Metody przetwarzania i analizy danych w R

Łukasz Wawrowski

Contents

Wprowadzenie

Literatura podstawowa:

- Przemysław Biecek - *Przewodnik po pakiecie R*
- Marek Gągolewski - *Programowanie w języku R. Analiza danych, obliczenia, symulacje.*
- Garret Grolemund, Hadley Wickham - *R for Data Science* (polska wersja)

Literatura dodatkowa:

- inne pozycje po polsku
- inne pozycje po angielsku

Internet:

- R-bloggers
- rweekly

Chapter 1

Wprowadzenie do analizy danych

Podstawowe cele w analizie danych:

- porównanie grup
- prognozowanie
- klasyfikacja

Bez względu na cel analizy jest kilka pojęć, które są wspólne.

1.1 Hipoteza statystyczna

Przypuszczenie dotyczące własności analizowanej cechy, np. średnia w populacji jest równa 10, rozkład cechy jest normalny.

Formułuje się zawsze dwie hipotezy: hipotezę zerową (H_0) i hipotezę alternatywną (H_1). Hipoteza zerowa jest hipotezą mówiącą o równości:

$H_0 : \bar{x} = 10$

Z kolei hipoteza alternatywna zakłada coś przeciwnego:

$H_1 : \bar{x} \neq 10$

Zamiast znaku nierówności ($\neq$) może się także pojawić znak mniejszości ($<$) lub większości ($>$).

1.2 Poziom istotności i wartość p

Hipotezy statystyczne weryfikuje się przy określonym poziomie istotności α, który wskazuje maksymalny poziom akceptowalnego błędu (najczęściej $\alpha = 0,05$).

Większość programów statystycznych podaje w wynikach testu wartość p. Jest to prawdopodobieństwo uzyskania obserwowanych wyników przy założeniu prawdziwości hipotezy zerowej.

Generalnie jeśli $p < \alpha$ - odrzucamy hipotezę zerową.

Krytyka wartości p

1.3 Testy parametryczne i nieparametryczne

Testy statystyczne dzielą się na dwie grupy:

- parametryczne, które wymagają spełnienia założeń, ale są dokładniejsze,
- nieparametryczne, które nie wymagają tylu założeń, ale są mniej dokładne.

Chapter 2

Regresja prosta

Na podstawie danych dotyczących informacji o doświadczeniu i wynagrodzeniu pracowników zbuduj model określający ‘widełki’ dla potencjalnych pracowników o doświadczeniu równym 8, 10 i 11 lat.

Chapter 3

Regresja wieloraka

Na podstawie danych dotyczących zatrudnienia opracuj model, w którym zmienną zależną jest bieżące wynagrodzenie. Jaka cecha ma największy wpływ na tę wartość?