AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Statystyka i analiza danych - W2:

Podstawy wnioskowania statystycznego
Zmienne losowe, rozkład prawdopodobieństwa.
Parametry rozkładu. Estymatory punktowe i przedziałowe. Weryfikacja hipotez statystycznych.

Dr Anna ADRIAN
Paw B5, pok 407
adan@agh.edu.pl

Plan

- Badania statystyczne
- Populacja i próba statystyczna -
- Zmienne losowe
- Rozkłady prawdopodobieństwa zmiennych losowych
  - dyskretnej
  - ciągłej
- Parametry rozkładu prawdopodobieństwa zmiennej losowej
- Estymatory parametrów rozkładu: punktowe i przedziałowe
- Dopasowanie rozkładu teoretycznego do rozkładu empirycznego

Rodzaje badań statystycznych

Badania pełne

obejmują wszystkie elementy populacji,
np. na przeglądzie uzębienia danego pacjenta
można określić dokładną liczbę zębów i ich stan.

Badania częściowe – badania elementów próbki statystycznej,
mają szerokie zastosowania i są:

- konieczne w przypadku populacji nieskończonej,
- stosowane w populacjach skończonych bardzo licznych
- stosowane w przypadkach badań niszczących

Populacja i próbka statystyczna

<u>Populacja</u> jest to zbiór wszystkich elementów (danych) reprezentujących analizowany problem (zjawisko)

Może to być zbiór skończony, przeliczalny lub nieprzeliczalny.

<u>Próbka statystyczna</u> – to każdy zespół elementów wylosowanych z populacji, inaczej: jest to podzbiór właściwy badanej populacji, będący podstawą wnioskowania statystycznego o populacji.

Wybór próbki statystycznej z populacji

Losowy dobór próbki polega na tym, że o fakcie znalezienia się poszczególnych elementów populacji w próbce statystycznej decyduje przypadek (los).

Powinny być spełnione następujące dwa warunki;

- każda jednostka populacji ma dodatnie, znane prawdopodobieństwo znalezienia się w próbce
- istnieje możliwość ustalenia prawdopodobieństwa znalezienia się w próbce dla każdego zespołu elementów populacji

Wybór próbki reprezentatywnej

Od próbki wymaga się reprezentatywności, czyli aby z przyjętą dokładnością reprezentowała strukturę populacji.

O reprezentatywności decydują dwa czynniki:

- Liczebność próbki (n)
- Sposób wyboru elementów populacji do próbki
  - Wybór celowy, o przynależności do grupy decyduje badacz, stopień reprezentatywności zależy wyłącznie od jakości selekcji
  - Wybór losowy- każdy element populacji ma jednakową szansę znalezienia się w próbce z takim samym prawdopodobieństwem, stopień reprezentatywności rośnie wraz ze wzrostem liczebności grupy.

  Stosowane są dwie techniki losowania:
    - Losowanie niezależne (zwrotne)
    - Losowanie zależne (bezzwrotne)

O źródłach błędów w badaniach statystycznych

Badania, zarówno pełne jak i częściowe, zawsze obciążone są błędami związanymi z:

- organizacją eksperymentu,
- niedokładnością pomiarową,
- przetwarzaniem wyników,
- w badaniach częściowych z niedokładnością odwzorowania struktury populacji w strukturę próbki

Zmienna losowa jest to funkcja rzeczywista X, określona na zbiorze zdarzeń elementarnych Ω

$$X: \Omega \rightarrow W$$

Zmienne losowe zwykle oznacza się dużymi literami z końca alfabetu : X, Y, Z.

Wartości zmiennych losowych zwykle oznacza się małymi literami z końca alfabetu: x,y,z.

Definiowanie zmiennej losowej jest to przypisanie wartości (liczbowych) zdarzeniom elementarnym

Z partii wyrobów zawierającej wyroby dobre i wyroby wadliwe losuję jeden wyrób, wtedy

$$\Omega = \{\omega_d , \omega_w \}$$

gdzie

ω_d- oznacza wylosowanie wyrobu dobrego

ω_w- oznacza wylosowanie wyrobu wadliwego

Określam zmienną losową X w następujący sposób:

$$X(\omega_d)=1 \qquad X(\omega_w)=0$$

Rozkład prawdopodobieństwa zmiennej losowej dyskretnej

Jeżeli w przedstawionym przykładzie, dotyczącym kontroli jakości wyrobów, 90% wyrobów było dobrych, natomiast 10% stanowiły wybraki, to możemy mówić o prawdopodobieństwie zdarzeń:

$$P(\{\omega : X(\omega)=0\}) = 0{,}1$$

$$P(\{\omega : X(\omega)=1\}) = 0{,}9$$

(jest to tzw. „dwupunktowy" rozkład prawdopodobieństwa)

Rozkład prawdopodobieństwa zmiennej losowej X jest zbiorem par {x, p}, gdzie x jest wartością zmiennej X,
p- prawdopodobieństwem wystąpienia wartości x.

Tablicowy zapis rozkładu prawdopodobieństwa zmiennej losowej X

x_i	0	1
p_i	0,1	0,9

Dystrybuanta zmiennej losowej

Dystrybuantą, $F_X(x_0)$, zmiennej losowej X jest prawdopodobieństwo zdarzenia, polegającego na tym, że zmienna ta przyjmie wartości mniejsze od x_0.

$$\mathbf{F_X(x_0) = P(X < x_0)}$$

Dystrybuanta jest funkcją:

- określoną na zbierze liczb rzeczywistych;
- o wartościach z przedziału [0-1];
- niemalejącą
- prawostronnie ciągłą

Dystrybuantę zmiennej losowej X oznaczamy zwykle jako F_X

$$F_X(x_0) = P_X((-\infty, x_0)) = P(X<x_0)$$

$$P([a,b]) = P(a \leq X < b) = F_X(b) - F_X(a)$$

Zastosowanie teorii w praktyce – wyznaczanie rozkładu zmiennej losowej

Z partii wyrobów losujemy 3 sztuki.

Na rysunku pokazano

- przestrzeń możliwych zdarzeń
- sposób określania zmiennej losowej

Rozkład prawdopodobieństwa i dystrybuanta zmiennej losowej X

$p_1=P(\ X=0)=1/8, \qquad p_2=P(\ X=1)=3/8, \ \ldots\ldots$

Rozkład prawdopodobieństwa zmiennej losowej X				
x_i	0	1	2	3
p_i	1/8	3/8	3/8	1/8
F(x)	0	1/8	1/2	7/8

Dystrybuanta

$F_X(0) = P_X((-\infty,0)) = P(X<0) = 0$

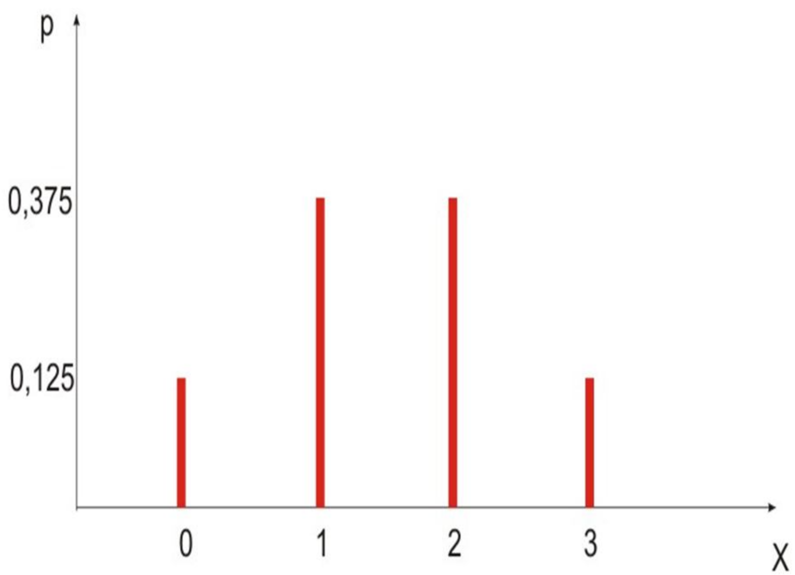
$F_X(1) = P_X((-\infty,1)) = P(X<1) = P(X=0) = 1/8$

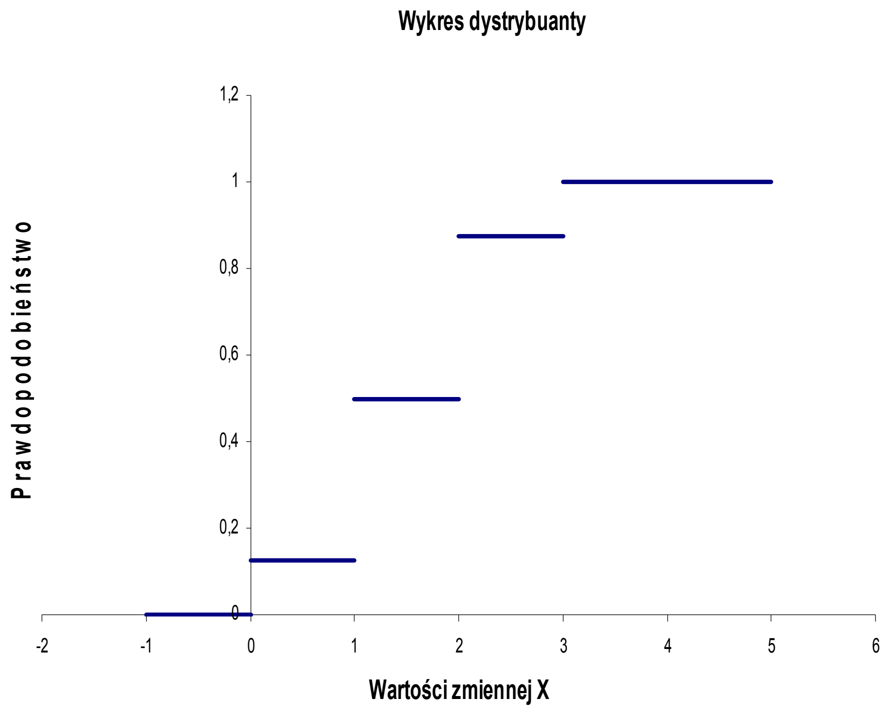
$F_X(2) = P_X((-\infty,2)) = P(X<2) = 1/8+3/8 = 4/8$

$F_X(3) = P_X((-\infty,3)) = P(X<3) = 1/8+3/8 +3/8 = 7/8$

$F_X(4) = P_X((-\infty,4)) = P(X<4) = 1$

Wykresy rozkładu prawdopodobieństwa i dystrybuanty zmiennej losowej dyskretnej (skokowej)





Parametry rozkładu zmiennej losowej -

Wartość oczekiwaną [nadzieję matematyczną / wartość przeciętną], zmiennej losowej X oznacza się E(X) i określa w następujący sposób

$$E(X) = \sum_{i=0}^{n} x_i p_i$$

- Wariancja zmiennej losowej

$$D^2(X) = E[X-E(X)]^2 = E(X^2)-E(X)^2$$

- Odchylenie standardowe: $D(X) = \sqrt{D^2(X)}$

- współczynnikiem zmienności : V = D(X)/E(X)

Przykład jak prosto obliczyć wartość oczekiwaną i wariancję

x_i	0	1	2	3	Σ
p_i	0,125	0,375	0,375	0,125	
x_i*p_i	0	0,375	0,75	0,375	1,5
x_i^2*p_i	0	0,375	1,5	1,125	3

$$E(X) = 1,5$$

$$D^2(X)=E\,(X^2) - (E(X))^2 = 3 - (1,5)^2 = 0,75$$

Rozkład Bernoullego – dwumianowy zmiennej losowej dyskretnej

- Prawdopodobieństwo odniesienia k sukcesów w n doświadczeniach $p_n(k)$,

- Jeżeli
  p-oznacza prawdopodobieństwo sukcesu w pojedynczym doświadczeniu,
- wtedy
  $p_n(k)$ obliczamy z wzoru Bernoullego

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

Wartość oczekiwaną i wariancję obliczamy z wzorów

$E(X) = np$ $D^2(X) = npq$

Rozkład normalny zmiennej losowej ciągłej

Rozkład normalny, zwany również rozkładem Gaussa-Laplace'a jest najczęściej spotykanym rozkładem zmiennej losowej ciągłej.

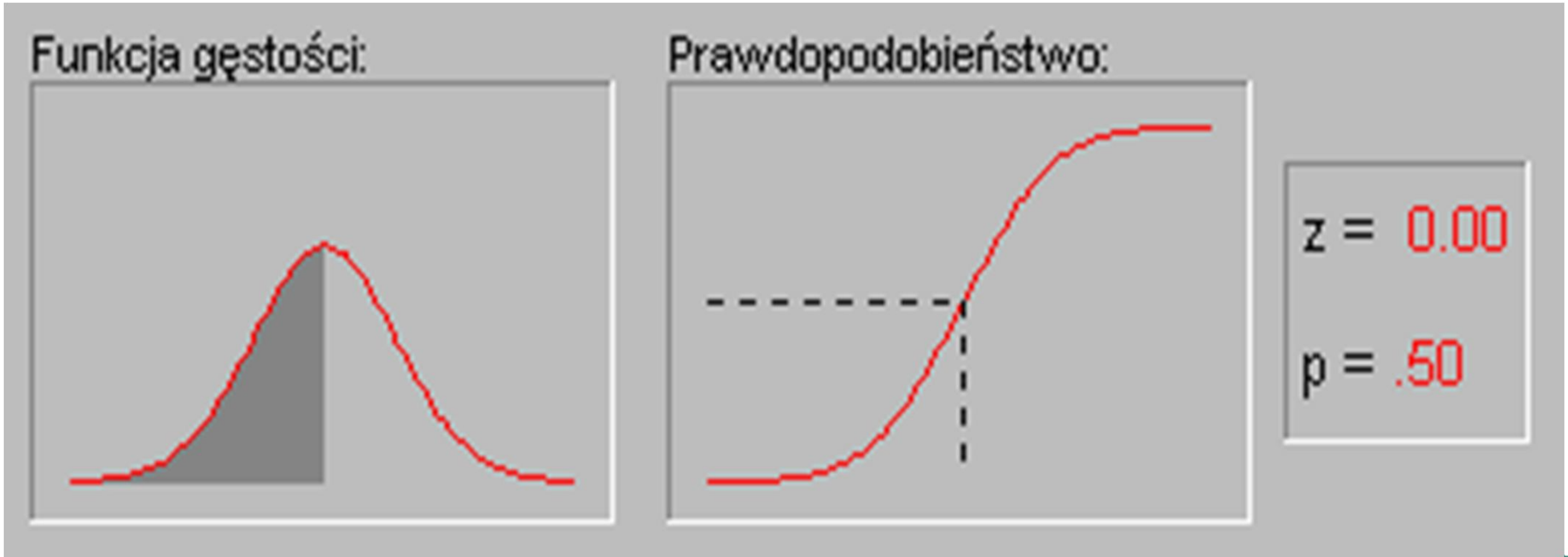
Fakt, że zmienna losowa ciągła X ma rozkład normalny o wartości oczekiwanej μ i odchyleniu standardowym σ , zapisujemy N (μ ,σ)

Funkcja gęstości w rozkładzie normalnym jest określona w R i ma postać

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

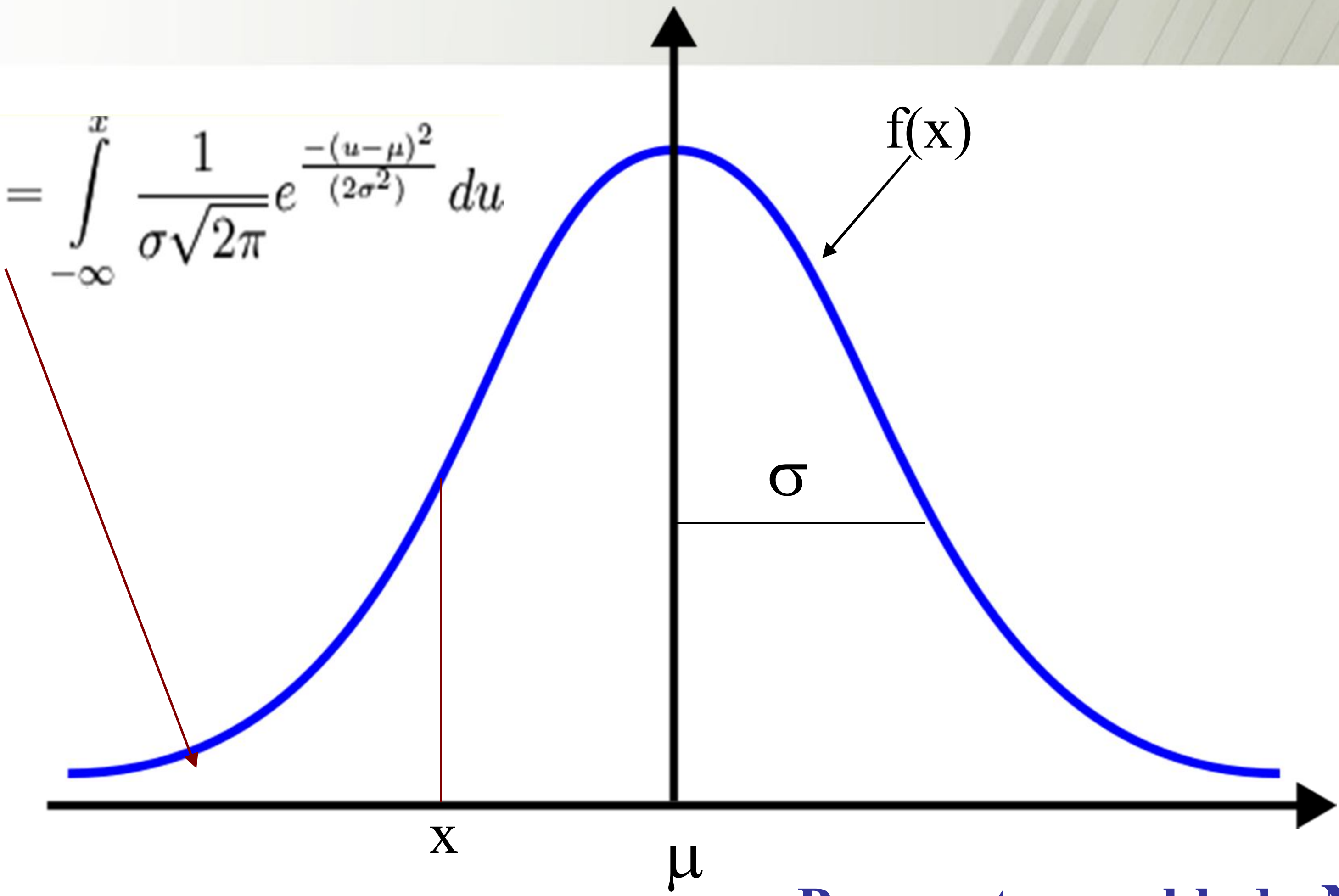
Rozkład normalny

W rozkładzie normalnym zarówno funkcja gęstości jak i dystrybuanta (prawdopodobieństwo) są określone dla wszystkich rzeczywistych wartości zmiennej X.



Rozkład normalny – wykres funkcji gęstości i interpretacja

$$P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(u-\mu)^2}{(2\sigma^2)}} du$$



Parametry rozkładu N(μ,σ),

μ - Wartość oczekiwana

σ^2 - Wariancja

Cechy charakterystyczne funkcji gęstości rozkładu normalnego

Funkcja gęstości w rozkładzie normalnym:

- jest symetryczna względem prostej $x = \mu$
- w punkcie $x = \mu$ osiąga wartość maksymalną
- ramiona funkcji mają punkty przegięcia dla $x = \mu - \sigma$ oraz $x = \mu + \sigma$

Kształt funkcji gęstości zależy od wartości parametrów: μ, σ :

- parametr μ decyduje o przesunięciu krzywej,
- parametr σ decyduje o „smukłości” krzywej.

Rozkład normalny - standaryzacja

Standaryzacja polega na sprowadzeniu dowolnego rozkładu normalnego **N(μ,σ)**, o danych parametrach μ i σ do **rozkładu standaryzowanego** (modelowego) o wartości oczekiwanej $\mu = 0$ i odchyleniu standardowym σ = 1.

Zmienną X zastępuje się **zmienną standaryzowaną U**

$$u = \frac{x - \mu}{\sigma}$$

która ma **rozkład N(0,1)**

Wtedy otrzymujemy następujące zależności :

f(x)→φ(u), F(x) →Φ(u), czyli:

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Własności dystrybuanty standaryzowanego rozkładu normalnego:

$$P(U \le u) = \Phi(u)$$

$$P(U \le -u) = \Phi(-u) = 1 - \Phi(u)$$

$$P(U > u) = 1 - P(U \le u) = 1 - \Phi(u)$$

$$P(U > -u) = \Phi(u)$$

Zadanie:

Wzrost kobiet w pewnej populacji ma rozkład normalny N(165,15). Oznacza to, iż zmienna losowa jaką jest wzrost kobiet ma rozkład normalny ze średnią równą 165 cm i odchyleniem standardowym równym 15 cm.

Jaki jest udział w populacji kobiet o wzroście:

a) do 160 cm,
b) w przedziale 165-170 cm,
c) powyżej 175 cm
d) dokładnie 150 cm

Rozwiązanie: a) do 160 cm

$$P(X \le 160) = P\left(\frac{X-165}{15} \le \frac{160-165}{15}\right) = P(U \le -0{,}33) =$$

$$= \phi(-0{,}33) = 1 - \phi(0{,}33) = 1 - 0{,}6293 = 0{,}3707$$

Zadanie: rozwiązanie

b) w przedziale 165-170 cm

$$P(165 < X \leq 170) = P\left(\frac{165-165}{15} < \frac{X-165}{15} \leq \frac{170-165}{15}\right) =$$

$$= P(0 < U \leq 0{,}33) = \phi(0{,}33) - \phi(0) = 0{,}6293 - 0{,}5 = 0{,}1293$$

c) powyżej 175 cm.

$$P(X > 175) = P\left(\frac{X-165}{15} > \frac{175-165}{15}\right) = P(U > 0{,}67) =$$

$$= 1 - P(U \leq 0{,}67) = 1 - \phi(0{,}67) = 1 - 0{,}748571 = 0{,}251429$$

d) dokładnie 150 cm.

$$P(X = 150) = P(150 \leq X \leq 150) = F(150) - F(150) = 0$$

Wnioskowanie statystyczne
Estymacja

Podstawowym narzędziem szacowania nieznanego parametru rozkładu zmiennej losowej jest estymator obliczony na podstawie próby statystycznej, który służy do wnioskowania na temat wartości danego parametru w całej populacji. Np. dla wartości oczekiwanej najlepszym estymatorem jest średnia arytmetyczna, albo średnia ważona.

Stosuje się :

- **Estymację punktową** – która ma zastosowanie gdy, na podstawie danych z próby, chcemy **ustalić liczbową wartość określonego parametru** rozkładu cechy w całej populacji (wzory na estymatory punktowe wartości średniej, wariancji, czy odchylenia standardowego, podano w wykładzie 1)
- **Estymacja przedziałowa** polega na wyznaczeniu granic przedziału liczbowego, w którym, z określonym prawdopodobieństwem, zawiera się wartość szacowanego parametru

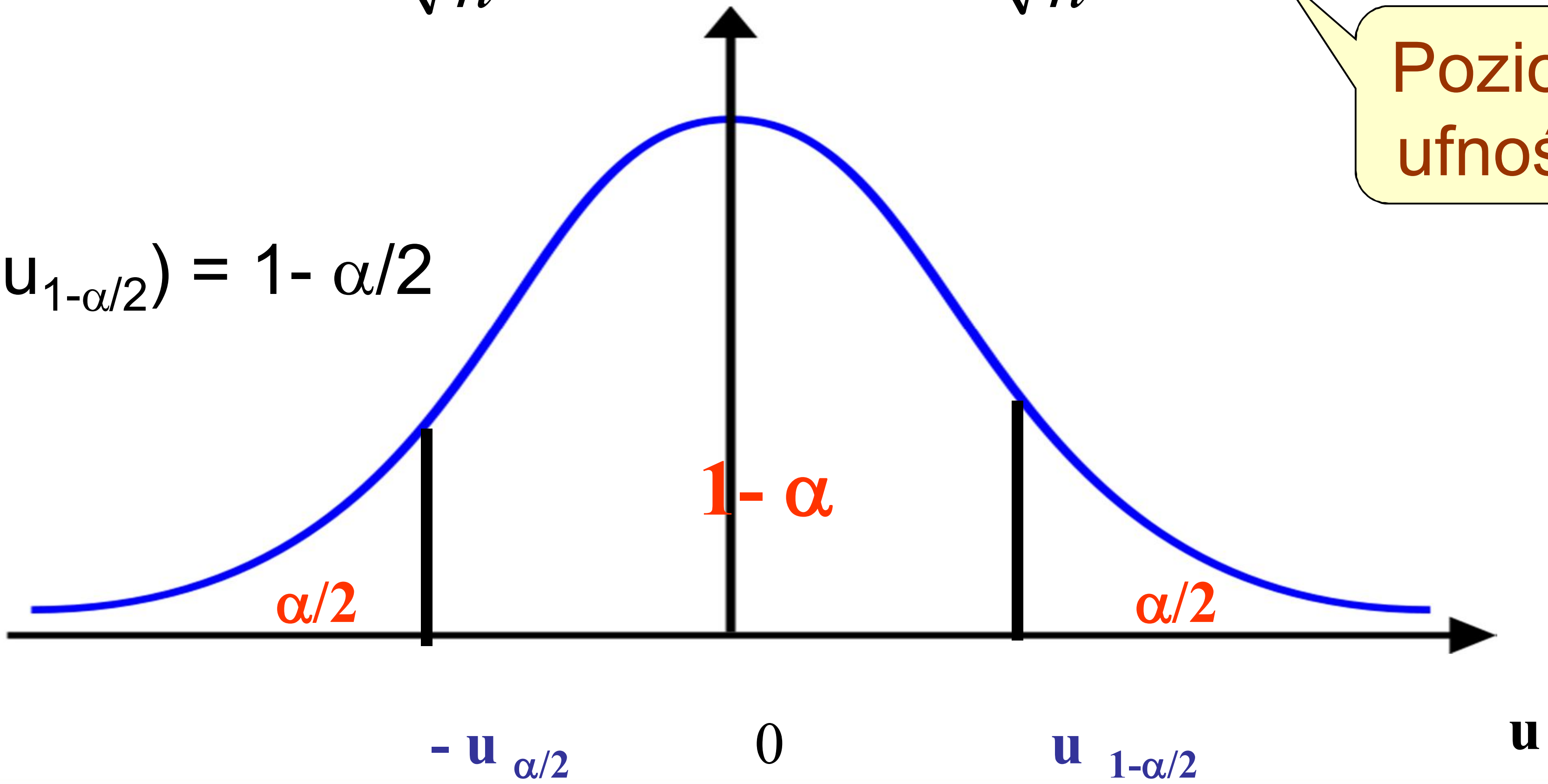
Przedział ufności dla wartości oczekiwanej

gdy znane jest odchylenie standardowe σ

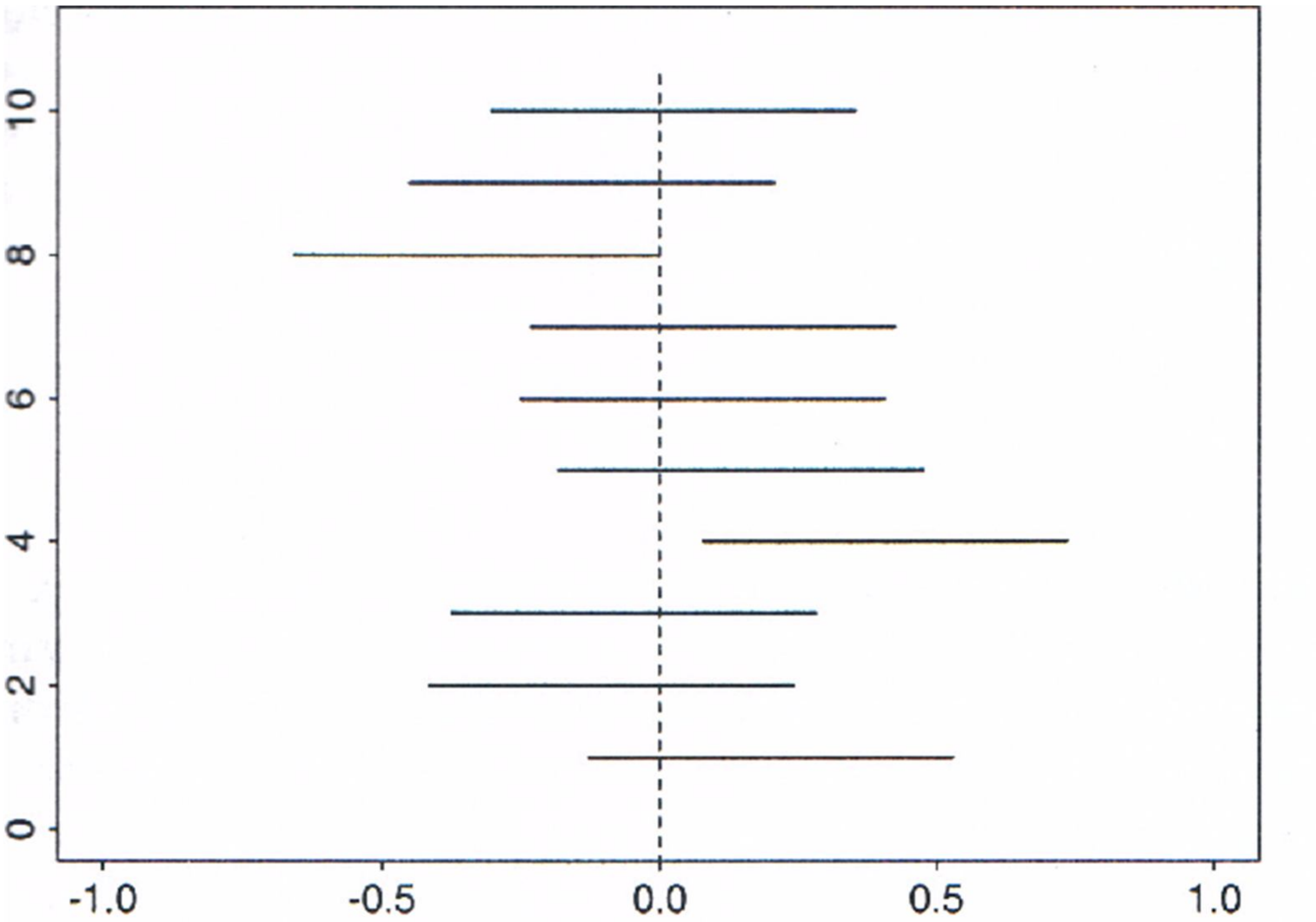
$$P(\overline{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Poziom ufności

$\Phi(u_{1-\alpha/2}) = 1 - \alpha/2$



Praktyczna realizacja przedziałów ufności dla μ, dla prostych prób losowych o licznościach n=25, z rozkładu N (0,1) dla poziomu ufności $1-\alpha = 0.9$



Problem minimalnej liczności próby

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

Długość przedziału ufności wynosi

$$2u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Żądamy by maksymalny błąd oszacowania nie przekraczał zadanej z góry wartości d

$$u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le d$$

Z tej relacji wynika, że

$$n \ge \frac{\left(u_{1-\frac{\alpha}{2}\alpha}\,\sigma\right)^2}{d^2}$$

Zadanie

- Wykonujemy pomiary grubości płytki metalowej.

  Jak dużą liczbę pomiarów (n) należy przeprowadzić, aby prawdopodobieństwem (ufnością) wynoszącym 0,95 maksymalny błąd oceny nie przekraczał 0,02 mm. Zakładamy, że odchylenie standardowe błędów pomiarów σ=0,1

Przedział ufności dla wartości oczekiwanej, gdy odchylenie standardowe jest <u>nieznane</u>

Estymatorem μ, uzyskanym MNW jest średnia arytmetyczna, nie znamy σ, musimy zatem wybrać statystykę, która od σ nie zależy

$$t = \frac{\overline{X} - m}{S} \sqrt{n - 1}$$

Statystyka t ma rozkład Studenta z n-1 stopniami swobody,

nie zależy od parametru σ ale od parametru S,

S jest odchyleniem standardowym obliczonym z próby.

Przedział ufności dla wartości oczekiwanej, gdy odchylenie standardowe jest <u>nieznane</u>

Przedział ufności dla wartości oczekiwanej ma wtedy postać

$$P(\overline{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n-1}} < m < \overline{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n-1}}) = 1 - \alpha$$

- gdzie wartość $t_{\alpha,n-1}$, jest kwantylem rzędu α, z n-1 stopniami swobody
- Długość przedziału wynosi $2\, t_{\alpha,n-1} S/\sqrt{n-1}$

Przedział ufności dla wartości oczekiwanej, gdy **nieznany jest rozkład w populacji**

- W praktyce często nie znany jest rozkład cechy w populacji i brak jest podstaw do przyjęcia, że jest on normalny.
- Wiadomo, że średnia arytmetyczna wyznaczona z próby o dowolnym rozkładzie jest zmienną losową o rozkładzie N(m, σ/√n) , dlatego

$$P(\overline{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

- Nieznane σ można przybliżyć obliczonym z dużej próby odchyleniem standardowym S

$$P(\overline{X} - u_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}) = 1 - \alpha$$

Zadanie

- Dokonano 10 pomiarów ciśnienia wody na ostatnim piętrze bloku 15 piętrowego i okazało się, że średnie ciśnienie wynosiło 2,21 podczas gdy wariancja wyniosła 4,41. Znaleźć liczbowe wartości krańców przedziałów ufności dla wartości oczekiwanej przyjmując poziom ufności
- $1-\alpha = 0{,}95$
- $1-\alpha = 0{,}90$
- $1-\alpha = 0{,}98$

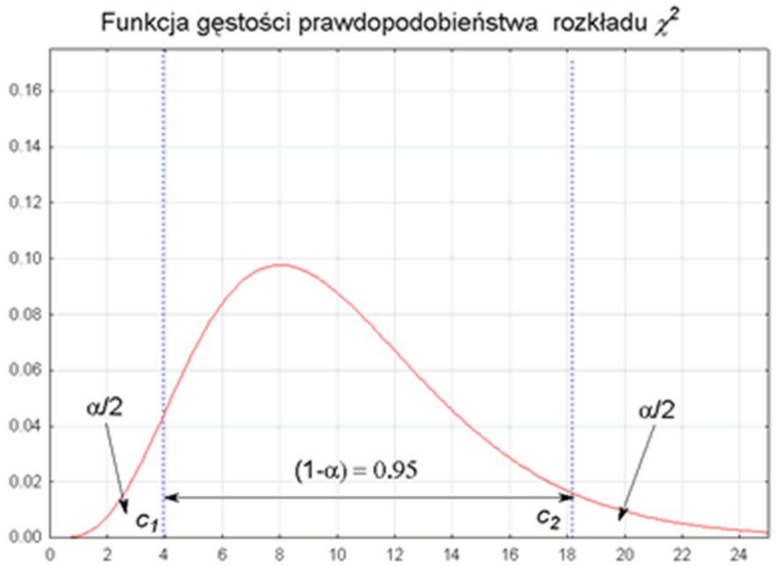
Przedział ufności dla wariancji w populacji normalnej

- Przedział jest zbudowany w oparciu o statystykę $\chi^2 = ns^2/\sigma^2$, która ma rozkład χ^2 o n-1 stopniach swobody.

- W rozkładzie χ^2 określa się dwie wartości , spełniające odpowiednio równości

$$P(\chi^2 \geq \chi^2_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2}$$

$$P(\chi^2 \geq \chi^2_{\frac{\alpha}{2},n-1}) = 1 - \frac{\alpha}{2}$$



Tablice rozkładu $\chi 2$

Poziom istotności / Stopnie swobody	Dystrybuanta odwrotna F^{-1} rozkładu chi-kwadrat dla ustalonej liczby stopni swobody oraz poziomu istotności												
	0,999	0,995	0,99	0,98	0,975	0,95	0,9	0,1	0,05	0,025	0,02	0,01	0,005
1	0,0000	0,0000	0,0002	0,0006	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	5,4119	6,6349	7,8794
2	0,0020	0,0100	0,0201	0,0404	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	7,8241	9,2104	10,5965
3	0,0243	0,0717	0,1148	0,1848	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	9,8374	11,3449	12,8381
4	0,0908	0,2070	0,2971	0,4294	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	11,6678	13,2767	14,8602
5	0,2102	0,4118	0,5543	0,7519	0,8312	1,1455	1,6103	9,2363	11,0705	12,8325	13,3882	15,0863	16,7496
6	0,3810	0,6757	0,8721	1,1344	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	15,0332	16,8119	18,5475
7	0,5985	0,9893	1,2390	1,5643	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	16,6224	18,4753	20,2777
8	0,8571	1,3444	1,6465	2,0325	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	18,1682	20,0902	21,9549
9	1,1519	1,7349	2,0879	2,5324	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	19,6790	21,6660	23,5893
10	1,4787	2,1558	2,5582	3,0591	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	21,1608	23,2093	25,1881
11	1,8338	2,6032	3,0535	3,6087	3,8157	4,5748	5,5778	17,2750	19,6752	21,9200	22,6179	24,7250	26,7569
12	2,2141	3,0738	3,5706	4,1783	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	24,0539	26,2170	28,2997
13	2,6172	3,5650	4,1069	4,7654	5,0087	5,8919	7,0415	19,8119	22,3620	24,7356	25,4715	27,6882	29,8193
14	3,0407	4,0747	4,6604	5,3682	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	26,8727	29,1412	31,3194
15	3,4825	4,6009	5,2294	5,9849	6,2621	7,2609	8,5468	22,3071	24,9958	27,4884	28,2595	30,5780	32,8015
16	3,9417	5,1422	5,8122	6,6142	6,9077	7,9616	9,3122	23,5418	26,2962	28,8453	29,6332	31,9999	34,2671
17	4,4162	5,6973	6,4077	7,2550	7,5642	8,6718	10,0852	24,7690	27,5871	30,1910	30,9950	33,4087	35,7184
18	4,9048	6,2648	7,0149	7,9062	8,2307	9,3904	10,8649	25,9894	28,8693	31,5264	32,3462	34,8052	37,1564
19	5,4067	6,8439	7,6327	8,5670	8,9065	10,1170	11,6509	27,2036	30,1435	32,8523	33,6874	36,1908	38,5821
20	5,9210	7,4338	8,2604	9,2367	9,5908	10,8508	12,4426	28,4120	31,4104	34,1696	35,0196	37,5663	39,9969
21	6,4467	8,0336	8,8972	9,9145	10,2829	11,5913	13,2396	29,6151	32,6706	35,4789	36,3434	38,9322	41,4009
22	6,9829	8,6427	9,5425	10,6000	10,9823	12,3380	14,0415	30,8133	33,9245	36,7807	37,6595	40,2894	42,7957
23	7,5291	9,2604	10,1957	11,2926	11,6885	13,0905	14,8480	32,0069	35,1725	38,0756	38,9683	41,6383	44,1814
24	8,0847	9,8862	10,8563	11,9918	12,4011	13,8484	15,6587	33,1962	36,4150	39,3641	40,2703	42,9798	45,5584
25	8,6494	10,5196	11,5240	12,6973	13,1197	14,6114	16,4734	34,3816	37,6525	40,6465	41,5660	44,3140	46,9280
26	9,2222	11,1602	12,1982	13,4086	13,8439	15,3792	17,2919	35,5632	38,8851	41,9231	42,8558	45,6416	48,2898
27	9,8029	11,8077	12,8785	14,1254	14,5734	16,1514	18,1139	36,7412	40,1133	43,1945	44,1399	46,9628	49,6450
28	10,3907	12,4613	13,5647	14,8475	15,3079	16,9279	18,9392	37,9159	41,3372	44,4608	45,4188	48,2782	50,9936
29	10,9861	13,1211	14,2564	15,5745	16,0471	17,7084	19,7677	39,0875	42,5569	45,7223	46,6926	49,5878	52,3355
30	11,5876	13,7867	14,9535	16,3062	16,7908	18,4927	20,5992	40,2560	43,7730	46,9792	47,9618	50,8922	53,6719

Przedział ufności dla wariancji w populacji normalnej

- Z podanych wzorów wynika, że

$$P(\chi^2_{1-\frac{\alpha}{2},n-1} < \chi^2 < \chi^2_{\frac{\alpha}{2},n-1}) = 1-\alpha \quad ; \quad P(\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{nS^2}{\sigma_2} < \chi^2_{\frac{\alpha}{2},n-1}) = 1-\alpha$$

- Po przekształceniu których otrzymujemy przedział ufności dla wariancji

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

Zadanie

- Odchylenie standardowe σ błędu przyrządu pomiarowego jest nieznane. Zakładamy, że rozkład błędów pomiarów jest rozkładem normalnym.
- Przeprowadzono n= 10 pomiarów i otrzymano następujące wyniki
  {7; 7,5; 8,5; 8; 6; 7,5; 6,5; 5;5 7,5; 6 }
- Wyznaczyć liczbowe wartości krańców przedziałów ufności dla
  - Wartości oczekiwanej
  - Dla odchylenia standardowego
- Na poziomie ufności $1-\alpha = 0,95$

Przedziały ufności dla proporcji p

- Opierając się na częstości $\hat{p}$ skonstruujemy przedziały ufności dla proporcji p. Jeśli próba losowa niezależnych zmiennych o rozkładzie punktowym P(X=1)=1-P(X=0) = p jest dostatecznie liczna, by móc skorzystać z przybliżenia rozkładem N(0,1) , statystyki

(*) $$\left(\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \right)$$

- Wówczas

$$P\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq u_{1-\frac{\alpha}{2}} \right) \approx 1 - \alpha$$

Zastosowanie

- Agencja badająca w 2000 roku opinie Polaków na podstawie 1000 elementowej próby stwierdziła, że 57% popiera wejście Polski do Unii.
- Uznając, ze mamy do czynienia z rozkładem dwupunktowym skonstruujemy przedział ufności na poziomie 0,95 dla proporcji Polaków popierających wejście Polski do UE
  - Próba o n=1000 jest dostatecznie liczna by skorzystać ze rozkładu statystyki (*)
  - Przedział 95% ufności to [0,54,0,60], natomiast wielkość $\sqrt{0{,}57(1-0{,}57)/1000} = 0{,}00156$ można uznać za błąd standardowy otrzymanej częstości, w ujęciu procentowym wynosi on około 1,6%

Przedział ufności dla proporcji p

$$P\left(\hat{p} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1-\alpha$$

Ważne jest aby pamiętać jakie są minimalne wymagania na liczność próby n i proporcję p, by móc rozkład podanej w (*) statystyki przybliżać rozkładem N(0,1)

Podstawowe etapy procesu weryfikacji hipotez statystycznych:

1. Sformułowanie hipotez H_0 i H_1
2. Przyjęcie odpowiedniego poziomu istotności α oraz liczebności próby
3. Określenie obszaru krytycznego i obszaru przyjęcia sprawdzanej hipotezy H_0
4. Wybór testu weryfikującego H_0 i wyliczenie statystyki testowej
5. Podjęcie decyzji weryfikacyjnej

1. Sformułowanie hipotez H_0 i H_1
Parametryczne testy istotności

Test dla wartości średniej w populacji generalnej

Hipoteza sprawdzana (zerowa) dotyczy określonego parametru, np wartości oczekiwanej m:

- H_0: $m=m_0$

przy jednej z hipotez alternatywnych

- H_1: $m \neq m_0$ lub H_1: $m > m_0$ lub H_1: $m < m_0$

- Hipoteza H_0 : o równości średnich z n - elementowej próby i w populacji będzie zweryfikowana na podstawie wyników próby losowej.

- Za sprawdzian hipotezy H_0 przyjmuje się określoną statystykę, zwaną także funkcją testową.

- Dla wartości oczekiwanej będzie to średnia arytmetyczną uzyskanych wyników z próby losowej.

2. Przyjęcie odpowiedniego poziomu istotności α oraz liczebności próby

Przy podejmowaniu decyzji weryfikującej hipotezy możemy popełnić dwa rodzaje błędów

Decyzja	Hipoteza H_0	
	prawdziwa	fałszywa
odrzucić	błąd I rodzaju	decyzja trafna
	α	$1-\beta$
nie odrzucić	decyzja trafna	błąd II rodzaju
	$1-\alpha$	β

Rodzaje błędów popełnianych przy weryfikacji hipotez statystycznych

- Błąd I rodzaju polega na odrzuceniu hipotezy zerowej, mimo że jest prawdziwa.
- Przyjmowany w procesie weryfikacji hipotezy poziom istotności jest równy prawdopodobieństwu popełnienia błędu I rodzaju, zwykle α=0.05lu 0.01
- Błąd II rodzaju polega za przyjęciu za prawdziwą hipotezy H_0 gdy ona w rzeczywistości jest fałszywa.

Przykład

H_0- oskarżony jest niewinny

H_1 - oskarżony jest winien

Błąd I rodzaju : sąd skazał niewinnego: H_0 prawdziwa, ale ją odrzucono

Błąd II rodzaju: sąd uwolnił winnego: H_1 prawdziwa, a przyjęto H_0,

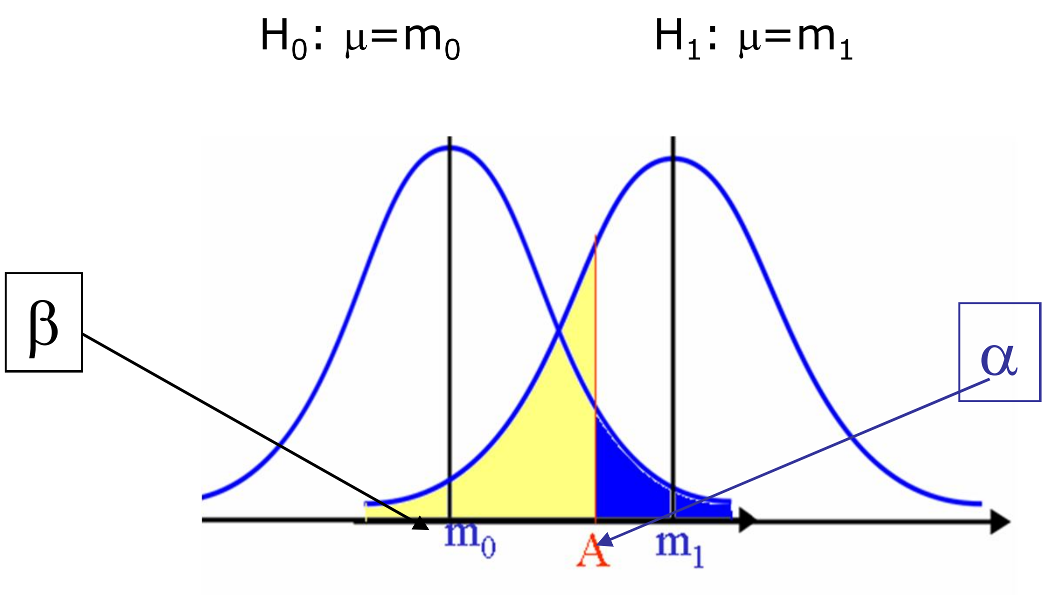
Tu błąd I rodzaju jest znacznie bardziej dotkliwy, dlatego należy zminimalizować prawdopodobieństwo jego popełnienia (czyli dostarczyć „niezbitych” dowodów)

Związek pomiędzy błędami I i II rodzaju: zmniejszanie wartości α pociąga wzrost wartości β

H_0: $\mu = m_0$ H_1: $\mu > m_1$

Przy przyjętym poziomie istotności α, obszar krytyczny obejmuje wartości średnie $\geq A$, gdy $P(x \geq A) = \alpha$

Dla określenia obszaru β przyjmiemy następujący zestaw hipotez H_0: $\mu = m_0$ H_1: $\mu = m_1 > m_0$



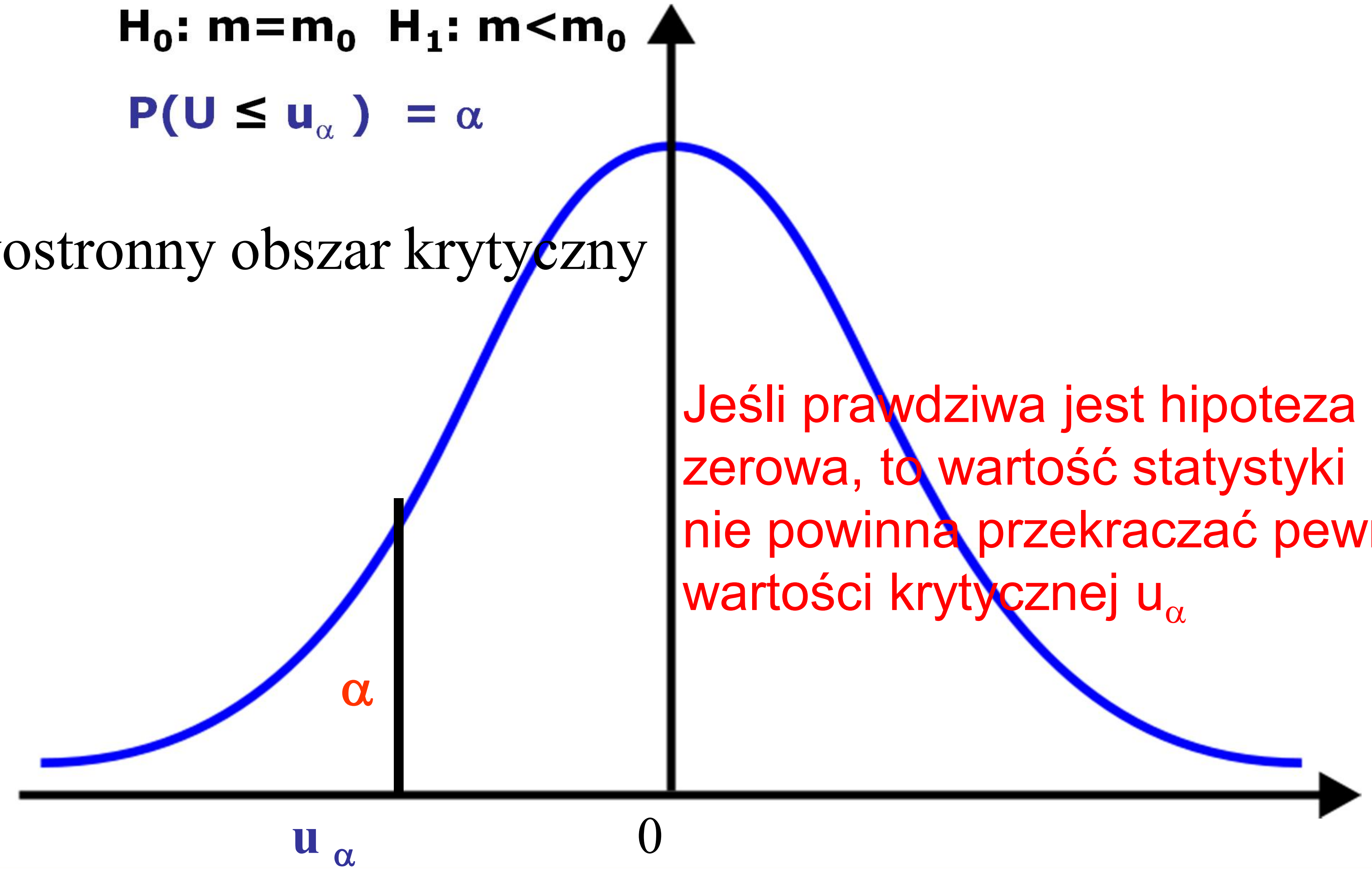
3. Określenie obszaru krytycznego i obszaru przyjęcia sprawdzanej hipotezy H_0

H_0: $m=m_0$ H_1: $m<m_0$

$P(U \leq u_\alpha) = \alpha$

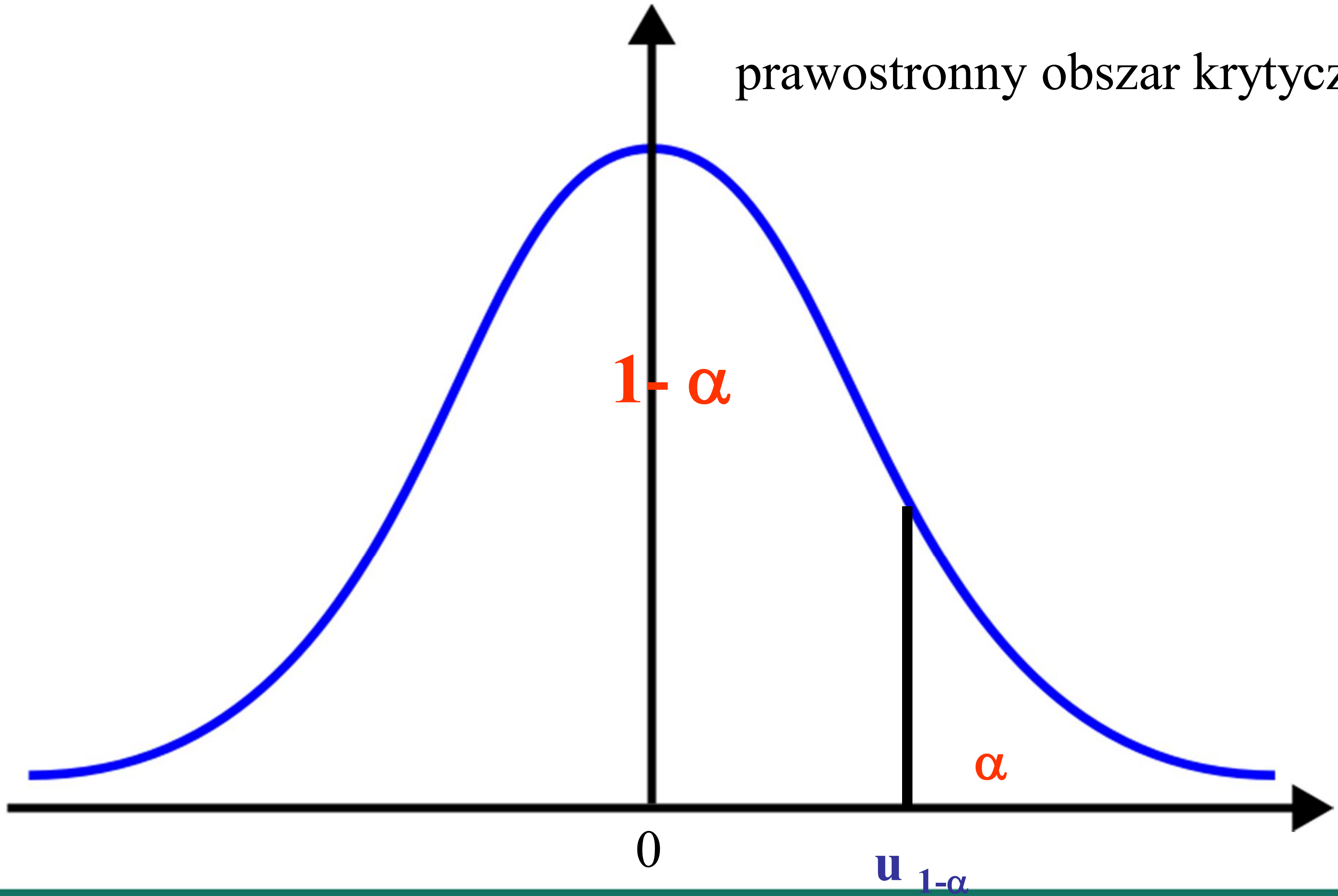
lewostronny obszar krytyczny

Jeśli prawdziwa jest hipoteza zerowa, to wartość statystyki nie powinna przekraczać pewne wartości krytycznej u_α



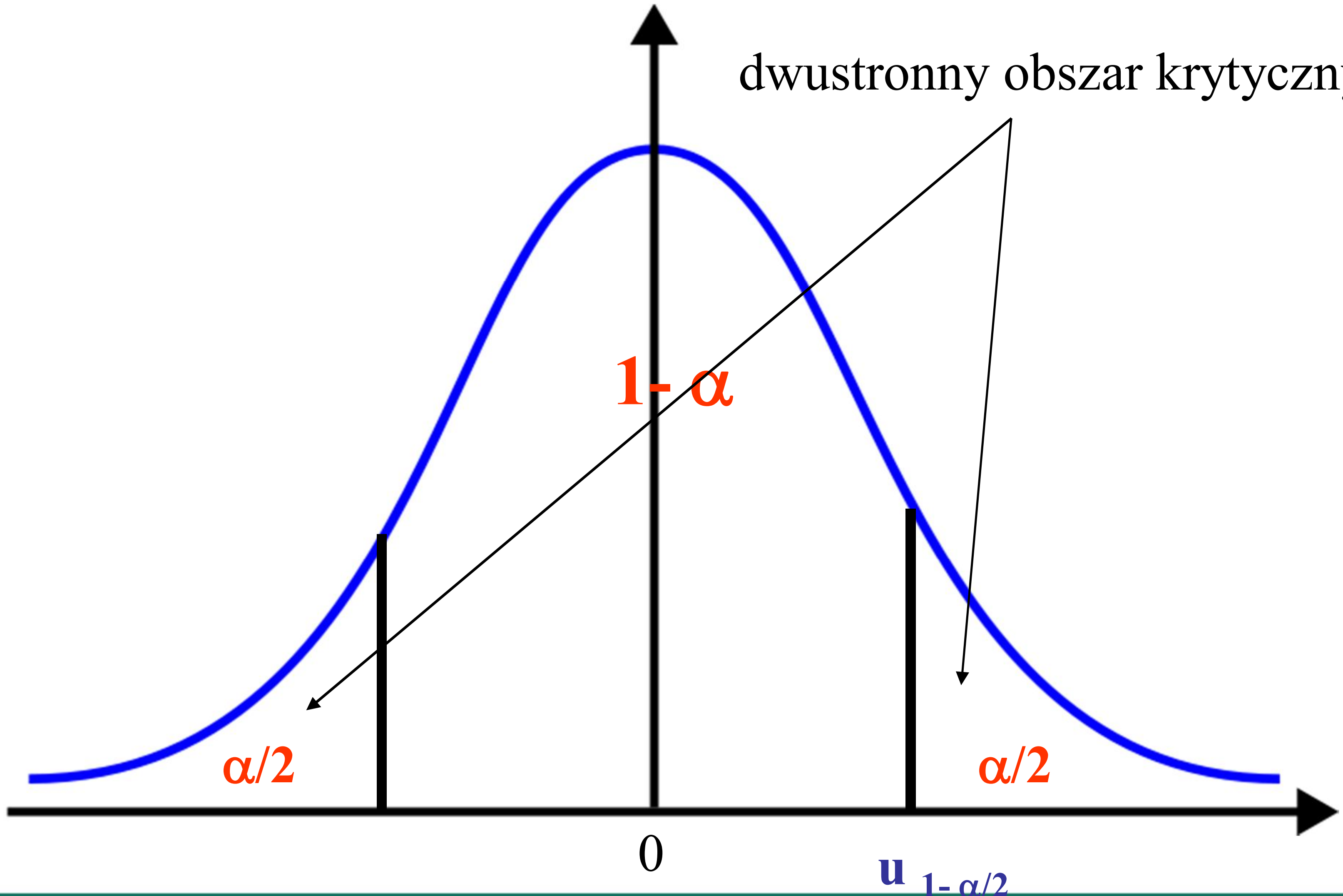
H_0: $m=m_0$ H_1: $m>m_0$

$P(U \geq u_\alpha) = \alpha$



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

$P(|U| \geq u_{1-\alpha/2}) = \alpha$



4. Wybór testu weryfikującego H_0 i wyliczenie statystyki testowej

Rozważamy rozkład średnich z n-elementowej próby, jest to rozkład $N(m_0, \sigma/\sqrt{n})$, o ile hipoteza H_0 jest prawdziwa

Stąd statystyka U , określona wzorem

$$U = \frac{\bar{x} - m_o}{\sigma} \sqrt{n}$$

ma rozkład N (0,1),

- Jeśli prawdziwa jest hipoteza zerowa , to wartość statystyki U nie powinna przekraczać pewnej wartości krytycznej u_α
- α oznacza obszar zbiór nietypowych wartości statystyki testowej pod warunkiem prawdziwości hipotezy zerowej

Funkcje testowe dla dużej próby i dla małej, gdy nieznana jest wartość wariancji w populacji

Duża próba

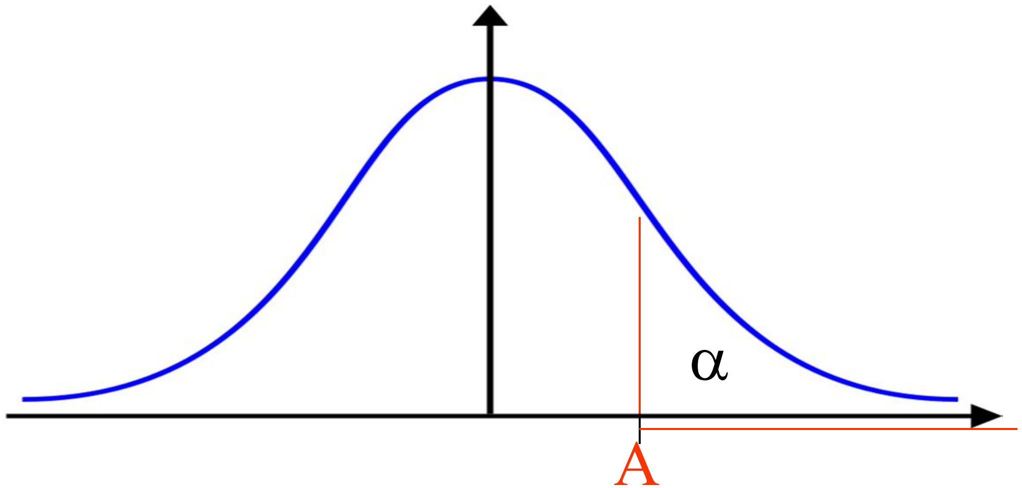
$$U = \frac{\overline{x} - m_o}{s}\sqrt{n}$$

Mała próba

$$t = \frac{\overline{x} - m_o}{s}\sqrt{n-1}$$

Podstawa do podjęcia decyzji weryfikacyjnej

- Jeżeli obliczona wartość funkcji testowej znajdzie się w obszarze krytycznym (np. f >A) , hipotezę H_0 należy odrzucić i przyjąć hipotezę H_1
- W programach komputerowych decyzję podejmuje się na następującej podstawie
  - jeśli $p < \alpha \Rightarrow H_0$ odrzucamy, przyjmujemy H
  - jeśli $p \geq \alpha \Rightarrow$ nie ma podstaw do odrzucenia H_0



Podstawowe twierdzenia dotyczące zmiennych o rozkładzie Studenta

1. **Jeżeli zmienne losowe** $X_1, X_2, \ldots, X_n$ mają jednakowy rozkład prawdopodobieństwa, który jest rozkładem normalnym o średniej m i wariancji σ^2, to zmienna **t** określona wzorem

$$t = \frac{\overline{X} - m}{s} \cdot \sqrt{n}$$

gdzie $\overline{X}$ jest wartością średnią z próby, zaś s - odchyleniem standardowym z próby - ma **rozkład t-Studenta** o v=n-1 stopniach swobody (niezależny od wartości wariancji w populacji σ^2).

2. **Jeżeli dwie próby** o liczebnościach n_1 oraz n_2, wartościach średnich $\overline{X}_1$ oraz $\overline{X}_2$ i wariancjach wyznaczonych z próby s_1^2 oraz s_2^2 zostały wylosowane z populacji mających taki sam rozkład normalny, to zmienna **t** określona wzorem:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)}$$

ma rozkład t-Studenta o $v = n_1 + n_2 - 2$ stopniach swobody.

Weryfikacja hipotezy o wariancji w rozkładzie normalnym

H_0: $(\sigma^2 \leq \sigma^2_0)$ przy H_1: $(\sigma^2 > \sigma^2_0)$

Przyjmujemy poziom istotności α

i wiemy, że statystyka $\frac{nS_n^2}{\sigma_0^2}$ ma rozkład chi-kwadrat o n-1stopniach swobody.

Skoro, gdy H_0 jest prawdziwa, zachodzi równość ,

$$P\left(\frac{nS_n^2}{\sigma_0^2} > \chi_\alpha^2\right) = \alpha$$

Zatem hipotezę H_0 odrzucamy, na rzecz H_1, ilekroć stwierdzimy (na podstawie obliczeń), że zaszła nierówność

$$\frac{nS_n^2}{\sigma_0^2} > \chi_\alpha^2$$

Weryfikacja hipotezy o wariancji w rozkładzie normalnym

- Błąd pomiaru odległości za pomocą radaru ma rozkład normalny. Przeprowadzono 10 pomiarów tej samej znanej odległości i otrzymano następujące wartości błędów

k	1	2	3	4	5	6	7	8	9	10
s_k [km]	0,115	-0,250	0,180	-0,060	-0,120	0,010	-0,050	0,075	-0,150	-0,250

suma błędów	-0,500
średni błąd	-0,050
wariancja błędów	0,0216

Na poziomie istotności α=0,05 zweryfikować hipotezę , że wariancja błędu nie przekracza 0,0125.

Odczytane z tablic chi kwadrat dla n-1=9 stopni swobody =16,919

Obliczam wartość funkcji testowej

$$\frac{nS_n^2}{\sigma_0^2} = \frac{10 * 0{,}0216}{0{,}0125} = 17{,}276 > 16{,}919$$

H_0 należy odrzucić

Etapy wnioskowania statystycznego

<u>obliczenia własne</u>

1. postawienie hipotezy zerowej
2. wybór testu i sprawdzenie spełnienia założeń
3. **obliczenie wartości funkcji testowej**
4. **ustalenie (odczytanie z tablic) wartości krytycznych dla danego poziomu istotności**
5. podjęcie decyzji o przyjęciu lub odrzuceniu hipotezy H_0
6. interpretacja otrzymanych wyników

z <u>użyciem pakietu STATISTICA</u>

1. postawienie hipotezy zerowej
2. wybór testu i sprawdzenie spełnienia założeń
3. **wprowadzenie danych**
4. podjęcie decyzji o przyjęciu lub odrzuceniu hipotezy H_0
5. interpretacja otrzymanych wyników

Przykład realizowany z pomocą pakietu STATISTICA

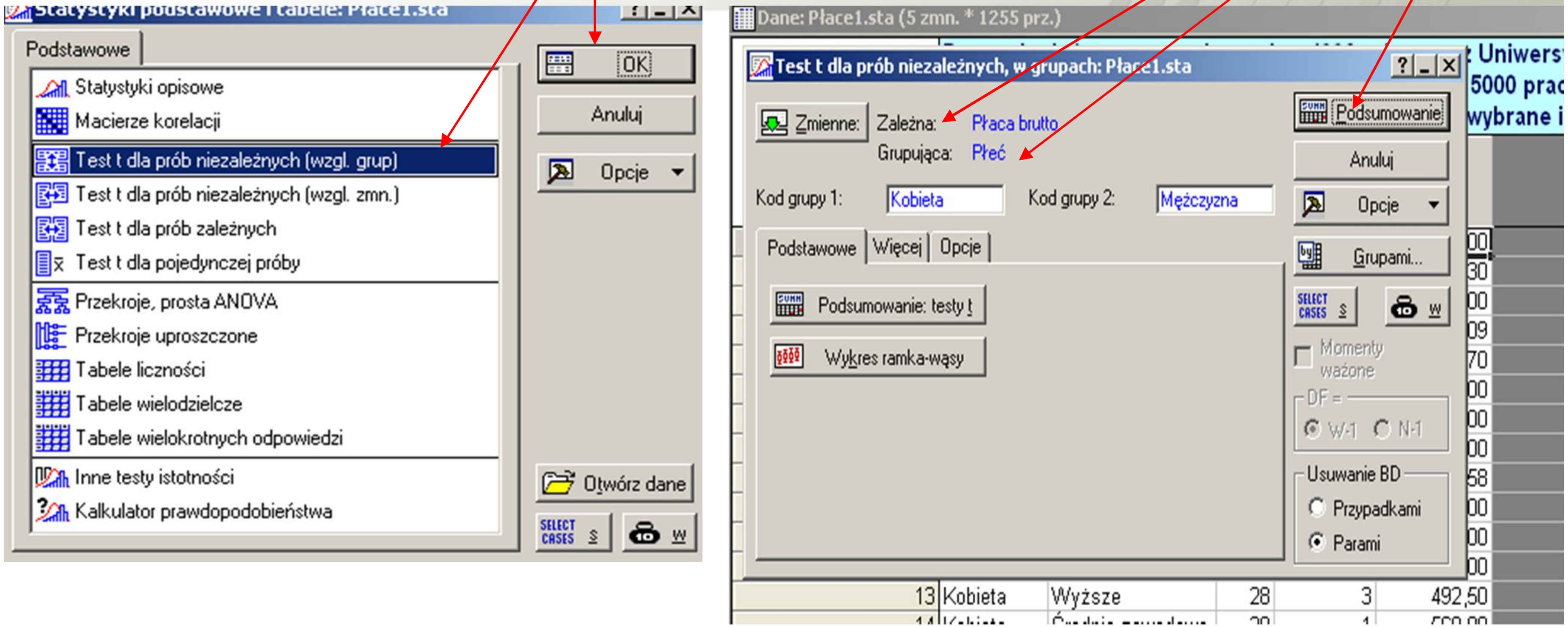
- Dane z badań przeprowadzonych w 1996 roku dotyczące zarobków Polaków.
- Ankiety wysłano do 5000 pracowników wylosowanych przez GUS.
- Ankiety zwróciło 1255 osób. Arkusz zawiera następujące informacje o badanych osobach
  - Płeć
  - Wykształcenie
  - Wiek
  - Staż pracy
  - Płaca brutto

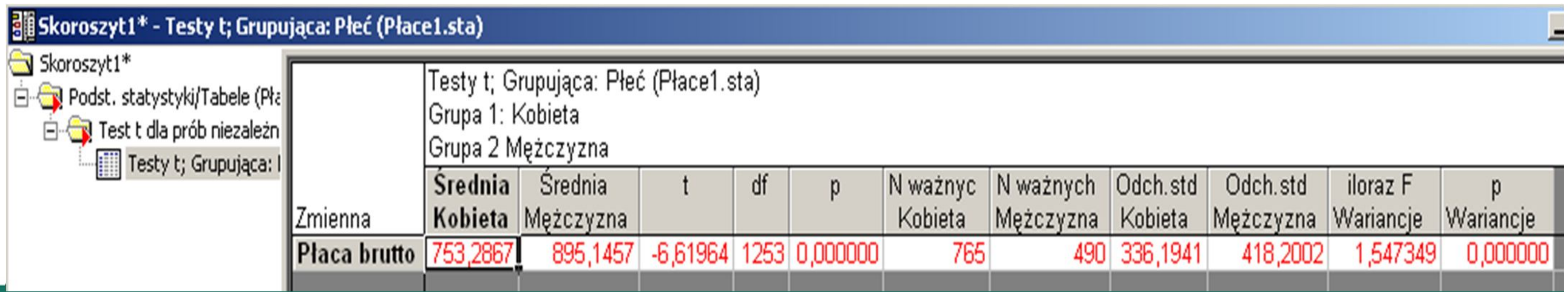
Stawiam pod wątpliwość twierdzenie, że płeć nie ma wpływu na wysokość zarobków w Polsce, jeśli by tak było to nie powinno być różnic pomiędzy średnimi wartościami zarobków kobiet i mężczyzn.

Hipotezą zerową jest zdanie: Zarobki mężczyzn i kobiet nie różnią się

$H_0 : m_1 = m_2$ przy hipotezie alternatywnej $H_1 : m_1 \neq m_2$,

Obliczenia w programie Statistica





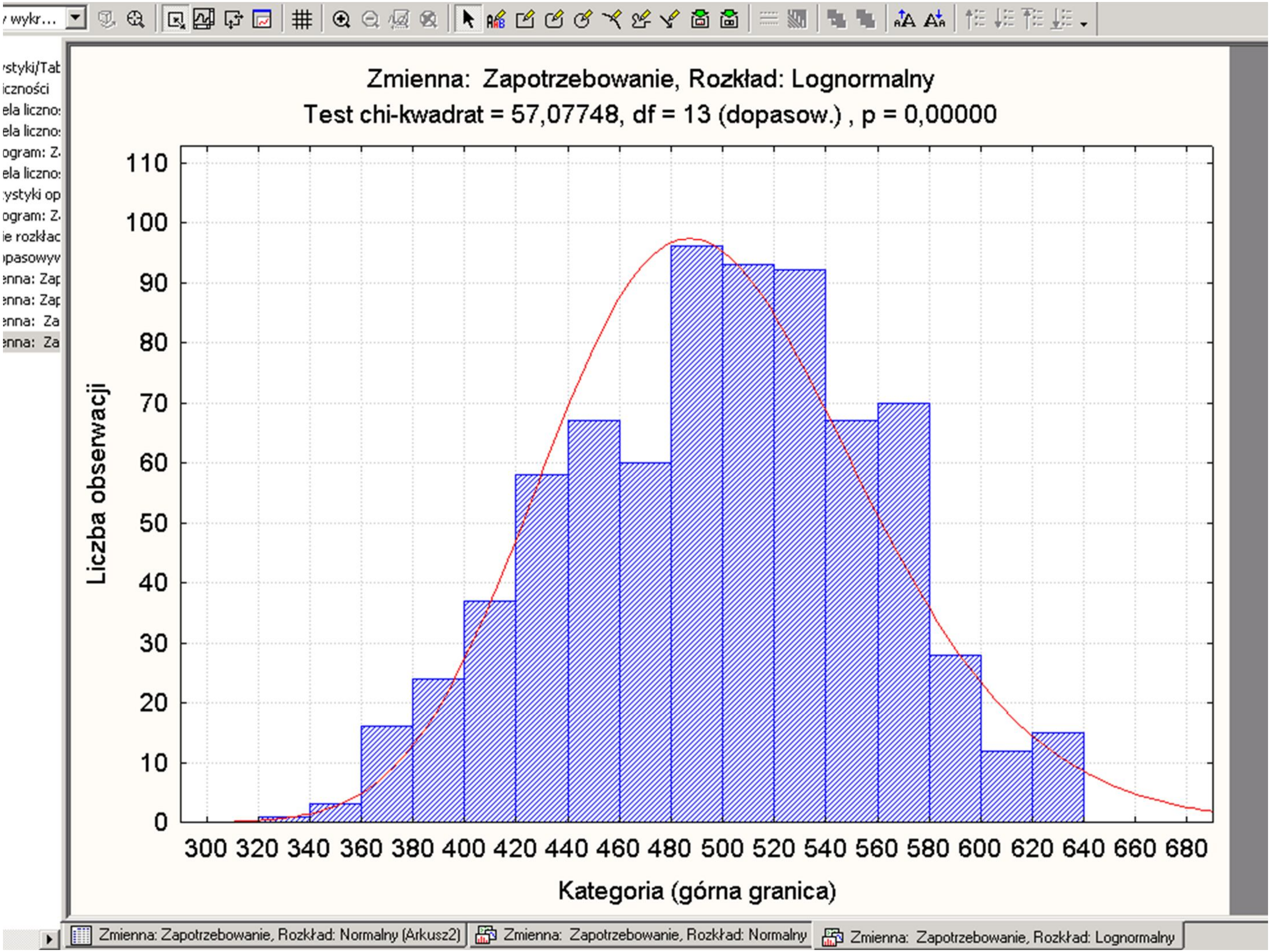
Testy t; Grupująca: Płeć (Place1.sta)
Grupa 1: Kobieta
Grupa 2 Mężczyzna

Zmienna	Średnia Kobieta	Średnia Mężczyzna	t	df	p	N ważnyc Kobieta	N ważnych Mężczyzna	Odch.std Kobieta	Odch.std Mężczyzna	iloraz F Wariancje	p Wariancje
Płaca brutto	753,2867	895,1457	-6,61964	1253	0,000000	765	490	336,1941	418,2002	1,547349	0,000000

Weryfikacja hipotez dotyczących postaci nieznanego rozkładu - Testy zgodności .

Podstawowe działania:

- Konstrukcja rozkładu empirycznego (najlepiej kilku rozkładów o różnej liczbie klas)
- Ocena podobieństwa rozkładu empirycznego do określonego rozkładu teoretycznego – postawienie hipotezy zerowej.
- Przyjęcie odpowiedniej statystyki, która może służyć za test do weryfikacji hipotezy zerowej



Zmienna: Zapotrzebowanie, Rozkład: Lognormalny
Test chi-kwadrat = 57,07748, df = 13 (dopasow.) , p = 0,00000
Liczba obserwacji
0
10
20
30
40
50
60
70
80
90
100
110
300 320 340 360 380 400 420 440 460 480 500 520 540 560 580 600 620 640 660 680
Kategoria (górna granica)
Zmienna: Zapotrzebowanie, Rozkład: Normalny (Arkusz2)
Zmienna: Zapotrzebowanie, Rozkład: Normalny
Zmienna: Zapotrzebowanie, Rozkład: Lognormalny

Test χ^2 Pearsona

- Suma kwadratów różnic $(n_j - n*p_j)$ tzn.

$$\sum_{j=1}^{r+1} \left(n_j - np_j\right)^2$$

może być miarą zgodności rozkładu zaobserwowanego w próbce z rozkładem hipotetycznym

- K. Pearson udowodnił, że statystyka

$$\chi^2 = \sum_{j=1}^{r+1} \frac{\left(n_j - np_j\right)^2}{np_j} \qquad (*)$$

ma, gdy $n \to \infty$, rozkład chi-kwadrat o r stopniach swobody

Test χ^2 Pearsona

- Statystyka określona wzorem (*), znana jest pod nazwą ***test χ^2 Pearsona.***
- Statystyka ta nie zależy od postaci dystrybuanty cechy X, a tylko od prawdopodobieństw $p_j = P(X \in I_j)$, przy czym podział na przedziały I_j jest zupełnie dowolny.
- Taki sam układ prawdopodobieństw $p1, p2, ..., p_{r+1}$ może odpowiadać wielu różnym rozkładom zarówno typu ciągłego jak i skokowego, stąd test χ^2 powinien być używany do weryfikowania hipotezy dotyczącej układu prawdopodobieństw a nie postaci rozkładu cechy X w populacji.
- W teście χ^2 ,
  - hipoteza zerowa dotyczy klasy wszystkich rozkładów dla których $P(X \in I_j) = p_j$,
  - hipoteza alternatywna obejmuje klasę wszystkich tych rozkładów, dla których co najmniej dla jednego j zachodzi $P(X \in I_j) \neq p_j$

Weryfikacja hipotezy o zgodności rozkładu empirycznego z teoretycznym

- Dla danej próbki statystyka χ^2 obliczona ze wzoru (*), będzie mieć taką samą wartość dla wielu różnych rozkładów.

- Przyjęcie hipotezy zerowej oznacza, że każdy rozkład należący do danej klasy może mieć zastosowanie do opisu zjawiska.

- Kierując się wiedzą o zjawisku, najczęściej wybiera się jeden z rozkładów należących do hipotezy zerowej, stąd często upraszcza się problem stosowania testu χ^2 formułując hipotezę zerową jako przypuszczenie, że cecha X ma w populacji rozkład określonej postaci (czyli o pisany konkretną dystrybuantą)

- Mając sprecyzowaną hipotezę zerową i wybrany test do weryfikacji dalej postępowanie przebiega jak w testach parametrycznych.

Algorytm realizacji testu χ^2 Pearsona

- Przyjmijmy poziom istotności α,
- Odczytać z tablic rozkładu χ^2 wartość krytyczną χ^2_α dla zadanej wartości α i r stopni swobody
- Obliczać wartość statystyki testowej χ^2,
- Porównać wartości $\chi^2_{obliczone}$ z wartością krytyczną χ^2_α
- Ponieważ

  zatem hipotezę H_0 odrzucamy ilekroć stwierdzimy, że

$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

H_0 przyjmujemy gdy

$$\chi^2_{obliczone} > \chi^2_\alpha$$

$$\chi^2_{obliczone} \leq \chi^2_\alpha$$

Zastosowania testu χ^2 –przykład1

- Przeprowadzono obserwacje dotyczące wypadków drogowych na określonym terenie, spowodowanych przez kierowców będących w stanie nietrzeźwym. Wyniki:

Pn	Wt	Śr	Cz	Pt	So	N
19	15	16	14	13	18	17

Na poziomie $\alpha = 0{,}05$ zweryfikować hipotezę, że dla każdego dnia tygodnia jest takie samo prawdopodobieństwo wypadku spowodowanego przez kierowcę będącego w stanie nietrzeźwym.

Wykonanie testu

- Dla α = 0,05 oraz r=6 stopni swobody znajduję w tablicach $\boldsymbol{\chi^2_\alpha}$ **= 12,592**

- obliczam wartość statystyki χ^2 według wzoru (*) , przy czym przyjmuję
  - $n=112$
  - $p_1=p_2=\ldots p_7=1/7$
  - $np_j=112/7=16$
  - liczności n_j biorę z tabelki i obliczam
- $\chi^2_{obliczone}=(9+1+0+4+9+4+1+)/16 = 1,75$

- Ponieważ $\chi^2_{obliczone} = 1,75 < \chi^2_\alpha = 12,592$, zatem nie ma podstaw do odrzucenia hipotezy zerowej, zatem utwierdzilismy sie w przekonaniu, że prawdopodobieństwo spowodowania wypadku na badanym terenie przez nietrzeźwego kierowcę jest jednakowe dla każdego dnia tygodnia.

Zastosowania testu χ^2 –przykład 2

STATISTICA - Płace.sta

Plik Edycja Widok Wstaw Format Statystyka Data Mining Wykresy Narzędzia Dane Okno Pomoc

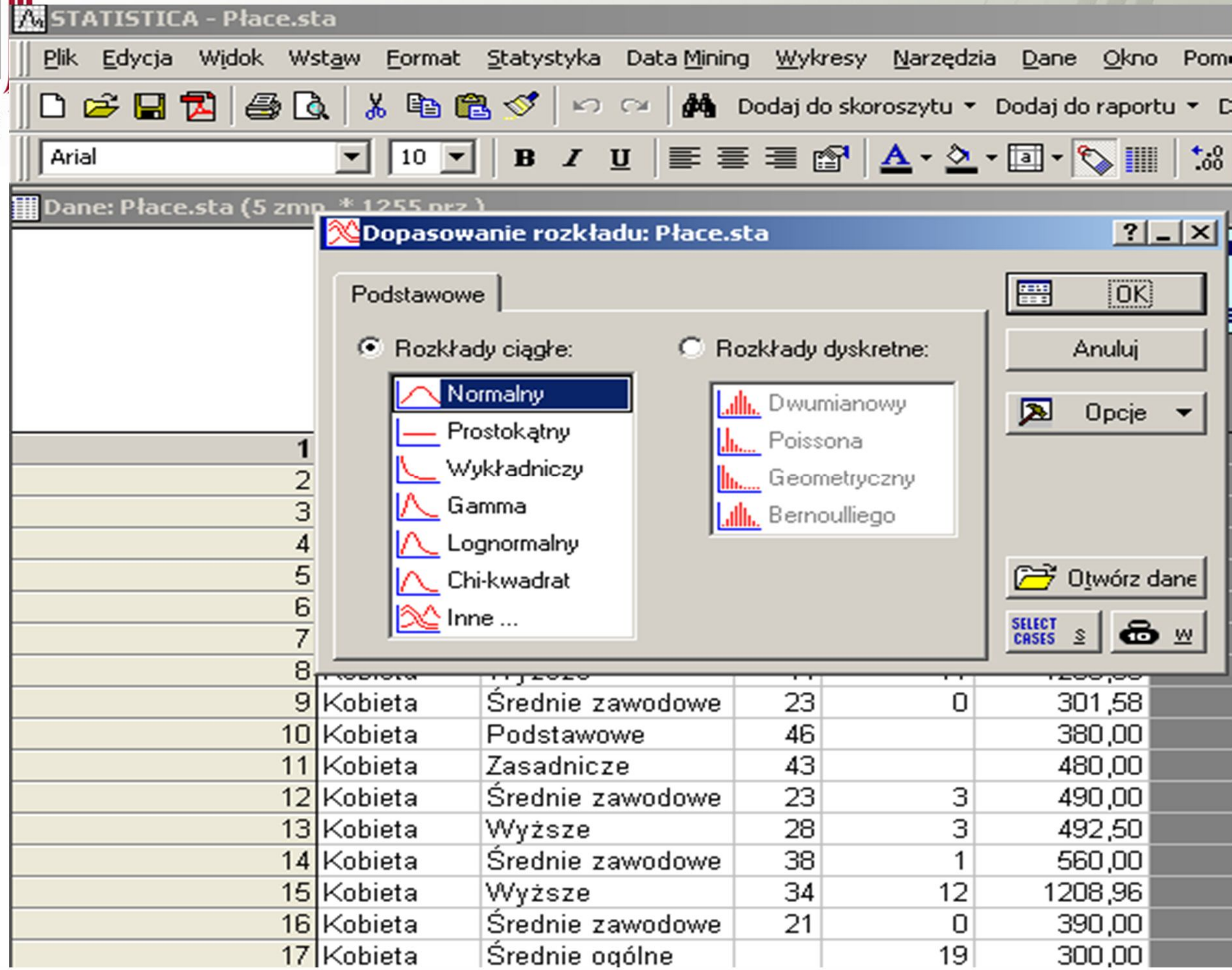
Dodaj do skoroszytu Dodaj do raportu Dodaj do MS

Arial 10

Dane: Płace.sta (5 zmn. * 1255 prz.)

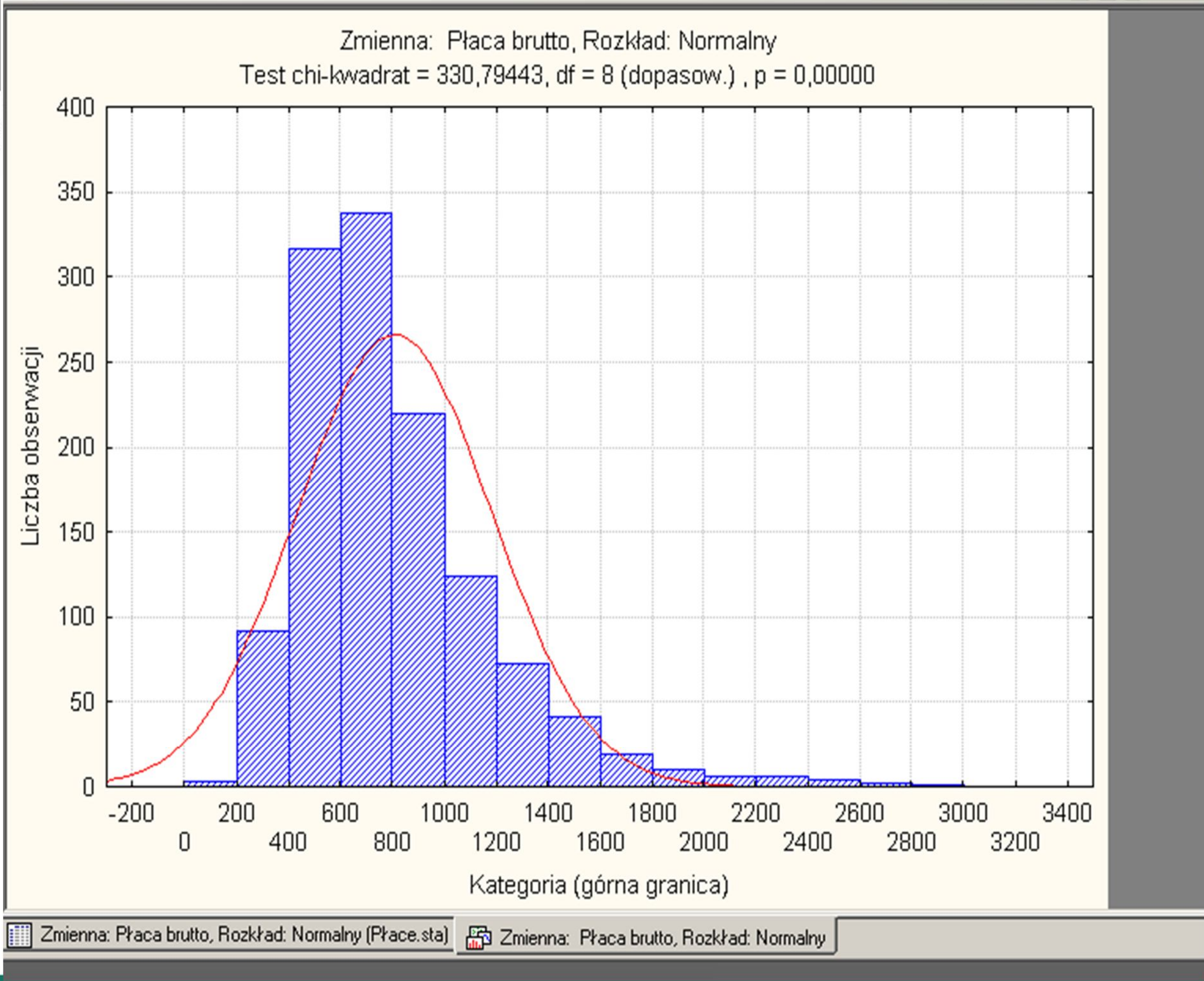
Dane z badań przeprowadzonych w 1996 roku przez Uniwersytet W
Ekonomiczną w Krakowie. Rozesłano ankiety do do 5000 pracownik
Ankiety zwróciło 1255 osób. Arkusz danych zawiera wybrane informa

	1 Płeć	2 Wykształcenie	3 Wiek	4 Staż pracy	5 Płaca brutto
1	Kobieta	Podstawowe	42		420,00
2	Kobieta	Wyższe	34	2	1268,30
3	Kobieta	Średnie zawodowe		17	862,00
4	Kobieta	Zasadnicze	44	25	662,09
5	Kobieta	Średnie zawodowe	38	5	542,70
6	Kobieta	Wyższe	45		816,00
7	Kobieta	Średnie zawodowe	48	3	883,00
8	Kobieta	Wyższe	44	11	1288,00
9	Kobieta	Średnie zawodowe	23	0	301,58
10	Kobieta	Podstawowe	46		380,00
11	Kobieta	Zasadnicze	43		480,00
12	Kobieta	Średnie zawodowe	23	3	490,00
13	Kobieta	Wyższe	28	3	492,50
14	Kobieta	Średnie zawodowe	38	1	560,00
15	Kobieta	Wyższe	34	12	1208,96
16	Kobieta	Średnie zawodowe	21	0	390,00
17	Kobieta	Średnie ogólne		19	300,00
18	Kobieta	Średnie zawodowe	48	1	380,00
19	Kobieta	Średnie zawodowe	39	2	400,00
20	Kobieta	Średnie zawodowe	22	3	445,00
21	Kobieta	Wyższe	26	2	569,62
22	Kobieta	Średnie zawodowe	24	5	590,00
23	Kobieta	Średnie zawodowe	24	5	795,00
24	Kobieta	Wyższe	30	5	1224,00
25	Kobieta	Wyższe	32	3	890,00


STATISTICA - Płace.sta
Plik Edycja Widok Wstaw Format Statystyka Data Mining Wykresy Narzędzia Dane Okno
Dodaj do skoroszytu Dodaj do raportu
Arial 10
Dane: Płace.sta (5 zmn * 1255 prz.)
Dopasowanie rozkładu: Płace.sta
Podstawowe
Rozkłady ciągłe:
Normalny
Prostokątny
Wykładniczy
Gamma
Lognormalny
Chi-kwadrat
Inne ...
Rozkłady dyskretne:
Dwumianowy
Poissona
Geometryczny
Bernoulliego
OK
Anuluj
Opcje
Otwórz dane
SELECT CASES
9 Kobieta Średnie zawodowe 23 0 301,58
10 Kobieta Podstawowe 46 380,00
11 Kobieta Zasadnicze 43 480,00
12 Kobieta Średnie zawodowe 23 3 490,00
13 Kobieta Wyższe 28 3 492,50
14 Kobieta Średnie zawodowe 38 1 560,00
15 Kobieta Wyższe 34 12 1208,96
16 Kobieta Średnie zawodowe 21 0 390,00
17 Kobieta Średnie ogólne 19 300,00

Zmienna: Płaca brutto, Rozkład: Normalny (Płace.sta)
Chi-kwadrat = 330,79443, df = 8 (dopasow.) , p = 0,00000

Górna Granica	**Obserw. Liczność**	Skumulow. Obserw.	Procent Obserw.	Skumul. % Obserw.	Oczekiwana Liczność	Skumulow. Oczekiwana	Procent Oczekiwana	Skumul. % Oczekiwana	Obserw. - Oczekiwana
<= 0,00000	0	0	0,00000	0,0000	19,9452	19,945	1,58926	1,5893	-19,9452
200,00000	3	3	0,23904	0,2390	46,6176	66,563	3,71455	5,3038	-43,6176
400,00000	92	95	7,33068	7,5697	107,8147	174,377	8,59081	13,8946	-15,8147
600,00000	316	411	25,17928	32,7490	189,2876	363,665	15,08268	28,9773	126,7124
800,00000	338	749	26,93227	59,6813	252,3054	615,971	20,10402	49,0813	85,6946
1000,00000	220	969	17,52988	77,2112	255,3372	871,308	20,34560	69,4269	-35,3372
1200,00000	124	1093	9,88048	87,0916	196,1940	1067,502	15,63299	85,0599	-72,1940
1400,00000	73	1166	5,81673	92,9084	114,4512	1181,953	9,11962	94,1795	-41,4512
1600,00000	41	1207	3,26693	96,1753	50,6846	1232,638	4,03861	98,2181	-9,6846
1800,00000	19	1226	1,51394	97,6892	17,0369	1249,674	1,35752	99,5757	1,9631
2000,00000	10	1236	0,79681	98,4861	4,3459	1254,020	0,34629	99,9219	5,6541
2200,00000	6	1242	0,47809	98,9641	0,8411	1254,862	0,06702	99,9890	5,1589
2400,00000	6	1248	0,47809	99,4422	0,1235	1254,985	0,00984	99,9988	5,8765
2600,00000	4	1252	0,31873	99,7610	0,0137	1254,999	0,00110	99,9999	3,9863
2800,00000	2	1254	0,15936	99,9203	0,0012	1255,000	0,00009	100,0000	1,9988
3000,00000	1	1255	0,07968	100,0000	0,0001	1255,000	0,00001	100,0000	0,9999
3200,00000	0	1255	0,00000	100,0000	0,0000	1255,000	0,00000	100,0000	-0,0000
<nieskończoność	0	1255	0,00000	100,0000	0,0000	1255,000	0,00000	100,0000	-0,0000

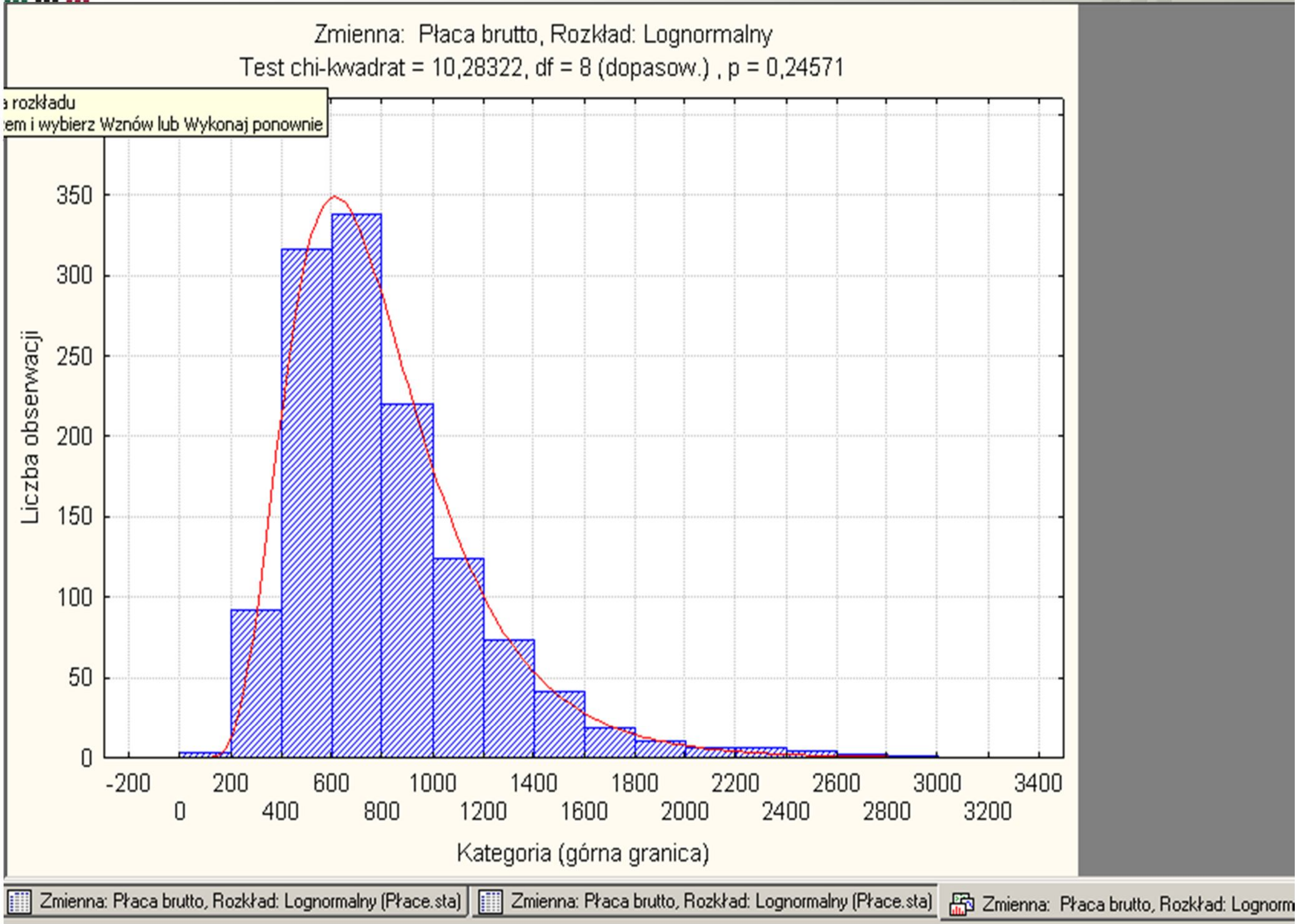


Zmienna: Płaca brutto, Rozkład: Normalny
Test chi-kwadrat = 330,79443, df = 8 (dopasow.) , p = 0,00000
Liczba obserwacji
Kategoria (górna granica)
Zmienna: Płaca brutto, Rozkład: Normalny (Płace.sta)
Zmienna: Płaca brutto, Rozkład: Normalny

Zmienna: Płaca brutto, Rozkład: Lognormalny (Płace.sta)
Chi-kwadrat = 10,28322, df = 8 (dopasow.) , p = 0,24571

vanie rozkładu (Płace.sta)
awym klawiszem i wybierz Wznów lub Wykonaj ponownie

Granica	Liczność	Obserw.	Procent Obserw.	Skumul. % Obserw.	Oczekiwana Liczność	Skumulow. Oczekiwana	Procent Oczekiwana	Skumul. % Oczekiwana	Obserw. - Oczekiwana
<= 0,00000	0	0	0,00000	0,0000	0,0000	0,000	0,00000	0,0000	0,0000
200,00000	3	3	0,23904	0,2390	1,4237	1,424	0,11344	0,1134	1,5763
400,00000	92	95	7,33068	7,5697	94,5011	95,925	7,52997	7,6434	-2,5011
600,00000	316	411	25,17928	32,7490	300,1623	396,087	23,91731	31,5607	15,8377
800,00000	338	749	26,93227	59,6813	327,7563	723,843	26,11604	57,6768	10,2437
1000,00000	220	969	17,52988	77,2112	233,7868	957,630	18,62843	76,3052	-13,7868
1200,00000	124	1093	9,88048	87,0916	138,6107	1096,241	11,04468	87,3499	-14,6107
1400,00000	73	1166	5,81673	92,9084	75,5729	1171,814	6,02174	93,3716	-2,5729
1600,00000	41	1207	3,26693	96,1753	39,7272	1211,541	3,16551	96,5371	1,2728
1800,00000	19	1226	1,51394	97,6892	20,6244	1232,165	1,64338	98,1805	-1,6244
2000,00000	10	1236	0,79681	98,4861	10,7102	1242,875	0,85340	99,0339	-0,7102
2200,00000	6	1242	0,47809	98,9641	5,6024	1248,478	0,44641	99,4803	0,3976
2400,00000	6	1248	0,47809	99,4422	2,9633	1251,441	0,23612	99,7164	3,0367
2600,00000	4	1252	0,31873	99,7610	1,5881	1253,029	0,12654	99,8430	2,4119
2800,00000	2	1254	0,15936	99,9203	0,8632	1253,892	0,06878	99,9117	1,1368
3000,00000	1	1255	0,07968	100,0000	0,4760	1254,368	0,03793	99,9497	0,5240
3200,00000	0	1255	0,00000	100,0000	0,2663	1254,635	0,02122	99,9709	-0,2663
<nieskończoność	0	1255	0,00000	100,0000	0,3652	1255,000	0,02910	100,0000	-0,3652


Zmienna: Płaca brutto, Rozkład: Lognormalny
Test chi-kwadrat = 10,28322, df = 8 (dopasow.) , p = 0,24571
a rozkładu
tem i wybierz Wznów lub Wykonaj ponownie
Liczba obserwacji
0
50
100
150
200
250
300
350
-200
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
Kategoria (górna granica)
Zmienna: Płaca brutto, Rozkład: Lognormalny (Płace.sta)
Zmienna: Płaca brutto, Rozkład: Lognormalny (Płace.sta)
Zmienna: Płaca brutto, Rozkład: Lognorm

Weryfikacja hipotezy o zgodności rozkładu empirycznego z teoretycznym

- Mając sprecyzowaną hipotezę zerową i wybrany test do weryfikacji dalej postępowanie przebiega jak w testach parametrycznych.
- Oblicza się wartość statystyki testowej, i porównuje z wartością krytyczną χ^2_α odczytaną z tablic rozkładu chi – kwadrat dla zadanej wartości α przy r stopniach swobody.
- Ponieważ

$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

- Zatem hipotezę H_0 odrzucamy ilekroć stwierdzimy, że

$$\chi^2_{obliczone} > \chi^2_\alpha$$