Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Statystyka i Opracowanie Danych

W7. Estymacja i estymatory

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

Estymacja parametryczna

- Podstawowym narzędziem szacowania nieznanego parametru jest estymator obliczony na podstawie próby. np. dla wartości oczekiwanej jest to średnia arytmetyczna.
- Liczba możliwych estymatorów konkretnego parametru rozkładu może być duża ale, bierze się pod uwagę tylko te, które posiadają określone właściwości (cechy).
- Estymator ma być zgodny, nieobciążony i najefektywniejszy.
- Ze względu na formę wyniku estymacji wyróżnimy

Estymacja punktowa –gdy szacujemy **liczbową wartość określonego parametru** rozkładu cechy w całej populacji

Estymacja przedziałowa –gdy wyznaczamy **granice przedziału liczbowego, w których,** z określonym prawdopodobieństwem, **mieści się prawdziwa wartość szacowanego parametru**.

Estymator NMW

- Def. Niech θ będzie liczbą rzeczywistą oznaczająca nieznany parametr populacji. Nieobciążonym estymatorem $\hat{\theta}(X_1, X_2, \ldots X_n)$ parametru θ, nazywamy estymatorem nieobciążonym o minimalnej wariancji (estymatorem NMW), jeśli wśród wszystkich nieobciążonych estymatorów szacowanego parametru, nie istnieje estymator, którego wariancja byłaby mniejsza dla jakiejś wartości θ.

Czyli dla wszystkich możliwych wartości θ i wszystkich nieobciążonych estymatorów

$$\tilde{\theta} = \tilde{T}(X_1, X_2, \ldots X_n)$$

$$Var(\hat{\theta}) \leq Var(\tilde{\theta})$$

Błąd średniokwadratowy estymatora

- Błędem średniokwadratowym estymatora $\hat{\theta}$, nazywamy wartość średnią kwadratu odległości $\mu_{(\hat{\theta}-\theta)^2}$

- Dla każdego estymatora $\hat{\theta}$ jego błąd średniokwadratowy jest sumą jego wariancji i kwadratu obciążenia

$$\mu_{(\hat{\theta}-\theta)^2} = \sigma^2_{\hat{\theta}} + (\mu_{\tilde{\theta}} - \theta)^2$$

- Błędem standardowym estymatora $\hat{\theta}$ parametru θ nazywamy dowolny estymator jego odchylenia standardowego

$$SE_{\tilde{X}} = \frac{S}{\sqrt{n}}$$

Przykłady estymatorów punktowych

Estymatorem zgodnym, nieobciążonym
i najefektywniejszym dla **wartości oczekiwanej**
w populacji jest średnia arytmetyczna

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

Mediana wyznaczona z próby jest nieobciążonym
ale mniej efektywnym od średniej arytmetycznej
estymatorem wartości oczekiwanej

Przykłady estymatorów punktowych

Niech m oznacza liczbę wyróżnionych elementów w próbie n elementowej (np. liczbę wyrobów wadliwych), wtedy statystyka będąca częstością w próbie

$$\overline{P} = \frac{m}{n}$$

jest estymatorem zgodnym, nieobciążonym i najefektywniejszym frakcji P w populacji

Przykłady estymatorów punktowych

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- S^2 jest estymatorem zgodnym ale obciążonym wariancji w całej populacji.

- Wskazówka: tego wzoru używamy obliczając wariancję z całej populacji, natomiast do estymacji na podstawie próbki należy wynik z próby pomnożyć przez współczynnik n/(n-1)

Obciążoność i nieobciążoność estymatora

Odchylenie standardowe dane wzorem

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

jest estymatorem obciążonym odchylenia standardowego w całej populacji, a nieobciążonym jest odchylenie obliczone z wzoru

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

Estymator obciążony wariancji

Estymator wariancji

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

Stąd obliczymy

$$E(s^2) = \frac{1}{n}\sum_{i=1}^{n} E(x_i^2) - E[(\overline{x})^2]$$

Obliczmy:

$$E[\overline{(x)^2}] = E(\frac{1}{n}\sum_{i=1}^{n} x_i)^2 = E(\frac{1}{n^2}\sum_{i=1}^{n} x_i^2 + \frac{1}{n^2}\sum_{j\neq k} x_j x_k) = \frac{1}{n^2}\sum_{i} E(x_i^2) + \frac{1}{n^2}\sum_{j\neq k} E(x_j x_k)$$

Zatem:

$$E(s^2) = \frac{1}{n} nE(x^2) - \frac{1}{n^2} nE(x^2) - \frac{n-1}{n}[E(x)]^2 = \frac{n-1}{n} D^2(x)$$

Estymator asymptotycznie nieobciążony

$$b(s^2) = \frac{n-1}{n} D^2(x) - D^2(X) = \frac{-1}{n} D^2(x)$$

$$\overline{s^2} = \frac{n}{n-1} s^2$$

$$E(\overline{s^2}) = E[\frac{n}{n-1} s^2] = \frac{n}{n-1}\frac{n-1}{n} D^2(X) = D_2(X)$$

Def: Estymator T_n jest asymptotycznie nieobciążony, jeśli

$$\lim_{n\to\infty} b(T_n) = 0$$

Stąd dla n $\to\infty$ przyjmuje się s^2 jako estymator wariancji

Estymacja przedziałowa polega na wyznaczeniu granic przedziału liczbowego, w którym, z określonym prawdopodobieństwem, równym (1-α), zawiera się wartość szacowanego parametru

Estymacja przedziałowa

$$P\left(\Theta_d (X_1,\ldots ,X_n) < \Theta < \Theta_g (X_1,\ldots ,X_n)\right) = 1-\alpha$$

- Losowy przedział (Θd ,Θg) nazywa się przedziałem ufności parametru Θ
- Granice przedziału ufności są funkcjami zmiennych losowych $X_1,\ldots ,X_n$

- $1-\alpha$ nazywamy poziomem ufności (lub współczynnikiem ufności)

Zwykle przyjmuje się $1-\alpha = 0,99$ lub 0,95 lub 0,90 w zależności od rozpatrywanego zagadnienia

Przedział ufności dla wartości oczekiwanej, gdy znane jest odchylenie standardowe

Cecha X ma w populacji rozkład normalny N(μ, σ), odchylenie standardowe σ jest znane.

Estymatorem wartości oczekiwanej μ, uzyskanym MNW jest średnia arytmetyczna, która jest zmienną losową o rozkładzie N(μ, σ/√n)

Po standaryzacji otrzymuję zmienną U o rozkładzie N(0,1)

$$U = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

gdzie:

n jest liczbą elementów z próby losowej

$\overline{X}$ oznacza średnią arytmetyczną obliczoną z próby losowej

σ odchylenie standardowe populacji

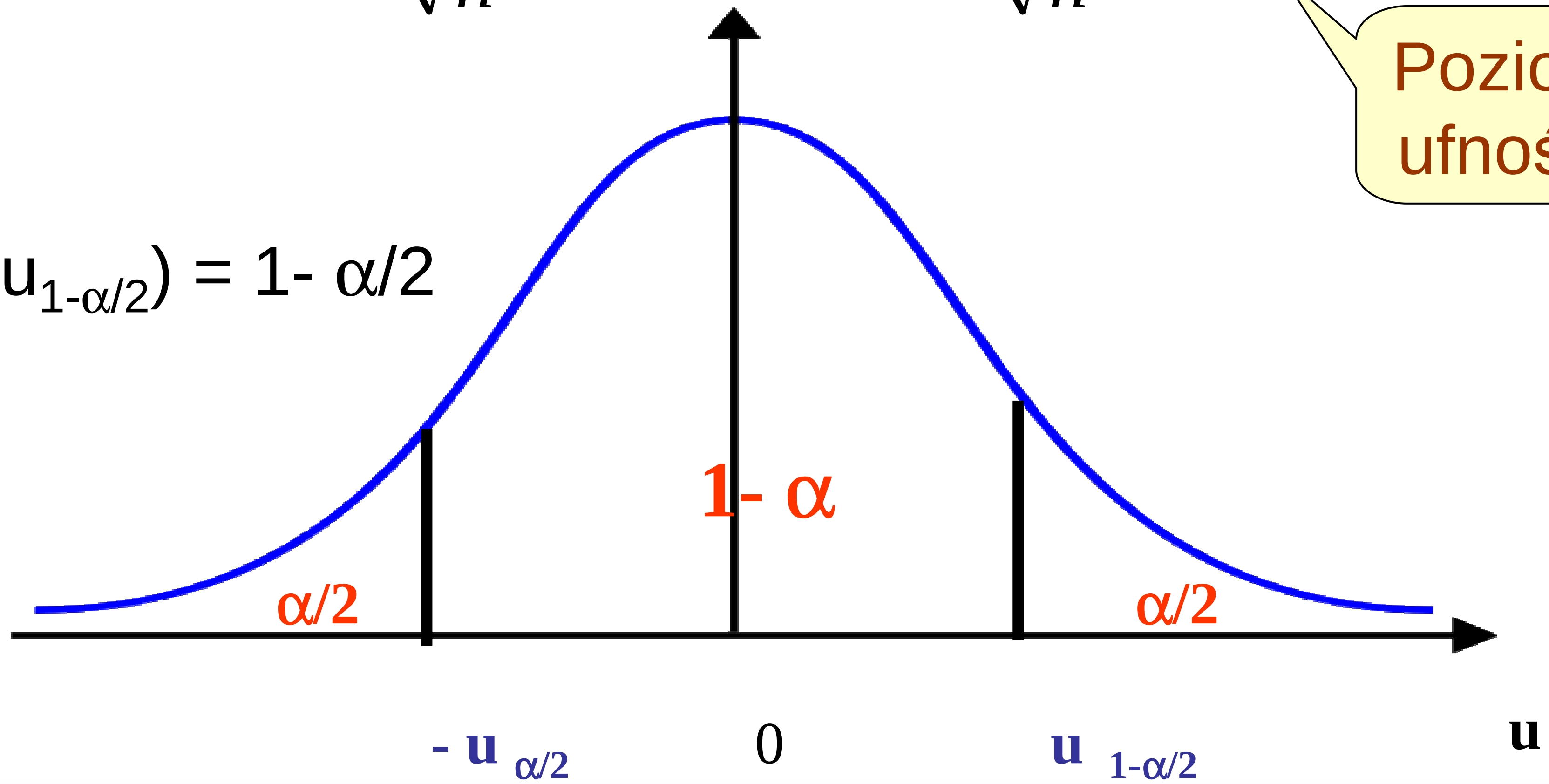
Przedział ufności dla wartości oczekiwanej

gdy <u>znane</u> jest odchylenie standardowe σ

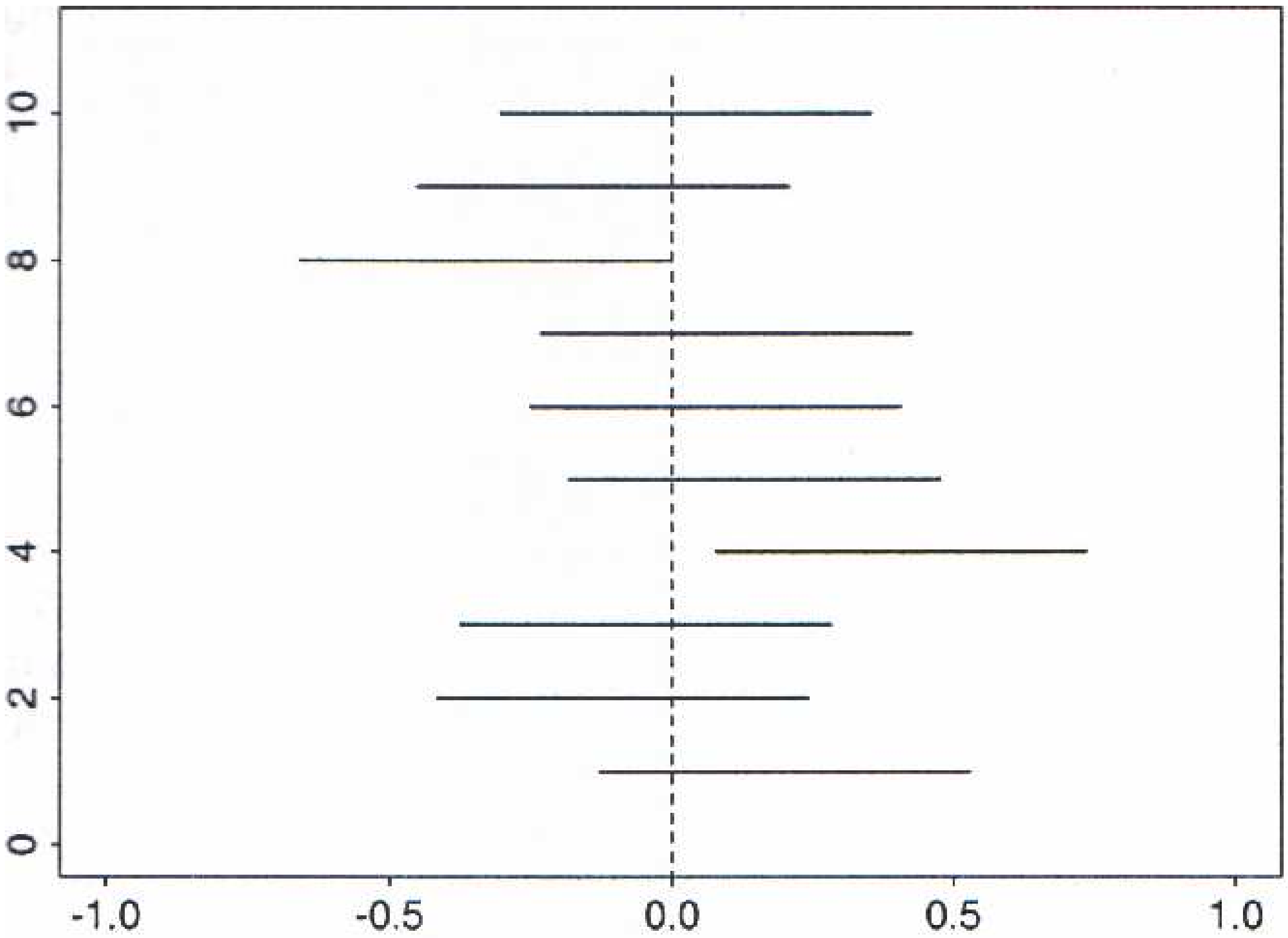
$$P\left(\overline{X} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

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

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

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

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

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 $\sigma=0.1$

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,

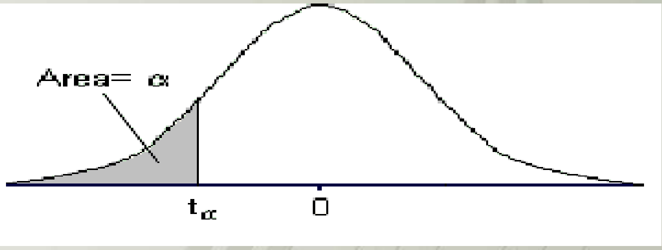
S jest odchyleniem standardowym obliczonym z próby.

Przedział ufności dla wartości oczekiwanej, gdy odchylenie standardowe jest nieznane

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 $2\, t_{\alpha,n-1} S/\sqrt{n-1}$



Kwantyle $t_{1-\alpha}(n)$, rzędu $1-\alpha$, rozkładu Studenta o n stopniach swobody

n	1-α									
	0.6	**0.75**	**0.9**	**0.95**	**0.975**	**0.99**	**0.995**	**0.9975**	**0.999**	**0.9995**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

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
- $1-\alpha = 0,95$
- $1-\alpha = 0,90$
- $1-\alpha = 0,98$

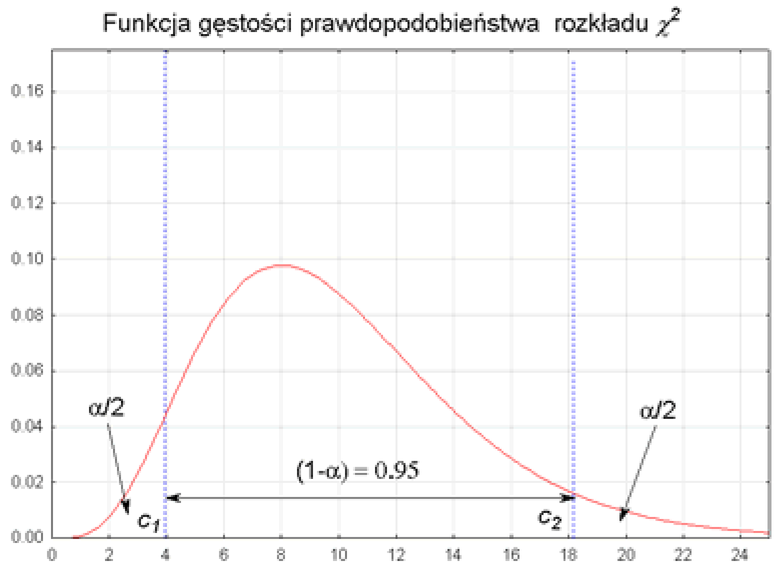
Przedział ufności dla wariancji w populacji normalnej

- 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

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

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



Dystrybuanta odwrotna F^{-1} rozkładu chi-kwadrat dla ustalonej liczby stopni swobody oraz poziomu istotności

Poziom istotności Stopnie swobody	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

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

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)

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 $1-\alpha = 0,95$