Inżynierskie zastosowania statystyki - Ćwiczenia nr 3

1 Wstęp teoretyczny

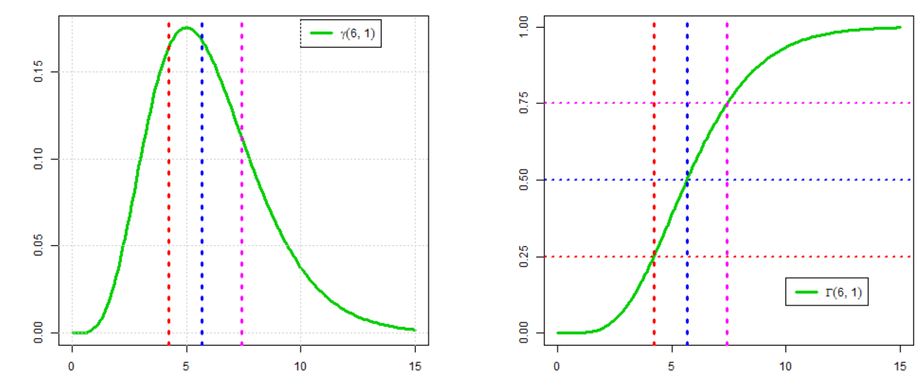
Kwantylem stopnia p, który oznaczamy przez q_p, nazywamy argument, dla którego wartość dystrybuanty jest równa p. Poniższe definicje są równoważne.

$$q_p : F\left(q_p\right) = p \tag{1}$$

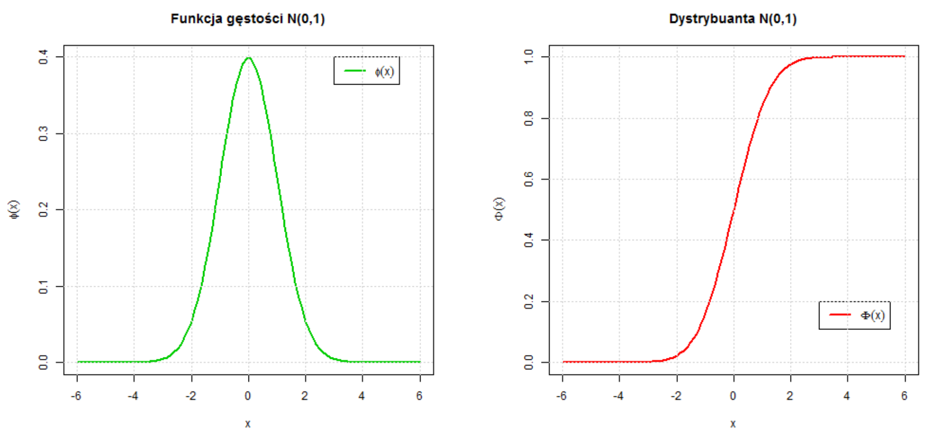
$$q_p : \int_{-\infty}^{p} f(u)\,du = p \tag{2}$$

$$q_p : \int_{p}^{+\infty} f(u)\,du = 1 - p \tag{3}$$

Kwartylami nazywamy kwantyle, które dzielą pole pod funkcją gęstości na cztery równe części. Na poniższych wykresach przedstawiono położenie kwartyli względem funkcji gęstości (po lewej) i dystrybuanty (po prawej).



Standaryzowany rozkład normalny $N(0, 1)$ to rozkład normalny o zerowej wartości średniej i wariancji równej 1. Dystrbuanta standaryzowanego rozkładu normalnego oznaczana jest przez $\Phi\left(x\right)$, natomiast funkcja gęstości przez $\phi\left(x\right)$.



Rozkład ten jest symetryczny względem swojej wartości średniej, stąd:

$$\Phi\left(-x\right) = 1 - \Phi\left(x\right) \tag{4}$$

Dla dowolnego rozkładu normalnego $N\left(\mu, \sigma^2\right)$ o dystrybuancie $F(x)$

$$F\left(x\right) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{5}$$

2 Zadania

2.1 Rozkład normalny

1. Korzystając z tego, że $\Phi\left(-x\right) = 1 - \Phi\left(x\right)$ oraz z faktu, że jeśli $F(x)$ jest dystrybuantą dowolnego rozkładu normalnego $N\left(\mu, \sigma^2\right)$ to $F\left(x\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ znajdź $F(5)$, $F(-1)$ oraz kwantyl rzędu 0.95 dla rozkładu normalnego $N(2, 9)$.

2. Niech $F(x)$ będzie dystrybuantą rozkładu normalnego $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$. Korzystając z tych samych własności rozkładu normalnego co w poprzednim zadaniu, wyprowadź wzór na prawdopodobieństwo $P\left((\mu - a, \mu + a)\right)$, gdzie $a \in \mathbb{R}_+$.

3. Zakładając, że pomiar ma rozkład normalny o wartości średniej $\mu = 123.4$ i wariancji $\sigma^2 = 25$, oblicz prawdopodobieństwo tego, że będzie on:
   (a) mniejszy od 120, (b) większy od 135, (c) będzie leżał w granicach od 117.4 do 129.4.

2.2 Rozkład wartości średniej

4. Artykuł "Średnia objętość płytek krwi u pacjentów z zespołem metabolicznym i jego związek z chorobą wieńcową" (Thrombosis Research [2007]: 245-250) zawiera dane sugerujące, że rozkład objętości płytek krwi u osób nie chorujących na chorobę wieńcową jest w przybliżeniu normalny ze średnią $\mu = 8.25$ i odchyleniem standardowym $\sigma = 0.75$.

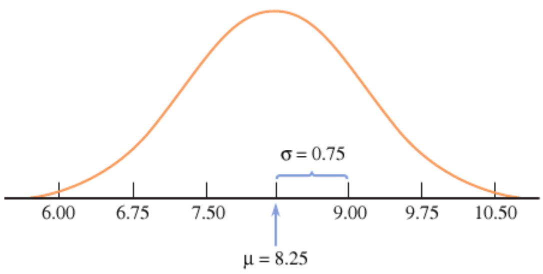


Figure 1: Rozkład normalny objętości płytek krwi z $\mu = 8.25$ i $\sigma = 0.75$

Na rysunku 1 pokazano krzywą Gaussa wyśrodkowaną w 8.25. Wartość odchylenia standardowego, 0.75, opisuje jak bardzo rozkład jest rozproszony względem wartości średniej.

Wybrano 500 prób $n = 5$ elementowych z tego rozkładu normalnego. Dla każdej z prób obliczono $\bar{x}$ i w wyniku otrzymano histogram 500 $\bar{x}$ wartości (patrz Rysunek 2(a)). Następnie procedurę powtórzono dla $n = 10$, $n = 20$ i $n = 30$ (patrz Rysunek 2(b)-(d))

   (a) Funkcję gęstości jakiego rozkładu przypominają histogramy na Rysunku 2?
   (b) Jaka jest wartość średnia odczytana z każdego z tych histogramów?
   (c) Jeżeli histogram byłby utworzony w oparciu o próby o nieskończonej liczbie elementów, jaka byłaby jego wartość średnia?
   (d) Jak zmienia się rozproszenie wartości wokół średniej wraz ze wzrostem liczby obserwacji w próbie?

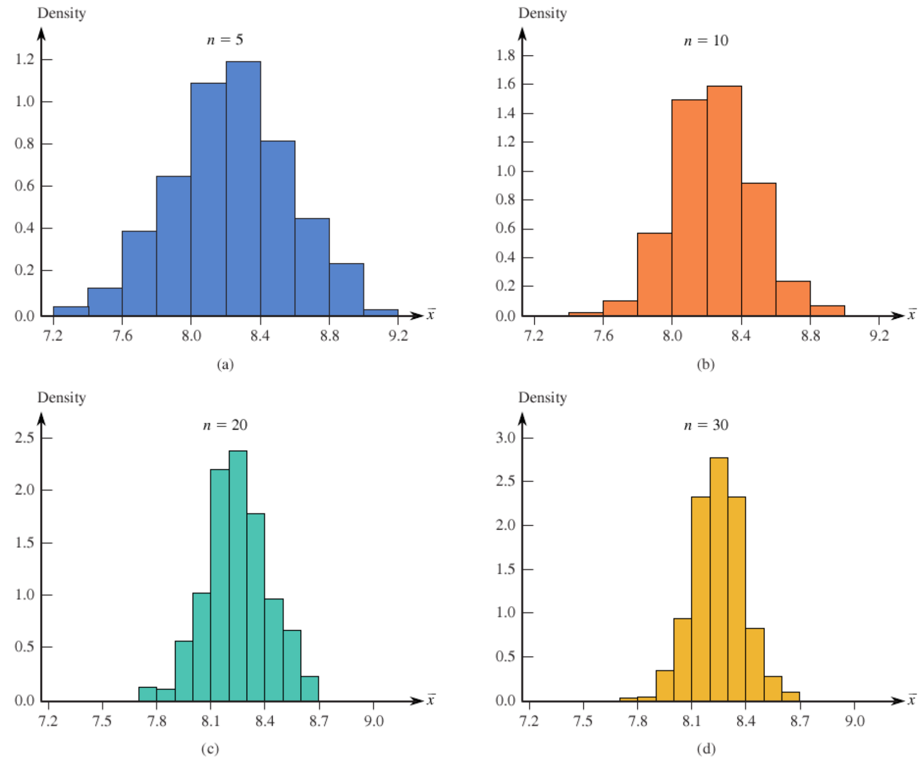


Figure 2: Histogramy częstości $\bar{x}$ obliczony z 500 prób n-elementowych

5. Rozpatrzmy własności rozkładu wartości średniej z próby $\bar{x}$ gdy rozkład prawdopodobieństwa jest skośny (i przez to niepodobny do rozkładu normalnego). Artykuł "Czy czas dogrywki w rozgrywkach NHL jest wystarczająco długi?" (American Statistician [2008]:151-154) podaje czas (w minutach) od początku meczu do pierwszego zdobytego gola w trakcie 281 meczów z dogrywką rozegranych w sezonie 2005-2006. Wykres 3 pokazuje histogram danych (odtworzony z artykułu). Histogram ma długi prawy (górny) ogon, co wskazuje, że w większości meczów pierwszy gol był zdobywany w pierwszych 20tu minutach, ale czasami był zdobywany znacznie później.

   Na podstawie histogramu obliczono, że wartość średnia tego rozkładu wynosi $\mu = 13$ min.

   Wybrano po 500 prób rozmiaru n, $n = 5, 10, 20, 30$ i utworzono cztery histogramy 500 wartości $\bar{x}$ (patrz Rysunek 4).

   (a) Czy histogramy na Rysunku 4 przypominają jakąś funkcję gęstości? Dla jakiego n?

   (b) Jaka jest (w przybliżeniu) wartość średnia odczytana z każdego z tych histogramów?

   (c) Jeżeli histogram byłby utworzony w oparciu o próby o nieskończonej liczbie elementów, jaka byłaby jego wartość średnia?

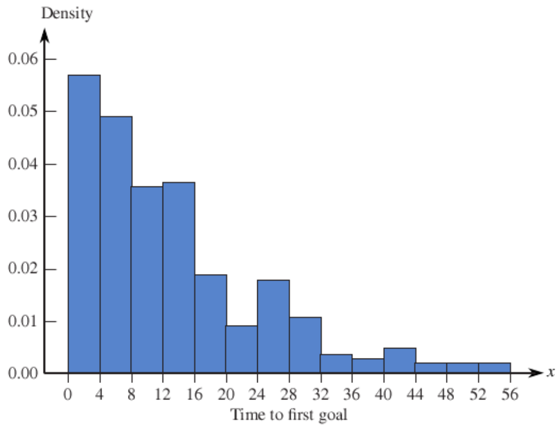


Figure 3: Rozkład populacji ($\mu = 13$)

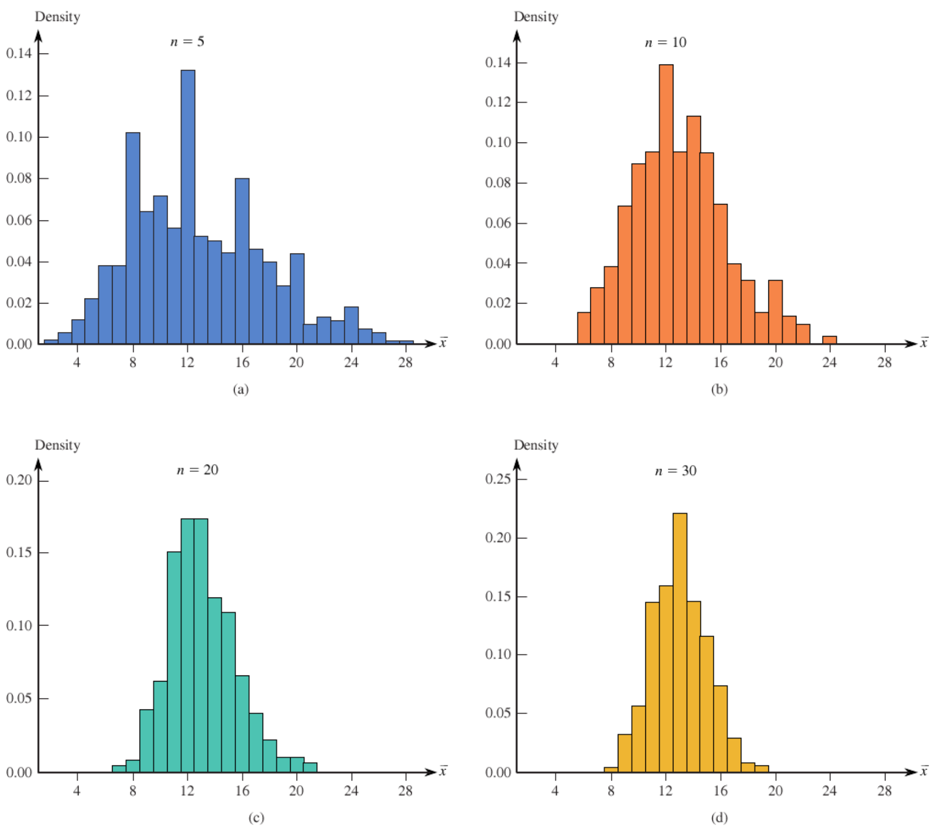


Figure 4: Cztery histogramy 500 wartości $\bar{x}$ obliczonych z n-elementowych prób