ANALIZA DANYCH ANKIETOWYCH, SEMESTR LETNI 2023/2024

Zadania do sprawozdania 1

Część I

zadanie 1. W pewnej dużej agencji reklamowej przeprowadzono ankietę mającą na celu ocenę poziomu satysfakcji z pracy. Wzięło w niej udział dwieście losowo wybranych osób (losowanie proste ze zwracaniem).

W pliku "ankieta.csv" umieszczono odpowiedzi na kilka z zadanych pytań:

- "W jakim dziale jesteś zatrudniony?" - zmienna **DZIAŁ** przyjmująca wartości: **HR** (Dział obsługi kadrowo-płacowej), **IT** (Dział utrzymania sieci i systemów informatycznych), **DK** (Dział Kreatywny) lub **DS** (Dział Strategii),
- "Jak długo pracujesz w firmie?" - zmienna **STAŻ** przyjmująca wartości: **1** (Poniżej jednego roku), **2** (Między jednym rokiem a trzema latami) lub **3** (Powyżej trzech lat),
- "Czy pracujesz na stanowisku menedżerskim?" - zmienna **CZY_KIER** przyjmująca wartości: **Tak** (Stanowisko menedżerskie) lub **Nie** (Stanowisko inne niż menedżerskie).
- "Jak bardzo zgadzasz się ze stwierdzeniem, że firma pozwala na elastyczne godziny pracy tym samym umożliwiając zachowanie równowagi między pracą a życiem prywatnym?" - zmienna **PYT_1** przyjmująca wartości: **-2** (zdecydowanie się nie zgadzam), **-1** (nie zgadzam się), **0** (nie mam zdania), **1** (zgadzam się), **2** (zdecydowanie się zgadzam).
- "Jak bardzo zgadzasz się ze stwierdzeniem, że twoje wynagrodzenie adekwatnie odzwierciedla zakres wykonywanych przez ciebie obowiązków?" - zmienna **PYT_2** przyjmująca wartości: **-2** (zdecydowanie się nie zgadzam), **-1** (nie zgadzam się), **1** (zgadzam się), **2** (zdecydowanie się zgadzam).

Dodatkowo w ramach metryczki ankietowani zostali poproszeni o wskazanie swojego wieku - zmienna **WIEK** przyjmująca wartości numeryczne, oraz wskazanie płci - zmienna **PŁEĆ** przyjmująca wartość **Kobieta** lub **Mężczyzna**.

Kilka tygodni później przeprowadzono rewizję wynagrodzeń, w wyniku której część pracowników otrzymała podwyżki. Ankietowanych biorących udział w badaniu poproszono wówczas o ponowną odpowiedź na pytanie dotyczące zadowolenia z wynagrodzenia - zmienna **PYT_3**.

1. Wczytaj dane i przygotuj je do analizy. Zadbaj o odpowiednie typy zmiennych, zweryfikuj czy przyjmują wartości zgodne z powyższym opisem, zbadaj czy nie występują braki w danych.
2. Utwórz zmienną **WIEK_KAT** przeprowadzając kategoryzację zmiennej **WIEK** korzystając z następujących przedziałów: do 35 lat, między 36 a 45 lat, między 46 a 55 lat, powyżej 55 lat.

3. Sporządź tablice liczności dla zmiennych: **DZIAŁ**, **STAŻ**, **CZY_KIER**, **PŁEĆ**, **WIEK_KAT**.

4. Sporządź wykresy kołowe oraz wykresy słupkowe dla zmiennych: **PYT_1** oraz **PYT_2**.

5. Sporządź tablice wielodzielcze dla par zmiennych: **PYT_1** i **DZIAŁ**, **PYT_1** i **STAŻ**, **PYT_1** i **CZY_KIER**, **PYT_1** i **PŁEĆ** oraz **PYT_1** i **WIEK_KAT**.

6. Sporządź tablicę wielodzielczą dla pary zmiennych: **PYT_2** i **PYT_3**.

7. Utwórz zmienną **CZY_ZADOW** na podstawie zmiennej **PYT_2** łącząc kategorie "nie zgadzam się" i "zdecydowanie się nie zgadzam" oraz "zgadzam się" i "zdecydowanie się zgadzam".

8. Korzystając z funkcji *mosaic* z biblioteki *vcd*, sporządź wykresy mozaikowe odpowiadające parom zmiennych: **CZY_ZADOW** i **DZIAŁ**, **CZY_ZADOW** i **STAŻ**, **CZY_ZADOW** i **CZY_KIER**, **CZY_ZADOW** i **PŁEĆ** oraz **CZY_ZADOW** i **WIEK_KAT**. Czy na podstawie uzyskanch wykresów można postawić pewne hipotezy dotyczące realicji między powyższymi zmiennymi? Spróbuj sformułować kilka takich hipotez.

Część II

zadanie 2. Zapoznaj się z biblioteką *likert* i dostępnymi tam funkcjami *summary* oraz *plot* (wykresy typu "bar", "heat" oraz "density"), a następnie zilustruj odpowiedzi na pytanie "Jak bardzo zgadzasz się ze stwierdzeniem, że firma pozwala na (...)?" (zmienna **PYT_1**) w całej badanej grupie oraz w podgrupach ze względu na zmienną **CZY_KIER**.

zadanie 3. Zapoznaj się z funkcją *sample* z biblioteki *stats*, a następnie wylosuj próbkę o liczności 10% wszystkich rekordów z pliku "ankieta.csv" w dwóch wersjach: ze zwracaniem oraz bez zwracania.

zadanie 4. Zaproponuj metodę symulowania zmiennych losowych z rozkładu dwumianowego. Napisz funkcję do generowania realizacji, a następnie zaprezentuj jej działanie porównując wybrane teoretyczne i empiryczne charakterystyki dla przykładowych wartości paramertów rozkładu: n i p.

zadanie 5. Zaproponuj metodę symulowania wektorów losowych z rozkładu wielomianowego. Napisz funkcję do generowania realizacji, a następnie zaprezentuj jej działanie porównując wybrane teoretyczne i empiryczne charakterystyki dla przykładowych wartości paramertów rozkładu: n i $\mathbf{p}$.

Część III oraz IV

zadanie 6. Napisz funkcję do wyznaczania realizacji przedziału ufności Cloppera-Pearsona. Niech argumentem wejściowym będzie poziom ufności, liczba sukcesów i liczba prób lub poziom ufności i wektor danych (funkcja powinna obsługiwać oba przypadki).

zadanie 7. Korzystając z funkcji napisanej w zadaniu 6. wyznacz realizacje przedziałów ufności dla prawdopodobieństwa, że pracownik jest zadowolony z wynagrodzenia w pierwszym badanym okresie oraz w drugim badanym okresie. Skorzystaj ze zmiennych **CZY_ZADW** oraz **CZY_ZADW_2** (utwórz zmienną analogicznie jak w zadaniu 1.7). Przyjmij $1-\alpha = 0.95$.

zadanie 8. Zapoznaj się z funkcjami *rbinom* z biblioteki *stats* oraz *binom.confint* z biblioteki *binom*.

zadanie 9. Przeprowadź symulacje, których celem jest porównanie prawdopodobieństwa pokrycia i długości przedziałów ufności Cloppera-Pearsona, Walda i trzeciego dowolnego typu zaimplementowanego w funkcji *binom.confint*. Rozważ $1-\alpha = 0.95$, rozmiar próby $n \in \{30, 100, 1000\}$ i różne wartości prawdopodobieństwa p. Wyniki umieść na wykresach i sformułuj wnioski, które dla konkretnych danych ułatwią wybór konkretenego typu przedziału ufności.

Część V

zadanie 10. Zapoznaj się z funkcjami *binom.test* oraz *prop.test* z biblioteki *stats*.

zadanie 11. Dla danych z pliku "ankieta.csv" korzystając z funkcji z zadania 10., przyjmując $1-\alpha = 0.95$, zweryfikuj następujące hipotezy i sformułuj wnioski:

1. Prawdopodobieństwo, że w firmie pracuje kobieta wynosi 0.5.
2. Prawdopodbieństwo, że pracownik jest zadowolony ze swojego wynagrodzenia w pierwszym badanym okresie jest większe bądź równe 0.7.
3. Prawdopodobieństwo, że kobieta pracuje na stanowisku menedżerskim jest równe prawdopodobieństwu, że mężczyzna pracuje na stanowisku menedżerskim.
4. Prawdopodobieństwo, że kobieta jest zadowolona ze swojego wynagrodzenia w pierwszym badanym okresie jest równe prawdopodobieństwu, że mężczyzna jest zadowolony ze swojego wynagrodzenia w pierwszym badanym okresie.
5. Prawdopodobieństwo, że kobieta pracuje w dziale obsługi kadrowo-płacowej jest większe lub równe prawdopodobieństwu, że mężczyzna pracuje w dziale obsługi kadrowo-płacowej.

zadanie 11. Wyznacz symulacyjnie moc testu dokładnego oraz moc testu asymptotycznego w przypadku weryfikacji hipotezy zerowej $H_0 : p = 0.9$ przeciwko $H_1 : p \neq 0.9$ przyjmując wartość $1-\alpha = 0.95$. Uwzględnij różne wartości alternatyw i różne rozmiary próby. Sformułuj wnioski.

Zadania dodatkowe

zadanie *1. Wyznacz granice asymptotycznego przedziału ufności dla prawdopodobieństwa sukcesu bazując na przekształceniu logit korzystając z metody delta. Zaimplementuj metodę oraz porównaj wyniki z funkcją *binom.confint*.