ANALIZA DANYCH – PODSTAWY STATYSTYKI

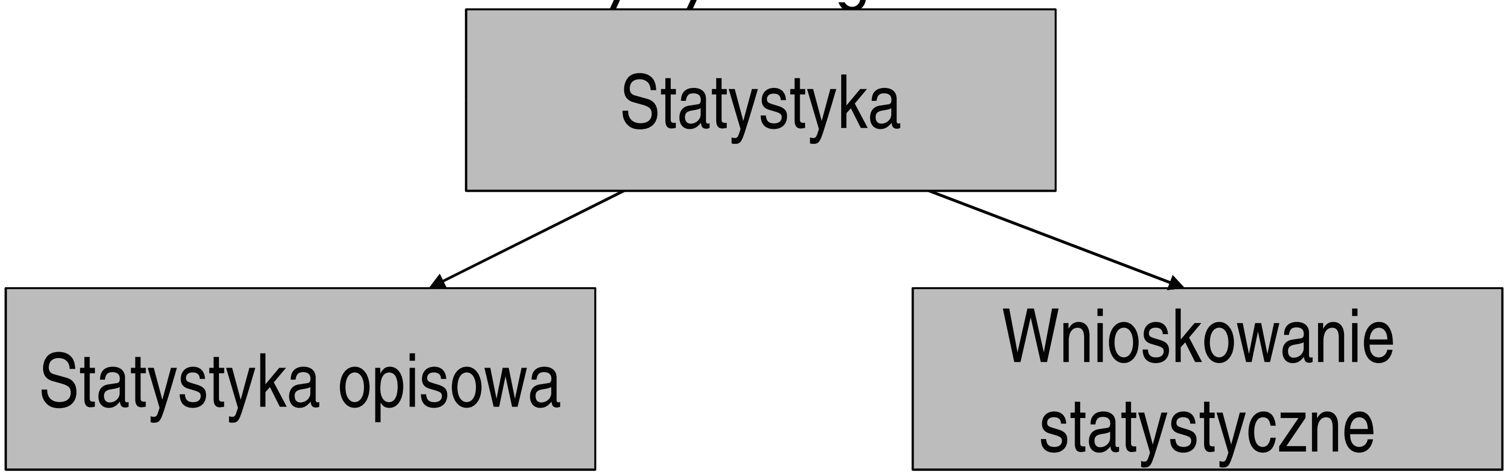
Dr Ewa Więcek-Janka, dr inż. Agnieszka Kujawińska

Celem analiz statystycznych jest:

- znalezienie prawidłowości kształtujących zjawiska:
  - badanie struktury kosztów produkcji
  - badanie zmian w poziomie i strukturze ludności na określonej przestrzeni i czasie
  - badanie związku pomiędzy stażem pracy a wydajnością pracowników
  - reklamą a obrotami
  - Inne.

Zbiór danych może być rozpatrywany:

- z punktu widzenia:
  - opisowego
  - wnioskowania statystycznego



Podział cech statystycznych:

Jednym z wielu podziałów cech jest:

mierzalne

1) wartości dają się wyrazić za pomocą liczb
2) wyrażone w różnych jednostkach: zł, tonach, sztukach, itd...

Skokowe

Ciągłe

niemierzalne

1) nie dają się zmierzyć
2) np. płeć, zawód, kolor..
3) opisane na skali nominalnej

nominalne

porządkowe

Miary statystyczne:

- miary położenia
- miary rozproszenia
- miary asymetrii
- miary koncentracji

Miary położenia (tendencji centralnej)

- **miary położenia ze zbioru:** percentyle
  - mediana
  - kwartyl I, III
- **średnia arytmetyczna**
- **dominanta**

Szczególne percentyle:

- **Mediana:**
  - leży w centrum zbioru w tym sensie, że połowa wyników znajduje się powyżej, a połowa poniżej jej wartości (**kwartyl 2**)
- **Kwartyle:**
  - **kwartyl 1: ozn. Q_1 (25% wyników leży poniżej tego percentyla)**
  - **kwartyl 3: Q_3 (75% wyników leży poniżej jego wartości)**

Pozostałe miary:

- **Dominanta (wartość modalna, moda)**
  - jest wartość, która w tym zbiorze występuje najczęściej

- **Średnia arytmetyczna (średnia klasyczna)**
  - zwaną także przeciętną jest to suma wartości wszystkich wyników podzielona przez ich liczbę

Średnia arytmetyczna - oznaczenia

średnia w próbie: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

średnia w populacji: $$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

Mediana

- dla zbioru o p… ch

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

- dla zbioru o nie… zbie danych

$$Me = x_{\frac{n+1}{2}}$$

Miary zróżnicowania:

- rozstęp (obszar zmienności)
- odchylenie przeciętne
- wariancja
- odchylenie standardowe

Miary rozrzutu: Rozstęp

- w zbiorze wyników obserwacji rozstępem nazywamy różnicę pomiędzy wartością największą i najmniejszą

$$R = x_{max} - x_{min}$$

- jest średnią arytmetyczną bezwzględnych różnic pomiędzy poszczególnymi wartościami cechy a wartością średnią

$$D_m = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Miary rozrzutu: Wariancja

- w zbiorze wyników wariancją nazywamy przeciętne kwadratowe odchylenie poszczególnych wyników od ich średniej

próba

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

populacja

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

Miary rozrzutu: Odchylenie standardowe

- pierwiastek kwadratowy z wariancji

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

Uwaga:

- w przypadku prób o liczebności n<30

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

Grupowanie danych

- **Szereg pozycyjny:** sortujemy dane rosnąco lub malejąco i zliczamy ile jest elementów o tej samej wartości lub cesze

- **Szereg rozdzielczy:** dane grupujemy w klasy, czyli przedziały o ustalonej wielkości

Algorytm postępowania

krok 1: zebrać dane

krok 2: ustalić rozstęp wartości $R=x_{max}-x_{min}$

krok 3: ustalić liczbę przedziałów *k* lub ze wzoru

$$k = 1+3,32*\log N$$

krok 4: podzielić rozstęp *R* przez liczbę przedziałów *k* (uzyskamy szerokość przedziałów *d*)

krok 5: wyznaczyć przedziały (lewostronnie domknięte lub prawostronnie-bądź konsekwentny!)

krok 6: przyporządkować dane do przedziałów

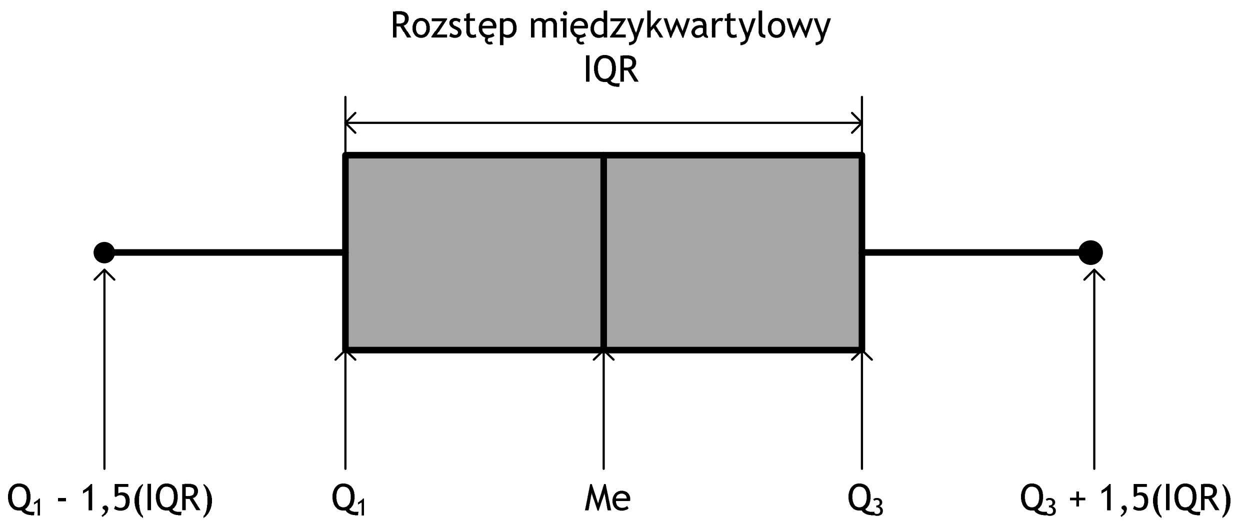
krok 7: histogram

Jak dobrać liczbę klas?

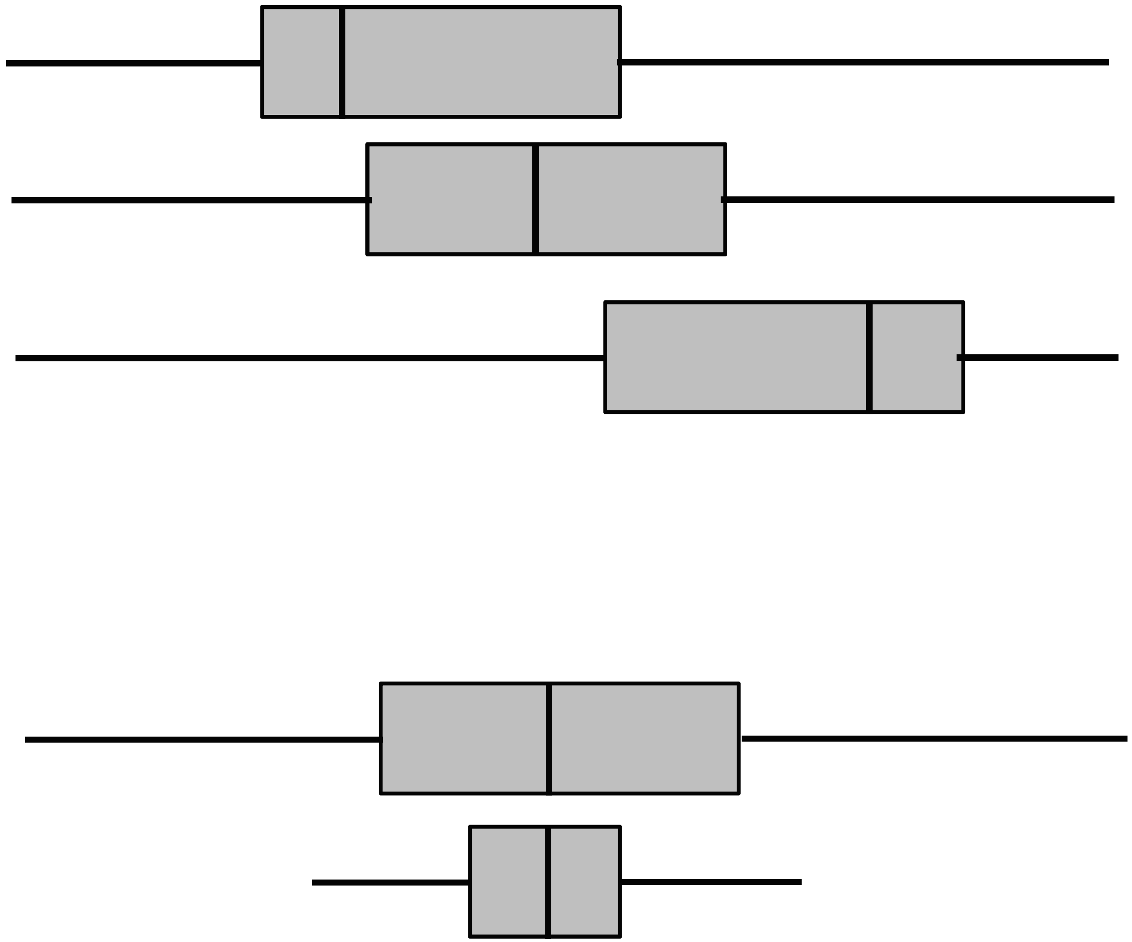
Liczność próbki n	Ilość przedziałów k
30 ÷ 50	6 ÷ 10
51 ÷ 100	7 ÷ 11
101 ÷ 200	8 ÷ 12
201 ÷ 500	9 ÷ 15

$$k = 1 + 3{,}32 * \log N$$

Wykresy RAMKA-WĄSY



Przykładowe wykresy ramka-wąsy



asymetria prawostronna

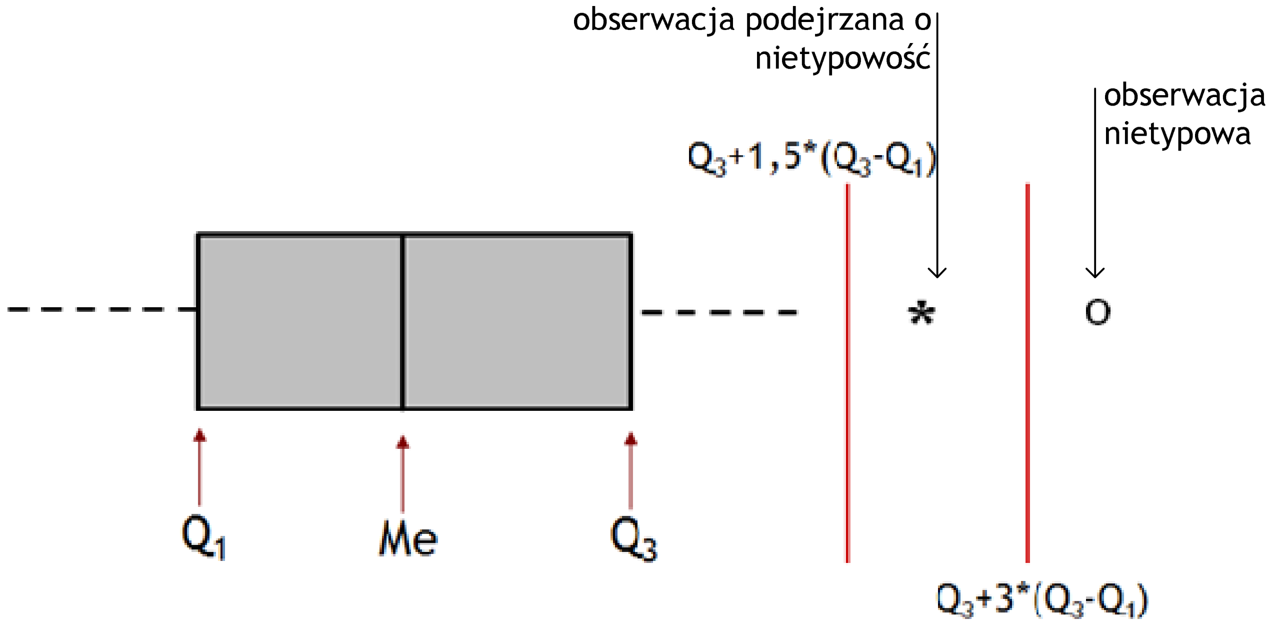
symetria

asymetria lewostronna

duża wariancja

mała wariancja

Wykres R-W pozwala na wykrycie obserwacji nietypowych!



Prawdopodobieństwo

Metody szacowania prawdopodobieństwa

- metoda oparta o klasyczną definicję prawdopodobieństwa,
- metoda empirycznej estymacji prawdopodobieństwa,
- metoda empirycznej estymacji prawdopodobieństwa subiektywnego.

Prawdopodobieństwo w ujęciu klasycznym

Ω – zbiór wszystkich zdarzeń elementarnych

A – zdarzenie losowe (podzbiór zdarzeń elementarnych)

k – liczba wyników, gdy zdarzenie A się pojawia

m – liczba wszystkich możliwych wyników

prawdopodobieństwo „a priori” **P(A) = k/m**

Podstawowe prawa rachunku prawdopodobieństwa:

1) $P(E_i) \geq 0$

2) $SP(E_i) = 1$

3) Jeżeli A^{-1} jest zdarzeniem przeciwnym do A (dopełnieniem) to $P(A) = 1 – P(A^{-1})$

Zmienne losowe i ich rozkłady

Zmienna losowa intuicyjnie - to zmienna, która przyjmuje wartości liczbowe z pewnego zbioru z określonym prawdopodobieństwem

Naukowo:jest to funkcja, która przy zajściu każdego zdarzenia losowego w przyjmuje konkretną wartość x(w), co zapisujemy:

$X: \omega \rightarrow x(\omega) \in R$

Przykładowo:

- jeśli doświadczenie polega na kontroli jakości 5 opon podlegających ocenie alternatywnej, to **zmienną losową** może być **liczba wadliwych opon, która może przyjąć wartość od 0 do 5**
- cecha, którą obserwujemy (mierzymy) jest zmienną losową

Zmienna losowa

dyskretna ciągła

Rozkład gęstości prawdopodobieństwa:

- zmienna losowa dyskretna

tablica, wzór lub wykres, który przyporządkowuje prawdopodobieństwa każdej możliwej wartości zmiennej

$P(X=x)=P(x) \geq 0$ dla każdego x

$\Sigma P(x_i) = 1$

- zmienna losowa ciągła

funkcja ciągła

$f(X=x)=f(x) \geq 0$ dla każdego x

$\int f(x)dx = 1$

Dystrybuanta zmiennej losowej

- Jest to funkcja określona wzorem

$$F(X) = P(X < x)$$

$$P(-\infty < X < x)$$

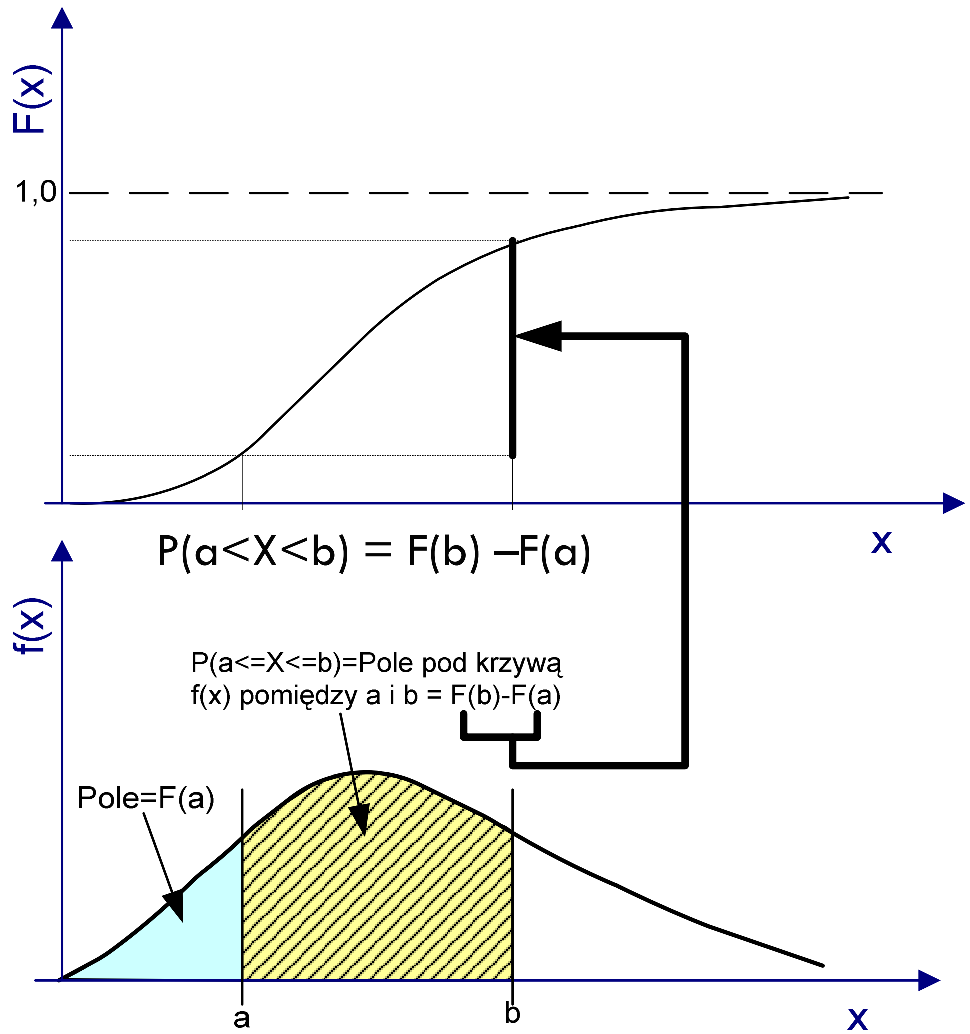
- dla zmiennej losowej skokowej:

$$F(x) = \sum_{x_i < x} p(x_i)$$

- dla zmiennej losowej ciągłej:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

Związek pomiędzy F(x) a f(x)



Rozkład normalny (Gaussa)

- Rozkład normalny jest rozkładem, do którego dąży m.in. rozkład dwumianowy gdy liczba doświadczeń n wzrasta

- Okazuje się, że rozkład normalny jest rozkładem granicznym wielu innych rozkładów, w sytuacjach gdy ujawniają się skutki różnych przypadkowych czynników pochodzących z różnych źródeł

Funkcja gęstości rozkładu normalnego

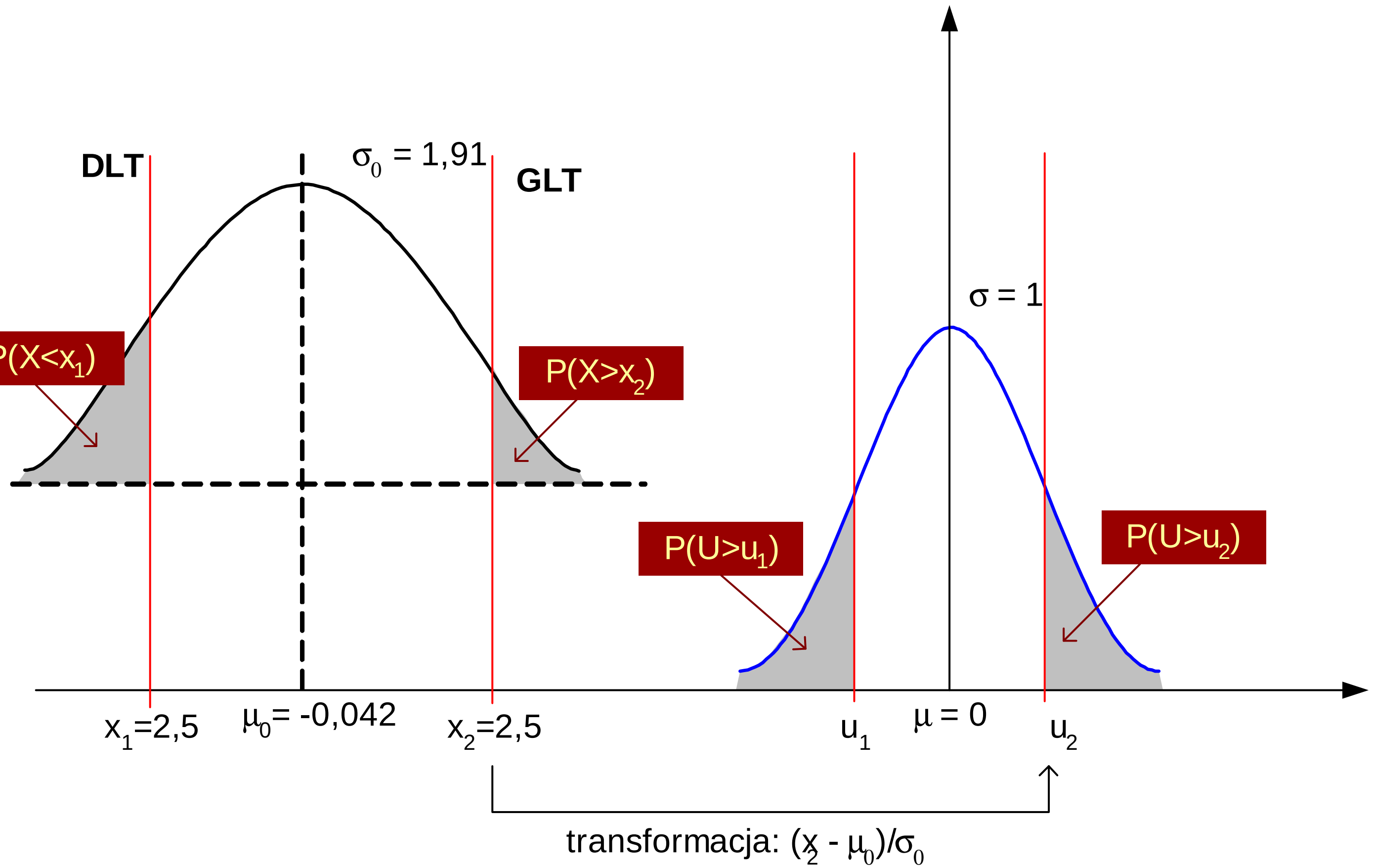
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standaryzowany rozkład normalny

- oznaczany: Z lub U
- zapisywany często: N(0, 1)

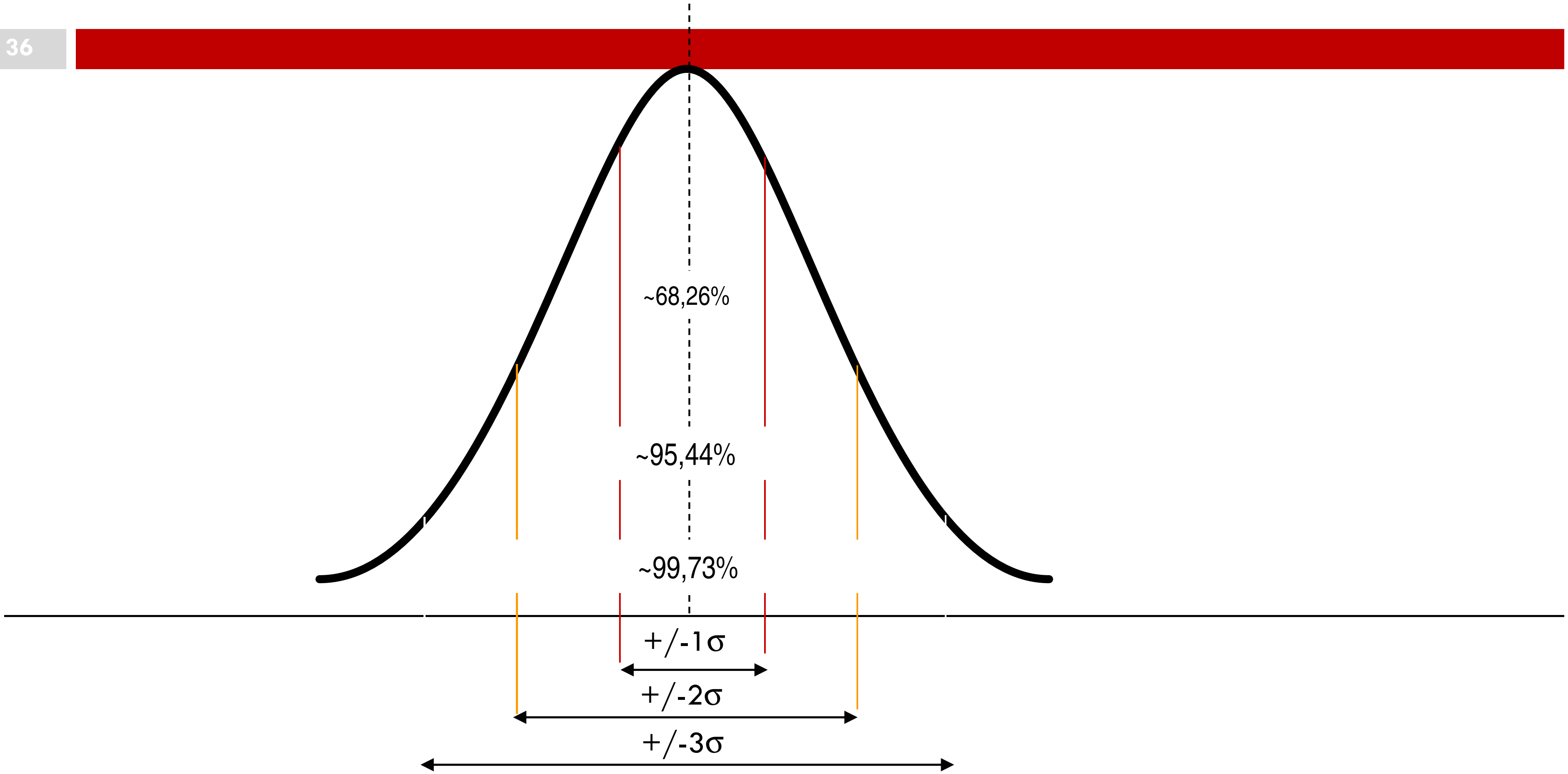
$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$

Przekształcenie



$P(X<x_1) = P(U<u_1)$
$P(X>x_2) = P(U>u_2)$

Właściwość rozkładu normalnego



WERYFIKOWANIE HIPOTEZ STATYSTYCZNYCH

Stosuje się dwie grupy testów:

- parametryczne i nieparametryczne
  - stosowanie pierwszych wymaga przyjęcia założeń o postaci rozkładu testowanej zmiennej losowej oraz znajomości wybranych statystyk
  - testy nieparametryczne takich założeń nie wymagają, ale nie są tak mocne jak parametryczne

Hipotezy statystyczne

- **Hipoteza statystyczna to każde przypuszczenie dotyczące rozkładu zmiennej losowej weryfikowane na podstawie n-krotnej realizacji tej zmiennej**

- Wyróżniamy:
  - Hipotezy
    - parametryczne i nieparametryczne
    - proste i złożone

Przykład: [źródło: Aczel 2000]

Firma rozwożąca paczki zapewnia, że średni czas dostarczenia przesyłki od drzwi klienta do odbiorcy wynosi 28 minut. By sprawdzić to stwierdzenie pobrano próbę n=100 przesyłek i obliczono średni czas dostawy 31,5 minut oraz odchylenie standardowe 5 minut. Czy zapewnienie firmy można uznać za nieprawdziwe?

$H_0 : \mu = 28$ $\qquad$ $1-\alpha = 0{,}95$

$H_1 : \mu \neq 28$ $\qquad$ $\sigma = 5$

zbudujmy 95% przedział ufności dla średniej:

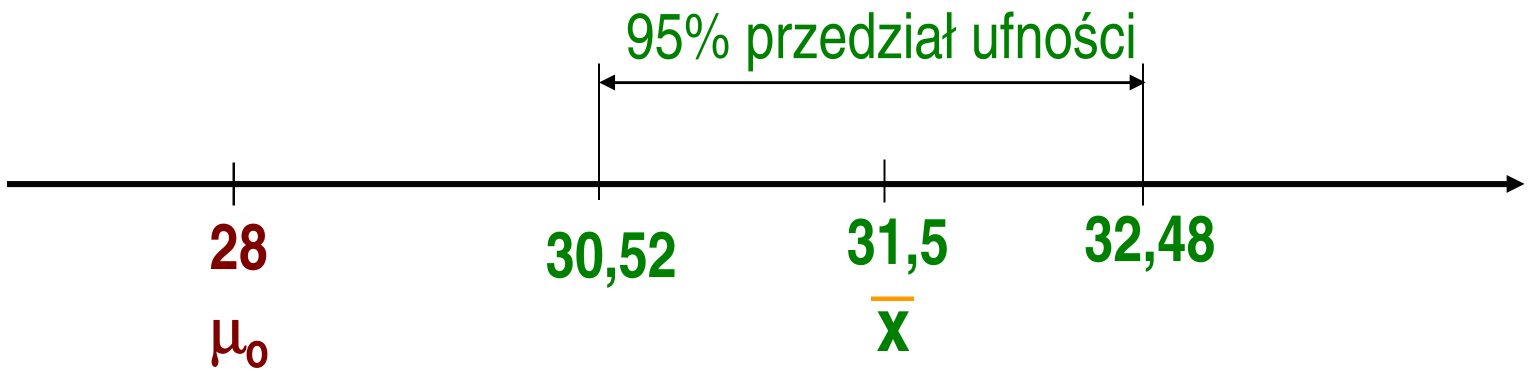
$$\bar{x} \pm u_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 31{,}5 \pm 1{,}96 \frac{5}{\sqrt{100}} = [30{,}52;\ 32{,}48]$$

Jeżeli mamy 95% ufności, że średni czas dostawy zawiera się w przedziale [30.52; 32.48] minuty, to mamy 95% zaufania, że czas ten nie znajdzie się poza tym przedziałem.

Wartość sprawdzana: 28 minut, leży poza tym przedziałem, zatem odrzucamy hipotezę zerową.

Czego uczy ww przykład?

Po pierwsze: przy weryfikowaniu testów można budować przedział ufności wokół wartości statystyki z próby i sprawdzać, czy weryfikowana wartość parametru należy do przedziału



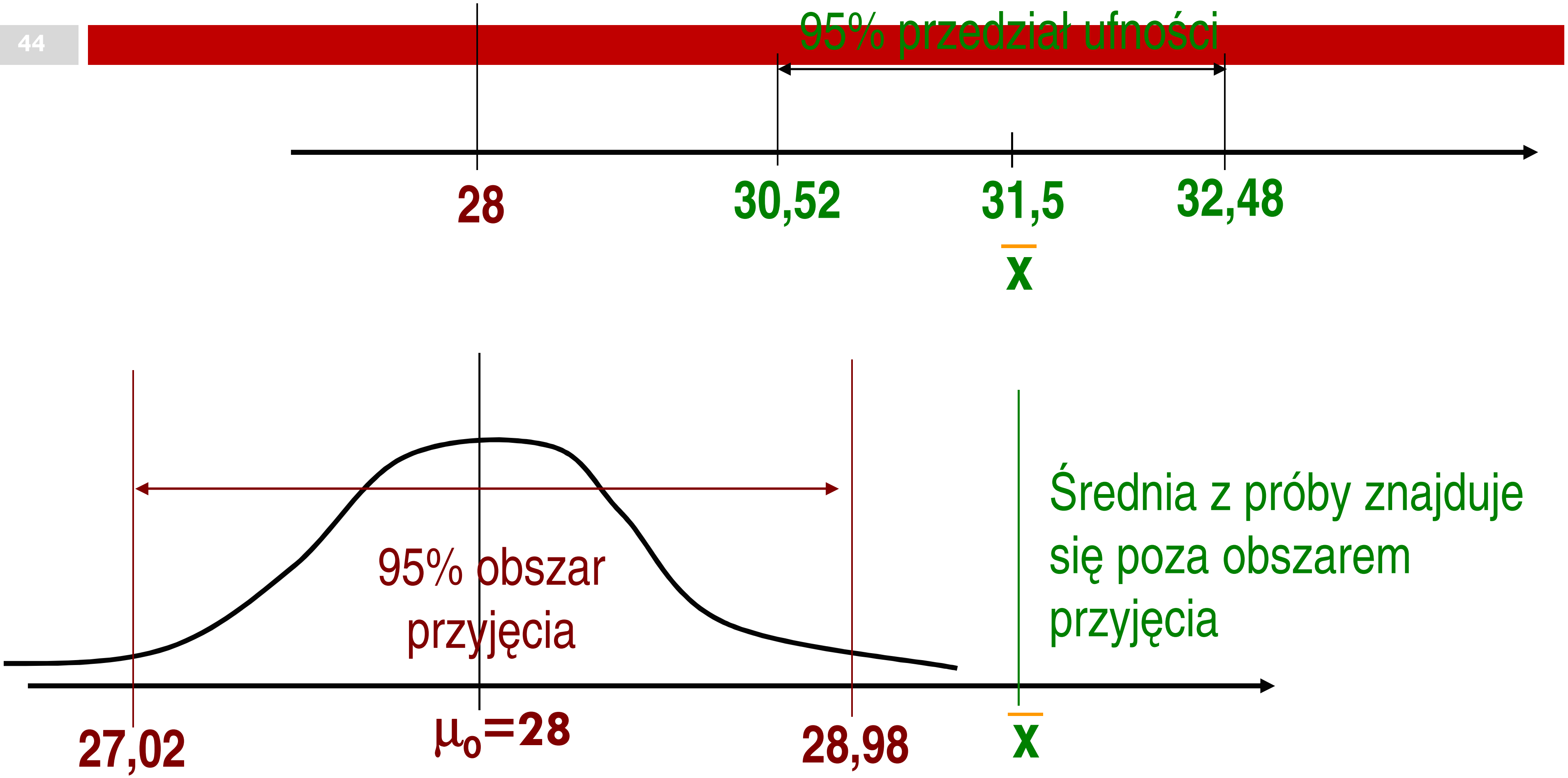
po drugie:

Można jako centrum traktować średnią populacji i sprawdzać wartość statystyki z próby względem przedziału ufności wokół parametru populacji

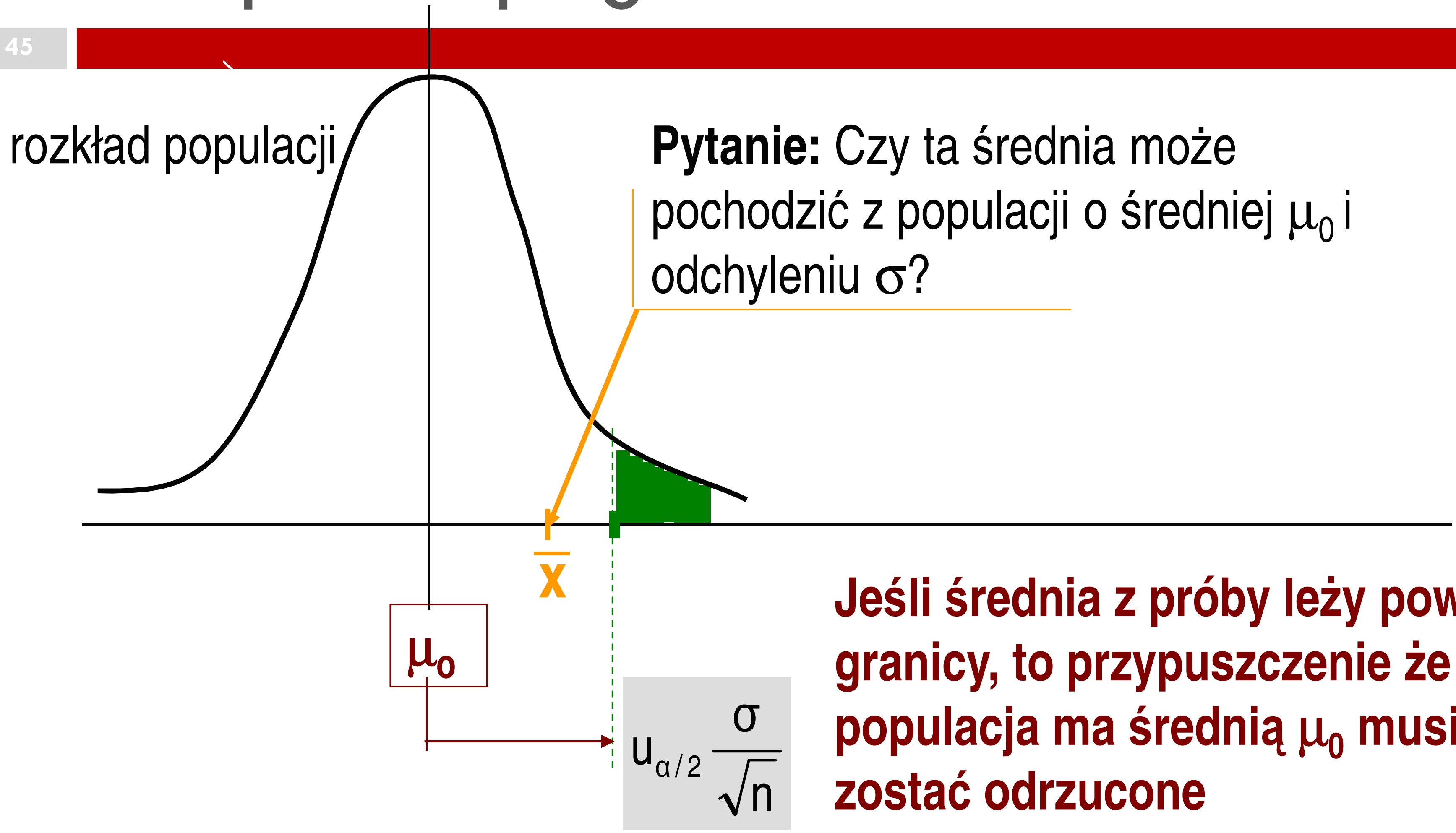
$$\mu_0 \pm 1{,}96\frac{s}{\sqrt{n}} = 28 \pm 1{,}96\frac{5}{\sqrt{100}} = [27{,}02;\ 28{,}98]$$

Wartość średnia z próby =31,5, zatem nie należy do przedziału ufności. Hipotezę zerową odrzucamy.

Interpretacja graficzna



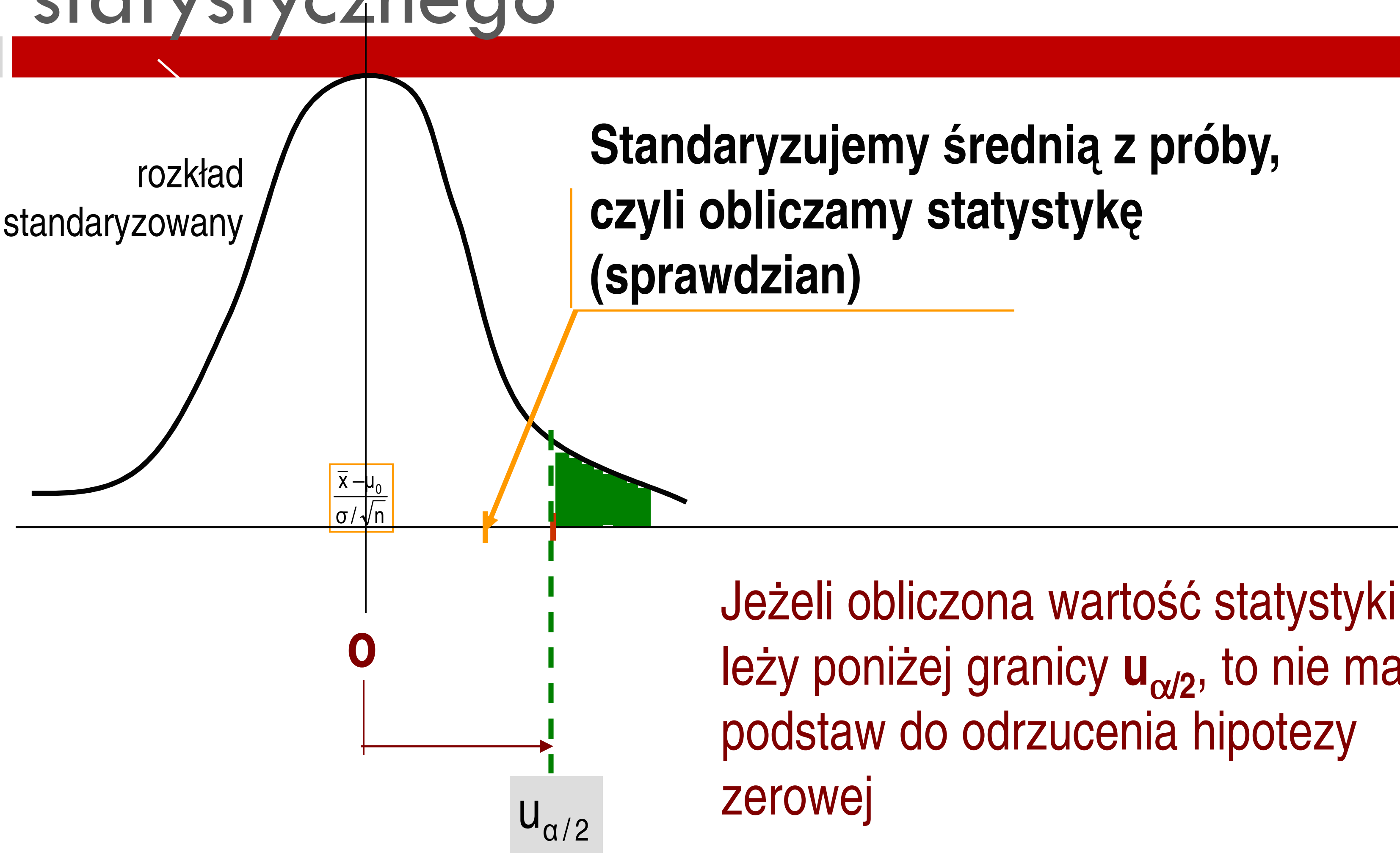
Interpretacja graficzna



Pytanie: Czy ta średnia może pochodzić z populacji o średniej μ_0 i odchyleniu σ?

Jeśli średnia z próby leży powyżej granicy, to przypuszczenie że populacja ma średnią μ_0 musi zostać odrzucone

Standaryzowana forma testu statystycznego



Standaryzujemy średnią z próby, czyli obliczamy statystykę (sprawdzian)

Jeżeli obliczona wartość statystyki leży poniżej granicy $\mathbf{u_{\alpha/2}}$, to nie ma podstaw do odrzucenia hipotezy zerowej

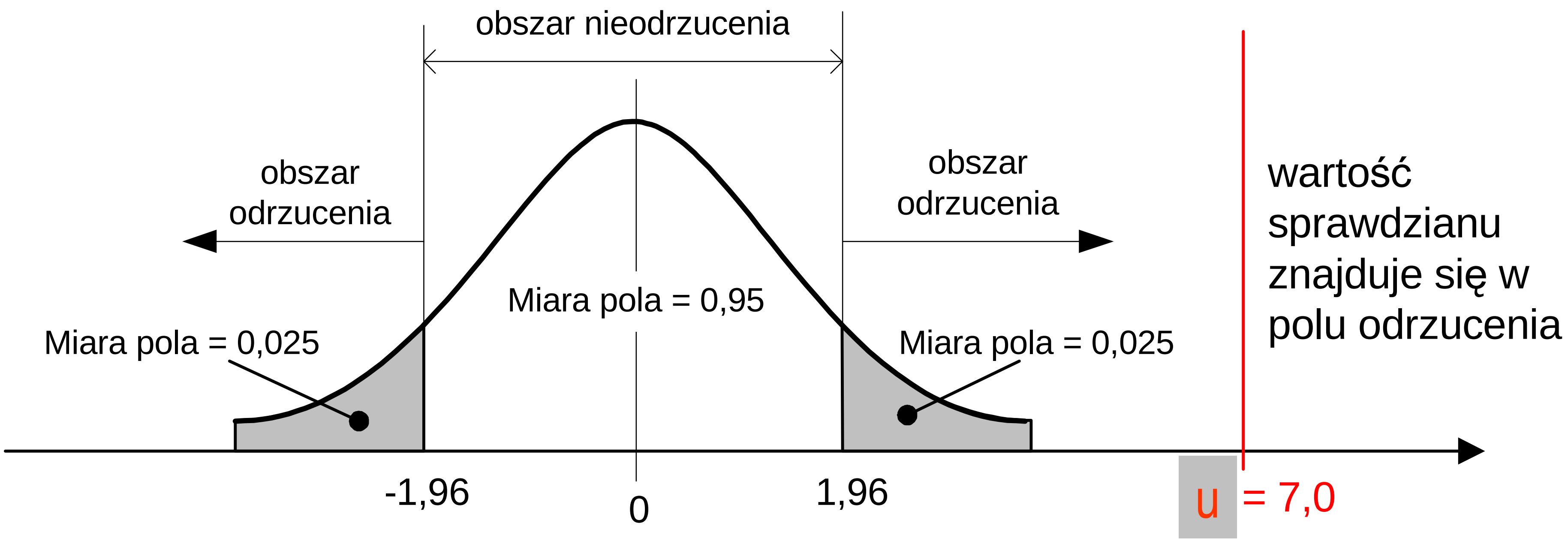
Wracając do przykładu:

$H_0: \mu = 28$

$H_1: \mu \neq 28$

$$u = \frac{31{,}5 - 28}{5/\sqrt{100}} = 7$$

Obszar krytyczny: **$R\alpha$ = (-∞; -1,96) ∪ (1,96; +∞)**



Testy jednostronne

- Wybór rodzaju testu podyktowany jest potrzebą działania
- Jeżeli działanie (np. korygujące) będzie podjęte, gdy parametr przekroczy pewną wartość a, to stosujemy **test prawostronny**:

$$H_0: \mu \leq a$$

$$H_1: \mu > a$$

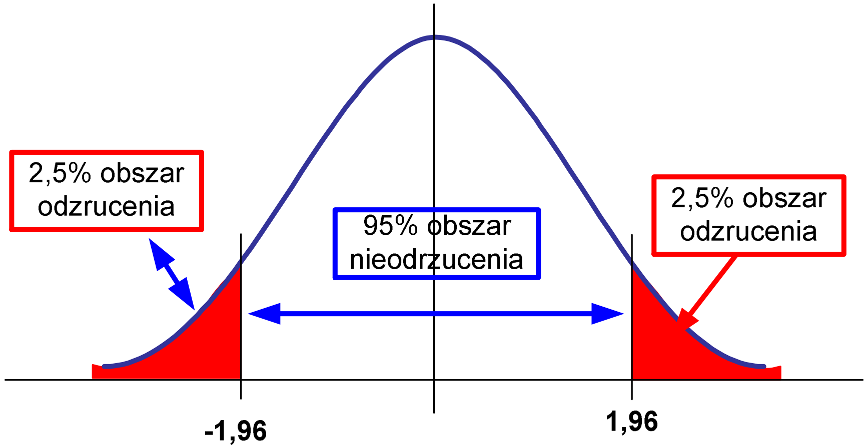
- Jeżeli działanie będzie podjęte, gdy parametr przyjmie wartość mniejszą niż a, to stosujemy **test lewostronny**:

$$H_0: \mu \geq a$$

$$H_1: \mu < a$$

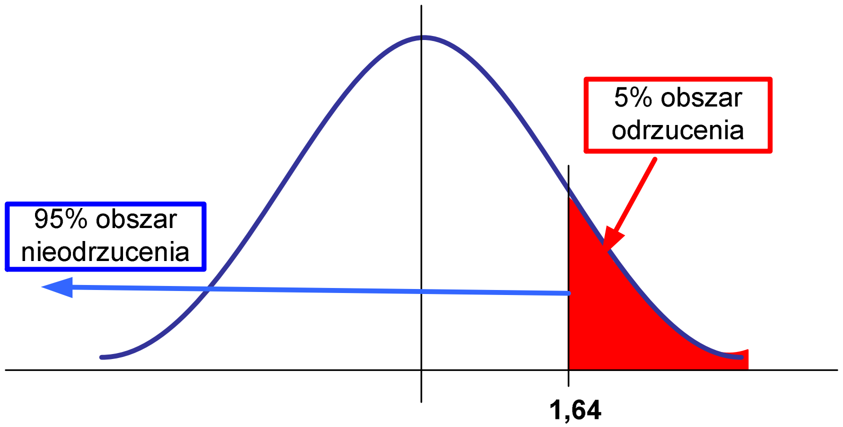
$H_0: \mu = a$

$H_1: \mu \neq a$



$H_0: \mu \leq a$

$H_1: \mu > a$



Prawdopodobieństwo błędu II-go rodzaju

- w testach zakładamy błąd α
- co z błędem β?

H_0: niewinna
H_1: winna

		Stan rzeczy	
		H_0	H_1
Decyzje	H_0	słuszna decyzja	β
	H_1	α	słuszna decyzja

Prawdopodobieństwo błędu II-go rodzaju

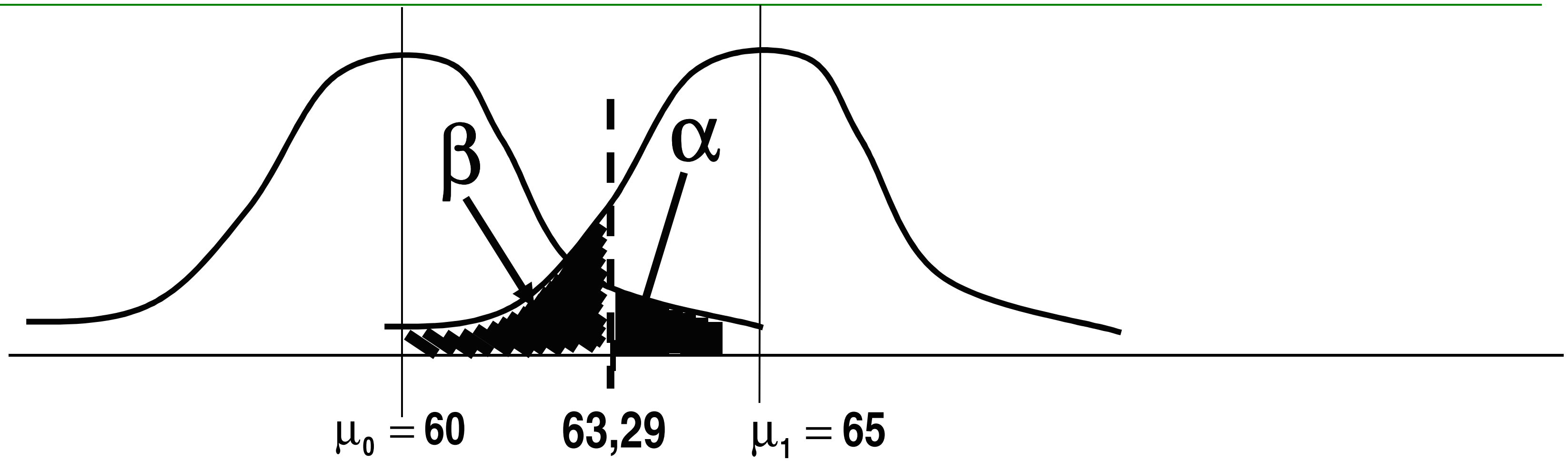
- niestety prawdopodobieństwo β jest trudne do wyznaczenia „a priori”,
- zależy ono od tego, którą z możliwych wartości przyjmie interesujący nas parametr,
- przykładowo dla testów dotyczących μ błąd β jest funkcją μ: $\beta(\mu)$.

Przykład wyznaczania β [źródło: Aczel 2000]:

H_0 = 60 **n = 100** **α = 0,05**

H_1 = 65 **σ = 20**

Mamy do czynienia z hipoteza prostą. Albo dojdziemy do wniosku, że średnia populacji jest równa 60, albo że jest równa 65.
W praktyce takie sytuacje zdarzają się rzadko.



$$C = \mu_0 + 1{,}645\frac{\sigma}{\sqrt{n}} = 63{,}29$$

Jakie jest prawdopodobieństwo β?

$$\alpha = P(\overline{X} > C / \mu = \mu_0)$$

$$\beta = P(\overline{X} < C / \mu = \mu_1)$$

α z góry ustalamy, zatem β:

$$\beta = P(\overline{X} < C / \mu = \mu_1) = P\left(\frac{\overline{X} - \mu_1}{\sigma / \sqrt{n}} < \frac{C - \mu_1}{\sigma / \sqrt{n}}\right) = P(U < -0{,}855) = 0{,}1963$$

Zatem prawdopodobieństwo β przyjęcia błędnej hipotezy, że średnia w populacji jest 60, podczas gdy w rzeczywistości wynosi 65, jest równe 0,1963.
Przeprowadzony test dopuszcza 5% ryzyko odrzucenia Ho gdy jest ona prawdziwa i 19,63% ryzyko przyjęcia Ho gdy jest ona fałszywa.

Moc testu

Mocą testu hipotezy statystycznej jest prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona fałszywa.

moc testu = 1-β

W przykładzie: **moc testu=1-0,1963=0,8037**
Mamy 80,37% szans, że odrzucimy Ho gdy średnia populacji jest równa 65, a nie 60.

Dla testów złożonych

przykładowo w przypadku testu jednostronnego

$H_0: \mu \leq 60$

$H_1: \mu > 60$

Jak zdefiniować moc testu w takiej sytuacji?

Moc testu = P(odrzucenia Ho/ Ho jest fałszywa)

W przykładzie Ho może być fałszywa na nieskończenie wiele sposobów: 61, 62, 67, 72.893 itd...

Własności mocy testu:

1. Moc zależy od odległości między wartością parametru zakładaną w hipotezie zerowej a prawdziwą wartością parametru. Im większa odległość tym większa moc.
2. Moc zależy od wielkości odchylenia standardowego w populacji. Im mniejsze odchylenie tym większa moc.
3. Moc zależy od liczebności próby. Im liczniejsza próba, tym większa moc.
4. Moc zależy od poziomu istotności testu. Im niższy poziom istotności tym mniejsza moc testu.

nie możemy kontrolować punktu 1 i 2
kształtujemy jedynie pkt. 3 i 4

Wartość p – co to takiego?

- to najniższy poziom istotności, przy którym hipoteza zerowa mogłaby być odrzucona przy otrzymanej wartości sprawdzianu

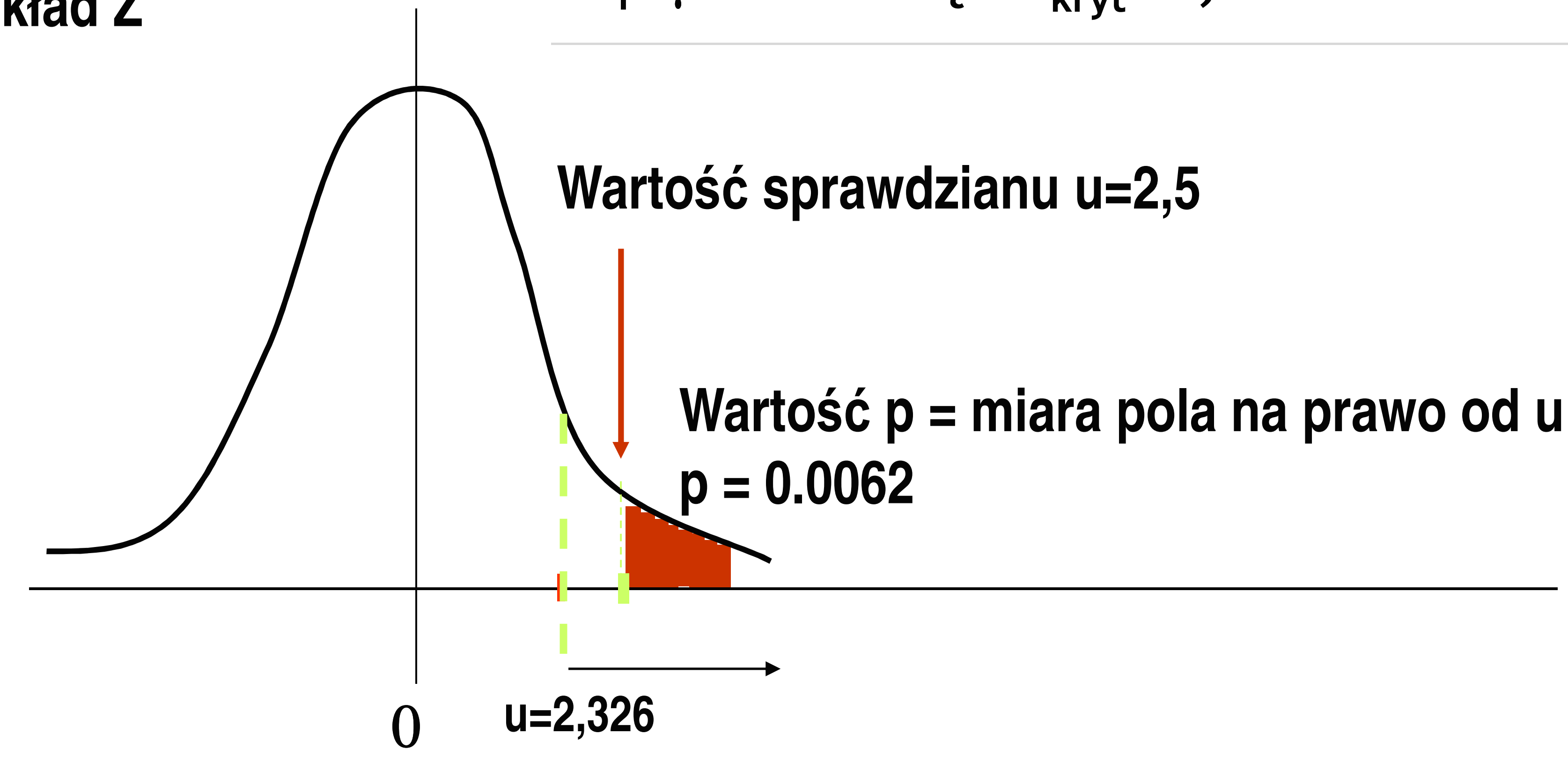
- to prawdopodobieństwo otrzymania takiej wartości sprawdzianu, jaką otrzymaliśmy przy założeniu, że hipoteza zerowa jest prawdziwa

Wartość p - co to takiego?

H_0: $\mu \leq 60$ $\alpha = 0.01$

H_1: $\mu > 60$ stąd u_{kryt}=2,326

rozkład Z



Interpretacja:

- jeśli otrzymana wartość sprawdzianu jest mało prawdopodobna przy założeniu, że Ho jest prawdziwa, to hipoteza Ho powinna być odrzucona

- jeśli otrzymana wartość sprawdzianu jest dosyć prawdopodobna (większa od 0.05; 0.1) to powinniśmy przyjąć hipotezę Ho

Wartość p

Jest czymś w rodzaju zindywidualizowanego poziomu istotności

Załóżmy, że wartość p dla wyznaczonego sprawdzianu wynosi 0.0002

Informacja dla użytkownika testu:

1) Ho musiałaby być odrzucona przy a=0.01
2) Ho musiałaby być odrzucona przy a=0.001 i przy wszystkich poziomach aż do 0.0002!!

Informacja zawarta w p=0.0002 jest bogatsza niż w stwierdzeniu, że Ho odrzucona na poziome α=0.05

ANALIZA DANYCH ANKIETOWYCH

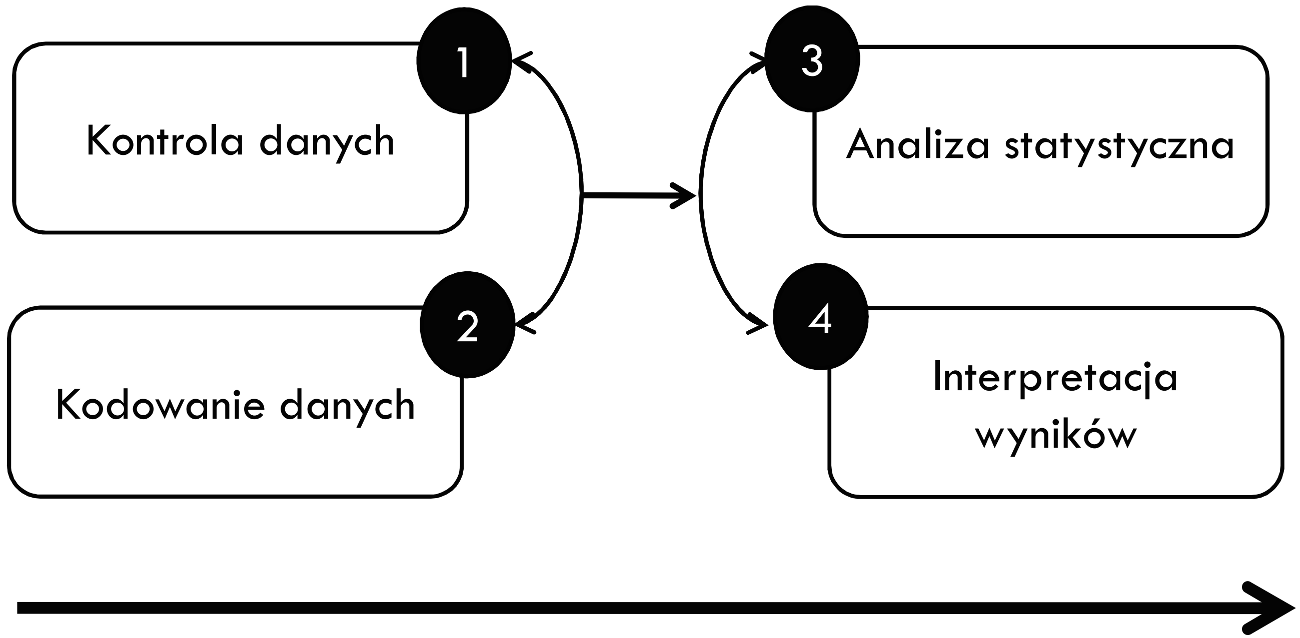
Proces badawczy

- Każdy proces badawczy składa się z etapów układających się w zamknięty cykl.
- Analiza danych jest jednym z elementów tak pojętego cyklu badawczego.
- Miejsce analizy danych w procesie badawczym przedstawia rysunek obok

Zastosowanie metod statystycznych..

- …wymaga odpowiedniego przygotowania danych surowych
- Dane surowe mogą mieć postać:
  - wypełnionych kwestionariuszy,
  - dzienników obserwacji,
  - dzienników panelowych,
  - zapisanych testów,
  - zapisów z pomiaru,
  - inną..
- Dane należy skontrolować i odpowiednio zakodować

Kontrola danych i kodowanie danych to etapy poprzedzające analizę danych



Najczęstsze błędy i braki dotyczące danych:

1) Brak czytelności i dokładności odpowiedzi
2) Pytania bez odpowiedzi, spowodowane m.in.:
   - Pominięcie przez prowadzącego wywiad całych stron arkusza wywiadu lub — w przypadku pomiarów ankietowych — przez respondenta
   - Odmowa odpowiedzi na niektóre pytania lub nie poddanie się pomiarowi
3) Pomiary fikcyjne (są to oszustwa świadomie dokonane przez osoby prowadzące pomiar.)
4) Odpowiedzi nieadekwatne (respondenci dają odpowiedzi nie związane z tematem pytania)

Najczęstsze błędy i braki dotyczące danych:

6) Sprzeczności i niezgodności:
   - Przykładem może być odpowiedź, z której wynika, że respondent nigdy nie słyszał o danym produkcie, podczas gdy w odpowiedzi na inne pytanie twierdzi, że używa tego produktu. O tym, która odpowiedź jest prawdziwa, można niekiedy wnioskować z innych odpowiedzi, ale wnioski te mogą być ryzykowne.

7) Odpowiedzi niekompletne lub niejednoznaczne
   - Niektóre odpowiedzi są niekompletne, nieczytelne lub niejasne i wieloznaczne. Niekompletną odpowiedź można w przybliżeniu określić i uzupełnić. Natomiast odpowiedzi niejednoznaczne lub nieokreślone są trudne do interpretacji i ewentualnej poprawy.

Kodowanie

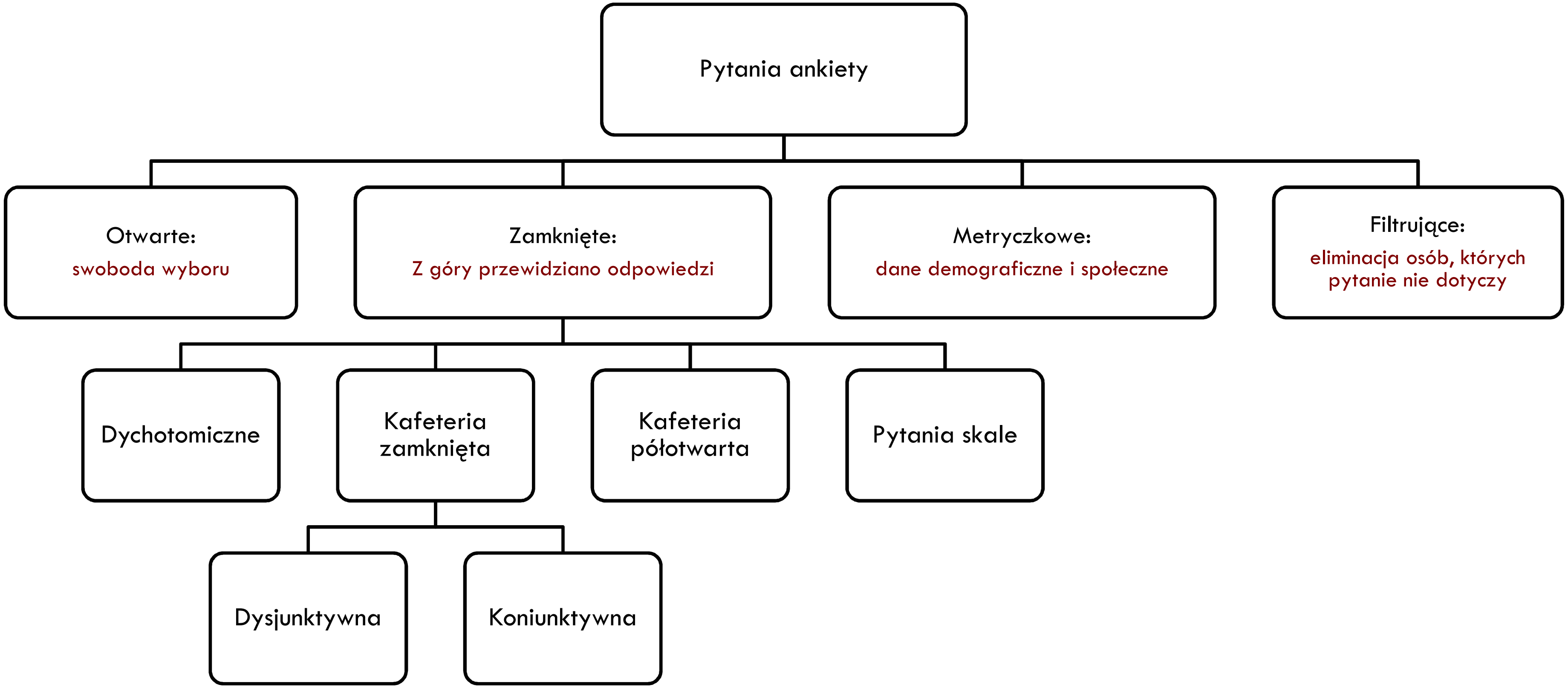
- Współcześnie **kodowanie odpowiedzi** w kwestionariuszach ma na celu **przeniesienie danych z instrumentu pomiarowego** (np. kwestionariusza ankiety) **do pamięci komputera** (arkusza kalkulacyjnego, bazy danych etc.).
- W tym kontekście kodowanie określić można jako przyporządkowanie symboli (liczb/kodów) danym zawartym w instrumentach pomiarowych

Etapy kodowania:

- Stworzenie instrukcji kodowania.
  - Przykład:

Nr w ankiecie	Nr zmiennej	Nazwa zmiennej	Etykieta zmiennej	Wartości (kody)
1	1	P1	Płeć	1 – kobieta 2 – mężczyzna
2	2	Z1	Wielkość firmy	1 – mała 2 – średnia 3 - duża
3	3	S1	Typ firmy	1 – produkcyjna 2 – usługowa 3 – handlowa 4 – produkcyjno-usługowa

- Sposób kodowania w istotnym stopniu zależy od rodzaju pytania i odpowiadających pytaniu odpowiedzi.
- W naukach społecznych wyróżnić można przynajmniej kilka rodzajów pytań. Ich typologię zawiera poniższy rysunek:



Pytania zamknięte – są pytaniami samokodującymi

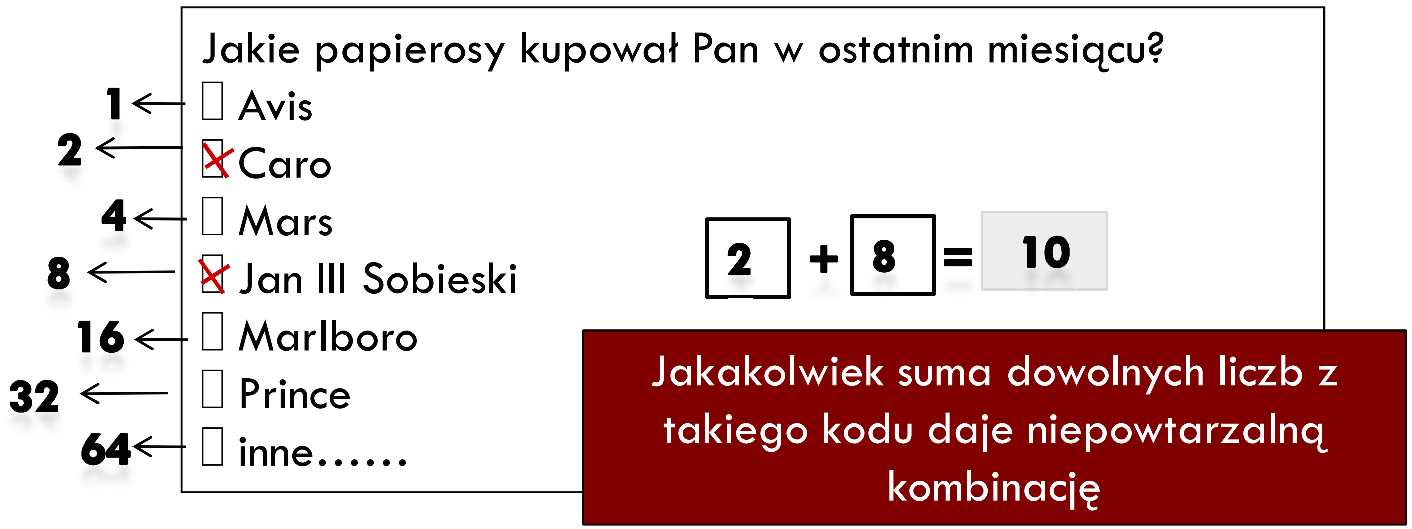
- Kodowanie pytań zamkniętych polega na przeniesieniu odpowiadającego danej odpowiedzi kodu (liczby) do bazy danych.
- Przykładowo dla pytań zamkniętych z jedną opcją wyboru. Każdej opcji przypisano cyfrę od 1 do 5:

Czy jest Pani zadowolona z dezodorantu Nivea?:

- Zdecydowanie tak → 1
- Raczej tak → 2
- Raczej nie → 3
- Zdecydowanie nie → 4
- Jeszcze nie mam wyrobionej opinii 5

Więcej opcji do wyboru – Kodowanie geometryczne

- Kod geometryczny to ciąg o wyrazie pierwszym równym 1 i o ilorazie równym 2.
- Są to następujące liczby (każda kolejna dwukrtonie większa od następnej): 1, 2, 4, 8, 16, 32, 64, …itd.



Więcej opcji do wyboru – Kodowanie binarne

- Kod geometryczny jest kłopotliwy, jeżeli jest dużo wariantów odpowiedzi
- Kodowanie binarne polega na wprowadzeniu do arkusza danych tylu zmiennych (kolumn), ile było wariantów odpowiedzi w danym pytaniu
- W kolumnach pojawiają się wówczas dwie wartości:
  - **0** – nie zaznaczenie odpowiedzi
  - **1** – wybranie odpowiedzi

Pytania otwarte

- W kodowaniu pytań tego typu badacz tworzy schemat kodowania nie przed podjęciem badań, lecz w ich trakcie, na podstawie reprezentatywnej próbki odpowiedzi na dane pytanie. Stąd, kodowanie tego rodzaju określić można jako indukcyjne.
- Tworzenie grup, kategorii i wskaźników kategorii

Nr w ankiecie	Nazwa zmiennej	Etykieta	Kod dla kategorii	Przykładowe odpowiedzi
2	5	Powody nie zaliczenia egzaminu	**0 = z powodu braku czasu**	„pracuję, nie mam czasu” „przy takiej liczbie egzaminów w sesji zabrakło mi czasu”
			1 = z powodu niezrozumienia tematu przedmiotu	„wykładowca prowadził tak wykład, że nie potrafię zrozumieć” „nie rozumiem wykładów, wkuję na pamięć”

Po etapie kodowania

- Następuje etap przygotowania do analiz statystycznych poprzez przygotowanie między innymi tablic wynikowych:
  - Jednodzielcze: służące do określenia prostych rozkładów częstotliwości występowania określonej jednej zmiennej
  - Dwudzielcze: ukazujące rozkłady dwóch zmiennych jednocześnie
  - Wielodzielcze: ukazujące rozkłady trzech (i więcej) zmiennych jednocześnie

Tablica jednodzielcza

- Tablice jednodzielcze ukazują nam częstotliwości z jakimi wystąpiły zawarte w kafeterii (w przypadku pytań zamkniętych) lub w kategoriach po kodowaniu (w przypadku pytań otwartych) odpowiedzi na odpowiednie pytania kwestionariusza
- Przykład:

	Proszę powiedzieć, w jakim stopniu uważasz następujące sprawy za ważne w Twoim życiu: **nauka**	
	Liczebność	%
bardzo ważne	223	44,2%
raczej ważne	271	53,8%
niezbyt ważne	6	1,2%
w ogóle nieważne	2	0,4%
trudno powiedzieć	2	0,4%
Ogółem	504	100,0%

Tabele dwudzielcze

- Tabele dwudzielcze prezentują liczebności (lub procenty) osób poklasyfikowanych według dwóch zmiennych jednocześnie

Płeć	Wykształcenie			
	Podstawowe	Średnie	Wyższe	Razem
Kobiety				
Mężczyźni				
Razem				

1. mężczyźni z podstawowym wykształceniem
2. kobiety z podstawowym wykształceniem
3. mężczyźni ze średnim wykształceniem
4. kobiety ze średnim wykształceniem
5. mężczyźni z wyższym wykształceniem
6. kobiety z wyższym wykształceniem

Procentowanie w tabelach dwudzielczych

Płeć	Wykształcenie			
	Podstawowe	Średnie	Wyższe	Razem
Kobiety				
Mężczyźni				
Razem	**100**	100	100	100

Płeć	Wykształcenie			
	Podstawowe	Średnie	Wyższe	Razem
Kobiety				**100**
Mężczyźni				100
Razem				100

Płeć	Wykształcenie			
	Podstawowe	Średnie	Wyższe	Razem
Kobiety				
Mężczyźni				
Razem				**100**

- Jeżeli z dwóch zmiennych jedna z nich jest zmienną niezależną w danym momencie analizy (jest przyczyną, jest zmienną wyjaśniającą, jest zmienną prognostyczną itp.), zaś druga jest zmienną zależną (skutkiem, zjawiskiem, wyjaśnianym, oczekiwanym efektem prognozy), to za podstawę do obliczeń procentowych (tj. za 100%) bierzemy liczebności podgrup poszczególnych wartości zmiennej niezależnej.
- Generalnie stwierdzić możemy, że:
  - **procentujemy zawsze w kierunku zmiennej niezależnej,**
  - **procenty czytamy zaś (porównujemy) zawsze w kierunku zmiennej zależnej**

Przykład:

- Jedno z pytań w ankiecie dotyczyło wartości wyznawanych w życiu (Liberalne-Konserwatywne), drugie pytanie dotyczyło chęci zakupu nieruchomości typu: „Dom jednorodzinny"
- Pytanie: czy istnieje zależność pomiędzy wartościami, którymi kierujemy się w życiu a chęcią zakupu „Domu jednorodzinnego?
- Tabela dwudzielcza z badań, poniżej:

Wartości	Lib.	Kons.
DOM **Nie**	137	102
DOM **Tak**	27	33

	Zmienne niezależne		
Wartości	Lib.	Kons.	Wiersz RAZEM
DOM Nie	137	102	239
DOM Tak	27	33	60
Kolumna RAZEM	164	135	299

Zmienne zależne: DOM Nie, DOM Tak

W tym przypadku przypuszczamy, że wartości społeczne mają wpływ na posiadanie określonego typu nieruchomości

Procenty w wierszach:

(137/239)*100%

Wartości	Lib.	Kons.	Wiersz RAZEM
DOM Nie	137 57,32%	102 42,68%	239 100%
DOM Tak	27 33,33%	33 66,67%	60 100%
Kolumna RAZEM	164	135	299

Procenty w kolumnach:

(137/239)*100%

Wartości	Lib.	Kons.	Wiersz RAZEM
DOM Nie	137 83,54%	102 75,55%	239
DOM Tak	27 16,46%	33 24,45%	60
Kolumna RAZEM	164 100%	135 100%	299

Procenty z całości:

(137/299)*100%

Wartości	Lib.	Kons.	Wiersz RAZEM
DOM Nie	137 45,82%	102 34,11%	239
DOM Tak	27 9,03%	33 11,04%	60
Kolumna RAZEM	164	135	299 100%

1. Przy prowadzeniu analizy danych za pomocą tabel dwudzielczych mamy następujące problemy do rozpatrzenia:
   a) Czy zaobserwowane różnice pomiędzy częstościami (w komórkach tabeli) są istotne statystycznie?
   b) Jaka jest siła związku pomiędzy zmiennymi?
   c) Czy relacje są pozorne czy rzeczywiste?

Odpowiedzi na problem a) daje między inymi test zgodności chi-kwadrat (χ^2)

- W teście chi-kwadrat stosowany jest następujący tok postępowania:
  - jest formułowane pewne przypuszczenie co do populacji przez określenie hipotezy zerowej i hipotezy alternatywnej,
  - są obliczane teoretyczne częstości występowania określonych zdarzeń, w założonych klasach ($i = 1, 2, ..., k$), tzn. takie, jakich należy spodziewać się przy założeniu prawdziwości hipotezy zerowej. Otrzymuje się w ten sposób oczekiwane liczności E_i (liczności teoretyczne, hipotetyczne) danych
    w różnych klasach,
  - zapisuje się zaobserwowane (empiryczne) liczności O_i danych należących do poszczególnych klas,

- oblicza się różnicę pomiędzy tym, co oczekiwane a tym, co zaobserwowane – z różnic tych oblicza się wartości statystyki testu chi-kwadrat:

$$\chi^2 = \sum_{i=0}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- porównuje się wartość statystyki obliczonej z punktami krytycznymi rozkładu chi-kwadrat i jeśli wartość obliczona:

$$\chi^2 > \chi^2_{\alpha;df}$$

- podejmuje się decyzję o odrzuceniu hipotezy zerowej,

  gdzie:

  α – poziom istotności,

  df – liczba stopni swobody, które są określane oddzielnie w każdej sytuacji; mogą na przykład być równe df = k – p – 1, gdzie: k to liczba klas, p to liczba parametrów szacowanych na podstawie próby.

Istotność różnic dla przykładu:

- Hipotezy:

  H_0: nie ma istotnych różnic w chęci kupienia domu pomiędzy grupą osób wzynających wartości liberalne a konserwatywne

  H_1: są istotne różnice w chęci kupienia domu pomiędzy grupą osób wzynających wartości liberalne a konserwatywne
- Zmienna niezależna: **Typ wartości**
- Zmienna zależna: **Chęć kupienia „Domu jednorodzinnego"**
- Tworzymy tabelę dwudzielczą (slajd następny):
  - W kolumnach umieszczono zmienną niezależną
  - W wierszach zmienną zależną
- Procent z częstości wyznaczamy po zmiennej niezależnej (kolumnach)

Wyznaczanie częstości teoretycznych

(164*239)/299= 131,09

Liczba liberałów, która nie kupiłaby domu,gdyby hipoteza zerowa była prawdziwa

Wartości	Lib.	Kons.	Wiersz RAZEM
DOM **Nie**	137 131,09	102 107	**239**
DOM **Tak**	27 32,90	33 27,09	**60**
Kolumna RAZEM	**164**	**135**	**299**

Istotność różnic:

$$\chi^2 = \frac{(137-131{,}09)^2}{131{,}09} + \frac{(102-107)^2}{107} + \frac{(27-32{,}09)^2}{32{,}09} + \frac{(33-27{,}09)^2}{27{,}09}$$

$\chi^2 = 2{,}69$

$\alpha = 0{,}05$

df=(w-1)(k-1)=1

$\chi_\alpha{}^2 = 3{,}84$

Nie ma podstaw do odrzucenia H_0. Brak jest związku miedzy wartościami a typem nabywanej nieruchomości

Siła związku

- Współczynnik kontyngencji C Pearsona:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

  - Jeżeli C = 0 to brak zależności
  - Górna granica zależy od liczby wierszy w tabeli i jest równa: $\sqrt{\frac{w-1}{w}}$

    - W przykładzie C=0,099, co wskazuje na stosunkowo słaby związek pomiędzy zmiennymi

Współczynnik V Cramera

- Niedogodność braku wartości maksymalnej dla współczynnika C Pearsona można pominąć stosując współczynnik **V Cramera:**

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

gdzie k – mniejsza z liczb kolumn lub wierszy

- Współczynnik przyjmuje wartości z przedziału <0, 1>

Dla zainteresowanych:

- Analiza korespondencji
- Wieloraka analiza korespondencji
- Analiza skupień

Literatura:

1) Aczel A., „Statystyka w zarządzaniu", PWN Warszawa 2000

1) Kaczmarczyk S., „Badania marketingowe, PWE Warszawa 2004

2) Churchill G. A., „Badania marketingowe," PWN Warszawa 2002

3) Kaden R.J., „Badania marketingowe" , PWE Warszawa 2008

4) Kędzior Z., Korcz K., „Badania marketingowe w praktyce", PWE warszawa 2008

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

WIEDZA DLA GOSPODARKI

UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY

PROJEKTOWANIE BADAŃ MARKETINGOWYCH

Ewa Więcek-Janka Agnieszka Kujawińska

Projekt współfinansowany ze środków Unii Europejskiej w ramach
Europejskiego Funduszu Społecznego