Rozkłady statystyk z próby

Statystyka

Rozkłady statystyk z próby

Próba losowa pobrana z populacji stanowi realizacje zmiennej losowej jak ciąg N zmiennych losowych (X_1, X_2,..... X_N) niezależnych i mających ten sam rozkład jak rozkład zmiennej losowej w populacji.

Statystyką z próby nazywamy zmienną losową (np. Z_N), będącą funkcją zmiennych X_1, X_2,..... X_N. Statystykami z próby są, na przykład, średnia arytmetyczna, wariancja oraz inne parametry.

Rozkład statystyki z próby zależy od rozkładu zmiennych losowych X_1, X_2,..... X_N i wielkości próby.

Rozkłady statystyk z próby

Próba	Parametr	Populacja
$\overline{x}$	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) – znana σ

$$\bar{x}_N \sim N(m, \frac{\sigma}{\sqrt{N}})$$

a po standaryzacji wyrażenie

$$\frac{\bar{x}_N - m}{\sigma} \cdot \sqrt{N} \sim N(0,1)$$

Średnia - wartość oczekiwana

$X \sim N(m, \sigma)$ - nie znana σ

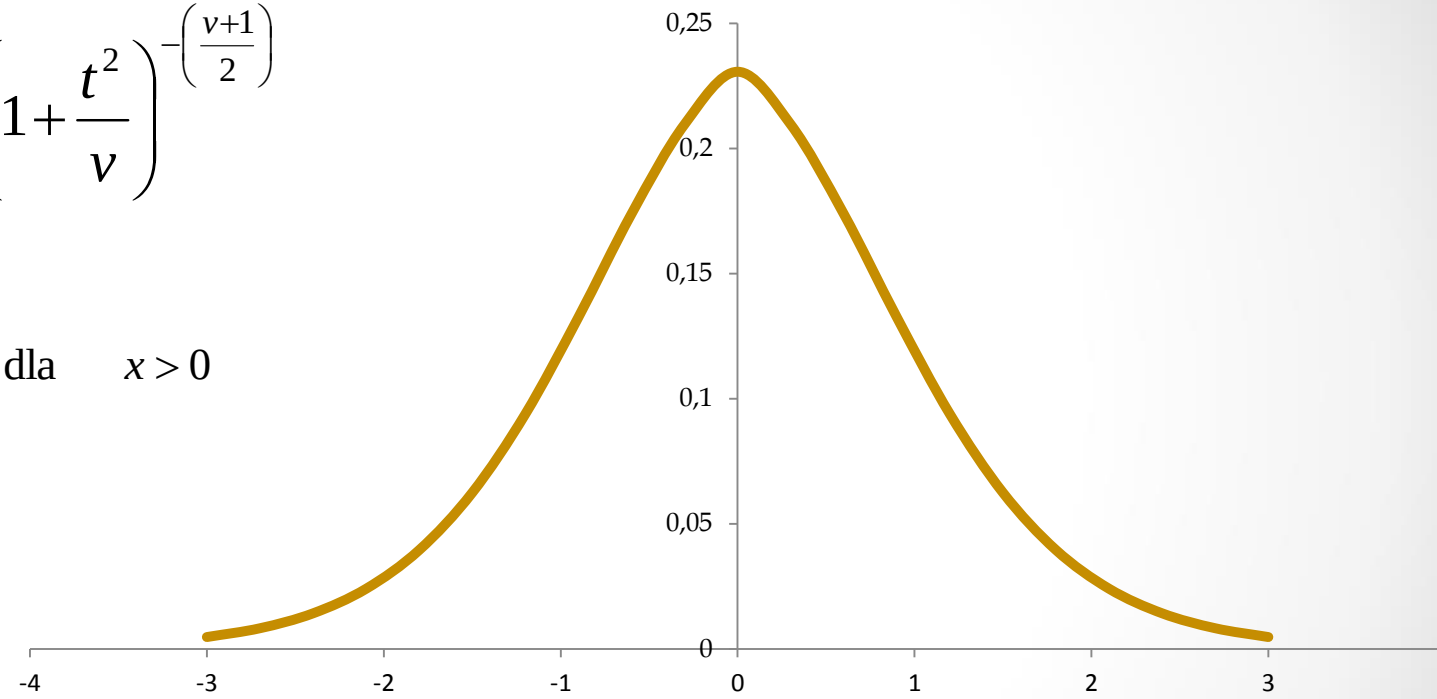
$$\frac{\bar{x}_N - m}{S} \cdot \sqrt{N} \sim t(N-1)$$

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

Rozkład t-Studenta

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

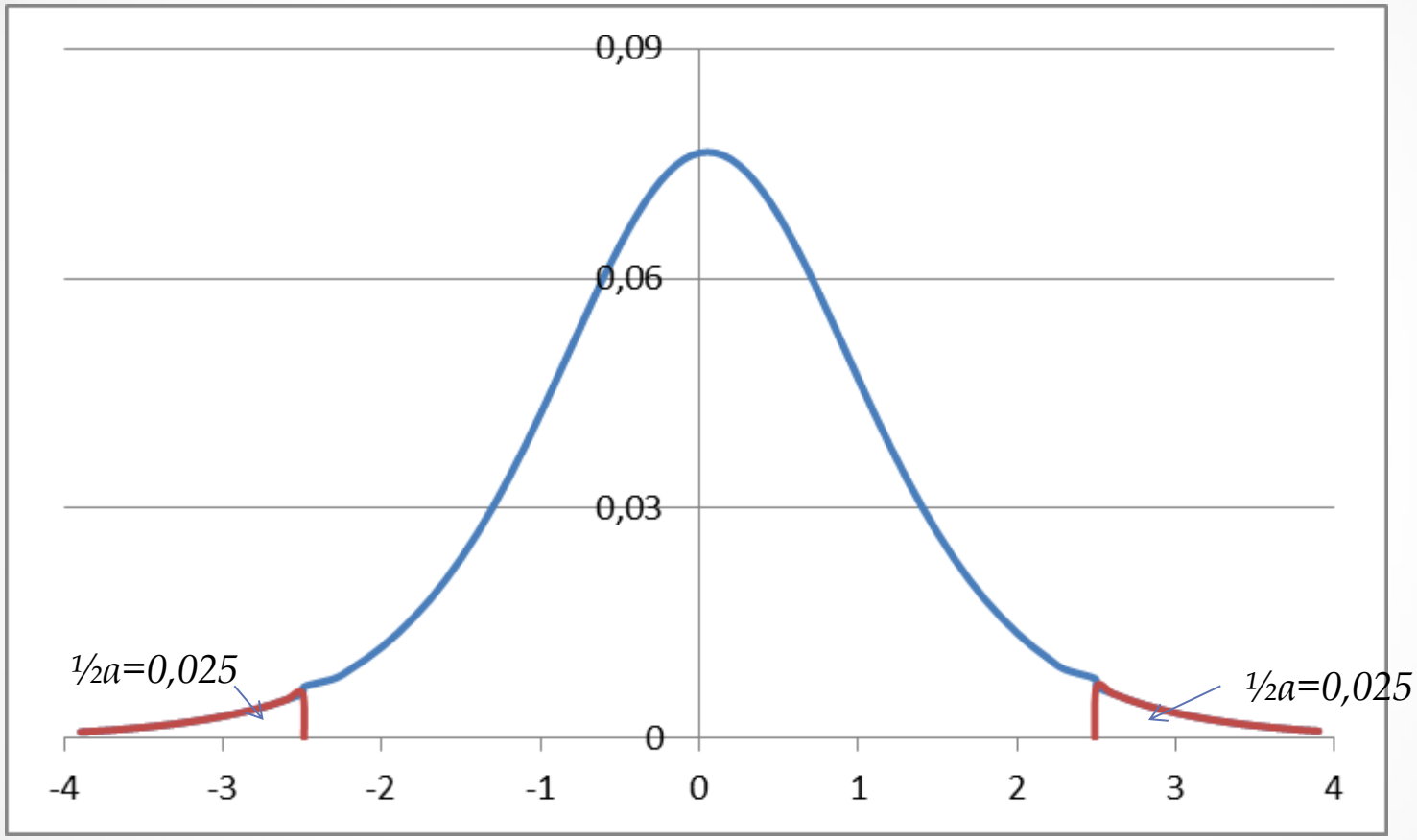
$$\Gamma(x)=\int_0^{\infty} t^{x-1}e^{-1}dt \qquad \text{dla} \qquad x>0$$



$$Et=0$$

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

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 $X_2 \sim N(m_2,\sigma_2)$ i znane są odchylenia standardowe obu rozkładów to różnica średnich prób ma rozkład normalny:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(m_1 - m_2; \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}\right)$$

a po standaryzacji

$$\frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim N(0;1)$$

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:

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

Przykład

Wysokość w kłębie koni rasy śląskiej ma rozkład normalny $X_1 \sim N(170 ; 5)$, a koni wielkopolskich $X_2 \sim N(168 ; 4)$. Obliczyć prawdopodobieństwo, że średnia arytmetyczna 9 elementowej próby wylosowanej z populacji koni śląskich jest większa o co najmniej 1 cm od średniej 16 elementowej próby wylosowanej z populacji koni wielkopolskich.

Poszukujemy: $P(\bar{x}_1 > \bar{x}_2 + 1) = P(\bar{x}_1 - \bar{x}_2 > 1)$

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2 = 170 - 168 = 2$$

$$D(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma^2{}_1}{N_1} + \frac{\sigma^2{}_2}{N_2}} = \sqrt{\frac{25}{9} + \frac{16}{16}} = \sqrt{34/9} \approx 1{,}944$$

$$P(\bar{x}_1 - \bar{x}_2 > 1) = 1 - F(\bar{x}_1 - \bar{x}_2 = 1) = 1 - F\left(U = \frac{1-2}{1{,}944}\right) = 1 - F(U = -0{,}514) =$$

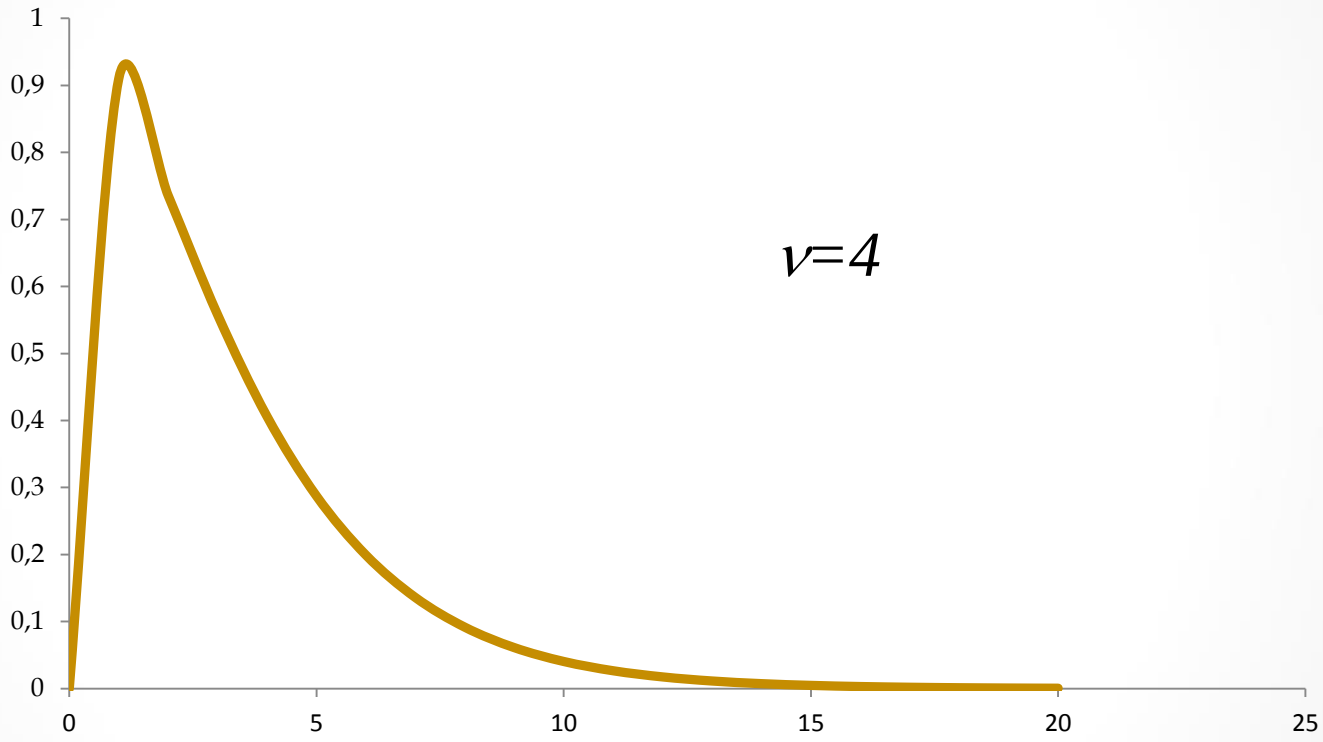
$$= 1 - (1 - F(U = 0{,}514)) = F(U = 0{,}514) = 0{,}6964$$

Wariancja w próbie – wariancja w populacji

Jeżeli zmienna losowa X ma rozkład normalny (X~N(m, σ)), 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



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

Przykład

Mierząc długość skór lisów zakłada się, że błąd pomiaru ma rozkład normalny N(μ=0 ; σ=0,5 cm). Obliczyć, jaka jest szansa, że wariancja w próbie złożonej z danych o długości dziesięciu skór nie przekroczy 0,15 cm^2.

statystyka $\frac{(N-1)\cdot S^2}{\sigma^2}$ zawierająca wariancję z próby i populacji

ma rozkład chi-kwadrat o 9 stopniach swobody

$$P(S^2 \le 0{,}15) = P\left(\frac{(N-1)\cdot S^2}{\sigma^2} \le \frac{9\cdot 0{,}15}{0{,}25}\right) = P(\chi^2_{N-1} \le 5{,}40) = F(\chi^2_{N-1} = 5{,}40) \approx 0{,}2$$

odchylenie standardowe w próbie i w populacji

Jeśli zmienna losowa X ma rozkład normalny (X~N(m,σ)) oraz próba jest duża, (licząca co najmniej 120 elementów), to odchylenie standardowe tej próby będzie miało rozkład normalny:

$$S \sim N(\sigma; \frac{\sigma}{\sqrt{2N}})$$

Dla nieznanego odchylenia standardowego populacji stosuje się przybliżenie:

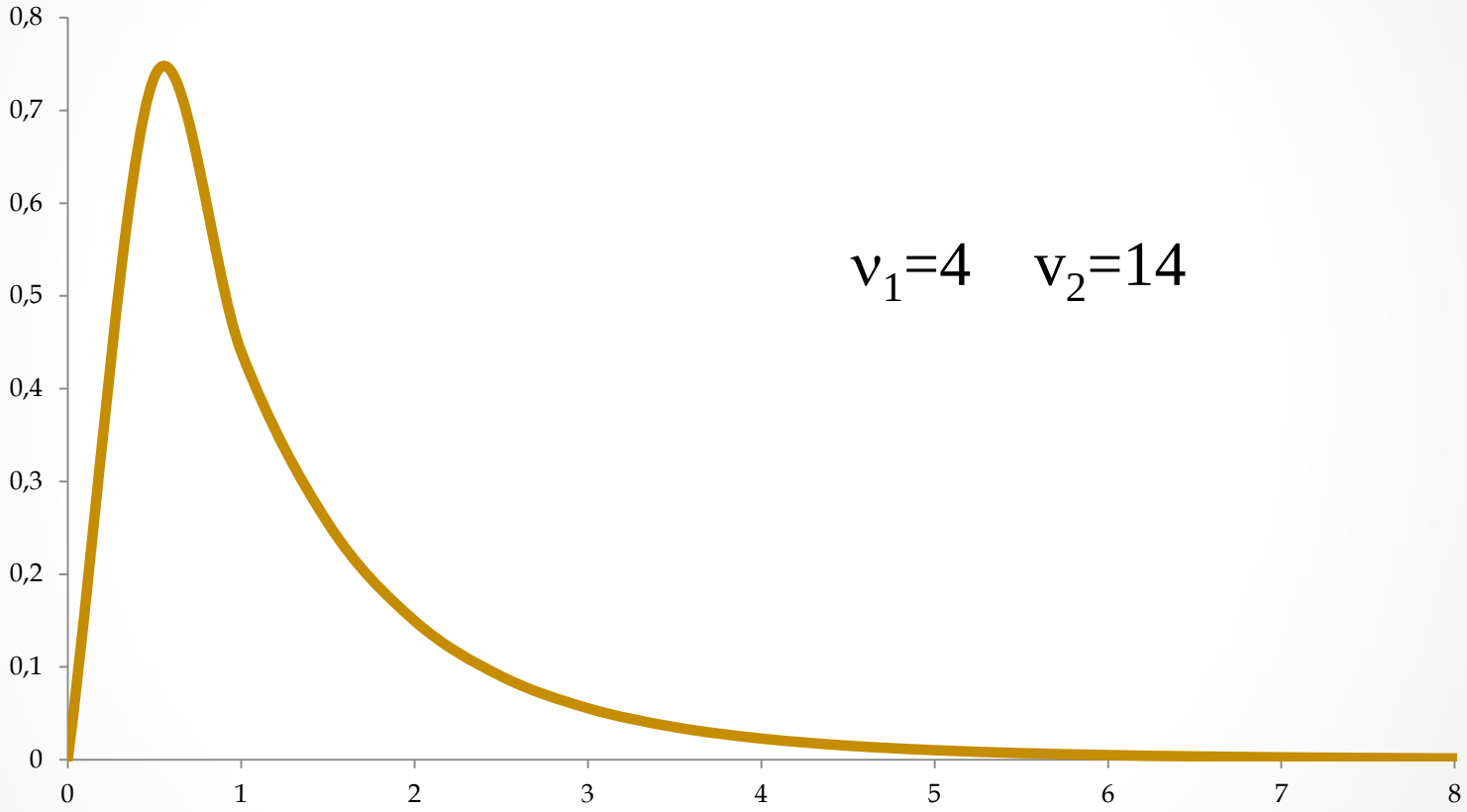
$$S \sim N(\sigma; \frac{S}{\sqrt{2N}})$$

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,

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F(\nu_1 = N_1 - 1; \nu_2 = N_2 - 1)$$

Rozkład F - Snedecora



ν_1=4 ν_2=14

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

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

Przykład

Wysokość w kłębie w populacji koni rasy śląskiej jest zmienną losową o rozkładzie normalnym $X_1 \sim N(170;5)$, a w populacji koni wielkopolskich $X_2 \sim N(168;4,47)$. Obliczyć prawdopodobieństwo, że wariancja 9-elementowej próby wylosowanej z populacji koni śląskich jest co najmniej pięciokrotnie większa niż wariancja 16-elementowej próby koni wielkopolskich

$$P\left(\frac{S_1^2}{S_2^2} > 5\right) = P\left(\frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} > 5 \cdot \frac{\sigma_2^2}{\sigma_1^2}\right) = P\left(F > 5 \cdot \frac{19,981}{25}\right) = P(F > 3,996) \approx 0,01$$

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:

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

Przykład

Wiadomo, że prawdopodobieństwo pojawienia się albinosa w populacji jest równe 0,06. Jaka jest szansa, aby wśród 200 młodych urodzonych na fermie pojawiło się co najmniej 15 albinosów.

$$P(w > 0{,}075) = P\left(\frac{w - p}{\sqrt{\frac{p \cdot (1-p)}{N}}} > \frac{0{,}075 - p}{\sqrt{\frac{p \cdot (1-p)}{N}}} \right) = P\left(U > \frac{0{,}075 - 0{,}06}{\sqrt{\frac{0{,}06 \cdot 0{,}94}{200}}} \right) =$$

$$= P(U > 0{,}893) = F(U = \infty) - F(U = 0{,}893) \approx 1 - 0{,}8141 = 0{,}1859$$

Bison bison