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 dyplomowej	dr inż. Andrzej Giniewicz		
	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*

pieczątka 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		
	Title, degree, name and surname	*Grade*	*Signature*

*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

Wstęp

Już na początku XVIII wieku ukazały się pierwsze prace związane z rozkładem dwumianowym (Bernoullego). Na przestrzeni tych kilku wieków teoria rachunku prawdopodobieństwa i statystyki znacząco się rozwinęła. *coś tu napisać o testowaniu ogólnie* W przypadku dyskretnym, najczęściej testowane są proporcje populacji. Albo czy dana próbka ma jakąś konkretną proporcję, albo czy dwie próbki mają tą samą. Obecnie 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

$$h(k; n, M, N) = P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \tag{1}$$

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

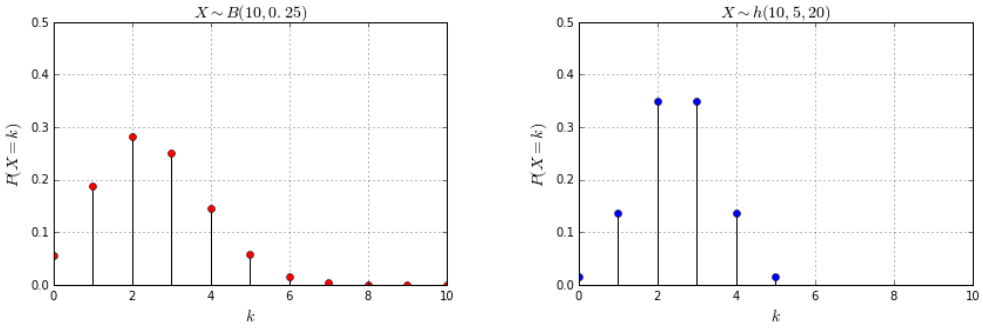
Zauważmy, że wzór funkcji h jest dość intuicyjny. Klasyczna definicja prawdopodobieństwa określa prawdopodobieństwo zajścia zdarzenia A jako iloraz liczby zdarzeń elementarnych w A przez liczbę zdarzeń elementarnych w Ω, czyli $P(A) = \frac{|A|}{|\Omega|}$. W tym przypadku zdarzeniem A jest to, że w 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 $\binom{N}{n}$. Także 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.

Tu bym chciała napisać o tym że rozkład hipergeometryczny daje bardzo podobne wyniki do rozkładu Bernoullego gdy populacja jest duża, a próbka stosunkowo mała. Jednakże w przeciwnym wypadku widać znaczne różnice między tymi rozkładami, szczególnie to, że wartość zmiennej losowej 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 poopulacji. 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.

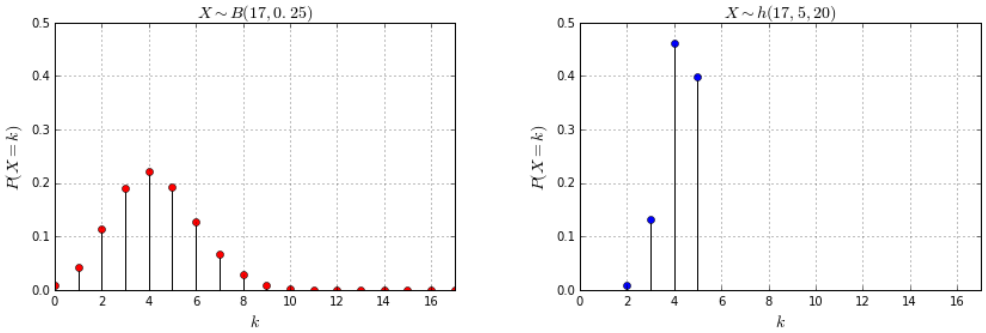
Rozważmy to na medycznym przykładzie. Załóżmy, że jest na świecie 20 osób, które są chore na jakąś bardzo rzadką chorobę oraz że 25% z nich ma szanse na wyzdrowienie. Chcemy zbadać ile osób spośród przebadanych może wyzdrowieć. Weźmy 3 różne próbki o wielkościach n równych odpowiednio 10, 17, 20. Możemy tę sytuację zamodelować rozkładem Bernoullego, 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 1-3 jest zobrazowana funkcja prawdopodobieństwa dla wymienionych przypadków.

Napiszę tu, że im większa próbka tym widać większą różnicę między funkcjami. Oraz zdecydowanie widać, że funkcja dla rozkładu hipergeometrycznego nie ma innych argumentów niż te, które są możliwe, natomiast funkcja dla rozkładu dwumianowego jest liczona również dla nieprawdopodobnych argumentów i wtedy nawet jej wartość jest niezerowa. Bardzo skrajnym przypadek jest ukazany na rysunku 3, gdy próbka jest równa populacji, czyli tak naprawdę wiemy już wszystko. Rysunek 3b idealnie obrazuje nam sytuację. Jest prawie pewne to że w 20 osobach będzie dokładnie 5, które mogą wyzdrowieć. Jednakże wykres 3a kompletnie nie pokazuje tego. Głównie ze względu



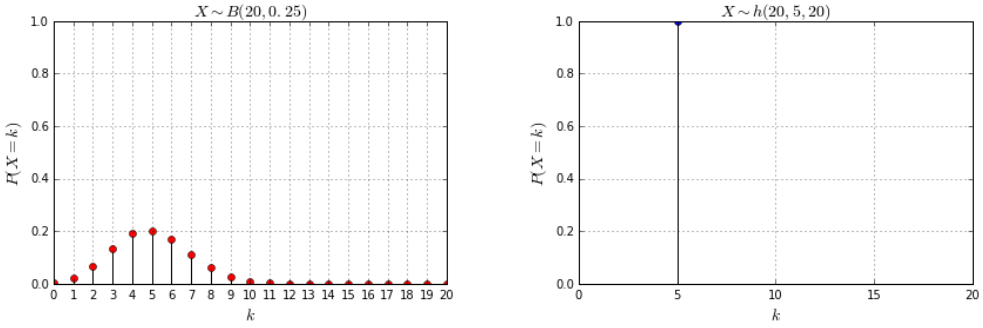
(a) (b)

Rysunek 1: Funkcja prawdopodobieństwa dla $n = 10$



(a) (b)

Rysunek 2: Funkcja prawdopodobieństwa dla $n = 17$



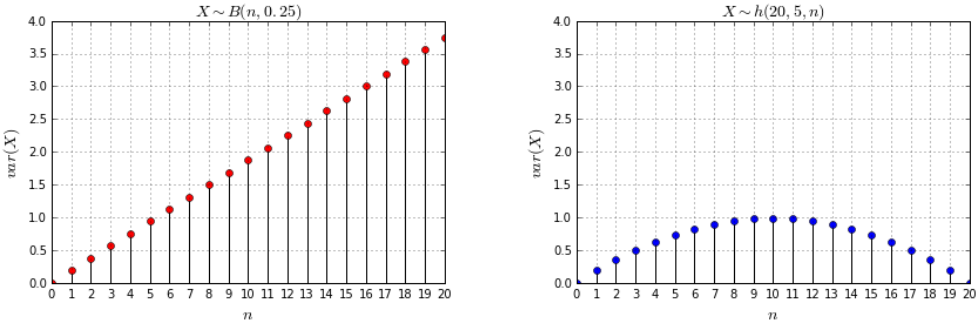
(a) (b)

Rysunek 3: Funkcja prawdopodobieństwa dla $n = 20$

na już wspomniane uwzględnianie nieprawdopodobnych argumentów, przez co prawdopodobieństwo rozkłada się na pozostałe przypadki. 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 4 widać bardzo wyraźnie różnice między analizowanymi rozkładami. W przypadku rozkładu dwumianowego wariancja stale rośnie wraz ze wzrostem próbki. Ostatecznie gdy już przebadamy wszystkim 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 już losowości. Taki rezultat daje nam wykres 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.



(a) (b)

Rysunek 4: Wariancja w zależności od rozmiaru próbki

Przedstawione zestawienie miało na celu pokazanie, że warto zająć się teorią testowania hipotez dla skończonej populacji. W określonych przypadkach rozkład hipergeometryczny daje dużo dokładniejszą informację o badanym przypadku niż przybliżenie rozkładem dwumianowym. Ponadto nauka w tym kierunku jest stosunkowo nowa i cały czas jeszcze się rozwija. 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.

Rozdział 1

Opis testów

Spis rysunków