

# ANALIZA DANYCH IŁOŚCIOWYCH Z WYKORZYSTANIEM R

Katarzyna Abramczuk, Jakub Rybacki

**WHEN R AND PYTHON**



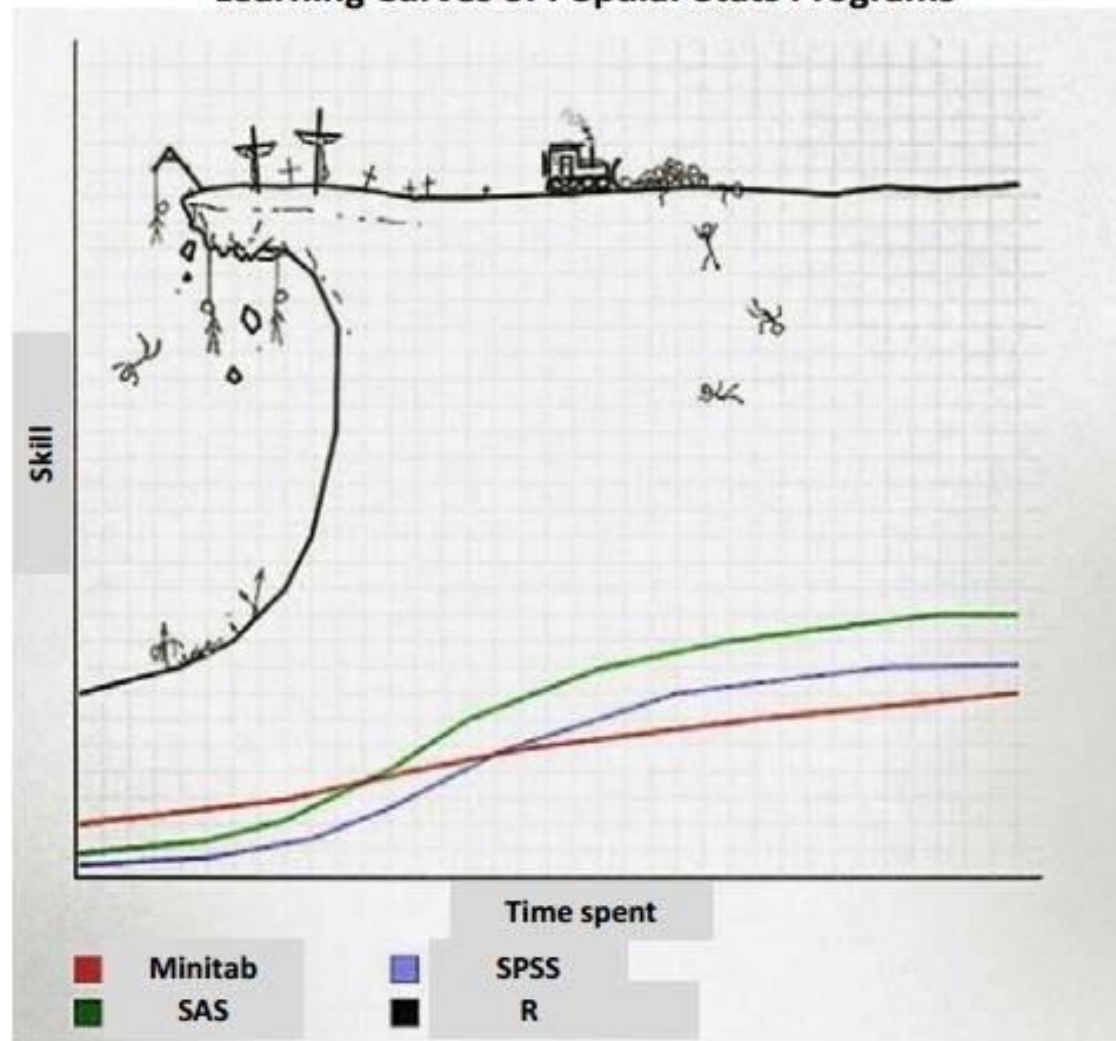
**JOIN FORCES**

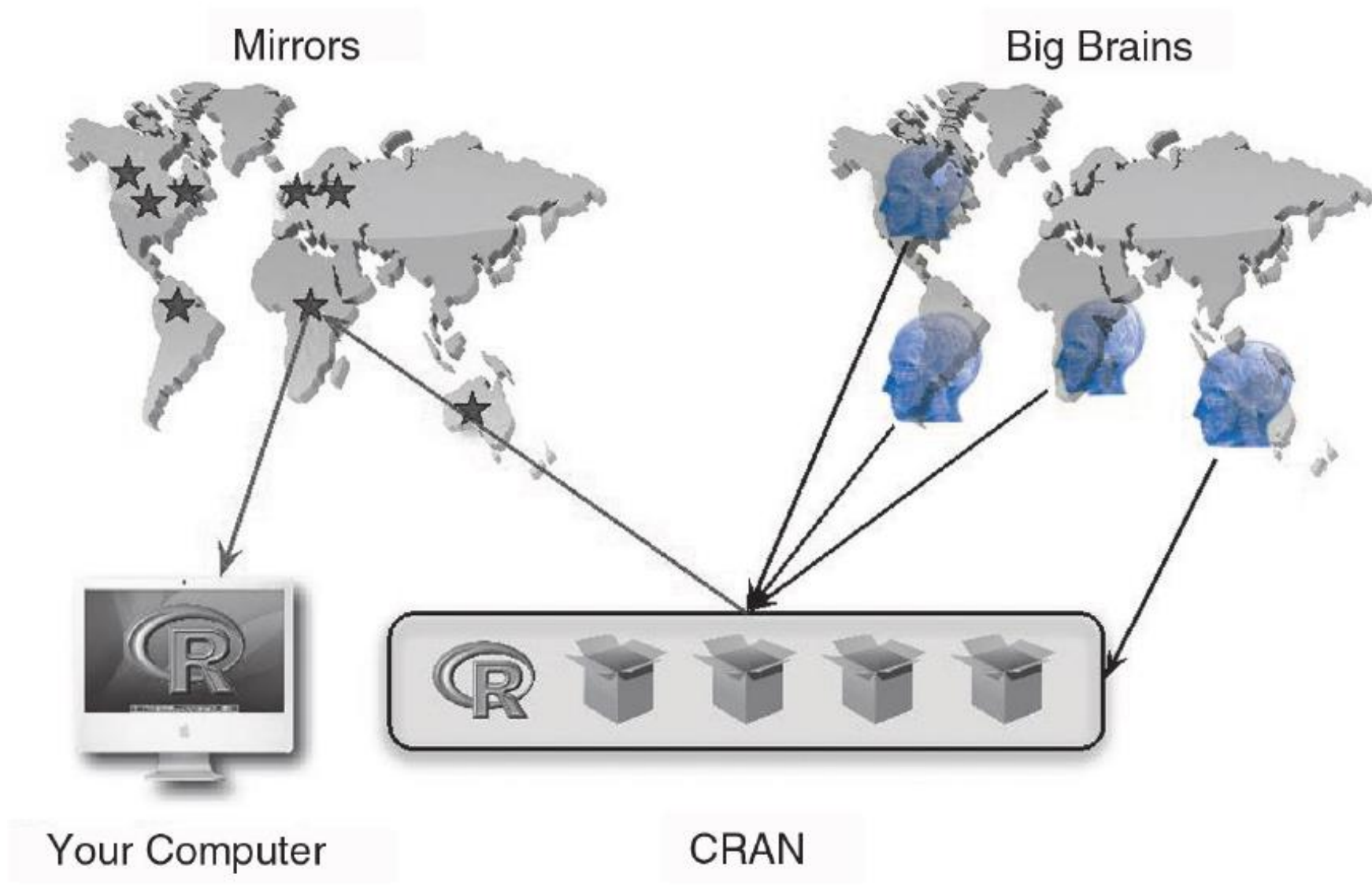
- The decision can be challenging because they **both Python and R have clear strengths.**
- **R is exceptional for Research:** Making visualizations, telling the story, producing reports, and making MVP apps with Shiny. From concept (idea) to execution (code), R users tend to be able to accomplish these tasks 3X to 5X faster than Python users, making them very productive for research.
- **Python is exceptional for Production ML:** Integrating machine learning models into production systems where your IT infrastructure relies on automation tools like Airflow or Luigi.





## Learning Curves of Popular Stats Programs







# Tidyverse

[Packages](#)[Blog](#)[Learn](#)[Help](#)[Contribute](#)

## R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

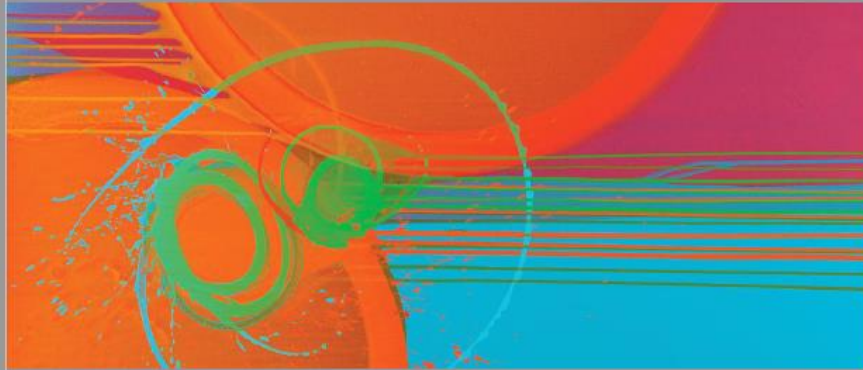


# Zarys programu

- Wprowadzenie do R
- Podstawowa eksploracja danych
- Podstawowa analiza współzmienności
- Wprowadzenie do wnioskowania statystycznego
- Prosta regresja liniowa
- Regresja wielokrotna
- Regresja logistyczna
- Porównywanie dwóch średnich
- ANOVA
- testy nieparametryczne
- analiza czynnikowa
- dane na słabych skalach
- modele ścieżkowe; metoda k-średnich; drzewa decyzyjne



# DISCOVERING STATISTICS USING R



ANDY FIELD | JEREMY MILES | ZOË FIELD

companion  
website







# Zasady zaliczenia

- Ocena końcowa z przedmiotu składa się z:
- 20% oceny za bieżące przygotowanie do zajęć;
  - aktywny udział w zajęciach, wykonywanie prac domowych
- 30% oceny z pracy nad projektem badawczym:
  - praca w dwu- lub trzyosobowych grupach, każda z grup będzie pracować nad projektem badawczym, polegającym na analizie wybranego zbioru danych ilościowych oraz prezentacji wyników
- 50% oceny z kolokwium końcowego:
  - kolokwium będzie się składało z 7-9 zadań
  - w trakcie kolokwium można korzystać z notatek



# Materiały do kursu

github.com/abramczuk/ADIWR

README.md

Update README.md

3 minutes ago

README.md

## Analiza Danych Ilościowych z wykorzystaniem R

Część materiałów na zajęcia *Analiza danych ilościowych z wykorzystaniem R* została przygotowana w ramach projektu *Program zintegrowanych działań na rzecz rozwoju Uniwersytetu Warszawskiego*, realizowanego w ramach programu operacyjnego *Wiedza Edukacja Rozwój*, oś priorytetowa III. Szkolnictwo wyższe dla gospodarki i rozwoju, działanie: 3.5 Kompleksowe programy szkół wyższych.

**Autorzy:** Katarzyna Abramczuk, Tomasz Żółtak, Agnieszka Karlińska, Jakub Rybacki

## Zarys programu

- Wprowadzenie do R
- Podstawowa eksploracja danych
- Podstawowa analiza współzmienności
- Wprowadzenie do wnioskowania statystycznego
- Prosta regresja liniowa
- Regresja wielokrotna

README

☆ 0 stars

👁 1 watching

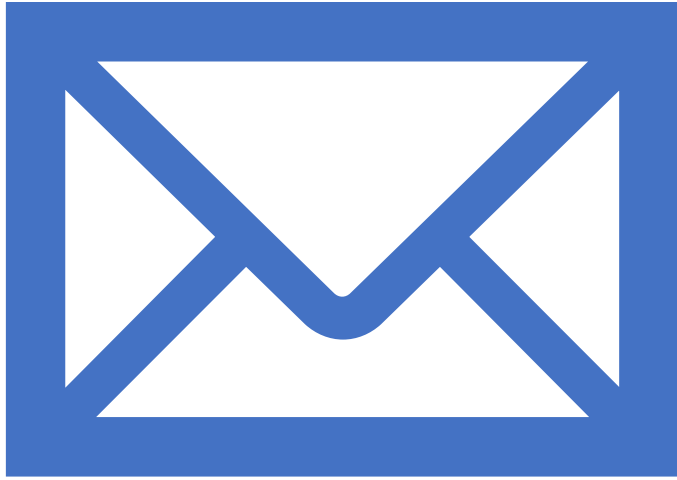
🍴 0 forks

### Releases

No releases published  
[Create a new release](#)

### Packages


No packages published  
[Publish your first package](#)



## Kontakt do prowadzących

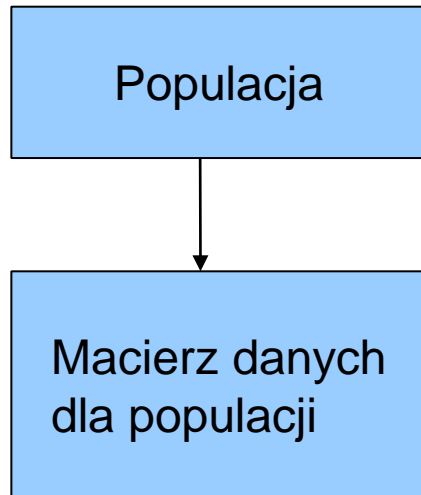
- [K.Abramczuk@uw.edu.pl](mailto:K.Abramczuk@uw.edu.pl)
- [Jakub.Rybacki@uw.edu.pl](mailto:Jakub.Rybacki@uw.edu.pl)

# Struktura badania statystycznego



Populacja

# Struktura badania statystycznego





# Struktura badania statystycznego

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

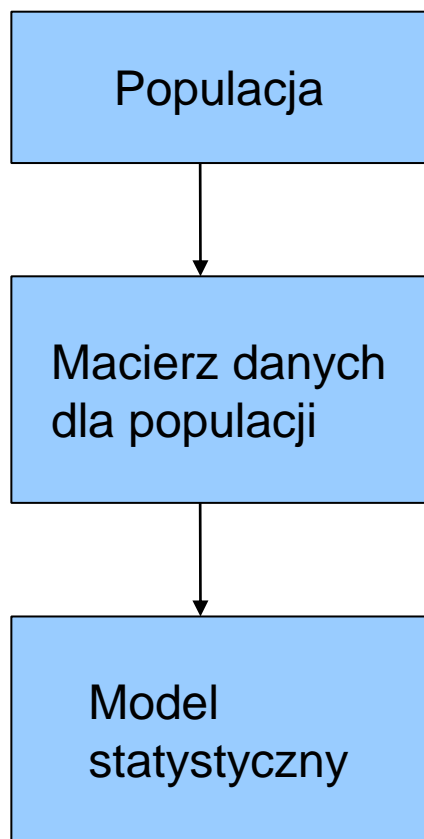
+ Go to file/function Addins

pisa x

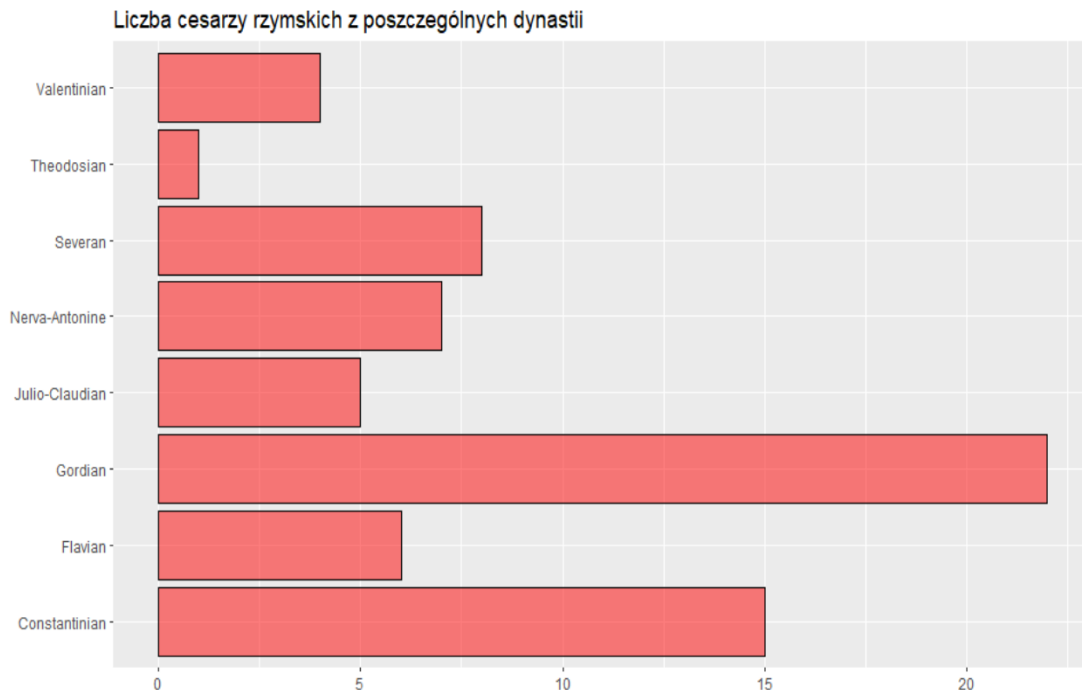
Filter

|    | id | schoolId | schoolType | sex    | scorePISAMath | scorePISARead | scorePISAScie | age      | scoreKKS | scoreKNS | scoreTMR |
|----|----|----------|------------|--------|---------------|---------------|---------------|----------|----------|----------|----------|
| 1  | 1  | 21       | LO         | female | 611.4486      | 522.1156      | 520.1351      | 17.05955 | 193      | 56       |          |
| 2  | 2  | 21       | LO         | female | 472.3196      | 478.7457      | 386.9010      | 17.64545 | 194      | 50       |          |
| 3  | 3  | 21       | LO         | female | 446.7525      | 471.6093      | 383.2587      | 17.89459 | 182      | 49       |          |
| 4  | 4  | 21       | LO         | male   | 539.9234      | 548.8076      | 481.0276      | NA       | NA       | NA       | /        |
| 5  | 5  | 21       | LO         | male   | 303.5452      | 357.4266      | 226.8285      | 17.47844 | 182      | 50       |          |
| 6  | 6  | 21       | LO         | male   | 406.1274      | 401.7657      | 311.9449      | 17.05955 | 187      | 51       |          |
| 7  | 7  | 21       | LO         | male   | 483.7699      | 480.4175      | 428.3090      | 17.31143 | 162      | 48       |          |
| 8  | 8  | 29       | LO         | female | 548.8641      | 606.8318      | 576.1126      | 17.73032 | 216      | 50       |          |
| 9  | 9  | 29       | LO         | female | 460.8693      | 493.0163      | 507.6744      | 17.80972 | 179      | 53       |          |
| 10 | 10 | 29       | LO         | female | 571.7647      | 598.6959      | 547.5487      | 17.64545 | 225      | 55       |          |
| 11 | 11 | 29       | LO         | female | 651.2894      | 616.8626      | 600.2673      | 18.73238 | 184      | 55       |          |
| 12 | 12 | 29       | LO         | male   | 468.2414      | 453.5346      | 423.8999      | 18.06160 | 153      | 47       |          |

# Struktura badania statystycznego

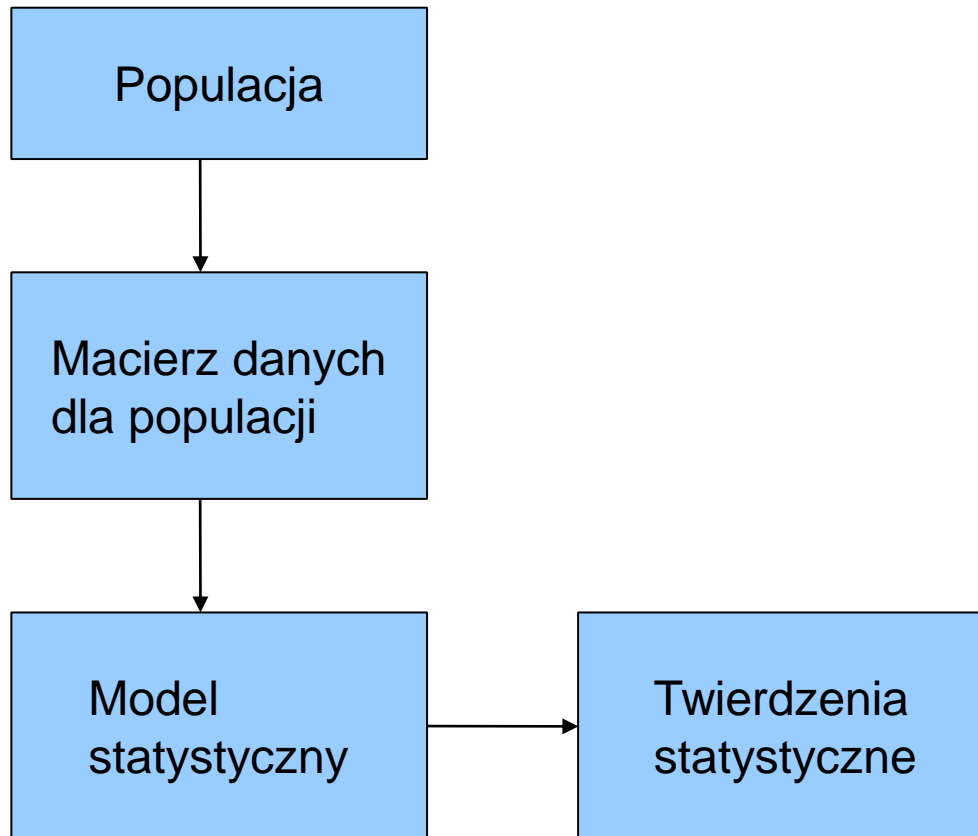


# Struktura badania statystycznego



| scoreTMR_5cat<br><chr> | n<br><int> | pct<br><dbl> |
|------------------------|------------|--------------|
| very low               | 8          | 7.766990     |
| low                    | 24         | 23.300971    |
| medium                 | 50         | 48.543689    |
| high                   | 17         | 16.504854    |
| very high              | 0          | 0.000000     |
| NA                     | 4          | 3.883495     |
| sum                    | 103        | 100.000000   |

# Struktura badania statystycznego

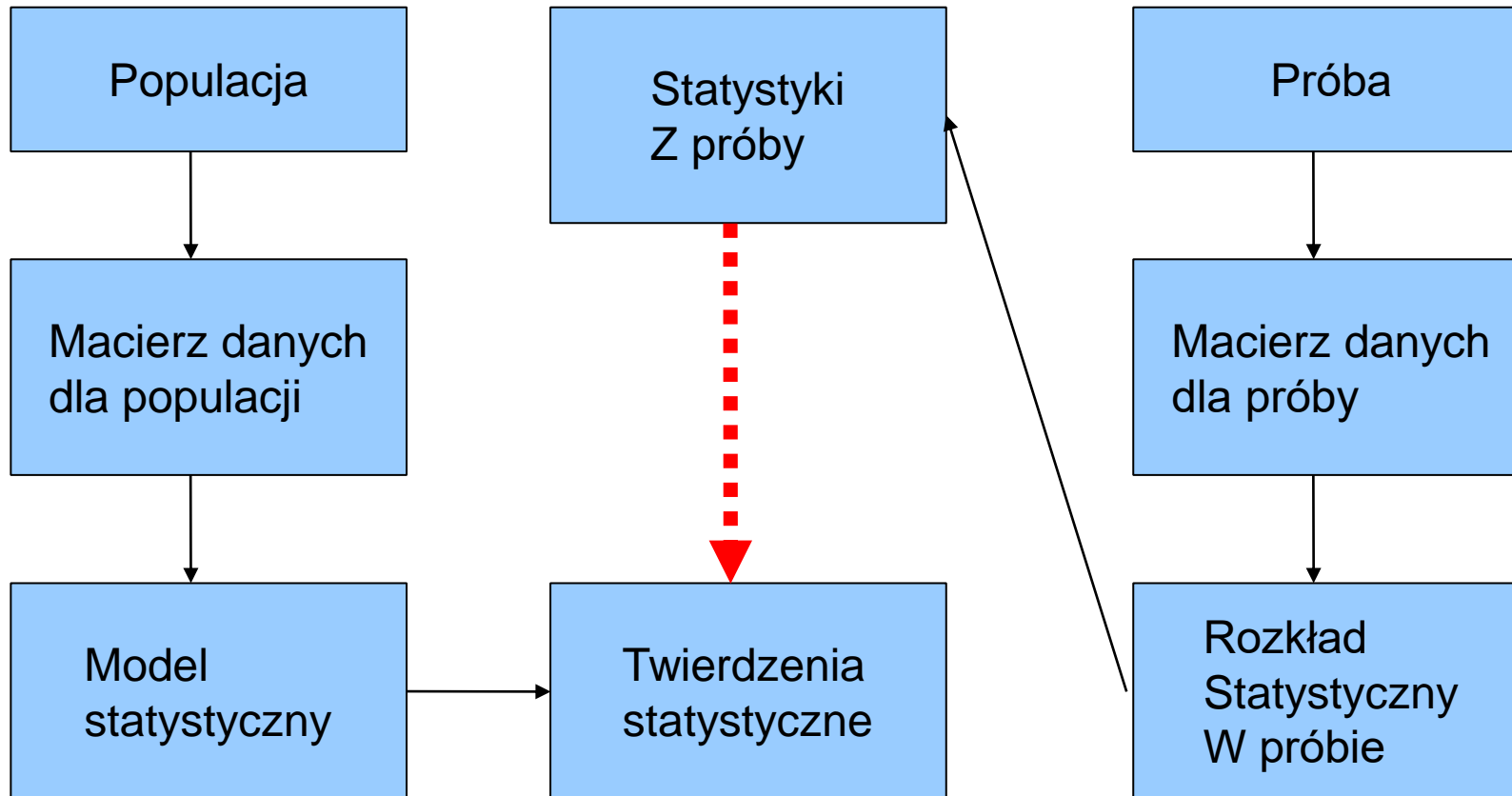




# Struktura badania statystycznego

- Średnia zarobków osób z wykształceniem wyższym jest większa niż średnia zarobków osób z wykształceniem średnim
- Osoby częściej korzystające z Internetu częściej chodzą do kina
- Osoby deklarujące większe zaufanie do innych sprawniej wykrywają oszustwa
- 40% małych firm nie przetrwa dłużej niż rok
- Rodziny, w których jest przynajmniej troje dzieci, rzadziej jeżdżą na wakacje
- Telewizory montowane na nocnej zmianie częściej są reklamowane przez klientów

# Struktura badania statystycznego



# Pomiar

- Konstrukcja skali pomiarowej polega na znalezieniu odpowiedniego modelu liczbowego. Relacje wewnątrz tego modelu powinny być izomorficzne w stosunku do pewnych relacji empirycznych.

5



4

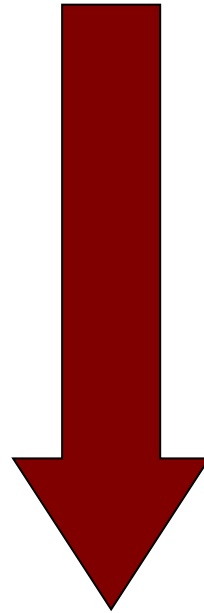


2,5



# Typy skal

- Nominalna
- Porządkowa
- Przedziałowa
- Stosunkowa
- Absolutna








Wymagania

Możliwości



# Skala nominalna

- Płeć, wyznanie, pochodzenie etniczne, marka kupowanego serka, preferowany edytor tekstowy ...

|   |   |  |   |   |   |
|---|---|--|---|---|---|
|  |  |  |  |  | = |
| 0   | 0   | 1  | 0   | 1   | ≠ |
| 1   | 1   | 0  | 1   | 0   |   |
| 1   | 1   | 2  | 1   | 2   |   |

# Skala porządkowa

- Poziom wykształcenia, większość skal ocen i postaw, częstotliwość pewnych czynności mierzona na skali od „bardzo często” do „nigdy” ...



3

30

30



2

20

10



1

10

1

= ≠

> <

# Skala przedziałowa

- Temperatura, data kalendarzowa, użyteczność ...



$= \neq$

$> <$

$$\frac{X_1 - X_2}{X_3 - X_4}$$

| Pon. | Wt.  | Śr.  | Czw. | Pt. |      |
|------|------|------|------|-----|------|
| 0    | 2    | -1   | 5    | 10  | st.C |
| 32   | 35,6 | 30,2 | 41   | 50  | st.F |

$$(10-5)/(2-0) = 2,5$$

$$(50-41)/(35,6-32) = 2,5$$

# Skala stosunkowa

- Staż pracy, wysokość zarobków, wzrost ...



$$\begin{array}{l} = \neq \\ > < \end{array} \quad \frac{X_1 - X_2}{X_3 - X_4} \quad \frac{X_1}{X_2}$$

Bolek

Lolek

2000

4000

PLN

$$4000/2000 = 2$$

560

1120

EUR

$$1120/560 = 2$$

820

1640

USD

$$1640/820 = 2$$



*CRAN*

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

*About R*

[R Homepage](#)

[The R Journal](#)

*Software*

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

*Documentation*

[Manuals](#)

[FAQs](#)

[Contributed](#)

## The Comprehensive R Archive Network

### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#) ([Debian](#), [Fedora/Redhat](#), [Ubuntu](#))
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

### Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2021-11-01, Bird Hippie) [R-4.1.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

ut R

cran.r-project.org

## RStudio Desktop 2021.09.2+382 - [Release Notes](#)

1. Install R. RStudio requires [R 3.0.1+](#).
2. Download RStudio Desktop. Recommended for your system:



Requires Windows 10 (64-bit)



## All Installers

Linux users may need to [import RStudio's public code-signing key](#) prior to installation, depending on the operating system's security policy.

RStudio requires a 64-bit operating system. If you are on a 32 bit system, you can use an [older version of RStudio](#).

 [rstudio.com/products/rstudio/download/#download](https://rstudio.com/products/rstudio/download/#download)

DZIĘKUJĘ