

Wykład 13. Zmienne losowe typu ciągłego

dr Mariusz Grządziel

styczeń 2014

Pole trapezu krzywoliniowego

Przypomnienie: figurę ograniczoną przez:

- wykres funkcji $y = f(x)$, gdzie f jest funkcją ciągłą;
- proste $x = a$, $x = b$, $a < b$,
- oś OX (tj. prostą $y = 0$)

będziemy nazywać *trapezem krzywoliniowym* (odpowiadającym funkcji f oraz odcinkowi $[a, b]$).

Pole tej figury można przedstawić w postaci całki:

$$\int_a^b f(x)dx.$$

Pole „nieograniczonego” trapezu krzywoliniowego-całka niewłaściwa

Problem: jesteśmy zainteresowani polem figury ograniczonej: wykresem funkcji $f(x) = e^{-x}$ oraz prostymi $y = 0$, $x = 0$. Pole tego obszaru można określić jako całkę:

$$\lim_{T \rightarrow \infty} \int_0^T f(x)dx.$$

Korzystając z faktu:

$$\int_0^T e^{-x}dx = -e^{-T} + e^0 = 1 - e^{-T}$$

znajdujemy, że granica ta jest równa 1.

Całka niewłaściwa z funkcji nieujemnej

Całkę niewłaściwą z funkcji nieujemnej f na półprostej $[a, \infty)$ można określić jako granicę:

$$\lim_{T \rightarrow \infty} \int_a^T f(t)dt$$

jeśli ona istnieje.

Analogicznie można całkę niewłaściwą z funkcji nieujemnej f na półprostej $(-\infty, b]$.

Całkę niewłaściwą z funkcji nieujemnej f na prostej $(-\infty, \infty)$ można określić jako granicę $\lim_{T \rightarrow \infty} \int_{-T}^T f(t)dt$, jeśli ona istnieje.

Zmienne losowe typu ciągłego

Definicja 1. Mówimy, że zmienna losowa X jest typu ciągłego, jeśli istnieje nieujemna funkcja g taka, że dla każdych liczb a i b spełniających warunek $-\infty \leq a < b \leq \infty$ zachodzi równość

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

Rozkład jednostajny na odcinku $[0, 1]$

Przykładem zmiennej losowej typu ciągłego jest rozkład jednostajny na odcinku $[0, 1]$ (oznaczenie: $U(0, 1)$). Jego funkcja gęstości u dana jest wzorem

$$u(x) = \begin{cases} 1, & \text{jeśli } 0 \leq x \leq 1, \\ 0 & \text{jeśli } x < 0 \text{ lub } x > 1. \end{cases}$$

Rozkład ten może opisywać np. czas oczekiwania na autobus A , odjeżdżający do miejscowości B co godzinę, przez pasażera C ; zakładamy, że C nie zna rozkładu jazdy dla tej linii i że przychodzi na przystanek w losowym momencie.

Rozkład jednostajny na odcinku $[0, 1]$ — przykład obliczeń

Czas oczekiwania na autobus- zmienna losowa $Y \sim U(0, 1)$. Prawdopodobieństwo $P(\frac{1}{3} < Y < \frac{1}{2})$ jest równe:

$$P\left(\frac{1}{3} < Y < \frac{1}{2}\right) = \int_{1/3}^{1/2} 1dx = [x]_{1/3}^{1/2} = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

Prawdopodobieństwa odpowiadające nierównościami ostrym i słabym

Dla zmiennej losowej X o rozkładzie typu ciągłego mamy:

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b).$$

Równość ta wynika z własności całki oznaczonej.

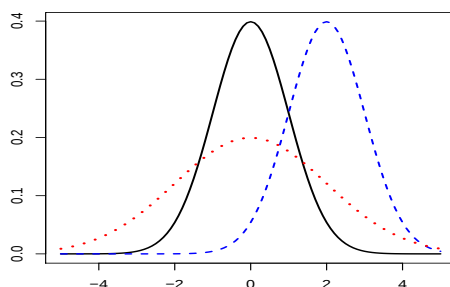
Rozkład normalny

Szczególnie ważnym w zastosowaniach jest rozkład normalny.

Definicja 2. Mówimy, że zmienna losowa X ma rozkład normalny z parametrami μ i σ , gdzie $\mu \in \mathbb{R}$ i $\sigma > 0$, jeżeli gęstość jej rozkładu jest określona wzorem:

$$\phi_{\mu, \sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Skrótowy zapis: $X \sim N(\mu, \sigma)$. Dla $\mu = 0$ i $\sigma = 1$ będziemy pisać zamiast $\phi_{0,1}(x)$ krótko $\phi(x)$.



Rysunek 1: Wykresy gęstości rozkładów normalnych: $N(0, 1)$ (linia ciągła), $N(0, 2)$ (linia „kropkowana”), $N(2, 1)$ (linia „kreskowana”).

Rozkład normalny — zastosowania

Wiele cech (zmiennych losowych) w życiu gospodarczym, w świecie przyrody ma rozkład zbliżony do normalnego.

Wynika to z tzw. centralnego twierdzenia granicznego, z którego wynika, że średnia $\frac{1}{n}(X_1 + X_2 + \dots + X_n)$, gdzie $X_1, X_2, \dots, X_n$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie, ma rozkład zbliżony do normalnego $N(\mu, \sigma)$ dla pewnych μ i σ . Dokładniejsze sformułowanie tego twierdzenia wymaga określenia wartości oczekiwanej i wariancji zmiennej losowej typu ciągłego.

Obliczanie prawdopodobieństw w rozkładzie normalnym- $N(0, 1)$

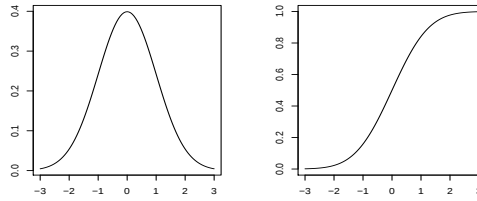
Dla $a < b$ prawdopodobieństwo $P(a < X < b)$, gdzie $X \sim N(0, 1)$ jest równe:

$$P(a < X < b) = \int_a^b \phi(x)dx = \Phi(b) - \Phi(a),$$

gdzie Φ jest określona przez:

$$\Phi(t) = \int_{-\infty}^t \phi(x)dx.$$

Funkcja Φ jest dystrybuantą rozkładu normalnego $N(0, 1)$. Funkcji Φ nie da się wyrazić za pomocą skończonej liczby działań na podstawowych funkcjach elementarnych — stąd potrzeba sporządzania tablic statystycznych zawierających wartości funkcji Φ (można je znaleźć w prawie każdym podręczniku statystyki).



Rysunek 2: Wykresy gęstości ϕ rozkładu normalnego (z lewej strony) $N(0, 1)$ i dystrybucyj rozkładu normalnego Φ (z prawej strony)

Własności funkcji Φ

Można pokazać, że $\Phi(0) = 0,5$ oraz $\Phi(t) = 1 - \Phi(-t)$ dla dowolnego t ; stąd można się ograniczyć do tablicowania funkcji Φ dla $t \geq 0$.

Obliczanie prawdopodobieństw w przypadku rozkładu normalnego $N(\mu, \sigma)$

Można pokazać, że jeśli $X \sim N(\mu, \sigma)$, to

$$\frac{X - \mu}{\sigma} \sim N(0, 1).$$

Stąd dla $a < b$ prawdopodobieństwo $P(a < X < b)$, $X \sim N(\mu, \sigma)$ jest równe:

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).$$

Obliczanie prawdopodobieństw w przypadku rozkładu normalnego — przykład

Niech X oznacza wzrost dorosłych mężczyzn w państwie A; zakładamy, że $X \sim N(177, 10)$. Chcemy obliczyć: (a) $P(174 < X < 182)$, (b) $P(X > 182)$. Obliczenia dla (a):

$$\begin{aligned} P(174 < X < 182) &= \Phi\left(\frac{182 - 177}{10}\right) - \Phi\left(\frac{174 - 177}{10}\right) \\ &= \Phi(0,5) - \Phi(-0,3) = \Phi(0,5) - (1 - \Phi(0,3)) = \\ &= \Phi(0,5) + \Phi(0,3) - 1 \approx 0,6915 + 0,6179 - 1 = 0,3094. \end{aligned}$$

Obliczenia dla (b) można przeprowadzić w analogiczny sposób, korzystając z równości:

$$P(X > 182) = 1 - P(X < 182) = 1 - \Phi\left(\frac{182 - 177}{10}\right) = 1 - \Phi(0,5).$$

Inne rozkłady ciągłe

Dowolna funkcja g spełniająca warunki:

- dziedziną funkcji g jest zbiór liczb rzeczywistych $\mathbb{R}$;
- $g(x) \geq 0$ dla $x \in \mathbb{R}$;
- $\int_{-\infty}^{\infty} g(x) = 1$.

jest funkcją gęstością pewnej zmiennej losowej; Poza rozkładem normalnym i rozkładem jednostajnym $U(0, 1)$ do opisu cech w życiu gospodarczym i naukach przyrodniczych stosuje się wiele innych rozkładów prawdopodobieństwa.

Gęstość emiryczna — przykład

Dane *normtemp*— zebrane w celu weryfikacji hipotezy mówiącej, że średnia wartość temperatury zdrowego człowieka jest równa 98,6 stopni w skali Fahrenheita (37,0 stopni w skali Celsjusza). Dane nt. temperatury i tętna (temperatura-stopnie Fahrenheita) Można je pobrać z odpowiedniego repozytorium a następnie zapisać do zbioru o nazwie np. **t** (tzw. „data frame”). Zbiór **t** składa się z trzech zmiennych: **temperature**, **gender** i **hr**. Aby uczynić naszą prezentację bardziej czytelną, zmieniamy nazwy zmiennych na odpowiednio: **temp**, **plec** i **tetno**. Odpowiednie polecenia systemu R są zapisane w pliku *t.R*. Wydruk tego pliku zamieszczamy poniżej (na następnym slajdzie):

Pobieranie zbioru danych z repozytorium systemu R

```
library(utils)
install.packages(c("xlsReadWrite", "UsingR"),
repo="http://cran.r-project.org")
library(xlsReadWrite)
library(UsingR)
t<-normtemp
names(t)<-c("temp", "plec", "tetno")
write.xls(t, 'c:/t.xls') # zbior t zapisany
                        # do pliku t.xls
```

System (pakiet) R można pobrać pod adresem <http://r.meteo.uni.wroc.pl/bin/windows/base/>

Pobieranie zbioru danych z repozytorium systemu R

```
> names(t)
[1] "temp" "plec" "tetno"
> t[1:10,]
      temp plec tetno
1  96.3     1    70
2  96.7     1    71
3  96.9     1    74
4  97.0     1    80
5  97.1     1    73
6  97.1     1    75
7  97.1     1    82
8  97.2     1    64
9  97.3     1    69
10 97.4     1    70
> sort(t$temp)
[1] 96.3 96.4 96.7 96.7 96.8 96.9 97.0 97.1
[9] 97.1 97.1 97.2 97.2 97.2 97.3 97.4 97.4
...
[129] 100.0 100.8
```

Szereg rozdzielczy

Dla zbioru danych liczbowych $\{y_1, y_2, \dots, y_N\}$ niech: $MIN1$ oznacza liczbę mniejszą od najmniejszej z liczb $\{y_1, y_2, \dots, y_N\}$ $MAX1$ oznacza liczbę większą lub równą od największej z liczb $\{y_1, y_2, \dots, y_N\}$ $MIN1$ i $MAX1$ mogą być odpowiednimi zaokrągleniami wartości, odpowiednio, minimalnej i maksymalnej naszego zbioru danych $MIN1 < MAX1$. Podzielmy odcinek $(MIN1, MAX1]$ na k przedziałów (zwanych klasami) o równej długości:

$$(x_0, x_1], (x_1, x_2], \dots, (x_{k-1}, x_k],$$

gdzie $x_0 = MIN1$, $x_k = MAX1$. Funkcję przyporządkowującą poszczególnym przedziałom liczbę elementów naszego zbioru danych do nich należących będziemy nazywać szeregiem rozdzielczym.

Liczbę klas k w szeregu rozdzielczym wyznaczamy korzystając z wzorów:

$$k \approx \log_2 n + 1 \quad \text{lub} \quad k \approx \frac{3\sqrt{n}}{4}.$$

Szereg rozdzielczy: dane NT

Przyjmujemy: $MIN1 = 96$, $MAX1 = 101$, $k = 10 \approx \frac{3\sqrt{130}}{4}$.

Zakładając, że dane dotyczące temperatury zdrowych ludzi znajdują się w zmiennej `t$temp` konstruujemy szereg rozdzielczy w środowisku R:

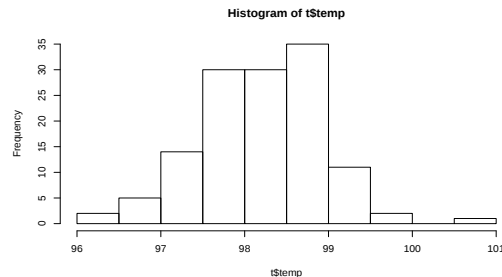
```
> table(cut(t$temp, breaks = c(96, 96.5, 97, 97.5, 98,
+ 98.5, 99, 99.5, 100, 100.5, 101)))
```

(96, 96.5]	(96.5, 97]	(97, 97.5]	(97.5, 98]
2	5	14	30

(98, 98.5]	(98.5, 99]	(99, 99.5]	(99.5, 100]
30	35	11	2
(100, 100.5]	(100.5, 101]		
0	1		

Histogram: dane NT

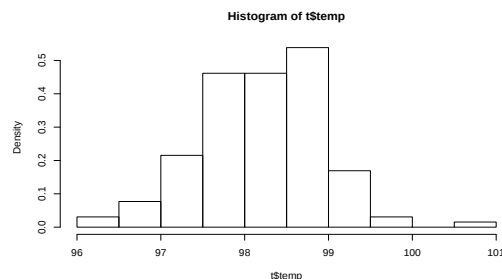
Wykres słupkowy odpowiadający szeregowi rozdzielczemu-histogram liczebności



Rysunek 3: Histogram dla danych NT odpowiadający szeregowi rozdzielczemu z poprzedniego slajdu

Histogram probabilistyczny i gęstość empiryczna - dane NT

Histogram probabilistyczny: histogram tak wyskalowany, aby "pole pod nim było równe 1": wysokości słupków: $\frac{2}{130 \times 0,5}, \frac{5}{130 \times 0,5}, \dots$; histogram probabilistyczny — przez niektórych definiowany jako funkcja przedziałami ciągła (stała), której wykres "pokrywa się" ze zdefiniowanym wyżej wykresem słupkowym; inna nazwa tak określonej funkcji: gęstość empiryczna.



Rysunek 4: Gęstość empiryczna dla danych NT

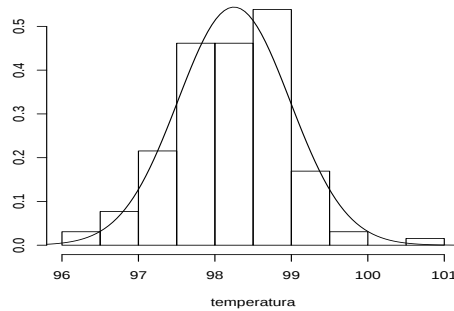
Gęstość empiryczna + krzywa normalna

Gęstość empiryczna —funkcja przedziałami ciągła

$$h(x) = \begin{cases} \frac{2}{130 \times 0,5}, & \text{dla } x \in (96; 96,5] \\ \frac{5}{130 \times 0,5}, & \text{dla } x \in (96,5; 97] \\ \vdots & \\ \frac{1}{130 \times 0,5}, & \text{dla } x \in (100,5; 101] \end{cases}$$

Fracja obserwacji należących do $(96; 97]$;

$$\int_{96}^{97} h(x) dx = \frac{1}{2} \left(\frac{2}{130 \times 0,5} + \frac{5}{130 \times 0,5} \right) = \frac{7}{130} \approx 0,054$$



Rysunek 5: Gęstość empiryczna dla danych NT z dołączonym wykresem gęstości normalnej z parametrami $\mu = 98,25$ i $\sigma = 0,733$.

Konstrukcja histogramu probabilistycznego (gęstości empirycznej) — przypadek ogólny

W ogólnym przypadku wysokość k -tego słupka histogramu probabilistycznego (wartość funkcji dla argumentów należących do k -tej klasy) jest równa $\frac{n_k}{nd}$, gdzie

- n_k jest liczebnością k -tej klasy,
- d jest długością klasy,
- n jest liczbą obserwacji.

Funkcja h — „histogram probabilistyczny” (gęstość empiryczna) — zdefiniowana wzorem:

$$h(x) = \begin{cases} \frac{n_k}{nd}, & x \text{ należy do } k\text{-tej klasy,} \\ 0, & x \text{ nie należy do żadnej z klas.} \end{cases}$$

Dziedziną funkcji h jest zbiór liczb rzeczywistych $\mathbb{R}$.

Lektura uzupełniająca

T. Bednarski, Elementy matematyki w naukach ekonomicznych. Oficyna ekonomiczna. Kraków 2004, str. 234–244.

Koronacki, J., Mielniczuk, J. Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT. Warszawa 2001, s. 111–118.