

1. Z rejestrów Narodowego Centrum Zdrowia (USA) wylosowano dane o dniu tygodnia, w którym rodziły się dzieci:

Dzień	nie	pon	wto	śro	czw	pią	sob
Liczba urodzin	77	110	124	122	120	123	97

źródło: Triola [595.12]

Zbadaj, czy dzień urodzin jest losowy. Skomentuj otrzymany wynik.

Przeprowadź analizę za pomocą statystyki χ^2 , diagnostykę według reszt Pearsona oraz za pomocą statystyki V Cramera.

2. Obliczono, jaką cyfrą zaczynają się losowo wybrane wpłaty na kampanię wyborczą, ujawnione przez pewną partię polityczną (USA).

Pierwsza cyfra	1	2	3	4	5	6	7	8	9
Liczba wpłat	52	40	23	20	21	9	8	9	30

źródło: Triola [597.23]

Czy zapisy wpłat na fundusz wyborczy były zgodne z rzeczywistością?

Przeprowadź analizę za pomocą statystyki χ^2 , diagnostykę według reszt Pearsona oraz za pomocą statystyki V Cramera.

Wskazówka. Zbiór rzeczywistych danych podlega prawu Benforda. Podaje ono rozkład pierwszych cyfr znaczących zbioru danych. Odpowiedzi na pytanie dlaczego prawdziwe jest to zdumiewające prawo, poszukaj np w internecie.

3. Przeprowadzono eksperyment o skuteczności wyciągu z *Echinacea angustifolia* w leczeniu skutków przeziębienia. Pacjenci zażywali przez pewien okres albo placebo, albo ekstrakt 20 % echinacea, albo ekstrakt 60 % echinacea. Następnie wszystkich zakażono rinowirusem, powodującym przeziębienie. Oto wyniki eksperymentu:

	Grupa terapeutyczna		
	Placebo	Ekstrakt 20 %	Ekstrakt 60 %
Chorzy	88	48	42
Zdrowi	15	4	10

źródło: Triola [598]

Czy leczenie echinacea jest skuteczne?

Przeprowadź analizę za pomocą statystyki χ^2 , diagnostykę według reszt Pearsona oraz za pomocą statystyki V Cramera.

4. Badano system oddechowy rodziców i jednego z ich dzieci. Sprawdzano kto kaszlał po przebudzeniu. Wyniki przedstawia tabela

Rodzic	Dziecko	
	Kaszele	Nie kaszele
Kaszele	29	104
Nie kaszele	172	5097

źródło: Triola [620.5]

Czy prawdopodobieństwo, że dziecko kaszele a rodzic nie kaszele i prawdopodobieństwo, że dziecko nie kaszele a rodzic kaszele są sobie równe? Jaki wniosek wynika z tych badań?

Wskazówka: Użyj testu McNemary

Vademecum

Statystyka χ^2 i V Cramera

Dana jest tablica zawierająca liczbę n_i pojawień się zdarzeń A_i :

Zdarzenia	A_1	A_2	$\dots$	A_k
Liczba pojawień	n_1	n_2	$\dots$	n_k

oraz tablica zawierająca oczekiwaną liczbę e_i pojawień się zdarzeń A_i przy założeniu pewnej hipotezy:

Zdarzenia	A_1	A_2	$\dots$	A_k
Oczekiwana liczba pojawień	e_1	e_2	$\dots$	e_k

Uwaga! Statystyk χ^2 i V Cramera można używać gdy wszystkie $e_i > 5$

Błąd względny obserwacji względem hipotezy wyraża się wzorem:

$$b_i = \frac{n_i - e_i}{e_i}$$

Oczekiwany kwadratowy błąd względny ma postać:

$$B^2 = \sum_{i=1}^k p_i b_i^2,$$

gdzie

$$p_i = \frac{e_i}{N},$$

$$N = \sum_{i=1}^k n_i$$

jest oszacowaniem hipotetycznego prawdopodobieństwa pojawienia się zdarzenia A_i .

Statystyka χ^2 wyraża się wzorem

$$\chi^2 = NB^2 = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}$$

Dla tablic o w wierszach i k kolumnach ($w \leq k$) statystyka V Cramera wyraża się wzorem:

$$V = \begin{cases} B & \text{jeżeli } w = 1 \\ \frac{B}{\sqrt{w-1}} & \text{jeżeli } w > 1 \end{cases}$$

Wartości krytyczne na poziomie 0.05 dla testu χ^2

Przybliżony (błąd $\leq 5\%$) wzór na wartość krytyczną dla przypadku $w = 1$.

$$\chi_{0.05}^2 = \begin{cases} \frac{16k+4}{9} & \text{jeżeli } 2 \leq k \leq 6 \\ \frac{5k+18}{4} & \text{jeżeli } k \geq 7 \end{cases}$$

Dla tablic o $w > 1$ wierszach i $k > 1$ kolumnach przybliżona wartość krytyczna wynosi $\chi_{0.05}^2 = w * k$.

Diagnostyka - reszty Pearsona

Wartości

$$r_i = b_i \sqrt{\frac{e_i}{1 - p_i}},$$

czasami definiowane jako

$$r_i = b_i \sqrt{e_i},$$

nazywają się *resztami Pearsona*. Mają one, dla dużych wartości e_i i gdy prawdziwa jest testowana hipoteza, rozkład normalny standardowy.

Stąd kryterium:

obserwacja o numerze i jest istotnie zawyżona (zanizowana), gdy reszta Pearsona jest większa (mniejsza) od 2 (-2).

obserwacja o numerze i jest bardzo istotnie zawyżona (zanizowana), gdy reszta Pearsona jest większa (mniejsza) od 2.6 (-2.6).

Wartości te odpowiadają kwantylom 0.975 i 0.995 standardowego rozkładu normalnego.

Interpretacja wartości testu V Cramera

V	zależność
- 0.15	nieistotna
- 0.20	słaba
- 0.25	umiarkowana
- 0.30	umiarkowanie mocna
- 0.35	mocna
- 0.40	bardzo mocna
- 0.45	wyjątkowo mocna
- 1.00	idealna

Prawo Benforda

Dla danych rzeczywistych prawdopodobieństwo, że pierwsza cyfra znacząca jest równa j ($j = 1, 2, \dots, 9$) wynosi

$$\log_{10} \frac{j+1}{j}$$

Test McNemary

Testu McNemary używa się do testowania hipotezy o równości prawdopodobieństw zdarzeń $A_1 \cap A'_2$ i $A'_1 \cap A_2$.

A_1 jest pewnym zdarzeniem, realizującym się w warunkach 1, A_2 jest tym samym zdarzeniem, ale realizującym się w warunkach 2.

Liczba obserwacji w próbie jest zamieszczona w tabeli:

	Warunki 2	
	A_2	A'_2
Warunki 1		
A_1	a	b
A'_1	c	d

Warunkiem stosowania testu McNemary jest $b + c \geq 10$.

Statystyka McNemary ma postać:

$$m = \frac{(|b - c| + 1)^2}{b + c}$$

Różnica prawdopodobieństw jest istotna, gdy $m > 3.841$ a bardzo istotna, gdy $m > 6.635$.

Wartości te odpowiadają kwantylom 0.95 i 0.99 rozkładu χ^2 z jednym stopniem swobody.