

ANALIZA DANYCH ANKIETOWYCH

Zadania do sprawozdania 1

1. Zaproponować i opisać badanie ankietowe dotyczące wybranego przez siebie tematu. Opis propozycji badania powinien zawierać: cel badania, definicję grupy docelowej, sposób zbierania danych, propozycje kwestionariusza, zawierającego: metryczkę i odpowiednie do wybranego tematu pytania (np. z wielokrotnymi odpowiedziami, ze skalą Likerta).
 2. Dane w pliku *Choroba.csv* zawierają następujące informacje o 196 osobach wybranych losowo z dwóch sektorów pewnego miasta: wiek, status ekonomiczny (1 – wysoki, 2 – średni, 3 – niski), sektor (1 – osoba mieszka w sektorze 1, 2 – osoba mieszka w sektorze 2), oszczędności (1 – posiada oszczędności, 0 – nie posiada oszczędności) oraz czy dana osoba jest chora (1), czy zdrowa (0).
 - (a) Sporządzić tablice licznosci dla zmiennych *Oszczed* oraz *Chory/Zdrowy* biorąc pod uwagę wszystkie dane.
 - (b) Sporządzić tabelę wielodzielczą uwzględniającą zmienną *Chory/Zdrowy* i *Sektor*.
 - (c) Sporządzić tabelę wielodzielczą uwzględniającą zmienną *Chory/Zdrowy* i *Status*.
 - (d) Przeprowadzić kategoryzację zmiennej *Wiek*.
 - (e) Sporządzić wykres kołowy i słupkowy dla zmiennej *Status*.
 - (f) Sporządzić skategoryzowane wykresy zmiennej *Chory/Zdrowy* przyjmując za zmienną kategoryzującą zmienną *Sektor*.
 - (g) Sporządzić wykresy mozaikowe odpowiadające wybranym zmiennym.
 3. Zapoznać się z funkcją *sample* (w pakiecie *stats*). Napisać fragment programu, którego celem jest wylosowanie próbki rozmiaru około $1/10$ liczby przypadków danej bazy danych (pewnej hipotetycznej lub z zadania 2.), ze zwracaniem oraz bez zwracania.
-
4. Przeprowadzić symulacje, których celem jest porównanie prawdopodobieństwa pokrycia i długości przedziałów ufności Cloppera-Pearsona, Walda i trzeciego dowolnego typu przedziału ufności zaimplementowanego w funkcji *binom.confint* pakietu *binom*. Uwzględnić poziom ufności 0.95, rozmiary próby $n \in \{30, 100, 1000\}$ i wartości prawdopodobieństwa $p \in \{0.1, 0.5, 0.8\}$. Wyniki zamieścić w tabelach i na rysunkach. Sformułować wnioski, które umożliwią praktykowi wybór konkretnego przedziału ufności do wyznaczenia jego realizacji dla konkretnych danych.
 5. Na podstawie danych zawartych w pliku *Choroba.csv*, wyznaczyć realizacje przedziałów ufności, na poziomie ufności 0.95, dla prawdopodobieństwa, że losowo wybrana

osoba z badanej populacji jest chora. Skorzystać z funkcji *binom.confint* i wyznaczyć realizacje wszystkich możliwych typów tych przedziałów, porównać te realizacje i ich długości, a następnie, w oparciu o sformułowane wnioski w rozwiązaniu zadania 4. wybrać jedną z nich (najlepszą w pewnym sensie).

6. Zapoznać się z funkcjami *binom.test* i *prop.test*.
7. Na podstawie danych zawartych w pliku *Choroba.csv*, na poziomie istotności 0.05, zweryfikować następujące hipotezy:
 - (a) prawdopodobieństwo, że losowo wybrana osoba z badanej populacji jest chora jest mniejsze bądź równe $1/2$,
 - (b) prawdopodobieństwo, że losowo wybrana osoba z sektora 1. jest chora jest równe prawdopodobieństwu, że losowo wybrana osoba z sektora 2. jest chora,
 - (c) powtórzyć punkt (a) i (b), ale dla osoby o średnim statusie ekonomicznym.

W każdym z powyższych punktów podać wartość poziomu krytycznego (p value) i sformułować odpowiedź.

8. Przeprowadzić symulacje, których celem jest
 - (a) oszacowanie prawdopodobieństwa błędu I-go rodzaju testu dokładnego i testu asymptotycznego, przy przyjętym poziomie istotności,
 - (b) porównanie mocy testu dokładnego i testu asymptotycznego (dla różnych wartości alternatyw),

w przypadku weryfikacji hipotezy zerowej $H_0 : \vartheta = 0.5$, przy hipotezie alternatywnej $H_1 : \vartheta \neq 0.5$, na poziomie istotności 0.05, gdzie ϑ jest prawdopodobieństwem sukcesu. Uwzględnić rozmiary próby $n \in \{30, 100, 1000\}$ i alternatywy $\vartheta \in \{0.3, 0.4, 0.6\}$. Wyniki zamieścić w tabelach i na rysunkach. Sformułować wnioski, które umożliwią praktykowi wybór jednego z dwóch powyższych testów dla konkretnych danych.

Alicja Jokiel-Rokita

1 października 2023