Testowanie hipotez i estymacja w sytuacji populacji skończonego rozmiaru

Kinga Kurowska
promotor: dr inż. Andrzej Giniewicz

Wydział Matematyki
Politechnika Wrocławska

Wrocław, 23.01.2017r.

Spis treści

Schemat pobierania obserwacji - nieskończona populacja

Pobieranie obserwacji to losowanie ze zwracaniem. Próbka pochodzi z rozkładu Bernoulliego o funkcji prawdopodobieństwa danej wzorem

$$b(k; n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ 0 \leq k \leq n, \tag{1}$$

gdzie n to rozmiar próbki, a p prawdopodobieństwo sukcesu.

Schemat pobierania obserwacji - skończona populacja

Pobieranie obserwacji to losowanie bez zwracania. Próbka pochodzi z rozkładu hipergeometrycznego o funkcji prawdopodobieństwa danej wzorem.

$$h(k; n, M, N) = P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \ L \leq k \leq U, \tag{2}$$

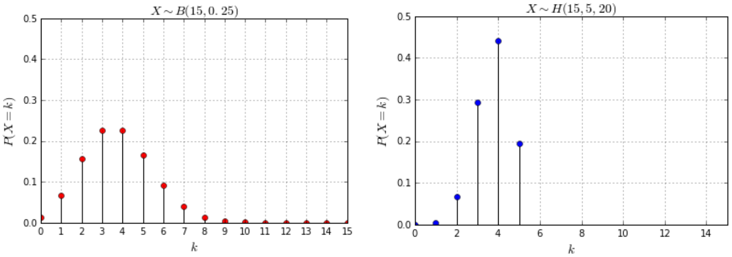
gdzie

$$L = \max\{0, M - N + n\}, \quad U = \min\{n, M\}. \tag{3}$$

Przy czym n to wielkość próbki, M liczba elementów z badaną cechą, a N to rozmiar populacji.

Porównanie rozkładów

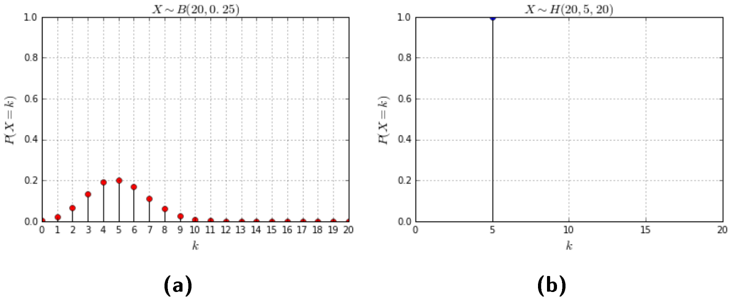
Załóżmy, że jest grupa 20 osób, które są chore na jakąś bardzo rzadką chorobę oraz że 25% z nich ma szanse na wyzdrowienie. Chcemy dowiedzieć się, ile osób spośród przebadanych może wyzdrowieć. Weźmy 2 różne próbki o wielkościach n równych odpowiednio 15 i 20.



(a) **(b)**

Rys. Funkcje prawdopodobieństwa $b(k; 15, 0.25)$ oraz $h(k; 15, 5, 20)$

Porównanie rozkładów



Rys. Funkcje prawdopodobieństwa $b(k; 20, 0.25)$ oraz $h(k; 20, 5, 20)$

Przedstawienie testów

Sformułowanie problemu

X_1 i X_2 - niezależne zmienne losowe. Wartości obserwacji oznaczmy k_1 i k_2 oraz proporcje w obserwacjach p_1 i p_2. Interesuje nas testowanie

$$H_0\colon p_1 = p_2 \quad \text{przeciwko} \quad H_1\colon p_1 \neq p_2. \tag{4}$$

Unormowana statystyka testowa to

$$Z_{X_1,X_2} = \frac{X_1/n_1 - X_2/n_2}{\sqrt{V_{X_1,X_2}}}, \tag{5}$$

gdzie V_{X_1,X_2} to estymator wariancji rozkładu zmiennej losowej $X_1/n_1 - X_2/n_2$, pod warunkiem prawdziwości H_0, w połączonej próbie.

Testy ze skończoną poprawką

Rozważmy zmienne losowe:

$$X_1 \sim \mathcal{H}(n_1, M_1, N_1),\ X_2 \sim \mathcal{H}(n_2, M_2, N_2). \tag{6}$$

Proporcje są równe $p_1 = M_1/N_1$, $p_2 = M_2/N_2$. Wariancja rozkładu $X_1/n_1 - X_2/n_2$ pod warunkiem $p_1 = p_2$ w połączonej próbie jest równa

$$V_{X_1,X_2} = \left(\frac{N_1 - n_1}{n_1(N_1 - 1)} + \frac{N_2 - n_2}{n_2(N_2 - 1)}\right)\left(\frac{X_1 + X_2}{n_1 + n_2}\right)\left(1 - \frac{X_1 + X_2}{n_1 + n_2}\right). \tag{7}$$

Test Z

Zakładamy, że rozważana statystyka $Z_{X_1,X_2} \sim \mathcal{N}(0,1)$, pod warunkiem prawdziwości H_0. Wtedy p-wartość wyraża się wzorem

$$P(|Z_{X_1,X_2}| \geq |Z_{k_1,k_2}|\ |H_0) \approx 2(1 - \Phi(|Z_{k_1,k_2}|)), \tag{8}$$

gdzie Φ to dystrybuanta rozkładu $N(0,1)$.

Test E

Test E opiera się o estymator p-wartości, którą w swoim artykule zaproponowali Krishnamoorthy i Thomson (2002) [1]

$$P(|Z_{X_1,X_2}| \geq |Z_{k_1,k_2}| \; |H_0) \approx$$
$$\approx \sum_{x_1=L_{x_1}}^{U_{x_1}} \sum_{x_2=L_{x_2}}^{U_{x_2}} h(x_1; n_1, \hat{M}_1, N_1) h(x_2; n_2, \hat{M}_2, N_2) \mathbb{1}\left(|Z_{x_1,x_2}| \geq |Z_{k_1,k_2}|\right), \tag{9}$$

przy czym $\hat{p} = (k_1 + k_2)/(n_1 + n_2)$, $\hat{M}_i = [N_i \hat{p}]$,
$L_{x_i} = \max\{0, \hat{M}_i - N_i + n_i\}$ oraz $U_{x_i} = \min\{n_i, \hat{M}_i\}$, $i = 1, 2$.

Test bez skończonej poprawki

Rozważamy zmienne losowe:

$$X_1 \sim \mathcal{B}(n_1, p_1),\ X_2 \sim \mathcal{B}(n_2, p_2). \tag{10}$$

Wariancja rozkładu $X_1/n_1 - X_2/n_2$ w łącznej próbie, pod warunkiem $p_1 = p_2$ jest równa

$$V_{X_1,X_2} = p(1-p)(1/n_1 + 1/n_2), \tag{11}$$

przy czym $p = (X_1 + X_2)/(n_1 + n_2)$.

Test Zb

Test Zb jest oparty o estymator p-wartości, który, zgodnie z artykułem Storer i Kim z 1990 roku, jest równy [4]

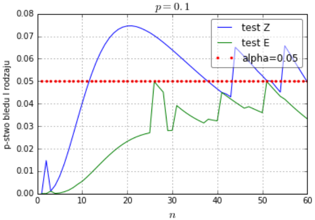
$$
\begin{aligned}
&P(|Z_{X_1,X_2}| \geq |Z_{k_1,k_2}| \; |H_0) \approx \\
&\approx \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} b(x_1; n_1, \hat{p}_1) b(x_2; n_2, \hat{p}_2) \mathbb{1} \left(|Z_{X_1,X_2}| \geq |Z_{k_1,k_2}| \right),
\end{aligned} \tag{12}
$$

gdzie $\hat{p} = (k_1 + k_2)/(n_1 + n_2)$.

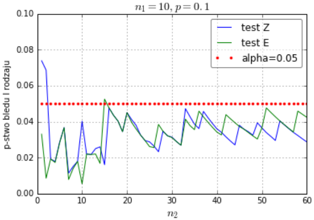
Analiza testów

Porównanie testów ze skończoną poprawką

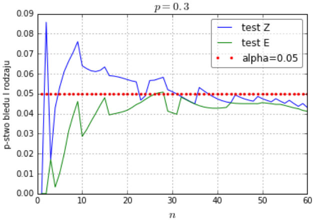
Rys. Prawdopodobieństwo błędu I rodzaju testów Z i E jako funkcja rozmiaru próbki n; $\alpha = 0.05$; $N_1 = N_2 = 100$



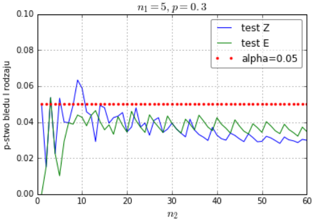
(a)



(b)

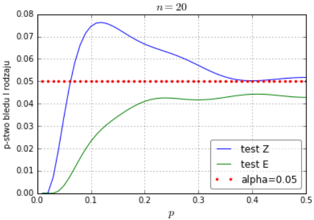


(c)

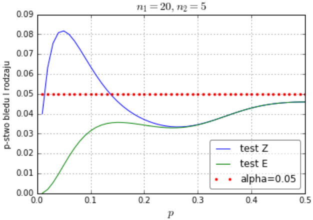


(d)

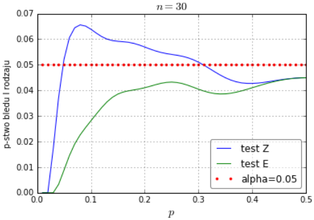
Rys. Prawdopodobieństwo błędu I rodzaju testów Z i E jako funkcja proporcji $p = M_1/N_1 = M_2/N_2$; $\alpha = 0.05$; $N_1 = N_2 = 100$



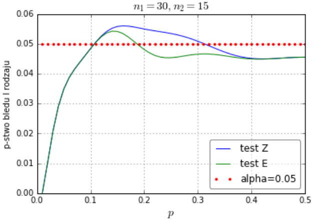
(a)



(b)

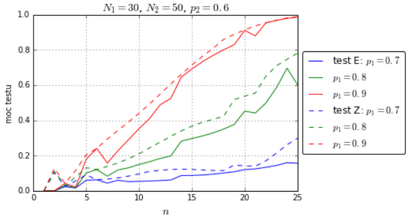


(c)

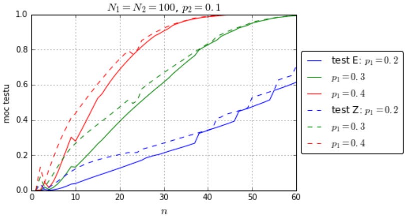


(d)

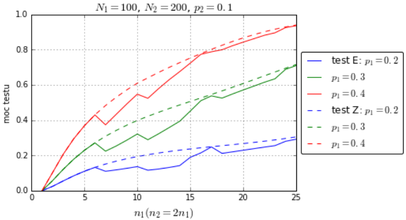
Rys. Moc testów Z i E jako funkcja rozmiaru próbki *n*



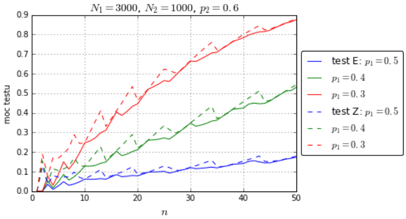
(a)



(b)



(c)

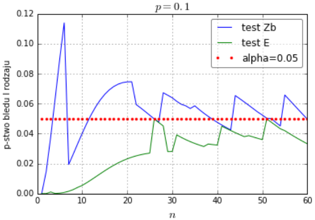


(d)

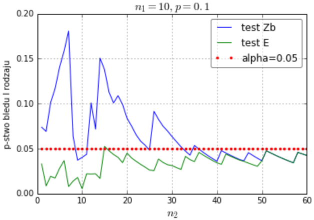
Analiza testów

Porównanie testu bez skończonej poprawki Zb z testem E

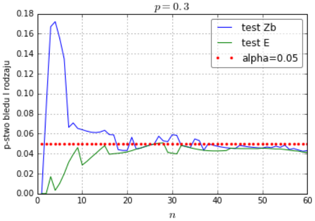
Rys. Prawdopodobieństwo błędu I rodzaju testów Zb i E jako funkcja rozmiaru próbki n; $\alpha = 0.05$; $N_1 = N_2 = 100$



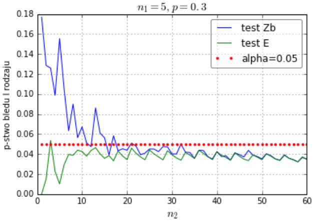
(a)



(b)

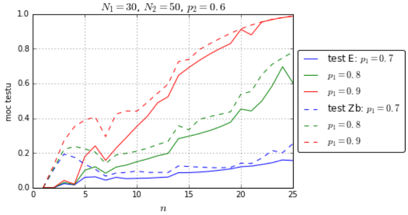


(c)

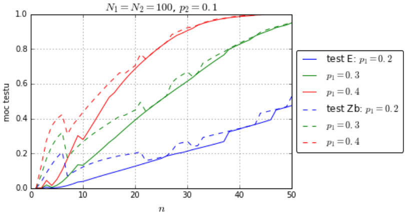


(d)

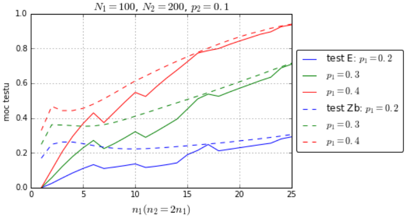
Rys. Moc testów Zb i E jako funkcja rozmiaru próbki *n*



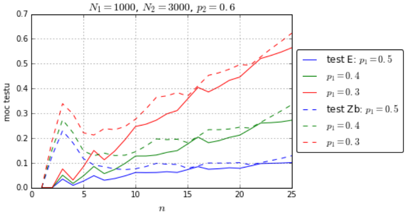
(a)



(b)



(c)



(d)

Wnioski

- Różnica między rozkładem dwumianowym i hipergeometrycznym jest najbardziej widoczna, gdy populacja jest mała albo obserwacja stanowi znaczną część populacji.

Wnioski

- Różnica między rozkładem dwumianowym i hipergeometrycznym jest najbardziej widoczna, gdy populacja jest mała albo obserwacja stanowi znaczną część populacji.
- Test Z nie utrzymuje poziomu istotności α, ze względu na zastosowanie aproksymacji rozkładem normalnym do statystyki testowej.

Wnioski

- Różnica między rozkładem dwumianowym i hipergeometrycznym jest najbardziej widoczna, gdy populacja jest mała albo obserwacja stanowi znaczną część populacji.
- Test Z nie utrzymuje poziomu istotności α, ze względu na zastosowanie aproksymacji rozkładem normalnym do statystyki testowej.
- Test Zb również przekracza poziom istotności α, szczególnie dla małych próbek.

Wnioski

- Różnica między rozkładem dwumianowym i hipergeometrycznym jest najbardziej widoczna, gdy populacja jest mała albo obserwacja stanowi znaczną część populacji.
- Test Z nie utrzymuje poziomu istotności α, ze względu na zastosowanie aproksymacji rozkładem normalnym do statystyki testowej.
- Test Zb również przekracza poziom istotności α, szczególnie dla małych próbek.
- Stosowanie testu bez skończonej poprawki w sytuacji małej populacji wiąże się z dużymi błędami.

Wnioski

- Różnica między rozkładem dwumianowym i hipergeometrycznym jest najbardziej widoczna, gdy populacja jest mała albo obserwacja stanowi znaczną część populacji.
- Test Z nie utrzymuje poziomu istotności α, ze względu na zastosowanie aproksymacji rozkładem normalnym do statystyki testowej.
- Test Zb również przekracza poziom istotności α, szczególnie dla małych próbek.
- Stosowanie testu bez skończonej poprawki w sytuacji małej populacji wiąże się z dużymi błędami.
- Test E jest dobry w przypadku małej populacji - nie wykracza znacząco powyżej poziomu istotności oraz jego moc jest niewiele mniejsza od mocy testów Z i Zb.

Bibliografia

- K. Krishnamoorthy, J. Thomson. „Hypothesis testing about proportions in two finite populations”. The American Statistician, 56(1):215–222, 2002.
- J. P. Buonaccorsi. „A note on confidence intervals for proportions in finite populations”. The American Statistician, 41(3):215–218, 1987.
- W. Wang. „Exact optimal confidence intervals for hypergeometric parameters”. Journal of the American Statistical Association, 110(512):1491–1499, Dec. 2015.
- B. E. Storer, C. Kim. „Exact properties of some exact test statistics for comparing two binomial proportions”. Journal of the American Statistical Association, 85:146–155, 1990.
- E. L. Lehmann. „Testowanie hipotez statystycznych”. PWN, 1968.