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 N, co jest zabezpieczeniem przed wzięciem próbki większej od 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 nieskończoności, 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. Aby ukazać te różnice przeanalizowałam wariancję obu rozkładów dla konkretnego przypadku.

Weźmy populację osób

Materiałów o testach hipotez z wykorzystaniem rozkładu hipergeometrycznego jest już dużo mniej. Możliwe, że wynika to z szczególnych przypadków, w których zastosowanie skończonej poprawki jest uzasadnione oraz tego że nauka w tym kierunku jest stosunkowo nowa i cały czas jeszcze się rozwija.

Bibliografia

[1] F. Y. Edgeworth. On the value of a mean as calculated from a sample. *Journal of the Royal Statistical Society*, 81(4):624–632, Jul. 1918.