Zadania statystyczne z egzaminu: Wstęp do Statystycznej Analizy Danych, styczeń 2019

Odpowiedzi.

7a) 220
7b) 200
7c) 5550 (lub 5045.45)
7d) 74.50 (lub 71.03)

8a) 1.75
8b) 0.0401
8c) TAK
8d) 0.0802
8e) NIE
8f) $[49.79, 53.71]$

9a) $[0.16, 0.24]$
9b) 2.31
9c) 0.0208
9d) TAK

11a) 6.67
11b) 5.9915
11c) TAK
11d) $e^{-3.33}$ (przybliżona wartość to 0.036)

Rozwiązania.

7a) Średnia wartość ceny mieszkania: $\bar{X} = \frac{X_1+\cdots+X_{11}}{11} = \frac{2420}{11} = 220.$

7b) Porządkując ceny mieszkania w sposób niemalejący, otrzymujemy: $120; 150; 170; 190; 195; 200; 225; 235; 245; 300; 390$. Przy 11 obserwacjach najbardziej środkową obserwacją jest obserwacja szósta licząc od najmniejszej - jest to mediana, czyli $Me = 200$.

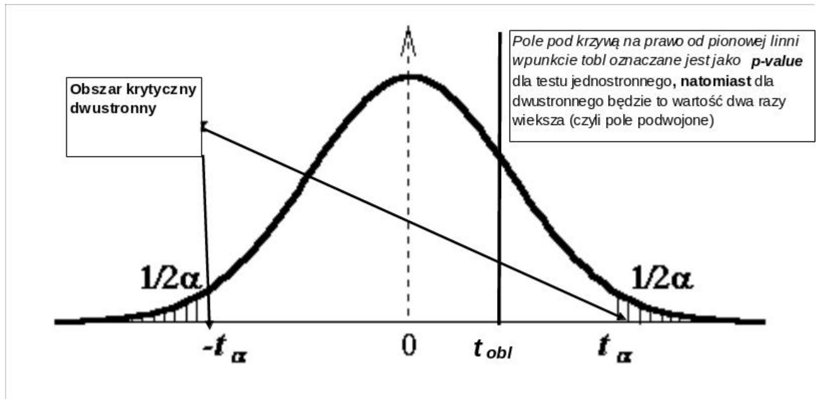
7c) Wariancja (wersja nieobciążona) ceny mieszkania: $S^2 = \frac{1}{10}\sum_{i=1}^{11}(X_i - \bar{X})^2 = \frac{1}{10}(100^2 + 70^2 + 50^2 + 30^2 + 25^2 + 20^2 + 5^2 + 15^2 + 25^2 + 80^2 + 170^2) = \frac{55500}{10} = 5550.$

Uwaga. Skoro w zadaniu nie jest powiedziane, o którą wersję wariancji chodzi, dopuszczalne jest podanie wariancji zgodnie z wersją obciążoną, czyli
$\widehat{S}^2 = \frac{1}{11}\sum_{i=1}^{11}(X_i - \bar{X})^2 = \frac{55500}{11} \approx 5045.45.$

7d) Odchylenie standardowe: $S = \sqrt{S^2} = \sqrt{\frac{55500}{10}} \approx 74.50$ lub $\widehat{S} = \sqrt{\widehat{S}^2} = \sqrt{\frac{55500}{11}} \approx 71.03.$

8a) Przy testowaniu średniej wartości μ dla próbki z rozkładu normalnego o nieznanej wariancji stosujemy test t-Studenta ze statystyką testową $T(X) = \sqrt{n}\frac{\bar{X}-\mu_0}{S}$, gdzie $\mu_0 = 50$ jest testowaną wartością średniej występującej w hipotezie H_0, $\bar{X}$ jest średnią z próby, $S = \sqrt{S^2}$, a S^2 jest nieobciążoną wersją wariancji z próby. Zatem $T = \sqrt{400} \cdot \frac{51.75-50}{20} = 1.75.$

8b) Liczenie p-wartości zależy od tego, jaka jest postać hipotezy alternatywnej H_1. Skoro hipoteza alternatywna jest hipotezą jednostronną (przy czym prawostronną), to p-wartość jest równa prawdopodobieństwu tego, że przy założeniu prawdziwości hipotezy H_0 statystyka testowa, jako zmienna losowa, jest większa od zaobserwowanej wartości tej statystyki, czyli 1.75 (jest to t_{obl} na wykresie).



Jeśli hipoteza H_0 jest prawdziwa, to statystyka testowa ma rozkład t-Studenta o $400-1=399$ stopniach swobody. Jak jest napisane w Uwadze do tego zadania, rozkład ten jest bardzo zbliżony do standardowego rozkładu normalnego, dlatego możemy korzystać z tablicy *Wartości dystrybuanty rozkładu normalnego standardowego.* Zatem p-wartość wynosi $\mathbb{P}=\mathbb{P}(T(X)>t_{obl})=1-\mathbb{P}(T(X)\leqslant t_{obl})\approx 1-\Phi(t_{obl})$, gdzie $\Phi(\cdot)$ jest dystrybuantą rozkładu $N(0,1)$. Więc, $\mathbb{P}\approx 1-\Phi(1.75)=1-0.9599=0.0401$.

8c) Skoro zaobserwowana wartość statystyki testowej $t_{obl}>t_{0.95}(399)\approx z_{0.95}=1.65$ (równoważnie p-wartość $\mathbb{P}<\alpha=0.05$), odrzucamy H_0.

8d) Teraz hipoteza alternatywna jest hipotezą dwustronną, więc p-wartość jest równa podwojonemu prawdopodobieństwu tego, że przy założeniu prawdziwości hipotezy H_0 statystyka testowa, jako zmienna losowa, jest większa od zaobserwowanej wartości tej statystyki 1.75. Zatem p-wartość wynosi $\mathbb{P}=2\mathbb{P}(T(X)>t_{obl})\approx 2(1-\Phi(1.75))=2(1-0.9599)=0.0802$.

8e) Skoro zaobserwowana wartość statystyki testowej $t_{obl}<t_{0.975}(399)\approx z_{0.975}=1.96$ (równoważnie p-wartość $\mathbb{P}>\alpha=0.05$), nie mamy podstaw do odrzucenia H_0.

8f) Przedział ufności dla μ na poziomie ufności $1-\alpha$ ma postać $[\bar{X}-t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}},\bar{X}+t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}]$, gdzie $t_{1-\alpha/2}(n-1)$ jest kwantylem rzędu $1-\alpha/2$ rozkładu t-Studenta o $n-1$ stopniu swobody. Mamy $n-1=399,\ 1-\alpha=0.95$, więc $1-\alpha/2=0.975$ oraz $t_{0.975}(399)\approx 1.96$ (patrz Uwagę do tego zadania). Zatem szukany przedział ufności wynosi

$[51.75-1.96\frac{20}{\sqrt{400}},51.75+1.96\frac{20}{\sqrt{400}}]=[49.79,53.71]$.

9a) Estymator punktowy nieznanej frakcji p palących wynosi $\widehat{p}=\frac{80}{400}=0.2$. Przedział ufności (przybliżony) dla p na poziomie ufności $1-\alpha$ ma postać $[\widehat{p}-z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p}}{n}},\widehat{p}+z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p}}{n}}]$, gdzie $z_{1-\alpha/2}$ jest kwantylem rzędu $1-\alpha/2$ rozkładu normalnego standardowego $N(0,1)$. Dla $1-\alpha=0.95$ mamy $1-\alpha/2=0.975$ oraz $z_{0.975}=1.96$. Zatem przedział ufności dla p wynosi

$[0.2-1.96\sqrt{\frac{0.2(1-0.2)}{400}},0.2+1.96\sqrt{\frac{0.2(1-0.2)}{400}}]\approx[0.16,0.24]$.

9b) Przy testowaniu nieznanej frakcji statystyka testowa wynosi $Z(X)=\sqrt{n}\frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)}}$, gdzie $p_0=0.25$ to testowana wartość frakcji występująca w hipotezie H_0. Zatem $Z=\sqrt{400}\cdot\frac{0.2-0.25}{\sqrt{0.25(1-0.25)}}=\frac{4}{\sqrt{3}}\approx 2.31$.

9c) Hipoteza alternatywna jest hipotezą dwustronną, więc p-wartość jest równa podwojonemu prawdopodobieństwu tego, że przy założeniu prawdziwości hipotezy H_0 statystyka testowa, jako zmienna losowa, jest większa od zaobserwowanej wartości tej statystyki, czyli 2.31. Zatem p-wartość wynosi $\mathbb{P}=2\mathbb{P}(Z(X)>z_{obl})\approx 2(1-\Phi(2.31))=2(1-0.9896)=0.0208$.

9d) Skoro zaobserwowana wartość statystyki testowej $z_{obl}>z_{0.975}=1.96$ (równoważnie p-wartość $\mathbb{P}<\alpha=0.05$), odrzucamy H_0.

11a) Zakładamy, że mamy do czynienia ze zmienną losową (wybór marki), która przyjmuje wartości A, B, C. Statystyka testowa testu zgodności χ^2 ma postać $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$, gdzie k to liczba wartości, które przyjmuje rozważana zmienna losowa, $\{n_i\}$ to obserwowane liczby występowania poszczególnych wartości zmiennej losowej w próbie, n to rozmiar próby, $\{p_i\}$ to testowane prawdopodobieństwa występujące w hipotezie H_0.

W naszym przypadku $k = 3, n_1 = 20, n_2 = 30, n_3 = 40, n = 20 + 30 + 40 = 90, p_1 = p_2 = p_3 = \frac{1}{3}$. Zatem $\chi^2 = \frac{(20-30)^2}{30} + \frac{(30-30)^2}{30} + \frac{(40-30)^2}{30} = \frac{20}{3} \approx 6.67$.

11b) W przybliżeniu statystyka χ^2 ma rozkład $\chi^2(k-1)$ (chi-kwadrat o $k-1$ stopniu swobody); u nas $k - 1 = 2$. Kwantyl $\chi^2_{0.95}(2)$ znajdziemy w *Tablice kwantyli rozkładu chi-kwadrat*: $\chi^2_{0.95}(2) = 5.9915$.

11c) Skoro zaobserwowana wartość statystyki testowej $\chi^2_{obl} \approx 6.67$ jest większa od wartości kwantyla $\chi^2_{0.95}(2) = 5.9915$, odrzucamy H_0.

11d) W teście zgodności χ^2 hipoteza alternatywna jest zawsze postaci: *testowane prawdopodobieństwa są inne niż w hipotezie* H_0, więc p-wartość zawsze liczy się tak, jak w teście t-Studenta dla jednostronnej (prawostronnej) hipotezy alternatywnej. Czyli, p-wartość jest równa prawdopodobieństwu tego, że przy założeniu prawdziwości hipotezy H_0 statystyka testowa, jako zmienna losowa, jest większa od zaobserwowanej wartości tej statystyki, czyli 6.67.

Jeśli hipoteza H_0 jest prawdziwa, to statystyka testowa w przybliżeniu ma rozkład $\chi^2(2)$. Zatem p-wartość wynosi $\mathbb{P} = \mathbb{P}(\chi^2 > \chi^2_{obl}) = 1 - \mathbb{P}(\chi^2 \leqslant \chi^2_{obl}) \approx 1 - F(\chi^2_{obl})$, gdzie $F(\cdot)$ to dystrybuanta rozkładu $\chi^2(2)$. Zgodnie ze Wskazówką do tego zadania, $\mathbb{P} \approx 1 - (1 - e^{-6.67/2}) = e^{-6.67/2} = e^{-3.33}$. W przybliżeniu, $e^{-3.33} \approx 0.036$.