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

ADTRANS
ADVANCED DATA ANALYSIS FOR TRANSPORTATION AND MOBILITY ACADEMIC TEACHERS

Nonparametric tests

Joanna Wachnicka

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.

Chi-squared goodness-of-fit test

Model: From a population, a large independent sample is drawn, and the results are divided into r mutually exclusive classes with sizes n_i in each class, such that $n = \sum n_i$. In this way, an empirical distribution is obtained. Based on the results of this sample, the hypothesis H_0 should be tested, stating that the general population has a hypothetical distribution of a given type.

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(n_i - np_i\right)^2}{np_i}$$

Where:

- n_i — empirical frequency of class i
- np_i — theoretical (hypothetical) frequency of class i
- p_i — probability that a random variable with the given hypothetical distribution takes values belonging to class i

Chi-squared goodness-of-fit test

- The statistic χ^2 under the assumption that H_0 is true has an asymptotic χ^2 distribution with $r - k - 1$ degrees of freedom.
- The critical region is constructed as the right tail based on the χ^2 distribution.
- The computed value of the statistic χ^2 is compared to the critical value χ^2_α read from tables:
  - If $\chi^2 \geq \chi^2_\alpha$,H_0 should be rejected.
  - If $\chi^2 < \chi^2_\alpha$,there are no grounds to reject H_0.

Example

A random sample $n = 200$ of independent observations of monthly food expenditures among families in Gdańsk produced the following distribution of these expenditures (in thousands of PLN):

Expenses	Number of families
1.0 - 1.4	15
1.4 - 1.8	45
1.8 - 2.2	70
2.2 - 2.6	50
2.6 - 3.0	20

At the significance level $\alpha = 0.05$, verify the hypothesis that the distribution of food expenditures is normal.

Kolmogorov λ goodness-of-fit test

Model: The general population has a continuous distribution with cumulative distribution function $F(x)$. A population sample of n elements (at least several dozen) is randomly and independently drawn. Based on the results of this sample, the hypothesis $H_0: F(x) = F_0(x)$,where $F_0(x)$ is the hypothetical cumulative distribution function, should be tested.

- Sort the sample results in ascending order or group them into intervals with right endpoints x_j and sizes n_j.
- For each x_j ,determine the empirical cumulative distribution function $F_n(x_k) = \frac{n_{sk}}{n}$) n_{sk} is the cumulative frequency from the start up to x_k :$n_{sk} = \sum_{j \le k} n_j$.(
- From the hypothetical distribution, for each x_j ,determine the theoretical cumulative distribution function $F(x)$ and compute the absolute value of the difference $F_n(x) - F(x)$ between the empirical and theoretical cumulative distribution functions.
- Compute the statistic: $D = \sup | F_n(x) - F(x) |$ and $\lambda = D\sqrt{n}$.
- Read the critical Kolmogorov value λ_α from the tables and compare it with the empirical value λ. If $\lambda \geq \lambda_\alpha$,the hypothesis H_0 should be rejected.

Kolmogorov-Smironov test

Model: Two general populations have distributions with continuous cumulative distribution functions $F_1(x)$ and $F_2(x)$. Two large independent samples with sizes n_1 and n_2 are drawn from each population. Based on the results of these samples, the hypothesis that both samples come from the same population, i.e., $H_0: F_1(x) = F_2(x)$, should be tested.

- Group the sample results into narrow intervals with the same endpoints x_j.
- For each x_j, calculate the empirical cumulative distribution functions $F_{n_1}(x) = \frac{n_{1,sk}}{n_1}$ and $F_{n_2}(x) = \frac{n_{2,sk}}{n_2}$ ($n_{1,sk}$, $n_{2,sk}$ are the cumulative frequencies up to x_j in both samples).
- Calculate the value of the statistic: $D^* = \sup | F_{n_1}(x) - F_{n_2}(x) |$ and $\lambda = D^*\sqrt{n}$, where $n = \frac{n_1 n_2}{n_1 + n_2}$.
- Read the Kolmogorov critical value λ_α from tables and compare it with the empirical value λ. If $\lambda \geq \lambda_\alpha$, the hypothesis H_0 should be rejected.

Signs Test

Model: The data are two general populations with continuous cumulative distribution functions $F_1(x)$ and $F_2(x)$ .From these populations, an equal number of pairs of corresponding elements (n) is drawn independently. Based on the results of these samples, the hypothesis that both samples come from the same population, i.e., H_0: $F_1(x) = F_2(x)$,should be tested.

- Examine the sign of the difference in pairs of values from these samples and find the number of signs, whichever is fewer; denote this number as r.
- From the sign distribution table, read for significance level α the number of result pairs r_α and compare it with the obtained number of signs r.
- If $r \leq r_\alpha$,reject H_0.

Example

To determine whether vocational training increases employee productivity, a sample of $n = 14$ employees was randomly selected from a certain plant, and their average work productivity before and after the training was examined. The results (as the number of pieces produced per hour) were obtained:

Before	52	220	125	84	150	92	94	125	78	265	187	113	63	146
After	68	242	120	107	159	80	115	162	90	241	197	101	85	180

Using the sign test at the significance level $\alpha = 0.1$, verify the hypothesis that work productivity before and after the training is the same.

Example

We want to check whether a certain drug lowers blood pressure in patients with hypertension. We randomly selected 20 patients and measured their blood pressure before and after administration of the drug. At the significance level $\alpha = 0.05$, using the sign test, verify the hypothesis that the blood pressure of the patients is the same before and after administration of the drug.

Before	320	200	340	240	200	300	240	290	180	210	250	300	180	270	290	200	280	190	220	290
After	270	180	260	250	150	260	200	310	150	220	270	260	200	240	250	160	210	160	170	220

Distribution of the number of signs

n	α				n	α				n	α				n	α			
	0,01	0,05	0,1	0,25		0,01	0,05	0,1	0,25		0,01	0,05	0,1	0,25		0,01	0,05	0,1	0,25
1					24	5	6	7	8	47	14	16	17	19	69	23	25	27	29
2					25	5	7	7	9	48	14	16	17	19	70	23	26	27	29
3				0	26	6	7	8	9	49	15	17	18	19	71	24	26	28	30
4				0	27	6	7	8	10	50	15	17	18	20	72	24	27	28	30
5			0	0	28	6	8	9	10	51	15	18	19	20	73	25	27	28	31
6		0	0	1	29	7	8	9	10	52	16	18	19	21	74	25	28	29	31
7		0	0	1	30	7	9	10	11	53	16	18	20	21	75	25	28	29	32
8	0	0	1	1	31	7	9	10	11	54	17	19	20	22	76	26	28	30	32
9	0	1	1	2	32	8	9	10	12	55	17	19	20	22	77	26	29	30	32
10	0	1	1	2	33	8	10	11	12	56	17	20	21	23	78	27	29	31	33
11	0	1	2	3	34	9	10	11	13	57	18	20	21	23	79	27	30	31	33
12	1	2	2	3	35	9	11	12	13	58	18	21	22	24	80	28	30	32	34
13	1	2	3	3	36	9	11	12	14	59	19	21	22	24	81	28	31	32	34
14	1	2	3	4	37	10	12	13	14	60	19	21	23	25	82	28	31	33	35
15	2	3	3	4	38	10	12	13	14	61	20	22	23	25	83	29	32	33	35
16	2	3	3	5	39	11	12	13	15	62	20	22	24	25	84	29	32	33	36
17	2	4	4	5	40	11	13	14	15	63	20	23	24	26	85	30	32	34	36
18	3	4	5	6	41	11	13	14	16	64	21	23	24	26	86	30	33	34	37
19	3	4	5	6	42	12	14	15	16	65	21	24	25	27	87	31	33	35	37
20	3	5	5	6	43	12	14	15	17	66	22	24	25	27	88	31	34	35	38
21	4	5	6	7	44	13	15	16	17	67	22	25	26	28	89	31	34	36	38
22	4	5	6	7	45	13	15	16	18	68	22	25	26	28	90	32	35	36	39
23	4	6	7	8	46	13	15	16	18										

Series tests

Model I: Given is a general population with any distribution. A sample of n elements is taken from the population. Check the hypothesis that this is a random sample.

- From the sequence of results of the sample, sorted by the order of drawing, calculate the median.
- Assign to each sample result x_i the symbol a if $x_i < Me$ or b if $x_i > Me$) results equal to the median can be omitted).
- Obtain a sequence consisting of symbols a and b (e.g., aabbbbababb) with a certain number of runs k (here $k = 6$ runs) and numbers of a and b elements n_1 and n_2.
- From the runs distribution tables, read two critical values k_1 and k_2) k_1 for $\frac{\alpha}{2}$,k_2 for $1 - \frac{\alpha}{2}$.(
- Compare k with k_1 and k_2 .If either inequality $k \leq k_1$ or $k \geq k_2$ is satisfied, reject H_0.

Example

- A certain machine turns rolls of a specific diameter. 16 pieces were successively sampled for technical inspection and the following diameter measurement results were obtained (in mm): 8.8; 9.2; 10.1; 10.0; 9.7; 10.6; 11.8; 10.5; 12.1; 11.5; 9.9; 12.6; 12.4; 12.8; 13.0; 12.7. At the significance level $\alpha = 0.1$, verify the hypothesis that the sample selection was random.
- In order to estimate the average number of residents of blocks of flats on a certain street, 17 blocks of flats were selected for the sample and the following numbers of residents were obtained: 143, 136, 140, 132, 120, 115, 108, 102, 105, 95, 90, 86, 84, 79, 75, 71, 67. At the significance level of 0.1, verify the hypothesis that the selection of blocks for the sample was random.

Series Distribution Tables

n_2 \ n_1	$\alpha=0,05$ 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	$\alpha=0,95$ 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	n_1 / n_2
2		4	2
3		5 6	3
4	2	5 6 7	4
5	2 2 3	5 7 8 8	5
6	2 3 3 3	5 7 8 9 10	6
7	2 3 3 4 4	5 7 8 9 10 11	7
8	2 2 3 3 4 4 5	5 7 9 10 11 12 12	8
9	2 2 3 4 4 5 5 6	5 7 9 10 11 12 13 13	9
10	2 3 3 4 5 5 6 6 6	5 7 9 10 11 12 13 14 15	10
11	2 3 3 4 5 5 6 6 7 7	5 7 9 11 12 13 14 14 15 16	11
12	2 3 4 4 5 6 6 7 7 8 8	5 7 9 11 12 13 14 15 16 16 17	12
13	2 3 4 4 5 6 6 7 8 8 9 9	5 7 9 11 12 13 14 15 16 17 17 18	13
14	2 3 4 5 5 6 7 7 8 8 9 9 10	5 7 9 11 12 13 15 16 16 17 18 19 19	14
15	2 3 4 5 6 6 7 8 8 9 9 10 10 11	5 7 9 11 13 14 15 16 17 18 18 19 20 20	15
16	2 3 4 5 6 6 7 8 8 9 10 10 11 11 11	5 7 9 11 13 14 15 16 17 18 19 20 20 21 22	16
17	2 3 4 5 6 7 7 8 9 9 10 10 11 11 12 12	5 7 9 11 13 14 15 16 17 18 19 20 21 21 22 23	17
18	2 3 4 5 6 7 8 8 9 10 10 11 11 12 12 13 13	5 7 9 11 13 14 15 17 18 19 20 20 21 22 23 23 24	18
19	2 3 4 5 6 7 8 8 9 10 10 11 12 12 13 13 14 14	5 7 9 11 13 14 15 17 18 19 20 21 22 22 23 24 24 25	19
20	2 3 4 5 6 7 8 9 9 10 11 11 12 12 13 13 14 14 15	5 7 9 11 13 14 16 17 18 19 20 21 22 23 24 24 25 26 26	20

Series tests

Model II: There are two general populations with arbitrary distributions of the studied characteristic. Two samples with sizes n_1 and n_2 are drawn from the population. You should test the hypothesis that the distributions of both samples do not differ (H_0 :both samples come from one population).

- Arrange the results of both samples into one sequence of ascending values.
- Mark elements from one sample as a, from the other as b.
- Obtain a sequence made up of symbols a and b (e.g., aabbbbababb) with a certain number of runs k (here $k = 6$ runs) and numbers of a and b elements (n_1 and n_2.(
- From the runs distribution tables, read the critical value k_α.
- Compare k with k_α) left-sided critical region). If the inequality $k \leq k_\alpha$ holds, reject H_0.

Example

- A sample of 6 children from two classes was drawn. For children from class A, the intelligence test results were: 110, 112, 115, 98, 130, 123. For children from class B: 88, 135, 140, 138, 95, 125. Using the runs test, verify the hypothesis that both samples come from a population with a specific distribution of intelligence quotient ($\alpha = 0.05$).

- A quiz was conducted in two groups of students. For sample A, the results were: 4, 3, 5, 2, 4, 6, 4, 5; for sample B, the results were: 1, 10, 8, 9, 9, 10, 8, 7. At the significance level of 0.05, use the runs test to verify the hypothesis that the distribution of results is the same for both groups.

Sources:

- Jerzy Greń: Mathematical Statistics. Models and Problems. Fourth revised edition. State Scientific Publishers. Warsaw, 1974.
- Kot, S. M., Jakubowski, J., Sokołowski, A.: Statistics. Second revised edition. Difin Publishing. Warsaw, 2011.
- Stanisz, A.: An Accessible Course in Statistics with Application of STATISTICA PL Using Examples from Medicine. Volume 1. Basic Statistics. StatSoft Polska Publishing. Kraków, 2006.
- Tables of the cumulative distribution of the normal distribution
- Tables of the Student's t distribution
- Tables of the χ^2 distribution