ANALIZA DANYCH ANKIETOWYCH

Zadania do sprawozdania 1

Dane zawarte w pliku *personel.csv* nie są danymi rzeczywistymi, ale wygenerowanymi w taki sposób, żeby można było je wykorzystać do zadań we wszystkich sprawozdaniach. Załóżmy, że zostały one uzyskane w wyniku badań ankietowych losowo (losowanie proste ze zwracaniem) wybranych dwustu pracowników pewnej wielkiej korporacji. Pytania ankietowe były następujące (w nawiasach podane są przyjęte nazwy zmiennych i symbole kodowania odpowiedzi w zbiorze danych).

1. (D) Pracuję
   (a) w dziale zaopatrzenia (Z),
   (b) w dziale produkcyjnym (P),
   (c) w dziale sprzedaży (w tym marketingu) (S),
   (d) w dziale obsługi kadrowo-płacowej (O).
2. (S) Pracuję na stanowisku kierowniczym (tzn. kieruję pracą więcej niż pięciu osób)
   (a) tak (1),
   (b) nie (0).
3. (A) Atmosfera w miejscu pracy jest bardzo dobra
   (a) zdecydownie się nie zgadzam (-2),
   (b) nie zgadzam się (-1),
   (c) trudno powiedzieć (0),
   (d) zgadzam się (1),
   (e) zdecydowanie się zgadzam (2).
4. (W1 i W2) Jestem zadowolona/y ze swojego wynagrodzenia
   (a) zdecydownie się nie zgadzam (-2),
   (b) nie zgadzam się (-1),
   (c) zgadzam się (1),
   (d) zdecydowanie się zgadzam (2).

Metryczka

1. (P) Płeć
   (a) kobieta (K),
   (b) mężczyzna (M).

2. (Wiek) Wiek
   (a) do 25 lat (1),
   (b) od 26 do 35 lat (2),
   (c) od 36 do 50 lat (3),
   (d) powyżej 50 lat (4).
3. (Wyk) Wykształcenie
   (a) zawodowe (1),
   (b) średnie (2),
   (c) wyższe (3).

W zbiorze danych *personel.csv* odpowiedź na pierwsze pytanie ankietowe znajduje się w pierwszej kolumnie danych o nazwie D, a na drugie pytanie ankietowe - w drugiej kolumnie o nazwie S. Tym samym pracownikom zadano pytanie 3. i 4. ankiety po kolejnym roku pracy, w którym zorganizowano dwa spotkania integracyjne, dlatego w zbiorze danych mamy kolumny zmiennych o nazwach A1 i A2, odpowiadające ocenie atmosfery w pracy odpowiednio w pierwszym badanym okresie i drugim badanym okresie oraz W1 i W2, odpowiadające ocenie wynagrodzenia odpowiednio w pierwszym badanym okresie i drugim badanym okresie. Odpowiedzi na pytania metryczki zamieszczone są w kolejnych kolumnach o nazwach P (płeć), Wiek, Wyk (wykształcenie).

Część I

1. Sporządzić tablice liczności dla zmiennych *A1* oraz *W1* biorąc pod uwagę wszystkie dane, jak również w podgrupach ze względu na zmienną *D* (dział), *P* (płeć) i *Wyk* (wykształcenie).
2. Sporządzić tabelę wielodzielczą uwzględniającą zmienną *W1* i *P*, *W1* i *S* oraz *A1* i *D*.
3. Sporządzić wykres kołowy i słupkowy dla zmiennej *W1* i *W2*.
4. Korzystając z funkcji *mosaic* biblioteki *vcd*, sporządzić wykresy mozaikowe odpowiadające parom zmiennych
   (a) *D* i *A1*,
   (b) *D* i *W1*,
   (c) *S* i *P*,
   (d) *P* i *W1*.

   Uwaga. Każdy rysunek i tabela musi być opisana i musi być do niej odwołanie w tekście sprawozdania.

Część II

5. 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), ze zwracaniem oraz bez zwracania.

6. Zapoznać się z funkcjami *plot.likert*, *summary.likert*, *likert.bar.plot*, *likert.density.plot* biblioteki *likert*. Następnie korzystając z tych funkcji zilustrować dane dotyczące oceny atmosfery w pracy w całej badanej grupie i w podgrupach ze względu na dział i ze względu na płeć.

7. Napisać deklarację funkcji, której wartością będzie realizacja przedziału ufności Cloppera-Pearsona, na zadanym poziomie ufności (argumentem funkcji może być wektor danych lub liczba sukcesów i liczba prób oraz oczywiście poziom ufności $1-\alpha$). Następnie korzystając z tej funkcji wyznaczyć realizacje przedziałów ufności na poziomie ufności dla prawdopodobieństwa, że pracownik jest zadowolony ze swojego wynagrodzenia w pierwszym badanym okresie (zliczamy odpowiedzi "zgadzam się" i "zdecydowanie zgadzam się" zmiennej *W1*) w całej badanej grupie pracowniczej i w podgrupach ze względu na dział, w którym pracuje i ze względu na stanowisko. Sprawdzić, czy uzyskane realizacje przedziałów ufności wyznaczone w oparciu o napisaną procedurę są takie same jak odpowiednie wartości funkcji *binom.confint* pakietu *binom*.

Część III

8. Zaproponować algorytm generowania liczb z rozkładu dwumianowego i udowodnić, że jest poprawny. Napisać program do generowania tych liczb zgodnie z zaproponowanym algorytmem. (W pakiecie R dostępna jest funkcja *rbinom*.)

9. 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 różne wartości prawdopodobieństwa p. Wyniki zamieścić w tabelach i na rysunkach. Sformułować wnioski, które umożliwią praktykowi wybór konkretnego przedziału ufności do wyznaczenia jego realizacji dla konkretnych danych.

Część IV

10. Zapoznać się z funkcjami *binom.test* i *prop.test*.

11. Korzystając z funkcji z zadania 10., na poziomie istotności 0.05, zweryfikować następujące hipotezy i napisać odpowiednie wnioski.

    (a) prawdopodobieństwo, że w korporacji pracuje kobieta wynosi 0.5,

    (b) prawdopodbieństwo, że pracownik jest zadowolony ze swojego wynagrodzenia jest większe bądź równe 0.8,

(c) prawdopodobieństwo, że kobieta pracuje na stanowisku kierowniczym jest równe prawdopodobieństwu, że mężczyzna pracuje na stanowisku kierowniczym,

(d) prawdopodobieństwo, że kobieta jest zadowolona ze swojego wynagrodzenia jest równe prawdopodobieństwu, że mężczyzna jest zadowolony ze swojego wynagrodzenia,

(e) prawdopodobieństwo, że kobieta pracuje w dziale obsługi kadrowo-płacowej jest większe lub równe prawdopodobieństwu, że mężczyzna pracuje w dziale obsługi kadrowo-płacowej.

Alicja Jokiel-Rokita

24 lutego 2023