

III. Wnioskowanie statystyczne -wykład

Struktura

- 3.1. Podstawy rachunku prawdopodobieństwa, prawdopodobieństwo i zmienna losowa
- 3.2. Podstawy metody reprezentacyjnej - próba, populacja, podstawowe schematy losowania
- 3.3. Wybrane rozkłady zmiennych losowych
- 3.4. Prawa wielkich liczb, twierdzenia graniczne , statystyki z próby i ich rozkłady
- 3.5. Wnioskowanie statystyczne - podejście klasyczne – estymacja, metody pozyskiwania estymatorów
- 3.6. Wnioskowanie statystyczne - podejście klasyczne – testowanie hipotez
- 3.7. Przykłady empiryczne

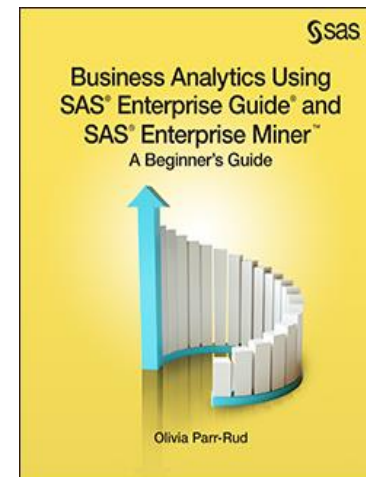
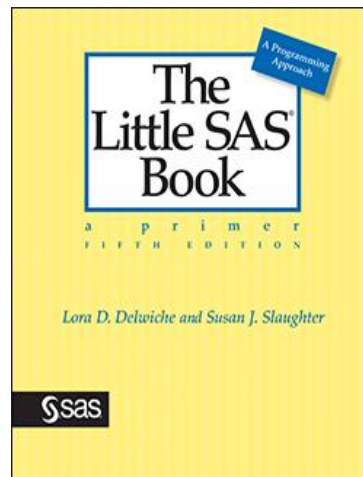
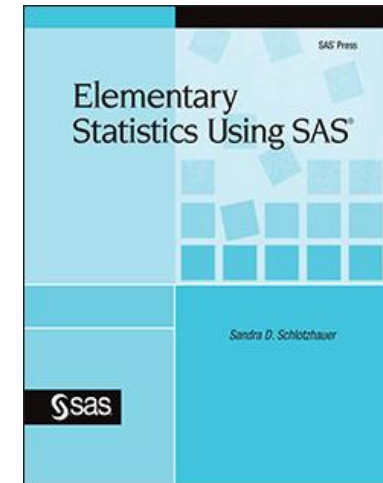
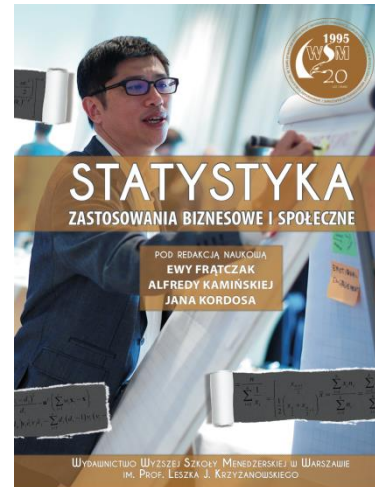
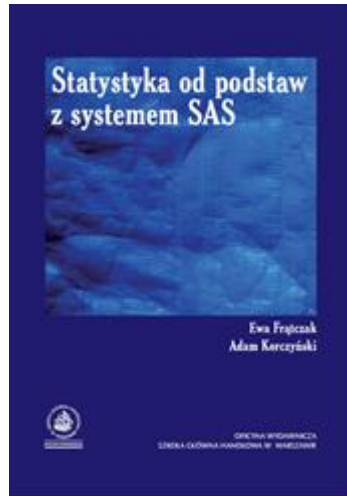
Literatura:

E. Frątczak, A. Korczyński, *Statystyka od podstaw z systemem SAS*, SGH, Warszawa 2013.

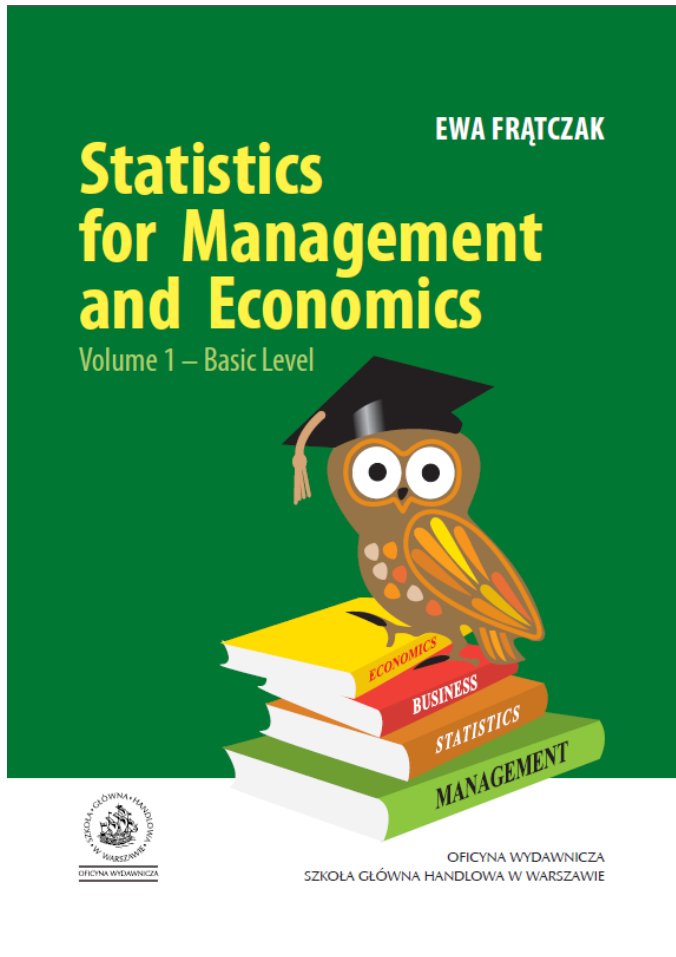
J.B.Davis, *Statistics Using SAS. Enterprise Guide*, SAS Publishing, SAS Institute Inc., Cary, NC 2007.

G. Keller., *Managerial Statistics*, 9th international ed., South-Western CENGAGE Learning, UK, USA 2012.

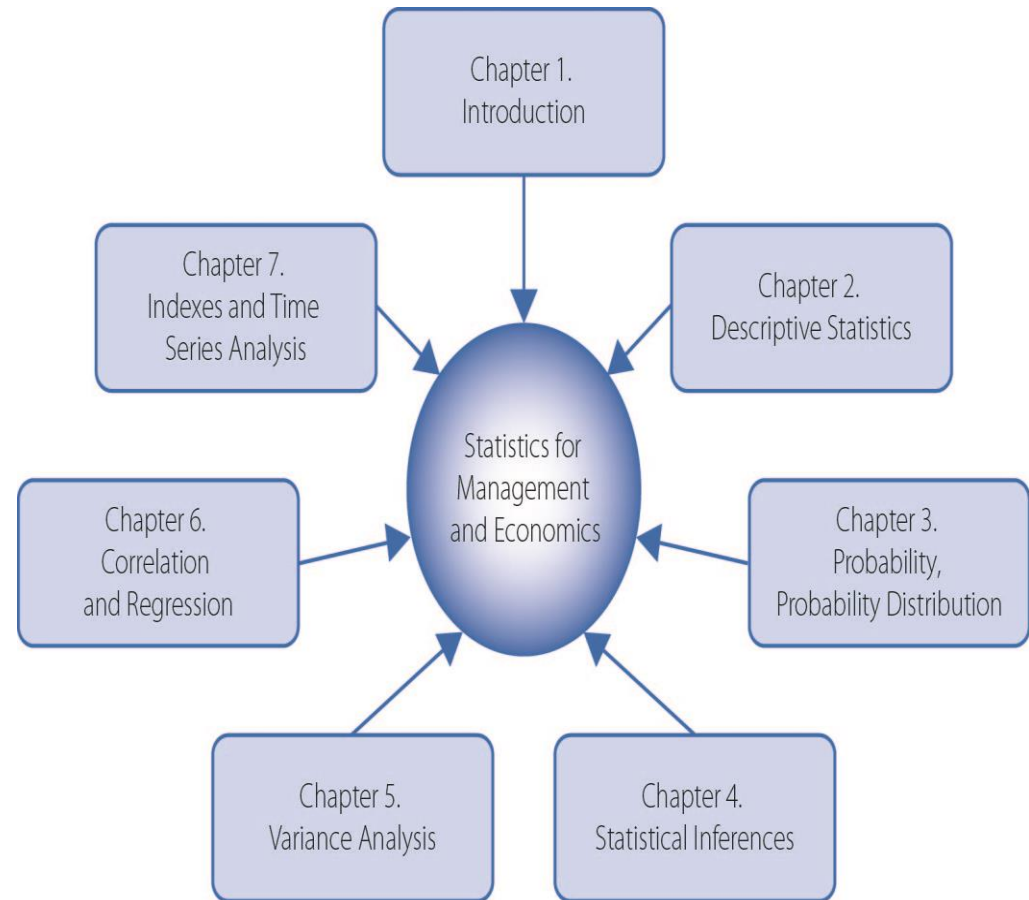
Literatura c.d.



Literatura c.d.



Picture of the book "Statistics for Management and Economics"



3.1. Podstawy rachunku prawdopodobieństwa, prawdopodobieństwo - zmienna losowa

Wnioskowanie statystyczne, to jest uogólnianie wyników z próby losowej na populację generalną, bazuje na elementarnych podstawach rachunku prawdopodobieństwa. Stąd na początku zostaną podane, w olbrzymim skrócie definicje najważniejszych pojęć dotyczących zdarzeń losowych i prawdopodobieństwa. **Pojęcia te to: doświadczenie losowe, zdarzenia elementarne, zdarzenia losowe, prawdopodobieństwo zdarzenia losowego.**

Doświadczenie losowe jest doświadczeniem, którego wyniku nie można przewidzieć przed wykonaniem doświadczenia. Można podać wiele przykładów takich doświadczeń w życiu codziennym, w przyrodzie i technice. Wynikami doświadczeń losowych mogą być:

- wynik rzutu monetą, liczba wyrzuconych oczek w rzucie kostką,
- liczba klientów w danym sklepie w określonym dniu,
- liczba wypadków samochodowych w woj. Mazowieckim w ciągu miesiąca, itp..

Innymi słowy: **doświadczeniem losowym jest zjawisko, w którym istnieje jakakolwiek niepewność.**

Zdarzenia elementarne - dla każdego doświadczenia losowego można wyróżnić zdarzenia, za pomocą których możemy opisać wszystkie wyniki tego doświadczenia. Te zdarzenia nazywamy zdarzeniami elementarnymi. Tworzą one zbiór zdarzeń elementarnych.

Wybór zbioru zdarzeń elementarnych jest umowny i decyduje o charakterze doświadczenia losowego.

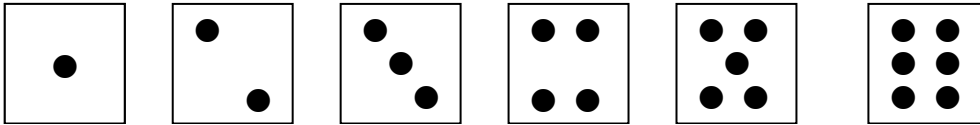
Na przykład w doświadczeniu losowym polegającym na jednokrotnym rzucie monetą przyjmujemy dwa zdarzenia elementarne:

- wypadł orzeł (O),
- wypadła reszka (R).

Zdarzenia losowe są to podzbiory zbioru zdarzeń elementarnych. Będziemy oznaczać je dużymi literami alfabetu (ewentualnie z subskryptami) np. A, B .

Przykład

Dla rzutu kostką zbiór zdarzeń elementarnych składa się z sześciu elementów:



Będziemy je oznaczać symbolami: 1, 2, 3, 4, 5, 6. Formalnie, zdarzeniami nie są liczby, ale poszczególne ścianki kostki.

Przykładami zdarzeń losowych są:

A - wypadła parzysta liczba oczek $A = \{2, 4, 6\}$,

B - wypadła nieparzysta liczba oczek $B = \{1, 3, 5\}$,

C - wypadła liczba oczek większa niż 4 $C = \{5, 6\}$,

D - wypadła liczba oczek mniejsza niż 2 $D = \{1\}$,

E - wypadła liczba oczek mniejsza niż 0 $E = (\text{zbiór pusty})$,

F - wypadła liczba oczek mniejsza niż 7 $F = \{1, 2, 3, 4, 5, 6\}$.

Zdarzenia elementarne wchodzące w skład danego zdarzenia losowego A (zbioru) nazywamy *zdarzeniami sprzyjającymi* zdarzeniu A. W podanym przykładzie zdarzeniami sprzyjającymi zdarzeniu $C = \{5, 6\}$ są:

5 (wypadnięcie pięciu oczek) i 6 (wypadnięcie sześciu oczek).

Zdarzenie F = nazywamy *zdarzeniem pewnym*. Zdarzenie E = nazywamy *zdarzeniem niemożliwym*.

Ponieważ zdarzenia losowe są utożsamiane z pewnymi podzbiorami zbioru zdarzeń elementarnych, istnieje odpowiedniość działań na zdarzeniach i działań na zbiorach.

Podstawowe działania to:

$A \cup B$ - alternatywa zdarzeń (zaszło zdarzenie A lub zdarzenie B)

$A \cap B$ - koniunkcja zdarzeń (zaszło jednocześnie zdarzenie A i zdarzenie B)

A' - zdarzenie przeciwne do A (nie zaszło zdarzenie A)

$A \setminus B$ - różnica zdarzeń (zaszło zdarzenie A i nie zaszło zdarzenie B)

$A \subset B$ - implikacja (jeżeli zaszło zdarzenie A , to na pewno zaszło zdarzenie B)

Bazując na danych z podanego wcześniej przykładu możemy zapisać następujące działania:

$$A \cup B = \{1, 2, 3, 4, 5, 6\} = \Omega$$

$$A \cup C = \{2, 4, 5, 6\}$$

$$A \cap B = \emptyset$$

$$C \cap D = \emptyset$$

$$B \cap C = \{5\}$$

$$A' = \{1, 3, 5\} = B$$


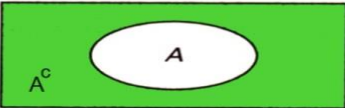
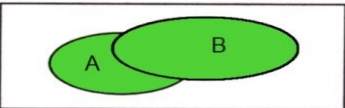
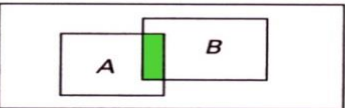
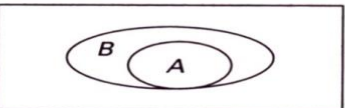
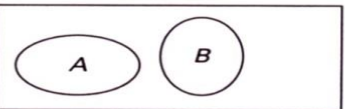
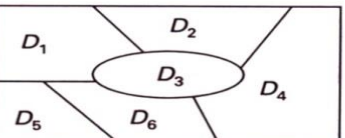
$$C' = \{1, 2, 3, 4\}$$

$$B \setminus C = \{1, 3\}$$

$$D \subset B$$

Mówimy, że zdarzenia A i B *wyłączają się* (są wykluczające, rozłączne), jeżeli $A \cap B = \emptyset$ (ich koniunkcja jest zdarzeniem niemożliwym).

Venn Diagram (J.Venn, 1834 – 1923) –istota rachunku działań na zdarzeniach

Concept	Notation	Venn diagram	Meaning for events
Empty set	\emptyset	-----	Cannot occur
Sample space	Ω		Occurs certainly
Complement	A^c		A does not occur
Union	$A \cup B$		At least one of the events A and B occurs
Intersection	$A \cap B$		Both A and B occur
Subset	$A \subset B$		If A occurs, then B occurs
Disjoint	$A \cap B = \emptyset$		A and B cannot occur jointly
Partition	D_1, \dots, D_s		Exactly one of the events D_1, \dots, D_s occurs

Podstawowe definicje prawdopodobieństwa

W literaturze przedmiotu istnieje kilka definicji prawdopodobieństwa. Poniżej zostaną przytoczone dwie: klasyczna (Laplace'a) i aksjomatyczna (Kołmogorowa). Reasumując dotychczasowe rozważania z zakresu podstawowych pojęć wprowadźmy następujące oznaczenia:

A, B, C, \dots lub A_1, A_2, \dots, A_n - zdarzenie (zdarzenia elementarne - podzbiory Ω)

Ω - przestrzeń wszystkich zdarzeń elementarnych

A' - zdarzenia przeciwne do zdarzeń ze zbioru A

Zdarzenie $F = \Omega$ - zdarzenie pewne

Zdarzenie $E = \emptyset$ - zdarzenie niemożliwe

$|\Omega|, |A|, |B| \dots$ - liczebność zbiorów $\Omega, A, B \dots$

Według klasycznej definicji sformułowanej przez Laplace'a prawdopodobieństwem P zdarzenia A nazywamy stosunek liczby zdarzeń elementarnych sprzyjających realizacji zdarzenia A do liczby wszystkich zdarzeń elementarnych jednakowo możliwych i wzajemnie wykluczających się.

$$P(A) = \frac{|A|}{|\Omega|}, \quad P(A) = p, \quad 0 \leq p \leq 1$$

Według aksjomatycznej definicji, prawdopodobieństwo definiuje się następująco:

1. Każdemu zdarzeniu losowemu A odpowiada określona liczba $P(A) = p$ zwana prawdopodobieństwem realizacji zdarzenia A , liczba p przyjmuje wartości z przedziału $(0;1)$: $0 \leq P(A) \leq 1$
2. Prawdopodobieństwo zdarzenia pewnego równa się 1, $P(F) = 1$.
3. Jeżeli $(A_1, A_2, A_3, \dots, A_n)$, jest ciągiem skończonym lub nieskończonym zdarzeń losowych parami wykluczających się to prawdopodobieństwo sumy tych zdarzeń jest równe sumie prawdopodobieństw tych zdarzeń.
4. $P(A_1 + A_2 + A_3 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$

Z aksjomatów wyprowadzono następujące wnioski:

- prawdopodobieństwo zdarzenia niemożliwego $= 0$, co zapisuje się $P(E) = 0$,
- prawdopodobieństwo zdarzenia A i przeciwnego do $A = 1$ co zapisuje się:
 $P(A) + P(A') = 1$, co po przekształceniach daje: $P(A') = 1 - P(A)$

ZMIENNA LOSOWA

Zmienna losowa jest odpowiednikiem cechy statystycznej w rozkładzie empirycznym. Odmiany cechy statystycznej nazywane są wariantami, natomiast w odniesieniu do zmiennej losowej używa się sformułowania -realizacja zmiennej losowej. Sformułowanie definicji zmiennej losowej opiera się na podstawowych pojęciach rachunku prawdopodobieństwa, takich jak: zdarzenie losowe, zdarzenie elementarne, pojęcie prawdopodobieństwa.

Przykład

Wykonujemy doświadczenie polegające na rzucie monetą, przy założeniu, że jest ona symetryczną. Zbiór możliwych zdarzeń elementarnych to: pojawienie się orła lub pojawienie się reszki, {orzeł, reszka}. Każdemu z tych zdarzeń elementarnych można przyporządkować liczbę rzeczywistą, np. wyrzuceniu orła liczbę 0, wyrzuceniu reszki liczbę 1. Przy założeniu, że moneta jest rzetelna, prawdopodobieństwo wyrzucenia reszki równa się prawdopodobieństwu wyrzucenia orła = 0,5.

Generalnie zmienną losową nazywać będziemy każdą funkcję rzeczywistą określoną na zbiorze zdarzeń elementarnych. Zmienne losowe mogą być typu skokowego lub ciągłego.

Zmienna losowa skokowa

Zmienna losowa X jest typu skokowego, jeżeli może przyjmować skończoną lub nieskończoną, ale przeliczalną liczbę wartości.

Niech zmienna losowa X typu skokowego przyjmuje wartości: x_1, x_2, \dots z prawdopodobieństwem : p_1, p_2, \dots

Zbiór prawdopodobieństw realizacji zmiennej losowej postaci:

$$P(X = x_i) = p_i \quad \text{nosi nazwę funkcji rozkładu prawdopodobieństwa zmiennej losowej typu skokowego.}$$

Funkcja ta spełnia następujące warunki:

1. $\sum_{i=1}^k p_i = 1$ - gdy zmienna losowa X przyjmuje skończoną liczbę k wartości,
2. $\sum_{i=1}^{\infty} p_i = 1$ - gdy zmienna losowa X przyjmuje nieskończoną liczbę wartości.

Warto podkreślić, że argumentami funkcji prawdopodobieństwa są wartości zmiennej losowej, natomiast wartościami tej funkcji prawdopodobieństwa realizacji poszczególnych wartości zmiennej losowej.

Jeśli zbiór wartości zmiennej losowej jest zbiorem skończonym funkcje można zapisać następująco:

$X=x_i$	x_1	x_2	x_3	...	x_k
$P(X=x_i)$	p_1	p_2	p_3	...	p_k

Drugą charakterystyką rozkładu zmiennej losowej skokowej jest dystrybuanta, określana symbolem $F(x)$, gdzie:

$$F(X) = \sum_{x_i \leq x} P(X = x_i) = P(X \leq x_i) = \sum_{x_i \leq x} p_i \quad -\infty < X < \infty$$

Dystrybuanta zmiennej losowej spełnia następujące warunki:

1. $0 \leq F(x) \leq 1$ $\lim_{x \rightarrow -\infty} F(x) = 0$ lub $F(-\infty) = 0$
 $\lim_{x \rightarrow \infty} F(x) = 1$ lub $F(\infty) = 1$
2. $F(x)$ jest funkcją niemalejącą, przedziałami stałą, tzn. jeśli $x_1 < x_2$, to $F(x_1) \leq F(x_2)$
3. $F(x)$ jest funkcją prawostronnie ciągłą, ma skończoną lub przeliczalną ilość punktów nieciągłości. Jeżeli przyjmiemy, że zbiór wartości zmiennej losowej jest skończony i został uporządkowany według wzrastających wartości, to dystrybuantę możemy zapisać następująco:

$$F(X) = \begin{cases} 0 & \text{dla } x \leq x_1 \\ p_1 & \text{dla } x_1 \leq x < x_2 \\ p_1 + p_2 & \text{dla } x_2 \leq x < x_3 \\ \dots & \dots \\ 1 & \text{dla } x \geq x_k \end{cases}$$

PARAMETRY ROZKŁADU ZMIENNEJ LOSOWEJ SKOKOWEJ

Rozkład zmiennej losowej skokowej jest określony, gdy podana jest funkcja rozkładu prawdopodobieństwa lub dystrybuanta. Podstawowymi parametrami rozkładu są : wartość oczekiwana - $E(X)$ i wariancja – $D^2(X)$. Wyznacza się je według następujących formuł:

$$E(X) = \sum_{i=1}^k x_i p_i$$

lub

$$E(X) = \sum_{i=1}^{\infty} x_i p_i$$

$$D^2(X) = \sum_{i=1}^k [x_i - E(X)]^2 p_i$$

$$D^2(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 p_i$$

$$D(X) = \sqrt{D^2(X)}$$

Wariancja charakteryzuje stopień rozproszenia (rozrzutu) wartości zmiennej losowej wokół wartości oczekiwanej.

Zmienna losowa ciągła

Zmienną losową X będziemy określać jako zmienną ciągłą, jeśli może przyjmować wartości z pewnego przedziału liczbowego, co oznacza, że zbiór zdarzeń elementarnych, na którym określono taką zmienną jest nieskończony, nieprzeliczalny. Opis rozkładu zmiennej losowej ciągłej przebiega odmiennie niż rozkładu zmiennej losowej skokowej, a podstawowe znaczenie ma funkcja gęstości.

Funkcje gęstości $f(x)$ zmiennej losowej ciągłej definiuje się następująco:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x}$$

i interpretuje się ją jako średnią liczbę prawdopodobieństwa przypadającą na jednostkę długości przedziału $(x, x + \Delta x)$, przy założeniu, że rozpiętość tego przedziału dąży do 0.

Znając $f(x)$ można obliczyć przybliżoną wartość prawdopodobieństwa $P(x < X \leq x + \Delta x)$ przy dostatecznie małym Δx

$$P(x < X \leq x + \Delta x) \cong f(x)\Delta x$$

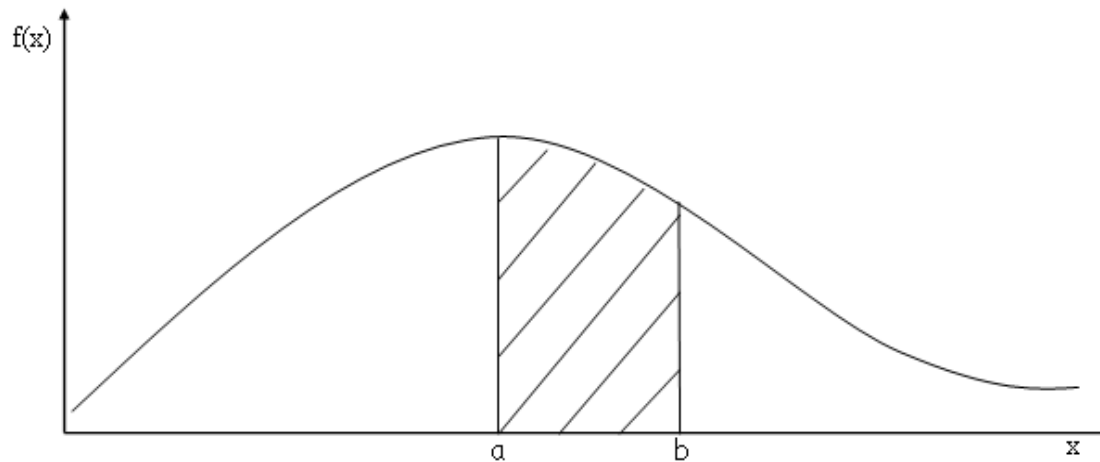
Funkcja gęstości posiada następujące własności:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X < \infty) = 1$

Istota opisu rozkładu zmiennej losowej X typu ciągłego przy pomocy funkcji gęstości $f(x)$ polega na tym, że prawdopodobieństwo realizacji tej zmiennej losowej w dowolnym przedziale (a, b) , gdzie $a < b$ można przedstawić przy pomocy całki oznaczonej z funkcji gęstości $f(x)$ w granicach całkowania (a, b) .

Graficzną interpretacją tego prawdopodobieństwa jest pole obszaru ograniczonego wykresem funkcji $f(x)$, osią odciętych i prostymi $x = a$ i $x = b$.

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$



Jeśli chcemy określić $P(X = a) = P(a \leq x \leq a) = \int_a^a f(x) dx = 0$

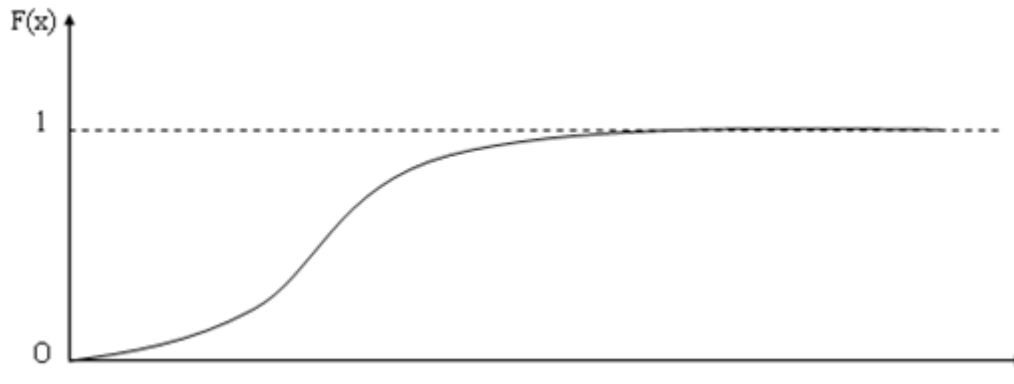
Kolejną charakterystyką rozkładu zmiennej losowej skokowej jest dystrybuanta. Dystrybuantę zmiennej losowej X typu ciągłego określa się następująco:

$$F(x) = \int_{-\infty}^x f(t) dt, \quad \text{gdzie } f(t) \text{ jest funkcją gęstości zmiennej losowej } X.$$

Dystrybuanta zmiennej losowej X posiada następujące własności:

1. $F(x)$ jest funkcją niemalejącą i ciągłą.
2. $0 \leq F(x) \leq 1$,przy czym $F(-\infty) = 0$
3. daje się przedstawić jako całka z funkcji gęstości $F(x) = \int_{-\infty}^x f(t) dt$

Przykładowy wykres dystrybuanty zmiennej losowej typu ciągłego:



Jeżeli wyznaczamy prawdopodobieństwo zmiennej losowej w przedziale

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx = F(b) - F(a)$$

PARAMETRY ROZKŁADU ZMIENNEJ LOSOWEJ CIĄGŁEJ

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

gdzie:

$f(x)$ – funkcja gęstości

$$D^2(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)dx$$

dx – pochodna tej funkcji

$$D(X) = \sqrt{D^2(X)}$$

Zestawienie podstawowych mierników dla rozkładu empirycznego (próby) i parametrów dla populacji generalnej

próba

populacja

w_i

p_i

\bar{X}

$E(X) ; m$

$S^2(x) D^2(X);$

σ^2

$S(X)$

$D(X) ; \sigma$

$G(x_i)$

$F(x_i)$

n

N

PODSTAWOWE WŁASNOŚCI ŚREDNIEJ I WARIANCJI

Średnia (nadzieja matematyczna) zmiennej losowej $E(X)$ i wariancja $D^2(X)$ posiadają określone własności.

WŁASNOŚCI $E(X)$ – WARTOŚCI OCZEKIWANEJ:

1. Wartość średnia stałej C jest równa stałej, co można zapisać:

$$E(C) = C$$

Stałą C można traktować jako zmienną losową przyjmującą wartość C z prawdopodobieństwem równym 1.

$$P(X = C) = 1, \text{ stąd } E(C) = 1 \cdot C = C.$$

2. Wartość średnia iloczynu zmiennej losowej X i stałej C równa się iloczynowi tej stałej i wartości oczekiwanej zmiennej losowej, co można zapisać:

$$E(CX) = C E(X),$$

$$E(CX) = \sum_{i=1}^k C x_i p_i = C \sum_{i=1}^k x_i p_i = C E(X)$$

3. Wartość oczekiwana sumy zmiennych losowych X_1 i X_2 jest równa sumie wartości oczekiwanych tych zmiennych, co można zapisać:

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

4. Jeśli X_1 i X_2 są niezależnymi zmiennymi losowymi to wartość średnia iloczynu tych zmiennych równa się iloczynowi wartości średnich tych zmiennych co można zapisać:

$$E(X_1 \cdot X_2) = E(X_1) \cdot E(X_2)$$

5. Wartość oczekiwana sumy zmiennej losowej i stałej równa się wartości oczekiwanej zmiennej plus stała, co można zapisać:

$$E(X+C) = E(X) + C$$

WŁASNOŚCI WARIANCJI - $D^2(X)$

1. **Wariancja stałej** C równa jest 0, co można zapisać:

$$D^2(C) = 0$$

dowód: $D^2(C) = (C - C)^2 \cdot 1 = 0$

2. **Wariancja iloczynu** stałej C i zmiennej losowej jest równa iloczynowi wartości wariancji zmiennej losowej i kwadratu stałej, co można zapisać:

$$D^2(CX) = C^2 \cdot D^2(X)$$

3. **Wariancja sumy** niezależnych zmiennych losowych X_1 i X_2 jest równa sumie wariancji zmiennych losowych, co można zapisać:

$$D^2(X_1 + X_2) = D^2(X_1) + D^2(X_2)$$

4. **Wariancja** da się przedstawić za pomocą momentów:

$$D^2(X) = E(X^2) - [E(X)]^2, D(X) = \sqrt{D^2(X)}$$

3.2. Podstawy metody reprezentacyjnej - próba, populacja, podstawowe schematy losowania

Podstawą wnioskowania statystycznego są wyniki badań statystycznych oparte na próbach losowych. Podstawowe znaczenie dla teorii wnioskowania statystycznego (estymacji i weryfikacji hipotez statystycznych) posiada koncepcja próby losowej i statystyki z próby.

Jeśli rozpatruje się zmienną losową, która w populacji generalnej ma rozkład dany dystrybuantą $F(x)$, to n -elementową próbę losową zapiszemy jako ciąg n -niezależnych zmiennych losowych (X_1, X_2, \dots, X_n) , z których każda ma taki sam rozkład dany dystrybuantą $F(x)$.

Jeśli dokonuje się n -krotnych obserwacji, to otrzymamy konkretną realizację próby losowej, którą można zapisać jako: (x_1, x_2, \dots, x_n) , gdzie x_1, x_2, \dots, x_n są wartościami odpowiadającymi poszczególnym elementom próby. W obu wypadkach używa się pojęcia próba.

Próba losowa (wyniki próby losowej) jest podstawą do wnioskowania o parametrach bądź rozkładzie zmiennej losowej w populacji generalnej.

Z reguły wnioskowanie statystyczne wymaga, aby próba była pobrana metodą ze zwracaniem tzn. wylosowana w sposób niezależny.

Metoda reprezentacyjna jest działem statystyki matematycznej którego przedmiotem są zagadnienia doboru próby losowej ze zbiorowości generalnej oraz metody uogólniania wyników badania próbnego na całą populację.

Sposób losowego doboru próby określa tzw. schemat losowania.

Natomiast szacowaniem wartości nieznanymi parametrów populacji na podstawie wyników uzyskanych z próby zajmuje się teoria estymacji statystycznej.

Optymalna wielkość próby: próba spełniająca dwa warunki:

1. Minimalizuje koszt badania dla określonej precyzji
2. Dostarcza ocen o maksymalnej precyzji przy ustalonym koszcie badania

Dobór jednostek do próby:

- losowy
- celowy (nielosowy), np. typowy, kwotowy

Kolejne określenia to: jednostka losowania i jednostka badania, liczebność próby, rodzaj błędów: losowe i nielosowe, plan losowania, itd.

Plan losowania (wyboru) próby obejmuje:

1. określenie i wybór operatu losowania próby
2. określenie i opracowanie schematu losowania próby
3. określenie minimalnej liczebności próby

Operat losowania:

1. Operat losowania to wykaz jednostek badania bądź ich zespołów zwanych jednostkami losowania. Jeśli wszystkie jednostki losowania są jednocześnie jednostkami badania, wówczas losowanie jest indywidualne. W przeciwnym przypadku losowanie nazywa się zespołowym.
2. W losowaniu jednostopniowym stosuje się jeden operat losowania, a w przypadku losowania wielostopniowego - wiele operatów losowania (na każdym stopniu operat losowania ograniczamy do jednostek wylosowanych do próby poprzedniego stopnia).
3. Aby zminimalizować błędy nielosowe operat losowania winien charakteryzować się następującymi właściwościami:
 - odpowiedniość dla celów badania,
 - aktualność,
 - kompletność,
 - identyfikowalność
 - jednokrotność występowania jednostek losowania,
 - jednoznaczna przynależność do określonej subpopulacji.
4. Systematyczna aktualizacja operatu losowania jest niezbędna, gdyż operaty losowania oparte są zazwyczaj na wynikach spisów (lub rejestracji) odzwierciedlających pewien przeszły stan faktyczny, który nie zawsze musi się pokrywać ze stanem obecnym.
5. Operat losowania powinien również obejmować wszystkie jednostki badanej populacji, przy czym każda jednostka badania powinna w nim figurować tylko jeden raz.

6. Operatami losowania mogą być wyniki rejestracji, spisów, dokumenty z bieżącej ewidencji zjawisk gospodarczych, społecznych itd.
7. Operatem losowania mogą być też mapy lub szkice terenowe z zaznaczonymi granicami obszarów tworzących jednostki losowania danego stopnia.
8. Zamiast przeprowadzać losowanie rekordów z dużego zbioru zawierającego wszystkie zmienne, wygodnie jest skorzystać z małego zbioru zawierającego wykaz identyfikatorów rekordów oraz zmienne warstwujące do losowania.

Schemat losowania próby obejmuje: system losowania oraz schemat (sposób) losowania

Podstawowe schematy losowania próby

1. Losowanie proste
2. Losowanie warstwowe
3. Losowanie systematyczne
4. Losowanie 2 stopniowe , wielostopniowe
5. Inne schematy losowania

Losowanie proste: jednostki losowane są z całej populacji, która została odpowiednio ponumerowana (od 1 do N). Wyróżnia się dwa podstawowe rodzaje:

- **ze zwracaniem jednostek wylosowanych** (losowanie zwrotne), lub
- **bez zwracania jednostek wylosowanych** (bezzwrotne)

Jednostki zazwyczaj wybierane są przy wykorzystaniu jakiegoś procesu randomizacji (liczby losowe z tablic lub generowane przez komputer).

Zachowana jest zasada jednakowego prawdopodobieństwa wyboru dla wszystkich jednostek w czasie losowania.

Losowanie warstwowe – składa się z czterech etapów.

E.1. Cała populacja dzielona jest na oddzielne grupy, zwane warstwami,

E.2. Z każdej warstwy losowane są niezależne próbki

E.3. W każdej warstwie oddzielnie szacuje się parametry, średnią i frakcje, które potem są ważone odpowiednimi wagami,

E.4. Wariancje dla ocen obliczane są w warstwach w zależności od rodzaju losowania z wykorzystaniem właściwego systemu wag.

Losowanie warstwowe może być: proporcjonalne lub optymalne:

Losowanie warstwowe charakteryzuje się określonymi własnościami:

1/ warstwowanie **redukuje wariancję** dla danego układu mierzonego zarówno wielkością próbki jak i kosztu. Wariancje mogą być zredukowane zarówno przez wykorzystanie losowania proporcjonalnego jak i lokalizacji "optymalnej". Są to powody uzasadnione w teorii, lecz w praktyce inne trzy powody mogą okazać się ważniejsze;

2/ warstwowanie może być zastosowane ze względów **bezpieczeństwa, wygody, zabezpieczenia**, jakie może dać wybór losowy;

3/ warstwowanie ułatwia lokalizację w domenach żądanej liczebności próbki, często proporcjonalnie, lecz specjalnie, gdy wymagana jest lokalizacja nieproporcjonalna;

4/ warstwowanie **ułatwia wykorzystanie różnych metod i procedur** dla zróżnicowania frakcji losowania w próbce.

Losowanie systematyczne - ten schemat losowania charakteryzuje się:

1. Na początku obliczany jest „**interwał losowania**” jako po ustaleniu żądanej liczebności próbki n i liczebności populacji N . (przyjmuje się zwykle l jako liczbę całkowitą).

Wybór systematyczny z losowym początkiem:

2. Wybiera się losowy początek r (liczba losowa z przedziału od 1 do l), a następnie dodaje się interwał l do wyznaczenia liczb $r, r+l, r+2l$, itd., które tworzą próbę wynoszącą N/l elementów.

Ograniczenia

Losowanie systematyczne jest powszechnie używane jako alternatywa losowania prostego i proporcjonalnego. Jest prostsze do zastosowania i nadzoru niż procedury wyboru próbki. Istnieje jednak niebezpieczeństwo popełnienia błędu „cyklicznego”, tj. gdy uporządkowanie jednostek losowania na wykazie, z którego dokonywany jest wybór próby, związane jest z interwałem losowania (np. wybór systematyczny próby 10-procentowej z wykazu pracowników podzielonych na brygady 10-cio osobowe.).

Losowanie dwustopniowe – schemat losowania dwustopniowego polega na wylosowaniu k grup, gdzie każda grupa losowana jest z takim samym prawdopodobieństwem wyboru. Następnie z każdej k grup losowana jest bezzwrotna próba prosta.

3.3. Wybrane rozkłady zmiennych losowych.

WYBRANE ROZKŁADY ZMIENNEJ LOSOWEJ SKOKOWEJ

ROZKŁAD ZERO-JEDYNKOWY (0-1)

Zmienna losowa X ma rozkład 0-1, jeśli przyjmuje wartość 1 z prawdopodobieństwem p , gdzie $0 < p < 1$, oraz wartość 0 z prawdopodobieństwem $q = 1 - p$. Funkcje rozkładu prawdopodobieństwa można zapisać:

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

lub

x_i	0	1
$p(x_i)$	$1 - p$	p

$$F(X) = \begin{cases} 0 & \text{dla } x < 0 \\ 1 - p & \text{dla } 0 \leq x \leq 1 \\ 1 & \text{dla } x \geq 1 \end{cases}$$

$$E(X) = p$$

$$D^2(X) = pq$$

$$E(X) = \sum_{i=1}^2 x_i p_i = 1p + 0(1 - p) = p$$

$$D^2(X) = \sum_{i=1}^2 (x_i - E(X))^2 p_i = (0 - p)^2 (1 - p)^2 p + p(1 - p) = pq$$

ROZKŁAD DWUMIANOWY

Przykład

Rzucamy 3 razy monetą. Niech zmienną losową X będzie liczba wyrzuconych orłów. Zmienna losowa może przyjmować wartości: 0, 1, 2, 3. Przy trzech rzutach monetą możliwe do wystąpienia sytuacje to:

Sytuacja	Liczba orłów	Prawdopodobieństwo
OOO	3	1/8
OOR ORO ROO	2	3/8
RRO ROR ORR	1	3/8
RRR	0	1/8

Dzieląc liczbę sytuacji sprzyjających występowaniu poszczególnych wartości zmiennej losowej przez liczbę wszystkich możliwych sytuacji znajdujemy interesujące nas prawdopodobieństwo. Znalezione rozkład jest rozkładem dwumianowym. Konstrukcja rozkładu dwumianowego oparta jest na schemacie doświadczeń zwanym schematem Bernoulliego. Przeprowadzamy n niezależnych doświadczeń losowych. W rezultacie każdego doświadczenia może wystąpić zdarzenie A (sukces) z prawdopodobieństwem p i zdarzenie przeciwne do zdarzenia A (\bar{A}) - niepowodzenie - z prawdopodobieństwem $q = 1 - p$

$A A, \dots, A$
k razy

$\bar{A} \bar{A}, \dots, \bar{A}$
n-k razy

Zmienna losowa X (liczba doświadczeń, w których wystąpi zdarzenie A) jest zmienną przyjmującą wartości $k = 0, 1, 2 \dots n$.

Funkcja prawdopodobieństwa tej zmiennej ma postać:

$$P(X = k) = \binom{n}{k} p^k q^{n-k} \quad \text{gdzie} \quad \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Dystrybuanta

$$F(X) = P(X \leq x) = \sum_{k \leq x} \binom{n}{k} p^k q^{n-k}$$

Parametry rozkładu

$$E(X) = \sum_{k=0}^n k P(X = k) = \sum_{k=0}^n k \binom{n}{k} p^k q^{n-k} = np$$

$$D^2(X) = \sum_{k=0}^n (k - np)^2 P(X = k) = np(1 - p) = npq$$

$$D(X) = \sqrt{npq}$$

$$A(X) = \frac{1 - 2p}{D(X)}$$

Własności rozkładu

1. Rozkład dwumianowy jest rozkładem sumy n niezależnych zmiennych losowych o rozkładzie zero-jedynkowym.
2. Zmienna losowa w rozkładzie dwumianowym jest zmienną skokową przyjmującą wartości liczb całkowitych nieujemnych.
3. Jeżeli $p = q$, to rozkład jest symetryczny; jeśli $p \neq q$, to rozkład jest asymetryczny.
4. W statystyce rozkład dwumianowy występuje przy losowaniu zwrotnym elementów z populacji generalnej lub przy losowaniu bezzwrotnym z populacji nieograniczonej, jeśli wynik pojedynczego losowania jest zmienną losową o rozkładzie zero-jedynkowym.

Rozkład dwumianowy jest tablicowany, wartości funkcji prawdopodobieństwa oraz dystrybuanty można odczytać z tablic statystycznych. Rozkład posiada zastosowanie w statystycznej kontroli jakości małych liczebnie prób losowych.

Czasami bywają użyteczne informacje o rozkładzie prawdopodobieństwa częstości względnej pojawienia się sukcesu w schemacie Bernoulliego. Jeśli X jest zmienną losową o rozkładzie dwumianowym z parametrami n i p , to częstość względną sukcesu definiujemy jako:

$$W = \frac{X}{n}$$

Zmienna losowa tak określona przyjmuje wartości: $0, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \dots, 1$

$$P\left(W = \frac{k}{n}\right) = P\left(\frac{X}{n} = \frac{k}{n}\right) = P(X = k)$$

$$E(W) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} np = p$$

$$D^2(W) = D^2\left(\frac{X}{n}\right) = \frac{1}{n^2} D^2(X) = \frac{1}{n^2} np(1-p) = \frac{pq}{n}$$

ROZKŁAD POISSONA

Dana jest zmienna o rozkładzie dwumianowym

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

Założmy, że przy $n \rightarrow \infty$ p zmienia się w ten sposób, że $np = \lambda$, gdzie λ jest pewną stałą.

Wówczas:

$$\lim_{n \rightarrow \infty} P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Zmienna losowa X przyjmująca wartości $k = 0, 1, 2, \dots$ ma rozkład Poissona o parametrze λ , jeżeli jej funkcja prawdopodobieństwa wyrażona jest wzorem:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{dla } k=0, 1, 2, \dots \quad \lambda - \text{stała} > 0$$

Rozkład Poissona daje dobre przybliżenie wartości rozkładu dwumianowego, jeśli n jest dostatecznie duże, a p małe. Z reguły przyjmuje się, że:

- liczba doświadczeń (losowań) powinna być dostatecznie duża, $n > 30$ ($n > 100$).
- prawdopodobieństwo p powinno być małe, bliskie 0, $p < 0,1$.
- w literaturze przedmiotu rozkład Poissona określany jest jako rozkład "rzadkich zdarzeń".

Parametry rozkładu:

$$E(X) = \lambda \quad \lambda \text{ jest zarazem wartością średnią i wariancją}$$

$$D^2(X) = \lambda$$

Wskaźnik asymetrii

$$A(X) = \frac{1}{\sqrt{\lambda}}$$

Dystrybuanta

$$F(X) = \sum_{k \leq x} \frac{\lambda^k}{k!} e^{-\lambda}$$

Rozkład Poissona posiada tablice statystyczne.

Twierdzenie Poissona

Jeśli $\lim_{n \rightarrow \infty} p_n = 0$, przy czym $p_n = \frac{\lambda}{n}$, gdzie λ jest wielkością stałą, to: $\lim_{n \rightarrow \infty} P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}$

Wniosek: Z twierdzenia Poissona wynika, że jeśli prawdopodobieństwo sukcesu p jest małe, a n dostatecznie duże, to zachodzi przybliżona równość:

$$\binom{n}{k} p^k q^{n-k} = \frac{(np)^k}{k!} e^{-np}$$

Znaczenie praktyczne: przy dużych wartościach n nie jest wygodne liczenie prawdopodobieństwa ze wzoru Bernoulliego, gdy p jest małe, można posłużyć się wzorem Poissona przyjmując $\lambda = n * p$.

WYBRANE ROZKŁADY ZMIENNEJ LOSOWEJ CIĄGŁEJ

ROZKŁAD JEDNOSTAJNY

Najprostszym rozkładem zmiennej losowej ciągłej jest rozkład jednostajny. Zmienna losowa X , przyjmująca wartości z przedziału $\langle a, b \rangle$ ma rozkład jednostajny, jeśli jej funkcja gęstości określona jest wzorem:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{dla } a \leq x \leq b \\ 0 & \text{dla } x < a \text{ } x > b \end{cases}$$

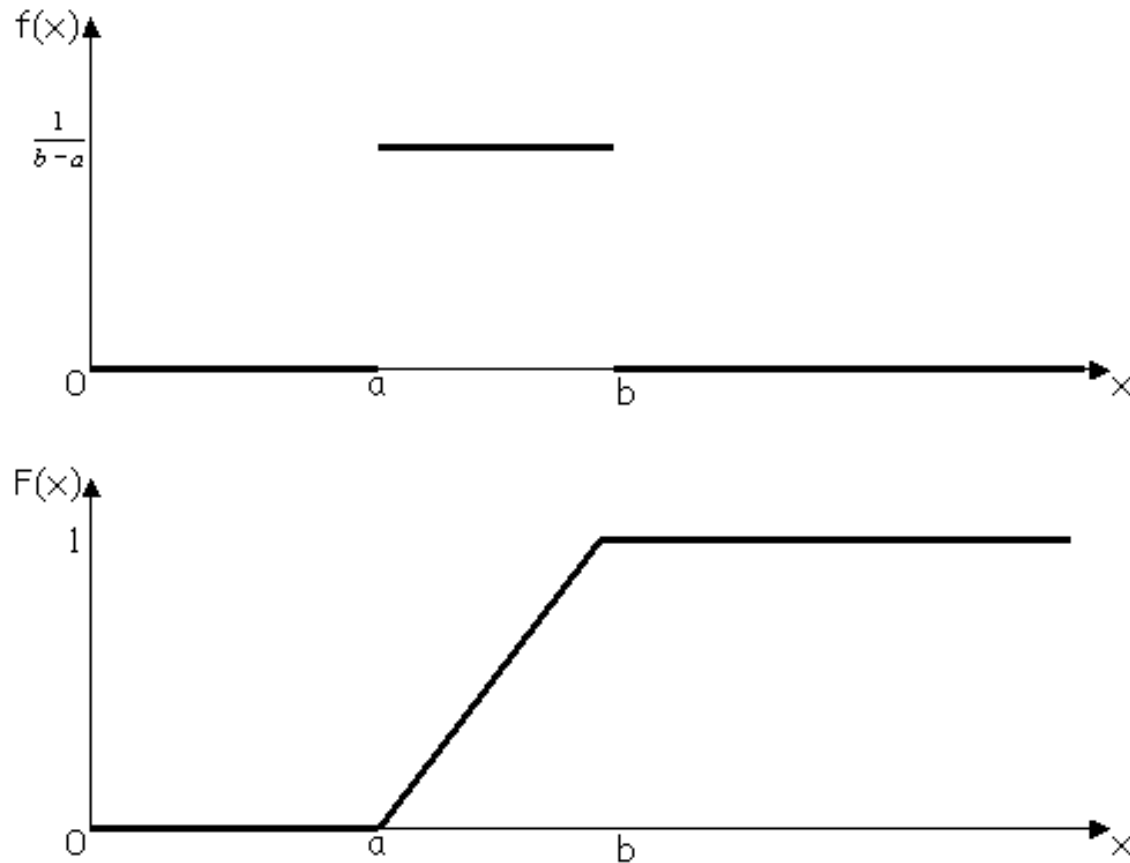
gdzie a i b są to dowolne stałe, przy czym $a < b$. Ze względu na postać wykresu funkcji gęstości rozkład ten nosi nazwę rozkładu prostokątnego.

Parametry rozkładu:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$$

$$D^2(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)dx = \int_a^b \left[x - \frac{a+b}{2} \right]^2 \frac{1}{b-a} dx = \frac{(a-b)^2}{12}$$

Wykres funkcji gęstości i dystrybuanty jest następujący:



ROZKŁAD NORMALNY

Jest podstawowym rozkładem zmiennej losowej ciągłej. Teoretyczne podstawy rozkładu wiążą się głównie z nazwiskami P.S. Laplace'a (1749-1827) oraz K.F. Gaussa (1777-1855). Stąd często w literaturze można spotkać określenia: rozkład normalny Gaussa-Laplace'a lub rozkład Gaussa. Z obserwacji zjawisk otaczającego nas świata wynika, że wiele z nich posiada rozkład zbliżony do rozkładu normalnego.

Zmienna losowa X posiada rozkład normalny, jeśli funkcja gęstości tej zmiennej dana jest wzorem:

inny zapis:

$$f(x) = \frac{1}{D(X)\sqrt{2\pi}} e^{-\left\{\frac{[x-E(x)]^2}{2D^2(X)}\right\}} \quad \text{inny zapis:} \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left\{\frac{[x-m]^2}{2\sigma^2}\right\}} \quad , \text{gdzie } -\infty < x < \infty$$

Odpowiednio $E(X) = m$, $D^2(X) = \sigma^2$, $D(X) = \sigma$ oznaczają wartość oczekiwaną, wariancję i odchylenie standardowe rozkładu.

Rozkład normalny jest określony całkowicie dwoma parametrami: $E(X)$ i $D(X)$, co można zapisać:

$$X \in N(E(X); D(X))$$

e i π są to stałe.

$e = 2,718$ podstawa logarytmów naturalnych,

$\pi = 3,142$ stosunek obwodu koła do jego średnicy.

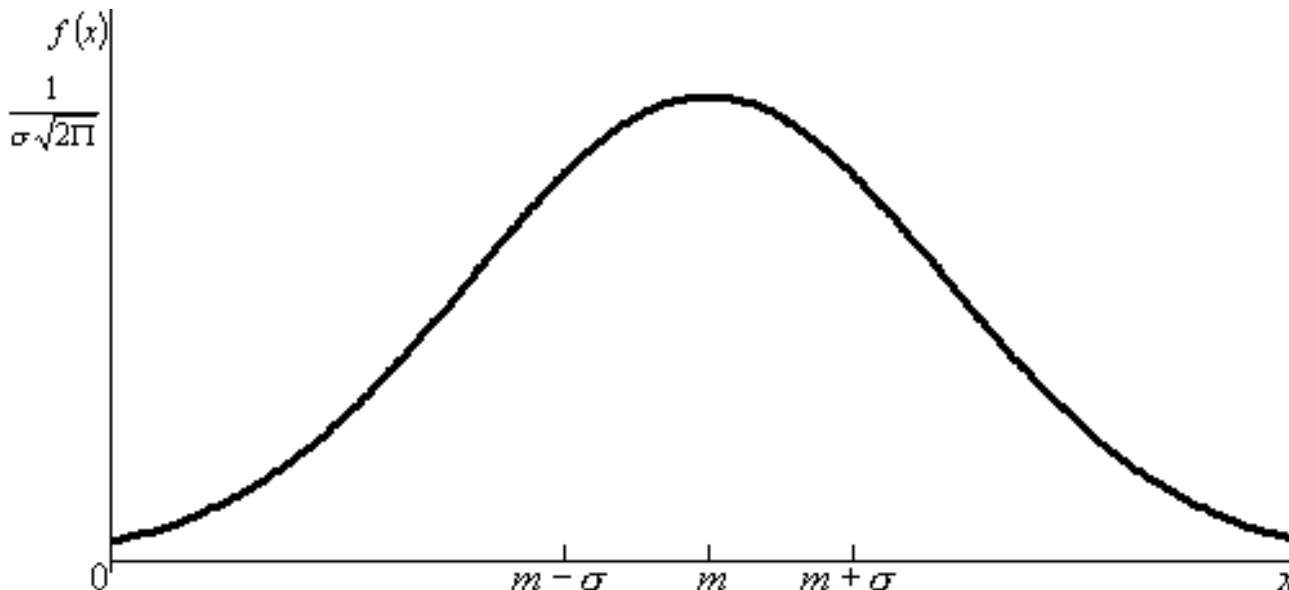
Parametry rozkładu:

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{D(X)\sqrt{2\pi}} e^{-\left\{\frac{[x-E(X)]^2}{2D^2(X)}\right\}} dx = m$$

$$D^2(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \frac{1}{D(X)\sqrt{2\pi}} e^{-\left\{\frac{[x-E(X)]^2}{2D^2(X)}\right\}} dx = \sigma^2$$

$$D(X) = \sigma$$

Funkcja gęstości rozkładu normalnego z parametrami m i σ



Wykres funkcji gęstości, określany jako krzywa normalna ma charakterystyczny kształt dzwonu.

Własności krzywej normalnej:

1. Jest symetryczna względem prostej $x = E(X)$

2. Osiąga maksimum równe $\frac{1}{D(X)\sqrt{2\pi}}$ dla $x = E(X)$

3. Posiada dwa punkty przegięcia dla $x = E(X)+D(X)$, $x = E(X)-D(X)$

Jak widać z wykresu wartość parametru $E(X)$ decyduje o położeniu krzywej względem osi X , natomiast własności nr 2 i 3 wskazują, że od wartości $D(X)$ zależy smukłość krzywej.

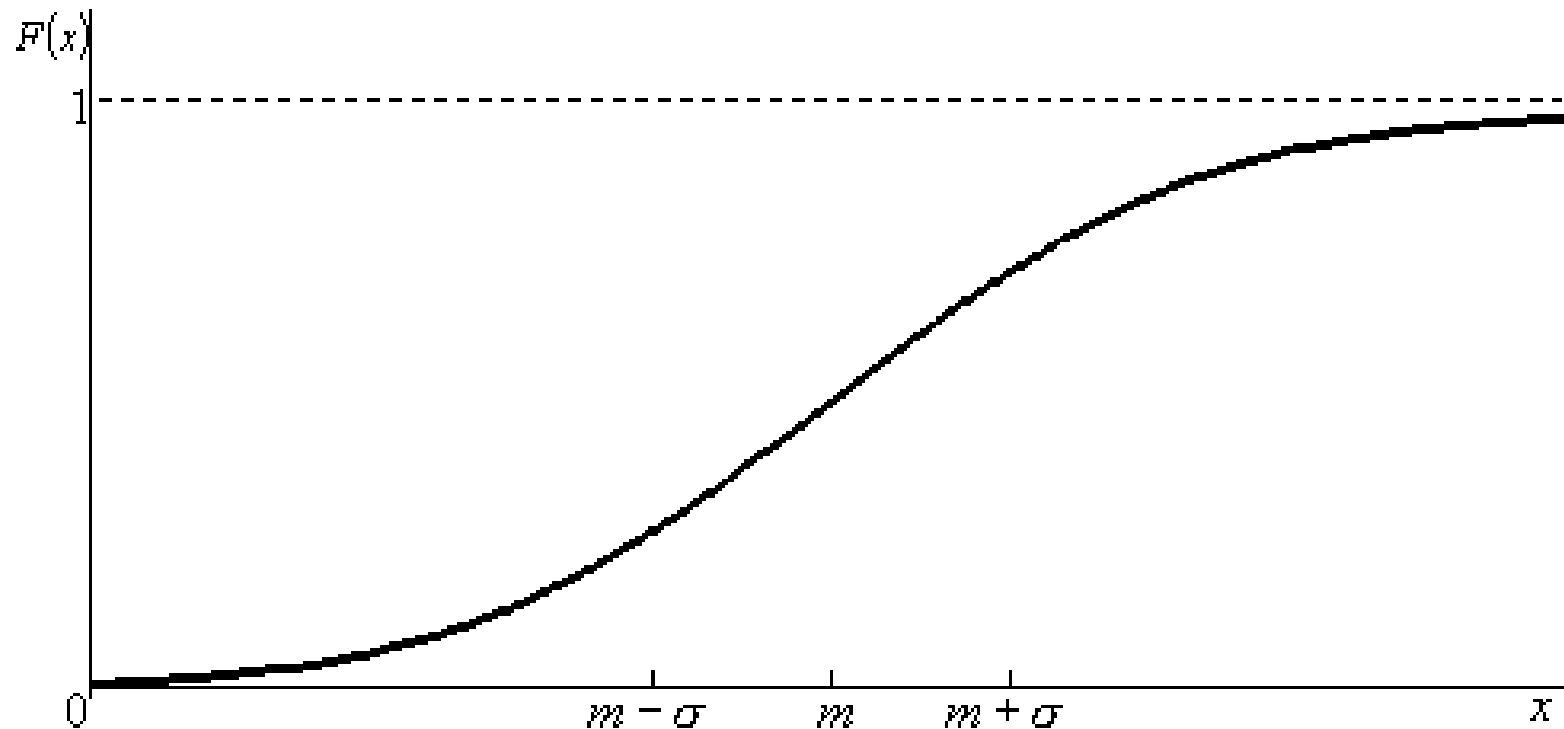
Poniższy wykres wskazuje jak wartości $E(X)$ i $D(X)$ wpływają na kształt i położenie krzywej.

Dystrybuanta rozkładu normalnego:

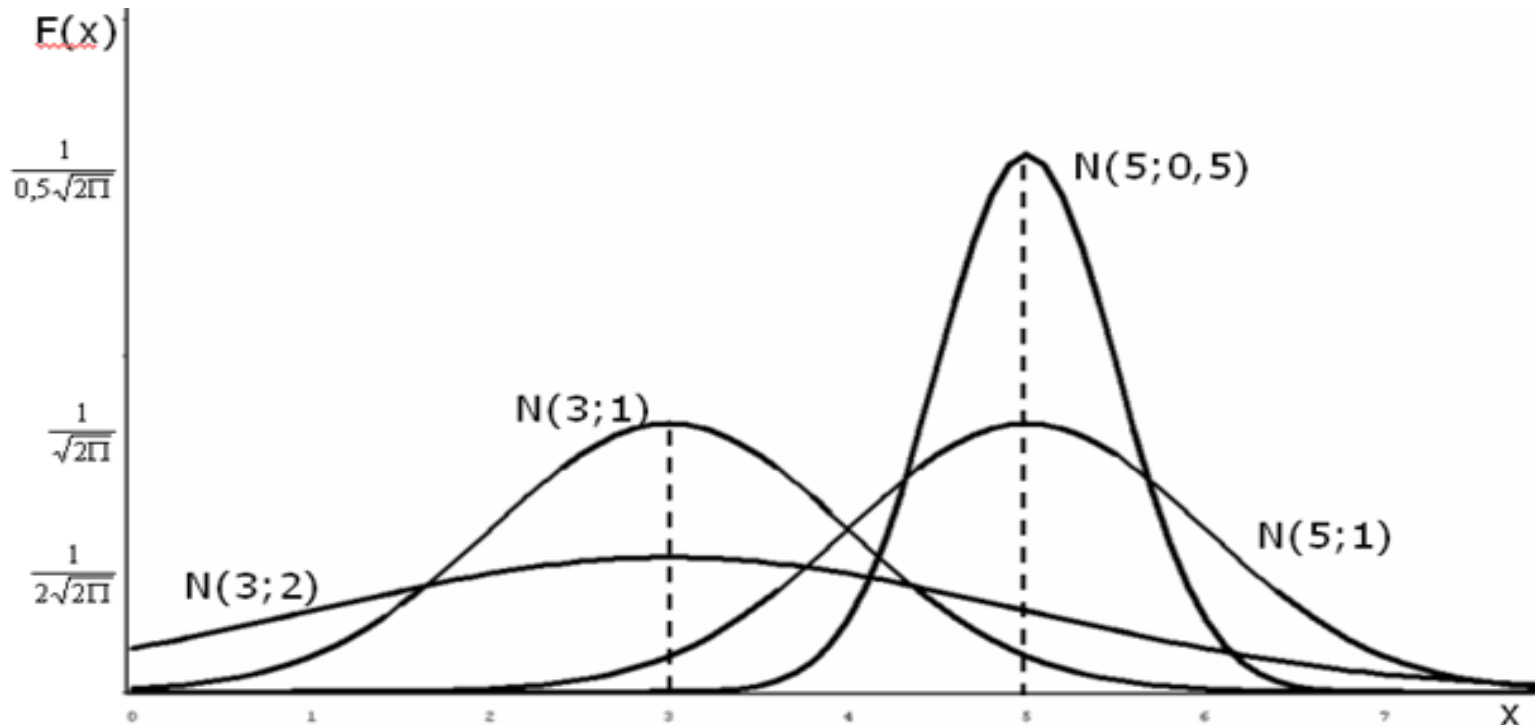
$$F(x) = \frac{1}{D(X)\sqrt{2\pi}} \int_{-\infty}^x e^{-\left\{\frac{[t-E(X)]^2}{2D^2(X)}\right\}} dt$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\left\{\frac{[t-m]^2}{2\sigma^2}\right\}} dt$$

Dystrybuanta rozkładu normalnego z parametrami m i σ



Wpływ parametrów m i σ na położenie i kształt krzywej normalnej



Źródło: J.Jóźwiak, J.Podgórski, Statystyka od podstaw, PWE, 2006, rys.6.8.,s.144

Ze względu na dużą użyteczność rozkładu normalnego jest on bardzo często stosowany do opisu zjawisk (rzeczywistości). Należy pamiętać, że obliczanie wartości prawdopodobieństw w przedziale (a, b) każdorazowo z wykorzystaniem rozkładu normalnego o dowolnych parametrach $E(X)$ i $D(X)$ jest bardzo żmudne, wymaga każdorazowo całkowania funkcji gęstości z odpowiednimi parametrami $E(X)$ i $D(X)$ w granicach całkowania (a, b) . Wielką pomocą jest możliwość sprowadzania każdego rozkładu normalnego o dowolnych parametrach $E(X)$ i $D(X)$ do postaci tzw. standardowego rozkładu normalnego, dla którego istnieją tablice funkcji gęstości i dystrybuanty.

Pojęcie zmiennej standaryzowanej.

Zmienna U definiowana jest następująco:

$$U = \frac{X - E(X)}{D(X)}$$

lub używając innych oznaczeń na średnią i odchylenie standardowe:

$$U = \frac{X - m}{\sigma}$$

$$E(U) = 0, \quad D^2(U) = D(U) = 1$$

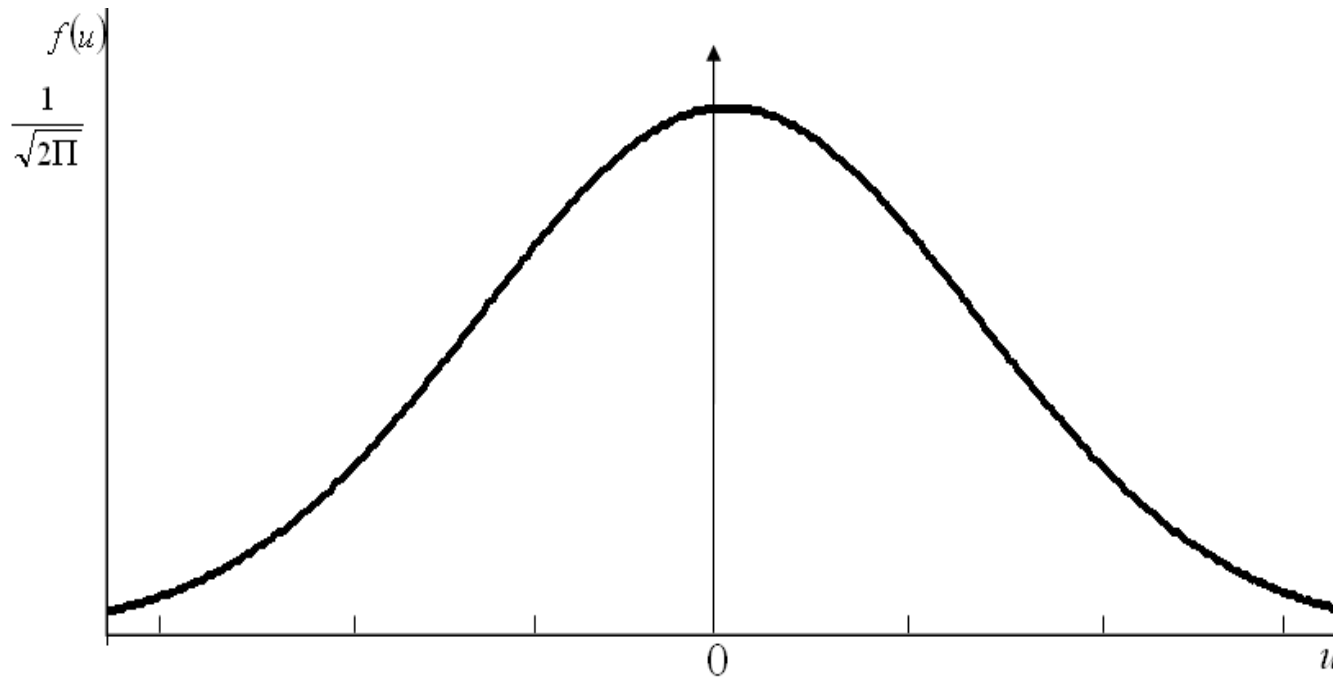
$$E(U) = E\left(\frac{X - E(X)}{D(X)}\right) = \frac{1}{D(X)} E[X - E(X)] = \frac{1}{D(X)} [E(X) - E(X)] = 0$$

$$D^2(U) = E(U^2) = E\left[\left(\frac{X - E(X)}{D(X)}\right)^2\right] = \frac{1}{D^2(X)} E[(X - E(X))^2] = \frac{1}{D^2(X)} D^2(X) = 1$$

Po standaryzacji dowolny rozkład normalny o parametrach $E(X)$ i $D(X)$ może być sprowadzony do rozkładu normalnego standaryzowanego, którego funkcja gęstości dana jest wzorem:

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \quad \text{zaś dystrybuanta} \quad F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{t^2}{2}} dt$$

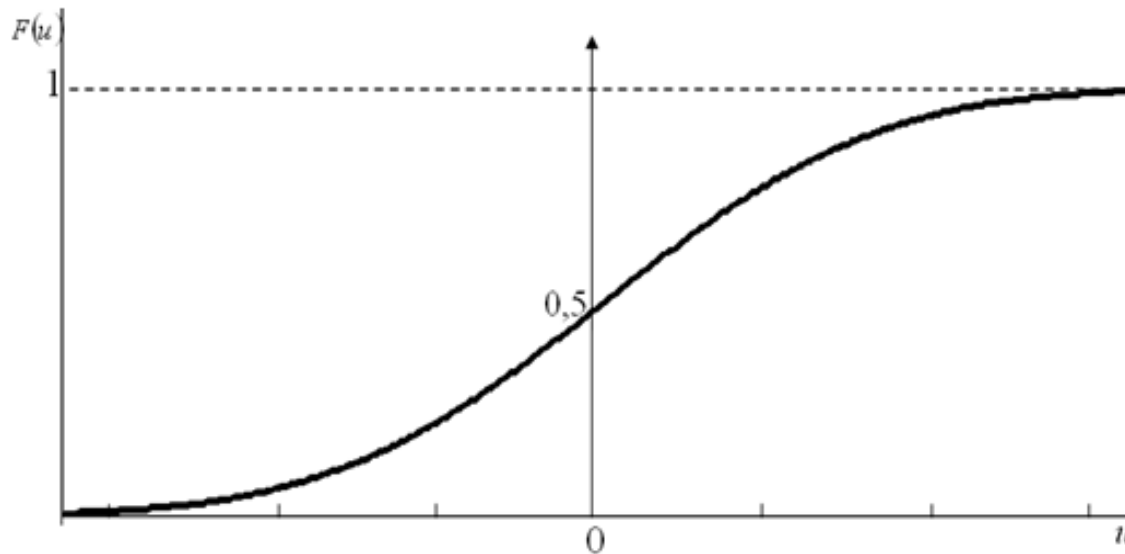
Wykres funkcji gęstości rozkładu normalnego $N(0;1)$



Własności krzywej:

1. Jest symetryczna względem prostej $u=0$.
2. Osiąga maksimum równe dla $u=0$.
3. Posiada dwa punkty przegięcia dla $u = 1$ i $u = -1$,

Wykres dystrybuanty rozkładu $N(0;1)$



Ze względu na symetrię funkcji gęstości względem prostej $u=0$ w tablicach podane są często

wartości obu funkcji tylko dla $u > 0$. Przy wyznaczaniu wartości $f(u)$ i $F(u)$ należy korzystać z następujących własności:

$$\begin{aligned} f(-u) &= f(u) \\ F(-u) &= 1 - F(u) \\ F(u) &= P(U \leq u) \end{aligned}$$

$$F(-\infty) = 0$$

$$F(\infty) = 1$$

$$F(-u) + F(u) = 1$$

$$P(u_1 \leq U \leq u_2) = F(u_2) - F(u_1)$$

$$P(U > u) = 1 - P(U \leq u)$$

Niech zmienna losowa X ma rozkład $N(E(X), D(X))$. Obliczyć prawdopodobieństwo, że zmienna losowa X przyjmie wartości różniące się od średniej $E(X)$ nie więcej niż:

- a) - jedno odchylenie standardowe,
- b) - dwa odchylenia standardowe,
- c) - trzy odchylenia standardowe.

a) odczytujemy z tablic dystrybucyj $N(0;1)$

$$\begin{aligned} P\{|X - E(X)| \leq D(X)\} &= P\{D(X) < X - E(X) \leq D(X)\} = P\left\{-\frac{D(X)}{D(X)} < \frac{X - E(X)}{D(X)} < \frac{D(X)}{D(X)}\right\} = \\ &= P\{-1 < U \leq 1\} = F(1) - F(-1) = F(1) - (1 - F(1)) = 2F(1) - 1 \end{aligned}$$

$$F(1) = 0,84135$$

$$2F(1) - 1 = 2 * 0,84135 - 1 = 0,6827 = 68,27\%$$

$$\begin{aligned} \text{b) } P\{|X - E(X)| \leq 2D(X)\} &= P\{-2D(X) < X - E(X) \leq 2D(X)\} = P\left\{-\frac{2D(X)}{D(X)} < \frac{X - E(X)}{D(X)} < \frac{2D(X)}{D(X)}\right\} = \\ &= P\{-2 < U \leq 2\} = F(2) - F(-2) = F(2) - (1 - F(2)) = 2F(2) - 1 \end{aligned}$$

$$F(2) = 0,97725$$

$$2F(2) - 1 = 2 * 0,97725 - 1 = 0,95450 = 95,45\%$$

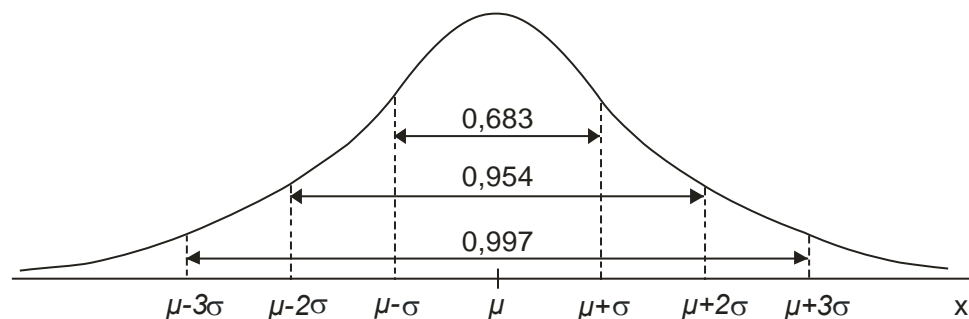
$$\begin{aligned} \text{c) } P\{|X - E(X)| \leq 3D(X)\} &= P\{-3D(X) < X - E(X) \leq 3D(X)\} = P\left\{-\frac{3D(X)}{D(X)} < \frac{X - E(X)}{D(X)} < \frac{3D(X)}{D(X)}\right\} = \\ &= P\{-3 < U \leq 3\} = F(3) - F(-3) = F(3) - (1 - F(3)) = 2F(3) - 1 \end{aligned}$$

$$2F(3) - 1 = 0,9973 = 99,73\%$$

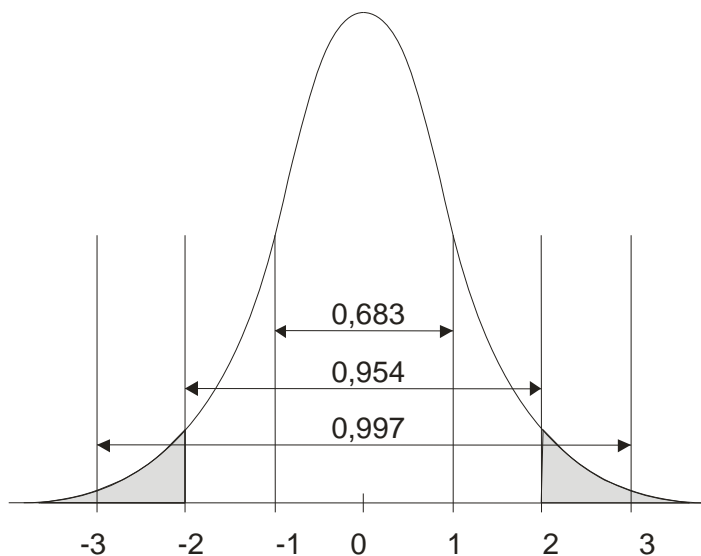
Wyniki przykładów (a), (b), (c) charakteryzują stopień skupienia się wartości zmiennej losowej wokół średniej w rozkładzie normalnym. Na ich podstawie można powiedzieć, że gdybyśmy obserwowali realizacje zmiennej losowej o dowolnym rozkładzie normalnym to około 68,3% obserwacji mieści się w granicach jednego odchylenia standardowego wokół średniej, około 95,5% obserwacji mieści się w granicach dwóch odchylen standardowych, około 99,7% w granicach trzech odchylen standardowych.

Wykresy funkcji gęstości rozkładu normalnego i normalnego standaryzowanego oraz reguła trzech sigm

$N(m, \sigma)$



$N(0,1)$



MOMENTY ROZKŁADU ZMIENNEJ LOSOWEJ

Rozkłady zmiennych losowych charakteryzowane są za pomocą parametrów zwanych momentami. Momenty mogą być zwykłe i centralne.

Momentem zwykłym (lub po prostu momentem) rzędu k , gdzie $k = 1, 2, \dots$ zmiennej losowej X nazywamy wartość oczekiwaną k -tej potęgi zmiennej X .

$$m_k = E(X^k) = \begin{cases} \sum x_i^k p_i & \text{dla zmiennej losowej skokowej} \\ \int_{-\infty}^{\infty} x^k f(x) dx & \text{dla zmiennej losowej ciągłej} \end{cases}$$

$E(X)$ jest momentem zwykłym rzędu pierwszego, $m_k = E(X)$

Momentem centralnym rzędu k , gdzie $k = 1, 2, \dots$ zmiennej losowej X nazywamy wartość oczekiwaną funkcji $g(X)$, gdzie: $g(X) = [X - E(X)]^k$ tej zmiennej, to znaczy

$$\mu_k = E[(X - E(X))^k] = \begin{cases} \sum [x_i - E(X)]^k p_i & \text{dla zmiennej losowej skokowej} \\ \int_{-\infty}^{\infty} [x - E(X)]^k f(x) dx & \text{dla zmiennej losowej ciągłej} \end{cases}$$

$D^2(X)$ jest drugim momentem centralnym zmiennej losowej X .

3.4. Prawa wielkich liczb, twierdzenia graniczne, statystyki z próby i ich rozkłady.

WYBRANE TWIERDZENIA GRANICZNE

Twierdzenia graniczne zarówno w teorii statystyki jak i w praktyce badań statystycznych mają ogromne znaczenie. Istotą ich jest to, że rozpatruje się ciąg zmiennych losowych (X_n) , których rozkłady przy wzroście n do nieskończoności mogą być zbieżne do pewnego rozkładu. Mówi się wtedy, że ciąg zmiennych losowych (X_n) ma graniczny (asymptotyczny) rozkład określonej postaci.

Twierdzenia graniczne formułują warunki, przy których istnieje dla ciągu zmiennych losowych asymptotyczny rozkład, oraz określają jaka jest postać rozkładu.

W twierdzeniach granicznych bada się dla dowolnego ciągu zmiennych losowych zbieżność odpowiadającego mu ciągu funkcji prawdopodobieństwa lub funkcji gęstości oraz ciągu dystrybuant.

Twierdzenia mówiące o zbieżności ciągu dystrybuant zmiennych losowych do dystrybuanty granicznej noszą nazwę INTEGRALNYCH TWIERDZEŃ GRANICZNYCH.

Twierdzenia mówiące o zbieżności funkcji prawdopodobieństwa zmiennych losowych skokowych lub o zbieżności funkcji gęstości zmiennych losowych ciągłych noszą nazwę LOKALNYCH TWIERDZEŃ GRANICZNYCH.

Twierdzenia graniczne mówiące o zbieżności ciągu dystrybuant do dystrybuanty rozkładu $N(0;1)$ noszą nazwę CENTRALNYCH TWIERDZEŃ GRANICZNYCH RACHUNKU PRAWDOPODOBIENSTWA. Odmianą klasę twierdzeń granicznych stanowią PRAWA WIELKICH LICZB, które dotyczą zbieżności ciągu zmiennych losowych do rozkładu jednopunktowego.

Pojęcie zbieżności stochastycznej

Mówimy, że ciąg zmiennych losowych (X_n) jest stochastycznie zbieżny do stałej c jeśli dla dowolnego $\varepsilon > 0$ spełniona jest zależność:

$$\lim_{n \rightarrow \infty} P\{|X_n - c| < \varepsilon\} = 1$$

Stochastyczna zbieżność ciągu zmiennych losowych (X_n) do stałej c oznacza inaczej, że gdy $n \rightarrow \infty$ to gęstość prawdopodobieństwa koncentruje się wokół wartości c . Inaczej, oznacza to, że rozkład zmiennej losowej (X_n) zmierza do rozkładu jednopunktowego, zatem rozkład jednopunktowy jest jej rozkładem granicznym.

Przytoczone zostaną wybrane twierdzenia, bez dokładnego wyprowadzenia dowodów - w formie uproszczonej.

PRAWA WIELKICH LICZB

1. PRAWO WIELKICH LICZB BERNOULLIE’GO

Jest najstarszym prawem wielkich liczb, udowodnione przez autora w 1713 roku, zwane złotym twierdzeniem. Jeśli (X_n) jest ciągiem zmiennych losowych o rozkładzie dwumianowym danych funkcją rozkładu prawdopodobieństwa:

$$P(X_n = k) = \binom{n}{k} p^k q^{n-k}$$

z parametrami n i p , to dla dowolnie małej liczby $\varepsilon > 0$ zachodzi:

co może być zapisane $\lim_{n \rightarrow \infty} P\left\{\left|\frac{X_n}{n} - p\right| < \varepsilon\right\} = 1$ również jako:

Zmienna X_n – jest $\lim_{n \rightarrow \infty} P\{|w_i - p_i| < \varepsilon\} = 1$ to liczba sukcesów w n doświadczeniach,

$\frac{X_n}{n}$ - częstość względna sukcesów w n doświadczeniach.

Z twierdzenia tego wynika, że wraz ze wzrostem liczby przeprowadzonych doświadczeń, z których każde może skończyć się zdarzeniem A (sukcesem) lub (porażką), zaobserwowana częstość realizacji zdarzenia A skupia się wokół pewnej stałej liczby, która to liczba jest prawdopodobieństwem realizacji zdarzenia A.

2. PRAWO WIELKICH LICZB CZEBYSZEWA

Zostało udowodnione w okresie 100 lat później po złotym twierdzeniu. Jeśli (X_n) jest ciągiem niezależnych zmiennych losowych posiadających jednakowy rozkład ze średnią $E(X)$ i wariancją spełniającą warunek:

$$\lim_{n \rightarrow \infty} D^2(X_n) = 0$$

to dla dowolnie małej liczby $\varepsilon > 0$ zachodzi:

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum X_i - E(X) \right| < \varepsilon \right\} = 1$$

Zależność ta oznacza, że średnia arytmetyczna n niezależnych zmiennych losowych o jednakowym rozkładzie różni się co do wartości bezwzględnej dowolnie mało od średniej w rozkładzie rozpatrywanych zmiennych losowych.

$$\lim_{n \rightarrow \infty} P\{|\bar{X} - E(X)|\} = 1$$

3. LOKALNE TWIERDZENIE POISSONA

Jeśli ciąg zmiennych losowych (X_n) jest ciągiem zmiennych losowych p rozkładzie dwumianowym z parametrami n i p tzn.

$$P(X_n = k) = \binom{n}{k} p^k q^{n-k} \quad \text{dla } k=0,1,\dots,n$$

przy czym parametr p jest funkcją n , gdzie: $p = \frac{\lambda}{n}$, gdzie λ jest stałą oraz $n=1,2,\dots$

wówczas:

$$\lim_{n \rightarrow \infty} P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Interpretacja powyższego twierdzenia jest następująca: Przy $n \rightarrow \infty$ funkcja prawdopodobieństwa rozkładu dwumianowego jest zbieżna do funkcji rozkładu prawdopodobieństwa Poissona, jeśli przy wzroście n parametr p maleje w ten sposób, że iloczyn $n * p$ jest stały i równy λ .

Należy podkreślić, że prawdopodobieństwo sukcesu $p = \frac{\lambda}{n}$ -przy dużym n jest bardzo małe, stąd mówi się, że rozkład Poissona jest rozkładem rzadkich zdarzeń.

4. TWIERDZENIE LOKALNE MOIVRE'A - LAPLACE'A

Jeśli (X_n) jest ciągiem zmiennych losowych o rozkładzie dwumianowym z parametrami n i p , tzn.:

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{dla } k=0,1,2,\dots,n$$

to dla ustalonego p zachodzi : $\lim_{n \rightarrow \infty} P\{x_n = k\} = \frac{1}{\sqrt{np(1-p)}\sqrt{2\pi}} e^{-\frac{(k-np)^2}{2np(1-p)}}$

co oznacza, że rozkład dwumianowy zmierza do rozkładu normalnego ze średnią $E(X) = np$ i odchyleniem standardowym. Z twierdzenia $D(X) = \sqrt{np(1-p)}$ wynika, że przy dużej liczbie doświadczeń (dużych n) wartość prawdopodobieństwa rozkładu dwumianowego można wyznaczyć przy pomocy funkcji gęstości rozkładu normalnego. W statystyce często obok zmiennej X_n , która posiada rozkład dwumianowy z parametrami np , rozpatruje się również zmienną definiowaną jako $W_n = \frac{X_n}{n}$.

Zmienna ta posiada rozkład dwumianowy ze średnią $E(W_n)=p$,

odchyleniem standardowym $D(W_n) = \sqrt{\frac{p(1-p)}{n}}$

Jeśli $n \rightarrow \infty$, to rozkład zmiennej W_n (na mocy twierdzenia de Moivre'a-Laplace'a) zmierza do rozkładu normalnego ze średnią p i odchyleniem standardowym $\sqrt{\frac{p(1-p)}{n}}$

5. INTEGRALNE TWIERDZENIE de MOIVRE'A-LAPLACE'A

Niech (X_n) będzie ciągiem zmiennych losowych, o rozkładzie dwumianowym z parametrami n i p . Niech (U_n) będzie ciągiem standaryzowanych zmiennych X_n , takich że:

$$U_n = \frac{X_n - np}{\sqrt{npq}}$$

Wtedy dla ciągu dystrybuant $F_n(u)$ zmiennych losowych U_n zachodzi:

$$\lim_{n \rightarrow \infty} F_n(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{z^2}{2}} dz$$

dla każdej wartości u

Twierdzenie to mówi, że ciąg dystrybuant standaryzowanych zmiennych o rozkładzie dwumianowym jest zbieżny do dystrybuanty rozkładu normalnego $N(0,1)$. Oznacza to, że dystrybuanta rozkładu normalnego $N(0,1)$ jest dystrybuantą graniczną (asymptotyczną) ciągu dystrybuant standaryzowanych zmiennych dwumianowych a rozkład normalny jest rozkładem granicznym (asymptotycznym) rozkładu dwumianowego.

6. CENTRALNE TWIERDZENIE LINDEBERGA-LEVY'EGO

Rozpatrzmy ciąg niezależnych zmiennych losowych (X_n) o jednakowym rozkładzie, a więc identycznych wartościach oczekiwanych $E(X_n) = E(X)$ oraz wariancjach $D^2(X_n) = D(X)$

Określmy przez S zmienną daną następującym wzorem:

$$S_n = \sum_{k=1}^n X_k, \text{ gdzie } E(S_n) = nE(X), D^2(S_n) = nD^2(X)$$

Oznaczmy zmienną standaryzowaną T_n , jako:
$$T_n = \frac{S_n - nE(X)}{D(X)\sqrt{n}}$$

Jeśli (X_n) jest ciągiem niezależnych zmiennych losowych o identycznych rozkładach i skończonej wariancji, to ciąg dystrybuant $F_n(t)$ zmiennych losowych T_n danych wzorem:

spełnia:

$$T_n = \frac{S_n - nE(X)}{D(X)\sqrt{n}} \quad \lim_{n \rightarrow \infty} F_n(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{z^2}{2}} dz$$

dla każdej wartości t .

Oznacza to, że ciąg zmiennych losowych (T_n) jest zbieżny do rozkładu $N(0,1)$.

Z powyższego twierdzenia wynika, że zmienna losowa S_n ma asymptotyczny rozkład Normalny

$$N(nE(X); D(X)\sqrt{n}).$$

STATYSTYKI Z PRÓBY I ICH WYBRANE ROZKŁADY

Narzędziem tego wnioskowania są funkcje zmiennych losowych X_1, X_2, \dots, X_n , charakteryzujące próbę losową i nazywane statystykami z próby. Ogólnie zapisywać je będziemy jako:

$$T_n = T(X_1, X_2, \dots, X_n)$$

Statystyką z próby jest np. średnia z próby \bar{X}_n , wariancja z próby S_n^2 , gdzie:

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n},$$

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

Statystyka z próby jako funkcja zmiennych losowych jest zmienną losową i posiada określony rozkład. Przy wnioskowaniu statystycznym, w sytuacji kiedy chcemy wykorzystać daną statystykę, niezbędna jest znajomość rozkładu tej statystyki. Rozkład statystyki z próby można ustalić w oparciu o rozkład badanej cechy w populacji generalnej. Wyróżnia się dwa typy rozkładów: **dokładne i graniczne**.

DOKŁADNE ROZKŁADY STATYSTYK Z PRÓBY - dla ustalonego n należy określić rozkład statystyki $T_n = T(X_1, X_2, \dots, X_n)$. Z reguły liczba obserwacji jest mała i mówi się Q tzw. rozkładach małych prób.

GRANICZNE ROZKŁADY STATYSTYK Z PRÓBY - rozkład T_n , gdy $n \rightarrow \infty$ (rozkłady asymptotyczne)

Wybrane rozkłady statystyk z próby

Rozkładem chi – kwadrat (χ^2) nazywamy rozkład zmiennej losowej, która jest sumą kwadratów n niezależnych zmiennych losowych X_1, \dots, X_n o jednakowym rozkładzie normalnym $N(0,1)$

$$\chi^2 = \sum_{i=1}^n X_i^2$$

Funkcja gęstości:

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Liczba stopni swobody $\nu = n$

$$E(\chi^2) = n \quad D^2(\chi^2) = 2n$$

Uwaga

Rozkład chi – kwadrat ma **własność addytywności**

Rozkładem t – Studenta nazywamy rozkład zmiennej losowej

$$t = \frac{X}{\sqrt{Y}} \sqrt{n}$$

gdzie X ma rozkład $N(0,1)$, Y ma rozkład chi – kwadrat z $\nu = n$ stopniami swobody oraz zmienne losowe X i Y są niezależne.

Funkcja gęstości:

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad t \in \mathbf{R}$$

Liczba stopni swobody

$$\nu = n$$

$$E(t) = 0 \quad D^2(t) = \frac{n}{n-2} \quad n > 2$$

Rozkładem Fishera – Snedecora (F - Snedecora) nazywamy rozkład zmiennej losowej

$$F = \frac{\chi_{n_1}^2 / n_1}{\chi_{n_2}^2 / n_2}$$

gdzie $\chi_{n_1}^2$ i $\chi_{n_2}^2$ są niezależnymi zmiennymi losowymi o rozkładach chi-kwadrat z n_1 i n_2 stopniami swobody.

Funkcja gęstości:

$$f(z) = \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_2}{n_1}\right)^{\left(\frac{n_2}{2}\right)} z^{\left(\frac{n_1}{2}-1\right)} \left(z + \frac{n_2}{n_1}\right)^{-\frac{n_1+n_2}{2}} \quad z > 0$$

$$E(F) = \frac{n_2}{n_2 - 2} \quad n_2 > 2 \quad D^2(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} \quad n_2 > 4$$

Wybrane dokładne rozkłady statystyk z próby

Rozkład średniej arytmetycznej z próby z populacji normalnej

Twierdzenie

Jeśli zmienne losowe X_1, \dots, X_n są niezależne i o jednakowym rozkładzie normalnym $N(m, \sigma)$, to zmienna losowa

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

ma rozkład normalny ze średnią $E(\bar{X}) = m$ i odchyleniem standardowym

$$D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Rozkład średniej arytmetycznej z próby z populacji normalnej z nieznanym odchyleniem standardowym

Twierdzenie

Jeśli X_1, \dots, X_n są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym $N(m, \sigma)$, to statystyka

$$t = \frac{\bar{X} - m}{S} \sqrt{n} \quad \left(t = \frac{\bar{X} - m}{\tilde{S}} \sqrt{n-1} \right)$$

gdzie $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$, $\left(\tilde{S} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \right)$

ma rozkład t – Studenta z $\nu = n - 1$ stopni swobody.

Rozkład różnicy średnich arytmetycznych z dwóch prób z populacji normalnych przy znanych odchyleniach standardowych

Niech będą dane dwie populacje o rozkładach normalnych $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$ z których pobiera się próby liczące odpowiednio n_1 i n_2 elementów.

Wiemy, że $\bar{X}_1 \sim N\left(m_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$ a $\bar{X}_2 \sim N\left(m_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$

Wówczas
$$\bar{X}_1 - \bar{X}_2 \sim N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Rozkład różnicy średnich arytmetycznych z dwóch prób z populacji normalnych przy nieznanach (ale jednakowych) odchyleniach standardowych

Niech będą dane dwie populacje o rozkładach normalnych $N(m_1, \sigma)$ i $N(m_2, \sigma)$ z których pobiera się próby liczące odpowiednio n_1 i n_2 elementów, następnie wyznaczamy średnie \bar{X}_1 i \bar{X}_2 oraz wariancje S_1^2 i S_2^2 oraz średnia ważoną z obu prób

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Wówczas statystyka

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

ma rozkład t – Studenta z $\nu = n_1 + n_2 - 2$ stopniami swobody.

Rozkład wariancji z próby z populacji normalnej

Twierdzenie

Jeśli X_1, \dots, X_n jest prostą próbą losową z populacji o rozkładzie normalnym, to statystyka

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \quad \left(\chi^2 = \frac{n\tilde{S}^2}{\sigma^2} \right)$$

ma rozkład chi – kwadrat o $\nu = n - 1$ stopniach swobody.

Rozkład ilorazu wariancji z prób z dwóch populacji normalnych

Założmy, że z dwóch niezależnych populacji o rozkładzie normalnym z dowolnymi średnimi oraz wariancjami σ_1^2 i σ_2^2 pobiera się próby liczące odpowiednio n_1 i n_2 elementów.

Wówczas statystyka

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

ma rozkład F – Snedecora $\nu_1 = n_1 - 1$ i $\nu_2 = n_2 - 1$ stopniach swobody.

Wartości rozkładu F-Snedecora są tablicowane

Z rozkładów granicznych statystyk z próby korzysta się, kiedy rozkład populacji nie jest znany.

Wyznaczanie rozkładu granicznego nie wymaga na ogół żadnych założeń co do postaci rozkładu populacji generalnej, natomiast wymagana jest **duża liczebność próby**.

Niech zmienna X ma rozkład dwumianowy z parametrami n i p . W praktyce korzysta się często ze statystyki $W = \frac{X}{n}$ (będącej częstością sukcesów w n doświadczeniach), która posiada rozkład dwumianowy z parametrami:

$$E(W) = p \text{ oraz } D(W) = \sqrt{\frac{p(1-p)}{n}}.$$

Z twierdzenia Moivre'a-Laplace'a wynika, że $W = \frac{X}{n}$

ma graniczny rozkład normalny

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Czyli, gdy **n jest dostatecznie duże**, można przyjąć, że W posiada w przybliżeniu rozkład normalny.

Jeśli zmienne X_1 i X_2 mają rozkłady dwumianowe z parametrami p_1 i p_2 , to statystyka postaci:

$$W_1 - W_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

przy $n_1 \rightarrow \infty$ i $n_2 \rightarrow \infty$ ma **graniczny rozkład normalny** ze średnią
i odchyleniem standardowym

$$E(W_1 - W_2) = p_1 - p_2$$

$$D(W_1 - W_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Można wnioskować o tym w oparciu o własność addytywności rozkładu normalnego.

Niech zmienna X ma dowolny rozkład ze średnią m i odchyleniem standardowym σ .

Z twierdzenia Lindeberga-Levy'ego wynika, że rozkład średniej z próby

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

zmierza przy $n \rightarrow \infty$ do rozkładu normalnego z wartością oczekiwaną m i odchyleniem standardowym $\frac{\sigma}{\sqrt{n}}$

czyli do rozkładu $N(m; \frac{\sigma}{\sqrt{n}})$

Niech zmienne X_1 i X_2 mają dowolne rozkłady z parametrami, odpowiednio, m_1 i σ_1 , m_2 i σ_2 .

Różnica średnich z próby \overline{X}_1 \overline{X}_2

ma przy $n_1 \rightarrow \infty$ i $n_2 \rightarrow \infty$ **graniczny rozkład normalny z parametrami**

$$m_1 - m_2 \quad \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

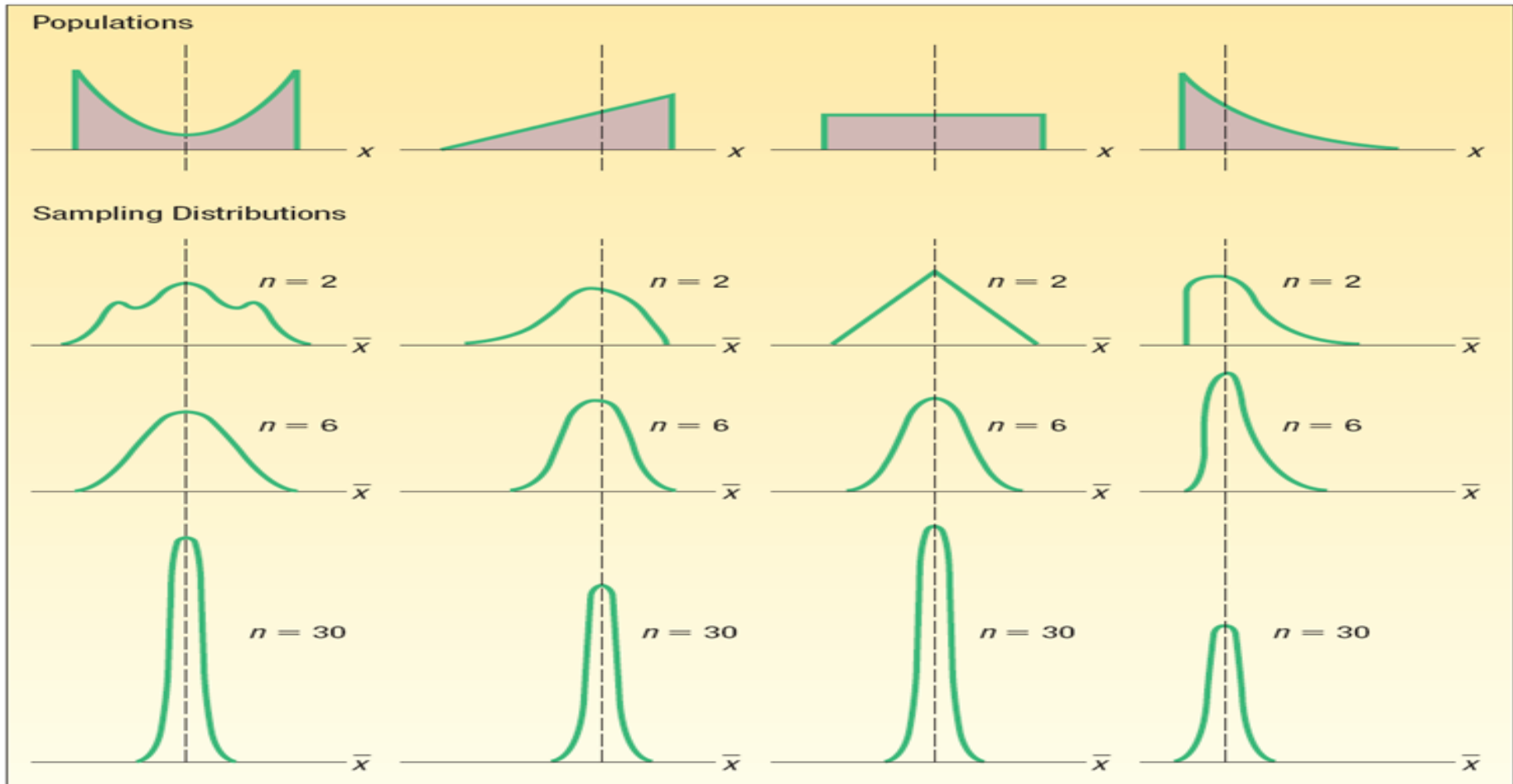
czyli rozkład $N(m_1 - m_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

**Istotę twierdzeń granicznych najlepiej oddaje
CETRALNE TWIERDZENIE GRANICZNE**

CENTRAL LIMIT THEOREM

**if all samples of particular size are selected from any
Population, the sampling distribution of the sample mean is
approximately a normal distribution. This approximation
improves with larger sample**

Istotę twierdzeń granicznych najlepiej oddaje CETRALNE TWIERDZENIE GRANICZNE



Istotę twierdzeń granicznych najlepiej oddaje CETRALNE TWIERDZENIE GRANICZNE

Properties of the Sampling Distribution

1. The mean of the sampling distribution = mean of the sampled population:

$$\mu_{\bar{X}} = \mu$$

2. The standard deviation of the sampling distribution¹²

$$= \frac{\text{Standard deviation of sampled population}}{\text{Square root of sample size}}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

(See footnote.¹³)

3. If the population distribution is normal, then so is the sampling distribution of \bar{X} .

- 4. The Central Limit Theorem** If the population distribution is not necessarily normal, and has mean μ and standard deviation σ , then, for sufficiently large¹⁴ n , the sampling distribution of \bar{X} is approximately normal, with mean

$$\mu_{\bar{X}} = \mu$$

and standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

3.5. Wnioskowanie statystyczne - podejście klasyczne – estymacja, metody pozyskiwania estymatorów

Informacje wprowadzające

Estymacja (teoria estymacji) jest podstawowym działem wnioskowania statystycznego. Stanowi ona zbiór metod pozwalających na wnioskowanie o postaci rozkładu populacji generalnej (tzn. o wartości parametrów rozkładu lub o jego postaci funkcyjnej). Teoria estymacji wiąże się z nazwiskami: K.Pearsona, R.A. Fishera, J. Neymana.

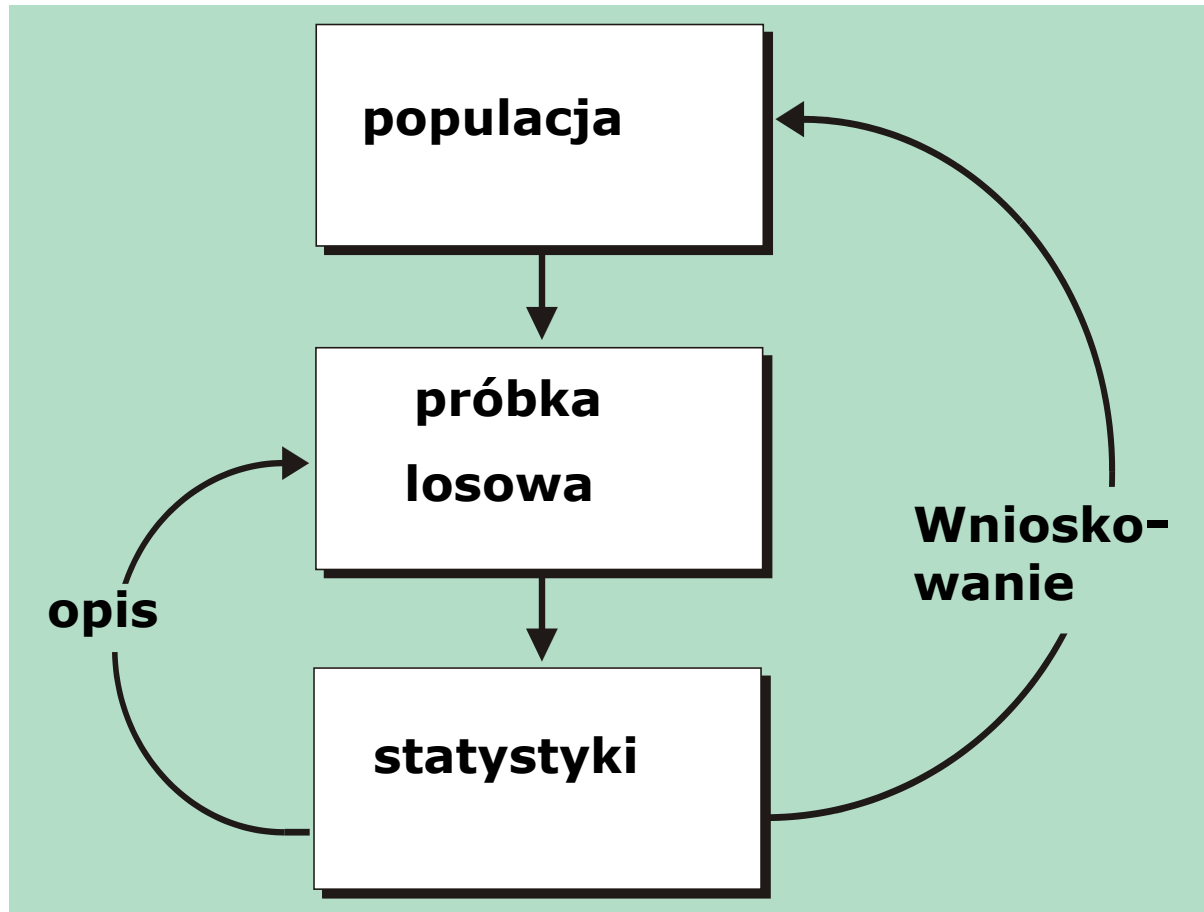
Estymacja polega na szacowaniu wartości parametrów ewentualnie postaci rozkładu zmiennej losowej w populacji generalnej na podstawie rozkładu empirycznego uzyskanego dla próby.

Jeśli szacuje się tylko wartości parametrów rozkładu populacji generalnej, mówimy o **estymacji parametrycznej**.

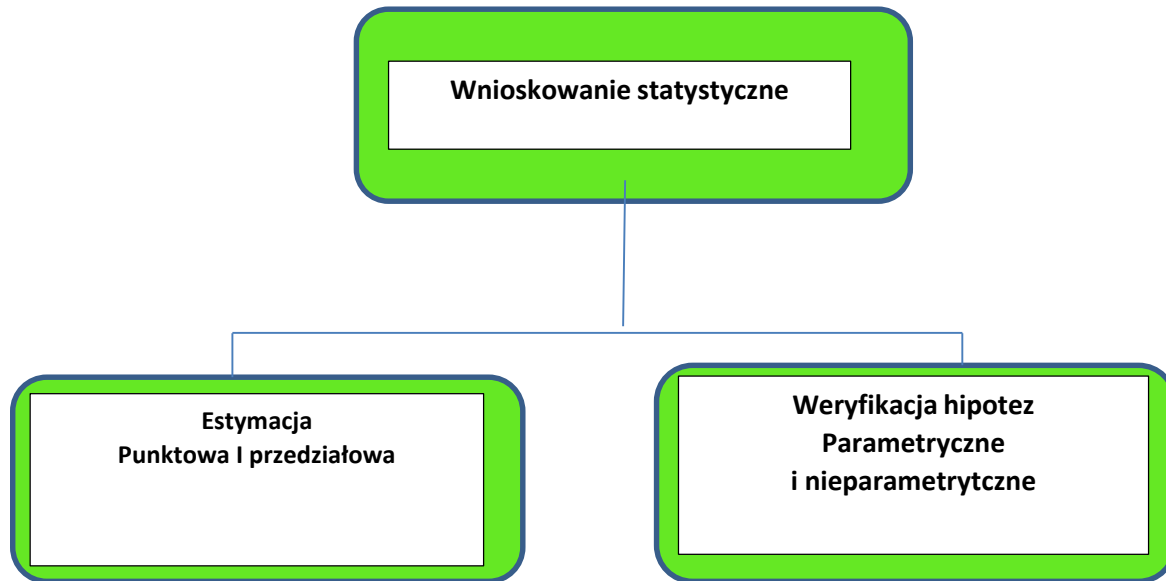
Jeśli szacowanie dotyczy również postaci funkcyjnej rozkładu populacji generalnej, mówimy o **estymacji nieparametrycznej**.

W praktyce stosuje się częściej estymację parametryczną, do oceny postaci funkcyjnej rozkładu powrócimy przy kolejnym dziale wnioskowania statystycznego tj. weryfikacji hipotez statystycznych.

Istota wnioskowania statystycznego – podstawowe pojęcia



Wnioskowanie statystyczne – ujęcie klasyczne



3.5. Wnioskowanie statystyczne - podejście klasyczne – estymacja, metody pozyskiwania estymatorów.

Informacje wprowadzające

Estymacja (teoria estymacji) jest podstawowym działem wnioskowania statystycznego. Stanowi ona zbiór metod pozwalających na wnioskowanie o postaci rozkładu populacji generalnej (tzn. o wartości parametrów rozkładu lub o jego postaci funkcyjnej). Teoria estymacji wiąże się z nazwiskami: K.Pearsona, R.A. Fishera, J. Neymana.

Estymacja polega na szacowaniu wartości parametrów ewentualnie postaci rozkładu zmiennej losowej w populacji generalnej na podstawie rozkładu empirycznego uzyskanego dla próby.

Jeśli szacuje się tylko wartości parametrów rozkładu populacji generalnej, mówimy o **estymacji parametrycznej**.

Jeśli szacowanie dotyczy również postaci funkcyjnej rozkładu populacji generalnej, mówimy o **estymacji nieparametrycznej**.

W praktyce stosuje się częściej estymację parametryczną, do oceny postaci funkcyjnej rozkładu powrócimy przy kolejnym dziale wnioskowania statystycznego tj. weryfikacji hipotez statystycznych.

Estymacja parametryczna obejmuje:

1. estymację punktową,
2. estymację przedziałową.

Ad.1. W estymacji punktowej za ocenę wartości parametru przyjmuje się konkretną wartość otrzymaną na podstawie próby losowej.

Ad.2. W estymacji przedziałowej szacuje się z wykorzystaniem odpowiednich reguł przedział liczbowy, który z określonym prawdopodobieństwem pokryje wartość szacowanego parametru.

Podstawowe pojęcia

Przyjmujemy następujące oznaczenia:

θ - parametr populacji generalnej, jest wielkością stałą i jednocześnie nieznaną.

T_n - estymator parametru θ .

Estymatorem T_n parametru θ w populacji generalnej nazywamy statystykę z próby:

$T_n = t(x_1, x_2, \dots, x_n)$, która służy do oszacowania parametru θ .

Estymator definiowany jako statystyka z próby, jest zmienną losową i jako zmienna posiada określony rozkład. Rozkład estymatora T_n jest uzależniony od rozkładu zmiennej losowej X w populacji generalnej.

t_n - ocena parametru θ

$t_n = t(x_1, x_2, \dots, x_n)$ - jest konkretną wartością liczbową, jaką przyjmuje estymator T_n parametru dla realizacji próby (x_1, x_2, \dots, x_n) .

Ocena t_n - jest realizacją zmiennej losowej T_n .

Ocena parametru $\theta(t_n)$ jest tą wielkością, jaką w estymacji punktowej przyjmuje się za oszacowanie wartości parametru θ .

Ponieważ szacunku parametru dokonuje się na podstawie próby losowej, zatem istnieje możliwość popełnienia błędu.

Błędem szacunku (estymacji) parametru θ nazywać będziemy różnicę pomiędzy estymatorem a wartością parametru co można zapisać: $T_n - \theta = d$, błąd szacunku jest zmienną losową a za miarę tego błędu przyjmuje się wyrażenie:

$$\Delta = E(T_n - \theta)^2$$

T_n - estymator

$D(T_n)$ - wariancja estymatora T_n

$D(T_n)$ - odchylenie standardowe estymatora T_n , które jest nazywane **średnim błędem (standardowym błędem) szacunku** parametru θ

Wyrażenie: $\frac{D(T_n)}{\theta}$ nazywać będziemy **względny błąd szacunku**.

$D(T_n)$ - powszechnie przyjęło się traktować jako podstawowy parametr określający dokładność estymacji punktowej.

Podstawowe własności estymatorów.

Aby oceny parametrów mogły być jak najbardziej trafne estymatory powinny spełniać określone własności. Do podstawowych własności estymatora(ów) zaliczamy:

1. nieobciążoność (asymptotyczna nieobciążoność),
2. zgodność,
3. efektywność (asymptotyczna efektywność),
4. wystarczalność zwana inaczej dostatecznością.

Nieobciążoność

Definicja 1.

Mówimy, że estymator T_n jest nieobciążonym estymatorem parametru θ , jeśli spełniona jest relacja $E(T_n) = \theta$. Spełnienie tego warunku oznacza, że estymator T_n ma rozkład ze średnią równą wartości szacowanego parametru. Jeśli szacujemy parametr θ przy pomocy estymatora nieobciążonego, to przy dużej liczbie prób, średnia uzyskanych ocen będzie bliska θ .

Definicja 2.

Mówimy, że estymator T_n parametru θ jest asymptotycznie nieobciążony jeżeli spełniona jest nierówność:

$\lim_{n \rightarrow \infty} b(T_n) = 0$ lub $\lim_{n \rightarrow \infty} E(T_n) = \theta$. Estymator $S^2(x)$ wariancji $D^2(X)$ w populacji generalnej posiada własność asymptotycznej nieobciążoności.

Zgodność

Ta własność estymatora wiąże się z dużą liczebnością próby.

Definicja 3.

Estymator T_n parametru θ jest zgodny, jeżeli spełnia relację dla dowolnego $\varepsilon > 0$ $\lim_{n \rightarrow \infty} P\{|T_n - \theta| < \varepsilon\} = 1,0$.

Estymator T_n parametru θ jest zgodny, jeżeli podlega działaniu prawa wielkich liczb, tzn. jeśli jest stochastycznie zbieżny do szacowanego parametru.

Z prawa wielkich liczb Czebyszewa wynika, że średnia arytmetyczna z próby \overline{X}_n jest zgodnym estymatorem wartości oczekiwanej w populacji generalnej.

$$\lim_{n \rightarrow \infty} P\{|\overline{X}_n - E(X)| < \varepsilon\} = 1,0 \quad \text{dla } \varepsilon > 0$$

Pomiędzy dwoma omówionymi własnościami (nieobciążonością i zgodnością) zachodzą następujące związki:

1. Jeśli estymator T_n parametru θ jest zgodny to jest asymptotycznie nieobciążony. Twierdzenie odwrotne nie jest prawdziwe.

2. Jeśli estymator T_n parametru θ jest nieobciążony lub asymptotycznie nieobciążony oraz jego wariancja spełnia $\lim_{n \rightarrow \infty} D^2(T_n) = 0$ to T_n jest estymatorem zgodnym.

Ostatnie twierdzenie jest użyteczne w sprawdzaniu zgodności estymatora np.: \overline{X}_n estymator $E(X)$, jeśli

$$D^2(\overline{X}) = \frac{D^2(X)}{n} \quad ; \quad \text{to} \quad D^2(\overline{X}_n) \rightarrow 0 \quad ; \quad \text{jeśli } n \rightarrow \infty$$

Dla jednego parametru θ populacji generalnej może istnieć kilka estymatorów. Powstaje wtedy problem wyboru odpowiedniego estymatora, estymatora najlepszego. Uzasadnione jest, aby za najlepszy estymator uznać ten, który charakteryzuje się najmniejszym rozrzutem wartości w stosunku do wartości parametru. Za najlepszy wśród estymatorów nieobciążonych uznajemy ten, który ma najmniejszą wariancję, przechodzimy tym samym do następnej własności estymatora efektywności.

Efektywność

Definicja 4.

Estymator T_n parametru θ jest najefektywniejszy, jeśli wśród estymatorów nieobciążonych ma najmniejszą wariancję. Jeśli dany jest zbiór wszystkich nieobciążonych estymatorów $T_n^1, T_n^2, \dots, T_n^r$

parametru θ , to estymator T_n^* , który ma w tym zbiorze najmniejszą wariancję, tzn.: $D^2(T_n^*) \leq D^2(T_n^i)$, $i=1, 2, \dots, r$, nazywamy najefektywniejszym estymatorem parametru θ

Miarą efektywności estymatora danego jest stosunek wariancji estymatora najefektywniejszego do wariancji estymatora danego, co można zapisać:

$$e(T_n^i) = \frac{D^2(T_n^*)}{D^2(T_n^i)} \quad \text{wyrażenie to określa się efektywnością estymatora.}$$

Efektywność estymatora najefektywniejszego jest równa jedności, w pozostałych wypadkach $0 < e < 1$.

Przy określaniu efektywności estymatora należałoby znać wszystkie estymatory nieobciążone szacowanego parametru i ich wariancję lub wiedzieć czemu jest równa wariancja estymatora najefektywniejszego.

W celu wyznaczeniu wariancji estymatora najefektywniejszego, można skorzystać z twierdzenia zwanego nierównością RAO-CRAMERA, które mówi, że przy pewnych ogólnych warunkach wariancja $D^2(T_n)$ dowolnego nieobciążonego estymatora parametru θ spełnia relację:

$$D^2(T_n) \geq \frac{1}{nE\left[\frac{\partial \log f(x, \theta)}{\partial \theta}\right]^2} = D^2(T_n^*)$$

gdzie $f(x, \theta)$ oznacza funkcję gęstości, lub funkcję prawdopodobieństwa rozkładu populacji generalnej. Dla określenia wagi estymatora najefektywniejszego, niezbędna jest znajomość postaci rozkładu populacji generalnej.

Asymptotyczna efektywność

Definicja 5.

Mówimy, że estymator T_n parametru θ jest asymptotycznie najefektywniejszy, jeśli $\lim_{n \rightarrow \infty} e(T_n) = 1,0$

tzn. jeżeli liczebność próby dąży do nieskończoności, wariancja $D^2(T_n)$ estymatora T_n przyjmuje wartości coraz bliższe wartości $D^2(T_n)$ najefektywniejszego estymatora.

Wystarczalność zwana dostatecznością

Mówimy, że estymator T_n parametru θ jest wystarczający (dostateczny), jeżeli zawiera wszystkie informacje jakie na temat tego parametru θ występują w próbie, i żaden inny estymator nie może dać dodatkowych informacji o szacowanym parametrze. Estymator dostateczny nie zawsze istnieje.

Estymator dostateczny, to taki, który dostarcza najwięcej informacji o danym parametrze wśród wszystkich możliwych estymatorów tego parametru.

Inaczej estymator wystarczający jest tak zbudowany, że żaden inny estymator nie może dostarczyć więcej informacji o szacowanym parametrze θ .

Metody uzyskiwania estymatorów.

Istnieje kilka metod uzyskiwania estymatorów. Z reguły własności estymatorów zależą również od metody uzyskiwania estymatorów.

Do najczęściej stosowanych metod należą:

- metoda momentów,
- metoda największej wiarygodności,
- metoda najmniejszych kwadratów,
- inne metody (metoda minimalnej straty, metoda najmniejszej odległości itp.)

Najstarsza metoda szacowania, **metoda momentów** (metoda analogii) jest najprostszą metodą. Idea tej metody polega na tym, że momenty (zwykle lub centralne) można przedstawić jako pewne funkcje parametrów rozpatrywanego rozkładu. Praktycznie istota tej metody polega na tym, że do szacowania parametrów populacji generalnej używamy odpowiednich momentów z próby lub ich funkcji. Tak więc, jeśli chcemy oszacować wartość średnią, to ponieważ jest to pierwszy moment, jako estymatora używamy pierwszego momentu z próby. Chcąc oszacować wariancję, biorąc pod uwagę, że jest to drugi moment centralny, jako estymatora używamy drugiego momentu z próby.

Metoda ta odznacza się dużą prostotą, jednakże estymatory uzyskiwane przy pomocy metody momentów nie mają na ogół wszystkich pożądanych własności. Z reguły metoda ta dostarcza estymatorów obciążonych i charakteryzujących się niewielką efektywnością.

Wyjątek stanowi tu średnia arytmetyczna jako estymator wartości oczekiwanej, która bez względu na rozkład zmiennej losowej w populacji generalnej ma wszystkie pożądane własności dobrego estymatora, czyli: zgodność, nieobciążoność, dostateczność. A jeśli zmienna losowa X ma rozkład $N(m, \delta)$ to średnia arytmetyczna jest również najefektywniejszym estymatorem wartości oczekiwanej m .

Metoda największej wiarygodności (konceptcja sformułowana przez R.A. Fishera w latach 20 tych XX wieku).

Jest ona jedną z najbardziej rozpowszechnionych metod estymacji. Punktem wyjścia jest określenie funkcji wiarygodności. Jeśli zmienna losowa X jest typu ciągłego to funkcja wiarygodności ma postać:

$$L(\theta) = f(x_1, \theta)f(x_2, \theta) \dots f(x_n, \theta).$$

Wyznaczenie estymatora θ metodą największej wiarygodności polega na znalezieniu maksimum funkcji $L(\theta)$ przy ustalonej wartości próby (x_1, x_2, \dots, x_n) . Przy szukaniu maksimum funkcji $L(\theta)$ wygodniej posługiwać się logarytmem tej funkcji, gdyż funkcja $\log L(\theta)$ a szukanie maksimum $\log L(\theta)$ jest na ogół łatwiejsze. Szukanie estymatorów odbywa się przy pomocy metod rachunku różniczkowego. Do wyznaczenia estymatora według metody największej wiarygodności potrzebna jest jedynie znajomość postaci funkcji gęstości lub funkcji prawdopodobieństwa rozkładu populacji generalnej.

Estymatory uzyskane metodą największej wiarygodności nie zawsze są nieobciążone, ale mają szereg innych własności. Przy pewnych ogólnych założeniach estymatory (metody największej wiarygodności) parametru θ charakteryzują się następującymi własnościami:

- są zgodne,

- mają asymptotyczny rozkład normalny o wartości oczekiwanej θ i wariancji równej:

$$\frac{1}{nE\left[\frac{\partial \log f(x, \theta)}{\partial \theta}\right]^2}$$

- są one co najmniej asymptotycznie nieobciążone i asymptotycznie najefektywniejsze,

- jeśli istnieje dostateczny estymator parametru θ to jest on estymatorem MNW,

- jeśli istnieje najefektywniejszy estymator parametru θ to jest on uzyskany metodą największej wiarygodności MNW.

Metoda najmniejszych kwadratów, (za twórców uważa się A.M. Lagendre'a, K.F. Gaussa, A.A. Markowa).

Ten typ estymatorów tzn. estymatorów uzyskanych MNK ma szczególne zastosowanie w analizie regresji.

Estymacja punktowa i przedziałowa. Przedziały ufności.

Przy estymacji punktowej uzyskaną w oparciu o losową próbę (x_1, x_2, \dots, x_n) wartość estymatora T_n przyjmujemy jako ocenę nieznanego parametru θ .

Przy estymacji punktowej ocenom parametru nie przypisujemy określonego prawdopodobieństwa, które wyrażałoby nasze zaufanie do tych ocen. Prawdopodobieństwo, że z próby uzyskamy wartość estymatora równą dokładnie wartości parametru jest równe 0.

Idea przedziałowej estymacji parametru wywodzi się z faktu, że nawet stosowanie estymatorów dobrych o wszystkich pożądanych własnościach nie gwarantuje tego, aby ocena parametru pokrywała się z rzeczywistą jego wartością. Zatem pożądane staje się, aby przy powtarzalnych próbach losowych wartość parametru znajdowała się w określonym podzbiorze zbioru możliwych wartości tego parametru z ustalonym prawdopodobieństwem.

Należy wyróżnić dwie podstawowe teorie dotyczące przedziałowej estymacji parametrów, a mianowicie:

- teorię przedziałów ufności Jerzego Sławy - Neymana (po prostu przedziały ufności Neymana),
- teorię przedziałów fiducyjnych Rolanda A. Fishera,
- teorię przedziałów T. Bayesa (zwana bayesowską).

Zasadnicza różnica między metodami estymacji Neymana i Fishera a metodą w ujęciu Bayesa, polega na uznaniu parametrów za zmienne losowe, a ich rozkłady prawdopodobieństwa można ustalić z góry przed uzyskaniem wyników z próby.

Zajmować będziemy się pierwszą z teorii tzn. przedziałami ufności Neymana. W ujęciu Neymana, parametr θ jest wielkością stałą, znany jest rozkład zmiennej losowej X w populacji generalnej dany dystrybuantą $F(x, \theta)$ oraz dany jest rozkład estymatora T_n , dany funkcją gęstości $h(t_n, \theta)$. Ustalając współczynnik ufności $P = 1 - \alpha$ określamy także funkcje $c_1(\theta)$ i $c_2(\theta)$ parametru θ , tak że:

$$P\{c_1(\theta) < T_n < c_2(\theta)\} = \int_{c_1(\theta)}^{c_2(\theta)} h(t_n, \theta) dt_n = 1 - \alpha$$

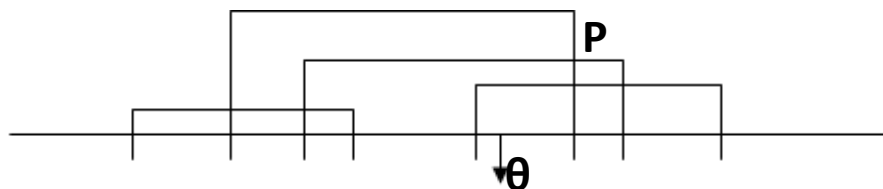
a po odpowiednich przekształceniach otrzymuje się:

$$P\{T_n^1 < \theta < T_n^2\} = \int_{c_1(\theta)}^{c_2(\theta)} h(t_n, \theta) dt_n = 1 - \alpha$$

Otrzymany w ten sposób przedział $[T_n^1; T_n^2]$ nazywamy przedziałem ufności a jego krańce odpowiednio, górną i dolną granicę przedziału ufności.

$P = 1 - \alpha$ nazywamy **współczynnikiem ufności**. Przyjmuje się Jako ex ante, subiektywnie, dowolnie duże prawdopodobieństwo zwykle bliskie jedności (nie niższe jednak niż 0,9).

Przedział liczbowy o krańcach $[T_n^1; T_n^2]$ jest jednym z możliwych do otrzymania przedziałów liczbowych, które z określonym prawdopodobieństwem $P = 1 - \alpha$ pokryją wartość szacowanego parametru populacji generalnej.



Przedział ufności parametru θ nazywamy centralnym, jeśli spełniona jest relacja:

$$P\{T_n^1 < \theta\} = P\{\theta < T_n^2\} = 1 - \frac{\alpha}{2}$$

Oznacza to, że wartości T_n^1 i T_n^2 są położone symetrycznie względem parametru θ .

Jeśli dla dużej próby estymator T_n parametru θ ma rozkład normalny, to centralny przedział ufności ma postać: $P\{-u_\alpha < U < u_\alpha\} = 1 - \alpha$, gdzie $U = \frac{T_n - \theta}{D(T_n)}$. Po odpowiednim przekształceniu tej zależności otrzymujemy:

$$P\{T_n - u_\alpha D(T_n) < \theta < T_n + u_\alpha D(T_n)\} = 1 - \alpha$$

Do tak ogólnej postaci wzoru podstawiając za T_n oraz $D(T_n)$ odpowiednie wzory na estymatory można podać postacie przedziałów ufności dla określonych parametrów populacji generalnej.

Ważne!

1. Końcówki przedziałów ufności Neymana T_n^1 i T_n^2 są zmiennymi losowymi.
2. Rozpiętość przedziału zależy od: poziomu współczynnika ufności $P=1-\alpha$, (przy czym im wyższy współczynnik, tym dłuższy przedział, oraz od odchylenia standardowego estymatora, czyli średniego błędu oceny $D(T_n)$, przy czym im większy błąd oceny, tym dłuższy przedział).
3. Im większe P tym szerszy przedział ufności, tym mniejsza dokładność estymacji.

Miarą stopnia dokładności (precyzji) oszacowania przy estymacji przedziałowej jest długość przedziału ufności: $d = T_n^2 - T_n^1$. Połowa przedziału zwana jest pół przedziałem ufności, inaczej dopuszczalnym lub maksymalnym błędem bezwzględnym lub inaczej tolerancją.

$$\Delta_{\Gamma_n} = \frac{1}{2} [T_n^2 - T_n^1] = |u_\alpha| D(T_n) \quad \text{maksymalny błąd szacunku, bezwzględna precyzja oszacowania parametru } \theta.$$

$$\delta_{\Gamma_n} = |u_\alpha| V(T_n) 100 \quad \text{względna precyzja oszacowania parametru } \theta.$$

Numeryczne przykłady estymacji wybranych parametrów rozkładów jednej zmiennej ograniczymy do:

- wartości oczekiwanej (średniej),
- prawdopodobieństwa realizacji zdarzenia π ,
- wariancji δ^2 ,
- odchylenia standardowego δ ,

Estymatorami wyróżnionych czterech parametrów są funkcje:

- średnia arytmetyczna (\bar{x}) , przy estymacji m ,
- częstość empiryczna (w_i) , przy estymacji p_i ,
- wariancja z próby $S^2(x)$, przy estymacji δ^2 ,
- odchylenie standardowe z próby $S(x)$ przy estymacji δ .

Przy czym w zależności od rodzaju estymatora i wielkości próby nie zawsze muszą to być rozkłady normalne, z reguły są różne.

Przedziały ufności dla średniej, wariancji i prawdopodobieństwa.

MODEL 1

Przedział ufności dla średniej m w populacji normalnej ze znany odchyleniem standardowym.

Niech X ma w populacji generalnej rozkład $N(m, \delta)$, gdzie δ jest nieznane, natomiast δ jest znane. Z populacji tej pobieramy n elementową próbę losową, w oparciu o wyniki, której budujemy przedział ufności dla średniej m z wartością $P=1-\alpha$. Estymatorem m jest średnia

$$\bar{x} = \frac{1}{n} \sum x_i, \text{ estymator ten ma rozkład } N\left(m, \frac{\delta}{\sqrt{n}}\right)$$

Standaryzując zmienną otrzymujemy $U = \frac{\bar{x} - m}{\delta} \sqrt{n} \in N(0,1)$ dla określonego α definiujemy u_α jako wartość w standardowym rozkładzie normalnym, która spełnia nierówność

$$P\{|U| \geq u_\alpha\} = \alpha \quad P\{-u_\alpha < U < u_\alpha\} = \left\{ -u_\alpha < \frac{\bar{x} - m}{\delta} < u_\alpha \right\} = 1 - \alpha$$

Przedział po odpowiednim przekształceniu ma postać:

$$P\left\{ \bar{x} - u_\alpha \frac{\delta}{\sqrt{n}} < m < \bar{x} + u_\alpha \frac{\delta}{\sqrt{n}} \right\} = 1 - \alpha$$

MODEL 2

Przedział ufności dla średniej m w populacji normalnej z nieznanym odchyleniem standardowym.

Zmienna X ma w populacji generalnej rozkład normalny $N(m, \delta)$, gdzie zarówno średnia jak i odchylenie standardowe są nieznanne. Losujemy n elementową próbę losową, w oparciu o wyniki tej próby konstruujemy przedział ufności przyjmując ustaloną wartość P .

Estymatorem jest $\bar{x} = \frac{1}{n} \sum x_i$, rozkład tego estymatora nie może być wyznaczony, ponieważ nie jest znana wartość δ . Wiadomo z rozkładu statystyki z próby, że statystyka $t = \frac{\bar{x} - E(X)}{S(X)} \sqrt{n-1}$, gdzie $S(x)$ jest odchyleniem standardowym z próby posiada rozkład t-Studenta o $n-1$ stopniach swobody, który jest niezależny od δ w populacji generalnej.

Dla ustalonej wartości P , α odczytujemy z tablic t-Studenta wartości $t_{\alpha,s}$. Dla określonego t spełnione są relacje:

$$P\{-t_{\alpha,n-1} < t < t_{\alpha,n-1}\} = 1 - \alpha \quad \text{podstawiając}$$

$$P\left\{t_{\alpha,n-1} < \frac{\bar{x} - m}{\delta} \sqrt{n-1} < t_{\alpha,n-1}\right\} = 1 - \alpha \quad \text{po przekształceniu otrzymujemy:}$$

$$P\left\{\bar{x} - t_{\alpha,s} \frac{S(X)}{\sqrt{n-1}} < m < \bar{x} + t_{\alpha,s} \frac{S(X)}{\sqrt{n-1}}\right\} = 1 - \alpha$$

MODEL 3

Przedział ufności dla średniej w populacji o nieznanym rozkładzie.

W praktyce, często nie znany jest rozkład zmiennej w populacji generalnej i brak jest podstaw od tego, aby przyjąć, że jest on normalny. Oznacza to, że budowa przedziałów w tej sytuacji może być oparta tylko na danych liczebności próby.

Zakładamy, że próba losowa (x_1, x_2, \dots, x_n) pochodzi z populacji o dowolnym rozkładzie z wartością $E(X)$ i ze znanym odchyleniem standardowym δ .

Konstruujemy $\bar{x} = \frac{1}{n} \sum x_i$ dla wartości przyjętego $1 - \alpha$.

Na podstawie rozkładu granicznego statystyki z próby stwierdzamy, że \bar{x} wyznaczona z próby pochodzącej z populacji o dowolnym rozkładzie posiada graniczny rozkład normalny $N\left(m, \frac{\delta}{\sqrt{n}}\right)$ a

statystyka $U = \frac{\bar{x} - m}{\delta} \sqrt{n} \in N(0,1)$

$$P\left\{-u_{\alpha} < \frac{\bar{x} - m}{\delta} \sqrt{n} < u_{\alpha}\right\} = 1 - \alpha \quad \text{po przekształceniu:} \quad P\left\{\bar{x} - u_{\alpha} \frac{\delta}{\sqrt{n}} < m < \bar{x} + u_{\alpha} \frac{\delta}{\sqrt{n}}\right\} = 1 - \alpha$$

Jeżeli δ jest nieznane a próba jest duża to można założyć, że $S(x) = \delta$. Stąd przedział ufności dla m w populacji o dowolnym rozkładzie z nieznanym δ wyznaczamy:

$$P\left\{\bar{x} - u_{\alpha} \frac{S(X)}{\sqrt{n}} < m < \bar{x} + u_{\alpha} \frac{S(X)}{\sqrt{n}}\right\} = 1 - \alpha$$

MODEL 4.

Przedział ufności dla wariancji w populacji normalnej.

Niech zmienna X ma w populacji generalnej rozkład $N(m, \delta)$ gdzie oba parametry są nieznane. Z populacji tej wylosowano niezależnie do próby n elementów (n małe, $n < 30$). W konstrukcji przedziału ufności dla wariancji korzysta się z rozkładu statystyki χ^2 postaci:

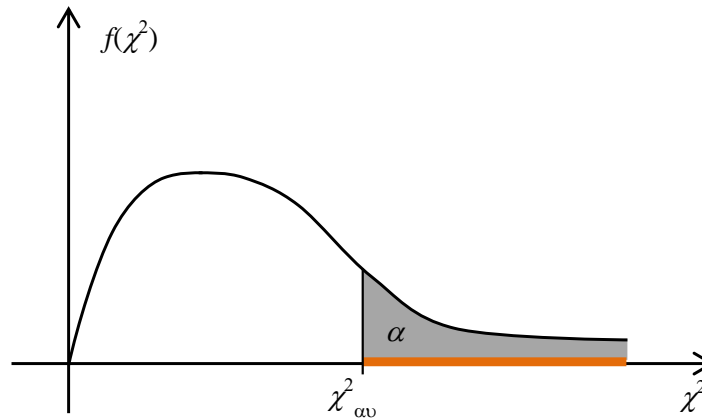
$$chi^2 = \frac{nS^2}{\delta^2}$$

Przedział ufności konstruuje się następująco:

$$P\left\{\frac{nS^2(X)}{chi^2_{\frac{\alpha}{2}, n-1}} < \delta^2 < \frac{nS^2(X)}{chi^2_{1-\frac{\alpha}{2}, n-1}}\right\} = 1 - \alpha$$

Pierwiastkując końce przedziału, można otrzymać przedział ufności dla odchylenia standardowego. Należy podkreślić, że przedział ufności nie jest symetryczny względem $S^2(x)$ (patrz poniższy rysunek).

Rozkład χ^2



MODEL 5.

$$P(\chi^2 \geq \chi^2_{\alpha}) = \alpha$$

Przedział ufności dla odchylenia standardowego.

Populacja generalna ma rozkład $N(m, \delta)$ lub zbliżony do normalnego o nieznanach parametrach: m i δ . Z populacji tej wylosowano niezależnie dużą liczbę n elementów (n co najmniej kilkadziesiąt). W oparciu o wyniki próby losowej wyznaczono wartość odchylenia standardowego $S(x)$.

Przedział ufności dla odchylenia standardowego δ w populacji generalnej jest określony wzorem:

$$P\left\{\frac{S(X)}{1 + \frac{u_\alpha}{\sqrt{2n}}} < \delta < \frac{S(X)}{1 - \frac{u_\alpha}{\sqrt{2n}}}\right\} = 1 - \alpha$$

u_α - jest wartością zmiennej normalnej standaryzowanej, po przekształceniu otrzymuje się:

$$P\left\{S(X) - u_\alpha \frac{S(X)}{\sqrt{2n}} < \delta < S(X) + u_\alpha \frac{S(X)}{\sqrt{2n}}\right\} = 1 - \alpha$$

MODEL 6.

Przedział ufności dla p_i (frakcji).

Populacja ma rozkład dwumianowy z parametrem p , gdzie $p = W = \frac{X}{n} \left(p = \frac{m}{n} \right)$

Z populacji wylosowano dużą próbę ($n > 100$). Przedział ufności dla frakcji elementów wyróżnionych jest postaci:

$$P\left\{\frac{m}{n} - u_\alpha \sqrt{\frac{\frac{m}{n} \left(1 - \frac{m}{n}\right)}{n}} < p < \frac{m}{n} + u_\alpha \sqrt{\frac{\frac{m}{n} \left(1 - \frac{m}{n}\right)}{n}}\right\} = 1 - \alpha$$

gdzie m - liczba elementów wyróżnionych w próbie.

Minimalna liczebność próby.

Dla dwóch najczęściej szacowanych parametrów populacji, tj. średniej m i frakcji p można otrzymać wzory na minimalną liczebność próby do oszacowania tych parametrów z góry żadaną dokładnością.

1. Populacja generalna ma rozkład normalny (m, δ) bądź zbliżony do normalnego, wariancja δ^2 jest znana. Jeśli przy szacowaniu średniej m na podstawie próby złożonej z n -elementów niezależnych pomiarów żądamy, aby przy ustalonym współczynniku ufności $1-\alpha$ maksymalny błąd szacunku średniej m (określany jako połowa przedziału ufności) nie przekroczył z góry danej liczby d , to niezbędną do tego celu liczebność próby n oblicza się według wzoru:

$$u_{\alpha} \frac{\delta}{\sqrt{n}} \leq d \Rightarrow n \geq \frac{u_{\alpha}^2 \delta^2}{d^2}$$

2. Populacja generalna ma rozkład dwupunktowy z parametrem p . Należy oszacować przedział dla parametru p tak, aby przy współczynniku ufności $1-\alpha$ maksymalny błąd szacunku wskaźnika struktury p nie przekroczył danej z góry liczby d . Minimalną liczebność próby, jeśli znany jest spodziewany rząd wielkości frakcji p , wyznacza się według formuły:

$$u_{\alpha} \cdot \sqrt{\frac{p(1-p)}{n}} \leq d \Rightarrow n \geq \frac{u_{\alpha}^2 p(1-p)}{d^2}$$

2a. Jeśli nieznan jest rząd wielkości wskaźnika struktury, to przyjmuje się, że iloczyn $pq=1/4$ i wzór na liczebność próby jest następujący:

$$n \geq \frac{u_{\alpha}^2}{4d^2}$$

3.6. Wnioskowanie statystyczne - podejście klasyczne – testowanie hipotez.

Informacje wprowadzające, podstawowe pojęcia.

Teoria weryfikacji (testowania) hipotez statystycznych jest drugim obok teorii estymacji działem wnioskowania statystycznego.

Hipotezą statystyczną jest każde przypuszczenie dotyczące postaci rozkładu populacji generalnej lub parametrów rozkładu zmiennej losowej. Stąd pierwszy podział hipotez statystycznych na: hipotezy nieparametryczne i hipotezy parametryczne.

Każdą hipotezę (przypuszczenie) weryfikujemy (sprawdzamy) w oparciu o próbę losową pobraną z tej populacji. Należy podkreślić, że dla hipotez statystycznych nie można z całą pewnością udowodnić ich prawdziwości lub fałszywości, gdyż wnioski opierane są jedynie na podstawie próby losowej z populacji.

Zapiszmy hipotezę statystyczną ogólnie:

$$H: F(X) \in \Omega$$

gdzie $F(x)$ jest dystrybuantą rozkładu populacji a Ω jest pewnym zbiorem dystrybuant, zwanym zbiorem hipotez statystycznych, z których jedna z nich jest przedmiotem badania.

Hipotezy statystyczne dzielą się na pewne typy:

Hipoteza statystyczna H nazywa się hipotezą parametryczną, jeżeli zbiór Ω hipotez dopuszczalnych zawiera hipotezy różniące się jedynie wartościami parametrów θ , od których zależy rozkład populacji określony dystrybuantą $F(x)$.

Hipotezy parametryczne dotyczą jedynie sądów o wartościach parametrów w ustalonym typie rozkładu populacji.

Hipotezę statystyczną H nazywa się hipotezą nieparametryczną, jeżeli elementy zbioru Ω hipotez dopuszczalnych mogą różnić się nie tylko wartościami parametrów θ , ale i postacią funkcyjną (typem) rozkładu.

Jeżeli w zapisanej hipotezie $H: F(X) \in \Omega$ zbiór hipotez dopuszczalnych składa się tylko z jednego elementu, to hipotezę taką nazywamy hipotezą prostą, natomiast gdy zbiór Ω zawiera więcej niż jedna hipotezę dopuszczalną to hipotezę H nazywa się hipotezą złożoną. Przykłady:

$H: \lambda = 2$ gdzie λ jest parametrem rozkładu Poissona (hipoteza parametryczna prosta);

$H: p \geq 0,5$ gdzie p jest parametrem rozkładu dwupunktowego (hipoteza parametryczna złożona);

$H: F(X) \in \Omega_N$ gdzie Ω_N jest zbiorem dystrybuant wszystkich rozkładów normalnych (hipoteza nieparametryczna złożona).

Hipotezy statystyczne weryfikujemy (sprawdzamy) konfrontując wyniki z próby wylosowanej z populacji z treścią hipotezy, tzn. Sprawdzamy czy wyniki z próby popierają stawianą hipotezę czy też jej przeczą. Jeżeli wyniki z próby losowej popierają postawioną hipotezę, to hipotezę tą przyjmujemy, gdy zaś wyniki z próby nie zgadzają się z nią (przeczą jej) to sprawdzoną hipotezę odrzucamy.

Narzędziem do weryfikacji hipotez statystycznych na podstawie wyników próby jest **test statystyczny**. W zależności od tego czy test służy do weryfikacji hipotezy parametrycznej czy też nieparametrycznej nazywany jest testem parametrycznym lub nieparametrycznym.

Test statystyczny - jest to reguła postępowania, która każdej możliwej próbie losowej przyporządkowuje decyzję przyjęcia lub odrzucenia sprawdzonej hipotezy.

W zależności od postaci postawionej hipotezy zerowej (tzn. bezpośrednio sprawdzanej) oraz postaci hipotezy alternatywnej (tzn. konkurencyjnej w stosunku do hipotezy zerowej) sposób budowy testu jest różny.

W procedurze testowania hipotez statystycznych, w oparciu o wyniki próby losowej, przy podejmowaniu decyzji istnieje możliwość popełnienia błędu. Możliwe błędne decyzje (dwie) przyjęło się w statystyce nazywać błędami pierwszego i drugiego rodzaju.

Schemat możliwych do przyjęcia decyzji ilustruje poniższe zestawienie:

Decyzja o hipotezie H_0 \ Hipoteza zerowa H_0	PRAWDZIWA	FAŁSZYWA
PRZYJAĆ	X decyzje prawidłowe	błąd II-go rodzaju $P = B$
ODRZUCIĆ	błąd I-go rodzaju $P = \alpha$	X decyzje prawidłowe

Błędem I-go rodzaju (α) nazywamy odrzucenie sprawdzanej hipotezy wtedy, gdy jest ona prawdziwa. Błędem II-go rodzaju (P) nazywamy przyjęcie sprawdzanej hipotezy wtedy, gdy jest ona fałszywa.

Z powyższej definicji wynika, że przy weryfikowaniu hipotezy statystycznej danym testem niemożliwe jest jednoznaczne popełnienie obu tych błędów. Na błąd pierwszego rodzaju jesteśmy narażeni tylko wówczas, gdy test doprowadza do decyzji o odrzuceniu hipotezy. Na błąd drugiego rodzaju narażamy się jedynie wtedy, gdy decydujemy się przyjąć sprawdzaną hipotezę.

Poziom istotności - prawdopodobieństwo popełnienia błędu pierwszego rodzaju w postępowaniu testującym hipotezę. Poziom istotności oznacza się zwykle symbolem α i przyjmuje się jego wartość z góry zwykle jako małe prawdopodobieństwo. Do najczęściej przyjmowanych poziomów istotności należą prawdopodobieństwa : 0,1; 0,05; 0,01; 0,001.

W klasycznej teorii weryfikacji hipotez statystycznych testy (obszar odrzucenia) konstruuje się w ten sposób, aby zminimalizować prawdopodobieństwo popełnienia błędu II-go rodzaju β przy ustalonym z góry poziomie prawdopodobieństwa popełnienia błędu I-go rodzaju α .

Tak zbudowane testy nazywa się najmocniejszymi, czyli test statystyczny nazywamy testem najmocniejszym, jeżeli oparty jest on na takim obszarze krytycznym, dla którego przy z góry danym prawdopodobieństwie α błędu I-go rodzaju, prawdopodobieństwo błędu II-go rodzaju jest najmniejsze.

Mocą testu nazywamy prawdopodobieństwo podjęcia prawidłowej decyzji polegającej na odrzuceniu sprawdzanej hipotezy, gdy jest ona fałszywa. Z definicji tej wynika, że miarą mocy testu jest wyrażenie: $M = 1 - \beta$, tak więc minimalizacja prawdopodobieństwa β jest równoważna maksymalizacji mocy testu.

W teorii weryfikacji hipotez statystycznych znane są twierdzenia (zasady) konstrukcji testów, które spełniają określone własności i tak są zasady konstruowania testów : najmocniejszych, jednostajnie najmocniejszych, zgodnych, nieobciążonych i o innych własnościach.

My zajmować się będziemy jedynie testami istotności przy weryfikacji hipotezy zerowej.

Test istotności - najczęściej używany w praktyce statystycznej typ testu pozwalający na odrzucenie hipotezy z małym ryzykiem popełnienia błędu (mierzonym) poziomem istotności.

Ze względu na to, że w teście uwzględnia się jedynie błąd I-go rodzaju, a nie rozpatruje się szansy popełnienia błędu drugiego rodzaju, to w wyniku zastosowania testu możliwa jest do podjęcia decyzja:

- odrzuć hipotezę zerową lub

- nie ma podstaw do odrzucenia hipotezy zerowej (co nie oznacza jej przyjęcia).

1. **Parametryczny test istotności** - test istotności weryfikujący hipotezę zerową precyzującą wartość parametru w ustalonym typie rozkładu populacji generalnej.
2. **Nieparametryczny test istotności** - test istotności dla hipotezy zerowej precyzującej ogólny typ postaci rozkładu populacji generalnej.

Rodzaje hipotez statystycznych, zasady konstrukcji testów, etapy weryfikacji

Kryteria stosowane przy klasyfikacji hipotez statystycznych mogą być różne. Stąd wyróżnia się:

- hipotezy proste i złożone,
- hipotezy zerowe i alternatywne (wśród hipotez alternatywnych: jednostronne i dwustronne),
- hipotezy parametryczne i nieparametryczne.

Przykładowy zestaw hipotez zerowych i alternatywnych dla hipotez parametrycznych jest następujący:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta > \theta_0$$

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta < \theta_0$$

Ogólne zasady konstruowania testów istotności.

Po sformułowaniu hipotezy zerowej i alternatywnej (H_0 i H_1), na podstawie wyników próby losowej (x_1, x_2, \dots, x_n) wyznacza się pewną statystykę Z_n (sprawdzian hipotezy), wyznacza się rozkład tej statystyki przy założeniu prawdziwości H_0 . Mając rozkład tej statystyki można wyznaczyć taki obszar jej wartości Λ , aby dla z góry ustalonego α (małego prawdopodobieństwa) spełniony był warunek :

$$P\{Z_n \in \Lambda\} = \alpha$$

Obszar Λ nazywa się obszarem krytycznym testu, ilekroć wartość statystyki Z_n dla konkretnej próby tj. Z_n znajdzie się w nim, to podejmuje się decyzję odrzucenia H_0 na korzyść H_1 . Natomiast, jeżeli otrzymana z konkretnej próby wartość statystyki Z_n nie należy do obszaru krytycznego Λ , to nie ma podstaw do odrzucenia H_0 , co nie jest równoważne z jej przyjęciem.

Powiedziano o testach dość dużo, ale nic nie wspomniano o regułach wyboru testów, statystyki Z_n , będącej sprawdzianem hipotezy. Otóż w statystyce w teorii testów istotności na ogół nie są formułowane ogólne zasady określania postaci tej statystyki. Dobór sprawdzianu jest w znacznym stopniu intuicyjny tzn. w celu zweryfikowania hipotezy o wartości oczekiwanej w populacji generalnej: $m ; [E(X)]$, za sprawdzian przyjmuje się średnią arytmetyczną z próby, a następnie z zachowaniem uprzednio wyznaczonych reguł postępowania ocenia się czy różnica pomiędzy hipotetyczną wartością $m [E(X)]$, a wartością uzyskaną z próby jest istotna, tzn. nieprzypadkowa. Jeżeli tak to H_0 należy odrzucić. Postać testów istotności może być różna.

Etapy weryfikacji hipotez statystycznych:

1. Sformułowanie hipotezy zerowej i alternatywnej.
2. Wybór testu i wyznaczenie jego wartości na podstawie próby.
3. Przyjęcie α i wyznaczenie wartości krytycznej testu, zbudowanie obszaru odrzucenia H_0 .
4. Podjęcie decyzji

Wybrane parametryczne testy istotności, podstawowe modele.

Test istotności dla wartości średniej w populacji generalnej.

Test istotności dla dwóch wartości oczekiwanych.

Test istotności dla wariancji w populacji generalnej.

Test istotności dla dwóch wariancji w populacji generalnej.

Test istotności dla wskaźnika struktury i dla dwóch wskaźników struktury

Testy istotności dla wartości średniej w populacji generalnej.

MODEL 1.

Populacja generalna ma rozkład $N(m, \sigma)$ przy czym odchylenie standardowe jest znane. Na podstawie wyników n -elementowej próby losowej pobranej z tej populacji należy sprawdzić hipotezę H_0 postaci:

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0$$

Na podstawie wyników próby losowej ustalamy wartość statystyki $\bar{x}_n = \frac{1}{n} \sum x_i$, wiemy, że jeśli H_0 jest prawdziwa to statystyka \bar{X}_n ma rozkład $N(m, \frac{\sigma}{\sqrt{n}})$. Ustalamy wartość zmiennej normalnej standaryzowanej

U według ogólnego wzoru:

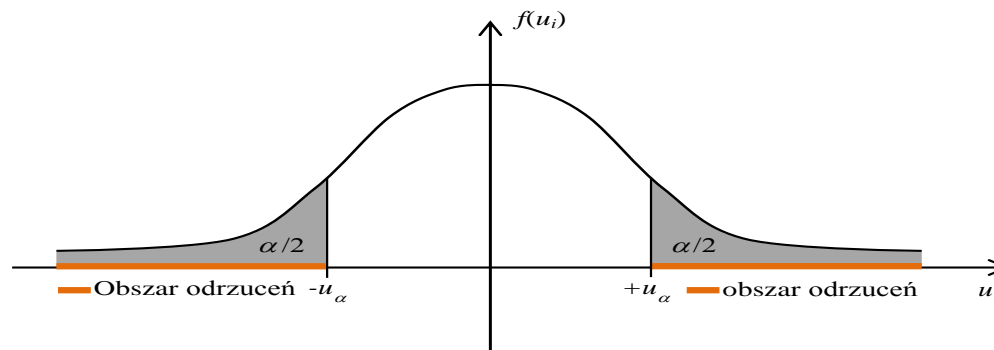
$$U = \frac{\bar{X} - m_0}{\sigma} \sqrt{n}$$

Jeśli H_0 jest prawdziwa to statystyka ta ma rozkład $N(0,1)$. Z tablic tego rozkładu wyznacza się taką wartość krytyczną u_α , aby dla z góry założonego małego prawdopodobieństwa α (poziom istotności) zachodziła równość:

$$P = \{ |U| \geq u_\alpha \} = \alpha$$

Zbiór wartości U określony nierównością $|U| > u_\alpha$ jest obszarem krytycznym tego testu tzn., jeżeli z próby otrzymamy taką wartość u , że $|u| > u_\alpha$ to hipotezę H_0 odrzucamy, gdy zaś $|u| < u_\alpha$ to nie ma podstaw do odrzucenia H_0 (tzn. wyniki z próby nie dają podstaw do odrzucenia H_0).

Graficzna prezentacja obszaru krytycznego.



Dwustronny obszar krytyczny Λ (tak zdefiniowany obszar nazywa się także symetrycznym lub centralnym). Hipoteza alternatywna może być formułowana jako hipoteza jednostronna, tzn.

$$H_o : m = m_0$$

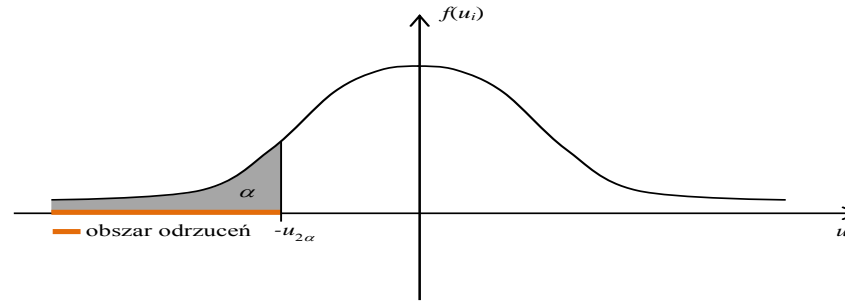
$$H_1 : m < m_0$$

$$H_1 : m > m_0$$

wówczas stosujemy test istotności z tzw. lewostronnym obszarem krytycznym określanym równością:

$$P = \{U \leq -u_{2\alpha}\} = \alpha$$

lewostronny obszar krytyczny tego testu obejmuje wartości z przedziału $(-\infty, -u_{2\alpha} >$

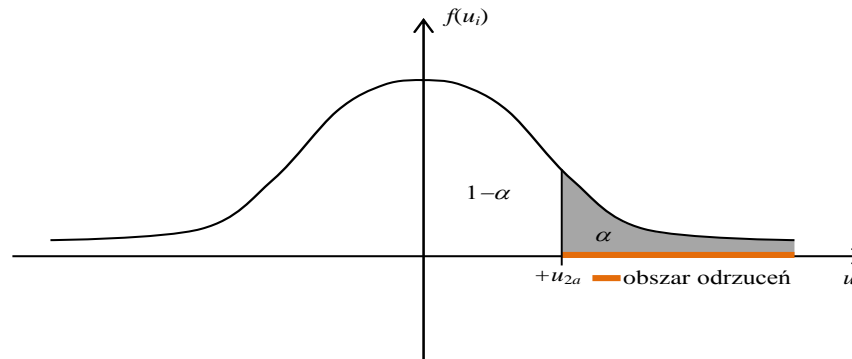


Lewostronny obszar krytyczny Λ .

Przy hipotezie alternatywnej: $H_1 : m > m_0$

stosujemy test istotności z tzw. prawostronnym obszarem krytycznym, określonym:

$$P = \{U \geq u_{2\alpha}\} = \alpha$$



Prawostronny obszar krytyczny Λ .

MODEL 2.

Populacja generalna ma rozkład $N(m, \sigma)$ przy czym odchylenie standardowe σ jest nieznane.

W oparciu o wyniki n -elementowej próby losowej należy zweryfikować hipotezę H_0 :

$$H_0 : m = m_0 \qquad H_1 : m \neq m_0$$

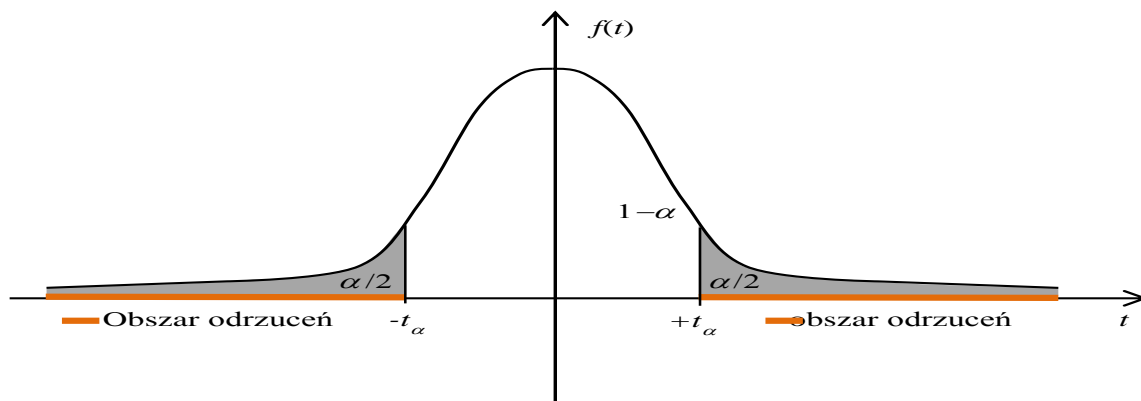
Test istotności dla powyżej sformułowanej hipotezy jest następujący: Z wyników próby oblicza się wartość statystyki \bar{x} oraz $S(x)$ a następnie wartość statystyki t :

$$t = \frac{\bar{X} - m_0}{\frac{S}{\sqrt{n-1}}}$$

Statystyka ta przy założeniu prawdziwości H_0 ma rozkład t-Studenta o $n-1$ stopniach swobody. Z tablicy tego rozkładu, dla ustalonego poziomu α i $n-1$ stopni swobody odczytuje się taką wartość t , że:

$$P = \{ |t| \geq t_{\alpha, n-1} \} = \alpha$$

Relacja powyższa wyznacza dwustronny obszar krytyczny testu, a reguły podejmowania decyzji są takie same jak poprzednio tzn., gdy t z próby przyjmie taką wartość że $|t| \geq t_\alpha$, H_0 możemy odrzucić, jeśli zachodzi $|t| < t_\alpha$ stwierdza się, że brak jest podstaw do odrzucenia hipotezy zerowej.



Analogicznie jak poprzednio, w zależności od postaci hipotezy alternatywnej można określić także lewo lub prawostronny obszar krytyczny, obejmujący odpowiednie wartości:

- lewostronny $(-\infty, -t_{2\alpha})$
- prawostronny $(t_{2\alpha}, +\infty)$

MODEL 3.

Populacja ma rozkład dowolny z nieznanymi parametrami, ale skończonej wariancji. Chcemy zweryfikować hipotezę dotyczącą wartości oczekiwanej m w tej populacji na podstawie wyników dużej próby losowej (n co najmniej rzędu kilku dziesiątków).

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0$$

Korzystamy z twierdzeń granicznych i wiemy, że przy dużym dostatecznie n średnia z próby jako statystyka ma graniczny rozkład normalny, oraz odchylenie standardowe z próby jest stochastycznie zbieżne do odchylenia standardowego w populacji.

Jeśli zatem prawdziwe jest H_0 , to średnia ma asymptotyczny rozkład normalny, $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$ natomiast

statystyka $U = \frac{\bar{X} - m_0}{s} \sqrt{n}$ ma asymptotyczny rozkład $N(0, 1)$.

Obszar krytyczny testu, korzystając z rozkładu granicznego statystyki, można określić następująco:

$$P\{|U| \geq u_\alpha\} = \alpha$$

Test istotności dla dwóch wartości oczekiwanych.

MODEL 1.

Badamy dwie populacje o rozkładach normalnych: $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$. Odchylenia standardowe tych populacji są znane. W oparciu o wyniki prób losowych pobranych z tych populacji o liczebnościach: n_1 i n_2 należy zweryfikować hipotezę postaci:

$$H_o : m_1 = m_2$$

$$H_1 : m_1 \neq m_2$$

Testem weryfikującym hipotezy H_0 jest statystyka U postaci:

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Statystyka ta przy założeniu prawdziwości H_0 ma rozkład normalny $N(0,1)$, obszar krytyczny dla H_0 określa się

$$P\{|U| \geq u_\alpha\} = \alpha$$

Dla hipotezy alternatywnej $:<$ stosuje się test z lewostronnym obszarem krytycznym zaś dla hipotezy alternatywnej $H, : m: >$ stosuje się test z prawostronnym obszarem krytycznym $P\{U \leq u_\alpha\} = \alpha$.

Wartość krytyczną testu dla jednostronnej hipotezy alternatywnej należy odczytać z rozkładu $N(0,1)$ dla 2α .

MODEL 2.

Badamy dwie populacje o rozkładach normalnych: $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$, przy czym odchylenia standardowe tych populacji nie są znane, ale zakładamy, że są one jednakowe tj. $\sigma_1 = \sigma_2$. Dla wyników prób losowych o małych liczebnościach odpowiednio: n_1 i n_2 należy zweryfikować hipotezę:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Na podstawie wyników prób losowych wyznacza się wartości średnich i wariancji, które wykorzystuje się do wyznaczenia wartości testu. Testem weryfikacyjnym hipotezy H_0 jest statystyka postaci:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)}$$

Statystyka t przy założeniu prawdziwości H_0 ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody. Obszar krytyczny przy wykorzystaniu tablic z rozkładu t-Studenta wyznacza się: $P\{|t| \geq t_{\alpha, n_1 + n_2 - 2}\} = \alpha$. Wartość krytyczną z tablic odczytuje się dla danego poziomu istotności α i liczby stopni swobody $n_1 + n_2 - 2$.

Przy hipotezach alternatywnych: $H_1 : m_1 > m_2$ lub $H_1 : m_1 < m_2$ określa się odpowiednio prawostronne i lewostronne obszary krytyczne. Przy odczytaniu wartości krytycznej testu wykorzystuje się informacje o liczbie stopni swobody i podwojonej wartości α .

MODEL 3.

Badamy dwie populacje generalne mające rozkłady normalne lub inne, ale o nieznanym lecz skończonym wariancjach σ_1^2, σ_2^2 . Na podstawie wyników danych prób losowych (n_1 i n_2 co najmniej kilku dziesiątków) weryfikuje się hipotezę:

$$H_0 : m_1 = m_2$$

$$H_1 : m_1 \neq m_2$$

Test dla H_0 buduje się analogicznie jak w modelu 1, korzystając z rozkładu $N(0,1)$.

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Test istotności dla wariancji w populacji generalnej.

MODEL 1.

Populacja generalna ma rozkład normalny $N(m, \sigma)$ o nieznanach parametrach m i σ . Z populacji tej wylosowano niezależnie n -elementową próbę, w oparciu o wyniki której należy zweryfikować hipotezę:

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_0 : \sigma^2 > \sigma_0^2$$

gdzie σ_0^2 jest hipotetyczną wartością wariancji σ^2 . Testem weryfikującym H_0 jest statystyka postaci:

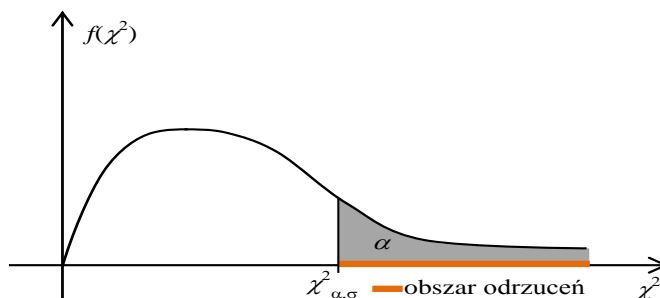
$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2$$

Tak określona statystyka (test) ma rozkład χ^2 o $n-1$ stopniach swobody. Dla ustalonego z góry poziomu α i dla $n-1$ stopni swobody z tablic χ^2 odczytać należy wartość krytyczną testu tak, aby spełniona była równość:

$$P\{\chi^2 \geq \chi_{\alpha, n-1}^2\} = \alpha$$

Nierówność $\chi^2 \geq \chi_{\alpha, n-1}^2$ określa prawostronny obszar krytyczny. Oznacza to, że jeśli z porównania wartości empirycznej χ^2 i wartości odczytanej z tablic $\chi_{\alpha, n-1}^2$ zachodzi nierówność $\chi^2 \geq \chi_{\alpha, n-1}^2$ hipotezę H_0 należy odrzucić na korzyść H_1 , w przeciwnym przypadku nie ma podstaw do odrzucenia H_0 .

Graficznie obszar odrzucenia H_0 przedstawia się następująco:



Test istotności dla dwóch wariancji w populacji generalnej.

MODEL 1.

Dane są dwie populacje generalne o rozkładach normalnych odpowiednio: $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$ gdzie, parametry tych rozkładów są nieznane. Z populacji tych wylosowano niezależnie dwie próby o liczebnościach odpowiednio: n_1 i n_2 elementów. Na podstawie wyników tych prób należy zweryfikować hipotezę:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 > \sigma_2^2$$

$$S_1^2 > S_2^2$$

Zgodnie z postaciami hipotezy alternatywnej populacje ustawia się tak, aby $S_1^2 > S_2^2$. Jeśli prawdziwa jest H_0 , to znaczy to statystyka postaci:

$$F = \frac{n_1 S_1^2 / (n_1 - 1)}{n_2 S_2^2 / (n_2 - 1)}$$

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

ma rozkład F-Snedecora o $v_1 = (n_1 - 1)$, $v_2 = (n_2 - 1)$ stopniach swobody. Obszar krytyczny określa się: $P\{F \geq F_\alpha\} = \alpha$, gdzie F_α jest wartością krytyczną odczytaną z tablic rozkładu F -Snedecora. Jeśli $F \geq F_\alpha$ to H_0 należy odrzucić, jeśli $F < F_\alpha$ nie ma podstaw do odrzucenia H_0 .

Często statystykę F buduje się wykorzystując zamiast wariancji s^2 wariancję \hat{s}^2 wyznaczaną według formuły:

$$\hat{s}^2 = \frac{n-1}{n} s^2$$

Wówczas statystyka F ma postać: $F = \frac{\hat{s}_1^2}{\hat{s}_2^2}$. Tak określona statystyka przy prawdziwości

H_0 posiada rozkład F -Snedecora z liczbą stopni swobody $v_1 = (n_1 - 1)$ i $v_2 = (n_2 - 1)$.

Test istotności dla wskaźnika struktury i dwóch wskaźników struktury.

Test istotności dla wskaźnika struktury.

MODEL 1.

Populacja generalna ma rozkład dwupunktowy z parametrem p (tj. frakcją elementów wyróżnionych). Z populacji tej wylosowano dużą próbę ($n > 100$). W oparciu o wyniki tej próby należy zweryfikować hipotezę:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

W oparciu o wyniki próby losowej wyznaczamy wskaźnik struktury: $\hat{p} = \frac{X}{n}$, gdzie X jest zmienną losową - liczbą wyróżnionych elementów w próbie. Statystyka W ma rozkład $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Jeśli H_0 jest prawdziwa tj. $p = p_0$, to W z próby ma asymptotyczny rozkład normalny $N\left(p_0; \sqrt{\frac{p_0(1-p_0)}{n}}\right)$

Testem, weryfikującym H_0 jest statystyka postaci:
$$U = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

która to statystyka posiada asymptotyczny rozkład $N(0;1)$. Obszar odrzucenia H_0 określa się: $P\{|U| \geq$

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

$$U = \frac{W_1 - W_2}{\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n}}}$$

Testem weryfikującym H_0 jest statystyka postaci:

Która posiada asymptotyczny rozkład normalny $N(0,1)$. Obszar odrzucenia H_0 określa się:

$P\{|U| \geq u_\alpha\} = \alpha$, gdzie $W_1 = \frac{X_1}{n}$, $W_2 = \frac{X_2}{n}$, $n = \frac{n_1 n_2}{n_1 + n_2}$, $\tilde{p} = \frac{m_1 + m_2}{n_1 + n_2}$. Dla jednostronnych hipotez alternatywnych: $H_1: p_1 > p_2$ lub $H_1: p_1 < p_2$ buduje się jednostronne obszary krytyczne odczytując wartość testów dla podwojonej wartości α .

Wybrane testy nieparametryczne

Nieparametryczne testy istotności służą do weryfikacji hipotezy dotyczącej rozkładu badanej cechy w populacji generalnej nie precyzując wartości parametrów tego rozkładu, ale nie tylko. Do testów nieparametrycznych zalicza się:

- testy losowości służą do weryfikacji hipotezy, że próba ma charakter losowy,
- testy niezależności służące do weryfikacji hipotezy o niezależności dwóch zmiennych losowych,
- testy zgodności służące do weryfikacji hipotezy o postaci funkcyjnej rozkładu populacji generalnej.

Z tej grupy testów będziemy zajmować się jedynie testami zgodności.

W grupie testów zgodności wyodrębnia się dwie podgrupy:

podgrupa 1. - testy zgodności służące do weryfikacji hipotezy o postaci funkcyjnej rozkładu populacji danej zwykle dystrybuantą.

podgrupa 2. - testy zgodności służące do weryfikacji hipotezy, że dystrybuanty dwóch lub więcej zmiennych losowych są identyczne, czyli, populacje mają taki sam rozkład.

Nazwa testów zgodności pochodzi stąd, że w pierwszej podgrupie testów, przy pierwszym z wymienionych typów hipotez, których weryfikacji służą, sprawdza się zgodność rozkładu empirycznego z próby z rozkładem hipotetycznym. W drugiej podgrupie sprawdza się zgodność dwóch lub więcej rozkładów empirycznych z próby.

Uwaga!

Testy nieparametryczne nie wymagają założeń co do rozkładu populacji generalnej.

Test zgodności χ^2 .

Sformułujmy hipotezę zerową:

$H_0: F(x) = F_0(x)$ co oznacza, że populacja generalna ma rozkład określony pewną dystrybuantą

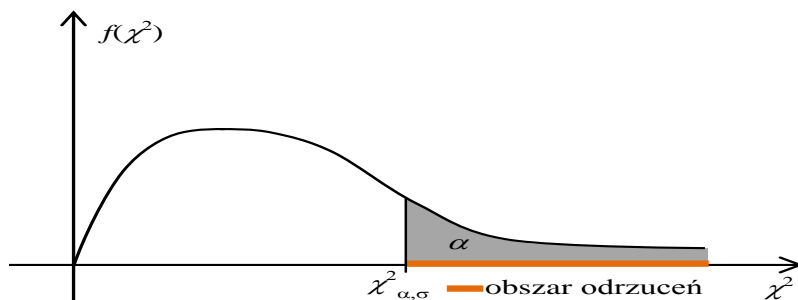
$F_0(X)$, wobec hipotezy alternatywnej:

$$H_1 : F(x) \neq F_0(x)$$

Z populacji losujemy dużą próbę (dużą, bo korzystać będziemy z rozkładu granicznego statystyki). Wyniki z próby porządkujemy w rozkład empiryczny w r -rozłącznych klas. Liczebności poszczególnych klas oznaczać będziemy jako n_i . Jeżeli H_0 jest prawdziwa, tzn. rozkład populacji generalnej opisany jest dystrybuantą $F_0(X)$, należy obliczyć prawdopodobieństwo że badana zmienna losowa przyjmie wartość z i -tej klasy. Gdyby hipoteza zerowa była prawdziwa, to liczebności w poszczególnych klasach powinny wynosić $n_i p_i$ - nazwijmy je umownie liczebnościami teoretycznymi, jako $\hat{n}_i = n_i p_i$; $i = 1, 2, \dots, r$, gdzie n - ogólna liczebność próby.

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

Podstawa, do konstrukcji miary zgodności rozkładu empirycznego z hipotetycznym jest różnica pomiędzy n_i i \hat{n}_i . Do oceny zgodności rozkładów stosuje się statystykę o postaci: $P\{\chi^2 \geq \chi_{\alpha,s}^2\} = \alpha$; która przy założeniu prawdziwości hipotezy zerowej ma asymptotyczny rozkład χ^2 o stopniach swobody $(r-k-1)$, gdzie r jest liczbą klas, natomiast k oznacza liczbę parametrów rozkładów, które zostały oszacowane na podstawie próby. Jeżeli prawdziwa jest H_0 , obszar krytyczny jest określony: $\chi_{\alpha,s}^2$, gdzie α - poziom istotności, s - wartość krytyczna wyznaczona z rozkładu . Jeżeli $\chi_{emp}^2 \geq \chi_{\alpha}^2$ tzn. że różnica jest statystycznie istotna i H_0 należy odrzucić. Obszar krytyczny $(\chi_{\alpha,s}^2, +\infty)$.



Testy zgodności stanowią dość liczną klasę testów istotności, i służą do sprawdzania hipotez nieparametrycznych - zwane są nieparametrycznymi testami istotności. Do tych testów zalicza się: test zgodności λ - Kołmogorowa, test zgodności Kołmogorowa - Smirnowa.

Test zgodności λ - Kołmogorowa służy do weryfikacji hipotezy, że populacja ma rozkład ciągły dany dystrybuantą $F_0(x)$.

$$H_0 : F(x) = F_0(X) \quad H_1 : F(x) \neq F_0(X)$$

Dla rozkładu empirycznego wyznacza się dystrybuantę empiryczną $F_0(x)$ a następnie określa się różnice pomiędzy dystrybuantą empiryczną i teoretyczną. Miarą zgodności dla dwóch dystrybuant jest statystyka:

$D = \sup_x |F_n(x) - F_0(x)|$ i kolejno na jej podstawie określa się statystykę: $\lambda = \sqrt{n}$. Tak określona statystyka przy prawdziwości H_0 ma asymptotyczny rozkład λ - Kołmogorowa. Obszar krytyczny określa się: $P\{\lambda \geq \lambda_\alpha\} = \alpha$. Jeśli $\lambda \geq \lambda_\alpha$, H_0 należy odrzucić, w przeciwnym przypadku brak podstaw do odrzucenia H_0 .

3.7. Przykłady empiryczne

Przykład 1.

Zmienna losowa X przyjmuje wartości: 2, 4, 6, 8 z jednakowymi prawdopodobieństwami. Określ funkcję rozkładu prawdopodobieństwa, wyznacz i zinterpretuj wartości parametrów. Wyznacz wartości dystrybuanty, sporządź jej wykres, podaj przykładowe interpretacje.

Przykład 2. Random variable X has the following mass probability function:

$X=x_i$	1	2	3	4	5	Total
$p(x_i)$	1/16	4/16	6/16	4/16	1/16	

1. Check, whether a given distribution is a probability distribution
2. Present graphically the probability mass function
3. Determine the distribution parameters and provide their interpretation
[$E(X)$; $D^2(X)$; $D(X)$; $V(X)$]
4. Determine a set of cumulative distribution functions, and make its graph
5. Determine the following probabilities:
 $P(X > 4)$; $P(X \leq 3)$; $P(X \geq 2)$; $P(2 \leq X < 5)$
6. Find the value of the following variables:
 X^2 ; $e^{(X)}$; $\log(1+X^2)$
7. Find the following expected values:
 $E[X^2]$; $E[(e)^X]$; $E[\log(1+X^2)]$.

Przykład 3 . The probability that a randomly chosen person in EU is 60 or older is approximately 0.2. Assuming, if X is the number of age 60+ in sample of 6:

1. Construct the probability mass function and it's plot
2. Determine a set of cumulative distribution functions, and make its graph
3. Determine the following probabilities:
 $P(X > 3)$; $P(X \leq 3)$; $P(X \geq 2)$; $P(2 < X \leq 6)$
4. Determine the distribution parameters based on definition

Przykład 4.

Bank's clients arriving to bank per minute on average.

1. Find the probability that in given minute exactly 3 people will arrive, assuming the value of $\lambda = 2$.
2. Generate the entire distribution for X and cumulative distribution function.
3. Get the entire distribution using Excel formula
4. Both functions presents in graphical forms.
5. Indicate the basic parameters of the distribution.
6. Calculate the values of the following probabilities: $P(X \leq 4)$; $P(X > 2)$; $P(X = 5)$.

Solution: Poisson distribution with $\lambda = 2$

Przykład 5.

Rzucamy 1 raz kością do gry. Interesuje nas wyrzucenie sześciu oczek. Prawdopodobieństwo sukcesu wynosi $1/6$, niepowodzenia $5/6$. Oznaczając sukces liczbą 1 a niepowodzenie liczbą 0 określ rozkład 0-1 – wyznacz dystrybuantę i parametry rozkładu.

Przykład 6.

Wadliwość wyrobu (tzn. przeciętny procent braków) wynosi 3%. Jakie jest prawdopodobieństwo, że w partii liczącej 100 sztuk znajdzie się nie więcej niż 2 sztuki złe.

Przykład 7.

Rozkłady miesięcznych wynagrodzeń pracowników pewnego banku w tys. zł mają w przybliżeniu rozkład normalny o parametrach $N(5,5; 2)$. Oblicz rachunkowo i przedstaw graficznie wartości następujących prawdopodobieństw:

$$P(X > 5,5); P(3,5 < X \leq 7,5); P(X \leq 8,5)$$

Przykład 8. Normal Distribution

The amounts of money request on home loan application at KMF Federal Savings follow the normal distribution with a mean of \$100.000 and standard deviation \$10.000. A loan applications is received last week. What is the probability:

1. The amount request is \$120.000 or more?
2. The amount request is between \$80.000 and \$120.000 ?
3. The amount request is \$70.000 or less ?
4. The amount request is $|X - m| \leq \$20.000$?

Calculate a value of probabilities

Przykład 9. Exponential Distribution

The time required to repair a computer in an office (Building M, WSE) is an exponential random variable with rate $\lambda = 0.5$ downs/hour.

1. What is the probability that a repair time exceeds 2 hours?
2. What is the probability that the repair time will take at least 4 hours given that the repair computer man has been working on the computer for 3 hours?

Przykład.10.

Czas czekania w kolejce w banku jest zmienną losową o rozkładzie normalnym z odchyleniem standardowym 2 min. Połowa klientów czeka na obsługę poniżej 6,9 min. Oblicz prawdopodobieństwo stania w kolejce dłużej niż średni czas oczekiwania i jednocześnie krócej niż 10, 5 min.

Przykład.11.

Zmienna losowa X w populacji generalnej ma rozkład normalny $N(100,50)$. Z populacji tej wylosowano niezależnie n elementową próbę losową $n=25$. Określ rozkład średniej jako statystyki z próby i podaj jego parametry.

Przykład.12.

Na 800 zbadanych wypadków drogowych okazało się, że 320 zostało spowodowanych nadmierną szybkością. Na poziomie $\alpha = 0,05$ zweryfikuj hipotezę, że % wypadków spowodowanych nadmierną szybkością jest równy 35%.

Przykład 13.

Z jakim prawdopodobieństwem oszacowano frakcję (odsetek) dorosłych Polaków, którzy uważają, że polityka gospodarcza rządzącej koalicji jest właściwa, jeśli otrzymany przedział ma końce: 0,777 - 0,822, a oszacowany został na podstawie 1200 elementowej próby losowej, wśród której 960 respondentów miało zdanie na tak.

Czy liczebność 1200 osób w próbie byłaby konieczna do oszacowania przedziału ufności, gdyby przyjąć maksymalny błąd szacunku równy 3%.

Przykład 14.

Jak liczna powinna być próba dokumentów księgowych pobranych do kontroli przez biegłego w czasie badania bilansu firmy, aby przy współczynniku ufności 0,95 oszacować % dokumentów wadliwych, jeśli nie chcemy pomylić się o więcej niż o 4%?

Przykład 15.

Postanowiono oszacować frakcję klientów banku BIGCITY, którzy są niezadowoleni z poziomu świadczonych usług.

a) Ilu klientów winna liczyć próba, aby oszacować odsetek klientów z błędem bezwzględnym nie większym niż 0,10 (przy poziomie ufności 0,90)?

b) Czy odpowiedź zmieni się, jeśli wcześniej przeprowadzono badanie pilotażowe, w którym wstępnie oszacowano ?

Przykład 16.

Mean – point and interval estimates, hypothesis testing

You are given an Excel file that contains information on daily average exchange rate of euro to polish zloty (ERU/PLN) for period 2nd January – 10th September 2013 published by the Polish National Bank (data available at <http://www.nbp.pl/home.aspx?c=/ascx/archa.ascx>). Below, there is a part of the file presented (in total, there are 175 observations):

	A	B
1	Date	1 EUR
2	20130102	4.0671
3	20130103	4.0770
4	20130104	4.1248
5	20130107	4.1218
6	20130108	4.1263
7	20130109	4.1192
8	20130110	4.0760
9	20130111	4.0996
10	20130114	4.1231
11	20130115	4.1151
12	20130116	4.1280
13	20130117	4.1178
14	20130118	4.1294
15	20130121	4.1762
16	20130122	4.1700
17	20130123	4.1591
18	20130124	4.1964
19	20130125	4.1903
20	20130128	4.1805
21	20130129	4.1969

Empirical

results:

$$n = 175$$

$$\bar{x} = 4.2$$

$$\sigma = 0.0649$$

Calculate the point estimate of the mean of the daily average exchange rate.

Calculate the interval estimate of the mean of the daily average exchange rate, assuming $P = 1 - \alpha = 0.95$.

What is the maximal absolute error of the mean estimation?

What is the minimal sample size that will ensure maximal absolute error at the level of 0.005 with $P = 0.95$?

Verify the hypothesis that the daily average exchange rate of EUR/PLN is equal to 4 against the alterative hypothesis stating that it is higher than 4, assuming significance level equal to 0.05.

Assumption: For the calculations assume the large sample (as $n > 30$) and unknown standard deviation in the population.

Przykład 17. Proportion – point and interval estimates, hypothesis testing

A pharmaceutical company conducted a clinical trial to find out if the new drug is effective. The company enrolled 300 patients and 240 patients were cured. Based on the information: calculate point estimate of the proportion of patients in population that would be cured with the new drug.

Calculate interval estimate of the proportion of patients that would be cured with the new drug, assuming $P = 1 - \alpha = 0.95$

Is the sample of 240 patients sufficient in order to estimate with 2% precision the proportion of patients in population that would be cured with the new drug, knowing that the results from the sample provided information at the level $w_i = 0.8$ with $P = 0.90$

Verify the hypothesis that the new drug will cure the illness in 60% of the patients in whole population assuming significance level 0.05.

Przykład 18.

Based on the random sample of printed books the Printing Office presents the following information about the number of mistakes per page. Fit a Poisson distribution and test of goodness of fit.

Number of mistakes per page	0	1	2	3	4	Total
Number of pages	200	100		20	5	325

Przykład 19.

The distribution of the rent for studio apartment is $N(1\,000\text{ PLN}; 100\text{ PLN})$ in city A and $N(1\,200\text{ PLN}; 200\text{ PLN})$ in city B. Calculate the probability that in the random sample of 20 studio apartments in city A and 25 such apartments in city B:

- the average rent in city A is at least 100 PLN lower than the average rent in city B
- the variance of rent in city B does not exceed $10\,000\text{ (PLN)}^2$?

Part of solution:

Ad. 1.

X_A – rent for studio apartment in city A (in a population), $X_A \sim N(1\,000; 100)$

X_B – rent for studio apartment in city B (in a population), $X_B \sim N(1\,200; 200)$

\bar{X}_A – average rent for studio apartment in city A (in a sample, $n_A = 20$)

\bar{X}_B – average rent for studio apartment in city B (in a sample, $n_B = 25$)

$P(\bar{X}_A + 100 \leq \bar{X}_B)$ – the probability that the average rent in the sample for city A is at least 100 PLN lower than in city B

Both populations are normal, with known standard deviations and difference in both sample means is normally distributed:

Przykład 20.

Przy kontroli pracy dwóch central telefonicznych w losowo dobranym dniu stwierdzono, że na 200 połączeń w centrali A 16 było pomyłkowych, natomiast na 100 połączeń w centrali B złych połączeń było 10. Na poziomie $\alpha=0,05$ zweryfikuj hipotezę, że % złych połączeń jest jednakowy w obu centralach.

Przykład 21.

W dwóch grupach firm: krajowych i zagranicznych przeprowadzono badanie zadłużeń z tytułu kredytów na działalność eksportową. Dla losowych prób: $n_1 = 150$ firm krajowych i $n_2 = 100$ firm zagranicznych otrzymano następujące informacje o wysokości zadłużenia:

firmy krajowe

średnia = 15,8 mln. zł

 $S_1(x) = 3,5$ mln zł

firmy zagraniczne

średnia = 35,0 mln zł

 $S_2(x) = 15,0$ mln zł

Czy na poziomie $\alpha = 0,01$ można uznać, że średni poziom zadłużenia w obu typach firm jest taki sam?

Przy jakim innym poziomie istotności decyzja weryfikacyjna może ulec zmianie?



Notatki