Analiza danych ankietowych
Lista 1-3

1. Sklasyfikować każdą zmienną jako nominalną, porządkową, interwałową albo proporcjonalną.

   a) Preferencje polityczne w Wielkiej Brytanii: Partia Pracy, Partia Konserwatywna, Socjaldemokracja.

   b) Ocena lęków: brak, łagodne, umiarkowane, ciężkie, bardzo ciężkie.

   c) Przeżycie pacjenta w miesiącach.

   d) Lokalizacja kliniki: Londyn, Boston, Madison, Rochester, Montreal.

   e) Odpowiedź na chemioterapię: całkowita eliminacja guza, częściową redukcja, stabilność, postęp.

   f) Ulubiony napój: woda, sok, mleko, napój bezalkoholowy, piwo, wino.

   g) Ocena poziomu zapasów firmy: zbyt niski, prawie dobry, za wysoki.

2. Zaprojektować badania ankietowe dotyczące wybranego przez siebie tematu. Projekt powinien zawierać:

   a) cel badań;

   b) definicję grupy docelowej;

   c) sposób zebrania danych;

   d) propozycję kwestionariusza, zawierającego: metryczkę, pytanie z wielokrotnymi odpowiedziami, pytanie ze skalą Likerta oraz pytanie ze skalą dyferencjału semantycznego (maksymalna liczba wszystkich zmiennych – 10).

   e) wygenerować losowo wyniki 100 ankiet i wpisać je do skoroszytu;

   f) wykonać analizę danych i zaprezentować wyniki.

3. Wygenerować 1000 liczb według rozkładu zerojedynkowego z dowolnie ustalonym parametrem π. Następnie utworzyć 100 podgrup 10-elementowych:

$$x_{i1}, ..., x_{i10}, \quad i = 1, ..., 100.$$

   Niech $\overline{x}_i = \frac{1}{10}\sum_{j=1}^{10} x_{ij}$, gdzie $i = 1, ..., 100$. Narysować histogram liczb $\overline{x}_i$ po odpowiednim unormowaniu. Sprawdzić zgodność z rozkładem normalnym (np. stosując test Shapiro-Wilka). Następnie porównać podstawowe wskaźniki (średnia, odchylenie standardowe, mediana oraz kwartle dolny i górny) wyznaczone ze 100-elementowej próby i wskaźniki z rozkładu standardowego normalnego.

4. Niech $X_1, ..., X_n$ będzie prostą próbą losową z rozkładu dwumianowego $Bin(1, \pi)$. Skonstruować test ilorazu wiarogodności, test Walda i test wynikowy do testowania

$$H_0 : \pi = \frac{1}{3} \text{ przeciwko } H_a : \pi \neq \frac{1}{3}.$$

5. Korporacja zatrudniająca ponad 2000 pracowników ma zamiar wybudować parking. Panuje przekonanie, że ponad 60% pracowników przyjeżdza do pracy samochodem. Sprawdzić czy to przekonanie jest prawdziwe, jeśli spośród 250 losowo wybranych osób, 206 zadeklarowało, że przjeżdża do pracy swoim samochodem. Przyjąć poziom istotności $\alpha = 0.01$.

6. Ojciec Mendel wyhodował nasiona potomków grochu o genotypie Aa, następnie wysiał 8023 takie nasiona i otrzymał 2001 roślin grochu zielonego (o genotypie aa). Stosując test dla proporcji sprawdzić słuszność prawa Mendla, to znaczy, że proporcja ziaren grochu zielonego wynosi 0.25.

7. Pewne ugrupowanie polityczne było przekonane, że poparcie Polaków dla wejścia ich kraju do UE nigdy nie przekroczy 53%. Przeprowadzona w czerwcu 2000r. ankieta wśród 1000 dorosłych Polaków dała 57% poparcie starań Polski do UE. Czy hipoteza wspomnianego ugrupowania była słuszna?

8. Niech $\mathbf{Y} = (Y_1, ..., Y_k) \sim Mult(n, \pi)$, gdzie $\pi = (\pi_1, ..., \pi_k)$.

   a) Pokazać, że

$$Cov(Y_i, Y_j) = \begin{cases} -n\pi_i\pi_j & \text{gdy} \quad i \neq j \\ n\pi_i(1-\pi_i) & \text{gdy} \quad i = j \end{cases}$$

   b) Znaleźć $Corr(Y_i, Y_j)$ i pokazać, że w przypadku $k = 2$, $Corr(Y_1, Y_2) = -1$.

9. Genotypy AA, Aa oraz aa pojawiają się z prawdopodobieństwami $\{\theta^2, 2\theta(1-\theta), (1-\theta)^2\}$. Dla wielomianowej próby o rozmiarze n uzyskano liczności $\{n_1, n_2, n_3\}$.

   a) Wyznaczyć logarytm funkcji wiarogodności i estymator ML parametru θ.

   b) Wyznaczyć informację Fishera $\mathcal{I}(\theta)$ a następnie uzyskać asymptotyczny błąd standardowy estymatora $\hat{\theta}$.

10. Niech $\mathbf{Y} = (Y_1, Y_2, ..., Y_k)$ będzie wektorem k niezależnych zmiennych losowych o rozkładzie Poissona z parametrami μ_j, $j = 1, ..., k$, odpowiednio. Wówczas warunkowym rozkład $\mathbf{Y}$ pod warunkiem $\sum_{j=1}^{k} Y_j = n$ jest $Mult(n, \pi)$, gdzie $\pi = (\pi_1, ..., \pi_k)$ i $\pi_j = \mu_j/(\sum_j \mu_j)$.

11. Niech $X_1, ..., X_n$ będzie prostą próbą losową z rozkładu Poissona $\mathcal{P}(\mu)$. Skonstruować test ilorazu wiarogodności, test Walda i test wynikowy do testowania

$$H_0 : \mu = 2 \text{ przeciwko } H_a : \mu \neq 2.$$

12. Niech $y_1, ..., y_n$ są obserwacjami próby prostej z rozkładu Poissona $\mathcal{P}(\mu)$. Skonstruować asymptotyczne przedziały ufności dla μ:

    a) metodą Walda;

    b) metodą statystyki wynikowej;

    c) metodą statystyki ilorazu wiarogodności.

13. Wykorzystując aproksymację $\frac{\sqrt{n}(\hat{\pi}-\pi)}{\sqrt{\pi(1-\pi)}} \sim \mathcal{N}(0, 1)$ wyznaczyć przedział ufności parametru π w rozkładzie $Bin(n, \pi)$ (do obliczeń można wykorzystać pakiet matematyczny lub Wolfram Alpha). Jest to oczywiście przedział ufności oparty na statystyce wynikowej. Stosując uzyskany wzór wyznaczyć przedział ufności dla parametru π, jeśli wylosowano wartość $y = 5$ z rozkładu $Bin(50; \pi)$. Porównać uzyskany wynik ze standardowym przedziałem ufności Walda. Przyjąć 95% poziom ufności.

14. Asymptotyczne przedziały ufności typu Walda można konstruować wykorzystując następujący wzór

$$\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{\mathcal{I}(\hat{\theta})}},$$

    gdzie $I(\hat{\theta}) = -E(L''(Y; \hat{\theta}))$. Funkcja $I(\theta)$ nazywana jest informacją Fishera. Wykorzystując powyższy wzór wyznaczyć przedział ufności dla estymatora $\frac{Y}{n}$, gdzie $Y \sim Bin(n, \pi)$.

15. Stosując przekształcenie logistyczne wyznaczyć przedział ufności parametru π rozkładu $Bin(20, \pi)$, jeśli wylosowano liczbę $x = 18$. Porównać uzyskany wynik ze standardowym przedziałem ufności dla rozkładu dwumianowego. Przyjąć 95% poziom ufności.

16. Ilu respondentów trzeba zbadać, aby na poziomie ufności 99% całkowity błąd oszacowania proporcji nie przekraczał 0,01.

17. Porównujemy nowy lek z lekiem wcześniej stosowanym (standardowym). Niech π oznacza prawdopodobieństwo, że nowy lek zostanie oceniony lepiej. Chcemy oszacować π oraz testować hipotezę $H_0 : \pi = 0.5$ przeciwko $H_a : \pi \neq 0, 5$. W 20 niezależnych obserwacjiach, nowy lek jest lepszy za każdym razem.

    a) Znajdź i narysuj funkcję wiarogodności. Podaj wartość estymatora ML parametru π.

    b) Przeprowadź test Walda i skonstruuj tą metodą 95% przedział ufności dla π. Czy jest on rozsądny?

    c) Przeprowadź test wynikowy, podaj P-wartość. Utwórz 95% przedział ufności.

    d) Przeprowadź test ilorazu wiarogodności i skonstruuj tą metodą 95% przedział ufności.

    e) Skonstruuj dokładny test dwumianowy i 95% przedział ufności (Cloppera-Pearsona).

    f) Załóżmy, że badacze potrzebowali wystarczająco dużej próbki, aby oszacować prawdopodobieństwo wyboru nowego leku z dokładnością 0,05, na pozionie ufności 0,95. Jak duża powinna być próbka, jeśli prawdziwym prawdopodobieństwem jest 0.9?

18. Wykorzystując związek rozkładów beta i F-Snedecora pokazać, że przedział Cloppera-Pearsona (1934) ma postać

$$\left[1 + \frac{n-y+1}{yF_{2y,2(n-y+1)}(1-\alpha/2)}\right]^{-1} < \pi < \left[1 + \frac{n-y}{(y+1)F_{2(y+1),2(n-y)}(\alpha/2)}\right]^{-1},$$

    gdzie $F_{a,b}(c)$ jest kwantylem z rozkładu F-Snedecora rzędu $1 - c$.

19. Na jednym wykresie słupkowym porównać długości przedziałów ufności dla nieznanego parametru π skonstruowanych: metodą Walda, statystyki wynikowej (*score*) oraz Cloppera-Pearsona dla jedenastu wartości $y_i = 2i$; $i = 0, 1, ..., 10$ wylosowanych z rozkładu $Bin(20; \pi)$ (przyjąć poziom ufności 95%).

20. Przedstaw na rysunku prawdopodobieństwa pokrycia 95% przedziałami ufności: Walda, statystyki wynikowej (*score*) i Cloppera-Pearsona dla zmieniającego się parametru π w rozkładzie $Bin(20; \pi)$.

21. Wśród losowo wybranych 200 respondentów przeprowadzono badania poziomu wykształcenia w województwie dolnośląskim. Uzyskano następujące wyniki:

    - wyższe - 43
    - średnie zawodowe i policealne - 52
    - średnie ogólnokształcące - 16
    - zasadnicze zawodowe - 64
    - podstawowe i niepełne podstawowe - 25

    Wyznaczyć przedziały ufności dla odsetka ludności posiadających wykształcenie:

    a) podstawowe lub zawodowe

    b) średnie.

22. Zaprogramować formatkę w Excelu do wyznaczania jednoczesnych przedziałów ufności (metoda Golda) dla parametrów $\pi_1, ..., \pi_4$ rozkładu wielomianowego $Mult(n, (\pi_1, ..., \pi_4))$. Uwzględnić dane wejściowe: n - liczba ankiet, $y_1, ..., y_4$ - liczba odpowiedzi na każde z pytań, $1 - \alpha$- poziom ufności w procentach. Ponadto zaprogramować formatkę do wyznaczania osobnych przedziałów ufności dla tych samych danych.

23. Wykorzystując formatki wyznaczyć jednoczesne i osobne przedziały ufności parametrów $\pi_1, ..., \pi_4$ rozkładu $Mult(100, (\pi_1, ..., \pi_4))$, jeśli w badaniach ankietowych w pytaniu z czterema odpowiedziami uzyskano następujące wyniki: $x_1 = 15$; $x_2 = 50$; $x_3 = 10$; $x_4 = 25$. Przyjąć 95% poziom ufności.

24. Pobrać ze strony: http://www.parlamentarny.pl/sondaze/ najnowsze wyniki sondaży partii politycznych. Wykorzystując formatkę dla jednoczesnych przedziałów ufności zbadać, które relacje (większe, mniejsze poparcie) pomiędzy wynikami partii są istotne statystycznie na poziomie ufności 99%. Przyjąć wielkość próby $n = 1000$.

25. Zaprojektować ankietę do badań omnibusowych zawierającą trzy warianty odpowiedzi do zbadania potencjału rynku rowerów w Polsce w sezonie 2017r, (sprzedaż od maja do września). Przyjąć liczebność próby $n = 1000$ oraz rozkład odpowiedzi: $y_1 = 40$ (planuje kupić), $y_2 = 900$ (na pewno nie kupi), $y_3 = 60$ (nie wie czy kupi). Do wyliczenia jednoczesnych przedziałów ufności wykorzystać formatkę z Zadania 22. Przyjąć 95% poziom ufności. Na podstawie wyliczonych przedziałów oszacować potencjał rynku rowerów w 2017 roku w różnych wariantach (optymistyczny,realny, pesymistyczny). Grupa wiekowa respondentów od 10 do 50 lat. Potrzebne dane do oszacowania znaleźć w Internecie.

26. Używając nierówności Bonferroniego wykazać, że prawdopodobieństwo tego, że k przedziałów ufności uzyskanych metodą Goodmana jednocześnie zawiera wszystkie parametry $\pi_1, ..., \pi_k$ rozkładu $Mult(n, (\pi_1, ..., \pi_k))$, dla dużych rozmiarów prób, wynosi co najmniej $1 - \alpha$.

27. Korzystając z dowolnego pakietu matematycznego (np. Wolfram Alpha) wyznaczyć jednoczesne przedziały ufności dla danych z Zadania 23 metodą Quesenberry'ego i Hursta oraz metodą Goodmana. Porównać wyniki.

28. Korzystając z dowolnego pakietu matematycznego (np. Mathematica, MATLAB, itp.) narysować obszar ufności dla parametrów π_1, π_2 rozkładu $Mult(30, (\pi_1, \pi_2, \pi_3))$ jeśli wylosowano następujące wyniki $y_1 = 5$, $y_2 = 15$.