

Analiza danych ankietowych
Lista 4-5

1. Zaprogramować formatki w Excelu do wyznaczania:
 - a) przedziałów ufności dla różnicy dwóch proporcji w przypadku dużych prób (dane wejściowe: zaobserwowane liczby sukcesów, liczebności prób, poziom istotności w %),
 - b) przedziałów ufności dla różnicy dwóch proporcji w przypadku małych prób (dane wejściowe: zaobserwowane liczby sukcesów, liczebności prób, poziom istotności w %).
2. Zaprogramować formatkę w Excelu do testowania hipotez niezależności zmiennych losowych za pomocą testu χ^2 , których zaobserwowane wartości wpisane są do tabeli 2×2 . Dane wejściowe: tabela z wynikami, poziom istotności w %. Dane wyjściowe: komunikat, czy hipotezę odrzucamy, czy nie.
3. Spośród sprzedanych 50 telewizorów pewnego typu w okresie gwarancyjnym, 8 wymagało naprawy, natomiast spośród 35 telewizorów drugiego typu naprawy gwarancyjnej wymagało 5.
 - a) Czy można na poziomie istotności $\alpha = 0.05$ uznać za słuszną hipotezę, że frakcje napraw gwarancyjnych telewizorów tych dwóch typów są jednakowe, jeśli alternatywną jest hipoteza, że procent telewizorów naprawianych w okresie gwarancyjnym dla pierwszego typu jest wyższy niż dla drugiego?
 - b) Wykorzystując formatkę z Zadania 1 wyznaczyć przedział ufności dla różnicy proporcji (frakcji).
 - c) Wykorzystując formatkę z Zadania 2 wykonać dla powyższych danych test niezależności χ^2 .
4. Rozważyć test porównujący dwie proporcje π_1 i π_2 ($H_0 : \pi_1 = \pi_2$) oparty na statystyce

$$Z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}.$$

Założyć, że rozmiary prób są takie same i wynoszą $n_i = n/2$ dla $i = 1, 2$. Przyjąć poziom istotności $\alpha = 0.05$. Ponadto założyć, że oczekuje się dla obu prób wartości proporcji $\pi_i \approx 0.6$.

- a) Obliczyć moc testu dla $n = 50$ i $n = 200$, jeśli $\pi_1 - \pi_2 = 0.1$ oraz $\pi_1 - \pi_2 = 0.2$.
 - b) Wyznaczyć n takie, że prawdopodobieństwo popełnienia błędu II rodzaju (β) wynosi 0.10, jeśli prawdziwe wartości proporcji π_1 i π_2 wynoszą około 0.60 i 0.70, odpowiednio.
5. Używając metody delta pokazać, że przedział ufności Walda dla *logit* funkcji parametru π rozkładu dwumianowego jest postaci

$$\log[\hat{\pi}/(1 - \hat{\pi})] \pm z_{\alpha/2}/\sqrt{n\hat{\pi}(1 - \hat{\pi})}.$$

6. Wyznaczyć przybliżoną wariancję $\sigma^2(\log \widehat{RR})$ estymatora względnego ryzyka oraz przybliżoną wariancję $\sigma^2(\log \widehat{OR})$ estymatora ilorazu szans.
7. (*Własności ilorazu szans.*) Dana jest tablica prawdopodobieństw 2×2

X	Y	
	y_1	y_2
x_1	π_{11}	π_{12}
x_2	π_{21}	π_{22}

i odpowiadający jej iloraz szans θ . Pokazać, że prawdziwe są nierówności:

$$\begin{aligned} \theta > 1 &\Leftrightarrow P(Y = y_1|X = x_1) > P(Y = y_1|X = x_2), \\ \theta > 1 &\Leftrightarrow P(X = x_1|Y = y_1) > P(X = x_1|Y = y_2), \\ \theta < 1 &\Leftrightarrow P(Y = y_1|X = x_1) < P(Y = y_1|X = x_2), \\ \theta < 1 &\Leftrightarrow P(X = x_1|Y = y_1) < P(X = x_1|Y = y_2). \end{aligned}$$

8. Udowodnić, że dla każdego $\theta > 0$ i dla każdych $0 < p < 1$ i $0 < q < 1$ istnieje tablica prawdopodobieństw 2×2

X	Y	
	y_1	y_2
x_1	π_{11}	π_{12}
x_2	π_{21}	π_{22}

taka, że jej iloraz szans jest równy θ i taka, że $\pi_{1+} = \pi_{11} + \pi_{12} = p$ oraz $\pi_{+2} = \pi_{12} + \pi_{22} = q$. Wyznaczyć taką tablicę dla $\theta = 1.5$, $p = 0.2$, $q = 0.6$.

9. Obliczyć iloraz szans θ dla danych Pearsoina (1906)¹ o rozwoju umysłowym i fizycznym uczniów.

Rozwój fizyczny	Rozwój umysłowy	
	dobry	zły
dobry	581	561
zły	209	351

- a) zilustrować na podstawie tych danych nierówności, opisane w zadaniu 7, zastępując odpowiednie prawdopodobieństwa przez ich częstości. Co te nierówności oznaczają?
- b) Przedstawić tę tablicę w postaci standaryzowanej i narysować odpowiadający jej wykres kołowy.
- c) Zastosować test χ^2 i test oparty na ilorazie krzyżowym θ dla testowania hipotezy niezależności dla tych danych.
- d) Znaleźć 95% przedział ufności dla θ .
10. Zaprogramować formatki w Excelu do wyznaczania:
- a) przedziałów ufności dla względnego ryzyka (dane wejściowe: wartości a, b, c, d z tabeli dwudzielczej, poziom istotności w %),
- b) przedziałów ufności dla ilorazu szans (dane wejściowe: wartości a, b, c, d z tabeli dwudzielczej, poziom istotności w %).
11. Stowarzyszenie konsumentów na wniosek użytkowników pewnego modelu smartfona zamówiło badania, których celem było porównanie ryzyka popsucia się tego telefonu z gwarancją fabryczną (3 lata), z ryzykiem popsucia się takiego telefonu z wykupioną dodatkową gwarancją (5 lat). Klienci podejrzewali, że ryzyko popsucia się telefonu z gwarancją fabryczną jest dwa razy większe, niż telefonów z wykupioną dodatkową gwarancją. Agencja badawcza uzyskała następujące wyniki:
 telefony z 3-letnią gwarancją: popsute - 30, bezawaryjne - 120;
 telefony z 5-letnią gwarancją: popsute - 4, bezawaryjne - 112.
 Wykonać odpowiedni test, aby odpowiedzieć na pytanie, czy podejrzenia klientów są słuszne. Przyjąć poziom ufności 95%.
12. W celu zbadania skuteczności pewnego leku na gripę podano ten lek grupie 260 pacjentów. Po tygodniu obliczono szansę na wyzdrowienie w tej grupie i uzyskano wynik - 12. W grupie kontrolnej liczącej 180 pacjentów szansa wyzdrowienia po tygodniu wyniosła 8. Wykorzystując formatkę z Zadania 10 wyznaczyć przedział ufności dla ilorazu szans powrotu do zdrowia pacjentów, którym podano lek i pacjentów, którzy leku nie otrzymali. Przyjąć poziom ufności 99%.
13. W celu zbadania nowego leku na ból kręgosłupa wylosowano dwie niezależne grupy pacjentów, którzy skarżyli się na tą dolegliwość. W grupie A podano lek a w grupie B placebo. Uzyskano następujące wyniki:
 Grupa A - ból minął u 25 pacjentów, a u 35 nie było żadnych zmian.
 Grupa B - ból minął u 45 pacjentów, a u 120 nie było żadnych zmian.
 Wykorzystując cztery różne metody zbadać skuteczność leku na poziomie istotności 5%.
14. Wyniki tabeli przedstawiają jedno z pierwszych badań zależności raka płuc i paleniem tytoniu (patrz Agresti (2002), str. 41-43). Skonstruować (i zinterpretować) 95% przedziały ufności a) ilorazu szans, b) różnicy proporcji, c) względnego ryzyka między paleniem a zachorowaniem na raka płuc.

pali	rak płuc	
	TAK	NIE
TAK	688	650
NIE	21	59
razem	709	709

15. Zmienna losowa Y ma rozkład $\mathcal{H}(N, m, n)$. Sprawdzić, że $\sum_y P(Y = y) = 1$.
16. Dla danych dotyczących „zdolności przewidywań pewnej ekspertki co wiano jako pierwsze” (tabela 3.8, Agresti (2002)) obliczyć P -wartość oraz mid- P -wartość dokładnego testu Fishera w przypadku hipotezy jednostronnej $H_a : \theta > 1$.
17. Sporządzić formatkę w Excelu do wykonywania dokładnych testów Fishera.

¹ „On the relationship of intelligence to size and shape of head, and to other physical and mental characters”, *Biometrika* **5**, 105-146

18. Przeprowadzono na pacjentach testy trzech leków na katar oznaczonych L_1 , L_2 , L_3 . Uzyskano następujące wyniki:
- a) L_1 : $a = 3$ (lek był skuteczny), $b = 17$ (lek nie był skuteczny), $c = 5$ (placebo było skuteczne), $d = 5$ (placebo nie było skuteczne);
- b) L_2 : $a = 8$, $b = 2$, $c = 2$, $d = 6$;
- c) L_3 : $a = 9$, $b = 11$, $c = 1$, $d = 19$.
- Wykorzystując formatkę na poziomie istotności 5% zbadać skuteczność wszystkich leków. Dla danych z punktu c) przeprowadzić test dwustronny.
19. Pewne biuro turystyczne zleciło badania zależności pomiędzy wykształceniem turystów, a ich preferencjami dotyczącymi spędzania czasu podczas urlopu. Wyniki badań respondentów zostały przedstawione w poniższej tabeli. Na poziomie istotności 1% zbadać czy sposób wypoczynku jest zależny od poziomu wykształcenia turystów. Test przeprowadzić przy wykorzystaniu pakietu R .

miejsce i sposób wypoczynku	wykształcenie			
	podstawowe	średnie	wyższe	
morze	32	21	14	67
góry	15	27	28	70
rower	11	17	25	53
zwiedzanie	8	17	31	56
inne	20	15	19	54
	86	97	117	300