



6. Estymacja przedziałowa. CLT.

6.1 Centralne Twierdzenie Graniczne

6.1.1 Rozkład dwumianowy

Definicja 6.1 — Rozkład dwumianowy. Opisuje prawdopodobieństwo otrzymania dokładnie k sukcesów przy n niezależnych próbach Bernoulliego ze stałym prawdopodobieństwem sukcesu p .

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{(n-k)}$$

Problem 6.1 Ile wynosi średnia i odchylenie standardowe tego rozkładu?

6.1.2 Wartość oczekiwana i wariancja średniej arytmetycznej

Przypomnijmy sobie jedno z ważniejszych odkryć zeszłych zajęć: ile wynosi wartość oczekiwana i wariancja średniej arytmetycznej (jeszcze raz oklaski dla kolegów, którzy zdołali to udowodnić na tablicy).

Twierdzenie 6.1 — Wartość oczekiwana i wariancja średniej arytmetycznej. Niech X będzie zmienną losową o wartości oczekiwanej $\mathbb{E}[X] = \mu$ i wariancji $\mathbb{D}^2[X] = \sigma^2$ wtedy^a:

$$\mathbb{E}[\bar{X}_n] = \mu \quad \mathbb{D}^2[\bar{X}_n] = \frac{\sigma^2}{n} \quad \mathbb{D}[\bar{X}_n] = \frac{\sigma}{\sqrt{n}}$$

^aZauważ, że jest to tylko wprowadzenie oznaczeń, nie zakładamy tutaj żadnego rozkładu X . W szczególności nie zakładamy rozkładu normalnego

Potencjalnym zaskoczeniem jest to, że wariancja średniej arytmetycznej jest mniejsza niż wariancja samej zmiennej. Dlaczego tak jest? Wyobraźmy sobie rozkład wieku wszystkich ludzi, który oczywiście **nie** jest normalny. Wylosujmy wiele próbek składających

się ze 100 ludzi i dla każdej z nich policzymy średnią, a następnie z policzonych średnich stwórzmy rozkład¹.

❗ Zwróć uwagę, że zmieniamy tutaj perspektywę: mówiąc o rozkładzie X mieliśmy na myśli rozkład cechy w badanej populacji, a mówiąc o rozkładzie \bar{X} mamy na myśli rozkład średniej w populacji próbek! Tak, dobrze przeczytałeś, jeśli badamy średnią to jej populacją jest populacja średnich policzonych na wszystkich możliwych próbkach o danym rozmiarze n .

Zauważ, że w każdej z próbek znajdują się ludzie starzy i młodzi, ale każdy z nich zostanie uśredniony do pewnej środkowej wartości. Dlatego też, w powstałym rozkładzie będzie „brakowało” najbardziej skrajnych wartości – przez co wariancja będzie mniejsza. Zawsze mogą się nam też zdarzyć próbki które będą zawierały same najmniejsze (lub największe) wartości i one będą tworzyły lewy (i prawy) koniec rozkładu. Jednak, im większy będzie rozmiar próbki tym prawdopodobieństwo uzyskania takiej skrajnej próby będzie się szybko zmniejszać². Jest więc intuicyjnie zrozumiałe dlaczego wariancja średniej maleje wraz ze wzrostem próbki.

6.1.3 Twierdzenie graniczne Lindeberga-Levy’ego

Ćwiczenie 6.1 Wejdź na stronę: https://gallery.shinyapps.io/CLT_mean/ Na początku wybierz rozkład normalny i poeksperymentuj z różnymi wielkościami próbki – jak wpływa to na rozkład średniej? Następnie poeksperymentuj z rozkładem skośnym – co możesz zaobserwować przy rozkładzie o wysokiej skośności? ■

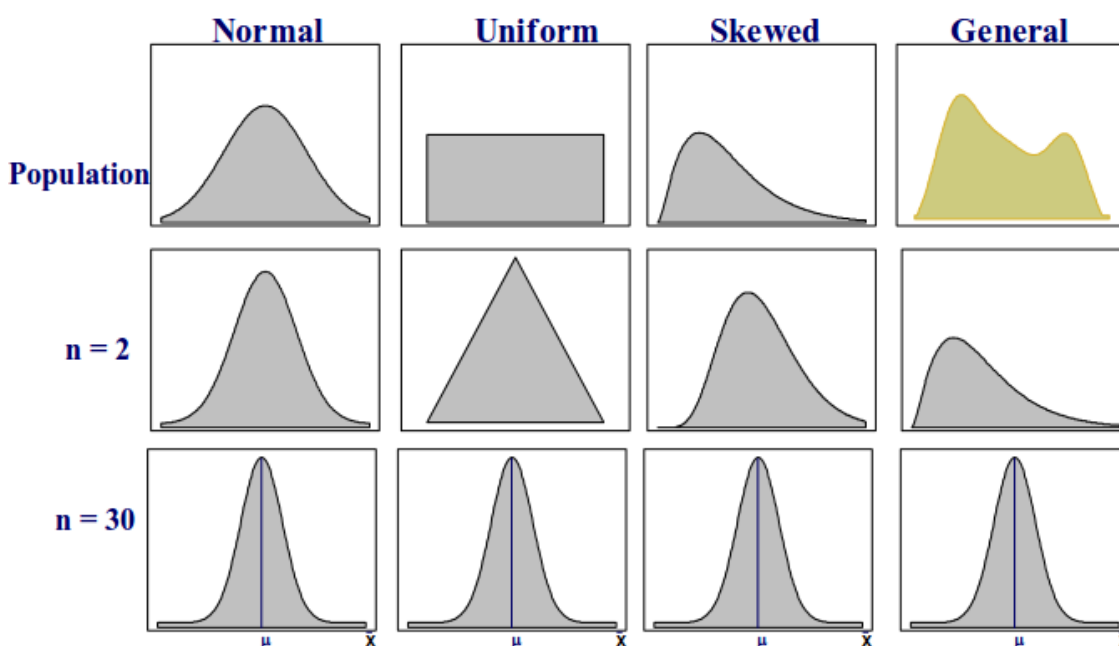
Ćwiczenie 6.2 Wejdź na stronę: https://gallery.shinyapps.io/CLT_prop/. Jaki rozkład ma wskaźnik struktury? Co stanie się gdy zwiększysz p do ekstremalnej wartości np. $p = 0.95$? ■

Twierdzenie 6.2 — Centralne Twierdzenie Graniczne (Lindeberga-Levy’ego) (ang. CLT, Central Limit Theorem). Niech X_1, X_2, \dots, X_n będzie ciągiem niezależnych zmiennych losowych o takim samym rozkładzie, ze skończoną wartością oczekiwaną μ oraz ze skończoną i niezerową wariancją σ^2 . Dodatkowo oznaczmy przez Z_n ustandaryzowaną zmienną $\bar{X}_n = n^{-1} \sum X_i$:

$$Z_n = \frac{\bar{X}_n - \mathbb{E}[\bar{X}_n]}{\sqrt{\mathbb{D}^2[\bar{X}_n]}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

¹Jest to konieczne, ponieważ chcemy zmierzyć wariancję średniej arytmetycznej, a średnia policzona na próbce jest tylko jedna (czyli liczenie wariancji byłoby bezsensu). W związku z tym aby otrzymać rozkład średniej arytmetycznej musimy wygenerować wiele próbek, policzyć z nich średnią i dopiero wtedy wygenerować rozkład tych średnich.

²W naszych rozważaniach najczęściej analizujemy populację o nieskończonym rozmiarze. Pomyśl jednak o 7,3 mld populacji ziemi i o próbach bez zwracania. Gdybyśmy rozważali populację średnich z 7,3 mld prób losowanych bez zwracania to wartość ta miałaby zerową wariancję. Dlaczego? Bo z 7,3 mld populacji jest możliwość uzyskania tylko jednej próby bez zwracania o rozmiarze 7,3 mld. Stąd taka średnia byłaby po prostu stałą.



Rysunek 6.1: Ilustracja Centralnego Twierdzenia Granicznego

Wtedy:

$$\lim_{n \rightarrow \infty} P(Z_n < z) = \Phi(z)$$

Twierdzenie 6.3 — Centralne Twierdzenie Graniczne dla średniej (uproszczone).

Dla dostatecznie dużej próby losowej prostej średnia arytmetyczna (\bar{X}_n) ma rozkład normalny z odpowiednią średnią i odchyleniem standardowym.

$$\text{Dla dużych } n : \quad \bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Istnieje także wersja Centralnego Twierdzenia Granicznego dla sumy.

Twierdzenie 6.4 — Centralne Twierdzenie Graniczne dla sumy (uproszczone).

Dla dostatecznie dużej próby losowej prostej suma ($\sum X_i$) ma rozkład normalny z odpowiednią średnią i odchyleniem standardowym.

$$\text{Dla dużych } n : \quad \sum_{i=1}^n X_i \sim N(n \cdot \mu, \sqrt{n} \cdot \sigma)$$

Omówmy jeszcze pokrótce założenia twierdzenia:

- obserwacje są niezależne.

To założenie spełniamy najczęściej poprzez losowe próbkowanie ze zwracaniem. Warto zauważyć, że dużo badań statystycznych jest realizowana na próbach bez zwracania (np. ta sama osoba raczej nie wypełnia tej samej ankiety dwa razy, nawet jeśli zostałaby wylosowana). W takiej sytuacji o ile próba w stosunku do populacji jest mała możemy uznać, że założenie jest spełnione. Częstym warunkiem jest

wymaganie aby rozmiar nie przekraczał 10% rozmiaru populacji.


- dostatecznie duża próba losowa.

Często w podręcznikach można znaleźć, że „dostatecznie duża próba losowa” to próba mająca powyżej 30 obserwacji. Nie jest to zawsze prawda! Jeżeli rozkład X jest bardzo skośny to rozmiar próby powinien być większy (im bardziej skośny tym rozmiar powinien być większy) – dlatego niektórzy podają, że $n > 100$ lub $n > 200$. Tak naprawdę, jeśli pracujesz z rozkładem mocno skośnym i masz wątpliwość czy przy danym rozmiarze próby rozkład średnich zaczął już zbiegać do normalnego powinieneś przeprowadzić dodatkową analizę sprawdzającą to założenie.

Takich heurystyk jest bardzo dużo np. dla rozkładu dwumianowego:

Definicja 6.2 Rozkład dwumianowy możemy przybliżyć rozkładem normalnym jeżeli liczba oczekiwanych sukcesów i porażek jest większa niż 5^a.

$$np > 5 \quad \wedge \quad n(1-p) > 5$$

^aNiektóre książki są bardziej zachowawcze i podają wartość 10. Osobiście proponuję następującą heurystykę: jeśli masz pod ręką komputer i możesz policzyć *dokładne* prawdopodobieństwo z rozkładu dwumianowego w  to po co stosować przybliżenia rozkładem normalnym? Jeśli obliczenie tego dokładnego prawdopodobieństwa na komputerze trwa dłużej niż kilka minut – zastosuj przybliżenie rozkładem normalnym.

■ **Przykład 6.1** Pewna fabryka produkuje wysokiej jakości smartwatch'e. Każdego dnia fabryka jest w stanie wyprodukować pewną liczbę smartwatch'y X , która ma jakiś rozkład o $\mu = 5$ i $\sigma = 3$. Firma ma szansę na bardzo opłacalne zamówienie na 1000 smartwatch'y, które musiałaby dostarczyć w przeciągu 8 miesięcy (≈ 240 dni). Jednakże, w przypadku zwłoki firma będzie musiała zapłacić tak dużą karę, że splajtuje. Dodatkowo wiesz, że zanim załatwisz wszystkie formalności i uruchomisz produkcję minie dokładnie 20 dni. Jako menadżer musisz podjąć decyzję czy przyjmujesz zamówienie.³ Dla uproszczenia przyjmij, że liczby wyprodukowanych zegarków w poszczególnych dniach są od siebie niezależne.

Aby podjąć taką decyzję najpierw spróbujmy oszacować ryzyko, że firma zapłaci karę. Ryzyko możemy oszacować poprzez obliczenie prawdopodobieństwa, że w przeciągu 220 dni nie uda się wyprodukować 1000 zegarków.

$$P\left(\sum_{i=1}^{220} X_i < 1000\right) = ?$$

Nasza próbka zawiera 220 wartości X , jeżeli więc nie spodziewamy się ekstremalnie skośnego rozkładu, możemy więc użyć Centralnego Twierdzenia Granicznego i przyjąć, że $\sum_{i=1}^{220} X_i$ ma rozkład normalny, pomimo tego, że nie wiemy jaki rozkład ma X .

$$\begin{aligned} P\left(\sum_{i=1}^{220} X_i < 1000\right) &= P\left(\frac{\sum_{i=1}^{220} X_i - 5 \cdot 220}{3 \cdot \sqrt{220}} < \frac{1000 - 5 \cdot 220}{3 \cdot \sqrt{220}}\right) \\ &= P\left(\frac{\sum_{i=1}^{220} X_i - 5 \cdot 220}{3 \cdot \sqrt{220}} < -2,24\right) \\ &= \Phi(-2,24) = 0.0125 \end{aligned}$$

³Zadanie jest wariacją zadania z kursu „Probabilistic Systems Analysis and Applied Probability” na MIT.

Cóż, dalej statystyka ci nie pomoże – czy podejmiesz ryzyko? ;)

Statystyka nie potrafi udowodniać rzeczy na 100%, a jedynie pozwala na oszacowanie prawdopodobieństwa przy użyciu danych. Na podstawie takich wyliczeń musisz samodzielnie podjąć decyzję o akceptacji (lub nie) stwierdzenia, warto więc odpowiedzieć sobie na pytanie: jak niskie powinno być prawdopodobieństwo abyś przyjął to zamówienie?

Problem 6.2 Na iPadzie masz 3000 piosenek. Średnia długość trwania piosenki to 3,45 minuty z $\sigma = 1,63$ minuty. Zrobiłem listę 100 piosenek, których będę słuchać w trakcie podróży z Poznania do Krakowa (6h). Jakie jest prawdopodobieństwo, że wystarczy mi piosenek do końca podróży?

6.1.4 ** Centralne Twierdzenie Graniczne - dlaczego rozkład normalny?

ADVANCED

Czasami trudno jest intuicyjnie zaakceptować fakt, że suma dużej liczby zmiennych losowych ma rozkład zbieżny do rozkładu normalnego. W tym celu pokażemy cztery wskazówki dlaczego tak jest, jednak aby zrobić to w przystępny sposób pozwolimy sobie na dodatkowe uproszczenia⁴. Oczywiście Centralne Twierdzenie Graniczne można udowodnić w sposób formalny, a dowód można bez problemu znaleźć np. w [3].

Po pierwsze, chwilowo założmy, że rzeczywiście suma znormalizowanych zmiennych losowych ($\mathbb{E}X = 0$ i $\mathbb{D}X = 1$) z próby losowej prostej zbiega do jakiegoś jednego hipotetycznego rozkładu prawdopodobieństwa $F(x)$. Przypomnijmy, że żeby wariancja sumy zmiennych losowych była taka sama jak wariancja tej zmiennej to musimy ją podzielić przez \sqrt{n} .

$$\mathbb{D}^2\left[\sum_{i=1}^n X_i\right] = n\sigma^2 \quad \mathbb{D}^2\left[\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right] = \frac{1}{n}\mathbb{D}^2\left[\sum_{i=1}^n X_i\right] = \frac{1}{n}n\sigma^2 = \sigma^2$$

Korzystając z naszego założenia wiemy, że $\frac{\sum_{i=1}^n X_i}{\sqrt{n}}$ zbiega do jakiegoś rozkładu F . Spróbujmy dodać dwie takie niezależne zmienne do siebie $\frac{\sum_{i=1}^n X_i}{\sqrt{n}} + \frac{\sum_{i=n+1}^{2n} X_i}{\sqrt{n}} = \frac{\sum_{i=1}^{2n} X_i}{\sqrt{n}}$ w rezultacie otrzymujemy zmienną, która na mocy naszego założenia, również dąży do takiego samego rozkładu co każda z tych dwóch sum! (bo też jest sumą zmiennych losowych) Jedna drobna różnica, że zbiega do rozkładu o innej wariancji... Aby zmienna ta zbiegała do rozkładu o takiej samej wariancji (równej 1), musielibyśmy ją znormalizować do $\frac{\sum_{i=1}^{2n} X_i}{\sqrt{2n}}$. Spodziewamy się, więc że $F + F = \sqrt{2}F$. Ciekawe jaki rozkład ma taką własność? Na podstawie twierdzenia z laboratoriów o rozkładzie normalnym (suma niezależnych normalnych jest normalna!): $N(0, 1) + N(0, 1) = N(0, \sqrt{1^2 + 1^2}) = N(0, \sqrt{2}) = \sqrt{2}N(0, 1)$. Widzimy więc, że jeśli taka własność istnieje to F może mieć właśnie rozkład normalny.

No właśnie, *jeśli* ta własność istnieje. Spróbujmy zaatakować ten problem. Założmy, że oprócz skończonej wariancji i wartości oczekiwanej (warunki CLT), X ma skończone wszystkie momenty. Jeśli są skończone to istnieje taka stała, która jest od nich większa. Oznaczmy $\mathbb{E}[X^3] \leq C_3$ i $\mathbb{E}[X^4] \leq C_4$ oraz przypomnijmy, że dla zmiennych niezależnych $\mathbb{E}[X_i X_j] = \mathbb{E}X_i \mathbb{E}X_j$, co dla zmiennych o średniej 0 przeradza się w $\mathbb{E}[X_i X_j] = 0$. Zwróć uwagę, że $\mathbb{E}[X_i^2] = \mathbb{E}[X_i X_i] \neq 0$, bo X_i nie jest niezależny od samego siebie!

⁴Jest to przetłumaczony i uproszczony opis z Filmus Y., *Two Proofs of the Central Limit Theorem* dostępny pod adresem <http://www.cs.toronto.edu/~yuvalf/CLT.pdf>

Ile wynosi drugi moment centralny (czyli wariancja) naszej sumy? Przy zerowej średniej momenty centralne są zwykłymi momentami (czyli tutaj $\mathbb{D}^2 X_i = \mathbb{E}[(X_i - \mu)^2] = \mathbb{E}X_i^2 = 1$).

$$\mathbb{E} \left[\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \right)^2 \right] = \frac{\sum_i \mathbb{E} X_i^2}{n} + \frac{\sum_{i \neq j} \mathbb{E}[X_i X_j]}{n} = \frac{n \mathbb{D}^2 X_i}{n} + \frac{\sum_{i \neq j} \cancel{\mathbb{E}[X_i]} \cancel{\mathbb{E}[X_j]}^0}{n} = 1 + 0 = 1$$

A więc wariancja tego rozkładu wynosi 1. Ile wynosi jego asymetria (przy $n \rightarrow \infty$)?

$$\begin{aligned} \mathbb{E} \left[\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \right)^3 \right] &= \frac{\sum_i \mathbb{E} X_i^3}{n^{3/2}} + 3 \frac{\sum_{i \neq j} \mathbb{E}[X_i^2 X_j]}{n^{3/2}} + \frac{\sum_{i \neq j \neq k} \mathbb{E}[X_i X_j X_k]}{n^{3/2}} \\ &= \frac{\sum_i \mathbb{E} X_i^3}{n^{3/2}} + 3 \frac{\sum_{i \neq j} \mathbb{E}[X_i^2] \mathbb{E}[X_j]^0}{n^{3/2}} + \frac{\sum_{i \neq j \neq k} \mathbb{E}[X_i]^0 \mathbb{E}[X_j]^0 \mathbb{E}[X_k]^0}{n^{3/2}} \\ &= \frac{\sum_i \mathbb{E} X_i^3}{n^{3/2}} + 0 + 0 \leq \frac{n C_3}{n^{3/2}} = \frac{C_3}{\sqrt{n}} \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

Hmm... A więc jest to rozkład o zerowej asymetrii⁵... Ciekawe, a ile wynosi jego kurtoza?

$$\begin{aligned} \mathbb{E} \left[\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \right)^4 \right] &= \frac{\sum_i \mathbb{E} X_i^4}{n^2} + 4 \frac{\sum_{i \neq j} \mathbb{E}[X_i^3 X_j]}{n^2} + 3 \frac{\sum_{i \neq j} \mathbb{E}[X_i^2 X_j^2]}{n^2} + 6 \frac{\sum_{i \neq j \neq k} \mathbb{E}[X_i^2 X_j X_k]}{n^2} \\ &\quad + \frac{\sum_{i \neq j \neq k \neq l} \mathbb{E}[X_i X_j X_k X_l]}{n^2} \\ &= \frac{\sum_i \mathbb{E} X_i^4}{n^2} + 0 + 3 \frac{\sum_{i \neq j} \mathbb{E}[X_i^2] \mathbb{E}[X_j^2]}{n^2} + 0 + 0 = \frac{\sum_i \mathbb{E} X_i^4}{n^2} + 3 \frac{\sum_{i \neq j} 1 \cdot 1}{n^2} \\ &= \frac{\sum_i \mathbb{E} X_i^4}{n^2} + 3 \frac{n(n-1)}{n^2} \leq \frac{n C_4}{n^2} + 3 \left(1 - \frac{1}{n} \right) \xrightarrow{n \rightarrow \infty} 0 + 3 = 3 \end{aligned}$$

Nie udało nam się dojść do tego czy suma znormalizowanych zmiennych losowych z próby losowej prostej o ograniczonych momentach dąży do tego samego rozkładu. Wiemy natomiast, że z całą pewnością wszystkie rozkłady tych sum dążą do rozkładu(ów) z wariancją 1, zerową asymetrią i kurtozą równą 3. Wnioski nasuwają się same, a niedowiarków zachęcam do udowadniania kolejnych momentów ;)

6.2 Estymacja przedziałowa

6.2.1 Analiza przedziałów ufności

Definicja 6.3 — Przedział ufności. Przedziałem ufności dla parametru θ na poziomie ufności $1 - \alpha$ nazywamy przedział $C_n = (a, b)$ gdzie a i b są funkcjami próby losowej takimi że

$$\forall_{\theta} P(\theta \in C_n) = 1 - \alpha$$

Najczęściej używa się wartości $\alpha = 5\%$.

⁵Rozkład symetryczny ma asymetrię 0, ale nie każdy rozkład o asymetrii 0 jest symetryczny!

Innymi słowy przedział ufności skonstruowany w ten sposób pokrywa prawdziwą wartość estymowanego parametru z prawdopodobieństwem $1 - \alpha$.

Zauważ, które części wzoru są zmiennymi losowymi: θ jest nieznanym (ale ustalonym) parametrem populacji, a przedział C_n jest zmienną losową. Z chwilą gdy skonstruujemy konkretny przedział ufności np. $c_n = (-5, 5)$ we wzorze $P(\theta \in c_n)$ mamy tylko konkretne wartości! Wynika z tego, że - zgodnie z definicją częstotliwościową - prawdopodobieństwo, że ten konkretny przedział pokrywa szukany parametr nie wynosi już $1 - \alpha$. Taki przedział albo pokrywa szukaną wartość albo jej nie pokrywa! Zauważ że wyrażenie $\theta \in c_n$ jest zwykłym wyrażeniem logicznym, które jest albo prawdziwe albo fałszywe.

Jak więc interpretować przedział ufności? Jeżeli powtórzyłbym eksperyment wielokrotnie (nieskończoną ilość razy) i dla każdego otrzymanego wyniku (próbki) stworzyłbym przedział ufności to tylko α procent z takich przedziałów nie pokrywałoby szukanej wartości. Jednak dany przedział, albo pokrywa daną wartość, albo nie (z prawdopodobieństwem 0% lub 100%).

Zwykle jednak nie powtarzamy danego eksperymentu (badania statystycznego) wiele razy, możemy więc przedział interpretować także w następujący sposób [3]. Prowadzisz firmę statystyczną: pierwszego dnia zbierasz dane i tworzysz dla parametru θ_1 przedział z poziomem ufności $1 - \alpha$. Następnego dnia zbierasz inne dane w celu estymacji innego parametru θ_2 (który nie jest w żaden sposób związany z θ_1) i tworzysz dla niego przedział ufności. Tę czynność kontynuujesz w kolejnych dniach funkcjonowania twojej firmy. W rezultacie $(1 - \alpha) \cdot 100$ procent skonstruowanych przez Ciebie przedziałów ufności będzie zawierało szukane parametry (pomylisz się w $\alpha \cdot 100$ procent przypadków). Możemy więc powiedzieć: „Mamy $(1 - \alpha) \cdot 100$ procent *ufności*, że prawdziwy parametr populacji jest pomiędzy ... i ...”. Nie możemy jednak powiedzieć, że prawdopodobieństwo tego, że prawdziwy parametr populacji jest w tym przedziale wynosi $(1 - \alpha)$. Twoja ufność a prawdopodobieństwo częstotliwościowe⁶ to nie to samo!

Jaki jest wobec tego sens znajdowania przedziału ufności, skoro wiemy, że najwyuczajniej w świecie szukana wartość jest lub nie jest w środku? „Kierujemy się tutaj praktyczną zasadą, według której zdarzenie o bardzo małym prawdopodobieństwie w jednym doświadczeniu praktycznie nie zachodzi [2]”. Dla przykładu rzucając jeden raz pięcioma monetami w praktyce nie liczymy się z możliwością, że wypadnie nam 5 orłów – pomimo tego, że zdarzenie to ma prawdopodobieństwo ok. 3%, a więc przeciętnie, przy powtarzaniu eksperymentu, w każdych 100 rzutach 3 razy zdarzy się taka sytuacja.

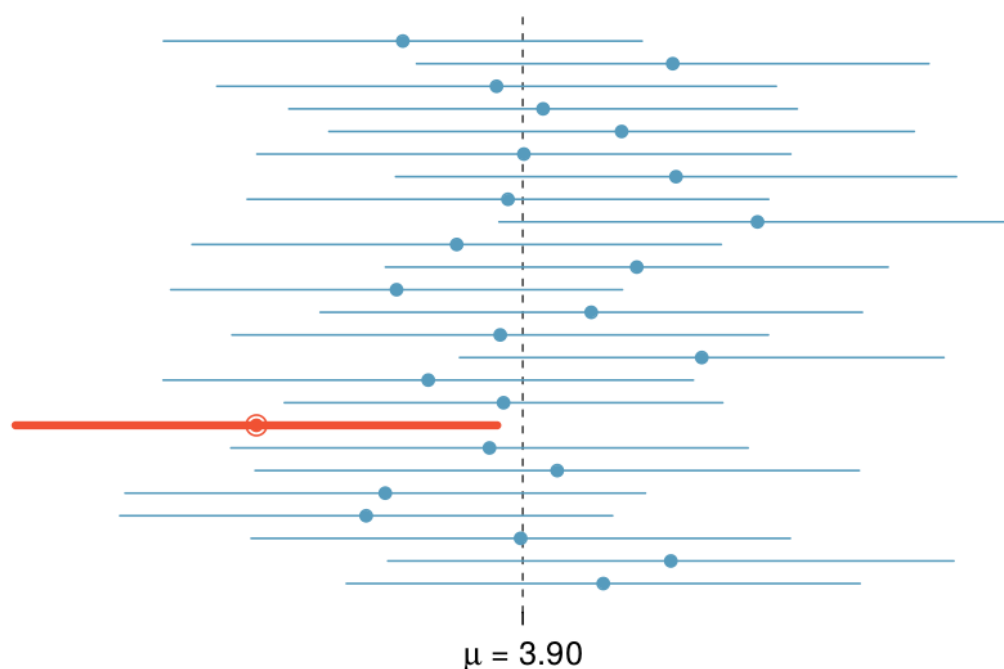
Problem 6.3 Które z poniższych stwierdzeń jest najlepszą interpretacją 95% przedziału ufności dla średniej?

- 95% rozkładu populacji jest zawarte w przedziale ufności
- Jeśli wzięliśmy wiele prób losowych i skonstruowaliśmy 95% przedział ufności dla każdej z nich to 95% tych przedziałów zawiera średnią populacji
- 95% przedział ufności ma 0.95 prawdopodobieństwo, że zawiera średnią populacji

Problem 6.4 Przeprowadzono badanie czasu dojazdu mieszkańców Poznania do pracy. W tym celu pobrano $n = 50$ elementową próbkę i uzyskano: średnią 24 minuty i 95% przedział ufności dla średniej czasu dojazdu: (17, 31). Które z poniższych stwierdzeń są prawidłowymi interpretacjami tego przedziału?

- 95% z tych 50 mieszkańców potrzebuje od 17 do 31 minut, aby dojechać do pracy

⁶Nawet ta interpretacja budzi wątpliwości statystyków Bayesowskich, którzy interpretują ufność (ang. *degree of believe*) jako prawdopodobieństwo!



Rysunek 6.2: Wizualizacja 25 przedziałów z 95% ufnością dla średniej wraz ze znaczoną prawdziwą wartością, która była estymowana. Tylko jeden z przedziałów nie zawiera prawdziwej wartości średniej.

- jest 95% prawdopodobieństwo, że prawdziwa średnia czasu dojazdu do pracy w Poznaniu jest pomiędzy 17 a 31 minut
- mamy 95% pewność, że prawdziwa średnia czasu dojazdu do pracy w Poznaniu jest pomiędzy 17 a 31 minut
- gdybyśmy powtórzyli to badanie wiele razy to 95% z uzyskanych średnich arytmetycznych byłoby pomiędzy 17 a 31 minut
- gdybyśmy powtórzyli to badanie wiele razy, i dla każdej z uzyskanych prób stworzylibyśmy przedział ufności to nasza średnia 24 minuty zawierałaby się w ok. 95% z nich
- gdybyśmy powtórzyli to badanie wiele razy, i dla każdej z uzyskanych prób stworzylibyśmy przedział ufności to średnia populacji zawierałaby się w ok. 95% z nich

6.2.2 Jak konstruować przedziały ufności?

Założmy, że konstruujemy przedział ufności dla jakiejś zmiennej ciągłej. Teoretycznie jest możliwe uzyskanie wielu różnych przedziałów ufności np. znaleźć takie a , że

$$P(\theta \in (-\infty, a)) = 1 - \alpha$$

albo

$$P(\theta \in (a, \infty)) = 1 - \alpha$$

albo takie a i b , że

$$P(\theta \in (a, b)) = 1 - \alpha$$

Chociaż jest możliwe konstruowanie różnych przedziałów ufności, najczęściej konstruujemy przedziały symetryczne (w sensie prawdopodobieństwa ;). To jest konstruujemy je w taki sposób, aby prawdopodobieństwo, że prawdziwa wartość parametru będzie poza przedziałem po lewej stronie (parametr będzie mniejszy niż lewy koniec przedziału) było takie samo jak prawdopodobieństwo, że wartość parametru będzie po prawej stronie przedziału. Czyli dla poniższych oznaczeń:

$$P(\theta \in C_n) = 1 - \alpha$$

$$P(\theta \notin C_n) = \alpha$$

$$P(\theta \notin C_n) = P(\theta \notin (a, b)) = P(\theta < a) + P(\theta > b) = \alpha$$

chcemy dobrać takie a i b aby $P(\theta < a) = P(\theta > b)$. Wynika z tego, że chcemy, aby oba te prawdopodobieństwa były równe $\frac{\alpha}{2}$.

Dodatkowym atutem takiego przedziału jest to, że jego szerokość (zakres wartości) dla zmiennej o rozkładzie normalnym jest najmniejsza wśród wszystkich przedziałów o takiej ufności. No właśnie, przejdźmy do konkretów: jak skonstruować taki przedział dla nieobciążonego estymatora o rozkładzie normalnym?

6.2.3 Przedziały ufności dla nieobciążonych estymatorów normalnych

Żałujemy, że mamy nieobciążony estymator pewnego parametru $\hat{\theta}$, który ma rozkład normalny z pewną prawdziwą średnią θ (jest to estymator nieobciążony) i pewnym odchyleniem standardowym $\mathbb{D}\hat{\theta}$ (w kontekście estymacji czasami zwanym błędem standardowym).

W takiej sytuacji zmienna $Z(\hat{\theta}) = \frac{\hat{\theta} - \theta}{\mathbb{D}\hat{\theta}}$ ma rozkład standardowy normalny. Dla jakiego przedziału $P(a < Z < b) = 1 - \alpha$? Jak uzasadniliśmy wyżej, chcielibyśmy uzyskać przedział symetryczny czyli chcemy, żeby $P(Z < a)$ było równe $P(Z > b)$, a dodatkowo z definicji przedziału ufności wiemy, że suma tych prawdopodobieństw powinna wynosić α .

$$P(Z < a) + P(Z > b) = 2P(Z < a) = \alpha$$

$$P(Z < a) = \frac{\alpha}{2}$$

Z symetryczności rozkładu normalnego wiemy, że $a = -b$, a wartość a możemy prosto znaleźć w tablicach. Dla przykładu: 95% przedział ufności będzie miał $\alpha = 5\%$ czyli odczytujemy wartość dla $\frac{\alpha}{2} = 2,5\%$, która wynosi -1.96 ($P(Z < -1.96) = 0.025$). Tę wartość odczytaną z tabeli dalej będziemy oznaczali jako $z_{\alpha/2}$ ($z_{0.025} = -1.96$).

Skonstruowaliśmy więc przedział ufności dla Z :

$$P(z_{\alpha/2} < Z < -z_{\alpha/2}) = 1 - \alpha$$

Ale przecież nas interesuje przedział ufności dla θ - przestąpmy do destandaryzacji.

$$\begin{aligned} P(z_{\alpha/2} < Z < -z_{\alpha/2}) &= P(z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\mathbb{D}[\hat{\theta}]} < -z_{\alpha/2}) \\ &= P(z_{\alpha/2} \mathbb{D}[\hat{\theta}] < \hat{\theta} - \theta < -z_{\alpha/2} \mathbb{D}[\hat{\theta}]) \\ &= P(-\hat{\theta} + z_{\alpha/2} \mathbb{D}[\hat{\theta}] < -\theta < -\hat{\theta} - z_{\alpha/2} \mathbb{D}[\hat{\theta}]) \\ &= P(\hat{\theta} - z_{\alpha/2} \mathbb{D}[\hat{\theta}] > \theta > \hat{\theta} + z_{\alpha/2} \mathbb{D}[\hat{\theta}]) \text{ /pomnóż przez -1/} \\ &= P(\hat{\theta} + z_{\alpha/2} \mathbb{D}[\hat{\theta}] < \theta < \hat{\theta} - z_{\alpha/2} \mathbb{D}[\hat{\theta}]) \text{ /przepisz od prawej do lewej/} \end{aligned}$$

Uzyskaliśmy więc ogólny wzór na przedział ufności konstruowany dla parametru estymowanego estymatorem nieobciążonym o rozkładzie normalnym:

$$(\hat{\theta} + z_{\alpha/2} \mathbb{D}[\hat{\theta}], \hat{\theta} - z_{\alpha/2} \mathbb{D}[\hat{\theta}])$$

Zauważ, że jest to niezwykle użyteczny wzór dla estymatorów największej wiarygodności: są one asymptotycznie (dla dużych prób) nieobciążone i asymptotycznie normalne. Innym przykładem estymatora, który cały czas będziemy używać, jest średnia arytmetyczna, która:

- dla próby z rozkładu normalnego zawsze ma rozkład normalny,
- dla dużych prób z dowolnego rozkładu (na mocy CLT) ma rozkład normalny

Definicja 6.4 Przy rozkładzie normalnym (lub przy dużej próbie), jeżeli wariancja jest znana to przedział ufności dla μ przy użyciu średniej arytmetycznej \bar{x} z $1 - \alpha$ poziomem ufności:

$$\left(\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

w skrócie:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

gdzie $z_{\alpha/2} = \Phi^{-1}(\frac{\alpha}{2})$. Wartość $z_{\alpha/2}$ nazywamy kwantylem.



Zauważ, że $z_{\alpha/2}$ ($\alpha/2$ zawsze jest mniejsza równa 0.5) będzie zawsze ujemny, więc w lewym końcu przedziału mamy *dodawanie*, a w prawym odejmowanie - co jest na pierwszy rzut oka nieintuicyjne!

Użyteczne wartości: dla 95% przedziału ufności $z_{0.025} = -1.65$, dla 99% przedziału ufności $z_{0.005} = -2.58$. Upewnij się, że potrafisz odczytać te wartości z tabeli rozkładu normalnego!

Problem 6.5 Jak będzie wyglądała powyższa definicja (6.4) dla sumy zmiennych losowych?

■ **Przykład 6.2** Pobrano próbę losową o licznosci $n = 16$ dla pewnej cechy o rozkładzie normalnym i znanej wariancji $\sigma^2 = 4$. Obliczono średnią arytmetyczną dla pobranej próby i uzyskano wynik $\bar{x} = 10$. Skonstruuj 90% przedział ufności.

Rozpocznijmy od ustalenia α : $1 - \alpha = 0,90$ czyli $\alpha = 0,10$.

Odczytajmy wartość $z_{\alpha/2}$ z tablic: $z_{0.05} = -1.65$. Podstawiamy do wzoru z definicji 6.4

$$\left(\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

i otrzymujemy⁷

$$\left(10 + (-1.65) \frac{2}{\sqrt{16}}, 10 - (-1.65) \frac{2}{\sqrt{16}} \right) = \left(10 - 1.65 \cdot \frac{1}{2}, 10 + 1.65 \cdot \frac{1}{2} \right) = (9.175, 10.825)$$

■

⁷Zwróć uwagę na pułapkę: w treści zadania podano wartość wariancji, a we wzorze mamy odchylenie standardowe

Problem 6.6 Wiemy, że X ma rozkład normalny $N(\mu, \sigma = 5)$. Skonstruuj 95% przedział ufności dla nieznanego parametru μ , mając do dyspozycji próbę losową z $n = 25$ obserwacjami, w której średnia wartość wyniosła 30.

Problem 6.7 Wiemy, że $X \sim N(\mu, \sigma = 20)$. Skonstruuj 99% przedział ufności dla nieznanego parametru μ , mając do dyspozycji próbę losową z $n = 100$ obserwacjami, w której średnia wartość wyniosła -6 .

Problem 6.8 Jak zmieni się przedział gdy zwiększę poziom ufności?

Problem 6.9 Algorytm rozpoznający automatycznie raka płuc na podstawie prześwietlenia podczas testu na 100 pacjentach popełnił 1 błąd. Lekarz przy ocenie prześwietlenia nie myli się w 98,4% przypadków. Podany wynik sugeruje, że algorytm jest lepszy niż lekarz. Skonstruuj przedział ufności, aby dowiedzieć się czy jest możliwe uzyskanie tak dobrego wyniku przez algorytm, nawet jeżeli w rzeczywistości jest on gorszy niż lekarz.

Problem 6.10 W przeprowadzanym badaniu statystycznym ($n = 100$) otrzymano szeroki przedział ufności dla wartości oczekiwanej. Aby uzyskać lepszą estymację chcielibyśmy go skrócić co najmniej dwukrotnie. O ile muszę zwiększyć rozmiar próby?

Ćwiczenie 6.3 Zbuduj przedziały ufności: <https://ophelia.cs.put.poznan.pl/webdav/ad/students/04/cw-2.xls>

6.3 Tworzenie własnych funkcji w R

Stworzenie własnej funkcji w **R** jest bardzo proste: wystarczy do nazwy funkcji przypisać jej kod rozpoczynając go od `function()`, a między nawiasami można umieścić nazwy argumentów. Dla przykładu mając funkcję obliczającą wariancję `var` możemy bardzo prosto zdefiniować funkcję obliczającą odchylenie standardowe:

```
sd <- function(x) sqrt(var(x))
```

Jeżeli chciał(a)byś napisać dłuższą funkcję, która nie zmieściłaby się w jednej linijce musiał(a)byś ująć ciało funkcji w nawiasy wąskie:

```
sd <- function(x){  
  sqrt(var(x))  
}
```

Zauważ, że w **R** nie ma konieczności używania odpowiednika słowa `return` z innych języków programowania. Po prostu ostatnia zewalutowana przez funkcję wartość jest traktowana jako jej wyjście. Natomiast, jeżeli `return` z jakiś powodów byłoby Ci potrzebne to możesz go użyć – przy czym w **R** jest ono funkcją: np. `return()` zwraca wartość `NULL`.

```
sd <- function(x) return(sqrt(var(x)))
```

Jeżeli chcesz zmienić już napisaną funkcję możesz wywołać funkcję `fix(nazwa_funkcji)`, która otworzy edytor z kodem źródłowym funkcji.


Ćwiczenie 6.4 Stwórz funkcję zwracającą przedział ufności dla średniej arytmetycznej dla danego w jej argumencie wektora. Przedział możesz zwrócić jako wektor dwóch liczb. ■

Oczywiście w  istnieją również konstrukcje warunkowe:

```
if (wyrazenie_logiczne) wyrazenie1 else wyrazenie2
```

Jeżeli chciałbyś użyć kilku wyrażeń w składni instrukcji warunkowej to musiał(a)byś, podobnie jak w przypadku funkcji, umieścić je w nawiasie wąsistym.

Ćwiczenie 6.5 Rozszerz implementację funkcji z poprzedniego ćwiczenia, tak aby przy małym rozmiarze wektora konstruowała ona przedział ufności używając kwantyli rozkładu t-Studenta. ■


W  istnieją również pętle np.

```
while (wyrazenie_logiczne) wyrazenie
```

```
for (nazwa in wyrazenie1) wyrazenie2
```



Pętla `for` jest zdefiniowana jako pętla typu `foreach` iterująca po wszystkich elementach sekwencji. Każdy kolejny element sekwencji (`wyrazenie1`) jest podstawiany do zmiennej `nazwa`. Z tego powodu, aby uzyskać iterowanie po liczbach całkowitych trzeba iterować po wynikach znanej Ci funkcji `seq()` lub użyć konstrukcji `od:do`.

Ćwiczenie 6.6 Wypisz liczby od 1 do 10 używając pętli `for` i funkcji `print()`. ■

 Spróbujmy sprawdzić działanie Centralnego Twierdzenia Granicznego dla znanego Tobie rozkładu χ^2 z $df = 1$. Aby zasymulować Centralne Twierdzenie Graniczne musimy najpierw pobrać bardzo wiele próbek z rozkładu populacji, następnie policzyć średnią na każdej z nich i stworzyć z nich rozkład.

1. Na początku utwórzmy sobie wektor w którym będziemy zbierać wyniki średnich policzonych na każdej z utworzonych przez nas prób. Załóżmy, że tych próbek będzie 1000. Aby nie wpisywać ręcznie tysiąca zer do wektora użyjemy funkcji `rep(element, ilosc_powtorzen)`, która tworzy wektor poprzez wielokrotne powtórzenie danego elementu.

```
sample_means <- rep(0, 1000)
```

 Aby zrobić to bardziej elegancko w języku  moglibyśmy zainicjalizować wektor wartościami `NA` („Not Available”) oznaczające wartość która jest brakująca/niedostępna. Alternatywnie można użyć konstruktora wektora wartości numerycznych `numeric(dlugosc_wektora)`.

2. Teraz napiszmy pętlę, która wygeneruje nam próbkę o pewnym rozmiarze z

rozkładu χ^2 , policzy z niej średnią i wpisze do przygotowanego wcześniej wektora `sample_means`. Ustalmy wielkość próbki na 30.

```
for(i in 1:iter){
  samp <- rchisq(30,df=1)
  sample_means[i] <- mean(samp)
}
```

3. Gotowe! Narysuj teraz histogram otrzymanych średnich oraz narysuj linie rozkładu normalnego, aby móc łatwo sprawdzić czy uzyskany rozkład jest rozkładem normalnym.

Przypomnienie: jeżeli $X \sim \chi_k^2$ to $\mathbb{E}[X] = k$, a $\mathbb{D}[X] = \sqrt{2k}$, gdzie k to liczba stopni swobody rozkładu.

4. Stwórz z kodu z poprzednich punktów tego ćwiczenia funkcję `plot_sampling_distr` z parametrami `n` licznosc generowanych próbek oraz `iter` liczba generowanych próbek, aby móc wywoływać ten kod wielokrotnie w wygodny sposób.
5. Wywołaj funkcję wielokrotnie z różnymi ustawieniami parametrów i sprawdź od jakich wartości argumentów funkcji rozkład średnich z próbek staje się normalny.

Ćwiczenie 6.7 Napisz funkcję, która dla wektora $(x_1, x_2, x_3, \dots, x_n)$ zwraca wektor $(x_1^1, x_2^2, x_3^3, \dots, x_n^n)$. ■

Ćwiczenie 6.8 Zaimplementuj poniższą ciągłą funkcję

$$f(x) = \begin{cases} x^2 + 2x + 3 & x < 0 \\ x + 3 & 0 \leq x < 2 \\ x^2 + 4x - 7 & 2 \leq x \end{cases}$$

i narysuj jej wykres od -3 do 3. ■

Ćwiczenie 6.9 Możesz sprawdzić które funkcje są rzeczywiście napisane w \mathbb{R} , a które wołają skompilowany kod napisany np. w C poprzez wpisanie jej nazwy bez nawiasów. Spróbuj dla funkcji obliczającej średnią `mean` i wariancję `var`. ■


Alternatywą do pisania pętli jest użycie funkcji `apply`, która aplikuje podaną funkcję do każdego wiersza lub kolumny ramki danych.


```
apply(ramka_danych, wymiar_do_iteracji, funkcja)
```


Na przykład poniższe wywołanie funkcji oblicza średnią dla każdej kolumny ramki `dataset`.

```
apply(dataset, 2, mean)
```


6.3.1 Wczytywanie danych do R

W powyższym tutorialu dane były już wczytane za Ciebie lub tworzyłeś je samodzielnie, poprzez wpisanie ich do konsoli. Dlatego na koniec warto wspomnieć o tym, że  posiada bardzo proste funkcje do importu danych z wielu pakietów statystycznych jak SPSS, SAS, Stata, a także z baz danych zarówno SQL (np. MySQL, Oracle) jak i NoSQL (np. MongoDB) oraz z innych źródeł np. z Twitter'a.

Jeżeli chcesz zaimportować plik w formacie tekstowym do  najprościej użyć do tego celu importera wbudowanego w RStudio. W prawym górnym okienku powinieneś odnaleźć przycisk „Import Dataset”, który uruchomi kreator proszący Cię o wskazanie pliku na dysku twardym, a następnie spróbuje automatycznie wykryć zasady formatowania pliku (np. jaki znak jest separatorem).

Oczywiście to samo możesz uzyskać poprzez wywołanie odpowiedniego polecenia w konsoli języka . W przypadku pliku tekstowego jest to `read.table` z odpowiednimi parametrami. W parametrach tej funkcji należy wyspecyfikować m.in. ścieżkę do pliku, informację czy pierwsza linijka zawiera nazwy kolumn (`header`) oraz separator poszczególnych pól (`sep`). Przykładowe wywołanie znajdziesz poniżej.

```
mydata <- read.table("c:/mydata.csv", header=TRUE, sep=",")
```

Z chwilą gdy ramkę danych (lub inny obiekt) masz już wczytaną do  możesz ją prosto zapisać do formatu RData:

```
save(obiekt, file="sciezka_do_pliku")
```

Taki plik możesz następnie prosto wczytać, aby w obszarze roboczym pojawiły się zapisane obiekty bez konieczności podawania separatorów, informacji o nagłówkach itd.

```
load("sciezka_do_pliku")
```

Literatura

Literatura powtórkowa

Opis estymatorów przedziałowych oraz podstawy testowania hipotez statystycznych można znaleźć w podrozdziale 2.3, 3.1, 3.2 książki [2], a także w 4 rozdziale dostępnej online książki [1].

Literatura dla chętnych

Dla chętnych polecam krótki filmik o testowaniu statystycznym <https://www.youtube.com/watch?v=ySPm3boe04c> oraz o bardzo ciekawym problemie, który przysporzył wielu statystykom bólu głowy <https://www.youtube.com/watch?v=mh1c7peG1Gg>.

Pytania sprawdzające zrozumienie

Pytanie 6.1 Czy rozumiesz i znasz Centralne Twierdzenie Graniczne (Lindeberga-Levy'ego)?

Pytanie 6.2 Rozwiąż zadania korzystając z Centralnego Twierdzenia Granicznego (takie jak przykład 5.1).

Pytanie 6.3 Skonstruuj przedział ufności dla podanej miary (sumy lub średniej). Czy wiesz co oznacza poziom ufności przedziału i umiesz go zinterpretować?

Pytanie 6.4 Jaki rozkład powstaje poprzez sumę dwóch zmiennych losowych o rozkładzie dwupunktowym?

Bibliografia

- [1] D.M. Diez, C.D. Barr, i M. Çetinkaya Rundel. *OpenIntro Statistics: Third Edition*. OpenIntro, Inc., 2015. ISBN 194345003X. URL openintro.org.
- [2] W. Kryszicki, J. Bartos, W. Dyczka, K. Królikowska, i M. Wasilewski. *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach. Część II: Statystyka matematyczna*. Wydawnictwo Naukowe PWN, 2002. ISBN 8301113847.
- [3] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer-Verlag New York, 2004.