Informacja wstępna

- Wykłady – 16 godzin; zajęcia laboratoryjne - 30 godzin.
- Forma zaliczenia wykładu: egzamin.
- Zaliczenie laboratoriów jest warunkiem dopuszczenia do egzaminu.
- Liczba punktów ECTS za kurs: **4**.

Spis literatury

1. Janina Jóźwiak, Jarosław Podgórski, *Statystyka od podstaw*, Polskie wydawnictwo ekonomiczne, Warszawa, 1998
2. Andrzej Stanisz. *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, Tom I,II,III, Kraków, 2007, StatSoft.
3. Tomasz Panek, *Statystyczne metody wielowymiarowej analizy porównawczej*, Szkoła Głowna handlowa w Warszawie, 2009.
4. Larose D., *Odkrywanie wiedzy z danych. Wprowadzanie do eksploracji danych*, PWN, 2006.
5. Mieczysław Korzyński, *Metodyka eksperymentu. Planowanie, realizacja i statystycznie opracowanie wyników eksperymentów technologicznych*, Wydawnictwo naukowo-techniczne Warszawa, 2006.

WSTĘP

Analiza danych — proces badania, filtracji, przekształcenia i modelowania danych w celu wyciągania cennej informacji i podjęcia decyzji.

Analiza danych ma mnóstwo aspektów i podejść, obejmuje różne metody w różnych dziedzinach nauki.

Jest to wieloprzedmiotowa dziedzina, która powstała i rozwija się na podstawie takich nauk jak statystyka, rozpoznawanie obrazów, eksploracja danych, sztuczna inteligencja, teoria baz danych i in.

Analiza danych

- nie może dać odpowiedzi na te pytania, które nie zostały zadane;
- nie potrafi zamienić analityka lecz tylko daje mu potężne narzędzie do ułatwienia i polepszenia jego pracy;
- różne narzędzia do analizy danych wymagają pewnej kwalifikacji użytkownika, więc oprogramowanie musi odpowiadać poziomowi przygotowywania użytkownika;
- potrzebny jest staranny wybór modelu i interpretacja zależności albo szablonów, które zostaną odzyskane, więc wymagana jest współpraca między ekspertem w dziedzinie przedmiotowej i specjalistą w dziedzinie analizy danych;
- pomyślna analiza wymaga jakościowej wstępnej obróbki danych.

Dane (1)

Dane to nieobrobiony materiał, który przedstawiony jest przez dostawców danych i wykorzystywany przez konsumentów w celu „kształtowania” informacji z wykorzystaniem tych danych.

- nieograniczona pojemność;
- różnorodność;
- konkretność i zrozumiałość wyników;
- proste narzędzia do obróbki danych.

Istnieje wiele różnych klasyfikacji rodzajów danych. W zależności od tego, w jakim celu zostały zgromadzone dane, można ich podzielić na ***pierwotne*** i ***wtórne***.

Dane (2)

Dane pierwotne – oryginalne informacje zbierane w ściśle określonym celu.

Zaletą danych uzyskanych ze źródeł pierwotnych jest ich wysoka skuteczność wynikająca z określenia celu odnoszącego się bezpośrednio do problemu badania natomiast wadą jest wysoki koszt i czasochłonność.

Dane wtórne – to takie, które już istnieją, zostały zebrane i opracowane w innym celu.

Zaletą jest ich niski koszt, a wadą ograniczona dokładność i dostosowanie do potrzeb, co wynika z odmiennego celu, dla którego były gromadzone.

Dane (3)

Rozróżniają dane ***jakościowe*** i ***ilościowe***.

Znaczna ilość informacji naukowej jest zapisywana w postaci liczb, co pozwala manipulować takimi danymi z użyciem statystycznych metod matematycznych. Takie dane są ***ilościowe***.

Istnieje ważna informacją, którą nie można zredukować do postaci liczb. Takie dane nazywa się ***jakościowe***. Werbalne pojęcia i relacje między nimi są mniej dokładne niż liczby i odpowiednie łącza.

W przeciwieństwie do badań ilościowych, w badaniach jakościowych nie istnieje powszechnie akceptowanej analizy danych.

Dane (4)

Większość surowych danych jest niejednorodna, niekompletna i zaszumiona. Aby dane były przydatne do celów analizy, muszą one przejść przez wstępną obróbkę.

ID Klienta	Kod pocztowy	Płeć	Dochód	Wiek	Stan cywilny	Kwota transakcji
1001	10048	M	75 000	C	M	5000
1002	J2S7K7	K	–40 000	40	W	4000
1003	90210		10 000 000	45	S	7000
1004	6269	M	50 000	0	S	1000
1005	55101	K	99 999	30	R	3000

Dane brakujące

Zastąpienie brakujących wartości:

- pewną stałą, określoną przez analityka;
- wartością średnią lub modalną;
- wartością wygenerowaną losowo z obserwowanego rozkładu zmiennej.

Zastępowanie brakujących wartości to ryzykowne przedsięwzięcie, a zyski muszą być skalkulowane z uwzględnieniem możliwego zafałszowania wyników.

Punkty oddalone

Punkty oddalone są skrajnymi wartościami, które znajdują się blisko granic zakresu danych lub są sprzeczne z ogólnym trendem pozostałych danych.

Metody identyfikacji punktów oddalonych:

- graficzna: histogramy, wykresy;
- standaryzacja;
- metoda rozstępu międzykwartylowego (*IQR*), który oblicza się jako różnica trzeciego i pierwszego kwartyli i może być zinterpretowane jako środkowe 50% danych; wartość danych jest punktem oddalonym, jeżeli jest położona przynajmniej o $1{,}5 \cdot IQR$ poniżej kwartyla pierwszego lub jest położona przynajmniej o $1{,}5 \cdot IQR$ powyżej kwartyla trzeciego.

Klasyfikacja (1)

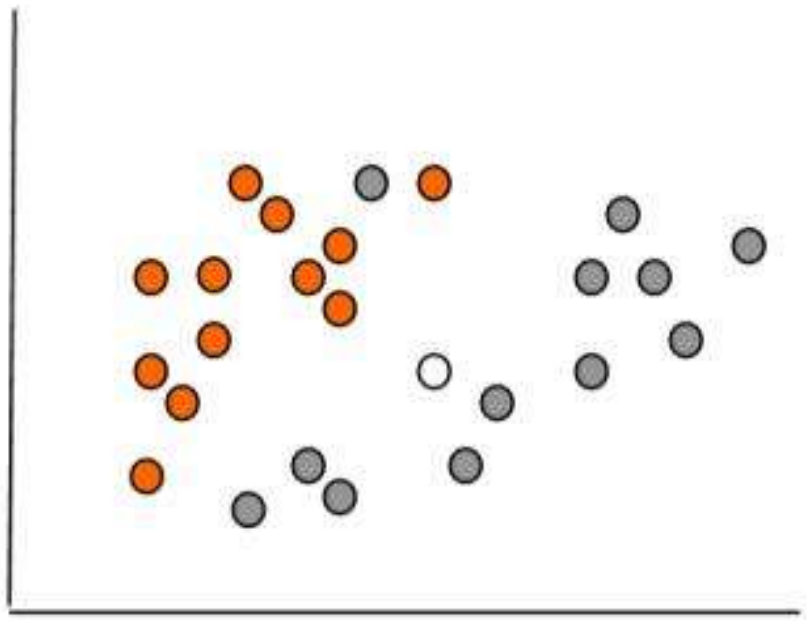
Klasyfikacja – sformalizowane zadanie, w którym jest zbiór obiektów podzielonych w pewny sposób na klasy.

Dany jest skończony zbiór obiektów, dla których wiadomo, do jakich klas one należą. Zbiór ten nazywa się próbką. Dla pozostałych obiektów nie wiadomo, do jakich klas one należą. Trzeba zbudować algorytm, który pozwoli klasyfikować dowolny obiekt z danego zbioru, czyli dokonać podział na grupy.

Klasyfikować obiekt oznacza to wskazać numer (lub nazwę) klasy, do której należy dany obiekt.

Klasyfikacja (2)

Klient	Wiek	Dochód	Klasa
1	18	25	1
2	22	100	1
3	30	70	1
4	32	120	1
5	24	15	2
6	25	22	1
7	32	50	2
8	19	45	2
9	22	75	1
10	40	90	2



Klasyfikacja (3)

W większości metod klasyfikacji zbiór danych początkowych rozbija się na dwa zbiory: ***uczący*** i ***testowy***.

Zbiór uczący wykorzysta się dla nauczania (budowania) modelu.

Zbiór testowy wykorzysta się w celu sprawdzania wiarogodności zbudowanego modelu. Podział na uczące i testowe zbiory wykonuje się za pomocą podziału próbki w pewnych proporcjach, np., zbiór uczący - dwie trzecie części danych i testowy - jedna trzecia część danych.

Jeżeli zaś próbka jest mała należy stosować specjalne metody, przy wykorzystaniu których zbiory uczący i testowy mogą częściowo się łączyć.

Klasteryzacja

Klasteryzacja – zadanie podziału próbki obiektów na podzbiory, zwane grupami (klasterami, klasami) w taki sposób, aby każda grupa zawierała podobne obiekty, a obiekty różnych grup znacznie się różniły pomiędzy sobą. Zadanie to jest podobne do klasyfikacji, lecz różni się tym, że klasy badanego zbioru danych nie są określone z góry.

Należy podkreślić, że w rezultacie wykorzystania różnych metod klasteryzacji mogą być otrzymane klastery różnych form, czyli mogą być otrzymane różne rezultaty, co jest właściwością działania tego lub owego algorytmu. Daną właściwość należy uwzględniać przy wyborze metody klasteryzacji.

Predykcja

Predykcja (***prognozowanie***) – zadanie przewidywania tego, w jaki sposób będą przebiegać w przyszłości procesy lub zdarzenia.

Ogólnie zadanie prognozowania doprowadza się do rozwiązania 2 zadań: ***wybór modelu prognozowania*** oraz ***analiza precyzyjności*** zbudowanej prognozy.

Wracając do przykładu o agencji turystycznej można powiedzieć, że określenie klasy do której należy klient jest rozwiązaniem zadania klasyfikacji, a prognozowanie dochodu, który przyniesie klient w przyszłości jest rozwiązaniem zadania prognozowania.

Wizualizacja (1)

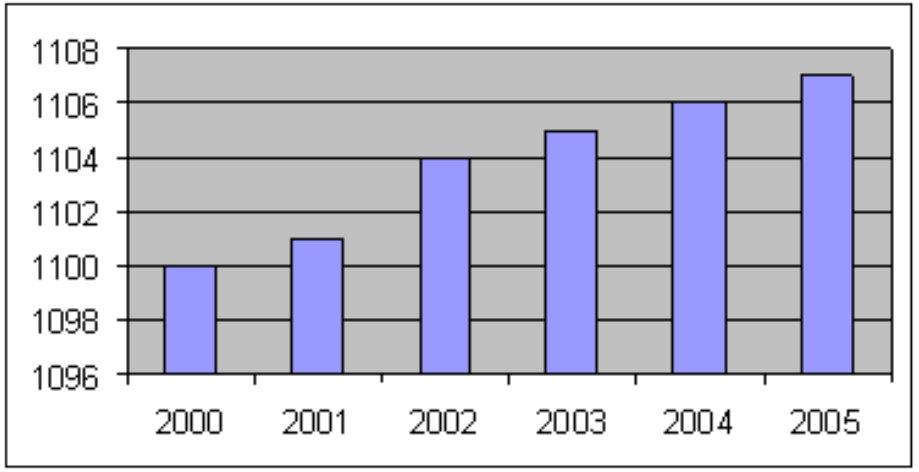
Wizualizacja – narzędzie, które pozwala obserwować końcowy rezultat obliczeń, organizować sterowanie procesem obliczeniowym i nawet cofnąć się z powrotem do danych początkowych, aby określić najlepszy kierunek dalszych działań.

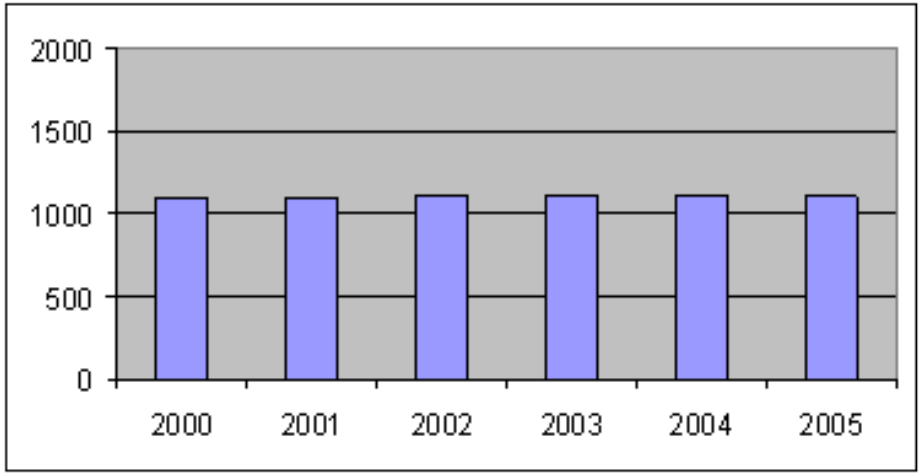
Schematy, diagramy, histogramy, wykresy itd.

Zaletą jest prawie zupełny brak specjalnego przygotowania użytkownika.

Wizualizacja (2)

Rok	Zysk
2000	1100
2001	1101
2002	1104
2003	1105
2004	1106
2005	1107





Czym jest statystyka? (1)

Statystyka to zbiór metod służących pozyskiwaniu, prezentacji i analizie danych. Ostatecznym celem stosowania tych metod jest otrzymanie, na podstawie zbioru danych, użytecznych, uogólnionych informacji na temat zjawiska, którego dane dotyczą.

Pozyskiwanie danych to proces zwany ogólnie ***badaniem statystycznym***, w ramach którego dokonuje się obserwacji statystycznej (pomiarów lub zliczania).

Prezentowanie danych to przedstawienie licznych zbiorów danych w postaci ułatwiającej ich ocenę i analizę. Można tu zastosować różne formy prezentacji, tabelaryczne i graficzne.

Czym jest statystyka? (2)

Podstawowe zadanie statystyki to jednak ***analiza*** i ***interpretacja*** danych. Analiza może sprowadzać się do sumarycznego opisu zbioru danych. Wykorzystywane do tego środki określa się mianem ***metod opisu statystycznego***. W wielu przypadkach zebranie wszystkich potencjalnych danych nie jest możliwe i należy wypowiadać się o badanym zjawisku na podstawie zebranych w odpowiedni sposób danych częściowych. Jest to przedmiotem tzw. ***statystyki matematycznej***, posiłkującej się metodami rachunku prawdopodobieństwa.

Populacja generalna

W statystyce matematycznej przyjęto się określać zbiorowość statystyczną mianem ***populacji generalnej*** lub ***zbiorowości generalnej***.

Jeśli zbiór elementów populacji generalnej jest skończony, to określamy ją jako ***skończoną***.

Jeśli zbiór elementów populacji jest nieskończony, to określamy ją jako ***nieskończoną***. Koncepcja populacji generalnej nieskończonej jest na ogół wynikiem myślenia teoretycznego i dotyczy raczej zjawisk, niż obiektów materialnych.

W praktyce zdarza się też, że zbiorowość generalna, chociaż skończona, jest tak liczna, że wygodniej jest traktować ją jako nieskończoną.

Cecha statystyczna

Elementy populacji generalnej mogą mieć różne właściwości, które podlegają obserwacji statystycznej. Właściwości te nazywamy ***cechami statystycznymi***.

Niektóre z tych właściwości mają charakter ilościowy (np., wiek, waga), nazywamy je cechami ***mierzalnymi***, inne mają charakter jakościowy (np., płeć, kolor oczu) i je nazywamy ***niemierzalnymi***.

Rozkład cechy

Elementy populacji generalnej różnią się na ogół między sobą wartościami rozpatrywanej cechy statystycznej; można więc mówić o ***rozkładzie cechy*** w populacji.

Celem badania jest na ogół poznanie rozkładu interesującej nas cechy w populacji generalnej oraz uzyskanie informacji o wartości charakterystyk (parametrów) tego rozkładu.

Podzbiór elementów populacji generalnej podlegających badaniu określa się mianem ***próby***.

Badanie próby

Wnioskowanie statystyczne: uogólnienie uzyskanych wyników na całą populację oraz szacowanie wielkości popełnionych przy tym błędów.

Zasadnicze typy problemów:

- ***estymacja*** (szacowanie) nieznanych wartości parametrów rozkładu cechy w populacji
- ***sprawdzanie słuszności hipotez***, dotyczących wartości parametrów rozkładu cechy w populacji, bądź postaci tego rozkładu.

Losowy dobór próby

Próba odzyskana w wyniku wylosowania, gdy prawdopodobieństwo określonego na danym etapie elementu nie zależy od wyników wcześniejszych etapów losowania nosi nazwę ***próba prosta***.

W odniesieniu do populacji nieskończonych pobranie próby prostej odbywa się przez przeprowadzenie serii *n* niezależnych eksperymentów z zachowaniem identycznych, dla każdego eksperymentu, warunków.

Rozkład empiryczny

Podstawą analiz statystycznych badanej cechy jest określenie tzw. ***empirycznego rozkładu cechy***, które polega na przyporządkowaniu uszeregowanym rosnąco wartościom, przyjmowanym przez tę cechę, odpowiednio zdefiniowanych częstości ich występowania.

Indywidualne wartości cechy X:

$$x_j, j = 1,2,...,n,$$

gdzie n jest liczebnością badanej zbiorowości (tzn. liczba jednostek lub pomiarów).

Przykład 1 (1)

0,3,1,1,2,2,0,0,3,5,0,1,2,2,1,1,0,1,1,1.

Badana cecha X - liczba błędów na jednej stronie tekstu, przyjmuje wartości całkowite 0,1,2,... Cechy tego typu, czyli cechy o wartościach ze zbioru przeliczalnego, nazywamy ***skokowymi***.

Wartości x_i cechy X uporządkowane niemalejące:

$$x_{\min} = x_1 < x_2 < \ldots < x_n = x_{\max}.$$

Liczbę jednostek zbiorowości, dla których cecha X przyjmuje wartość x_i, oznaczać będziemy n_i. Suma takich cząstkowych liczebności jest równa liczebności zbiorowości:

$$\sum n_i = n.$$

Przykład 1 (2)

Gdy poszczególnym wartościom x_i cechy X przyporządkowane są liczebności n_i, to w ten sposób określony jest ***rozkład empiryczny***.

Częstości:

$$p_i = \frac{n_i}{n}.$$

Dane uporządkowane i pogrupowane nazywane są ***szeregiem rozdzielczym***.

Przykład 1 (3)

liczba błędów x_i	liczba stron n_i	częstość stron p_i
0	5	0,25
1	8	0,40
2	4	0,20
3	2	0,10
4	0	0
5	1	0,05
$\sum$	20	1,00

Cecha ciągła

Badana cecha, która może przyjmować wartości rzeczywiste (wartości ze zbioru nieprzeliczalnego), określa się jako ***ciągła***.

W praktyce: cecha, przyjmująca dużo wartości.

Przedziałami klasowymi – przedział wartości cechy z przyporządkowaną liczebnością.

$\left(x_{i\max} - x_{i\min}\right)$ - ***rozpiętość przedziału*** (***szerokość klasy***).

Środek i-tego przedziału:

$$\overset{\circ}{x}_i = \frac{x_{i\min} + x_{i\max}}{2}.$$

Przykład 2

15, 37, 34, 9, 61, 24, 56, 52, 6, 35, 21, 46, 86, 40, 74, 39, 48, 55, 73, 92, 43, 78, 67, 30, 29

czas obsługi (w s) $y_{i\min} - y_{i\max}$	liczba klientów n_i	częstość klientów p_i
0 – 20	3	0,12
20 – 40	9	0,36
40 – 60	6	0,24
60 – 80	5	0,20
80 – 100	2	0,008
$\sum$	25	1,00

Dystrybuanta empiryczna

$$F_n(x) = \begin{cases} 0 & x < x_i \\ \sum_i p_i & x_i \le x < x_{i+1}, \quad i = 1,2,\ldots,(k-1), \\ 1 & x \ge x_k \end{cases}$$

gdzie $i = 1,2,\ldots,k$ - liczba klas.

Dystrybuanta empiryczna (prawdopodobieństwo skumulowanie) jest funkcją niemalejącą w przedziale $[0,\ 1]$.

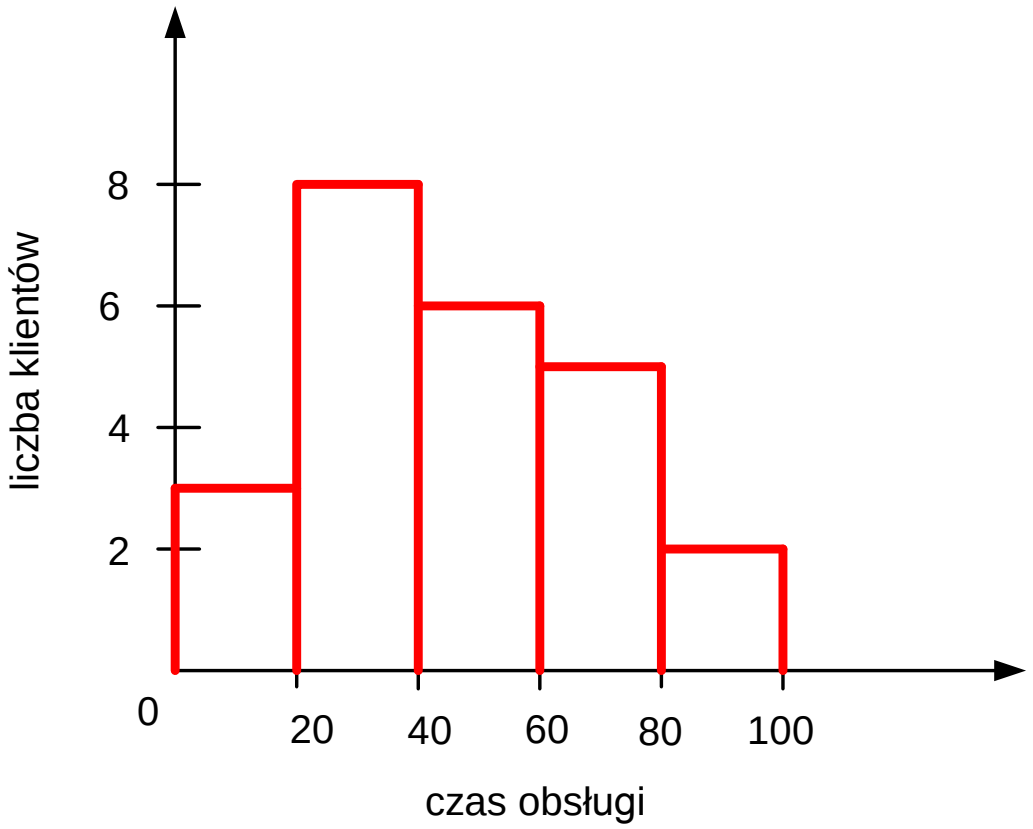
Przykład 2 cd

czas obsługi $y_{i\min} - y_{i\max}$	liczba klientów n_i	skumulowana liczba klientów $n(y_{i\max})$	częstość klientów p_i	dystrybuanta empiryczna $F_n(y_{i\max})$
0 – 20	3	3	0,12	0,12
20 – 40	9	12	0,36	0,48
40 – 60	6	18	0,24	0,72
60 – 80	5	23	0,20	0,92
80 - 100	2	25	0,008	1,00

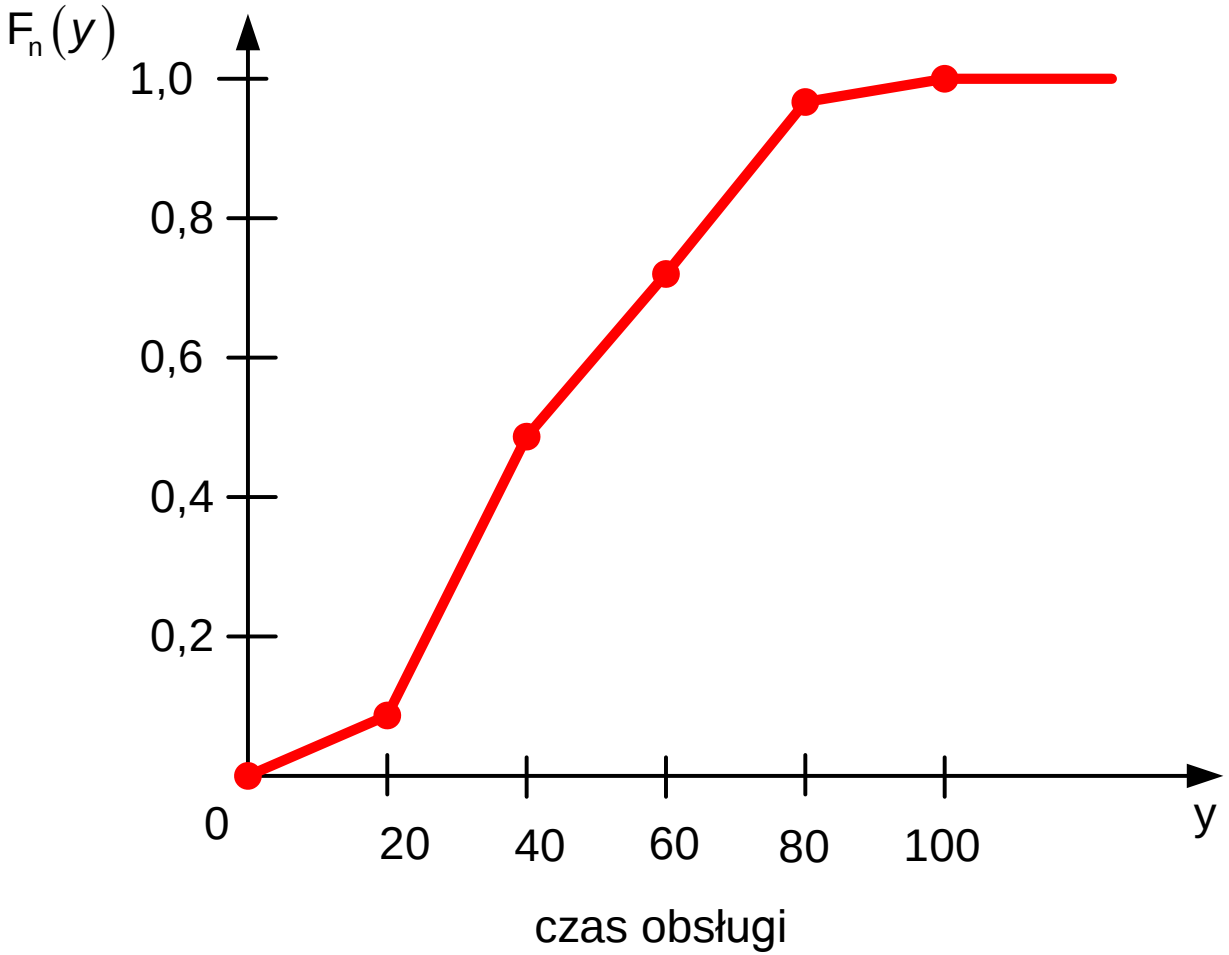
Np., $F(40) = 12$ oznacza, że 12 osób było obsługiwanych przy kasie co najwyżej 40 sekund, $F(60) = 0{,}72$ oznacza, że 72% osób było obsługiwane nie dłużej, niż 60 sekund.

Historgram

Histogram jest to zbiór prostokątów, których podstawy wyznaczony są na osi odciętych przez poszczególne podziały klasowe, natomiast wysokości są określone na osi rzędnych przez liczebności (lub częstości) odpowiadające poszczególnym przedziałom klasowym



Dystrybuanta empiryczna



Statystyki i parametry

Charakterystyki, opisujące w sposób syntetyczny właściwości rozkładu badanej cechy określają się jako ***statystyki***, gdy analizowane są dane próby losowej, lub jako ***parametry***, gdy analizowane są dane pełnej populacji.

Miary opisu rozkładu można podzielić na: miary położenia, miary zróżnicowania i miary asymetrii.

Średnia arytmetyczna

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j ,$$

gdzie $x_j, j = 1,..,n$ - indywidualne obserwacje w zbiorze danych, n - liczba obserwacji

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i ,$$

gdzie $x_i, i = 1,..,k$ są wyróżnionymi wartościami w rozkładzie, n_i - odpowiednie liczebności klasowe

Średnia arytmetyczna jest pewną abstrakcyjną wielkością i może przyjmować wartości w zbiorowości nie występujące.

Przykład 2 cd

$y_{i\min} - y_{i\max}$	$\overset{\circ}{y}_i$	n_i	$\overset{\circ}{y}_i n_i$
0 – 20	10	3	30
20 – 40	30	9	270
40 – 60	50	6	300
60 – 80	70	5	350
80 – 100	90	2	180
$\sum$	$\times$	25	1130

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} \overset{\circ}{y}_i n_i = \frac{1}{25}\left(30 + 270 + 300 + 350 + 180\right) = \frac{1130}{25} = 45{,}2$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{25}\left(15 + 37 + 34 + .. + 30 + 29\right) = \frac{1150}{25} = 46$$

Wariancja

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2$$

- dla zbioru danych indywidualnych.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_j - \bar{x}\right)^2 n_i$$

- w przypadku szeregu rozdzielczego.

Odchylenie standardowe:

$$s = \sqrt{s^2}\,.$$

$$s^2 = \frac{1}{25-1}\left[(15-46)^2 + (37-46)^2 + \ldots + (30-46)^2 + (29-46)^2\right] =$$

$$= 548{,}5$$

$$s = \sqrt{548{,}5} = 23{,}42$$

Kwantyl

Kwantylem rzędu p $(0 < p < 1)$ w rozkładzie empirycznym nazywamy taką wartość x_p cechy X, dla której jako pierwszej dystrybuanta empiryczna spełnia warunek:

$$F_n(x_p) \geq p.$$

Inaczej mówiąc taka wartość x_p, przed którą znajduje się $(100 \cdot p)\%$ elementów próbki.

Gdy $p = 0{,}25$, $p = 0{,}5$, $p = 0{,}75$, to kwantyle odpowiednio: ***dolny*** (pierwszy), ***środkowy*** (mediana, drugi), ***górny*** (trzeci).

Jeśli $p = 0{,}25$, a $x_{0{,}25} = 27$, to oznacza, że w rozkładzie empirycznym 25% wartości cechy są mniejsze bądź równe 27.

Dominanta

Dominantą (modą) w rozkładzie empirycznym nazywamy wartość cechy występującą w tym rozkładzie najczęściej, tzn. wartość, której odpowiada najwyższa liczebność (częstość).

Postać standaryzowana (1)

$$u = \frac{x - \bar{x}}{s}$$

Otrzymane w wyniku standaryzacji wartości wskazują, o ile odchyleń standardowych różnią się wartości cechy od średniej arytmetycznej.

Postać standaryzowana (2)

Średnia liczba punktów, zdobytych przez kandydatów zdających egzaminy wstępne na pewną uczelnię wyniosła 72, natomiast odchylenie standardowe 6 punktów.

Wartości standaryzowane, odpowiadające liczbie punktów: 60, 80, 72:

$$u = \frac{60-72}{6} = -2,\ u = \frac{80-72}{6} = 1\frac{1}{3},\ u = \frac{72-72}{6} = 0.$$

Liczba punktów 60 jest o dwa odchylenia standardowe niższa od średniej, liczba punktów 80 jest wyższa od średniej o $1\frac{1}{3}$ odchylenia standardowego, liczba punktów 72 pokrywa się ze średnią.

Postać standaryzowana (3)

Dla każdego zbioru danych, średnia dla wszystkich zmiennych standaryzowanych wynosi 0, a odchylenie standardowe 1.

Wartości standaryzowane u mogą wskazywać na nietypowe wartości cechy w zbiorze, zwane ***wartościami odstającymi***.

W praktyce wartość odstająca: $|u| > 3$.

Czyli wartości różniące się od średniej o więcej, niż 3 odchylenia standardowe.

Istnieje wiele miar opisu empirycznego rozkładu cechy, takich jak: współczynnik zmienności, rozstęp, asymetria, skośność. (Do samodzielnego zapoznania się).

Parametry rozkładu cechy w populacji

Przypuśćmy, że $x_1, x_2, ..., x_N$ są wartościami badanej cechy w skończonej populacji, gdzie liczbę elementów w populacji oznaczono przez N.

$$m = \frac{1}{N}\sum_{j=1}^{N} x_j$$ - średnia w populacji.

Odchylenie standardowe w populacji:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\left(x_j - m\right)^2 \qquad \sigma = \sqrt{\sigma^2}$$

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j - m\right)^2 \qquad S = \sqrt{S^2}$$

Zmienna losowa (1)

Podstawową metodologią statystyczną jest ***wnioskowanie o populacji generalnej na podstawie próby***. Najpierw jednak trzeba dysponować modelami, które mogą opisywać badaną zbiorowość. Takim bardzo ogólnym modelem do opisu zachowania się cechy w populacji jest tzw. ***zmienna losowa***. Jest to wielkość, która w wyniku „doświadczenia" przyjmuje różne wartości, przy czym przed doświadczeniem nie jesteśmy w stanie określić z absolutną pewnością, jaka wartość właśnie się pojawi (zrealizuje). Jesteśmy w stanie co najwyżej podać zbiór możliwych wartości, jakie mogą się pojawić, oraz odpowiadające im prawdopodobieństwa.

Zmienne losowe są oznaczane dużymi literami, a wartości, które zmienne losowe przyjmują – małymi literami.

Zmienna losowa (2)

Funkcja, opisująca sposób przyporządkowania prawdopodobieństw poszczególnym wartościom zmiennej losowej nazywa się ***rozkładem prawdopodobieństwa***.

Zmienne losowe: ***skokowe*** i ***ciągłe***.

Dystrybuanta podaje prawdopodobieństwo tego, że zmienna losowa przyjmuje wartość mniejszą od zadanej liczby:

$$F(x) = P(X \leq x).$$

Dystrybuanta jest niemalejąca funkcją zmiennej losowej X taka, że:

$$\text{jeśli } x \to -\infty, \text{ to } F(x) = 0,$$

$$\text{jeśli } x \to \infty, \text{ to } F(x) = 1.$$

Zmienna losowa (3)

Funkcja rozkładu prawdopodobieństwa (zmienna skokowa):

$$P\left(X = x_i\right) = p_i,$$

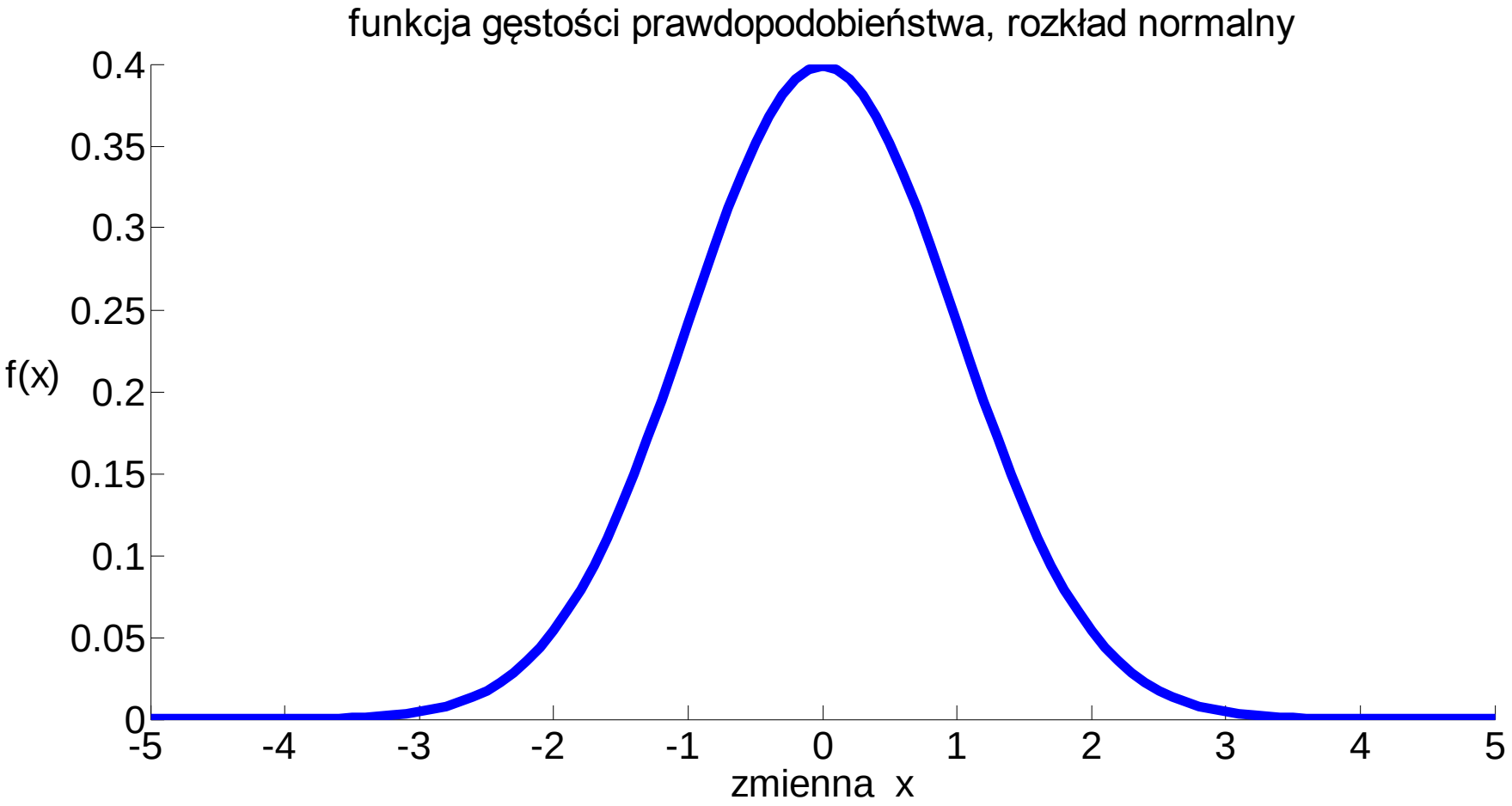
Funkcja gęstości prawdopodobieństwa (zmienna ciągła):

$$f\left(x\right) = \frac{dF\left(x\right)}{dx}.$$

Znając gęstość prawdopodobieństwa, można łatwo obliczyć dystrybuantę:

$$F\left(x\right) = \int_{-\infty}^{x} f\left(x\right)dx.$$

Rozkład normalny (1)



Rozkład normalny (2)

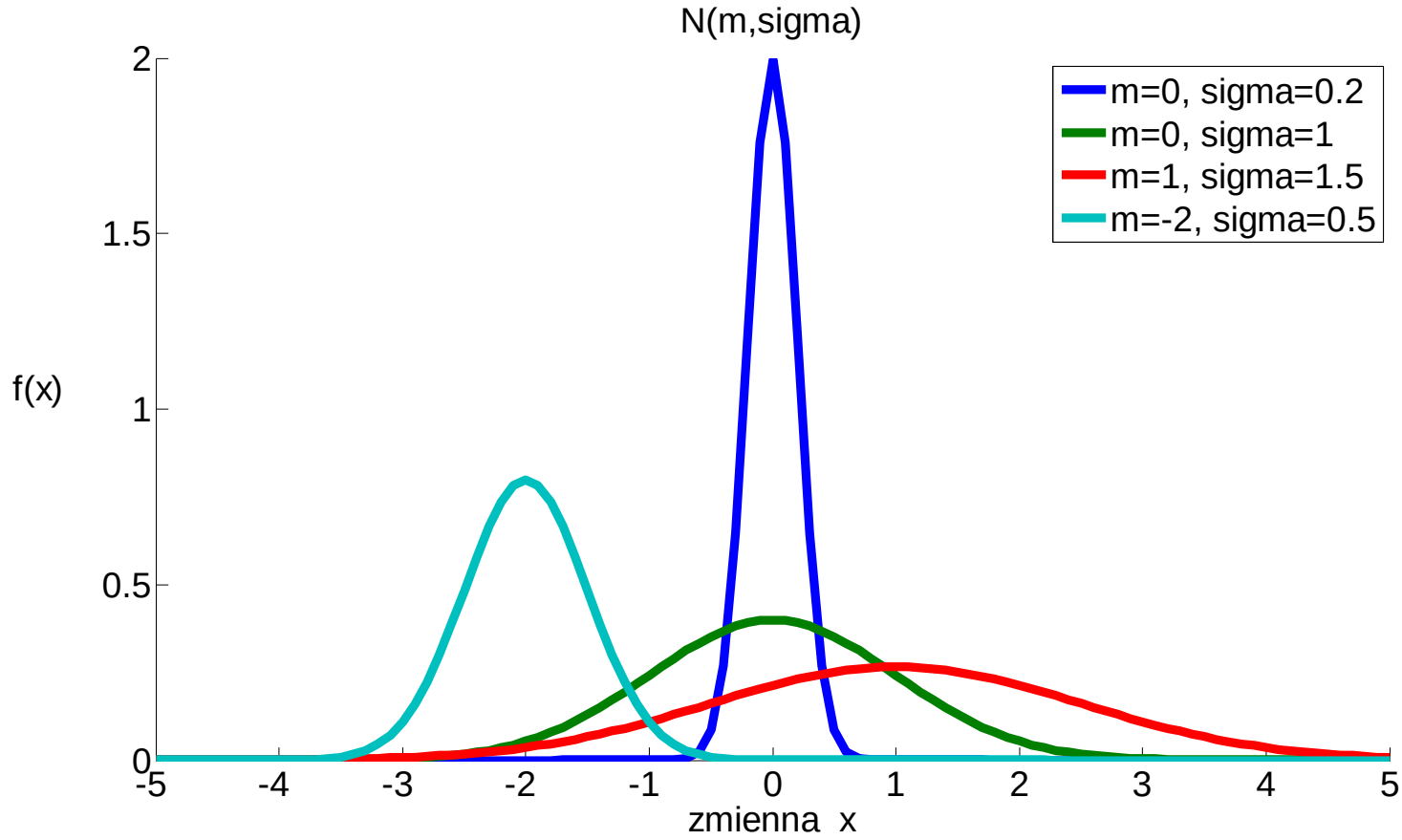
Funkcja gęstości prawdopodobieństwa:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Rozkład ten oznaczany jest także jako:

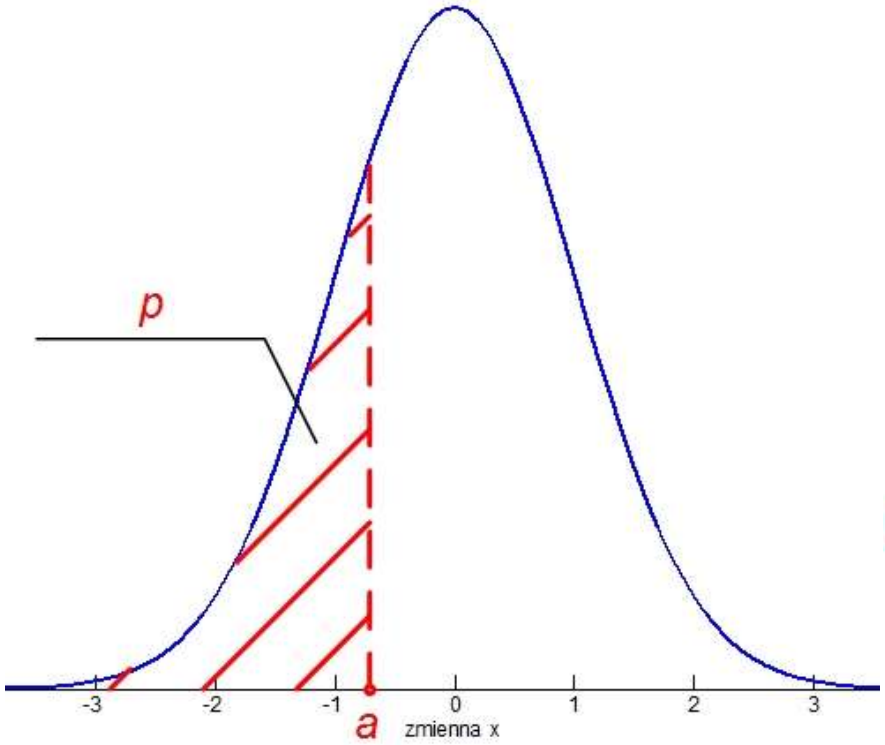
$$N(m,\sigma).$$

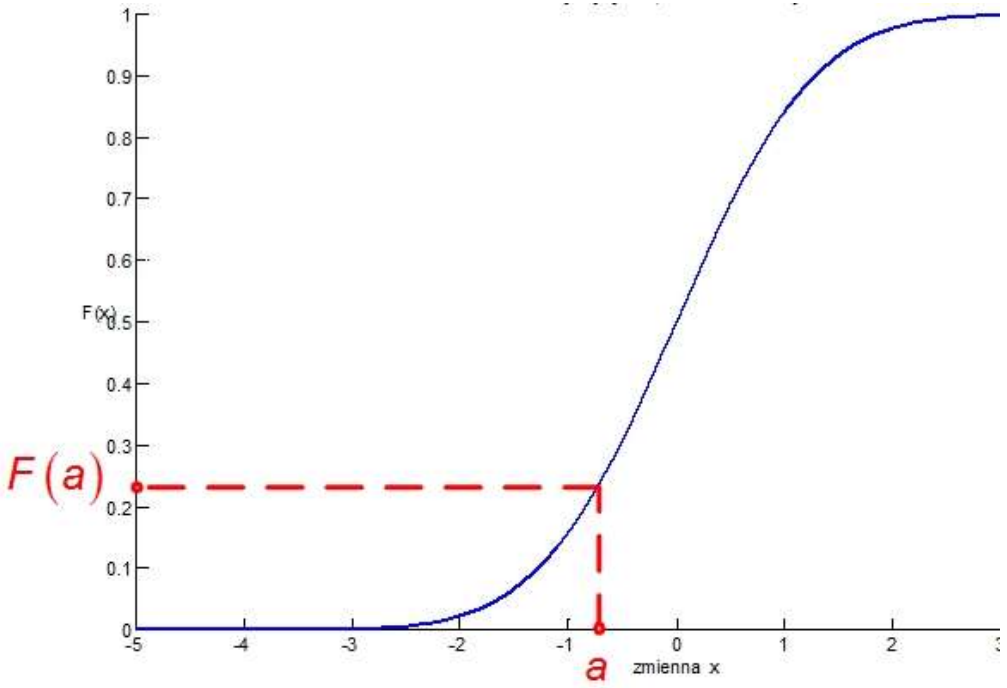
Rozkład normalny (3)



Rozkład normalny standaryzowany $N(0,1)$ jest stablicowany.

Funkcja gęstości a dystrybuanta





$$p = P(X \le a) = F(a)$$

Niektóre rozkłady zmiennych losowych

Zmienna losowa skokowa:

- rozkład dwumianowy
- rozkład Poissona

Zmienna losowa ciągła:

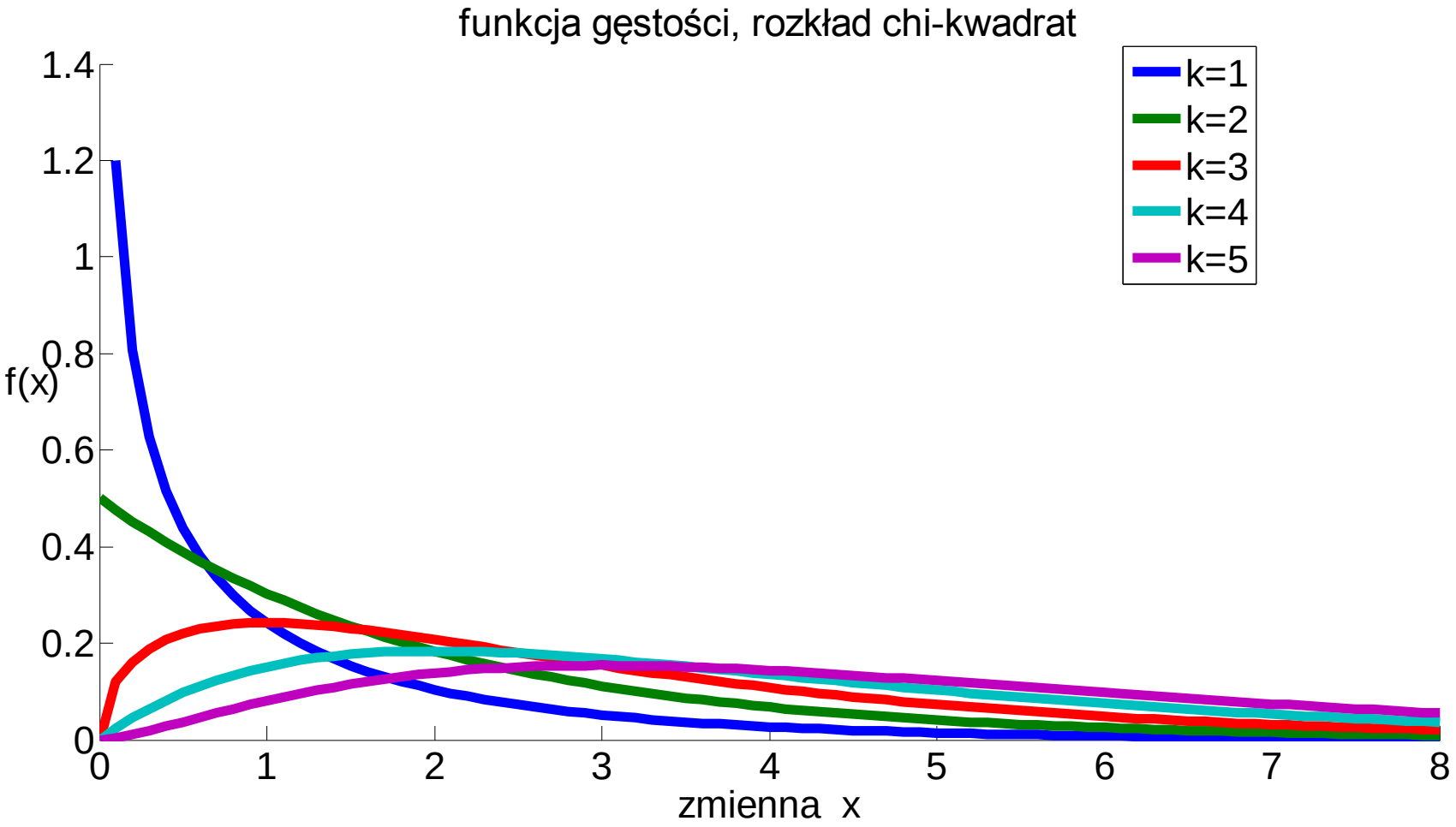
- rozkład normalny (Gaussa)
- rozkład χ^2 (chi-kwadrat)
- rozkład *t*-Studenta
- rozkład *F*-Fishera

Rozkład chi-kwadrat (1)

Rozkład χ^2 to rozkład zmiennej losowej, która jest sumą k kwadratów niezależnych zmiennych losowych o standardowym rozkładzie normalnym.

Liczba naturalna k nazywana jest ***liczbą stopni swobody*** zmiennej losowej.

Rozkład chi-kwadrat (2)



Rozkład chi-kwadrat jest zbieżny do rozkładu normalnego.

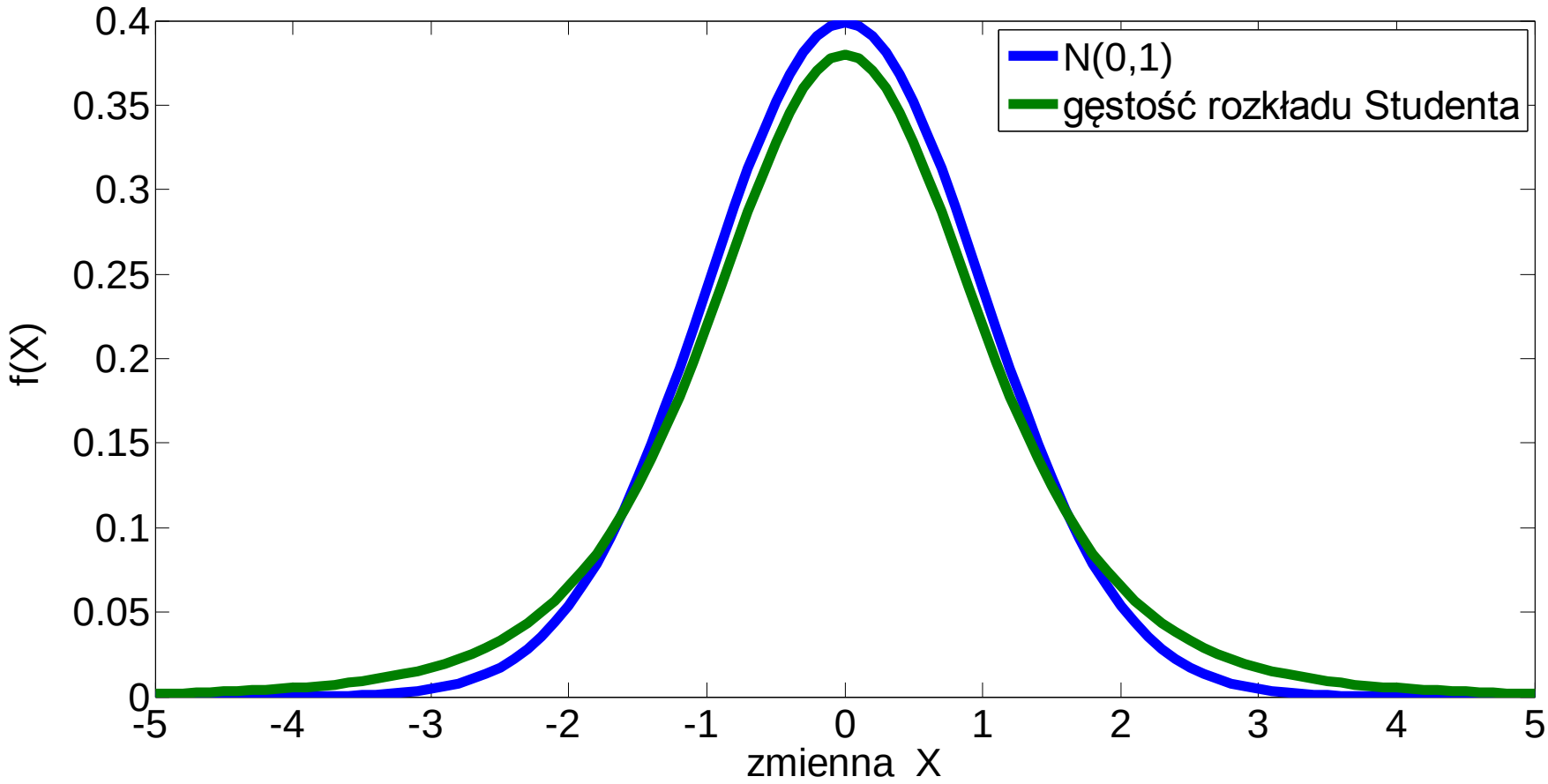
Rozkład Studenta (1)

Jeśli $X_1, X_2, ..., X_n$ są niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(m, \sigma)$ oraz $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ i $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, to zmienna losowa $t = \frac{\overline{X} - m}{s}\sqrt{n-1}$ ma rozkład Studenta z $(n-1)$ stopniami swobody.

Gosset (pod pseudonimem Student) umożliwił badanie średniej arytmetycznej z próby bez znajomości odchylenia standardowego σ.

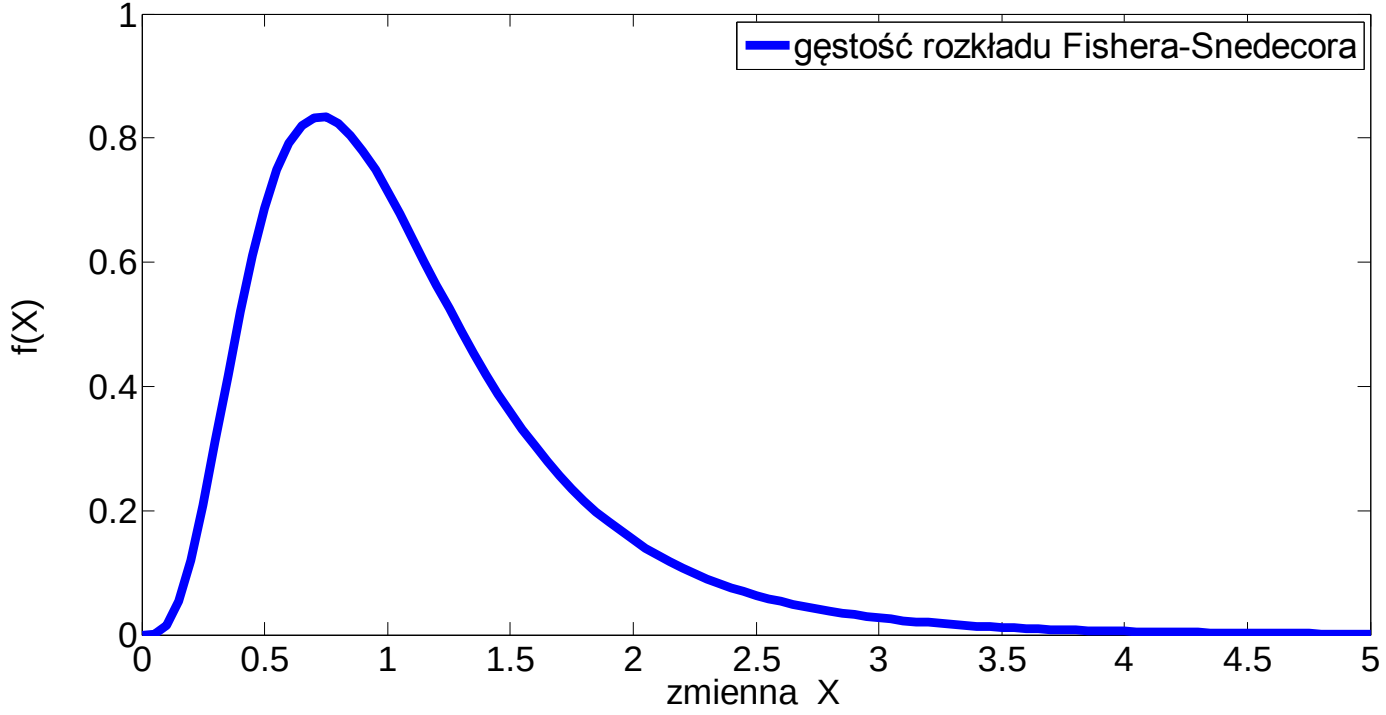
Rozkład Studenta jest zbieżny do rozkładu normalnego.

Rozkład Studenta (2)



Rozkład Fishera-Snedecora

Jeśli X_1^2 i X_2^2 są zmiennymi o rozkładzie χ^2 z n_1 i n_2 stopniami swobody odpowiednio, to zmienna $F = \frac{n_2 X_1^2}{n_1 X_2^2}$ ma rozkład Fishera-Snedecora z n_1 i n_2 stopniami swobody.



Wnioskowanie statystyczne

Często nie znamy ani typu rozkładu ani wartości parametrów.

Wnioskujemy o zbiorowości (populacji) na podstawie próby.

Poprawność wnioskowania zależy przede wszystkim od tego, czy próba dobrze reprezentuje analizowaną populację, czyli struktura próby jest jak najbardziej zbliżona do struktury populacji.

Reprezentatywność próby jest zapewniona, jeśli próba jest losowa.

Wnioskowanie statystyczne obejmuje dwie grupy metod: estymacje i weryfikacje hipotez statystycznych.

Estymacja (1)

Estymacja, czyli szacowanie - to „odgadywanie” rozkładu lub wartości parametrów na podstawie próby.

Estymacja rozkładu to estymacja ***nieparametryczna***. Najprostszą metodą jest tu obliczanie częstości oraz rysowanie histogramu, który pozwala wstępnie ocenić rozkład.

Estymacja parametryczna wykorzystuje pewne charakterystyki liczbowe wyliczane z próby. Są to estymatory, które zależą od wartości parametru populacji oraz od wyników próby. Ponieważ próba jest losowa, to i estymator jest zmienna losową posiadająca własny rozkład prawdopodobieństwa.

Estymacja (2)

Wymagane jest, aby estymatory były:

- zgodne, czyli w miarę wzrostu liczebności próby coraz precyzyjniej „odgadywały” szacowany parametr;
- nieobciążone – średnio „trafiające” w nieznany parametr;
- efektywne – zapewniające mały błąd estymacji;
- odporne – mało wrażliwe na błędy w danych.

Jeżeli nieznana wartość parametru jest równa wartości estymatora otrzymanej w próbie, to mamy do czynienia z ***estymacją punktową***.

Można jednakże wykorzystywać informacje o rozkładzie estymatora i konstruować tzw. ***przedziały ufności***, czyli przedziały liczbowe, o których z dużą ufnością (zazwyczaj 95%) możemy powiedzieć, że zawierają w sobie nieznaną, szukaną wartość parametru (***estymacja przedziałowa***).

Weryfikacja (1)

Weryfikacja hipotez statystycznych pozwala przy pomocy testu statystycznego zweryfikować hipotezę (sąd) o rozkładzie lub parametrze populacji.

Test statystyczny to procedura pozwalająca odrzucić badaną hipotezę z małym ryzykiem popełnienia błędu polegającego na odrzuceniu hipotezy prawdziwej. Ryzyko to mierzone jest tzw. poziomem istotności α, który przez większość badaczy przyjmowany jest na poziomie 0,05.

Badacz musi sformułować hipotezę zerową H_0 oraz hipotezę alternatywną H_1, zwaną niekiedy hipotezą badawczą.

Weryfikacja (2)

Testy parametryczne wymagają, aby rozkład badanej cechy był określonego typu (zazwyczaj normalny), a ***testy nieparametryczne*** wolne są już od takich założeń.

Podstawowym „narzędziem” w teście jest ***statystyka testowa***, której rozkład jest znany i w związku z tym jesteśmy w stanie ocenić, które wyniki (wartości statystyki) są mało prawdopodobne przy danej hipotezie zerowej.

Obecnie wszystkie statystyczne pakiety komputerowe podają ***wartość*** *p*.

Hipotezę zerową należy odrzucić, jeśli wartość *p* jest mniejsza od przyjętego poziomu istotności α. Jest to prosta reguła taka sama dla wszystkich testów statystycznych.

Testy parametryczne (1)

Dla testów parametrycznych:

1. H_0 : postaci $=$

2. $H_1 : \sim H_0$ (może być postaci $<$ $>$ $\neq$)

3. α

Podczas testowania hipotezy zerowej możemy popełnić jeden z dwóch błędów: *I rodzaju* - gdy odrzucamy hipotezę zerową H_0 w przypadku jej prawdziwości; *II rodzaju* - nie odrzucamy hipotezę zerową H_0 gdy jest ona fałszywa.

Testy parametryczne (2)

4. $p = \begin{cases} \Phi(T) & \text{dla hipotez lewostronnych} \\ 2 \cdot \min\{\Phi(T), 1-\Phi(T)\} & \text{dla hipotez obustronnych} \\ 1-\Phi(T) & \text{dla hipotez prawostronnych} \end{cases}$

T - wartość statystyki testowej, obliczonej na podstawie próby

$\Phi(T)$ - dystrybuanta w punkcie T, czyli prawdopodobieństwo, że wartość statystyki jest mniejsza bądź równa wartości statystyki testowej.

Uwaga: w przypadku rozkładów F i χ^2 dla hipotez obustronnych i prawostronnych $p = 1-\Phi(T)$!

Testy parametryczne (3)

5. $(p \leq \alpha \ \vee \ \ p > \alpha) \rightarrow$ wniosek

***p*-wartość** to najmniejszy poziom istotności, przy którym zaobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej.

Jeżeli $p \leq \alpha$, to odrzucamy hipotezę zerową H_0.

Jeżeli $p > \alpha$, to nie ma podstaw do odrzucenia hipotezy zerowej H_0.

Przykład 3

$$p=\begin{cases}\int_{-\infty}^{2,3} f(x)dx=\Phi(2,3)=98,93\% & \text{hipoteza lewostronna}\\ \int_{2,3}^{-\infty} f(x)dx=1-\Phi(2,3)=1,07\% & \text{hipoteza prawostronna}\\ 2\cdot\min\{98,93\%,1,07\%\}=2,14\% & \text{hipoteza obustronna}\end{cases}$$

Jeśli testujemy hipotezę o równości przeciwko hipotezy o nierówności, to wartość p wynosi 2,14%, czyli dla $2,14\% < \alpha$ hipotezę zerową H_0 należy odrzucić na korzyść hipotezy alternatywnej H_1.

Przykład 4 (1)

$\alpha = 0{,}05$

1. $H_0 : m = 250$ $H_1 : m < 250$, $u = -2{,}06$; rozkład normalny $N(0,1)$.

Hipoteza lewostronna, $p = \Phi(u) = \Phi(-2{,}06) = 0{,}0197$, rozkład normalny standaryzowany.

Wniosek: $p < \alpha$, odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej, czyli średnia w rozkładzie jest mniejsza, niż 250.

Przykład 4 (2)

2. $H_0 : m = 120$ $H_1 : m \neq 120$, $t = 1{,}67$; rozkład t-Studenta, liczba stopni swobody $n = 25$.

Hipoteza obustronna,

$$p = 2 \cdot \min\left\{\Phi(t), 1 - \Phi(t)\right\} = 2 \cdot \min\left\{0{,}9463;\ 0{,}0537\right\} = 0{,}1074.$$

Wniosek: $p > \alpha$, brak podstaw do odrzucania hipotezy zerowej, że średnia w rozkładzie jest równa 120.

Przykład 4 (3)

3. $H_0 : \sigma^2 = 0{,}6$ $H_1 : \sigma^2 > 0{,}6$, $\chi^2 = 16{,}5$; rozkład χ^2, liczba stopni swobody $n = 11$.

Hipoteza prawostronna,

$p = 1 - \Phi\left(\chi^2\right) = 1 - \Phi\left(16{,}5\right) = 1 - 0{,}8764 = 0{,}1236$, rozkład chi-kwadrat.

Wniosek: $p > \alpha$, brak podstaw do odrzucania hipotezy zerowej, że wariancja w rozkładzie jest równa 0,6.

Przykład 4 (4)

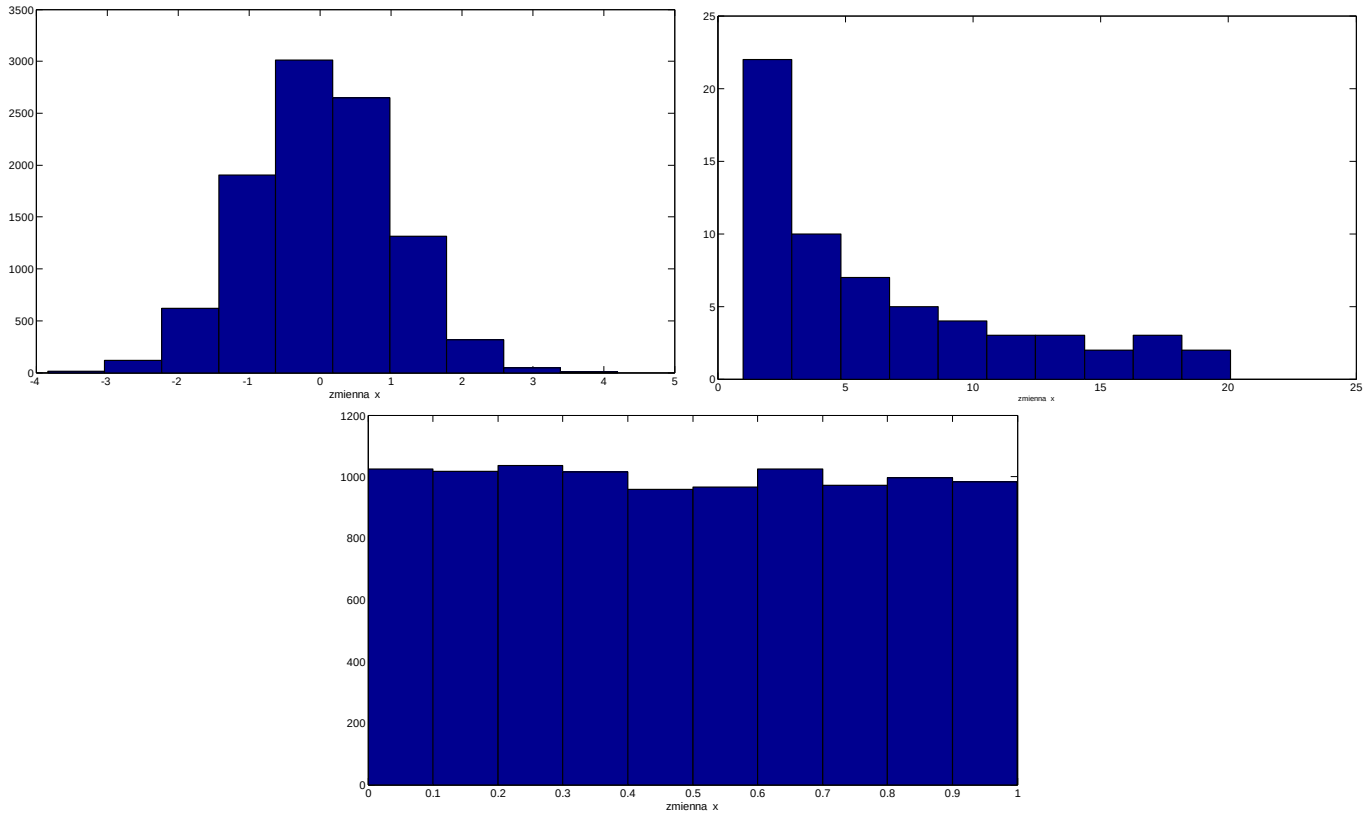
4. $H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$, $F = 1{,}78$; rozkład $F - Snedecora$, liczba stopni swobody $n_1 = 20$, $n_2 = 15$.

Hipoteza prawostronna,
$p = 1 - \Phi(F) = 1 - \Phi(1{,}78) = 1 - 0{,}8708 = 0{,}1292.$

Wniosek: $p > \alpha$, brak podstaw do odrzucania hipotezy zerowej, że wariancje w rozkładach są równe.

Test zgodności (1)

Założenie o rozkładzie można zrobić, wykorzystując histogram, np., normalny, wykładniczy, równomierny:



Założenia testu zgodności chi-kwadrat: minimalna liczebność próby jest równa 5 i niezależność grup.

Test zgodności (2)

1. H_0 : hipoteza o postaci rozkładu zmiennej.
2. $H_1 : \sim H_0$.
3. α.
4. Obliczenie statystyki chi-kwadrat: $\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i\right)^2}{np_i}$.

Liczba stopni swobody: $\nu = k - r - 1$, gdzie r - liczba parametrów w rozkładzie. W przypadku, gdy liczebności oczekiwane $n_i p_i$ są znane z góry $\nu = k - 1$.

5. $p = 1 - \Phi\left(\chi^2\right)$, gdzie $\Phi\left(\chi^2\right)$ - dystrybuanta rozkładu χ^2.

Porównanie p z α.

6. Wniosek.

Test zgodności (3)

rozkład	funkcja	parametry	liczba parametrów	estymatory	liczba st.sw.
Poissona	$p_k = \frac{\lambda^k e^{-\lambda}}{k!}$	λ	1	$\lambda = \bar{x}$	$\nu = k - 2$
równomierny	$p_i = \begin{cases} 0, x < a \\ \frac{1}{b-1}, a \le x \le b \\ 0, x > b \end{cases}$	a, b	2	$a = \bar{x} - \sqrt{3} \cdot s,$ $b = \bar{x} + \sqrt{3} \cdot s$	$\nu = k - 3$
normalny	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}$	m, σ	2	$m = \bar{x}$ $\sigma = s$	$\nu = k - 3$
wykładniczy	$f(x) = \lambda e^{-\lambda x}$	λ	1	$\lambda = \frac{1}{\bar{x}}$	$\nu = k - 2$

Przykład 5 (1)

Czy na poziomie istotności $\alpha = 0{,}05$ można sądzić, że rozkład dziennej liczby dostaw dla pewnego przedsiębiorstwa w ciągu 90 dni jest rozkładem Poissona?

liczba dostaw x_i	liczba dni n_i
0	19
1	29
2	17
3	14
4	11

$H_0 : X \text{ ma rozkład Poissona}$

$H_1 : X \text{ nie ma rozkładu Poissona}$

Przykład 5 (2)

$$\lambda = \bar{x} = \frac{0 \cdot 19 + 1 \cdot 29 + 2 \cdot 17 + 3 \cdot 14 + 4 \cdot 11}{(19 + 29 + 17 + 14 + 11)} = \frac{149}{90} = 1{,}656 \approx 1{,}7.$$

Prawdopodobieństwo $p_i = P(X = x_i)$, czyli:

$$p_0 = \frac{1{,}656^0 e^{-1{,}656}}{0!} = 0{,}191, \quad n \cdot p_0 = 90 \cdot 0{,}191 = 17{,}887$$

$$p_1 = \frac{1{,}656^1 e^{-1{,}656}}{1!} = 0{,}3162, \quad n \cdot p_1 = 90 \cdot 0{,}3162 = 28{,}4569$$

$$p_2 = \frac{1{,}656^2 e^{-1{,}656}}{2!} = 0{,}2617, \quad n \cdot p_2 = 90 \cdot 0{,}2617 = 23{,}5560$$

...

Przykład 5 (3)

dzienna liczba dostaw x_i	liczba dni n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
0	19	0,191	17,9	0,1909
1	29	0,3162	28,5	0,0104
2	17	0,2617	23,6	1,8246
3	14	0,1444	13,0	0,0770
4	11	0,0598	5,38	5,8697

$$\chi^2 = 0{,}1909 + \ 0{,}0104 + 1{,}8246 + 0{,}0770 + 5{,}8697 = 7{,}9726$$

$$\nu = k - r - 1 = 5 - 1 - 1 = 3$$

$$p = 1 - \Phi\left(\chi^2\right) = 1 - \Phi\left(7{,}97\right) = 1 - 0{,}9534 = 0{,}0466 < 0{,}05$$

Wniosek: odrzucamy H_0.

Przykład 6 (1)

Dla zmiennej X sprawdzić, czy ma ona rozkład normalny.

	x_i	n_i
1	94-100	3
2	100-106	7
3	106-112	11
4	112-118	20
5	118-124	28
6	124-130	19
7	130-136	10
8	136-142	2

Przykład 6 (2)

Estymowane parametry rozkładu:

$$m = \bar{x} = 119,2 \;\; \sigma = s = 9,35.$$

1. H_0 : rozkład zmiennej X jest normalny $N(119,2;\ 9,35)$.
2. H_1 : $\sim N(119,2;\ 9,35)$.
3. $\alpha = 0,05$.
4. Obliczenie statystyki chi-kwadrat: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$.

Przykład 6 (3)

$$p_i = P(a < X < b) = F(b) - F(a) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)$$

Przykładowo:

$$p_1 = P(94 < X < 100) = F(100) - F(94) =$$

$$= \Phi\left(\frac{100-119,2}{9,35}\right) - \Phi\left(\frac{94-119,2}{9,35}\right) =$$

$$= \Phi(-2,05) - \Phi(-2,69) = -0,4798 + 0,4964 = 0,0166$$

Przykład 6 (4)

	x_i	n_i	p_i	$n \cdot p_i$
1	94-100	3	0,0166	1,66
2	100-106	7	0,059	5,9
3	106-112	11	0,1416	14,16
4	112-118	20	0,2283	22,83
5	118-124	28	0,2472	24,72
6	124-130	19	0,1798	17,98
7	130-136	10	0,0878	8,78
8	136-142	2	0,0288	2,88

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i\right)^2}{np_i} = 3,2979$$

Liczba stopni swobody $\nu = 8 - 3 = 5$

Przykład 6 (5)

5. $p = 1 - \Phi(3,2979) = 1 - 0,3458 = 0,6542$, gdzie $\Phi(\chi^2)$ - dystrybuanta rozkładu χ^2 z 5 stopniami swobody.

6. Wniosek: na poziomie istotności 0,05 brak podstaw do odrzucania H_0 o normalnym rozkładzie.