STATYSTYKA

Twierdzenia graniczne

Rozkłady statystyk z próby

Twierdzenia graniczne

Jeżeli rozpatrujemy ciąg zmiennych losowych $\{X_1 ; X_2 ;...; X_n\}$, to zdarza się, że ich rozkłady przy $n \to \infty$ mogą być zbieżne do jednego rozkładu, który będzie nazywany rozkładem granicznym (asymptotycznym) dla tego ciągu zmiennych losowych.

„złote” twierdzenie Bernoulli’ego

inaczej tzw. prawo wielkich liczb,

wraz ze wzrostem liczby przeprowadzanych doświadczeń, z których każde może zakończyć się sukcesem lub porażką, zaobserwowana częstość sukcesu skupia się wokół pewnej stałej, równej prawdopodobieństwu sukcesu „p”

To twierdzenie stanowi podstawę statystycznej definicji prawdopodobieństwa

twierdzenie Moivre'a – Laplace'a

ciąg dystrybuant standaryzowanych zmiennych losowych o rozkładzie dwumianowym jest zbieżny do dystrybuanty rozkładu normalnego standaryzowanego N(0;1).

Z tego twierdzenia wynika, że ciąg niestandaryzowanych dystrybuant zm. losowych o rozkładzie dwumianowym jest zbieżny do rozkładu normalnego $N(np;\sqrt{np(1-p)})$

gdzie n oraz p są parametrami rozkładu dwumianowego

centralne twierdzenie graniczne Lindenberg'a-Levy'ego,

mówi, że zmienne, których wartości kształtują się pod wpływem złożenia wielkiej liczby czynników losowych mają rozkład normalny

Centralne twierdzenie graniczne

jeżeli X_k jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie i skończonej wariancji to ciąg zmiennych losowych zdąża przy n→∞

$$Z_n = \sum_{i=1}^{n} X_i$$

do rozkładu normalnego

$$N(nEX; \sqrt{n}DX)$$

Sumę można zastąpić średnią, wtedy ciąg zmiennych losowych

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

zdąża do rozkładu normalnego

$$N(EX; \frac{DX}{\sqrt{n}})$$

Rozkłady statystyk z próby

Rozkłady statystyk z próby

Próba	Parametr	Populacja
$\bar{x}$	średnia arytmetyczna –wartość oczekiwana	EX ; m
S^2	wariancja	D^2X ; σ^2
S	odchylenie standardowe	DX ; σ
w	częstość empiryczna - prawdopodobieństwo	p

Rozkłady statystyk z próby

Jeżeli znany jest rozkład statystyki z próby to na tej podstawie można szacować wartości nieznanych parametrów populacji. Znajomość rozkładów statystyk z próby jest zatem niezbędna we wnioskowaniu statystycznym.

Rozkłady statystyk z próby, w których parametrem jest liczba stopni swobody (zależna od liczebności próby) nazywane są dokładnymi i są wykorzystywane w przypadku małych prób.

Jeżeli znalezienie dokładnego rozkładu statystyki nie jest możliwe, wykorzystywane są rozkłady graniczne statystyk , ale wtedy wymagana jest duża próba.

Średnia - wartość oczekiwana

X~N(m,σ) - nie znana σ

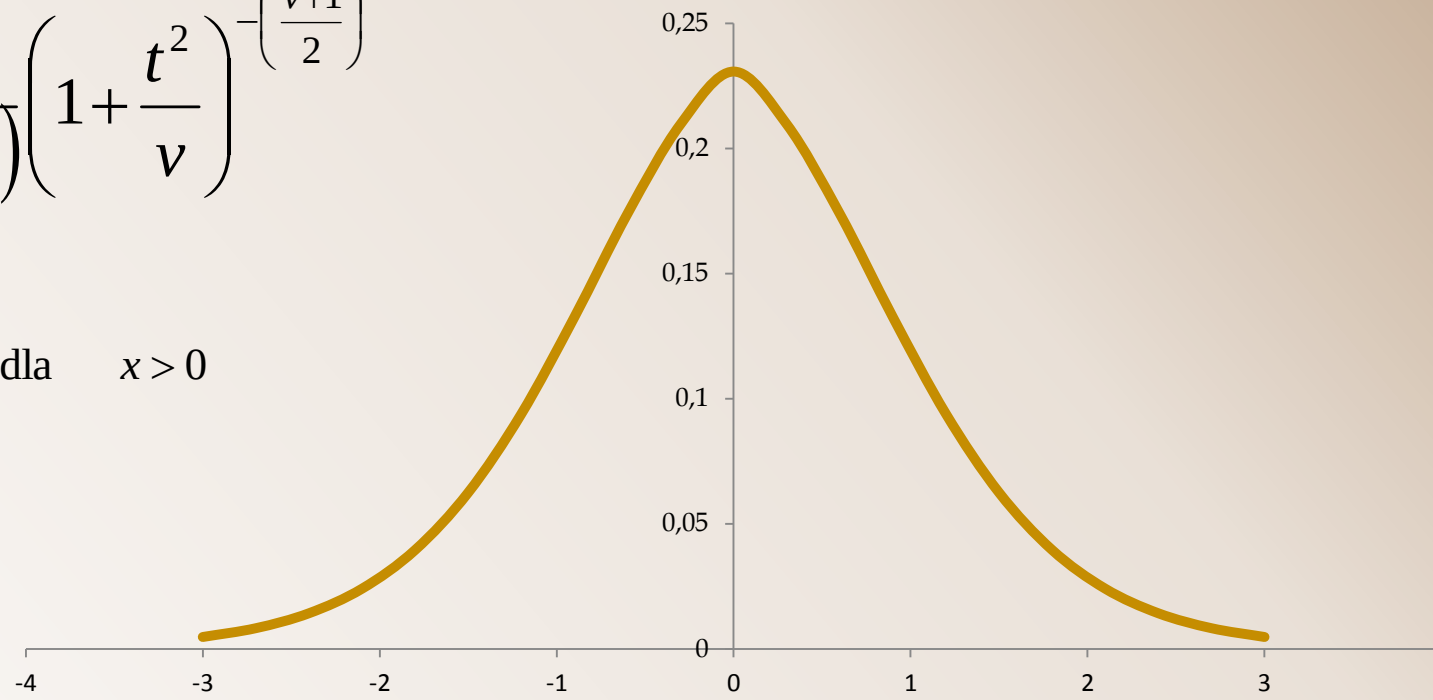
$$\frac{\bar{x}_N - m}{S} \cdot \sqrt{N} \sim t_{(N-1)}$$

ma rozkład **t**-Studenta z lss (v) = (N – *1)*

Rozkład t-Studenta

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{N\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

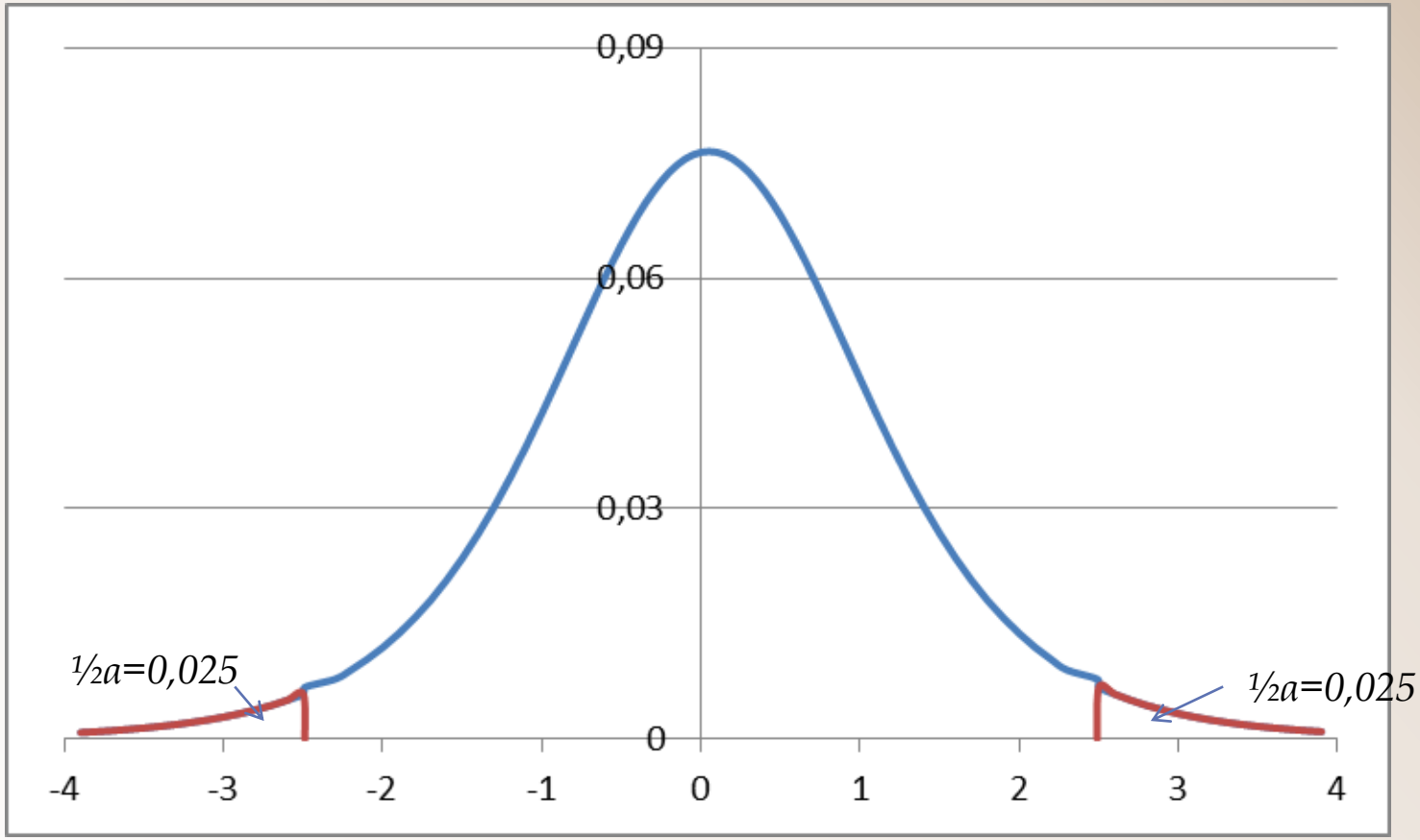
$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-1}dt \qquad \text{dla} \qquad x > 0$$



$$Et = 0$$

$$D^2 t = \frac{\nu}{\nu - 2} = \frac{N-1}{N-3}$$

Rozkład t-Studenta



F(t=2,447)=0,975

Średnia - wartość oczekiwana

Jeżeli zmienna ma dowolny rozkład to na mocy centralnego twierdzenia granicznego, dla dużych prób:

$$\bar{x}_{N\to\infty} \sim N(EX, \frac{DX}{\sqrt{N}})$$

różnica średnich – różnica wartości oczekiwanych

Jeśli $X_1 \sim N(m_1,\sigma_1)$ oraz $X_2 \sim N(m_2,\sigma_2)$ - odchylenia standardowe są nieznane, to wyrażenie zawierające różnicę średnich dwóch prób ma rozkład t-Studenta z liczbą st. swobody $\nu = N_1 + N_2 - 2$

$$\frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{S_{\bar{x}_1 - \bar{x}_2}} \sim t_{(N_1 + N_2 - 2)}$$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

różnica średnich – różnica wartości oczekiwanych

Jeśli zmienne X_1 oraz X_2 są zmiennymi losowymi o dowolnym rozkładzie to na mocy centralnego twierdzenia granicznego dla dużych prób rozkład różnicy dwóch średnich arytmetycznych jest rozkładem normalnym:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(EX_1 - EX_2 ; \sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}} \right)$$

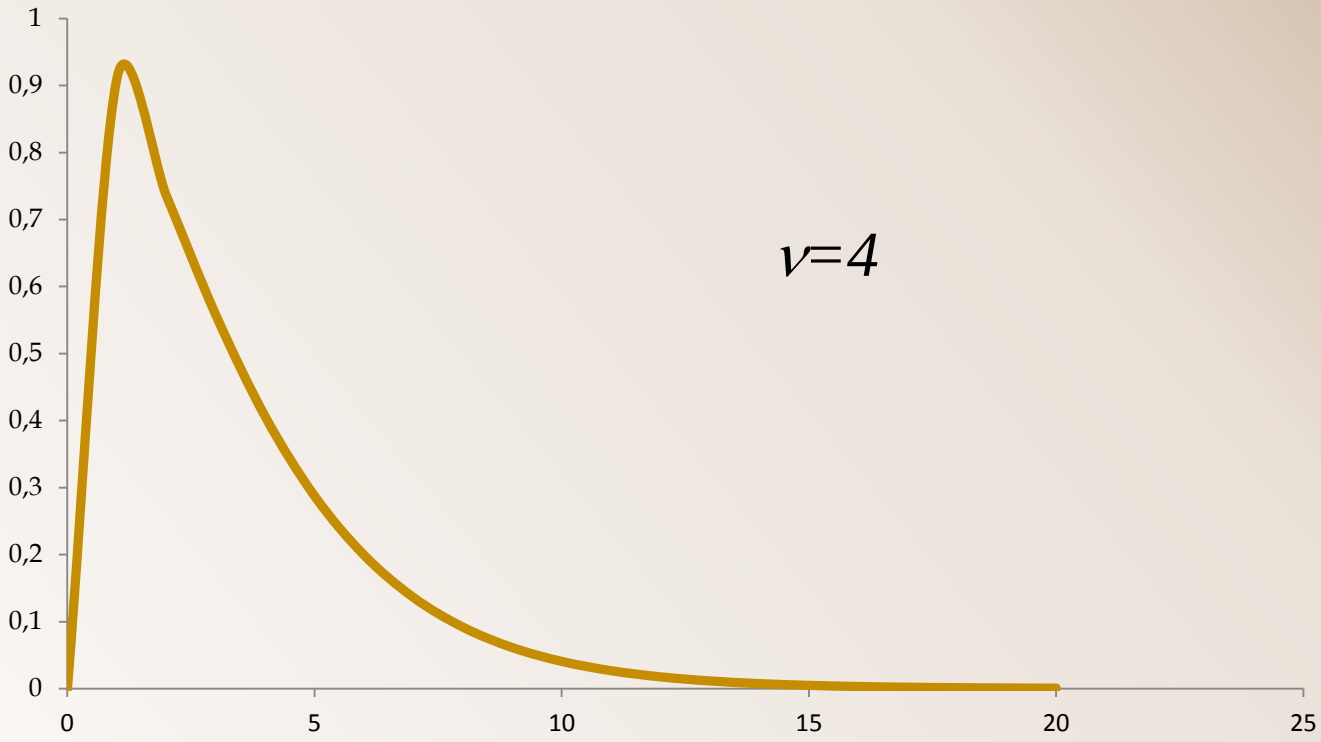
Wariancja w próbie – wariancja w populacji

Jeżeli zmienna losowa X ma rozkład normalny ($X \sim N(m, \sigma)$), to dla dowolnej N-elementowej próby poniższa statystyka

ma rozkład Chi-kwadrat Pearson'a

$$\frac{(N-1)\cdot S^2}{\sigma^2} \sim \chi^2_{(N-1)}$$

Rozkład chi-kwadrat



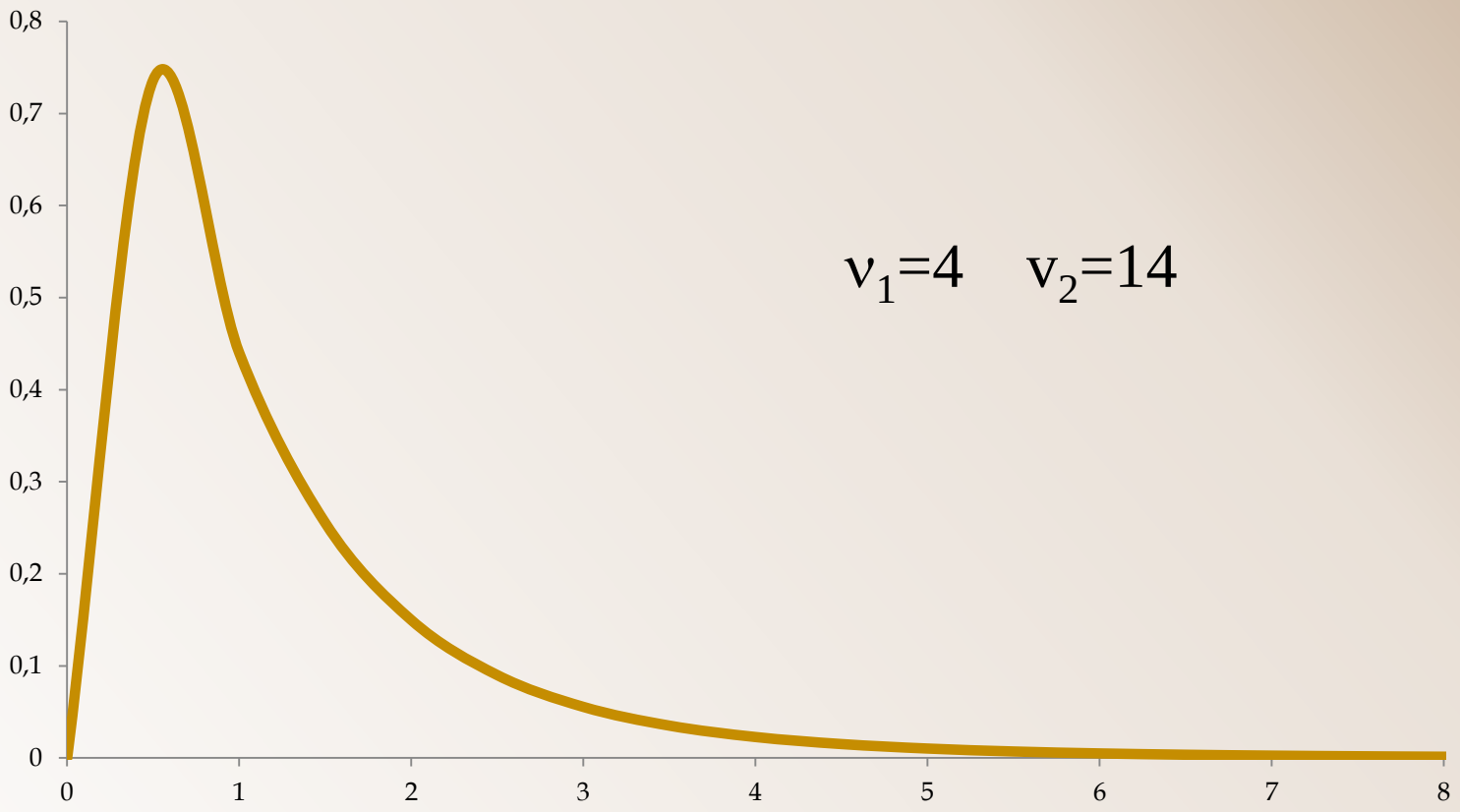
$E\chi^2 = \nu$ $\quad$ $D^2\chi^2 = 2\nu$

Iloraz wariancji dwóch prób i iloraz wariancji dwóch populacji

Jeśli zmienna losowa $X_1 \sim N(m_1,\sigma_1)$ oraz zmienna losowa $X_2 \sim N(m_2,\sigma_2)$ to iloraz wariancji dwóch prób o liczebnościach N_1 i N_2 pobranych z dwóch populacji ma rozkład F –Snedecora z lss : N_1-1 oraz N_2-1,

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(N_1 - 1;\, N_2 - 1)}$$

Rozkład F - Snedecora



$$EF = \frac{\nu_2}{\nu_2 - 2}$$

$$D^2F = \frac{2 \cdot \nu_2^{\,2} \cdot (\nu_1 + \nu_2 - 2)}{\nu_1 \cdot (\nu_2 - 2)^2 (\nu_2 - 4)}$$

Częstość empiryczna – prawdopodobieństwo

Jeżeli próba jest duża (co najmniej 100-120 elementów) i obserwujemy w niej cechę o rozkładzie dwupunktowym, to częstość empiryczna sukcesu $\left(w = \frac{m}{N} \right)$

na mocy omówionych twierdzeń granicznych, będzie miała rozkład normalny:

$$w \sim N\left(p; \sqrt{\frac{w \cdot (1-w)}{N}} \right)$$

a po standaryzacji:

$$\frac{w-p}{\sqrt{\frac{w \cdot (1-w)}{N}}} \sim N(0;1)$$

różnica częstości empirycznych – różnica prawdopodobieństw

Jeżeli próby pochodzące z dwóch populacji są duże (co najmniej 100-120 elementów w każdej) i w każdej populacji obserwujemy tę samą cechę o rozkładzie dwupunktowym, to różnica częstości empirycznych sukcesów (w_1-w_2), na mocy omówionych twierdzeń granicznych, będzie miała rozkład normalny:

$$w_1 - w_2 \sim N\left(p_1 - p_2 ; \sqrt{\overline{w} \cdot (1 - \overline{w}) \cdot \left(\frac{1}{N_1} + \frac{1}{N_2} \right)} \right) \quad \text{gdzie} \quad \overline{w} = \frac{m_1 + m_2}{N_1 + N_2}$$

Wnioskowanie

Estymacja parametrów populacji

Estymacja parametrów populacji

Estymator T_N parametru populacji θ to statystyka z próby (czyli funkcja elementów próby), która służy do oszacowania nieznanej wartości parametru θ.

Estymacja punktowa: $\hat{\theta} = T_N \pm S_{T_N}$

Estymacja przedziałowa $P(\hat{\theta} = T_N \pm S_{T_N}) = 1 - \alpha$

Metody uzyskiwania estymatorów:

- momentów
- największej wiarygodności

Własności estymatora

- **nieobciążoność**

Estymator nazywamy nieobciążonym, wtedy, gdy wartość oczekiwana estymatora jest równa estymowanemu parametrowi,

czyli $E(T_N) = \theta$

- **zgodność**

Estymator jest zgodny, gdy jest stochastycznie zbieżny do szacowanego parametru, czyli:

$$\lim_{N \to \infty} P\left\{ \left| T_N - \theta \right| < \varepsilon \right\} = 1$$

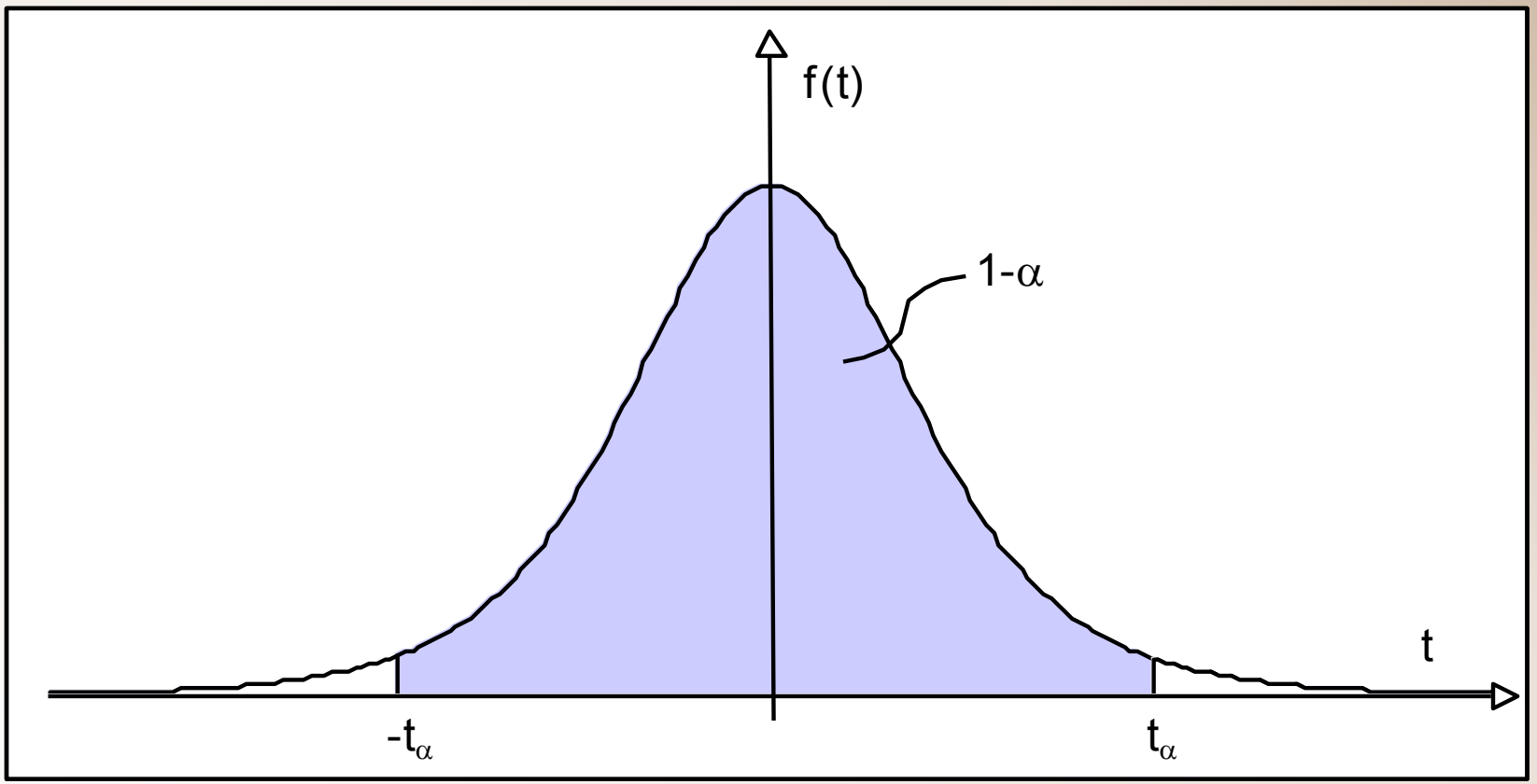
- **efektywność**

Estymator jest tym efektywniejszy im ma mniejszą wariancję

- **dostateczność**

Estymator jest dostateczny jeśli jest konstruowany na podstawie wszystkim informacji z próby

Konstrukcja przedziału ufności



$$P(-t_\alpha < t < t_\alpha) = 1 - \alpha$$

Konstrukcja przedziału ufności

$$P(-t_\alpha < t < t_\alpha) = 1 - \alpha$$

$$P(-t_\alpha < \frac{\bar{x} - m}{S} \cdot \sqrt{N} < t_\alpha) = 1 - \alpha$$

$$P(-t_\alpha \cdot \frac{S}{\sqrt{N}} < \bar{x} - m < t_\alpha \cdot \frac{S}{\sqrt{N}}) = 1 - \alpha$$

$$P(\bar{x} - t_\alpha \cdot \frac{S}{\sqrt{N}} < \hat{m} < \bar{x} + t_\alpha \cdot \frac{S}{\sqrt{N}}) = 1 - \alpha$$

Ocena precyzji oszacowania

Miarą precyzji estymacji jest współczynnik względnej precyzji, czyli iloraz połowy długości przedziału do wartości estymatora: .

$$\sigma_{\bar{x}} = t_{\alpha} \cdot \frac{S}{\bar{x} \cdot \sqrt{N}}$$

Przykład

Skonstruować 95% przedział ufności dla wartości oczekiwanej masy ciała cieląt w wieku 4 tygodni na podstawie losowo wybranej 9-cio elementowej próby, = 72 kg oraz S = 12 kg. Ocenić precyzję tego oszacowania.

$$P(72 - 2{,}306 \cdot \frac{12}{\sqrt{9}} < \hat{m} < 72 + 2{,}306 \cdot \frac{12}{\sqrt{9}}) = 1 - 0{,}05$$

$$P(72 - 9{,}224 < \hat{m} < 72 + 9{,}224) = 0{,}95$$

$$P(62{,}776 < \hat{m} < 81{,}224) = 0{,}95$$

$$\sigma_{\bar{x}} = 2{,}306 \cdot \frac{12}{72 \cdot \sqrt{9}} = 0{,}1281 = 12{,}81\%$$

Przykład

Czy skonstruowany 95% przedział ufności dla wartości oczekiwanej zawartości białka w mleku na podstawie losowo wybranej próby jest precyzyjny.

$$P(4{,}1 < \hat{m} < 5{,}3) = 0{,}95$$

$$\sigma_{\overline{x}} = \frac{(5{,}3 - 4{,}1) * 0{,}5}{(5{,}3 + 4{,}1)/2} = \frac{0{,}6}{4{,}7} = 12{,}7\%$$

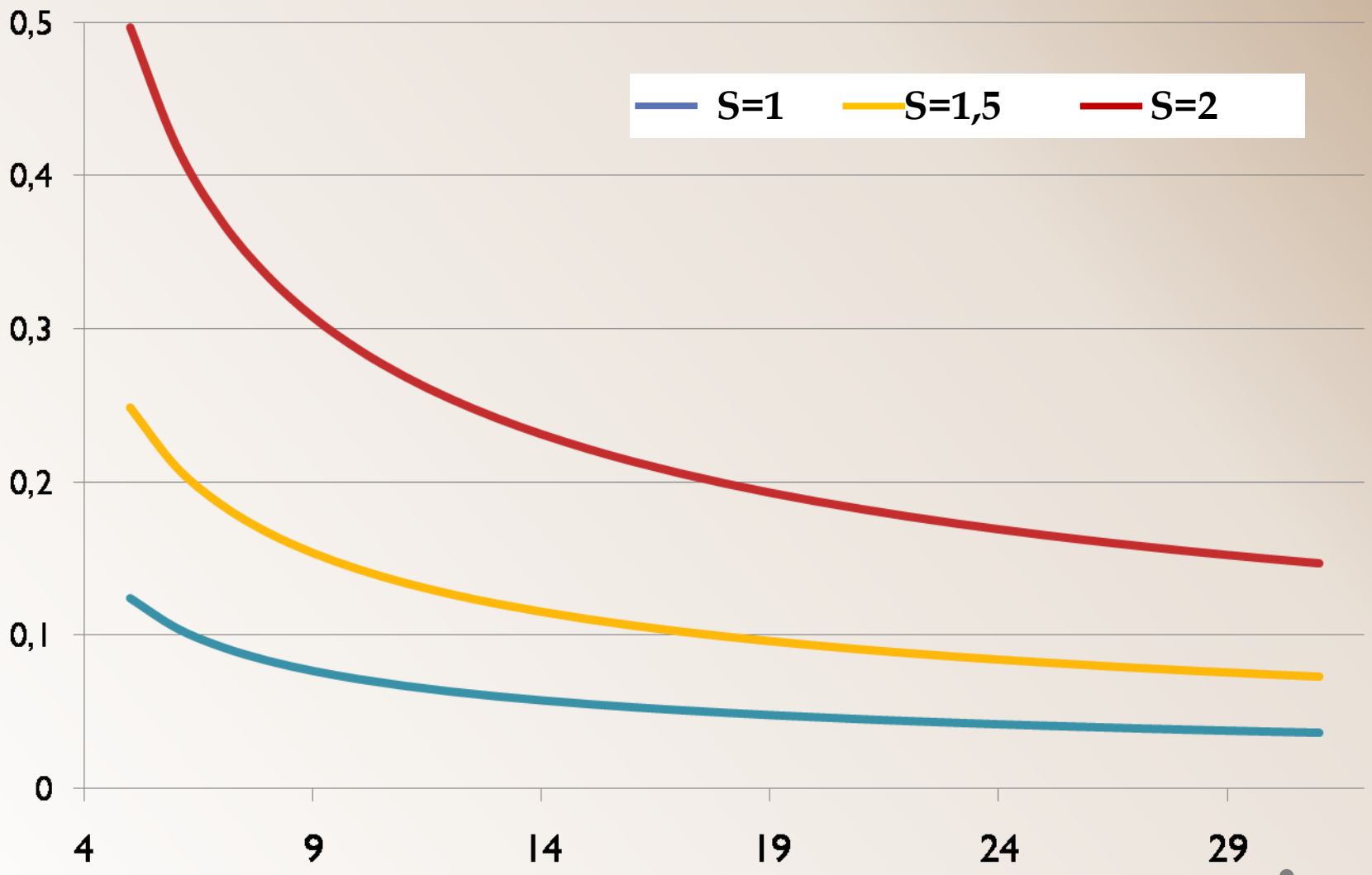
Przedział ufności dla prawdopodobieństwa

$$P\left(w - u_\alpha \cdot \sqrt{\frac{w \cdot (1-w)}{N}} < \hat{p} < w + u_\alpha \cdot \sqrt{\frac{w \cdot (1-w)}{N}}\right) = 1 - \alpha$$

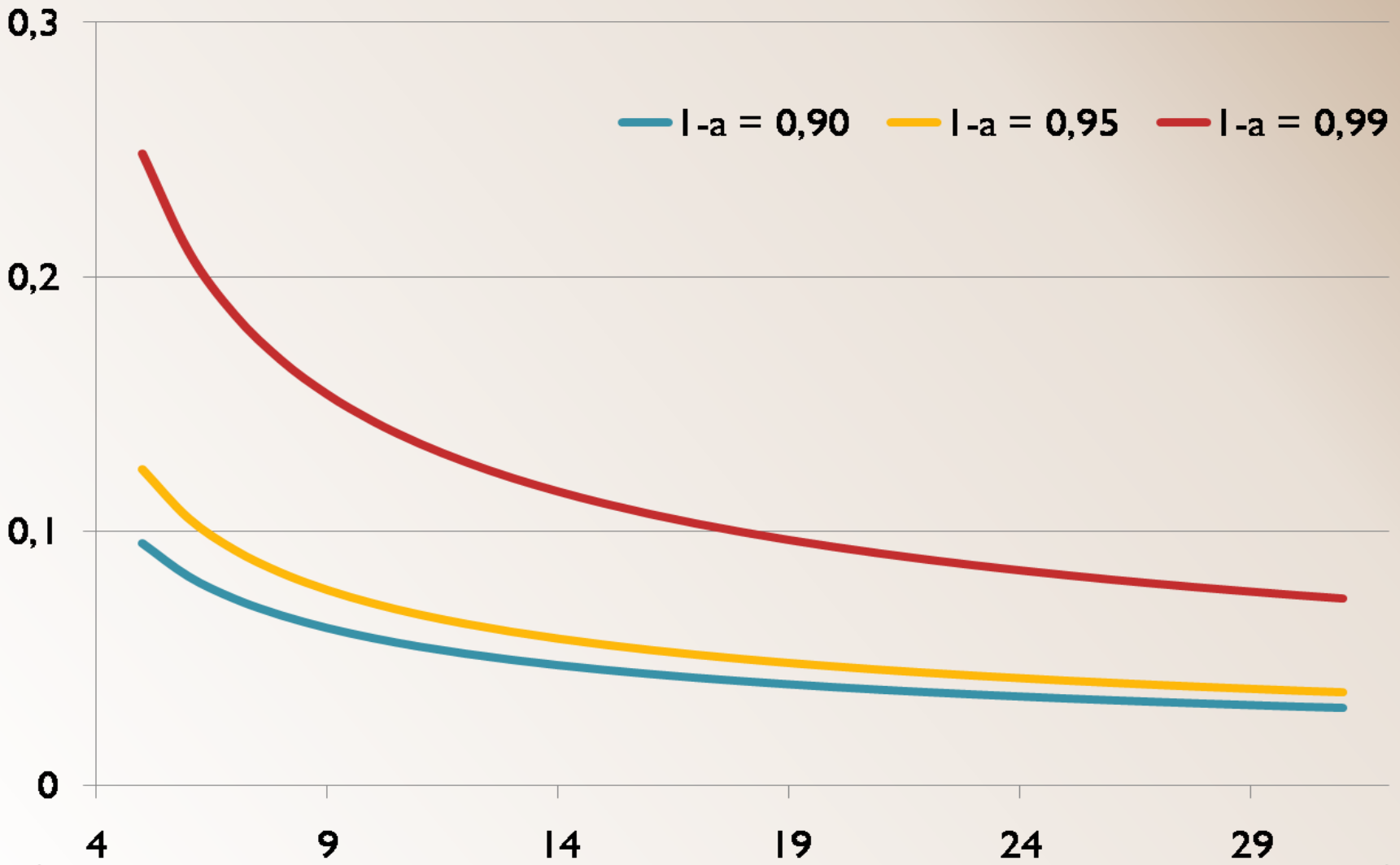
Precyzja oszacowania

$$\sigma_w = u_\alpha \cdot \sqrt{\frac{(1-w)}{w \cdot N}}$$

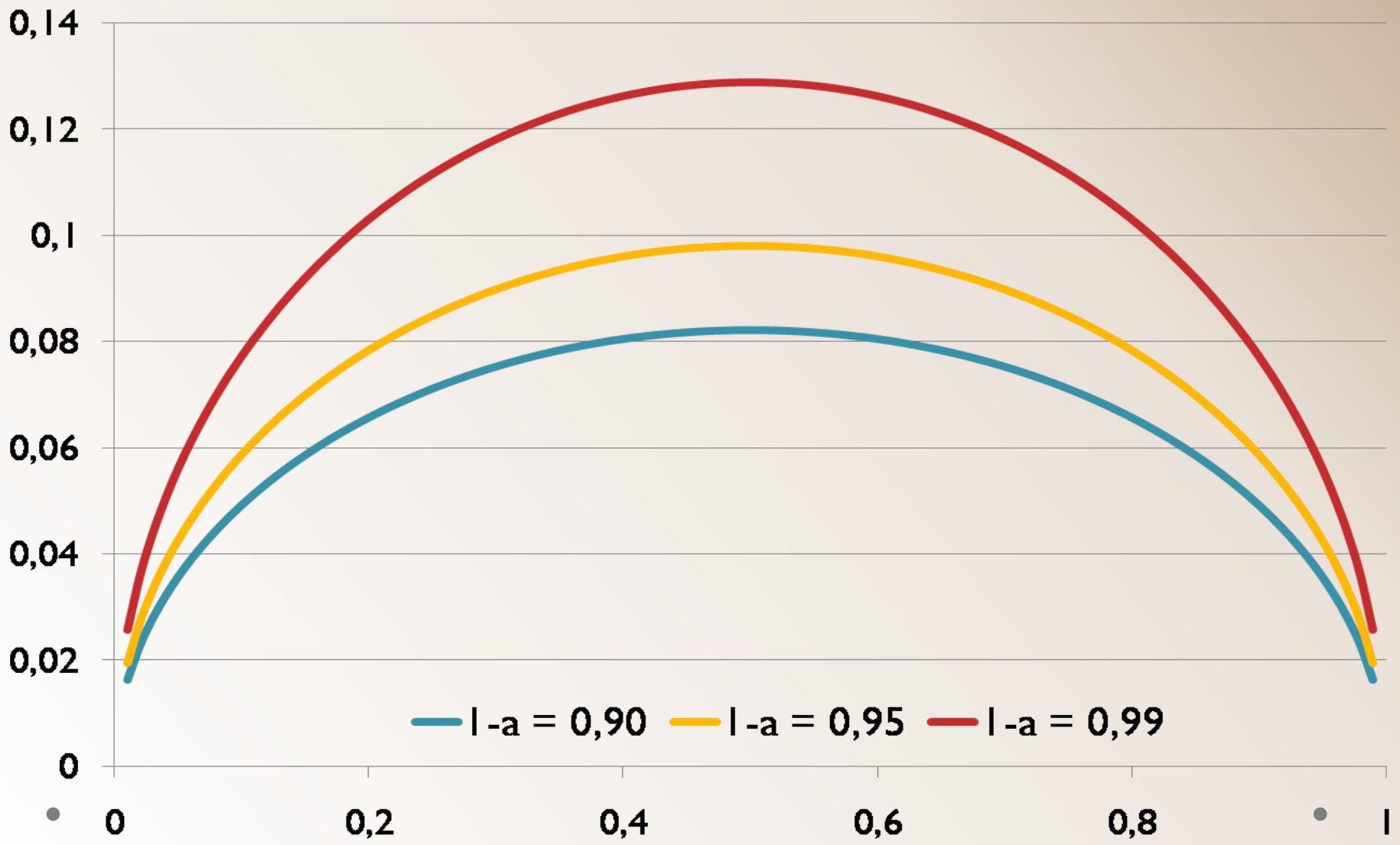
Precyzja przedziałowego oszacowania wartości oczekiwanej zależnie od liczebności próby i odchylenia standardowego



Precyzja przedziałowego oszacowania wartości oczekiwanej zależnie od liczebności i poziomu ufności



Długość przedziału ufności dla prawdopodobieństwa zależnie od częstości empirycznej w próbie



Wnioskowanie

Weryfikacja hipotez

Co nazywamy hipotezą

Każde stwierdzenie o parametrach rozkładu lub rozkładzie zmiennej losowej w populacji nazywać będziemy hipotezą statystyczną (H).

np.

Hipoteza statystyczna to dwie wartości oczekiwane dwóch populacji połączone jednym z operatorów: „=, >, <, ≠”.

Co nazywamy hipotezą

Hipoteza statystyczna jest zwykle formułowana na podstawie merytorycznego zagadnienia czy problemu.

np.

Hipoteza mówiąca, że obwód klatki piersiowej bydła rasy Shorthorn jest cechą o rozkładzie normalnym,

Hipoteza informująca, że wariancja masy ciała jelenia jest równa 400 kg^2

Podział hipotez

W zależności od ich treści hipotezy statystyczne możemy podzielić na

- parametryczne (mówiące o parametrach rozkładu)
- nieparametryczne (dotyczące rozkładu zmiennej losowej)

Podział hipotez

Hipotezy dzielimy też

- na weryfikowalne (oznaczane przez H_0) – takie do sprawdzenia prawdziwości których istnieją narzędzia
- nie weryfikowalne (oznaczane przez H_1)..

Weryfikacja Hipotezy

		Obiektywnie H_0	
		prawdziwa	fałszywa
decyzja	odrzucamy	błąd I rodzaju (α)	**właściwa**
	przyjmujemy	**właściwa**	błąd II rodzaju (β)

Testy istotności

		Obiektywnie H_0	
		prawdziwa	fałszywa
decyzja	odrzucamy	błąd I rodzaju (α)	**właściwa**
	~~przyjmujemy~~	~~**właściwa**~~	~~błąd II rodzaju (β)~~

Przebieg weryfikacji hipotezy

Sformułowanie hipotezy zerowej

H_0 : θ = A

Weryfikowana hipoteza z reguły zawiera znak równości między parametrem θ a założoną dla tego parametru wartością, co wynika z konstrukcji i wymagań testów statystycznych

Przebieg weryfikacji hipotezy

Dobranie hipotezy alternatywnej

H_1

Hipoteza alternatywna stanowi dopełnienie lub zaprzeczenie hipotezy zerowej. Mamy do wyboru jedną z trzech możliwości:

H_1 : $\theta \neq$ **A** (zaprzeczenie H_0)

albo

H_1 : $\theta >$ **A** lub **H_1 :** $\theta <$ **A** (uzupełnienie H_0)

Przebieg weryfikacji hipotezy

Próba

Wybór reprezentatywnej próby dla każdej populacji oraz scharakteryzowanie jej przy pomocy potrzebnych parametrów (średnia odch. standardowe itd.)

Przebieg weryfikacji hipotezy

Dobór testu statystycznego

Po wybraniu testu obliczana jest jego empiryczna wielkość : $\mathbf{test_{emp}}$

Wybór testu jest determinowany rodzajem hipotezy zerowej i próbą, głównie jej liczebnością

Przebieg weryfikacji hipotezy

Dobranie poziomu istotności α

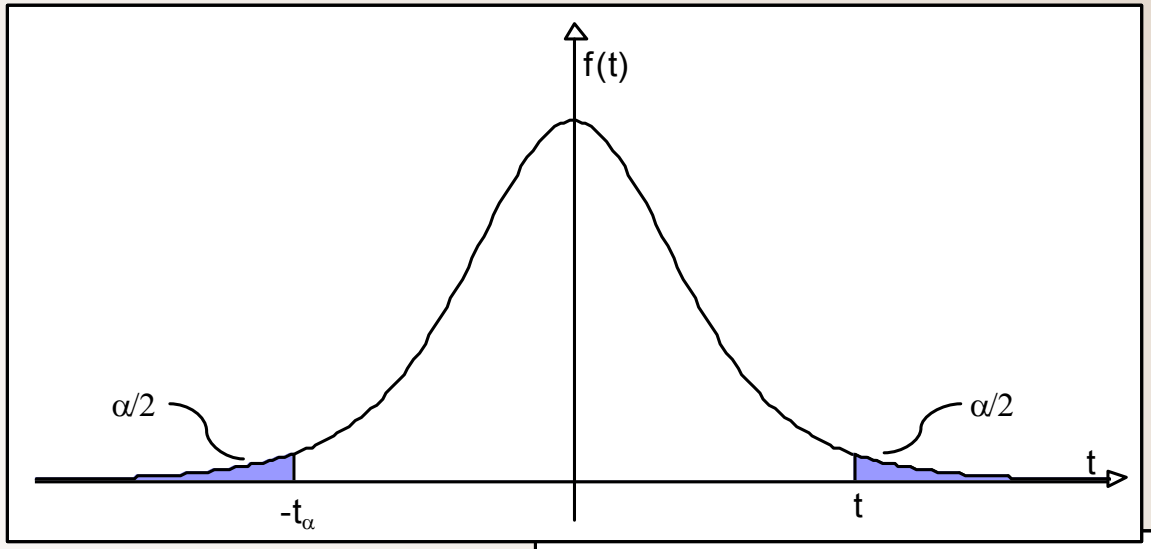
i określenie obszaru krytycznego ω

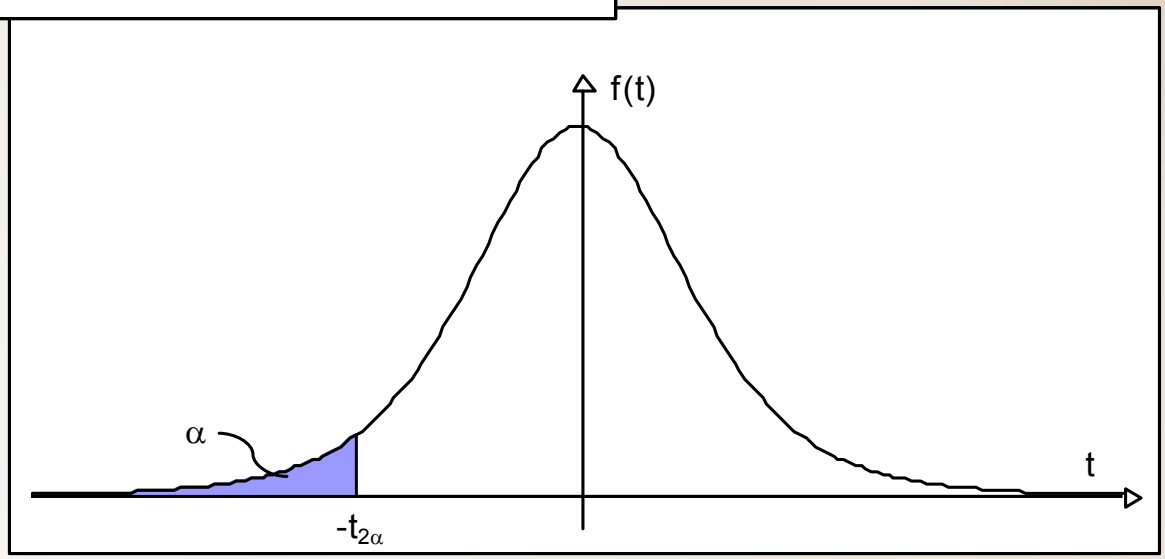
Obszar krytyczny to przedział lub suma przedziałów liczbowych wyznaczonych przez **wartości krytyczne.**

Pole pod funkcją rozkładu testu nad obszarem krytycznym jest równe poziomowi istotności α.

Położenie obszaru krytycznego zależy od wybranej uprzednio hipotezy alternatywnej

Obszar krytyczny





Obszar krytyczny

Rodzaj testu	granice obszaru krytycznego		
	dwustronnego	prawostronnego	lewostronnego
t-Studenta	$(-\infty ; -t_{\alpha}\rangle \cup \langle t_{\alpha} ; \infty)$	$\langle t_{2\alpha} ;\infty)$	$(-\infty ; -t_{2\alpha}\rangle$
u (normalny standaryzowany)	$(-\infty ; -u_{\alpha}\rangle \cup \langle u_{\alpha} ; \infty)$	$\langle u_{2\alpha} ;\infty)$	$(-\infty ; -u_{2\alpha}\rangle$
χ^2–Pearson'a	$(0 ; \chi^2_{1-\alpha/2}\rangle \cup \langle \chi^2_{\alpha/2} ; \infty)$	$\langle \chi^2_{\alpha} ;\infty)$	$(0 ; \chi^2_{1-\alpha}\rangle$
F–Snedecor'a	$(0 ; F_{1-\alpha/2}\rangle \cup \langle F_{\alpha/2} ; \infty)$	$\langle F_{\alpha} ;\infty)$	$(0 ; F_{1-\alpha}\rangle$

Podejmowanie decyzji

odrzucenie lub nie odrzucenie hipotezy zerowej H_0

Jeśli obliczona wartość testu ($test_{emp}$) należy do obszaru krytycznego hipoteza zastaje odrzucona

Przykład

H_0: m=6

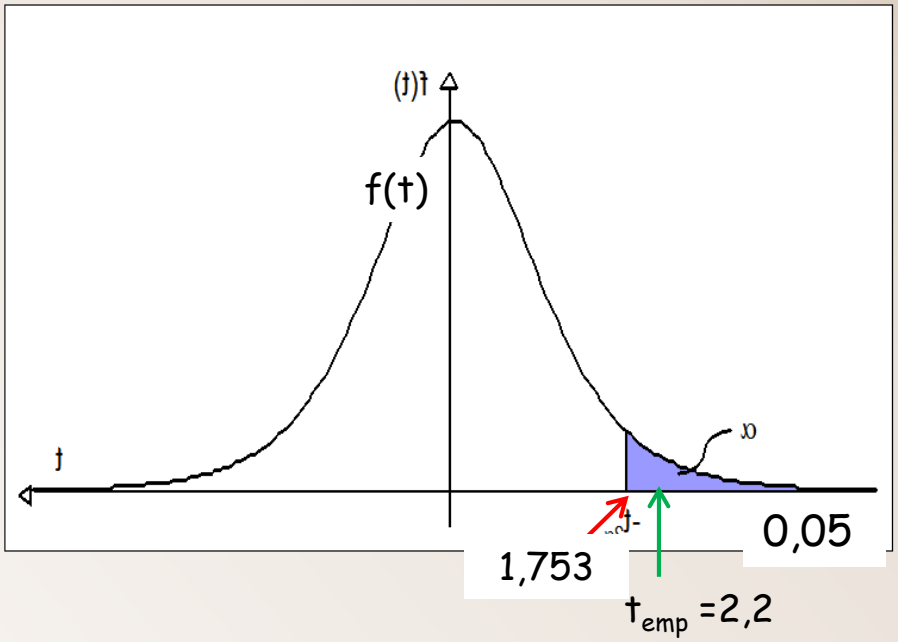
H_1: m>6

Próba: $\bar{x}$=7,1, N=16, S=2

t_{emp}=2,2

α=0,05

$t_{2\alpha;\,15}$ =1,753



„krytyczny" poziom istotności
dla α=0,022, $t_{2\alpha;\,15}$ = 2,2

Jaki test wybrać?

Test jest dopasowany do hipotezy zerowej oraz do cechy i próby. Bierze się pod uwagę wielkość próby i rozkład analizowanej cechy w populacji

Hipoteza dotycząca wartości oczekiwanej cechy:

$$\mathbf{H_0 : EX = EX_0}$$

Jeśli rozpatrywana cecha ma w populacji rozkład normalny, to hipotezę można zapisać (**$H_0 : m = m_0$**) i stosuje się test t-Studenta

$$t_{emp} = \frac{\bar{x} - m_0}{S}\sqrt{N}$$

Hipoteza dotycząca wartości oczekiwanej cechy:

$$\mathbf{H_0 : EX = EX_0}$$

Jeśli cecha nie ma rozkładu normalnego, to próba musi być duża i stosujemy test u

$$u_{emp} = \frac{\bar{x} - EX_0}{S}\sqrt{N}$$

Hipoteza dotycząca wartości oczekiwanych dwóch populacji

H_0: $EX_1=EX_2$

Jeśli cecha nie ma rozkładu normalnego, to próba musi być duża i stosujemy test u

$$u_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

Hipoteza dotycząca wartości oczekiwanych dwóch populacji

H_0: $EX_1 = EX_2$

Jeśli rozpatrywana cecha ma w populacji rozkład normalny, to hipotezę można zapisać (**$H_0 : m_1 = m_2$**) i stosuje się test t-Studenta

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2} \cdot \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}$$

Hipoteza dotycząca wartości oczekiwanych dwóch populacji

H_0: $m_1=m_2$

Warunkiem wyboru testu t-Studenta jest nieodrzucenie na poziomie istotności 0,05 Hipotezy o równości wariancji (H_0: $\sigma_1^2 = \sigma_2^2$)

W przeciwnym przypadku używamy test C-Cochrana

$$C_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

Z wartościami krytycznymi:

$$C_\alpha = \frac{t_{\alpha,\nu_1}\frac{S_1^2}{N_1} + t_{\alpha,\nu_2}\frac{S_2^2}{N_2}}{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}$$

gdzie $v_1=N_1-1$; $V_2=N_2-1$

Hipoteza dotycząca wariancji cechy w dwóch populacjach

$$\mathbf{H_0}: \sigma_1^2 = \sigma_2^2$$

Jeśli rozpatrywana cecha ma rozkład normalny, stosuje się test F-Snedecora

$$F_{emp} = \frac{S_1^2}{S_2^2}$$

Hipoteza dotycząca wariancji cechy w populacji

$$\mathbf{H_0}: \sigma^2 = \sigma_0^2$$

Jeśli rozpatrywana cecha ma w populacji rozkład normalny, stosuje się test chi-kwadrat

$$\chi^2_{emp} = \frac{(N-1)S^2}{\sigma_0^2}$$

Przykład

Pytanie – czy zawartość tłuszczu w mleku jest różna u kóz różnych ras?

Wylosowano po 13 kóz z każdej z dwóch ras, zmierzono zawartość tłuszczu w ich mleku i uzyskano: = 4,6%, S_1 = 0,15%, = 4,1%, S_2 = 0,19%

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

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

$$F_{emp} = \frac{(0{,}19)^2}{(0{,}15)^2} = 1{,}604$$

Obszar krytyczny ⟨2,69 ; ∞)

H_0 : $m_1 = m_2$

H_1 : $m_1 > m_2$

$$t_{emp} = \frac{4{,}6 - 4{,}1}{\sqrt{\frac{12 \cdot 0{,}0361 + 12 \cdot 0{,}0225}{24} \cdot \left(\frac{1}{13} + \frac{1}{13}\right)}} = 7{,}4473$$

Obszar krytyczny ⟨2,42 ; ∞)

Hipoteza dotycząca wartości prawdopodobieństwa sukcesu w populacji $\mathbf{H_0 : p = p_0}$

Hipotezę tę można weryfikować jedynie, gdy próba jest duża (N>100), gdyż stosowany test u wykorzystuje twierdzenie graniczne

$$u_{emp} = \frac{w - p_0}{\sqrt{\frac{p_0(1 - p_0)}{N}}}$$

Przykład

W celu sprawdzenia dokładności wskazań wagi zważono 5 razy tego samego psa i uzyskano następujące wyniki: 8,99 kg, 8,98 kg, 9,00 kg, 9,01 kg, 8,97 kg. Uznaje się wagę za wiarygodną, jeśli wariancja powtórzeń nie przekracza 0,0001 kg^2.

Należy zweryfikować

H_0 : σ^2 = 0,0001

H_1 : σ^2 > 0,0001 (prawostronna hipoteza alternatywna)

$$\bar{x} = \frac{1}{5} \cdot (8{,}99 + 8{,}98 + 9{,}00 + 9{,}01 + 8{,}97) = 8{,}99$$

$$S^2 = \frac{1}{4} \cdot \left[(8{,}99 - 8{,}99)^2 + (8{,}98 - 8{,}99)^2 + (9{,}00 - 8{,}99)^2 + (9{,}01 - 8{,}99)^2 + (8{,}97 - 8{,}99)^2\right] = 0{,}00025$$

$$\chi^2{}_{emp} = \frac{(5-1)0{,}00025}{0{,}0001} = 10$$

obszar krytyczny: ⟨9,488 ; ∞)

Hipoteza dotycząca wartości prawdopodobieństwa sukcesu w dwóch populacjach $\mathbf{H_0 : p_1 = p_2}$

Hipotezę tę można weryfikować gdy próby są duże, stosowany test u :

$$u_{emp} = \frac{w_1 - w_2}{\sqrt{\frac{\overline{w}(1-\overline{w})}{\overline{N}}}}$$

$$\overline{w} = \frac{m_1 + m_2}{N_1 + N_2} \qquad \overline{N} = \frac{N_1 \cdot N_2}{N_1 + N_2}$$

Hipotezy nieparametryczne

- O zgodności rozkładu cechy z rozkładem teoretycznym
- O niezależności dwóch cech
- O losowości wyboru próby

Hipotezy nieparametryczne

- O zgodności rozkładu cechy z rozkładem teoretycznym
- O niezależności dwóch cech
- O losowości wyboru próby

Hipoteza o zgodności rozkładu cechy z rozkładem teoretycznym

Rozkład teoretyczny pozwala wyznaczyć prawdopodobieństwa p_i w każdej klasie. Służą one do obliczenia teoretycznych liczebności (p_iN) a te porównywane są z empirycznymi (n_i). Testem do weryfikacji hipotez jest test zgodności Chi-kwadrat:

$$\chi^2_{emp} = \sum_{i=1}^{k} \frac{(n_i - p_i N)^2}{p_i N}$$

H_0: Rozkład grup krwi A:B:AB:0 jest jak 3:2:2:1

klasa	n_i	p_i	p_iN	$(n_i-p_iN)^2/p_iN$
A	53	0,375	50,25	0,150
B	25	0,250	33,50	2,157
AB	39	0,250	33,50	0,903
0	17	0,125	16,75	0,004
	134		Chi^2_{emp}=	**3,214**

α = **0,3598**

lss = 3

o zgodności rozkładu cechy z rozkładem teoretycznym

Test Kołmogorowa

Dla cech ciągłych i znanych parametrów rozkładu teoretycznego

$$\lambda_{emp} = \left(\sup\left|F_N(x_i) - F_{teor}(x_i)\right|\right) \cdot \sqrt{N}$$

$\lambda_{0,05} = 1,36;$
$\lambda_{0,01} = 1,63$

≠

Przykład

H_0 : X~N(15;2,5) (cecha X ma rozkład normalny o parametrach (15;2,5)

H_1 : X≠N(15;2,5) (cecha X ma rozkład normalny o parametrach 15;2,5)

wartość cechy	do 12	12 - 14	14 - 16	16 - 18	18 - 20	pow. 20
liczebność empiryczna (n_i)	10	26	56	64	30	14
częstość empiryczna (w_i)	0,05	0,13	0,28	0,32	0,15	0,07
dystrybuanta empiryczna $F_n(x_{i2})$	0,05	0,18	0,46	0,78	0,93	1
dystrybuanta teoretyczna $F(x_{i2})$	0,1151	0,3446	0,6554	0,8849	0,9773	1
moduł różnicy dystrybuant	0,0651	0,1646	**0,1954**	0,1049	0,0473	0

$$\lambda_{emp} = 0{,}1954 \cdot \sqrt{200} = 2{,}7634$$

Test Shapiro Wilka

$$W_{emp} = \frac{\left[\sum_{i=1}^{\left[\frac{N}{2}\right]} a_{N,i} \cdot (x_{(N-i+1)} - x_{(i)})\right]^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$a_{N,i}$ są wartościami odczytywanymi z tablic współczynników testu Shapiro-Wilka zależnymi od wielkości próby (N) oraz od numeru w uporządkowanej próbie (i).

Zapis $\left[\frac{N}{2}\right]$ oznacza część całkowitą liczby $\frac{N}{2}$

Współczynniki ($a_{N,i}$) testu Shapiro-Wilka

i \ N	3	4	5	6	7	8	9	10
1	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5700
2	,0000	,1677	,2413	,2806	,3031	,3164	,3244	,3291
3			,0000	,0875	,1401	,1743	,1976	,2141
4					,0000	,0561	,0947	,1224
5							,0000	,0399

i \ N	11	12	13	14	15	16	17	18	19	20
1	0,5601	0,5475	0,5359	0,5251	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734
2	,3315	,3325	,3325	,3318	,3306	,3290	,3273	,3253	,3232	,3211
3	,2260	,2347	,2412	,2460	,2495	,2521	,2540	,2553	,2561	,2565
4	,1429	,1586	,1707	,1820	,1878	,1939	,1988	,2027	,2059	,2085
5	,0695	,0922	,1099	,1240	,1353	,1447	,1524	,1587	,1641	,1686
6	,0000	,0303	,0539	,0727	,0880	,1005	,1109	,1197	,1271	,1334
7			,0000	,0240	,0433	,0593	,0725	,0837	,0932	,1013
8					,0000	,0196	,0359	,0496	,0612	,0711
9							,0000	,0013	,0303	,0422
10									,0000	,0140

Przykład

$(y_i - \hat{y}_i)$
-1,082
3,561
1,520
-0,878
-4,679
-3,561
-1,403
-2,036
1,638
1,281
-0,602
1,643
3,719
0,036
0,842

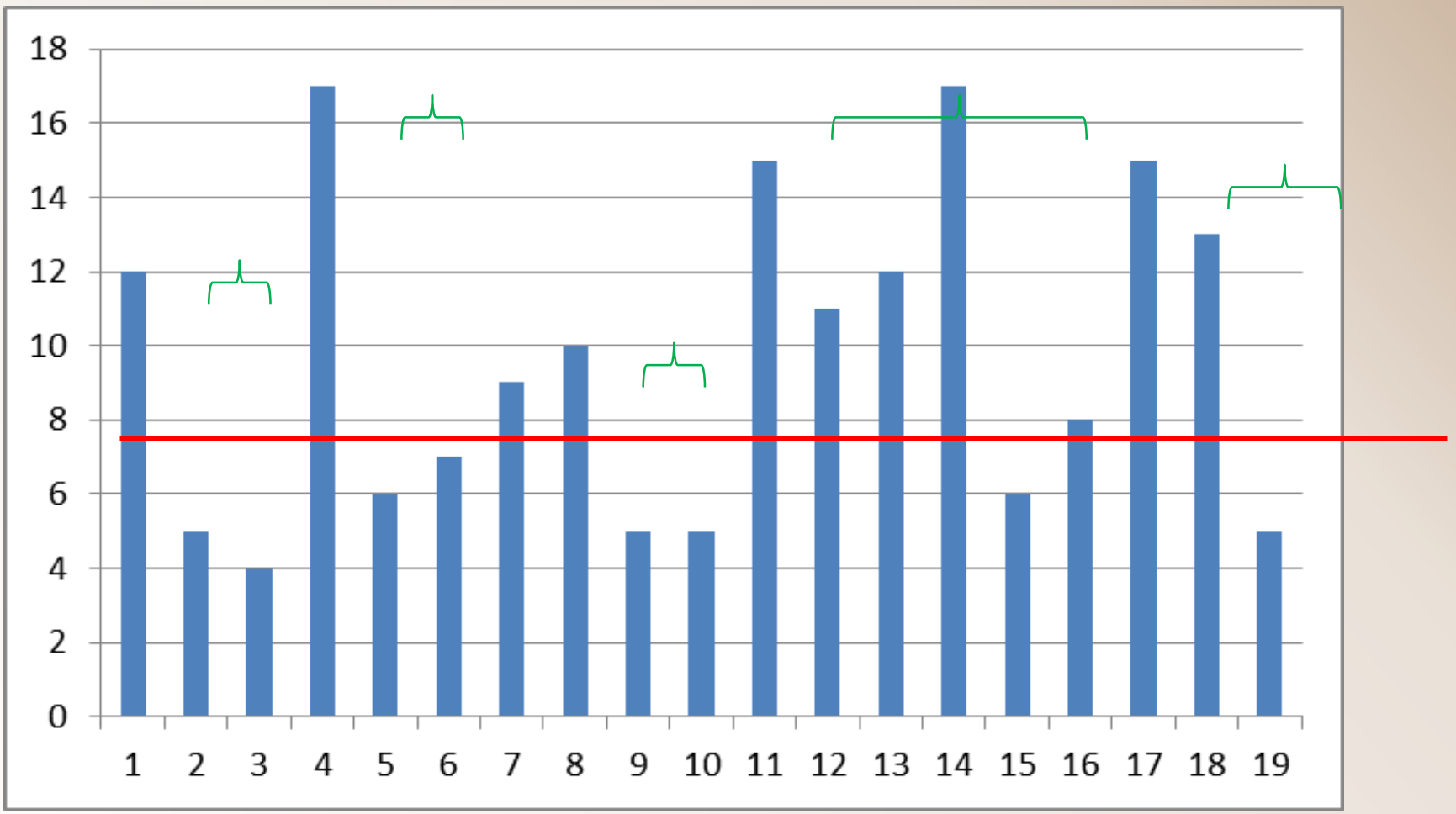
$$W_{emp} = \frac{1}{19{,}56}\big(0{,}5150\cdot(3{,}719+4{,}679)+0{,}3306\cdot(3{,}561+3{,}561)+0{,}2495\cdot(1{,}643+2{,}036)+$$
$$+0{,}1878\cdot(1{,}638+1{,}403)+0{,}1353\cdot(1{,}520+1{,}082)+0{,}0880\cdot(1{,}281+0{,}878)+$$
$$0{,}0433\cdot(0{,}842+0{,}602)\big) = \frac{(8{,}7730)^2}{19{,}56} = 3{,}935$$

$\omega = (0\ ;\ 0{,}881\rangle.$

Wartości krytyczne ($W_{\alpha,N}$) testu Shapiro-Wilka

N \ α	0,05	0,01	N \ α	0,05	0,01	N \ α	0,05	0,01
			11	0,850	0,792	21	0,908	0,873
			12	0,859	0,805	22	0,911	0,878
3	0,767	0,753	13	0,866	0,814	23	0,914	0,881
4	0,748	0,687	14	0,874	0,825	24	0,916	0,884
5	0,762	0,686	15	0,881	0,835	25	0,918	0,888
6	0,788	0,713	16	0,887	0,844	26	0,920	0,891
7	0,803	0,730	17	0,892	0,851	27	0,923	0,894
8	0,818	0,749	18	0,897	0,858	28	0,924	0,896
9	0,829	0,764	19	0,901	0,863	29	0,926	0,898
10	0,842	0,781	20	0,905	0,868	30	0,927	0,900

Losowość wyboru próby



Wartości krytyczne rozkładu serii $P(k \le k_{\alpha,n_1,n_2}) = \alpha$

α=0,05

n_2 \ n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3																			
4			2																
5		2	2	3															
6		2	3	3	3														
7		2	3	3	4	4													
8	2	2	3	3	4	4	5												
9	2	2	3	4	4	5	5	6											
10	2	3	3	4	5	5	6	6	6										
11	2	3	3	4	5	5	6	6	7	7									
12	2	3	4	4	5	6	6	7	7	8	8								
13	2	3	4	4	5	6	6	7	8	8	9	9							
14	2	3	4	5	5	6	7	7	8	8	9	9	10						
15	2	3	4	5	6	6	7	8	8	9	9	10	10	11					
16	2	3	4	5	6	6	7	8	8	9	10	10	11	11	11				
17	2	3	4	5	6	7	7	8	9	9	10	10	I1	11	12	12			
18	2	3	4	5	6	7	8	8	9	10	10	11	1I	12	12	13	13		
19	2	3	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14	
20	2	3	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14	14	15

α=0,025

n_2 \ n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3																			
4																			
5			2	2															
6		2	2	3	3														
7		2	2	3	3	3													
8		2	3	3	3	4	4												
9		2	3	3	4	4	5	5											
10		2	3	3	4	5	5	5	6										
11		2	3	4	4	5	5	6	6	7									
12	2	2	3	4	4	5	6	6	7	7	7								
13	2	2	3	4	5	5	6	6	7	7	8	8							
14	2	2	3	4	5	5	6	7	7	8	8	9	9						
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10					
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11				
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11			
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12		
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	
20	2	3	4	5	6	6	7	8	9	9	10	10	12	12	13	13	13	13	14

α=0,95

n_2 \ n_1	2	3	4	5	6	7	8	9	10	I1	12	13	14	15	16	17	18	19	20
2	4																		
3	5	6																	
4	5	6	7																
5	5	7	8	8															
6	S	7	8	9	10														
7	5	7	8	9	10	11													
8	5	7	9	10	11	12	12												
9	5	7	9	10	11	12	13	13											
10	5	7	9	10	11	12	13	14	15										
11	5	7	9	11	12	13	14	14	15	16									
12	5	7	9	11	12	13	14	15	16	16	17								
13	5	7	9	11	12	13	14	15	16	17	17	18							
14	5	7	9	11	12	13	15	16	16	17	18	19	19						
15	5	7	9	11	13	14	15	16	17	18	18	19	20	20					
16	5	7	9	11	13	14	15	16	17	18	19	20	20	21	22				
17	5	7	9	11	13	14	15	16	17	18	19	20	21	21	22	23			
18	5	7	9	11	13	14	15	17	18	19	20	20	21	22	23	23	24		
19	5	7	9	11	13	14	15	17	18	19	20	21	22	22	23	24	24	25	
20	5	7	9	11	13	14	16	17	18	19	20	21	22	23	24	24	25	26	26

α=0,975

n_2 \ n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	4																		
3	5	6																	
4	5	7	8																
5	5	7	8	9															
6	5	7	8	9	10														
7	5	7	9	10	11	12													
8	5	7	9	10	11	12	13												
9	5	7	9	11	12	13	13	14											
10	5	7	9	11	12	13	14	15	15										
11	5	7	9	11	12	13	14	15	16	16									
12	5	7	9	11	12	13	15	15	16	17	18								
13	5	7	9	11	13	14	15	16	17	18	18	19							
14	5	7	9	11	13	14	15	16	17	18	19	19	20						
15	5	7	9	11	13	14	15	17	17	18	19	20	21	21					
16	5	7	9	11	13	15	16	17	18	19	20	20	21	22	22				
17	5	7	9	11	13	15	16	17	18	19	20	21	22	22	23	24			
18	5	7	9	11	13	15	16	17	18	19	20	21	22	23	24	24	25		
19	5	7	9	11	13	15	16	17	19	20	21	22	22	23	24	25	25	26	
20	5	7	9	11	13	15	16	17	19	20	21	22	23	24	24	25	26	26	27

Grus grus