Zestaw 3

Statystyczna Analiza Danych

Emilia Wiśnios

25 marca 2021

1. **Rozmiar testu, moc testu a liczebność**

   Niech $\{X_1, ..., X_n\}$ będzie próbą losową z rozkładu normalnego $N(m, 2^2)$. Hipotezę $H_0 : m = 1$ przy alternatywie $H_1 : m = 3$ będziemy weryfikować, wykorzystując test o zbiorze krytycznym postaci $\{(x_1, x_2, ..., x_n) : \sum_{i=1}^{n} x_i > k_\alpha\}$.

   - Wyznacz k_α, aby otrzymać test o rozmiarze 0.05.
   - Jak dużą próbę losową należy pobrać, aby uzyskać test o mocy nie mniejszej niż 0.95?

   Rozwiązanie:

   $X_1, ..., X_n \sim N(\mu, 2^2), \quad H_0 : \mu = 1, \quad H_1 : \mu = 3$

   Zbiór krytyczny: $K = \{x : \sum x_i \geq k_\alpha\}$

   $$\sum x_i \sim (n\mu, n\sigma^2)$$

   $$P_{\mu=1}(x \in K) = P_{\mu=1}(\sum x_i \geq K_\alpha) = P_{\mu=1}\left(\frac{\sum x_i - n}{2\sqrt{n}} \geq \frac{k_\alpha - n}{2\sqrt{n}}\right) = 1 - \psi\left(\frac{k_\alpha - n}{2\sqrt{n}}\right) = 0.05$$

   $$\psi\left(\frac{k_\alpha - n}{2\sqrt{n}}\right) = 0.95$$

   $$\frac{k_\alpha - n}{2\sqrt{n}} = z_{0.95} = 1.65, \quad \Rightarrow \quad k_\alpha = 3.3\sqrt{n} + n$$

   Moc testu:

   $$P_{\mu=3}(\sum x_i \geq k_\alpha) = P_{\mu=3}(\sum x_i \geq 3.3\sqrt{n} + n) = P_{\mu=3}\left(\frac{\sum x_i - 3n}{2\sqrt{n}} \geq \frac{3.3\sqrt{n} + n - 3n}{2\sqrt{n}}\right) =$$

   $$1 - \psi(1.65 - \sqrt{n}) \geq 0.95$$

   $$\psi(1.65 - \sqrt{n}) \leq 0.05$$

   $$1.65 - \sqrt{n} \leq -20.95$$

   $$\sqrt{n} > 3.3$$

   $$n > (3.3)^2$$

2. **Rozmiar testu**

Wyhodowano nową odmianę pewnej rośliny. Hipotezę, że kiełkuje 70% sadzonek, wobec hipotezy alternatywnej, że kiełkuje więcej niż 70%, testowano na podstawie próbki 10 sadzonek. Hipotezę zerową odrzucamy, gdy wykiełkuje 8 lub więcej sadzonek.

- Czy rozmiar tego testu jest mniejszy niż 0.05?
- Jaki jest rozmiar innego testu, który odrzuca hipotezę zerową, jeśli wszystkie sadzonki wykiełkują?

Rozwiązanie: $H_0 : p = 0.7,\ H_1 : p > 0.7$

Zbiór krytyczny: $K = \{S_{10} \geq 8\}$

Rozmiar testu:

$$P_{H_0}(S_{10} \geq 8) = \binom{10}{8}(0.7)^8(0.3)^2 + \binom{10}{9}(0.7)^9(0.3) + \binom{10}{10}(0.7)^{10} = 0.38$$

Druga kropka:

$$K = \{S_{10} = 10\}$$

$$P_{H_0}(S_{10} = 10) = \binom{10}{10}(0.7)^{10} = 0.02..$$

3. **Test najmniejszy**

Populacja ma rozkład opisany funkcją gęstości postaci $f(x) = \beta e^{-\beta x}$ dla $x > 0$. Z tej populacji wylosowano dziesięcioelementową próbę:

$$1 \quad 0.8 \quad 1.7 \quad 5.5 \quad 1.9 \quad 8.1 \quad 2.6 \quad 2.5 \quad 1.4 \quad 2.4$$

Przestestuj wykorzystując test najmocniejszy przy poziomie istotności $\alpha = 0.05$ hipotezę zerową, że $\beta = \frac{1}{2}$, przeciw hipotezie alternatywnej, że $\beta = \frac{1}{3}$.

Rozwiązanie:

$H_0 : \beta = \frac{1}{2},\ H_1 : \beta = \frac{1}{3}$ Korzystając z lematu Neymana-Pearsona mamy:

$$\frac{f_{1/3}(x_1, ..., x_n)}{f_{1/2}(x_1, ..., x_n)} > c$$

$$\frac{\left(\frac{1}{3}\right)^{10} e^{\frac{1}{3}\sum x_i}}{\left(\frac{1}{2}\right)^{10} e^{\frac{1}{2}\sum x_i}} > c$$

Wyrzućmy stałe

$$\exp\left(\frac{1}{2}\sum x_i - \frac{1}{3}\sum x_i\right) > c$$

Z tego wyznaczamy nasz zbiór krytyczny:

$$K = (\sum x_i > c)$$

$$P_{1/2}(\sum x_i > c) = \alpha \quad \Rightarrow \quad c = \chi^2(2n, 1 - \alpha)$$

$$x_i \sim \exp(1/2) = T(1, 1/2)$$

$$\sum x_i = T(n, 1/2) \sim \chi^2(2n)$$

$$\chi^2(20, 0.95) = 31.41$$

$$\sum x_i = 27.9$$

Nie możemy odrzucić hipotezy zerowej.

4. **Test ilorazu wiarogodności dla rozkładu wielomianowego**

   Skonstruuj test ilorazu wiarogodności dla rozkładu wielomianowego. Przyjmij hipotezy

   - Zerową $H_0 : p = [p_1(\theta), ..., p_k(\theta)],\ \theta \in w_0$
   - Alternatywną $H_1 : p \neq [p_1(\theta), ..., p_k(\theta)],\ \theta \in w_0$, nie czyni żadnych założeń o prawdopodobieństwach p, poza $\sum_i p_i = 1$, $\Omega = \{[p_1, ..., p_k] \mid \sum_i p_i = 1\}$

   Rozwiązanie:

5. Pokazać, że przy hipotezie zerowej spełnionej, iloraz wiarogodności w teście dla rozkładu wielomianowego i statystyka w teście zgodności Pearsona są asymptotycznie równoważne.

   Rozwiązanie:

6. Zmienna losowa X ma gęstość

$$f_\theta(x) = \frac{1}{\theta}\delta_{(0,\theta)}(x),$$

   gdzie $\delta_{(0,\theta)}(x)$ to funkcja przyjmująca wartość 1 dla $x \in (0, \theta)$ i 0 dla x spoza tego przedziału, a θ jest nieznanym parametrem. Niech c będzie ustaloną dodatnią stałą. Test polega na tym że jeśli $X \geq c$, to należy przyjąć hipotezę alternatywną $H_1 : \theta = 4$, a gdy $X < c$, należy przyjąć hipotezę zerową, $H_0 : \theta = 2$. Obliczyć:

   (a) prawdopodobieństwo błędów pierwszego i drugiego rodzaju (α i β),

   (b) moc testu,

   (c) β, gdy $\alpha = 0.05$.

   Rozwiązanie:

   $X \sim U(0, \theta)$

   $H_0 :\ \theta = 2, \quad H_1 :\ \theta = 4$

   Zbiór krytyczny: $K = \{x \geq c\}$

   (a) Błąd pierwszego rodzaju:

$$P_{\theta=2}(X \geq c) = \begin{cases} 0 & c \geq 2 \\ \frac{2-c}{2} = 1 - \frac{c}{2} & c \in [0, 2] \end{cases}$$

   Błąd drugiego rodzaju:

$$P_{\theta=4}(X < c) = \frac{c}{4}$$

   (b) Moc testu - $1 - \beta$

   (c) $\alpha = 0.05$, to

$$1 - \frac{c}{2} = 0.05 \quad \Rightarrow \quad c = 1.9$$

   Zatem

$$\beta = \frac{1.9}{4}$$

7. Cena pewnego wyrobu jest różna w zależności od punktu sprzedaży. Przyjęto, że cena w wylosowanym punkcie sprzedaży jest zmienną losową o rozkładzie normalnym. Po obliczeniu średniej i wariancji z 15 wylosowanych punktów sprzedaży otrzymano $\overline{X} = 2.5, \hat{S}^2 = 1.8$. Na poziomie istotności $\alpha = 0.05$ zweryfikuj hipotezę, że wahania cen (mierzone wariancją) są równe 1, przeciwko hipotezie, że są większe od 1.

   Wskazówka: $\chi^2(0.95, 14) = 23.685$.

   Rozwiązanie:

8. Według teorii Profesora Genka, komórki macierzyste pewnego organizmu różnicują się na 5 typów dojrzałych komórek, z prawdopodobieństwami: $p_1 = 7/16, p_2 = 1/4, p_3 = p_4 = 1/8, p_5 = 1/16$. Przeprowadzono 496 niezależnych powtórzeń eksperymentu różnicowania i w 212 powtórzeniach powstała komórka typu 1, w 123 powstała komórka typu 2, w 62 typu 3, w 45 typu 4, oraz w 54 powtórzeniach powstały komórki typu 5. Testem χ^2 na poziomie istotności $\alpha = 0.01$ zweryfikować hipotezę H_0, że teoria Genka dobrze opisuje zjawisko zderzeń.

Wskazówka: $\chi^2(0.99, 4) = 13.277$.

Rozwiązanie:

	I	II	III	IV	V
Observed:	212	123	62	45	54
	p_1	p_2	p_3	p_4	p_5
Expected:	217	124	62	62	31

Hipoteza zerowa: zachodzą prawdopodobieństwa z treści, hipoteza alternatywna: któreś z prawdopodobieństw jest różne

$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

gdzie O_i - observed, E_i - expected.

$$\frac{(217-212)^2}{212} + \frac{1}{124} + \frac{0}{62} + \frac{17^2}{62} + \frac{23^2}{31} = ... + 17.06 > \chi^2(0.99, 4) = 13.277$$

Wpadamy w obszar krytyczny, zatem odrzucamy teorię Profesora Genka.

9. **Test istotności dla dwóch średnich, przy próbach sparowanych**

10 robotnikom wprowadzono gimnastykę w trakcie pracy. Notowano wyniki pracy przed i w trakcie eksperymentu. Zarząd fabryki chciałby wiedzieć, czy wyniki pracy polepszyły się dzięki gimnastyce. Wyniki pomiarów wydajności pracy i-tego pracownika przed eksperymentem x_{i1} oraz w trakcie eksperymentu x_{i2} podane są w sztukach na godzinę w tabelce poniżej.[1] Przyjmij hipotezę zerową $H_0 : \mu_R = 0$ (wydajność pracy przed i po jednakowa) oraz hipotezę alternatywną $H_1 : \mu_R \neq 0$. Zweryfikuj hipotezę zerową na poziomie istotności $\alpha = 0.05$.

i	x_{i1}	x_{i2}	$r_1 = x_{i1} - x_{i2}$	r_i^2
1	28	32	-4	16
2	27	30	-3	9
3	24	24	0	0
4	27	28	-1	1
5	26	28	-2	4
6	22	24	-2	4
7	30	29	1	1
8	26	24	2	4
9	25	27	-2	4
10	26	29	-3	9
$\sum$			-14	52

Rozwiązanie:

$R_1, ..., R_{10} \sim N(\mu_R, \sigma_R^2)$

[1] dane wzięte z książki J. Józwiak, J. Podgórski, *Statystyka od podstaw*

$H_0 : \mu_R = 0, \quad H_1 : \mu_R \neq 0$ Skorzystamy z T-testu

$$T = \frac{\overline{X} - \mu_R}{S}\sqrt{n-1}, \quad \mu_R = 0$$

$$S^2 = \frac{1}{n}\sum(X_i - \overline{X})^2 = \overline{X^2} - (\overline{X})^2$$

$$\overline{X} = -1.4, \quad \overline{X^2} = 5.2$$

$$S^2 = 5.2 - 1.96 = 3.24$$

$$S = 1.8$$

$$T = \frac{-1.4}{1.8} \cdot 3 = -2\frac{1}{3}$$

$$|T| > t_{n-1,1-\frac{\alpha}{2}}$$

$$2.33 = |T| \geq t_{9,0.975} = 2,262$$

Wpada do obszaru krytycznego, zatem mam podstawy do odrzucenia hipotezy zerowej (czyli gimnastyka poprawia wydajność).

10. Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu normalnego $N(\mu, \sigma^2)$ ze znaną wariancją.

- Skonstruuj test najmocniejszy dla hipotezy $H_0 : \mu = \mu_0$ przeciw alternatywie $H_1 : \mu = \mu_1 > \mu_0$ na poziomie istotności 0.05.
- Oblicz moc uzyskanego testu.
- Jaka powinna być długość próby, aby moc testu była większa od 0.95?

Rozwiązanie:

$X_1, ..., X_n \sim N(\mu, \sigma^2), \quad \sigma^2$ - znane

$H_0 : \mu = \mu_0, \; H_1 : \mu = \mu_1, \quad \mu_1 > \mu_0$

$H_0 : f_{\mu_0}(x_1, ..., x_n), H_1 : f_{\mu_1}(x_1, ..., x_n)$

$$\frac{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum(x_i - \mu_1)^2\right)}{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum(x_i - \mu_0)^2\right)} > c$$

$$\exp\left(\frac{\sum(x_i - \mu_1)^2 - \sum(x_i - \mu_0)^2}{2\sigma^2}\right) > c$$

$$\sum(x_i - \mu_1)^2 - \sum(x_i - \mu_0)^2 > c$$

$$\sum x_i(\mu_1 - \mu_0) > c$$

Zgodnie z założeniem $\mu_1 > \mu_0$ więc możemy bezpiecznie podzielić stronami

$$\sum x_i > c$$

$$P_{\mu_0}(\sum x_i > c) = P\left(\frac{\sum x_i - n\mu_0}{\sqrt{n}\sigma} > \frac{c - n\mu_0}{\sqrt{n}\sigma}\right) = 1 - \psi\left(\frac{c - n\mu_0}{\sqrt{n}\sigma}\right) = 0.05$$

$$\psi\left(\frac{c - n\mu_0}{\sqrt{n}\sigma}\right) = 0.95$$

$$\frac{c - n\mu_0}{\sqrt{n}\sigma} = z_{0.95} = 1.65$$

Na mocy lematu Neymana-Pearsona

$$c = 1.65\sqrt{n}\sigma + n\mu_0$$

Moc

$$P_{\mu_1}(\sum x_1 > 1.65\sqrt{n}\sigma + n\mu_0) = P_{\mu_1}\left(\frac{\sum x_i - n\mu_1}{\sqrt{n}\sigma} > \frac{1.65\sqrt{n}\sigma + n(\mu_0 - \mu_1)}{\sqrt{n}\sigma}\right) =$$

$$= 1 - \psi\left(1.65 + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right)$$

Ostatnia kropka:

$$1 - \psi\left(1.65 + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right) > 0.95$$

$$\psi\left(1.65 + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right) < 0.05$$

$$1.65 + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma} < -z_{0.95}$$

$$\sqrt{n} > \frac{3.3 \cdot \sigma}{\mu_1 - \mu_0}$$

11. W 100 laboratoriach przeprowadzono niezależnie taki sam test na poziomie istotności 0.05. Zakładając, że hipoteza zerowa jest prawdziwa, oblicz prawdopodobieństwo, że w przynajmniej jednym z laboratotoriów została ona odrzucona.

    Rozwiązanie: P (odkrycie laboratorium) $= 0.05$

    Szukamy P(którekolwiek laboratorium odkryje)

    $1 - P$(nie ma odkryć) $= 1 - (0.95)^{100} = 1 - 0.006$

12. **Rozkład T-Studenta**

    Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu normalnego $N(\mu, \sigma^2)$ oraz niech $\overline{X}$ będzie średnią empiryczną oraz $S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$. Pokaż, że $\overline{X}$ i S_n^2 są niezależne oraz, że S_n^2 ma rozkład $\chi^2(n-1)$.

    Rozwiązanie:

$$Q = \begin{bmatrix} \frac{1}{\sqrt{n}} & \dots & \dots & \frac{1}{\sqrt{n}} \\ & & & \end{bmatrix}$$

$$Q^T Q = QQ^T = I$$

$$X = [X_1, ..., X_n], \quad X \sim N(\mu \cdot \mathbf{1}, \sigma^2 \cdot I)$$

$$Y = QX, \quad Y \sim N(Q \cdot \mu \cdot \mathbf{1}, Q \cdot \sigma^2 \cdot I \cdot Q^T)$$

$$Y_1, ..., Y_n - \text{ nzal}$$

$$Y_1 = \frac{\sum X_i}{\sqrt{n}} = \sqrt{n}\,\overline{X}$$

$$\sum Y_i^2 = Y^T Y = ||Y||^2 = ||QX||^2 = \sum X_i^2$$

$$\sum_{\mathbf{i=2}}^{n} Y_i^2 = \sum X_i^2 - n\overline{X}^2 = \sum (X_i - \overline{X})^2 = f(Y_2, ..., Y_n)$$

$$\overline{X} = g(Y_1)$$

$$\overline{X} \text{ nzal z } (X_i - \overline{X})^2$$

$$Y_i = Q_i^T X$$
$$Q_i^T Q_1 = 0$$
$$Q_i^T 1 = 0$$
$$EY_i = 0, \quad i = 2, ..., n$$
$$\chi^2(n+1) \sim \frac{\sum(X_i - \overline{X})^2}{\sigma^2} = \sum_{i=2}^{n} \frac{Y_i^2}{\gamma^2} = \sum Z_i^2, \quad Z_i \sim N(0,1)$$
$$\gamma = Q \cdot \sigma^2 \cdot Q^T$$

13. **Test t-studenta**

Wiadomo, że jeśli X ma rozkład $N(0,1)$ oraz Z jest niezależne od X o rozkładzie $\chi^2(k)$ to zmienna losowa $T = \frac{X}{\sqrt{Z/k}}$ ma rozkład t-studenta o k stopniach swobody. Korzystając z powyższego faktu pokaż, że dla próby prostej $X_1, ..., X_n$ z $N(\mu, \sigma^2)$

$$\sqrt{n=1}\frac{\overline{X} - \mu}{S_n}$$

ma rozkład t-studenta o $n-1$ stopniach swobody

Rozwiązanie:

$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1)$

$$\frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{\sum(x_i-\overline{x})^2}{\sigma^2(n-1)}}} \sim t(n-1)$$

14. Każda ze sprzedanych suszarek do włosów pewnego typu z nieznanym prawdopodobieństwem θ zostanie już pierwszego dnia po zakupie zareklamowana przez kupującego z powodu ukrytej wady. Przypuszcza się, że prawdopodobieństwo zgłoszenia reklamacji później, niż pierwszego dnia po zakupie jest takie samo. Zweryfikuj tą hipotezę za pomocą testu χ^2 na poziomie istotności $\alpha = 0.05$ na podstawie danych dotyczących 1000 suszarek, z których 40 reklamowano pierwszego dnia, 60 reklamowano później, a pozostałe nie były reklamowane.

Wskazówka: Kwantyl rozkładu $\chi^2(0.95, 1) = 3.84$.

Rozwiązanie:

	brak	po 1 dniu	później
Observed:	900	40	60
	p_1	p_2	p_3
Expected:	900	50	50

Hipoteza zerowa: $p_1 = 1 - 2\theta, p_2 = p_3 = \theta$, hipoteza alternatywna: $p_2 \neq p_3$

$$\theta \to \hat{\theta_{MLE}}$$
$$L(\theta) = (1-2\theta)^{900}\theta^{40+60}$$
$$l(\theta) = \log L(\theta) = 900\log(1-2\theta) + 100\log(\theta)$$
$$l'(\theta) = -\frac{2 \cdot 900}{1-2\theta} + \frac{100}{\theta} = 0 \quad \Rightarrow \quad 1800\theta = 100 - 200\theta \quad \Rightarrow \quad \theta = \frac{1}{20}$$
$$\chi^2 = \frac{0}{900} + \frac{10^2}{50} + \frac{10^2}{50} = 4 > \chi^2(0.95, 1) = 3.84$$

Wpadamy w obszar krytyczny, zatem odrzucamy hipotezę zerową.