

Statystyka w analizie i planowaniu eksperymentu

Wykład 8

Testów proporcji i testów średnich ciąg dalszy

Przemysław Biecek

Dla 1 roku studentów Biotechnologii

Przypomnienie kilku faktów

$$X_i \sim \mathcal{N}(0, 1)$$

Zmienne X_i mają rozkład normalny.

$$Y_i = \sum_{i=1}^k X_i \sim \mathcal{N}(0, k)$$

Suma zmiennych normalnych ma rozkład normalny o odpowiedniej średniej i wariancji.

$$X_i/a \sim \mathcal{N}(0, 1/a^2)$$

Iloczyn liczby o rozkładzie normalnym i stałej ma rozkład normalny.

$$X_i^2 \sim \chi_1^2$$

Kwadrat liczby o rozkładzie normalnym ma rozkład χ^2 z jednym stopniem swobody.

$$Z = \sum_{i=1}^k X_i^2 \sim \chi_k^2$$

Jeżeli sumowanych jest więcej kwadratów to otrzymujemy zmienną o rozkładzie χ^2 o k stopniach swobody.

$$\frac{X}{\sqrt{Z/k}} \sim t_k$$

Iloraz zmiennej o rozkładzie normalnym i o rozkładzie χ_k^2 ma rozkład t-Studenta o k stopniach swobody.

$$\frac{Z_1/n_1}{Z_2/n_2} \sim F_{n_1, n_2}$$

Iloraz dwóch zmiennych o rozkładzie χ^2 ma rozkład F .

Zadanie:

Czy częstość występowania genotypu bb o fenotypie niebieskich oczu występuje w populacji z częstością $\frac{1}{4}$?

Eksperyment:

Sprawdzono kolory oczu 200 studentów z biotechnologii, 70 z nich miało niebieskie oczy.

Pytanie:

- Czy próba jest prawidłowo zebrana?
- Jeżeli jest to jak odpowiedzieć na Zadanie?

Test dla proporcji - duże próby

W dużych próbach rozkład częstości przybliżyć można rozkładem normalnym. Do testowania hipotezy

$$H_0 : p = p_0$$

gdzie p_0 zadana wartość, wykorzystać można test oparty na statystyce testowej

$$T(X) = n \frac{p - p_0}{\sqrt{p_0(1 - p_0)n}}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład normalny $\mathcal{N}(0, 1)$. Obszary krytyczne wyznacza się ze wzorów

- dla dwustronnej hipotezy alternatywnej
 $W_\alpha = (-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty)$
- dla lewostronnej hipotezy alternatywnej
 $W_\alpha = (-\infty, q_\alpha]$
- dla prawostronnej hipotezy alternatywnej
 $W_\alpha = [q_{1-\alpha}, \infty).$

$$p = 70/200 = 0.35$$

$$T(X) = 200 \frac{0.35 - 0.25}{\sqrt{0.25 * 0.75 * 200}} = 3.27$$

Decyzja?

Zadanie:

Czy częstość występowania genotypu *bb* u kobiet i u mężczyzn jest taka sama?

Eksperyment:

Sprawdzono kolory oczu 200 studentów z biotechnologii (120 kobiet i 80 mężczyzn), 70 z nich miało niebieskie oczy (odpowiednio 40k i 30m).

Pytanie:

- Czy próba jest prawidłowo zebrana?
- Jeżeli jest to jak odpowiedzieć na Zadanie?

W dużych próbach rozkład częstości przybliżyć można rozkładem normalnym. Do testowania hipotezy

$$H_0 : p_1 = p_2,$$

wykorzystać można test oparty na statystyce testowej

$$T_1(X) = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}.$$

lub

$$T_2(X) = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład normalny $\mathcal{N}(0, 1)$. Obszary krytyczne wyznacza się jak dla testu dla jednej próby.

$$p = 70/200 = 0.35$$

$$p_1 = 40/120 = 0.333$$

$$p_2 = 30/80 = 0.375$$

$$T(X) = \frac{0.042}{0.35 * 0.65 * (0.0083 + 0.0125)} = 0.72$$

Decyzja?

Zadanie:

Czy zmienność ocen ze statystyki wśród kobiet jest taka sama jak u mężczyzn?

Eksperyment:

Sprawdzono wyniki pierwszego kolokwium, $S_K^2 = 0.7$ a $S_M^2 = 0.5$.
Wyniki dla 50 kobiet i 20 mężczyzn.

Do testowania hipotezy

$$H_0 : \sigma_1^2 = \sigma_2^2$$

gdzie σ_i^2 to wariancja w grupie i , wykorzystuje się test oparty o statystykę testową

$$T(X) = \frac{S_1^2}{S_2^2}$$

(większą wariancję zawsze wpisujemy do licznika).

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład normalny $\mathcal{F}(n_1 - 1, n_2 - 1)$. Obszary krytyczne wyznacza się ze wzorów

- dla dwustronnej hipotezy alternatywnej !!!

$$W_\alpha = [f_{1-\alpha/2}^{n_1-1, n_2-1}, \infty)$$

- dla jednostronnej hipotezy alternatywnej

$$W_\alpha = [f_{1-\alpha}^{n_1-1, n_2-1}, \infty).$$

Wyliczona wartość statystyki testowej wynosi

$$T(x) = 0.7/0.5 = 1.4$$

Wartość krytyczna odczytana z tablic

$$f_{0.95}^{(49,19)} \approx 2$$

Decyzja?

Zadanie:

Czy liczba punktów z pierwszego kolokwium była większa niż na drugim?

Eksperyment:

hmmmm....

Nieparametryczny odpowiednik testu t Studenta. W wersji sparowanej hipoteza zerowa ma postać

$$H_0 : \theta = 0$$

gdzie θ to mediana różnic $d_i = Y_i - X_i$. Do testowania wykorzystuje się statystykę testową

$$S = \min(W^+, W^-)$$

gdzie

$$W^+ = \sum_{d_i > 0} r(d_i), \quad W^- = \sum_{d_i < 0} r(d_i)$$

a $r(d_i)$ to ranga wartości d_i wyznaczona wektorze wartości bezwzględnych $|d_i|$. Dla dużych prób ($n > 20$) statystykę S można przybliżyć rozkładem normalnym o średniej $\frac{n(n+1)}{4}$ i wariancji $\frac{n(n+1)(2n+1)}{24}$. Dla małych prób wartości krytyczne powinny być odczytywane z tablic.

W wyniku eksperymentu zaobserwowano następujące d_i

$$d = c(-2, -1, 0.5, 2, -1, 1.5, 2.5, 2.5)$$

$$r(|d|) = c(3.5, 6.5, 8, 3.5, 6.5, 5, 1.5, 1.5)$$

$$W^+ = 7 + 3.5 + 5 + 1.5 + 1.5 = 18.5$$

$$W^- = 3.5 + 6.5 + 6.5 = 16.5$$

$$S = 16.5$$

Odczytujemy kwantyle (0.05 dla alternatywy jednostronnej i 0.025 dla alternatywy dwustronnej)

$$q_{0.05}^8 = 6, \quad q_{0.025}^8 = 4$$

W pakiecie R kwantyl można odczytać korzystając z funkcji `qsignrank(kwantyl, n)`.

Porównajmy dochody 10 wylosowanych z populacji pracujących kobiet i mężczyzn, czy są one równe?

zarobki M = 1500, 2000, 3500, 5500, 10000

zarobki K = 1600, 1900, 2400, 4000, 5000

To nieparametryczny odpowiednik testu t Studenta.
Hipoteza zerowa ma postać

$$H_0 : \theta_X = \theta_Y$$

gdzie θ_X to mediana dla populacji X a θ_Y dla Y .
Do testowania wykorzystuje się statystykę testową

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 1_{X_i < Y_j}$$

Dla dużych prób ($n > 20$) statystykę U można przybliżyć rozkładem normalnym o średniej $\frac{n_1 n_2}{2}$ i wariancji $\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$.
Dla małych prób wartości krytyczne odczytujemy z tablic.

Test U Wilcoxona-Manna-Whitneya

zarobki M = 1500, 2000, 3500, 5500, 10000

zarobki K = 1600, 1900, 2400, 4000, 5000

Wyznaczamy wartość statystyki U

$$U = 1 + 1 + 2 + 3 + 3 = 10.$$

Odczytujemy kwantyl dla rozkładu statystyki testowej

$$q_{0.025}^{(5,5)} = 3, \quad q_{0.975}^{(5,5)} = 22.$$

W pakiecie R kwantyl można odczytać korzystając z funkcji `qwilcox(kwantyl, n1, n2)`.

Teraz spróbujemy przybliżyć statystykę testową rozkładem normalnym. Normalizujemy wynik statystyki testowej

$$z = (10 - 12.5) / \sqrt{(25 * 11 / 12)} = -0.11$$

Czy cechy kolor oczu i płeć są ze sobą zależne?

	K	M
niebieskie	30	8
brązowe	60	12

Do testowania hipotezy

$$H_0 : X \text{ niezależne od } Y$$

wykorzystuje się test oparty o statystykę testową

$$T = \sum \frac{(O - E)^2}{E} = \sum_{i=1}^k \sum_{j=1}^p \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

gdzie

$$E_{ij} = \frac{\sum_{i=1}^k n_{ij} \sum_{j=1}^p n_{ij}}{\sum_{i=1}^k \sum_{j=1}^p n_{ij}}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład

$\chi^2_{(k-1)(p-1)}$ ze $(k-1)(p-1)$ stopniami swobody.

Obszary krytyczne wyznacza się ze wzoru

$$W_\alpha = [\chi_{1-\alpha}^{2, (k-1)(p-1)}, \infty)$$

Czy kolor oczu i płeć są ze sobą zależne?

Obserwowane

	K	M	
niebieskie	30	8	38
brązowe	60	12	72
	90	20	110

Oczekiwane

	K	M	
niebieskie	31.1	6.9	38
brązowe	58.9	13.1	72
	90	20	110

$$T = 1.1^2/31.1 + 1.1^2/6.9 + 1.1^2/58.9 + 1.1^2/13.1 = 0.33$$

$$\chi_{0.95}^{2,1} = 3.84$$

Czy dziewczynki są bardziej podatne na chorobę niż chłopcy?
Zbadano grupę 110 par bliźniąt dwujajowych w których jedna osoba jest chora a druga zdrowa.

zdrowy / chory	K	M
K	a=30	b=8
M	c=60	d=12

Do testowania hipotezy

$$H_0 : b \text{ występuje równie często jak } c$$

wykorzystuje się test oparty o statystykę testową

$$T = \frac{(b - c)^2}{b + c}.$$

Przy prawdziwej hipotezie zerowej statystyka ta ma rozkład χ_1^2 z 1 stopniem swobody.

Obszary krytyczne wyznacza się ze wzoru

$$W_\alpha = [\chi_{1-\alpha}^{2,1}, \infty)$$

Test Kołomogorova-Smirnova

Do testowania hipotezy

$$H_0 : X \sim F$$

wykorzystuje się test oparty o statystykę testową

$$D_n = \sup_x |F_n(x) - F(x)|$$

gdzie $F_n(x)$ to dystrybuanta empiryczna zadana wzorem

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}.$$

$$\sqrt{n}D_n \xrightarrow{n \rightarrow \infty} \sup_t |B(F(t))|$$

Kwantyli rozkładu tej statystyki testowej najlepiej szukać w tablicach.

Jak wykonać omawiane testy w R?

- Test dla proporcji zaimplementowany jest w funkcji `prop.test()`,
- Test dla wariancji zaimplementowany jest w funkcji `var.test()`,
- Test dla parametrów przesunięcia zaimplementowany jest w funkcji `wilcox.test()`,
- Test χ^2 zaimplementowany jest w funkcji `chisq.test()`,
- Test McNemara zaimplementowany jest w funkcji `mcnemar.test()`,
- Test Kołomogorova-Smirnova χ^2 zaimplementowany jest w funkcji `ks.test()`,
- Dobry test normalności zaimplementowany jest w funkcji `shapiro.test()`.

Bardzo Ważna Tabelka

Stan faktyczny	Decyzja	
	przyjąć H_0 $\psi(x) = 0$	odrzuć H_0 $\psi(x) = 1$
H_0 prawdziwa	decyzja poprawna	błąd I rodzaju
H_0 fałszywa	błąd II rodzaju	decyzja poprawna

Moc

Moc testu określamy jako prawdopodobieństwo odrzucenia hipotezy zerowej, w sytuacji gdy jest ona fałszywa.

Moc zależy od:

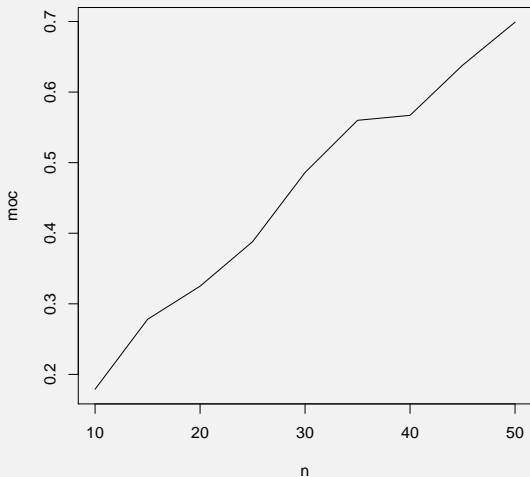
- przyjętego poziomu istotności,
- rozmiaru próby,
- różnicy pomiędzy alternatywą a hipotezą zerową.

Jak wyznaczyć moc?

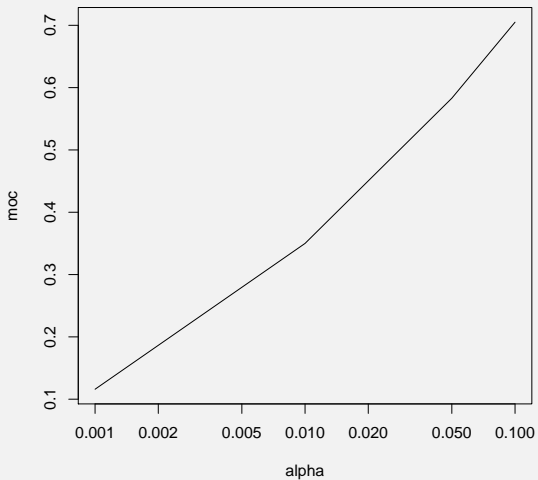
W R to jest proste!

```
> pwartosci = NULL
> for (i in 1:1000) {
>   x = rnorm(n)
>   y = rnorm(n)+0.5
>   pwartosci[i] = t.test(x,y)$p.value < 0.05
> }
> mean(pwartosci)
0.331
```

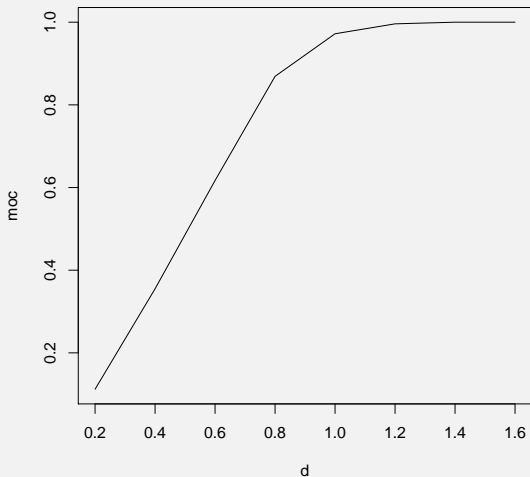
Moc w zależności od liczebności próby



Moc w zależności od poziomu istotności



Moc w zależności od różnic pomiędzy hipotezami



W rzeczywistych danych często zdarzają się brakujące obserwacje

- pomiar się nie powiódł a ze względów finansowych lub organizacyjnych nie jesteśmy w stanie go powtórzyć,
- jakiś pomiar przyjmuje ewidentnie błędną wartość, np. ciśnienie =350,
- operujemy na danych z innego źródła, które są niekompletne.

Co zrobić?

- Możemy usunąć cały przypadek w którym choć jeden pomiar jest brakujący, są plusy i minusy,
- Możemy wstawić za brakującą wartość wartość charakterystyczną dla zmiennej (średnią, medianą),
- Możemy przeprowadzić zbiór testów, wstawiając za brakującą wartość losową wartość, jedną z występujących w próbie.

Zobaczmy, jak wyglądają rzeczywiste analizy.

Co trzeba zapamiętać?

- Jak działa i po co jest test Wilcoxon?
- Jak działa i po co jest test U-Wilcoxon-Manna-Withneya?
- Jak działa i po co jest test χ^2 ?
- Jak działa i po co jest test proporcji?
- Jak działa i po co jest test F?
- Jak działa i po co jest test Kołomogorova Smirnova?
- Co to jest moc i po co nam to pojęcie?