30]$ (g),
- ✓ temperatura ciała człowieka zdrowego, $X \in [35, 37]$ (stopni C).

Probabilistyczne modele danych
Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

Typ danych, jakie posiadamy, ma wpływ na sposób obliczania prawdopodobieństwa.

Prawdopodobieństwo, że wartość zmiennej losowej znajduje się w pewnym przedziale, określa **funkcja gęstości prawdopodobieństwa**

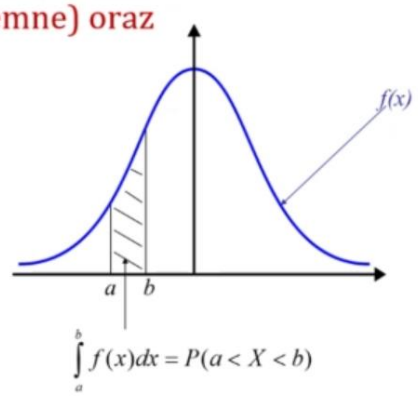
Probabilistyczne modele danych

Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

Definicja

Funkcją gęstości prawdopodobieństwa zmiennej losowej typu **ciągłego** nazywamy funkcję **f(x)**, określoną na zbiorze liczb rzeczywistych, taką że: f(x)≥0 (przyjmuje wartości nieujemne) oraz dla dowolnych a < b zachodzi

$$\int_a^b f(x)dx = P(a < X < b)$$



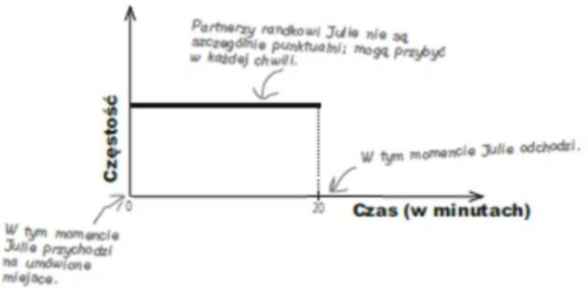
www.agh.edu.pl

Probabilistyczne modele danych

Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

Julia nie znosi czekania na osoby, z którymi się spotyka, dlatego przyjęła zasadę, że jeśli osoba nie pojawi się w ciągu 20 min, wraca do domu.

Przykładowy wykres czystości przedstawiający czas, przez jaki Julia musi czekać na rozpoczęcie każdej randki:



Źródło: Head First Statistics A Brain-Friendly

Probabilistyczne modele danych

Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

Czy czas oczekiwania zmianie się w sposób skokowy, czy ciągły?
Jak można wyznaczyć jego rozkład prawdopodobieństwa?

Nie możemy tutaj podać prawdopodobieństwa realizacji **każdej** pojedynczej wartości zmiennej losowej, ponieważ mamy do czynienia ze zmienną losową **ciągłą**.
Nawet nie potrafimy wymienić wszystkich możliwych realizacji.
Julia może czekać np. 5min, 5min i 10s, 5min i 10,5s.

Dlatego... skupimy się na obliczeniu **prawdopodobieństwa** przyjęcia przez zmienną losową, nie konkretnej/pojedynczej wartości, a wartości z określonego **przedziały** jej zmienności.

Źródło: Head First Statistics A Brain-Friendly

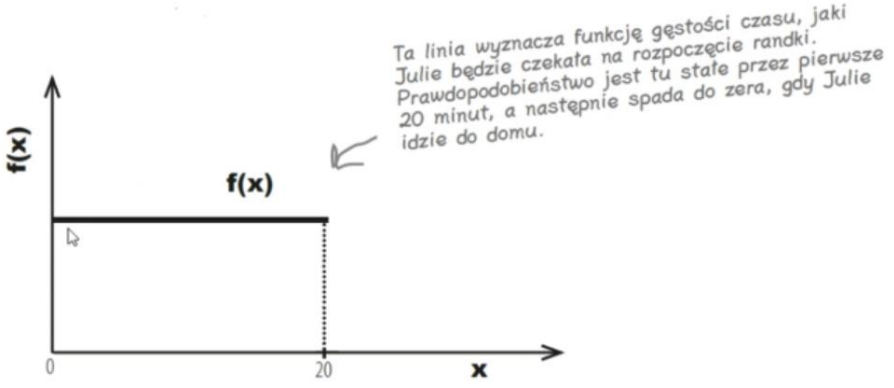
www.agh.edu.pl

Probabilistyczne modele danych

Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

Prawdopodobieństwo, że wartość zmiennej losowej znajduje się w pewnym przedziale, określa **funkcja gęstości prawdopodobieństwa.**

Od tej funkcji zależy kształt rozkładu.

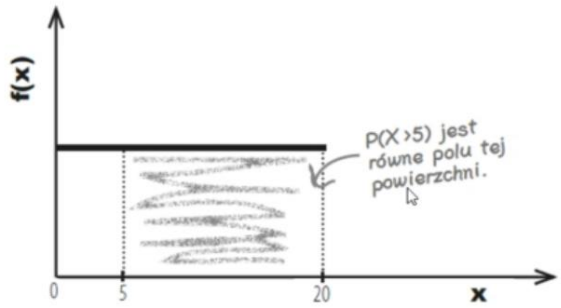


Źródło: Head First Statistics A Brain-Friendly

Znajdźmy prawdopodobieństwo, że Julia będzie czekała na swojego kolegę dłużej niż 5min ☺

czyli…

Obliczyć: $P(X > 5)$



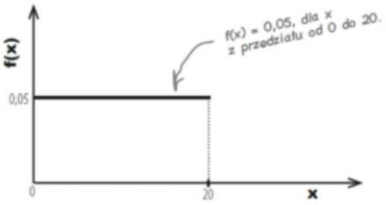
Źródło: Head First Statistics A Brain-Friendly

Probabilistyczne modele danych
Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

1. Pole pod krzywą wynosi 1
2. Znaleźć postać funkcji gęstości prawdopodobieństwa f(x)

Pole=1 $\Rightarrow$ $20 \cdot wysokość = 1 \Rightarrow wyskość = 0{,}05$

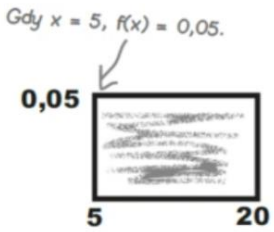
$f(x) = 0{,}05$ dla x z przedziału od 0 do 20



Źródło: Head First Statistics A Brain-Friendly

Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

$$P(X > 5) = (20 - 5) \cdot 0{,}05 = 0{,}75$$



A zatem....

Prawdopodobieństwo, że Julia będzie musiała czekać na partnera dłużej niż 5 min, wynosi 0,75.

Źródło: Head First Statistics A Brain-Friendly

Probabilistyczne modele danych

Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

- ✓ W przypadku **zmiennych losowych ciągłych** musimy obliczać **prawdopodobieństwo** jako **pole** powierzchni pod wykresem funkcji gęstości.
- ✓ Nie możemy (tak jak to miało miejsce w przypadku zmiennej losowej dyskretnej) zsumować prawdopodobieństwa każdej wartości z danego przedziału, ponieważ takich wartości jest nieskończenie wiele.

GDY MAMY DO CZYNIENIA ZE **ZMIENNYMI CIĄGŁYMI**, LICZYMY **PRAWDOPODOBIEŃSTWO** DLA WARTOŚCI **PRZEDZIAŁOWYCH**

www.agh.edu.pl

Funkcja gęstości prawdopodobieństwa zmiennej losowej ciągłej

Własności funkcji gęstości prawdopodobieństwa:

1. $f(x) \geq 0$ dla wszystkich x,
2. $\int_{-\infty}^{\infty} f(x)dx = 1,$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx.$

Dystrybuanta zmiennej losowej ciągłej

Dystrybuantą zmiennej losowej X nazywamy **funkcję**

$$F(b) = P(X \leq b).$$

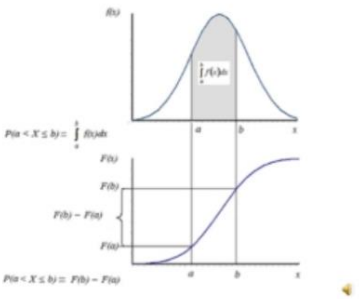
Jeśli X jest zmienną ciągłą o funkcji gęstości prawdopodobieństwa f(x), to:

$$F(b) = \int_{-\infty}^{b} f(x)dx.$$

Probabilistyczne modele danych
Związek dystrybuanty i gęstości zmiennej losowej ciągłej

Dla dowolnej funkcji f, będącej gęstością prawdopodobieństwa zachodzi zależność

$$\boldsymbol{F}(\infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$



Dla zmiennej losowej ciągłej zachodzi równość

$$P(a \le X \le b) = F(b) - F(a)$$

Stąd wynika, że:

$$P(X = x_i) = 0$$

ponieważ

$$P(X = a) = P(a \le X \le a) = F(a) - F(a) = 0$$

Probabilistyczne modele danych
Parametry rozkładu zmiennej losowej ciągłej

Wartość oczekiwaną (nadzieję matematyczną / wartość przeciętną), zmiennej losowej X oznacza się **E(X)** i określa w następujący sposób:

$$\boldsymbol{E(X)} = \int_{-\infty}^{+\infty} xf(x)dx$$

Wariancja zmiennej losowej

$$\boldsymbol{D^2(X)} = \int_{-\infty}^{+\infty} \{x - E(X)\}^2 f(x)dx$$

Odchylenie standardowe:

$$\boldsymbol{D(X)} = \sqrt{D^2(X)}$$

Wybrane rozkłady prawdopodobieństwa

ZMIENNA LOSOWA DYSKRETNA	ZMIENNA LOSOWA CIĄGŁA
1. Rozkład dwupunktowy	1. **Rozkład normalny**
2. Rozkład dwumianowy (Bernoulliego)	2. Rozkład jednostajny
3. Rozkład Poissona	3. Rozkład wykładniczy
	4. Rozkład t-studenta

Kluczowe punkty

- Prawdopodobieństwo ⇒ wiemy co jest w szufladzie;
  Statystyka ⇒ wnioskujemy o zawartości szuflady na podstawie próby;
- Statystyki są zbudowane na podstawie prawdopodobieństwa. Chociaż zazwyczaj nie mamy pełnych informacji o populacji, to używając twierdzeń i wyników dotyczących prawdopodobieństwa, możemy uzyskać wyniki statystyczne dotyczące całej populacji.
- Zmienna losowa to funkcja, która zdarzeniom elementarnym przypisuje liczby.
- Rozkład prawdopodobieństwa przypisuje każdej wartości zmiennej losowej dyskretnej prawdopodobieństwo jej realizacji.
- Gdy mamy do czynienia ze zmiennymi ciągłymi, liczymy prawdopodobieństwo dla wartości przedziałowych

AGH

www.agh.edu.pl

DZIĘKUJĘ ZA UWAGĘ ☺