Przypomnienie + R

Piotr Guzik, Lucjan Janowski, Krzysztof Rusek

October 10, 2017

Outline

1 Projekty

2 Losowość

3 Źródła danych

4 Typy zmiennych losowych

5 Parametry

6 Statystyka

Jak idzie?

Outline

1 Projekty

2 Losowość

3 Źródła danych

4 Typy zmiennych losowych

5 Parametry

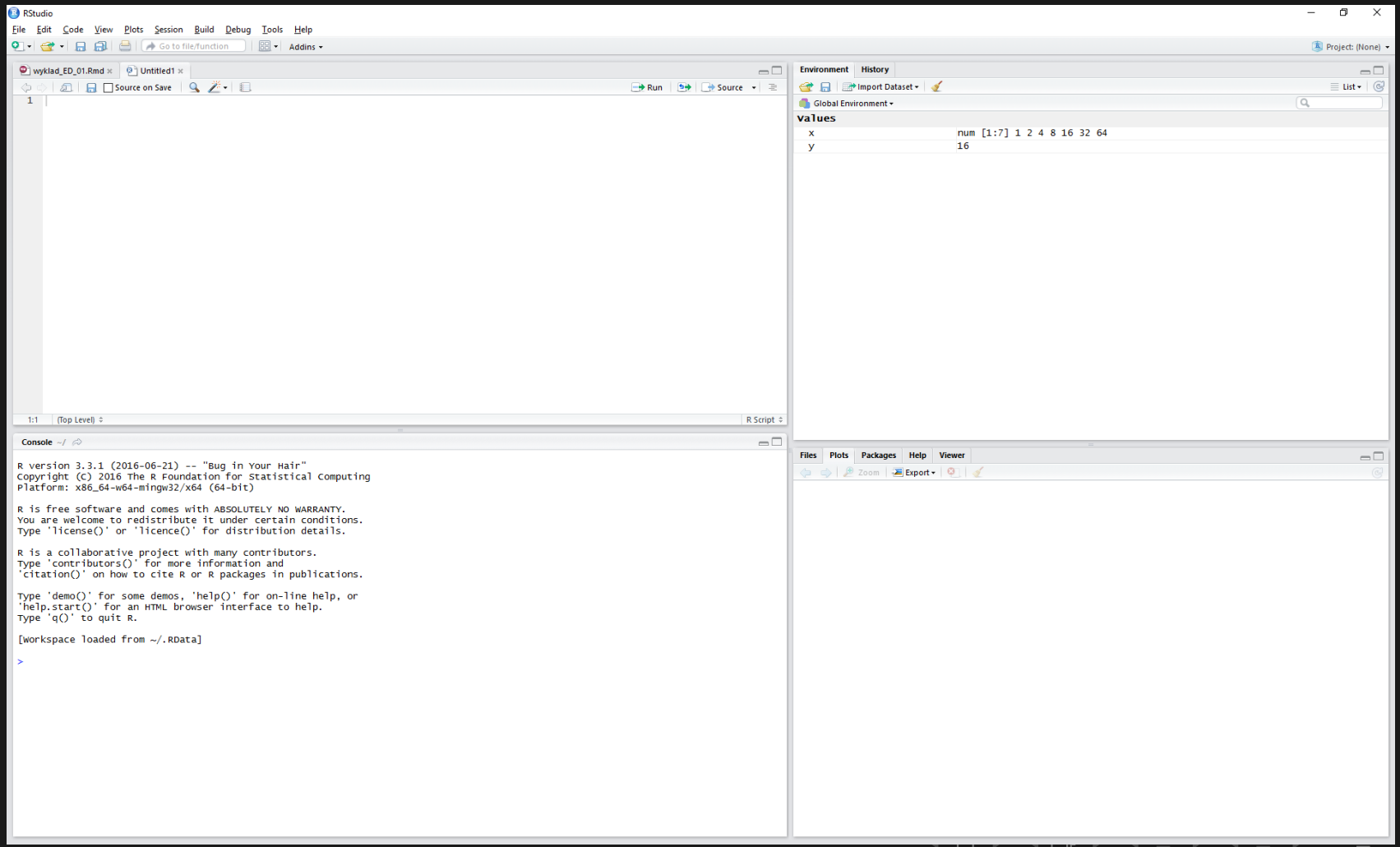
6 Statystyka

Losowość

Losowość jest cechą ograniczonej informacji lub samego świata. To co musimy pamiętać to, jakie są tego konsekwencje.

Rozważania na temat losowości rzutem monetą:
`https://www.youtube.com/watch?v=AYnJv68T3MM`
Oraz losowości czysto fizycznej:
`https://www.youtube.com/watch?v=zcqZHYo7ONs`

RStudio


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Addins
Project: (None)
wyklad_ED_01.Rmd
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment
History
Import Dataset
List
Global Environment
values
x num [1:7] 1 2 4 8 16 32 64
y 16
Console ~/
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Files
Plots
Packages
Help
Viewer
Zoom
Export

Moneta

W R możemy symulować monetę z wykorzystaniem kodu:

```
a = sample(0:1, 4, rep = TRUE, prob=c(0.4,0.6))
a
mean(a)
```

Moneta

W R możemy symulować monetę z wykorzystaniem kodu:

```
a = sample(0:1, 4, rep = TRUE, prob=c(0.4,0.6))
a
mean(a)
```

Małe szkoły w USA.

Moneta

W R możemy symulować monetę z wykorzystaniem kodu:

```
a = sample(0:1, 4, rep = TRUE, prob=c(0.4,0.6))
a
mean(a)
```

Dla innej liczności próbki mamy inne wyniki:

```
a = sample(0:1, 10000, rep = TRUE,
  prob=c(0.4,0.6)); mean(a)
```

Konsekwencje

- Rzadko obserwujemy samo zjawisko, raczej pewien czynnik losowy

Konsekwencje

- Rzadko obserwujemy samo zjawisko, raczej pewien czynnik losowy
- Podczas analizy danych musimy wiedzieć, jaka część to czynnik losowy i/lub czy obserwowane różnice wyników to wpływ losowości

Konsekwencje

- Rzadko obserwujemy samo zjawisko, raczej pewien czynnik losowy
- Podczas analizy danych musimy wiedzieć, jaka część to czynnik losowy i/lub czy obserwowane różnice wyników to wpływ losowości
- Podczas modelowania chcemy modelować część "deterministyczną" i ocenić czynnik losowy

Outline

1 Projekty

2 Losowość

3 Źródła danych

4 Typy zmiennych losowych

5 Parametry

6 Statystyka

Źródła danych

- Dane wewnętrzne firmy

Źródła danych

- Dane wewnętrzne firmy
- Kwestionariusze

Źródła danych

- Dane wewnętrzne firmy
- Kwestionariusze
- Programy lojalnościowe

Źródła danych

- Dane wewnętrzne firmy
- Kwestionariusze
- Programy lojalnościowe
- Dane zewnętrzne, zwłaszcza rządowe

Źródła danych

- Dane wewnętrzne firmy
- Kwestionariusze
- Programy lojalnościowe
- Dane zewnętrzne, zwłaszcza rządowe
- Dane makroekonomiczne

Źródła danych

- Dane wewnętrzne firmy
- Kwestionariusze
- Programy lojalnościowe
- Dane zewnętrzne, zwłaszcza rządowe
- Dane makroekonomiczne
- Udostępnione przez konkurencję

Źródła danych

- Dane wewnętrzne firmy
- Kwestionariusze
- Programy lojalnościowe
- Dane zewnętrzne, zwłaszcza rządowe
- Dane makroekonomiczne
- Udostępnione przez konkurencję
- Internet (blogi itd.)

Czytanie danych

Czyszczenie pamięci

```
rm(list=ls())
```

Czytanie plików CSV

```
met <- read.csv('resAg.csv', header = T)
```

Czytanie pliku on-line (pogromcydanych.icm.edu.pl)

```
read.table(file =
    "http://biecek.pl/MOOC/dane/koty_ptaki.csv",
    sep=";", dec=",", header=TRUE)
```

Podgląd danych

Czytanie danych

Czytanie z plików xlsx

```
piwa <- read.xls("http://kt.agh.edu.pl/
~janowski/PiwaWyniki.xlsx", 2)
```

Pakiety

Doinstalowywanie pakietów

```
install.packages("gdata")
```

Pakiet musi być uruchomiony!

```
library(gdata)
```

Odwołanie do tabel

Można jak do wektorów:

```
piwa[2,5]
piwa[,3]
piwa[10,]
```

Przez nazwę kolumny:

```
piwa$Tester
piwa$Rodzaj[30]
summary(piwa$Ocena)
```

API

Przykład z (pogromcydanych.icm.edu.pl) jak czytać zewnętrzne dane, do których mamy dostęp poprzez API:

```
library(SmarterPoland)
tsdtr210 <- getEurostatRCV("tsdtr210")
head(tsdtr210, 3)
summary(tsdtr210)
```

Zabawa danymi

Tu można zapoznać się z funkcją ggplot:
http://ggplot2.tidyverse.org/reference/ggplot.html

```
ggplot(data = tsdtr210[tsdtr210$geo == "PL",],
    aes(x = time, y = value,
      group = vehicle, colour = vehicle))
    + geom_line()
```

Zapisywanie

Do formatu rozumianego przez R

```
save(piwa, file="PierwszeWyniki.rda")
```

Odczytujemy to funkcją

```
load("PierwszeWyniki.rda")
```

Zapis do pliku cvs

```
write.csv(piwa, file="piwa.csv")
```

Coś innego z zapisu i odczytu?

Outline

1 Projekty

2 Losowość

3 Źródła danych

4 Typy zmiennych losowych

5 Parametry

6 Statystyka

Zmienne nominalne

- model samochodu, płeć

Zmienne nominalne

- model samochodu, płeć

Dla zmiennych nominalnych możemy

- Obliczyć częstość; ile aut danej marki ma wypożyczalnia
- Moda; w wypożyczalni najwięcej jest opli

Zmienne nominalne

- model samochodu, płeć

Dla zmiennych nominalnych możemy

- Obliczyć częstość; ile aut danej marki ma wypożyczalnia
- Moda; w wypożyczalni najwięcej jest opli

Dla zmiennych nominalnych NIE można

- Porządkować; *opel jest "większy" od audi !?!*
- *Obliczać średniej; średnia marka w wypożyczalni to gśdk !?!*
- *Regresji; jak zmienimy markę o trzy litery to wzrośnie spalanie o 0.3 litra !?!*

Zmienne nominalne

- model samochodu, płeć

Dla zmiennych nominalnych możemy

- Obliczyć częstość; ile aut danej marki ma wypożyczalnia
- Moda; w wypożyczalni najwięcej jest opli

Dla zmiennych nominalnych NIE można

- Porządkować; *opel jest "większy" od audi !?!*
- *Obliczać średniej; średnia marka w wypożyczalni to gśdk !?!*
- *Regresji; jak zmienimy markę o trzy litery to wzrośnie spalanie o 0.3 litra !?!*

Często w analizie zmienne nominalne są binaryzowane, czyli jedna zmienna nominalna zamieniana jest na $k - 1$ zmiennych binarnych.

Zmienne ilorazowe/absolutne

- spalanie, temperatura w stopniach Kelvina

Zmienne ilorazowe/absolutne

- spalanie, temperatura w stopniach Kelvina

Możemy przeprowadzać dowolne obliczenia.
Możemy liczyć średnie spalanie, zlogarytmować i ocenić jaki jest stosunek spalania w mieście do autostrady itd.

Zmienne porządkowe

- typy aut w wypożyczalni, jakość piwa

Zmienne porządkowe

- typy aut w wypożyczalni, jakość piwa

Dla zmiennych porządkowych możemy

- Obliczyć częstość; 40% wypożyczeń to grupa B
- Stosować metody rangowe; 80% wypożyczeń to auta z grupy C lub niższej

Zmienne porządkowe

- typy aut w wypożyczalni, jakość piwa

Dla zmiennych porządkowych możemy

- Obliczyć częstość; 40% wypożyczeń to grupa B
- Stosować metody rangowe; 80% wypożyczeń to auta z grupy C lub niższej

Dla zmiennych porządkowych NIE można

- Obliczać średniej; *średnie wypożyczone auto to grupa BC !?!*
- *Logarytmu; logarytm grupy B wynosi P ?!?*

Zmienne porządkowe

- typy aut w wypożyczalni, jakość piwa

Dla zmiennych porządkowych możemy

- Obliczyć częstość; 40% wypożyczeń to grupa B
- Stosować metody rangowe; 80% wypożyczeń to auta z grupy C lub niższej

Dla zmiennych porządkowych NIE można

- Obliczać średniej; *średnie wypożyczone auto to grupa BC !?!*
- *Logarytmu; logarytm grupy B wynosi P ?!?*

Często te zmienne traktowane są jak liczbowe.

Typy danych

Proszę wczytać dane o piwach. Jakie typy danych tu występują?

```
piwa <- read.xls("http://kt.agh.edu.pl/
~janowski/PiwaWyniki.xlsx", 2)

summary(piwa)
```

Typy danych

Proszę wczytać dane z metrykami zniekształceń wideo. Jakie typy danych tu występują?

```
met<-read.csv('resAg.csv')
summary(met)
```

Wczytanie danych z uwzględnieniem typów wygląda tak:

```
met2 = read.csv("resAg.csv", header=TRUE,
colClasses = c("character", "factor", "factor",
"numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric"))
summary(met2)
```

Outline

1 Projekty

2 Losowość

3 Źródła danych

4 Typy zmiennych losowych

5 Parametry

6 Statystyka

Wartość oczekiwana

To wartość, która niekoniecznie jest "oczekiwana" czyli najbardziej prawdopodobna. Dla rozkładu:

```
a = sample(c(1, 2, 4), 1000, rep = TRUE,
prob=c(0.5,0.25, 0.25))
mean(a)
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Mode(a)
```

Wartość oczekiwana

To wartość, która niekoniecznie jest "oczekiwana" czyli najbardziej prawdopodobna. Dla rozkładu:

```
a = sample(c(1, 2, 4), 1000, rep = TRUE,
prob=c(0.5,0.25, 0.25))
mean(a)
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Mode(a)
```

wartość oczekiwana wynosi ...
Najbardziej prawdopodobna wartość (moda) wynosi ...

Wartość oczekiwana

To wartość, która niekoniecznie jest "oczekiwana" czyli najbardziej prawdopodobna. Dla rozkładu:

```
a = sample(c(1, 2, 4), 1000, rep = TRUE,
prob=c(0.5,0.25, 0.25))
mean(a)
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Mode(a)
```

Wartość oczekiwana jest wynikiem spodziewanym w wielu uśrednionych losowaniach. $\lim_{n\to\infty} \frac{X_1+\cdots X_n}{n}$; obliczenia: $EX = \int_{-\infty}^{\infty} xf(x)dx$

Standardowe odchylenie

Standardowe odchylenie to miara rozrzutu zmiennej wokół średniej. Pozwala ocenić ryzyko w szansie otrzymania wartości oczekiwanej.

Jaka jest szansa na wylosowanie liczby większej od 3, jeżeli wartość oczekiwana oraz standardowe odchylenie wynoszą 1?

Standardowe odchylenie

Standardowe odchylenie to miara rozrzutu zmiennej wokół średniej. Pozwala ocenić ryzyko w szansie otrzymania wartości oczekiwanej.

Jaka jest szansa na wylosowanie liczby większej od 3, jeżeli wartość oczekiwana oraz standardowe odchylenie wynoszą 1?

Standardowe odchylenie NIE pozwala na oszacowanie zakresu zmiennej, czy nawet szansy na otrzymanie wartości w danym zakresie!

Standardowe odchylenie

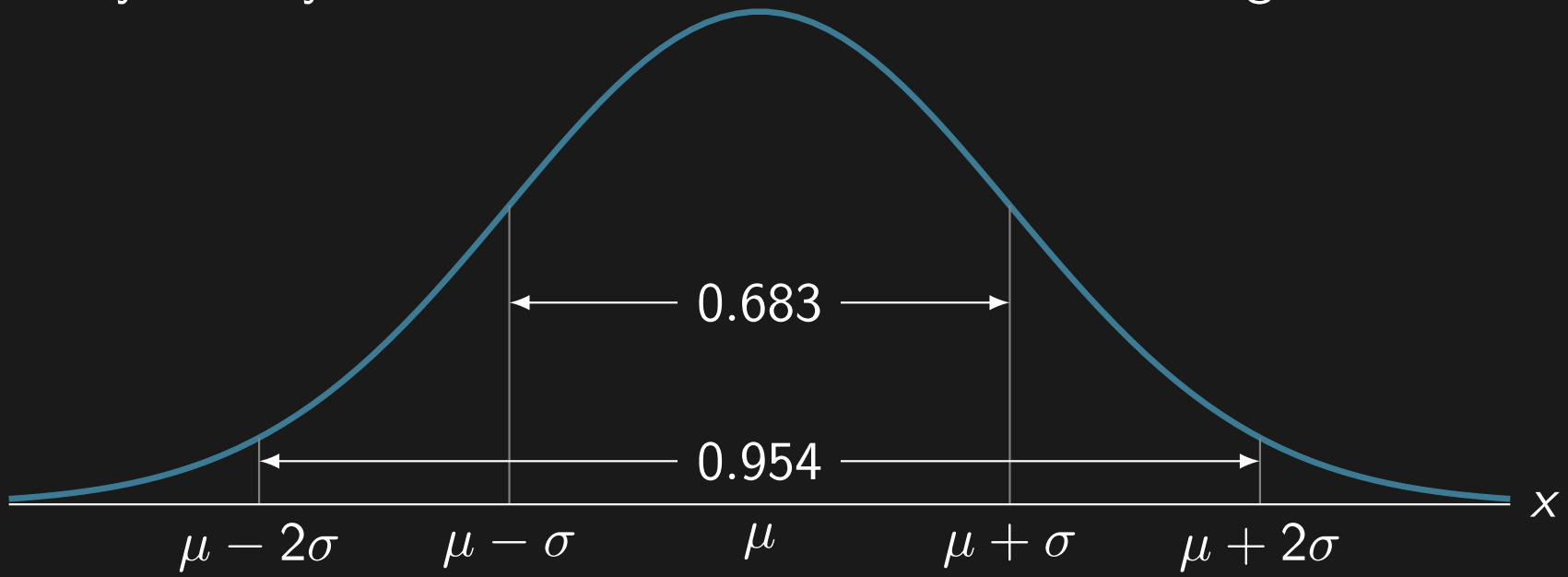
Proszę porównać prawdopodobieństwa otrzymania wartości w przedziale $(EX - \sigma_X, EX + \sigma_X)$ dla dwóch różnych rozkładów.

```
a = rexp(1000000, 2)
ld=mean(a)-sd(a)
lg=mean(a)+sd(a)
sum(a>ld & a<lg)

a = runif(1000000)
ld=mean(a)-sd(a)
lg=mean(a)+sd(a)
sum(a>ld & a<lg)
```

Standardowe odchylenie

Szacowanie prawdopodobieństwa z wykorzystaniem standardowego odchylenia wynika z odniesienia do rozkładu normalnego.



i tylko dla tego rozkładu to działa.

Korelacja

Korelacja mówi o liniowej zależności między zmiennymi.

```
a = 100
x = 1:1000
eps = rnorm(1000)
y = x + a*eps
plot(x, y)
cor(x, y)
```

Korelacja

Dla braku liniowej zależności będzie się mylić.

```
a = 100
x = 1:1000
eps = rnorm(1000)
y = x^2 + a^2*eps
plot(x, y)
cor(x, y)
```

Korelacja

Dla braku liniowej zależności może też kłamać.

```
a = 0.1
x=seq(0,1,0.001)
eps = rnorm(1001)
y=ifelse(runif(1001)>0.5, 1, -1)
*sqrt(1-x^2) + a*eps
plot(x, y)
cor(x, y)
```

Korelacja

To co wiemy:

Jeżeli korelacja jest niezerowa, to zmiana jednej zmiennej jest skojarzona (być może przypadkowo) ze zmianami drugiej zmiennej.

`http://tylervigen.com/spurious-correlations`

Jeżeli korelacja jest zerowa, to wiemy, że nie ma zależności liniowej, a nieliniowa musi być wyjątkowo wredna.

Outline

1 Projekty

2 Losowość

3 Źródła danych

4 Typy zmiennych losowych

5 Parametry

6 Statystyka

Rozkłady

Proszę narysować rozkłady (histogramy) dla różnych zmiennych.

```
hist(met$mean_Noise)
hist(met$mean_SA)
hist(met$mean_TA)
hist(met$mean_Blockiness)
hist(met$mean_Blockloss)
hist(met$mean_Blur)
```

Estymacja

W przypadku danych możemy być zainteresowani czy te dane pochodzą/sostały wylosowane z danego rozkładu. Można to sprawdzić w trzech etapach.

- Wybór rozkładu
- Estymacja parametrów
- Test dopasowania

Estymacja: przykład

Dopasowanie rozkładu Gamma do metryki SA

```
install.packages("fitdistrplus")
library(fitdistrplus)
met<-read.csv('resAg.csv')
fitdistr(met$mean_Blur, 'normal')
ks.test(met$mean_Blur, 'pnorm',
   mean=9.63, sd = 5.61)
```

Tu dokładne wyjaśnienie czemu taka procedura jest zła i jaka jest dobra: http://stats.stackexchange.com/questions/132652/how-to-determine-which-distribution-fits-my-data-best

Dopasowanie rozkładu

Proszę dopasować rozkład na podstawie tego tutoriala: `https://www.r-bloggers.com/fitting-distributions-with-r/`

Przedział ufności

Szansa, że prawdziwa wartość jest w przedziale ufności jest określona. Losowe są końce przedziału, a NIE to czy prawdziwa wartość jest w przedziale lub nie!

Testowanie hipotez

Testujemy czy dana wartość może być prawdziwą wartością parametru.

Mamy próbkę opóźnień serwera liczonych w milisekundach $3.1, 4.3, 2.1, 1.2, 5.2, 3.3, 4.5$ i zastanawiamy się czy wartość oczekiwana opóźnienia może wynosić 5 ms?

$H_0 : \mu = 5$
$H_1 : \mu \neq 5$
Musimy ustalić poziom istotności, czyli poziom dla którego zaakceptujemy, że nawet jeżeli próbka pochodzi z rozkładu o $\mu = 5$ to i tak uznamy, że tak nie jest. Zauważmy, że nawet dla rozkładu o parametrze $\mu = 10$ istnieje prawdopodobieństwo wylosowania zaobserwowanej próbki. Jest ono małe, ale nie zerowe.

Testowanie hipotez: przykład

Czy moneta jest symetryczna? Oddaliśmy 35 rzutów, 12 to orły, przyjmujemy poziom istotności 0.05.

```
prop.test(12, 35, p=0.5)
```

Interpretacja:

- p-value > 0.05; nie mamy podstaw do odrzucenia hipotezy alternatywnej
- p-value < 0.05; jest mało prawdopodobne żeby hipoteza zerowa była prawdziwa, więc ją odrzucamy