

# Próbki, zmienne losowe i ich rozkłady w R

Agnieszka Goroncy



**UNIWERSYTET  
MIKOŁAJA KOPERNIKA  
W TORUNIU**

Wydział Matematyki  
i Informatyki

W środowisku R istnieje możliwość **generowania próbek losowych** różnego rodzaju. Mogą to być

- próbki losowe wygenerowane na podstawie wektora zawierającego dane jakościowe bądź liczbowe, bez lub z powtórzeniami,
- próbki losowe pochodzące z konkretnego rozkładu prawdopodobieństwa.

# Funkcja `sample()`

Do wylosowania próbki na podstawie zadanego wektora danych służy funkcja **`sample()`**. Argumenty są następujące:

- pierwszym jest albo wektor źródłowy, z którego będziemy wybierać, albo liczba całkowita, określająca rozmiar próby źródłowej,
- rozmiar próby, którą chcemy uzyskać,
- `replace` - określa, czy pobieramy z powtórzeniami (`=TRUE`), czy bez (`=FALSE`, domyślnie),
- `prob` - opcjonalny wektor prawdopodobieństw wyboru poszczególnych elementów z wektora źródłowego.

# Funkcja `sample()` - przykłady

## Przykłady zastosowania:

```
> #wylosowanie 10 dużych liter alfabetu,  
> # możliwe powtórzenia  
> sample(LETTERS[1:26], 10, replace=TRUE)  
  
> # wylosowanie 5 różnych małych liter alfabetu  
> sample(letters[1:26], 5)  
  
> # wylosowanie 20 liczb z przedziału [1,1000],  
> # możliwe powtórzenia  
> sample(1000, 20, replace=TRUE)
```

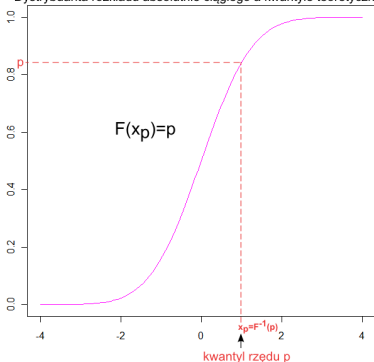
# Kwantyle teoretyczne

**Kwantylem rzędu  $p$** , gdzie  $0 \leq p \leq 1$ , rozkładu zmiennej losowej  $X$  nazywamy wartość  $x_p$ , dla której spełnione są nierówności

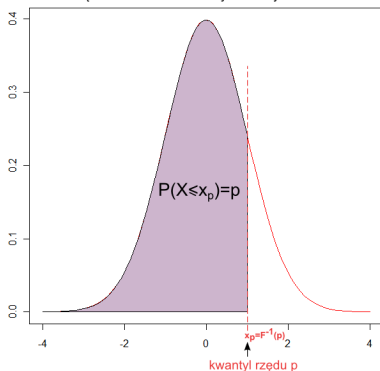
$$P(X \leq x_p) \geq p \quad \text{ i } \quad P(X \geq x_p) \geq 1 - p.$$

W przypadku rozkładów absolutnie ciągłych ta definicja się upraszcza:

Dystrybuanta rozkładu absolutnie ciągłego a kwantyle teoretyczne



Gęstość rozkładu a kwantyle teoretyczne



# Zmienne losowe

W R dostępny jest pakiet funkcji statystycznych (`stats`), które umożliwiają pracę ze zmiennymi losowymi pochodzącymi z różnych rozkładów. Najczęściej przydatne są funkcje, których nazwa składa się z przedrostka:

- **r** (random generation) - pozwala **wygenerować próbkę losową** pochodzącą z danego rozkładu,
- **d** (density) - pozwala wyznaczyć wartość **gęstości** w punkcie dla danego rozkładu,
- **p** (probability distribution) - pozwala wyznaczyć wartość **dystrybuanty** w punkcie dla danego rozkładu,
- **q** (quantile) - pozwala wyznaczyć wartość **funkcji kwantylowej** w punkcie dla danego rozkładu,

oraz nazwy rozkładu prawdopodobieństwa, np.:

- **norm** - rozkład **normalny**,
- **unif** - rozkład **jednostajny**,
- **exp** - rozkład **wykładniczy**,
- **gamma** - rozkład **gamma**,
- **beta** - rozkład **beta**,
- **chisq** - rozkład **Chi-kwadrat**,
- **t** - rozkład **t-Studenta**, itp.

Przykładowo, funkcja `pt()` pozwala wyznaczyć wartość dystrybuanty rozkładu t-Studenta w dowolnym punkcie, funkcja `rexp()` pozwala wygenerować liczby losowe pochodzące z rozkładu wykładniczego.

# Przykłady

- Wygenerowanie 20 liczb losowych z rozkładu wykładniczego z parametrem 2:  
`> rexp(20, rate=2)`
- Wygenerowanie 30 liczb losowych z rozkładu standardowego normalnego:  
`> rnorm(30, mean=0, sd=1)`  
`> # jeśli nie podamy parametrów rozkładu, R domyślnie`  
`> # ustawi parametry standardowe, czyli mean=0, sd=1:`  
`> rnorm(30)`
- Wartości dystrybuanty i gęstości dla rozkładu normalnego ze średnią 2 i odchyleniem standardowym 0.5, obliczone w punkcie 1:  
`> pnorm(1, mean=2, sd=0.5)`  
`> dnorm(1, mean=2, sd=0.5)`
- Mediana (kwantyl rzędu  $\frac{1}{2}$ ) rozkładu standardowego normalnego i rozkładu normalnego ze średnią 2 i odchyleniem standardowym 0.5:  
`> qnorm(1/2)`  
`> qnorm(1/2, mean=2, sd=1/2)`

# Wykresy dot. rozkładów prawdopodobieństwa

Klikając w link *R Tutorial - Rozkłady zmiennych losowych* można znaleźć różne przykłady dot. rozkładów zmiennych losowych.

Natomiast wiedzę na temat generowania liczb losowych i próbek z różnych rozkładów można uzupełnić klikając w link *Dane losowe*.

W R można generować wykresy m.in. **dystybuant i gęstości** znanych rozkładów. Służy do tego np. funkcja **plot()**.

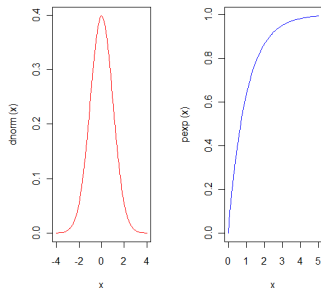
Pierwszym argumentem jest funkcja, której wykres chcemy uzyskać, np. `punif`, `dexp` itp. Kolejnymi są liczby określające dziedzinę funkcji.



# Wykresy - przykłady

## Przykłady:

```
> plot(dnorm,-4,4, col="red") #gęstość rozkładu normalnego  
> plot(pexp,0,5, col="blue") #dystrybuanta rozkładu  
> #wykładniczego
```



**UWAGA:** Sprawdź w R działanie funkcji `curve()` i wygeneruj za jej pomocą analogiczne wykresy.

# Rozkłady empiryczne

O dystrybuancie empirycznej patrz plik "Dystrybuanta empiryczna".

Przydatne funkcje powiązane z **dystrybuantą empiryczną**:

- `ecdf(x)` - pozwala wyznaczyć dystrybuantę empiryczną dla wektora obserwacji  $x$ ,
- `plot.ecdf()` - pozwala narysować dystrybuantę empiryczną

## Przykład:

```
> x=rnorm(10) # próba losowa
> # z rozkładu standardowego normalnego
> F_n<-ecdf(x) #funkcja  $F_n$  - dystrybuanta empiryczna
> F_n(0) #wartość dystrybuanty empirycznej w zerze
> F_n(x) #zwraca percentyle dla próby, czyli
> # wektor wartości dystrybuanty empirycznej
> # odpowiadających współrzędnym wektora x
```

Kolejną przydatną funkcją jest `knots()`, która pozwala wyznaczyć wektor punktów skoku dystrybuanty empirycznej. Jeżeli w próbce nie ma powtórzeń, wówczas rozmiar wektora jest równy rozmiarowi próbki, a punkty skoków pokrywają się z uporządkowanymi rosnąco współrzędnymi wektora  $x$ .

```
> knots(F_n)
```

Sprawdźmy, jak zachowa się funkcja podsumowująca podstawowe statystyki wywołana dla dystrybuanty empirycznej:

```
> summary(F_n)
```

# Dystrybuanta empiryczna - wykres

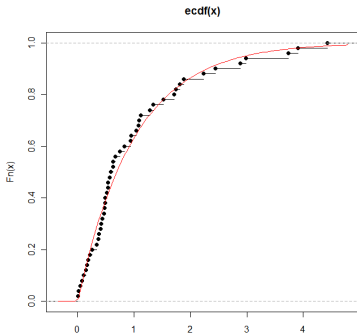
Narysujmy wykres dystrybuanty empirycznej dla próby losowej pochodzącej z rozkładu wykładniczego:

```
> y=rexp(50)
```

```
> plot.ecdf(y)
```

Czerwoną linią dodano dla porównania dystrybuantę rozkładu wykładniczego (nie zamykając okno z poprzednim wykresem):

```
> curve(pexp(x), add=TRUE, col="red")
```



# Wykres typu kwantyl-kwantyl

Wykresy typu kwantyl-kwantyl to bardzo użyteczne wykresy, które **pozwalają porównać rozkład próbki** z rozkładem teoretycznym bądź z rozkładem innej próbki.

Funkcja **qqnorm()** pozwala wygenerować wykres porównujący kwantyle empiryczne próby z kwantylami rozkładu normalnego. Ponadto, aby narysować na nim linię prostą, która przechodzi przez górny i dolny kwartyl, należy wywołać funkcję **qqline()**.

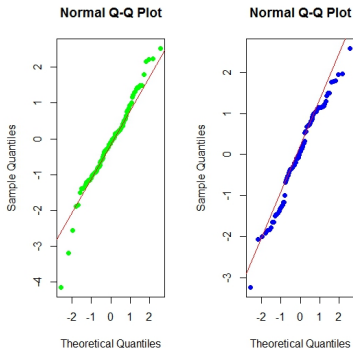
**UWAGA:** Ułożenie punktów wzdłuż prostej sugeruje normalność rozkładu próbki.

Funkcja **qqplot()** pozwala wygenerować wykres umożliwiający porównanie kwantyli empirycznych dwóch prób.

# Wykres kwantyl-kwantyl: przykład 1

**Przykład 1:** Porównamy kwantyle empiryczne stu elementowych próbek pochodzących z rozkładu t-Studenta z 10 i 100 stopniami swobody odpowiednio, z kwantylami rozkładu normalnego.

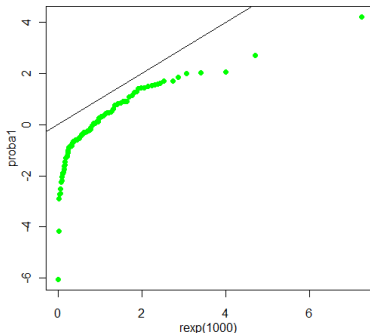
```
> proba1=rt(100,10)
> proba2=rt(100,100)
> par(mfrow=c(1,2))
> qqnorm(proba1,pch=19,col="green")
> qqline(proba1,col="red")
> qqnorm(proba2,pch=19,col="blue")
> qqline(proba2,col="red")
```



Jak widać na wykresie, im więcej stopni swobody, tym bardziej rozkład t-Studenta dopasowuje się do rozkładu normalnego.

## Wykres kwantyl-kwantyl: przykład 2

**Przykład 2:** Porównamy kwantyle empiryczne wygenerowanej poprzednio próbki z kwantylami standardowego rozkładu wykładniczego.






```
> qqplot(rexp(1000),proba1,  
+ col="green")  
> #rysujemy prostą y=x:  
> abline(0,1)
```

Położenie całego wykresu pod prostą sugeruje, że kwantyle rozkładu wykładniczego są bardziej zagęszczone w stosunku do kwantyli próbki. Wykres zupełnie nie pasuje do prostej  $\Rightarrow$  rozkład próbki nie jest wykładniczy.

Podobnie postępujemy, gdy chcemy porównać rozkłady dwóch próbek:

```
> qqplot(proba1,proba2)  
> abline(0,1)
```

-  Przemysław Biecek, **Przewodnik po pakiecie R**, Oficyna Wydawnicza GiS, Wrocław, 2011
-  Łukasz Komsta, **Wprowadzenie do środowiska R**
-  Joseph Adler, **R in a Nutshell**, O'Reilly Media, 2009