



Politechnika Wrocławska

Wydział Matematyki

Kierunek: matematyka stosowana

Specjalność: *nie dotyczy*

Praca dyplomowa — inżynierska

TESTOWANIE HIPOTEZ I ESTYMACJA W SYTUACJI POPULACJI SKOŃCZONEGO ROZMIARU

Kinga Kurowska

Słowa kluczowe:
testowanie hipotez
przedziały nieufności
zastosowanie w biologii i medycynie

Krótkie streszczenie:

Tu znajdzie się krótki streszczenie mojej pracy... Tu znajdzie się krótki streszczenie mojej pracy... Tu znajdzie się krótki streszczenie mojej pracy... Tu znajdzie się krótki streszczenie mojej pracy... Tu znajdzie się krótki streszczenie mojej pracy... Tu znajdzie się krótki streszczenie mojej pracy... Tu znajdzie się krótki streszczenie mojej pracy...

Opiekun pracy	dr inż. Andrzej Giniewicz		
dyplomowej	Stopień naukowy, imię i nazwisko	Ocena	Podpis

Do celów archiwalnych pracę dyplomową zakwalifikowano do: *

- a) kategorii A (akta wieczyste),
- b) kategorii BE 50 (po 50 latach podlegające ekspertyzie).

* niepotrzebne skreślić

pieczęć wydziałowa

Wrocław, rok 2017



Politechnika Wrocławska

Faculty of Pure and Applied Mathematics

Field of study: Applied Mathematics

Specialty: *not applicable*

Engineering Thesis

**HYPOTHESIS TESTING AND ESTIMATION
IN THE CASE OF FINITE POPULATION SIZE**

Kinga Kurowska

keywords:

hypothesis testing

tolerance and confidence regions

applications to biology and medical sciences

Short summary:

Here will be short summuary of my bachelor thesis... Here will be short summuary of my bachelor thesis... Here will be short summuary of my bachelor thesis... Here will be short summuary of my bachelor thesis... Here will be short summuary of my bachelor thesis...

Supervisor	dr inż. Andrzej Giniewicz		
	<i>Title, degree, name and surname</i>	<i>Grade</i>	<i>Signature</i>

*For the purposes of archival thesis qualified to: **

a) Category A (perpetual files)


b) Category BE 50 (subject to expertise after 50 years)

** Delete as appropriate*

stamp of the faculty

Wrocław, 2017

Spis treści

0	Wstęp	5
	1 Przedstawienie testów	9
1.1	Sformułowanie problemu	9
1.2	Test Z	9
1.3	Test E	10
1.4	Błąd I rodzaju	10
1.5	Moc testu	10
1.6	Test bez skończonej poprawki	11
	Spis rysunków	12

0



Wstęp

Początki teorii rachunku prawdopodobieństwa i statystyki sięgają XVI wieku. Na początku były to analizy rzutu kostką, czy prawdopodobieństwa błędów pomiarowych. Już w XVII Blaise Pascal formułuje i dowodzi własności trójkąta arytmetycznego oraz używa pojęcia kombinacji. A na początku XVIII wieku opublikowane zostają prace Jacoba Bernoullego, w których zawarł wiele swoich tez na temat prawdopodobieństwa. Przez te kilka wieków teoria rachunku prawdopodobieństwa i statystyki znacząco się wzbogaciła i rozwinęła. Powstał temat estymacji i testowania hipotez, który w naszych czasach jest zasadniczą domeną statystyki.

W przypadku dyskretnym najczęściej testowane są proporcje populacji. Albo czy dana próbka ma jakąś konkretną proporcję, albo czy dwie próbki mają tę samą. Znana jest powszechnie teoria dotycząca testowania hipotez, gdy populacja jest nieskończona (a raczej na tyle duża, że możemy ją w przybliżeniu uznać za nieskończoną). Wtedy proporcja to parametr p rozkładu dwumianowego. Jednak przypadek nieskończonej populacji nie wyczerpuje tematu testowania proporcji. Gdy populacja jest bardzo mała albo, gdy próbka jest niewiele mniejsza od całej populacji rozkład dwumianowy nie jest dobrym modelem. Tymczasem dobrze taką sytuację modeluje rozkład hipergeometryczny, ponieważ uwzględnia on rozmiar populacji. Załóżmy, że N będzie rozmiarem populacji, n rozmiarem próbki, a M ilością osobników w populacji z daną cechą (której proporcje będziemy testować). Wtedy zmienna losowa X z rozkładu hipergeometrycznego określa ilość osobników z daną cechą w próbce. Funkcja prawdopodobieństwa określona jest wzorem

$$h(k; n, M, N) = P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}, \quad (0.0.1)$$

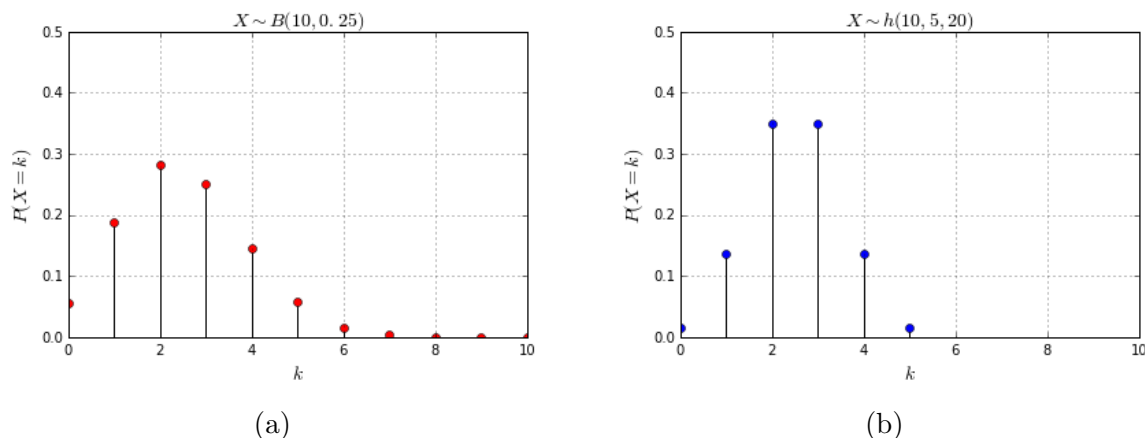
gdzie $L \leq k \leq U$, $L = \max\{0, M - N + n\}$ i $U = \min\{n, M\}$.

Zauważmy, że wzór funkcji h jest dość intuicyjny. Klasyczna definicja prawdopodobieństwa określa prawdopodobieństwo zajścia zdarzenia A jako iloraz liczby zdarzeń elementarnych w A przez liczbę zdarzeń elementarnych w Ω , czyli $P(A) = |A|/|\Omega|$. W tym przypadku zdarzeniem A jest to, że w próbce będzie k osobników z daną cechą. Zatem ilość zdarzeń elementarnych w A to kombinacje. Na ile różnych sposobów możemy wybrać k osobników z M wszystkich posiadających daną cechę $\binom{M}{k}$ razy możliwość wyborów pozostałych osobników z reszty populacji $\binom{N-M}{n-k}$. Natomiast zbiór wszystkich zdarzeń elementarnych w Ω to po prostu wybór losowej próbki n osobników z N -elementowej populacji $\binom{N}{n}$. Również ograniczenia nałożone na k są naturalne. Dolne ograniczenie L jest równe maksimum z 0 i $M - N + n$. Zatem może mieć tę drugą wartość, gdy jest ona

większa od zera. Zapiszmy to w ten sposób $n - (N - M) > 0$. Przenosząc na drugą stronę, otrzymujemy $n > N - M$. Taka postać jasno pokazuje, że jest to przypadek, w którym wielkość próbki przekracza ilość osobników w populacji bez badanej cechy. W konsekwencji czego mamy pewność, że w próbce będzie przynajmniej tyle osobników z daną cechą, ile wynosi różnica $n - (N - M)$. Ograniczenie górne jest mniej skomplikowane. Jest ono równe minimum z n i M , co jest oczywiste, że nie może być więcej osób w próbce z daną cechą niż w całej populacji. Cała ta analiza pokazuje, że rozkład hipergeometryczny jest ściśle związany z rozmiarem populacji.

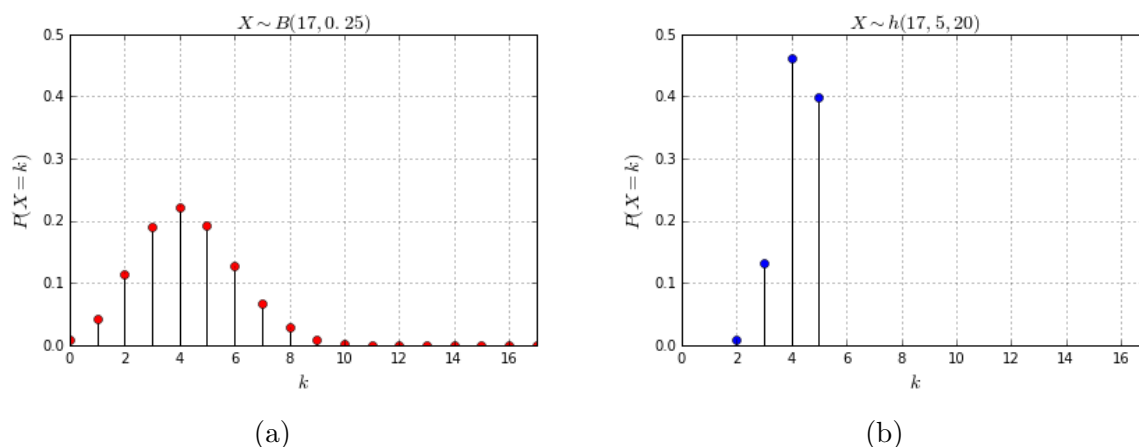
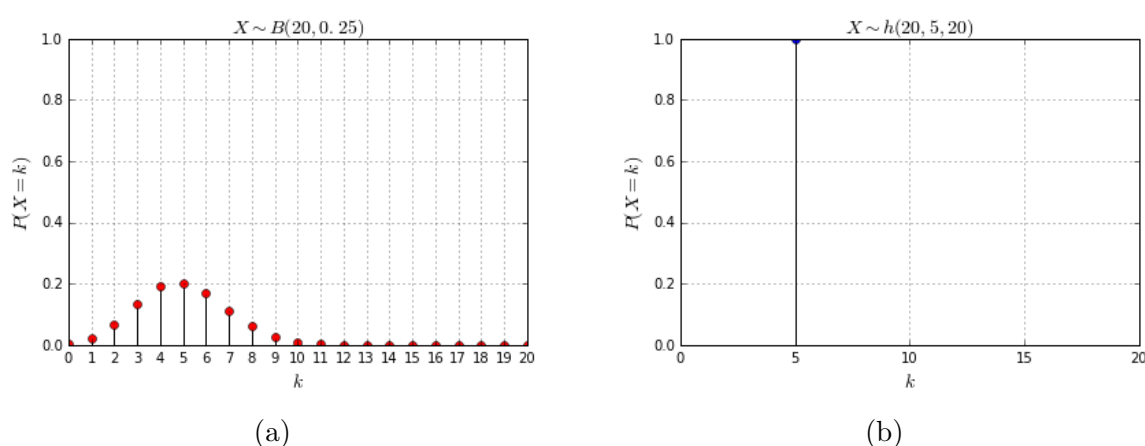
Rozkład hipergeometryczny daje bardzo podobne wyniki do rozkładu Bernoulliego, gdy populacja jest duża, a próbka stosunkowo mała. Jednakże w przeciwnym wypadku widać znaczne różnice między tymi rozkładami. Szczególnie to, że zmienna losowa z rozkładu dwumianowego może osiągać wartości od 0 do n , ponieważ podczas brania próbki zakładamy losowanie ze zwracaniem, czyli za każdym razem jest losowanie z takiej samej populacji. Natomiast w rozkładzie hipergeometrycznym kolejne losowania są od siebie zależne, prawdopodobieństwo wylosowania zmienia się w zależności od tego, co już wcześniej znalazło się w próbce.

Rozważmy to na medycznym przykładzie. Załóżmy, że jest na świecie 20 osób, które są chore na jakąś bardzo rzadką chorobę oraz że 25% z nich ma szansę na wyzdrowienie. Chcemy dowiedzieć się, ile osób spośród przebadanych może wyzdrowieć. Weźmy 3 różne próbki o wielkościach n równych odpowiednio 10, 17, 20. Możemy tę sytuację zamodelować rozkładem Bernoulliego, wtedy badana zmienna losowa będzie z rozkładu $B(n, 0.25)$. Drugim sposobem jest rozkład hipergeometryczny, wtedy zmienna losowa jest z rozkładu $h(n, 4, 20)$. Na rysunkach 0.1-0.3 jest zobrazowana funkcja prawdopodobieństwa dla wymienionych przypadków.



Rysunek 0.1: Funkcja prawdopodobieństwa dla $n = 10$

Im większa próbka tym widać większą różnicę między funkcjami obu rozkładów. Funkcja dla rozkładu hipergeometrycznego nie ma innych argumentów niż te, które są możliwe, natomiast funkcja dla rozkładu dwumianowego jest liczona również dla nieprawdopodobnych argumentów, ponadto wtedy jej wartość jest niezerowa. Bardzo skrajny przypadek jest ukazany na rysunku 0.3, gdy próbka równa się populacji, czyli tak naprawdę wiemy, że wszystko. Rysunek 0.3b idealnie obrazuje nam sytuację. Jest prawie pewne to, że w 20 osobach będzie dokładnie 5, które mogą wyzdrowieć. Jednakże wykres 0.3a kompletnie nie pokazuje tego. Głównie ze względu na już wspomniane uwzględnianie nieprawdopodobnych argumentów, przez co prawdopodobieństwo rozkłada się na pozostałe przypadki.

Rysunek 0.2: Funkcja prawdopodobieństwa dla $n = 17$ Rysunek 0.3: Funkcja prawdopodobieństwa dla $n = 20$ 

Warto również wspomnieć, że im większa próbka, tym prawdopodobieństwo, że $X = 5$ dla rozkładu Bernoullego jest coraz mniejsze (bo mamy więcej prób). Tymczasem dla rozkładu hipergeometrycznego jest wręcz odwrotnie, to prawdopodobieństwo rośnie, aż w końcu osiąga wartość 1. Co jest dużo bardziej logiczne, bo im więcej przebadaliśmy osobników, tym więcej wiemy o próbce i jest bardziej prawdopodobne, że jest w niej aż 5 szczęśliwych pacjentów.

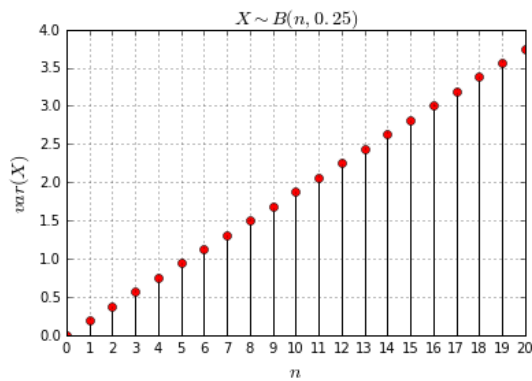


Spójrzmy jeszcze, jak wygląda wariancja dla obu przypadków. Dla rozkładu dwumianowego wariancja wyraża się wzorem $np(1 - p)$, a dla rozkładu hipergeometrycznego $n(M/N)(1 - M/N)(N - n)(N - 1)$. Na rysunku 0.4 widać bardzo wyraźnie różnice między analizowanymi rozkładami. W przypadku rozkładu dwumianowego wariancja stale rośnie wraz ze wzrostem próbki. Ostatecznie, gdy już przebadamy wszystkich chorych pacjentów, ma ona największą wartość. Myśląc zdroworozsądkowo, nie jest to poprawny wynik.

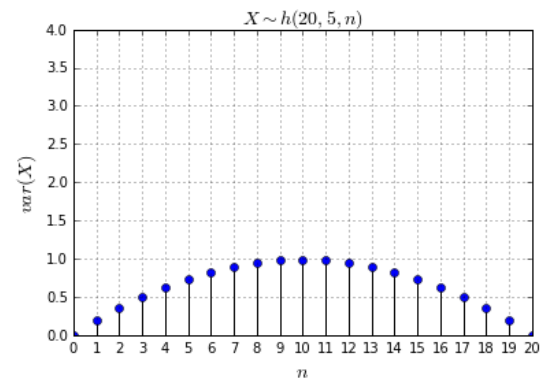


Raczej byśmy oczekiwali, że gdy przebadamy już wszystkich, wariancja osiągnie wartość 0, ponieważ nie ma wtedy losowości. Taki rezultat daje nam wykres 0.4b. Funkcja na początku rośnie, ale gdy wielkość próbki przekroczy połowę rozmiaru populacji, wariancja zaczyna maleć aż do zera. Odzwierciedla to fakt, że gdy coraz więcej wiemy o populacji, losowość uzyskanych wyników maleje.

Przedstawione zestawienie miało na celu pokazanie, że warto zająć się teorią testowania hipotez dla skończonej populacji. W określonych przypadkach rozkład hipergeometryczny



(a)



(b)


Rysunek 0.4: Wariancja w zależności od rozmiaru próbki

daje dużo dokładniejszą informację o badanym przypadku niż przybliżenie rozkładem dwumianowym. Ponadto zastosowanie tego typu testów ma duże znaczenie w medycynie, gdzie często rozważane populacje mają na tyle wyspecjalizowane cechy, że są uważane za małe. W dalszej części pracy opiszę dwa testy oparte o rozkład hipergeometryczny oraz przeanalizuję dla nich prawdopodobieństwo błędu I-rodzaju i moc testu. Porównam je także z testem wykorzystującym rozkładem Bernoullego.



1

Przedstawienie testów

 W tym rozdziale chciałabym opisać dwa testy wykorzystujące rozkład hipergeometryczny oraz test oparty na rozkładzie dwumianowym. Omówię także sposób liczenia prawdopodobieństwa błędu I rodzaju i mocy testu.

1.1 Sformułowanie problemu


Założmy, że X_1 i X_2 są niezależnymi zmiennymi losowymi o rozkładzie hipergeometrycznym $X_1 \sim h(n_1, M_1, N_1)$, $X_2 \sim h(n_2, M_2, N_2)$. Zaobserwowane wartości X_1 i X_2 oznaczmy odpowiednio k_1 i k_2 oraz proporcje $p_1 = M_1/N_1$, $p_2 = M_2/N_2$. Będę zajmować się testowaniem hipotez

$$H_0: p_1 = p_2 \quad \text{przeciwko} \quad H_1: p_1 \neq p_2, \quad (1.1.1)$$

w oparciu o (k_1, n_1, N_1) i (k_2, n_2, N_2) . Rozważmy unormowaną statystykę

$$Z_{X_1, X_2} = \frac{X_1/n_1 - X_2/n_2}{\sqrt{V_{X_1, X_2}}}, \quad (1.1.2)$$


gdzie estymator wariancji pod warunkiem zachodzenia H_0 jest równy




$$V_{X_1, X_2} = \left(\frac{N_1 - n_1}{n_1(N_1 - 1)} + \frac{N_2 - n_2}{n_2(N_2 - 1)} \right) \left(\frac{X_1 + X_2}{n_1 + n_2} \right) \left(1 - \frac{X_1 + X_2}{n_1 + n_2} \right). \quad (1.1.3)$$

Wartość statystyki dla k_1 i k_2 będę oznaczać jako Z_{k_1, k_2} . Jest ona wyliczana według powyższych wzorów, zamieniając X_1 i X_2 wartościami obserwacji.

1.2 Test Z

Ten test jest oparty na centralnym twierdzeniu granicznym, które mówi, że pod warunkiem H_0 w przybliżeniu rozważana statystyka Z_{X_1, X_2} jest z rozkładu normalnego standaryzowanego $N(0, 1)$. Wtedy p -wartość wyraża się wzorem 

$$P(|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}| | H_0) = 2(1 - \Phi(|Z_{k_1, k_2}|)), \quad (1.2.1)$$

 gdzie $\Phi()$ oznacza dystrybuantę rozkładu $N(0, 1)$. Test Z odrzuca hipotezę zerową, gdy p -wartość jest mniejsza od poziomu istotności α .

1.3 Test E

W tym przypadku opieramy się o rzeczywistą p -wartość, która jest równa



$$\begin{aligned} P(|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}| | H_0) &= E_{X_1, X_2}(\mathbf{1}(|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}|) | H_0) = \\ &= \sum_{x_1=L_1}^{U_1} \sum_{x_2=L_2}^{U_2} h(x_1; n_1, N_1 p, N_1) h(x_2; n_2, N_2 p, N_2) \mathbf{1}(|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}|), \end{aligned} \quad (1.3.1)$$

gdzie E_{X_1, X_2} to wartość oczekiwana łącznego rozkładu (X_1, X_2) , a p jest nieznaną wspólną proporcją pod warunkiem H_0 . Nie jest możliwe policzenie p -wartości wprost ze wzoru (1.3.1), ponieważ nie znamy parametru proporcji p . W artykule [1] zaproponowany jest estymator p -wartości



$$\begin{aligned} P(|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}| | H_0) &= \\ &= \sum_{x_1=L_{x_1}}^{U_{x_1}} \sum_{x_2=L_{x_2}}^{U_{x_2}} h(x_1; n_1, \hat{M}_1, N_1) h(x_2; n_2, \hat{M}_2, N_2) \mathbf{1}(|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}|), \end{aligned} \quad (1.3.2)$$

przy czym $\hat{p} = (k_1 + k_2)/(n_1 + n_2)$, $\hat{M}_i = [N_i \hat{p}]$, $L_{x_i} = \max\{0, \hat{M}_i - N_i + n_i\}$, $U_{x_i} = \min\{n_i, \hat{M}_i\}$, $i = 1, 2$. Test odrzuca H_0 wtedy, gdy p -wartość wyliczona wg wzoru (1.3.2) jest mniejsza od poziomu istotności α .



1.4 Błąd I rodzaju



Błąd I rodzaju to odrzucenie hipotezy zerowej, gdy jest ona prawdziwa. Prawdopodobieństwo tego błędu można wyliczyć, losując próbki z populacji, gdy $p_1 = p_2$ i sprawdzając, ile razy zostaną odrzucone. Dla dużej ilości próbek prawdopodobieństwo błędu I rodzaju powinno być w okolicy poziomu istotności testu.

1.5 Moc testu



Przypomnę, że moc testu to prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona nieprawdziwa. Toteż jest ona wyznacznikiem dobrego testu. Większa wartość mocy oznacza lepszy test.


Moc obu testów można wyliczyć, korzystając z funkcji prawdopodobieństwa rozkładu hipergeometrycznego. Dla testu Z pod warunkiem hipotezy alternatywnej H_1 moc jest równa



$$\sum_{k_1=L_1}^{U_1} \sum_{k_2=L_2}^{U_2} h(k_1; n_1, M_1, N_1) h(k_2; n_2, M_2, N_2) \mathbf{1}(|Z_{k_1, k_2}| > z_{1-\alpha/2}), \quad (1.5.1)$$

gdzie $L_i = \max\{0, M_i - N_i + n_i\}$ i $U_i = \min\{n_i, M_i\}$, a $z_{1-\alpha/2}$ oznacza kwantyl rozkładu normalnego standardowego rzędu $1 - \alpha/2$.

Tymczasem dla testu E moc zdefiniowana jest następująco




$$\sum_{k_1=L_1}^{U_1} \sum_{k_2=L_2}^{U_2} h(k_1; n_1, M_1, N_1) h(k_2; n_2, M_2, N_2) \times \\ \times \mathbb{1} \left(\sum_{x_1=L_{x_1}}^{U_{x_1}} \sum_{x_2=L_{x_2}}^{U_{x_2}} h(x_1; n_1, \hat{M}_1, N_1) h(x_2; n_2, \hat{M}_2, N_2) \mathbb{1} (|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}|) \leq \alpha \right), \quad (1.5.2)$$


gdzie wszelkie parametry oznaczają to samo co we wzorach (1.3.2) i (1.5.1).



1.6 Test bez skończonej poprawki




W tym przypadku zamiast rozkładu hipergeometrycznego używamy dwumianowego, a więc X_1 i X_2 są niezależnymi zmiennymi losowymi o rozkładzie Bernoullego $X_1 \sim B(n_1, p_1)$, $X_2 \sim B(n_2, p_2)$. Unormowana statystyka w tym przypadku przyjmuje postać



$$Z_{X_1, X_2} = \frac{X_1/n_1 - X_2/n_2}{\sqrt{V_{X_1, X_2}}}, \quad (1.6.1)$$

gdzie estymator wariancji pod warunkiem H_0 jest równy



$$V_{X_1, X_2} = \sqrt{p(1-p)(1/n_1 + 1/n_2)}, \quad (1.6.2)$$

gdzie $p = (X_1 + X_2)/(n_1 + n_2)$. Wartość statystyki Z_{k_1, k_2} jest wyliczana według powyższych wzorów, wstawiając obserwacje k_1 i k_2 w miejsca zmiennych losowych.


Ten test jest, podobnie jak omówiony wcześniej test Z, oparty na centralnym twierdzeniu granicznym. Czyli rozważana statystyka Z_{X_1, X_2} ma rozkład standardowy normalny $N(0, 1)$. Wtedy p -wartość wyraża się wzorem

$$P(|Z_{X_1, X_2}| \geq |Z_{k_1, k_2}| | H_0) = 2(1 - \Phi(|Z_{k_1, k_2}|)), \quad (1.6.3)$$



Test odrzuca hipotezę zerową, gdy p -wartość jest mniejsza od poziomu istotności α .

Prawdopodobieństwo błędu I rodzaju jest liczone analogicznie jak w przypadku poprzednich testów. Aczkolwiek moc testu jest liczona inaczej, ze względu na inny rozkład zmiennych losowych. W tym przypadku będzie ona równa



$$\sum_{k_1=0}^n \sum_{k_2=0}^n b(k_1; n_1, p_1) b(k_2; n_2, p_2) \mathbb{1} (|Z_{k_1, k_2}| > z_{1-\alpha/2}), \quad (1.6.4)$$

przy czym $b(k; n, p)$ oznacza funkcję prawdopodobieństwa rozkładu dwumianowego określoną wzorem

$$b(k; n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n. \quad (1.6.5)$$

Spis rysunków

0.1	Funkcja prawdopodobieństwa dla $n = 10$	6
0.2	Funkcja prawdopodobieństwa dla $n = 17$	7
0.3	Funkcja prawdopodobieństwa dla $n = 20$	7
0.4	Wariancja w zależności od rozmiaru próbki	8

Bibliografia



- [1] J. Thomson K. Krishnamoorthy. Hypothesis testing about proportions in two finite populations. *The American Statistician*, 56(1):215–222, 2002.