AdTrans project , 2023-1-PL01-KA220-HED-000158917

ADTRANS
ADVANCED DATA ANALYSIS FOR TRANSPORTATION AND MOBILITY ACADEMIC TEACHERS

Mathematical Statistics - Additional Parameters

Anna Gobis

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

Confidence interval for fraction (structure index)

- Model: The general population has a binomial distribution with parameter p, i.e., the elements of the population are divided into two classes, where the fraction of distinguished elements in the population is p, which is not a small fraction ($p > 0.05$). A large number n of elements were randomly sampled from the population ($n > 100$).
- Confidence interval for the structure index p of the general population:

$$P\left\{\frac{m}{n} - u_\alpha \cdot \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} < p < \frac{m}{n} + u_\alpha \cdot \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\right\} \approx 1 - \alpha$$

Where:

m — number of distinguished elements found in the sample,

n — size of the independently drawn sample,

u_α — value of the random variable U which has a standard normal distribution, read from tables for the value $1 - \frac{\alpha}{2}$

Example 5.1-5.2

5.1. We want to estimate what percentage of working people living in Gdańsk work remotely from home. To this end, an independent random sample of $n = 900$ people was taken. Based on the study, it was determined that in this sample there are $m = 300$ people who work remotely from home. Assuming a confidence coefficient equal to $1 - \alpha = 0.95$, construct a confidence interval for the percentage of the studied category of people working in Gdańsk.

5.2. Among students of the Medical Academy, an independent random sample of 150 students was drawn and asked whether they smoke cigarettes. 114 students replied that they do not smoke. Estimate, using the interval method, the percentage of non-smoking students at this university, assuming a confidence coefficient of 0.90.

Interval estimation for variance and standard deviation

Model I	Model II

Confidence interval for variance

- Model I – The general population has a normal distribution $N(m, \sigma)$ with unknown parameters m and σ. An independently drawn sample of small size n (less than 30) is taken from the population. From the sample, the value s^2 is calculated. Then, the confidence interval for the variance σ^2 of the general population is:

$$P\left\{\frac{ns^2}{c_2} < \sigma^2 < \frac{ns^2}{c_1}\right\} = 1 - \alpha, \quad \text{gdzie} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Where:

c_1, c_2 – values of the χ^2 variable determined from the χ^2 distribution table for $n - 1$ degrees of freedom and confidence coefficient $1 - \alpha$.

Tablica 6. Rozkład χ^2

n \ α	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	α / n
1	$0,0^3157$	$0,0^3628$	0,00393	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827	1
2	0,0201	0,0404	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815	2
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,268	3
4	0,297	0,429	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,465	4
5	0,554	0,752	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,517	5
6	0,872	1,134	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457	6
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322	7
8	1,646	2,032	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125	8
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877	9
10	2,558	3,059	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588	10
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264	11
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909	12
13	4,107	4,765	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528	13
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123	14
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697	15
16	5,812	6,614	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252	16
17	6,408	7,255	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790	17
18	7,015	7,906	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312	18
19	7,633	8,567	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820	19
20	8,260	9,237	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315	20
21	8,897	9,915	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797	21
22	9,542	10,600	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268	22
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728	23
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179	24
25	11,524	12,697	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620	25
26	12,198	13,409	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052	26
27	12,879	14,125	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476	27
28	13,565	14,847	16,928	19,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893	28
29	14,256	15,574	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302	29
30	14,953	16,306	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703	30

Chi-squared distribution tables

Here is the English translation:

- The value c_1 is found for $1 - \frac{\alpha}{2}$
- The value c_2 is found for $\frac{\alpha}{2}$

Confidence interval for variance

- Model II – The general population has a normal distribution $N(m, \sigma)$ or is approximately normal with unknown parameters m and σ. From the population, a large independent sample of n elements (at least several dozen) is drawn. From the sample, the value $s = \sqrt{s^2}$ is calculated. Then, an approximate confidence interval for the standard deviation σ of the general population is:

$$P\left\{\frac{s}{1+\frac{u_\alpha}{\sqrt{2n}}} < \sigma < \frac{s}{1-\frac{u_\alpha}{\sqrt{2n}}}\right\} \approx 1-\alpha, \quad \text{gdzie} \quad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

u_α – value of the random variable U having a standard normal distribution, read from tables for the value $1 - \frac{\alpha}{2}$

Examples 5.3-5.5

5.3. In a study of the durability of a certain material, $n = 4$ independent measurements of durability were carried out and the following results were obtained: 120, 102, 135, 115. A confidence interval for the variance σ^2 of the durability of this material should be constructed, assuming a confidence coefficient of $1 - \alpha = 0.96$.

5.4. In survey research, the monthly food expenses of Gdańsk households were examined. A sample of 632 households was drawn, among which the average monthly food expenses amounted to 1570 PLN, and the standard deviation of these expenditures was 224 PLN. Assuming a confidence coefficient of 0.90, construct a confidence interval for the standard deviation σ of food expenses of households in Gdańsk.

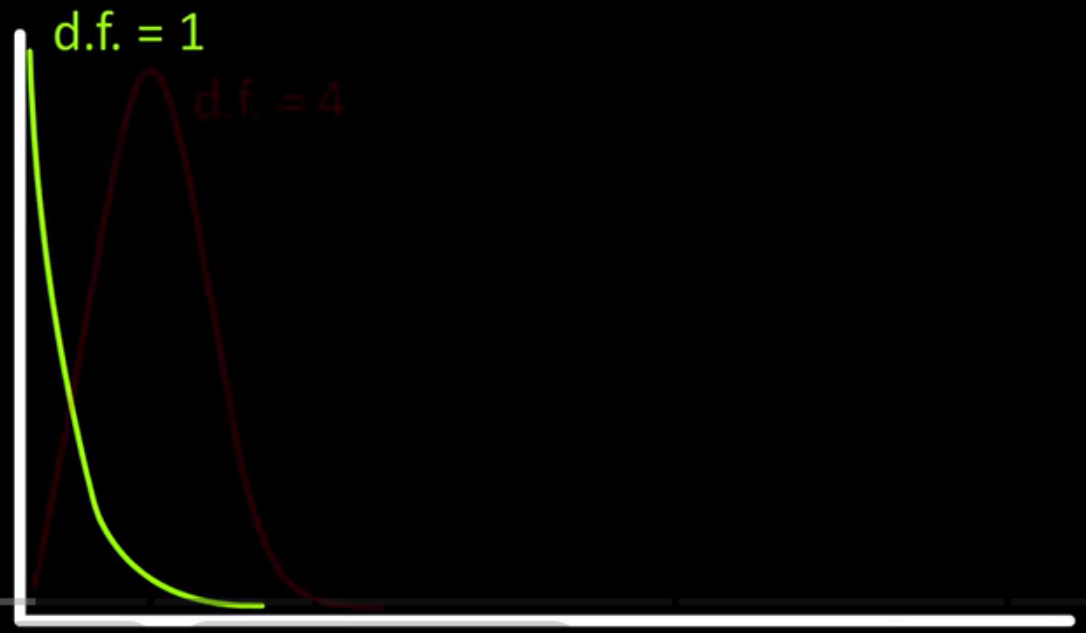
5.5. To estimate the accuracy of a certain measuring instrument, 5 independent measurements of the length of a certain segment were made, and the following results (in mm) were obtained: 15.15; 15.20; 15.04; 15.14; 15.22. Assuming a confidence coefficient of 0.98, construct a confidence interval for the unknown variance of measurements with this instrument.

Confidence Intervals For Variance And Standard Deviation

confidence intervals for variance and standard deviation
- use chi-square distribution

χ^2 chi-square symbol

family of curves based on the degrees of freedom



Odtwórz

0:26 / 7:36 Czym jest rozkład chi-kwadrat? >

https://www.youtube.com/watch?v=bLd0W2WnWqA

Independent tasks

1. Seven independent measurements of the initial velocity of a bullet fired from a certain gun were made, and the following results were obtained (in m/s): 605.4, 604.6, 605.4, 605.2, 607.0, 603.8, 603.3. Assuming a confidence coefficient of 0.98, estimate the standard deviation of the initial velocity of a bullet fired from this gun using the interval method.

2. In order to estimate the variation in the unit production cost of a certain article manufactured by different plants, a sample of $n = 80$ production plants was independently drawn, and the following results were obtained for the cost study (in PLN):

Unit Cost	Number of Plants
20 – 40	10
40 – 60	16
60 – 80	24
80 – 100	18
100 – 120	12

Assuming a confidence coefficient of 0.95, estimate by interval method the standard deviation of the unit production cost of this article.

Determining sample size

Model I	Model II	Model III

Determining sample size

Model I – The general population has a normal distribution $N(m, \sigma)$ or is approximately normal. The population variance σ^2 is known. We want to estimate the unknown mean value m of the population based on a sample of n independent measurements. If we require that, for a given confidence coefficient $1 - \alpha$, the maximum error in estimating the mean m (i.e., half the length of the confidence interval) does not exceed a predefined value d, the sample size is determined by the following formula:

$$n = \frac{u_\alpha^2 \sigma^2}{d^2}$$

Where:

u_α – value of the random variable U with a standard normal distribution, read from tables for the value $1 - \frac{\alpha}{2}$

d – allowable, predefined maximum error of estimating the mean m

Determining sample size

Model II – The general population has a normal distribution $N(m, \sigma)$. The population variance σ^2 is unknown, but the value of the sample statistic s^2 obtained from a small sample of n_0 elements is known. We want to estimate the unknown mean value m of the population based on a sample of n independent measurements. If we require that, with a given confidence coefficient $1 - \alpha$, the maximum error in estimating the mean m (i.e., half the length of the confidence interval) does not exceed a predefined value d, the sample size is determined by the following formula:

$$n = \frac{t_\alpha^2 s^2}{d^2}$$

Where:

t_α – the value of Student's t variable read from the tables for the given α and

$r = n - 1$ degrees of freedom

Determining sample size

Model III: The general population has a binomial distribution with parameter p, i.e., the population elements are divided into two classes, with the fraction of distinguished elements in the population being p. The parameter p should be estimated using the interval method so that, with confidence coefficient $1 - \alpha$, the maximum estimation error for the structure index p does not exceed a predetermined value d.

- If the expected order of magnitude of the estimated fraction p is known:

$$n = \frac{u_\alpha^2 p(1-p)}{d^2}$$

- If the order of magnitude of the estimated structure index p is not known:

$$n = \frac{u_\alpha^2}{4d^2}$$

Confidence Intervals about Population Proportions

https://www.youtube.com/watch?v=3ReWri_jh3M

Example 6.1

Determine how many independent observations should be included in a sample so that, based on it, the average time for a worker to perform a certain technical operation can be estimated with a maximum error of 20 seconds, assuming a confidence coefficient of 0.95. It is known that the time required to perform this technical operation follows a normal distribution $N(m, 40)$.

Example 6.2

We want to estimate the average mass of a certain chemical substance. How many independent experiments should be conducted so that, with a confidence coefficient of 0.95, the average mass can be estimated by the interval method with a maximum error of 0.01 grams, if the variance of a preliminary sample of 5 independent experiments was $s^2 = 0.0006$?

Example 6.3

Determine how many students from a given university should be independently sampled in order to estimate the percentage of students commuting to the university by public transport with a maximum error of 5%, assuming a confidence coefficient of 0.90. It is assumed that the estimated percentage of students commuting by public transport is around 70%.

Example 6.4

How many inhabitants of a certain city should be drawn independently into the sample in order to estimate the percentage of this city's inhabitants suffering from rheumatic diseases, if in estimating this percentage, which is of the order of 20%, we do not want to be mistaken by more than 5%? Assume a confidence coefficient of 0.95.

Example 6.5

How many cows of a certain breed should be independently selected from the sample in order to estimate the average daily milk yield of a cow of this breed with a maximum error of 0.5 l, if it is known that the standard deviation of the daily milk yield of cows of this breed is 2.5 l and the confidence coefficient is assumed to be 0.95?

Additional tasks with solutions

Task 1

The variation in the time needed to bind a book in a bookbinding shop was studied. Twenty orders were randomly selected and it was found that the average time needed to bind a book was 5 hours, with a variance of four hours. We assume that the distribution of the time needed for correction is a normal distribution. What result was obtained if the confidence coefficient was assumed to be $1 - \alpha = 0.90$?

Solution

n = 20 – Small sample

$\bar{x}$ = 5 godzin, $S^2 = 4$, $1 - \alpha = 0{,}90$,

N(m, σ)

$$P\left(\frac{20 \cdot 4}{\chi^2_{\frac{0,1}{2};19}} < \sigma^2 < \frac{20 \cdot 4}{\chi^2_{1-\frac{0,1}{2};19}}\right) = 0{,}9$$

$\chi^2_{0,05;19} = ?$ $\chi^2_{0,95;19} = ?$

Chi-squared distribution quantiles

k\alfa	0,06	0,05	0,04
17	26,8701	27,5871	28,4450
18	28,1370	28,8693	29,7451
19	29,3964	30,1435	31,0367
20	30,6489	31,4104	32,3206
21	31,8949	32,6706	33,5972

k\alfa	0,96	0,95	0,94
17	8,2878	8,6718	9,0083
18	8,9889	9,3905	9,7421
19	9,6983	10,1170	10,4833
20	10,4154	10,8508	11,2314
21	11,1395	11,5913	11,9858

$$\chi^2_{0,05;19} = 30{,}144$$

$$\chi^2_{0,95;19} = 10{,}117$$

Solution

$$P\left(\frac{80}{30{,}144} < \sigma^2 < \frac{80}{10{,}117}\right) = 0{,}9$$

$$P\left(2{,}654 < \sigma^2 < 7{,}907\right) = 0{,}9$$

$$\sigma^2 \in (2{,}654;\ 7{,}907)$$

$$\sigma \in (1{,}629;\ 2{,}812)$$

Task 2

48 wheat grains were selected and their protein content (in percent) was tested. The mean was 16.8[%] and the standard deviation was 2.1[%]. Find the 98% confidence interval for the variance of the protein content in wheat grains of the entire batch.

Solution

n = 48 – **Large sample**

$\bar{x}$ = 16,8%, S = 2,1%, 1 - α = 0,98 (czyli α = 0,02)

$$P\left(\frac{2,1}{1+\frac{u_\alpha}{\sqrt{2\cdot 48}}} < \sigma < \frac{2,1}{1-\frac{u_\alpha}{\sqrt{2\cdot 48}}}\right) = 0,98$$

$$\Phi(u_\alpha) = 1 - \frac{\alpha}{2} = 1 - 0,01 = 0,99 \qquad u_{0,99} = 2,33$$

u	0	0,01	0,02	0,03	0,04
2,00	0,9772	0,9778	0,9783	0,9788	0,9793
2,10	0,9821	0,9826	0,9830	0,9834	0,9838
2,20	0,9861	0,9864	0,9868	0,9871	0,9875
2,30	0,9893	0,9896	0,9898	0,9901	0,9904
2,40	0,9918	0,9920	0,9922	0,9925	0,9927

Solution

$$P\left(\frac{2,1}{1+\frac{2,33}{\sqrt{96}}} < \sigma < \frac{2,1}{1-\frac{2,33}{\sqrt{96}}}\right) = 0,98$$

$$P(1,697 < \sigma < 2,755) = 0,98$$

$$P(2,880 < \sigma^2 < 7,591) = 0,98$$

$$\sigma^2 \in (2,880\,;\,7,591)$$

Task 3

In a certain clinic, among a randomly selected 980 people who underwent X-rays, pathological changes were found in 100 people. Determine the 95% confidence interval for the fraction of sick people among all the people served by this clinic.

Solution

X = 100, n = 980, $1 - \alpha = 0.95$

$$\frac{100}{980} - u_\alpha \cdot \sqrt{\frac{100}{980}\left(1 - \frac{100}{980}\right)/980} < p < \frac{100}{980} + u_\alpha \cdot \sqrt{\frac{100}{980}\left(1 - \frac{100}{980}\right)/980}$$

$$\Phi(u_\alpha) = 1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975 \; u_{0.975} = 1.96$$

$$0.102 - 0.019 < p < 0.102 + 0.019$$

$$0.083 < p < 0.121$$

From 8.3% to 12.1% of the patients at this clinic are sick — with a probability of 0.95.

Solution

$$P\left(\frac{2,1}{1+\frac{2,33}{\sqrt{96}}} < \sigma < \frac{2,1}{1-\frac{2,33}{\sqrt{96}}}\right) = 0,98$$

$$P(1,697 < \sigma < 2,755) = 0,98$$

$$P(2,880 < \sigma^2 < 7,591) = 0,98$$

$$\sigma^2 \in (2,880\,;\,7,591)$$

Sources:

- **Jerzy Greń: Statystyka matematyczna. Modele i zadania. Wydanie czwarte uzupełnione. Państwowe wydawnictwo naukowe. Warszawa, 1974.**
- Kot, S. M., Jakubowski, J., Sokołowski, A.: Statystyka. Wydanie drugie poprawione. Wyd. Difin. Warszawa, 2011.
- Stanisz, A.: Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 1. Statystyki podstawowe. Wyd. StatSoft Polska. Kraków, 2006.
- **Tablice dystrybuanty rozkładu normalnego**
- **Tablice rozkładu *t Studenta***
- **Tablice rozkładu χ^2**
- http://dydaktyka.polsl.pl/roz6/amularczyk/Inne%20dokumenty/Statystyka%20matematyczna/Wyk%C5%82ad%204.pdf