Chi-square & Non–parametric tests

E. Papageorgiou, G. Katsouleas

University of West Attica

June 19, 2024

1. Chi square tests
   - Goodness of fit test
   - Kolmogorov–Smirnov test
   - χ^2 test of independence

2. Non–parametric tests
   - Sign Test
   - Wilkoxon signed–rank Test
   - Kruskal–Wallis Test

Chi square tests

Goodness of fit test

- A test to determine if some population has a specified theoretical distribution.

$$H_0 : \text{ The population has a specified theoretical distribution vs.}$$
$$H_1 : \text{ The population does NOT have the specified theoretical distribution.}$$

- The test is based on how good is the obtained fit between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.
- For example, we consider the tossing of a die. We hypothesize that the die is honest.

$$H_0 : f(x) = \frac{1}{6}, x = 1, \ldots, 6,$$
$$H_1 : \text{NOT } H_0.$$

Assume that the die is tossed 120 times and each outcome is recorded as follows:

x	1	2	3	4	5	6
Observed	18	22	30	21	17	12
Expected	20	20	20	20	20	20

- Test at a level of significance of 5% whether the die is unbiased.

Goodness of fit test (2)

- Observed frequencies: O_i
- Compute the expected frequencies under the null hypothesis:

$$E_i = np_i,$$

where

$$\begin{aligned} n &: \text{ sample size,} \\ p_i &: \text{ probability of value } x_i \ (i = 1, \ldots, k). \end{aligned}$$

- Compute the quantity

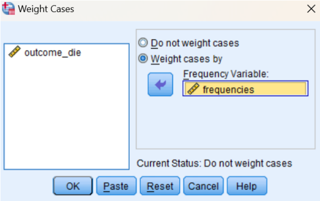
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim X^2_{k-1}, \text{ under } H_0.$$

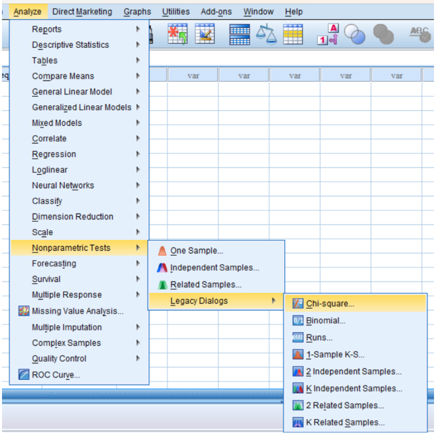
- If $\chi^2 > X_{k-1;1-\alpha}$, then H_0 is rejected at α% significance level.
- Use this test only if
  - (a) No more than 1/5 of the expected values are < 5.
  - (b) No expected value is < 1.
- $\chi^2 = \frac{(18-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(21-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(12-20)^2}{20} = 9.1$

$X_{k-1;1-\alpha} = X_{5;0.95} = 11.07 \longrightarrow$ fail to reject H_0.

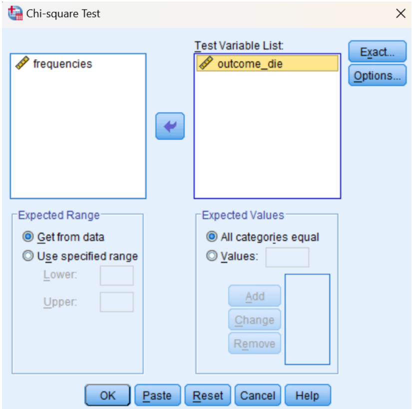
Goodness of fit test: Example (SPSS)

outcome_die	frequencies
1,00	18,00
2,00	22,00
3,00	30,00
4,00	21,00
5,00	17,00
6,00	12,00





Goodness of fit test: Example (SPSS), cont'd



outcome_die

	Observed N	Expected N	Residual
1,00	18	20,0	-2,0
2,00	22	20,0	2,0
3,00	30	20,0	10,0
4,00	21	20,0	1,0
5,00	17	20,0	-3,0
6,00	12	20,0	-8,0
Total	120		

Test Statistics

	outcome_die
Chi-Square	9,100[a]
df	5
Asymp. Sig.	,105

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 20,0.

Goodness of fit test: Continuous variable case

- When the X^2 test is applied to continuous variables, it is influenced by the grouping of the data. Hence, the X^2 goodness of fit test is preferred when we have categorical variables with finite state space.
  In SPSS, we cannot apply the X^2 goodness of fit test to continuous variables. In cases where continuous data is available, the Kolmogorov - Smirnov test is preferable. This test is based on the empirical cumulative distribution function.
- The X^2 test can be applied even when the parameters of the population distribution are unknown. In this case the degrees of freedom of the statistical test are reduced according to the number g of parameters under estimation.
  The unknown parameters of the distribution are estimated by the observations and in this case the test statistic has the following form:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim X^2_{k-g-1}, \text{ under } H_0$$

and H_0 is rejected when $\chi^2 > X_{k-g-1;1-\alpha}$.

Goodness of fit test: Example

- Diastolic blood-pressure measurements were collected at home in a community-wide screening program of 14.736 adults ages 3069 in East Boston, Massachusetts, as part of a nationwide study to detect and treat hypertensive people.
- We would like to assume these measurements came from an underlying normal distribution because standard methods of statistical inference could then be applied on these data.

Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	77.9	≥80, <90	4604	4478.5
≥50, <60	330	547.1	≥90, <100	2119	2431.1
≥60, <70	2132	2126.7	≥100, <110	659	684.1
≥70, <80	4584	4283.3	≥110	251	107.2
			Total	14,736	14,736

Goodness of fit test: Example

- Diastolic blood-pressure measurements were collected at home in a community-wide screening program of 14.736 adults ages 3069 in East Boston, Massachusetts, as part of a nationwide study to detect and treat hypertensive people.
- We would like to assume these measurements came from an underlying normal distribution because standard methods of statistical inference could then be applied on these data.

Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	77.9	≥80, <90	4604	4478.5
≥50, <60	330	547.1	≥90, <100	2119	2431.1
≥60, <70	2132	2126.7	≥100, <110	659	684.1
≥70, <80	4584	4283.3	≥110	251	107.2
			Total	14,736	14,736

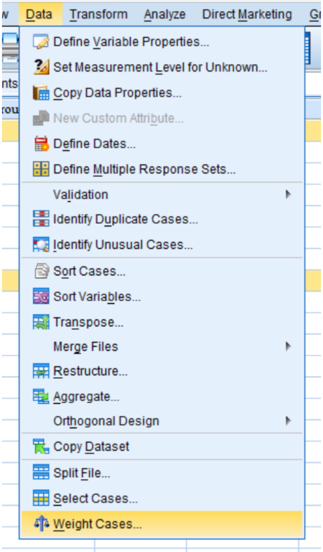
- Assume the mean and standard deviation of this hypothetical normal distribution are given by the sample mean and standard deviation, respectively ($\overline{x} = 80.68$, $s = 12.00$).
- The expected frequency within a group interval from a to b would then be given by:

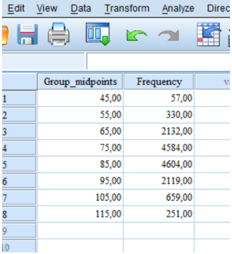
$$14.736 \left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \right],$$

where $\Phi(x) = P(X < x)$.

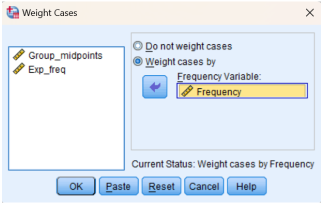
- The statistic $\chi^2 = \frac{(57-77.9)^2}{77.9} + \frac{(330-547.1)^2}{547.1} + \cdots + \frac{(251-107.2)^2}{107.2} = 350.2 \sim X^2_{k-g-1}$ under H_0, where $k = 8$ the number of groups and $g = 2$ the number of estimated parameters (internally specified model).

Goodness of fit test: Example (SPSS)

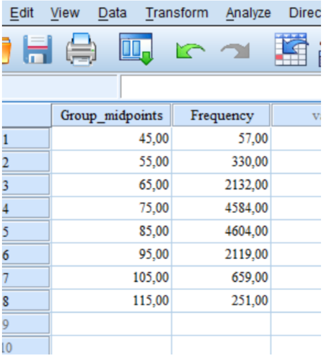


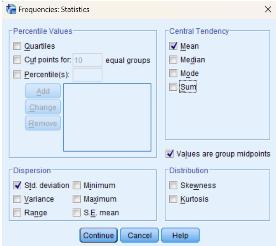


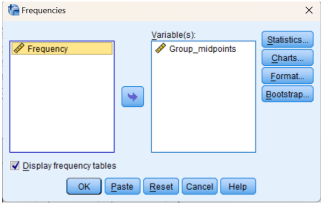
	Group_midpoints	Frequency
1	45,00	57,00
2	55,00	330,00
3	65,00	2132,00
4	75,00	4584,00
5	85,00	4604,00
6	95,00	2119,00
7	105,00	659,00
8	115,00	251,00
9		
10		



Goodness of fit test: Example (SPSS), cont'd







Statistics

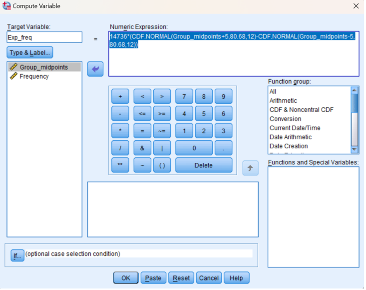
Group_midpoints

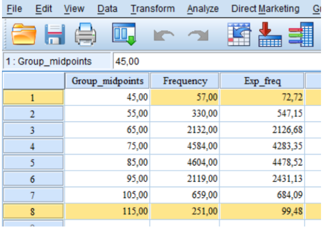
N	Valid	14736
	Missing	0
Mean		81,0125
Std. Deviation		12,12561

Group_midpoints

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	45,00	57	,4	,4	,4
	55,00	330	2,2	2,2	2,6
	65,00	2132	14,5	14,5	17,1
	75,00	4584	31,1	31,1	48,2
	85,00	4604	31,2	31,2	79,4
	95,00	2119	14,4	14,4	93,8
	105,00	659	4,5	4,5	98,3
	115,00	251	1,7	1,7	100,0
	Total	14736	100,0	100,0	

Goodness of fit test: Example (SPSS), cont'd





	Group_midpoints	Frequency	Exp_freq
1	45,00	57,00	72,72
2	55,00	330,00	547,15
3	65,00	2132,00	2126,68
4	75,00	4584,00	4283,35
5	85,00	4604,00	4478,52
6	95,00	2119,00	2431,13
7	105,00	659,00	684,09
8	115,00	251,00	99,48

- The discrepancy in the outer bins is due to the fact the original bins were not bounded by specific bounds (need to correct by hand).

Kolmogorov–Smirnov test

Empirical cumulative distribution function.

- Let $x_1, x_2, \ldots, x_n$ be a random sample.

$$F_n(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} I(x_i \leq x),$$

where $I(x_i \leq x)$ denotes the number of incidences of observations with $x_i \leq x$.

- If the sample is derived from the assumed distribution then the empirical cumulative distribution function should not differ significantly from the theoretical cdf.
- It holds

$$P\left(\lim_{n \longrightarrow \infty} |F_n(x) - F(x)| = 0\right) = 1, \text{ for all } x.$$

- The Kolmogorov - Smirnov test is based on the observed differences of $F_n(x)$ between and the theoretical cdf $F(x)$.

Kolmogorov–Smirnov test (2)

- K–S Test:

$$H_0 : F_n(x) = F(x), \text{ for all } x \in \mathbb{R} \text{ vs.}$$
$$H_1 : F_n(x) \neq F(x), \text{ for at least one } x.$$

- Let

$$D_n^+ = \sup \{F_n(x) - F(x)\}$$
$$D_n^- = \sup \{F(x) - F_n(x)\} .$$

- The test statistic is

$$D = \sup \{|F_n(x) - F(x)|\}$$
$$= \max \left\{D_n^+, D_n^-\right\}$$

- It is based on the maximum observed difference of the theoretical and the empirical cdfs.

Kolmogorov–Smirnov test (3)

- Under H_0, it holds

$$P\left(\sqrt{n}D < d\right) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2/d^2}, 0 \leq d \leq 1.$$

  This is true for any theoretical distribution assumed.
- H_0 is rejected at significance level α if $D > D_{n;\alpha}$, where $D_{n;\alpha}$ the value of the corresponding table.
- Kolmogorov - Smirnov test requires the theoretical distribution under the null hypothesis to be fully determined.
- If the theoretical distribution is not known, its parameters are estimated by the data. But, in this case, there are no tables to give the critical values, so simulations are needed to identify them.

Kolmogorov–Smirnov test: Example

Example:

- The following Table shows fasting blood glucose values ($mg/100ml$) for 36 nonobese, apparently healthy, adult males:

75	92	80	80	84	72
84	77	81	77	75	81
80	92	72	77	78	76
77	86	77	92	80	78
68	78	92	68	80	81
87	76	80	87	77	86

- Test whether the observations given above come from the normal distribution.

Kolmogorov–Smirnov test: Example

Example:

- The following Table shows fasting blood glucose values ($mg/100ml$) for 36 nonobese, apparently healthy, adult males:

75	92	80	80	84	72
84	77	81	77	75	81
80	92	72	77	78	76
77	86	77	92	80	78
68	78	92	68	80	81
87	76	80	87	77	86

- Test whether the observations given above come from the normal distribution.

Calculation of Empirical cdf $F_S(x)$:

x	Frequency	Cumulative Frequency	$F_S(x)$
68	2	2	.0556
72	2	4	.1111
75	2	6	.1667
76	2	8	.2222
77	6	14	.3889
78	3	17	.4722
80	6	23	.6389
81	3	26	.7222
84	2	28	.7778
86	2	30	.8333
87	2	32	.8889
92	4	36	1.0000
	36		

Calculation of Theoretical cdf $F_T(x)$:

x	$z = (x - 80)/6$	$F_T(x)$
68	−2.00	.0228
72	−1.33	.0918
75	−.83	.2033
76	−.67	.2514
77	−.50	.3085
78	−.33	.3707
80	.00	.5000
81	.17	.5675
84	.67	.7486
86	1.00	.8413
87	1.17	.8790
92	2.00	.9772

Kolmogorov–Smirnov test: Example (cont'd)

Calculation of Empirical cdf $F_S(x)$:

x	Frequency	Cumulative Frequency	$F_S(x)$
68	2	2	.0556
72	2	4	.1111
75	2	6	.1667
76	2	8	.2222
77	6	14	.3889
78	3	17	.4722
80	6	23	.6389
81	3	26	.7222
84	2	28	.7778
86	2	30	.8333
87	2	32	.8889
92	4	36	1.0000
	36		

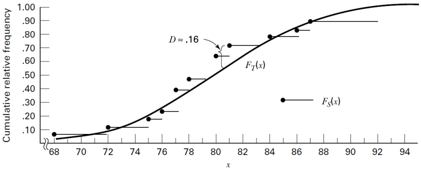
Calculation of Theoretical cdf $F_T(x)$:

x	$z = (x - 80)/6$	$F_T(x)$
68	−2.00	.0228
72	−1.33	.0918
75	−.83	.2033
76	−.67	.2514
77	−.50	.3085
78	−.33	.3707
80	.00	.5000
81	.17	.5675
84	.67	.7486
86	1.00	.8413
87	1.17	.8790
92	2.00	.9772

Calculation of $|F_S(x) - F_T(x)|$:

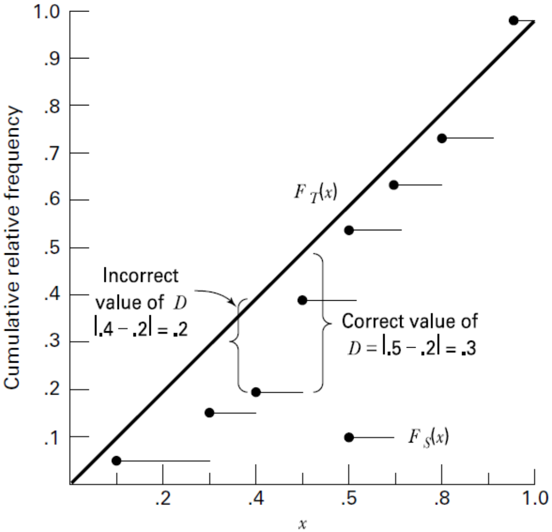
x	$F_s(x)$	$F_T(x)$	$\|F_s(x) - F_T(x)\|$
68	.0556	.0228	.0328
72	.1111	.0918	.0193
75	.1667	.2033	.0366
76	.2222	.2514	.0292
77	.3889	.3085	.0804
78	.4722	.3707	.1015
80	.6389	.5000	.1389
81	.7222	.5675	.1547
84	.7778	.7486	.0292
86	.8333	.8413	.0080
87	.8889	.8790	.0099
92	1.0000	.9772	.0228

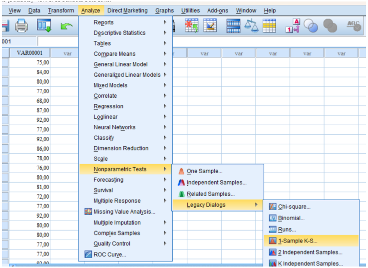
- The test statistic D may be computed algebraically, or it may be determined graphically by actually measuring the largest vertical distance between the curves of $F_S(x)$ and $F_S(x)$ on a graph:

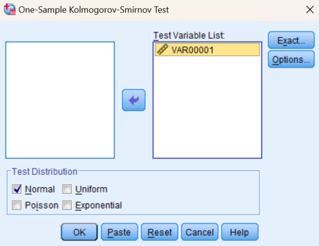


- Examination of the graphs reveals that $D \simeq 0.72 - 0.56 = 0.16$.
- *p*–value. Since we have a two-sided test, and since $0.1547 < 0.174$, we have $p > 0.20$.
- Therefore, we cannot reject H_0, i.e., the sample may have come from the specified distribution.

Kolmogorov–Smirnov test: Correct calculation of D statistic







One-Sample Kolmogorov-Smirnov Test

		VAR00001
N		36
Normal Parameters[a,b]	Mean	80,0833
	Std. Deviation	6,19850
Most Extreme Differences	Absolute	,163
	Positive	,163
	Negative	-,095
Kolmogorov-Smirnov Z		,981
Asymp. Sig. (2-tailed)		,291

a. Test distribution is Normal.

b. Calculated from data.

χ^2 test of independence

- The chi-square test can also be used to test the hypothesis of independence of two variables of classification.

$$H_0: \text{ Variables } A \text{ and } B \text{ are independent vs.}$$
$$H_1: \text{ Variables } A \text{ and } B \text{ are dependent.}$$

- Consider the two–way Table:

	B_1	B_2	$\cdots$	B_c
A_1	O_{11}	O_{12}	$\cdots$	O_{1c}
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}

- Note the following property:
  - H_0: The probability in the (i,j)–cell is equal to the product of the probabilities of being in the group–i of variable A and in group–j of variable B:

$$p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j}, \text{ for all } i,j.$$

  - H_1: Not H_0; i.e., $p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j}$, for at least one pair (i,j).

- Test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim X^2_{(r-1)(c-1)}, \text{ under } H_0.$$

- Assumptions:
  1. No more than $1/5$ of the cells have expected values < 5.
  2. No cell has an expected value < 1.

χ^2 test of independence: Example

- Suppose we want to investigate the relationship between age at first birth and development of breast cancer.

  In particular, we would like to know whether the effect of age at first birth follows a consistent trend, that is,
  (1) more protection for women whose age at first birth is < 20 than for women whose age at first birth is 25-29 and
  (2) higher risk for women whose age at first birth is ≥ 35 than for women whose age at first birth is 30-34.

  The data are presented in the following Table:

	Age at first birth					
Case–control status	<20	20–24	25–29	30–34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465

- We want to test for a relationship between age at first birth and casecontrol status.

χ^2 test of independence: Example

- Suppose we want to investigate the relationship between age at first birth and development of breast cancer.

  In particular, we would like to know whether the effect of age at first birth follows a consistent trend, that is,
  (1) more protection for women whose age at first birth is < 20 than for women whose age at first birth is 25-29 and
  (2) higher risk for women whose age at first birth is ≥ 35 than for women whose age at first birth is 30-34.

  The data are presented in the following Table:

Case–control status	Age at first birth <20	20–24	25–29	30–34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465

- We want to test for a relationship between age at first birth and casecontrol status.
- Compute the expected table for these data:

Case–control status	Age at first birth <20	20–24	25–29	30–34	≥35	Total
Case	416.6	1348.3	933.6	371.9	149.7	3220
Control	1325.4	4289.7	2970.4	1183.1	476.3	10,245
Total	1742	5638	3904	1555	626	13,465

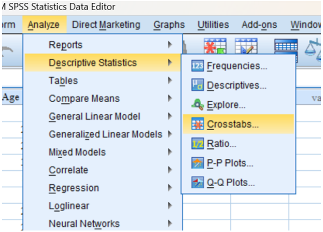
- Test statistic:

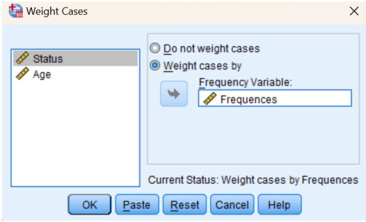
$$\chi^2 = \frac{(320 - 416.6)^2}{416.6} + \frac{(1206 - 1384.3)^2}{1384.3} + \cdots + \frac{(406 - 476.3)^2}{476.3} = 130.3 > 18.47 = \chi^2_{4;0.001}.$$

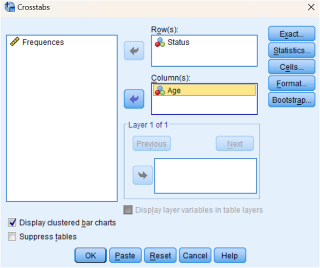
- H_0 is rejected.

χ^2 test of independence: Example (SPSS)

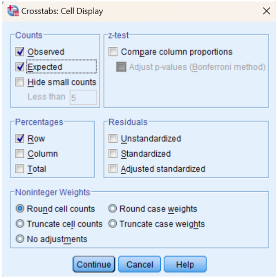
Status	Age	Frequences
Case	<20	320
Case	20-24	1206
Case	25-29	1011
Case	30-34	463
Case	>=35	220
Control	<20	1422
Control	20-24	4432
Control	25-29	2893
Control	30-34	1092
Control	>=35	406

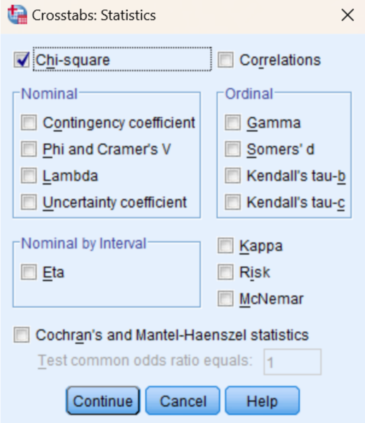






χ^2 test of independence: Example (SPSS), cont'd





χ^2 test of independence: Example (SPSS), cont'd

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Case-control status * Age at first birth	13465	100,0%	0	0,0%	13465	100,0%

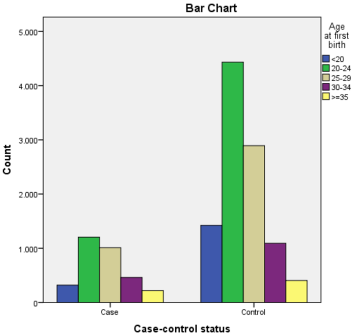
Case-control status * Age at first birth Crosstabulation

			Age at first birth					
			<20	20-24	25-29	30-34	>=35	Total
Case-control status	Case	Count	320	1206	1011	463	220	3220
		Expected Count	416,6	1348,3	933,6	371,9	149,7	3220,0
		% within Case-control status	9,9%	37,5%	31,4%	14,4%	6,8%	100,0%
	Control	Count	1422	4432	2893	1092	406	10245
		Expected Count	1325,4	4289,7	2970,4	1183,1	476,3	10245,0
		% within Case-control status	13,9%	43,3%	28,2%	10,7%	4,0%	100,0%
Total		Count	1742	5638	3904	1555	626	13465
		Expected Count	1742,0	5638,0	3904,0	1555,0	626,0	13465,0
		% within Case-control status	12,9%	41,9%	29,0%	11,5%	4,6%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	130,338[a]	4	,000
Likelihood Ratio	127,385	4	,000
Linear-by-Linear Association	129,002	1	,000
N of Valid Cases	13465		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 149,70.



Non–parametric tests

The Sign Test: Exact method ($n < 20$)

- Test the equality of the medians of two continuous dependent random variables X and Y (paired samples).

$$H_0 : \delta_X - \delta_Y = 0 \text{ vs.}$$
$$H_1 : \delta_X - \delta_Y \neq 0.$$

- Equivalently, this test can be rephrased as

$$H_0 : p = 0.5 \text{ vs.}$$
$$H_1 : p \neq 0.5.$$

where $p = P(X > Y)$.
- Test statistic: Let

$$R = \sum_{i=1}^{n} I(x_i > y_i).$$

- Under H_0, the random variable R follows the binomial distribution with parameters $(n, 0.5)$; i.e. $R \sim \mathscr{B}(n, 0.5)$.
Important: In computations, reduce sample size to exclude ties (if any).
- p–values: Thus, the probability $P(R > r)$ (i.e., the corresponding p–values) can be computed as follows:

$$p = \begin{cases} 2 \cdot \sum_{i=R}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n, & \text{if } R > \frac{n}{2}, \\ 2 \cdot \sum_{i=0}^{R} \binom{n}{i} \left(\frac{1}{2}\right)^n, & \text{if } R < \frac{n}{2}, \\ 1, & \text{if } R = \frac{n}{2}. \end{cases}$$

The Sign Test: Exact method example

Example.

- Suppose we wish to compare two different types of eye drops (A, B) that are intended to prevent redness in people with hay fever.
- Drug A is randomly administered to one eye and drug B to the other eye.
- The redness is noted at baseline and after 10 minutes by an observer who is unaware of which drug has been administered to which eye.
- We find that for 15 people with an equal amount of redness in each eye at baseline, after 10 minutes the drug A eye is less red than the drug B eye for 2 people ($d_i = x_i - y_i < 0$); the drug B eye is less red than the drug A eye for 8 people ($d_i > 0$); and the eyes are equally red for 5 people ($d_i = 0$).
- Assess the statistical significance of the results.

The Sign Test: Exact method example

Example.

- Suppose we wish to compare two different types of eye drops (A, B) that are intended to prevent redness in people with hay fever.
- Drug A is randomly administered to one eye and drug B to the other eye.
- The redness is noted at baseline and after 10 minutes by an observer who is unaware of which drug has been administered to which eye.
- We find that for 15 people with an equal amount of redness in each eye at baseline, after 10 minutes the drug A eye is less red than the drug B eye for 2 people ($d_i = x_i - y_i < 0$); the drug B eye is less red than the drug A eye for 8 people ($d_i > 0$); and the eyes are equally red for 5 people ($d_i = 0$).
- Assess the statistical significance of the results.

Answer.

- Because sample size $n = 10$ is small, the exact method must be used.
- Because

$$R = 8 > \frac{10}{2} = 5,$$

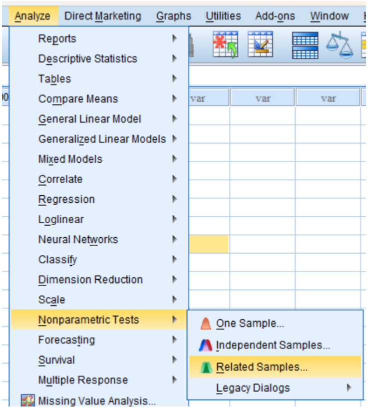
we have

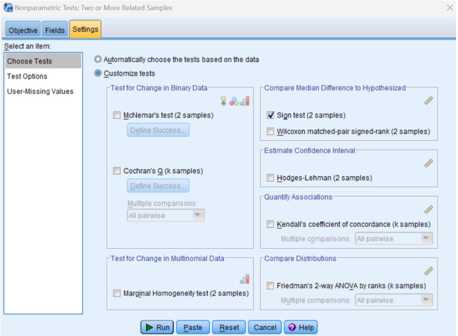
$$p = 2 \cdot \sum_{i=8}^{10} \binom{n}{i} \left(\frac{1}{2}\right)^{10} = 2 \cdot (P(C = 8) + P(C = 9) + P(C = 10)) = 2 \cdot (0.0439 + 0.0098 + 0.0010)$$
$$= 2 \cdot 0.0547 = 0.109$$

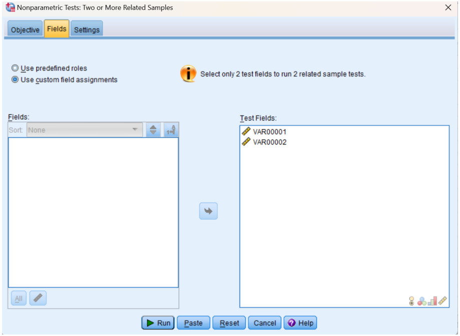
- This is not statistically significant. Thus, we accept H_0, that the two types of eye drops are equally effective in reducing redness in people with hay fever.

The Sign Test: Exact method example SPSS

	VAR00001	VAR00002
1	1,00	1,00
2	1,00	1,00
3	1,00	,00
4	1,00	,00
5	1,00	,00
6	1,00	,00
7	1,00	,00
8	1,00	,00
9	1,00	,00
10	1,00	,00
11	,00	1,00
12	,00	1,00
13	,00	1,00
14	,00	1,00
15	,00	1,00







Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between VAR00002 and VAR00001 equals 0.	Related-Samples Sign Test	,581[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1]Exact significance is displayed for this test.

The Sign Test: Normal theory method

- Test the equality of the medians of two continuous dependent random variables X and Y (paired samples).

$$H_0 : \delta_X - \delta_Y = 0 \text{ vs.}$$
$$H_1 : \delta_X - \delta_Y \neq 0.$$

- Equivalently, this test can be rephrased as

$$H_0 : p = 0.5 \text{ vs.}$$
$$H_1 : p \neq 0.5.$$

  where $p = P(X > Y)$.
- Let

$$R = \sum_{i=1}^{n} I(x_i > y_i).$$

- Under H_0, the random variable R follows the binomial distribution with parameters $(n, 0.5)$; i.e. $R \sim \mathscr{B}(n, 0.5)$.
  Thus, $E(R) = \frac{n}{2}$ and $Var(R) = \frac{n}{4}$ and the probability $P(R > r)$ (i.e., the corresponding p–values) can be computed.
- If n is large then $(n > 20)$, under H_0, $R \sim \mathscr{N}\left(\frac{n}{2}, \frac{n}{4}\right)$.
- In this case, the test statistic is

$$\frac{R - \frac{n}{2} - \frac{1}{2}}{\sqrt{\frac{n}{4}}} \sim \mathscr{N}(0, 1).$$

- p–value. Denoting $\Phi(z) = P(Z < z)$, we have

$$p = \begin{cases} 2 \cdot \left(1 - \Phi\left(\frac{R - n/2 - 1/2}{\sqrt{n/4}}\right)\right), & \text{if } R > n/2, \\ 2 \cdot \Phi\left(\frac{R - n/2 + 1/2}{\sqrt{n/4}}\right), & \text{if } R < n/2, \\ 1, & \text{if } R = n/2. \end{cases}$$

The Sign Test: Normal theory method example

Example.

- Suppose we want to compare the effectiveness of two ointments (A, B) in reducing excessive redness in people who cannot otherwise be exposed to sunlight.
- Ointment A is randomly applied to either the left or right arm, and ointment B is applied to the corresponding area on the other arm. The person is then exposed to 1 hour of sunlight, and the two arms are compared for degrees of redness.
- Suppose only the following qualitative assessments can be made:
  1. Arm A is not as red as arm B.
  2. Arm B is not as red as arm A.
  3. Both arms are equally red.
- Of 45 people tested with the condition, 22 are better off on arm A, 18 are better off on arm B, and 5 are equally well off on both arms.
- Can we decide whether this evidence is enough to conclude that ointment A is better than ointment B?

Answer.

- There are 40 untied pairs and $R = 18 < n/2 = 20$.
- Using the normal distribution, for $\alpha = 0.05$, the critical values are given by

$$c_2 = \frac{n}{2} + \frac{1}{2} + z_{\alpha/2}\sqrt{\frac{n}{4}} = \frac{40}{2} + \frac{1}{2} + 1.96 \cdot 3.162 = 26.7$$

and

$$c_1 = \frac{n}{2} - \frac{1}{2} - z_{\alpha/2}\sqrt{\frac{n}{4}} = \frac{40}{2} - \frac{1}{2} - 1.96 \cdot 3.162 = 13.3.$$

- Because $13.3 \leq R = 18 \leq 26.7$, H_0 is accepted using a two-sided test with $\alpha = 0.05$ and we conclude the ointments do not significantly differ in effectiveness.
- Also, we have $p = 2 \cdot \Phi\left(\frac{18-20+\frac{1}{2}}{\sqrt{40/4}}\right) = 2 \cdot \Phi(-0.47) = 2 \cdot 0.316 = 0.635.$

Wilkoxon signed–rank Test

$$H_0 : \delta_X - \delta_Y = 0 \text{ vs.}$$
$$H_1 : \delta_X - \delta_Y \neq 0.$$

- Wilcoxon signed-rank test step by step:
  1. Compute the ranks of the absolute differences $|x_i - y_i|$, ignoring cases where $x_i - y_i = 0$.
  2. Compute the sum of the ranks S_p that correspond to positive differences.
  3. If the sample size n is large, then

$$Z = \frac{\left| S_p - \frac{n(n+1)}{4} \right| - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{L} \frac{t_i^3 - t_i}{48}}} \sim \mathcal{N}(0,1),$$

under H_0.
Here, L denotes the number of cases for which we have equal observations and t_i the number of observations with the same rank.

Wilkoxon signed–rank Test: Example

Example.

- Previously, we assumed that the only possible assessment was that the degree of sunburn with ointment A was either better or worse than that with ointment B.
- Suppose instead that the degree of burn can be quantified on a 10-point scale, with 10 being the worst burn and 1 being no burn at all.
- We can now compute $d_i = x_i - y_i$, where x_i = degree of burn for ointment A and y_i = degree of burn for ointment B. If d_i is positive, then ointment, B is doing better than ointment A; if d_i is negative, then ointment A is doing better than ointment B.
- Difference in degree of redness between ointment A and ointment B arms after 10 minutes of exposure to sunlight:

	Negative		Positive	
$\lvert d_i \rvert$	d_i	f_i	d_i	f_i
10	−10	0	10	0
9	−9	0	9	0
8	−8	1	8	0
7	−7	3	7	0
6	−6	2	6	0
5	−5	2	5	0
4	−4	1	4	0
3	−3	5	3	2
2	−2	4	2	6
1	−1	4	1	10
		22		18

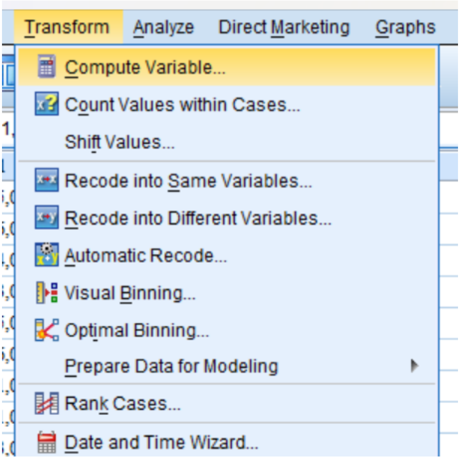
- Test whether the ointments are equally effective.

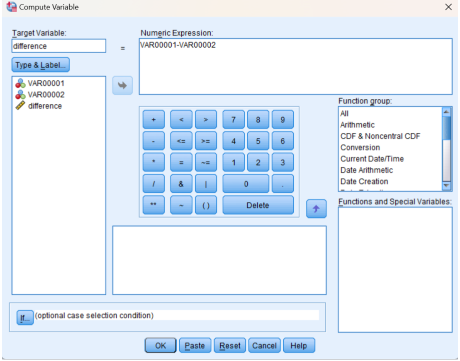
Wilkoxon signed–rank Test: Example ranks computation

$\lvert d_i \rvert$	Negative d_i	Negative f_i	Positive d_i	Positive f_i	Number of people with same absolute value	Range of ranks	Average rank
10	−10	0	10	0		—	—
9	−9	0	9	0	0	—	—
8	−8	1	8	0	1	40	40.0
7	−7	3	7	0	3	37–39	38.0
6	−6	2	6	0	2	35–36	35.5
5	−5	2	5	0	2	33–34	33.5
4	−4	1	4	0	1	32	32.0
3	−3	5	3	2	7	25–31	28.0
2	−2	4	2	6	10	15–24	19.5
1	−1	4	1	10	14	1–14	7.5
		22		18			
0	0	5					

- Arrange the differences d_i in order of absolute value.
- Count the number of differences with the same absolute value.
- Ignore the observations where $d_i = 0$, and rank the remaining observations from 1 for the observation with the lowest absolute value, up to n for the observation with the highest absolute value.
- If there is a group of several observations with the same absolute value, then find the lowest rank in the range $= 1 + R$ and the highest rank in the range $= G + R$, where $R =$ the highest rank used prior to considering this group and $G =$ the number of differences in the range of ranks for the group. Assign the average rank $=$ (lowest rank in the range $+$ highest rank in the range)/2 as the rank for each difference in the group.

Wilkoxon signed–rank Test: Example ranks computation (SPSS)





Wilkoxon signed–rank Test: Difference frequences computation (SPSS)

e | -1,00

VAR00001	VAR00002	difference
5,00	6,00	-1,00
6,00	7,00	-1,00
4,00	5,00	-1,00
3,00	4,00	-1,00
5,00	7,00	-2,00
6,00	8,00	-2,00
1,00	4,00	-3,00
1,00	4,00	-3,00
3,00	6,00	-3,00
5,00	9,00	-4,00
3,00	8,00	-5,00
4,00	9,00	-5,00
4,00	10,00	-6,00
2,00	8,00	-6,00
2,00	9,00	-7,00
1,00	8,00	-7,00
3,00	10,00	-7,00
,00	8,00	-8,00
3,00	6,00	-3,00
4,00	7,00	-3,00

Statistics

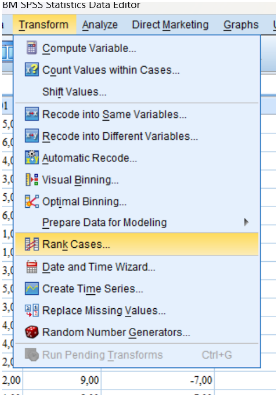
difference

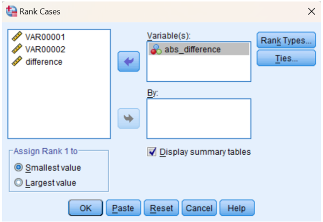
N	Valid	40
	Missing	0

difference

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-8,00	1	2,5	2,5	2,5
	-7,00	3	7,5	7,5	10,0
	-6,00	2	5,0	5,0	15,0
	-5,00	2	5,0	5,0	20,0
	-4,00	1	2,5	2,5	22,5
	-3,00	5	12,5	12,5	35,0
	-2,00	4	10,0	10,0	45,0
	-1,00	4	10,0	10,0	55,0
	1,00	10	25,0	25,0	80,0
	2,00	6	15,0	15,0	95,0
	3,00	2	5,0	5,0	100,0
	Total	40	100,0	100,0	

Wilkoxon signed–rank Test: Example ranks (SPSS)



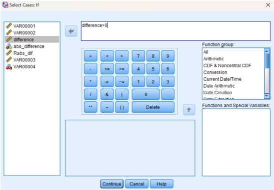


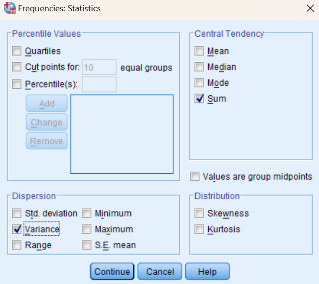
VAR00001	VAR00002	difference	abs_difference	Rabs_dif
5,00	6,00	-1,00	1,00	7,500
6,00	7,00	-1,00	1,00	7,500
4,00	5,00	-1,00	1,00	7,500
3,00	4,00	-1,00	1,00	7,500
5,00	7,00	-2,00	2,00	19,500
6,00	8,00	-2,00	2,00	19,500
1,00	4,00	-3,00	3,00	28,000
1,00	4,00	-3,00	3,00	28,000
3,00	6,00	-3,00	3,00	28,000
5,00	9,00	-4,00	4,00	32,000
3,00	8,00	-5,00	5,00	33,500
4,00	9,00	-5,00	5,00	33,500
4,00	10,00	-6,00	6,00	35,500
2,00	8,00	-6,00	6,00	35,500
2,00	9,00	-7,00	7,00	38,000
1,00	8,00	-7,00	7,00	38,000
3,00	10,00	-7,00	7,00	38,000
,00	8,00	-8,00	8,00	40,000
3,00	6,00	-3,00	3,00	28,000
4,00	7,00	-3,00	3,00	28,000

Wilkoxon signed–rank Test: Example (rank sum)

- Because the number of nonzero differences $(22 + 18 = 40) \geq 16$, the normal approximation method can be used. Compute the rank sum S_p for the people with positive d_i—that is, where ointment B performs better than ointment A, as follows:

$$S_p = 10 \cdot 7.5 + 6 \cdot 19.5 + 2 \cdot 28.0 = 75 + 117 + 56 = 248$$





	VAR00001	VAR00002	difference	abs_difference	Rabs_dif	filter_$
16	1,00	8,00	-7,00	7,00	38,000	Not Selected
17	3,00	10,00	-7,00	7,00	38,000	Not Selected
18	,00	8,00	-8,00	8,00	40,000	Not Selected
19	3,00	6,00	-3,00	3,00	28,000	Not Selected
20	4,00	7,00	-3,00	3,00	28,000	Not Selected
21	5,00	7,00	-2,00	2,00	19,500	Not Selected
22	3,00	5,00	-2,00	2,00	19,500	Not Selected
23	1,00	,00	1,00	1,00	7,500	Selected
24	3,00	2,00	1,00	1,00	7,500	Selected
25	2,00	1,00	1,00	1,00	7,500	Selected
26	2,00	1,00	1,00	1,00	7,500	Selected
27	9,00	8,00	1,00	1,00	7,500	Selected
28	8,00	7,00	1,00	1,00	7,500	Selected
29	8,00	7,00	1,00	1,00	7,500	Selected
30	7,00	6,00	1,00	1,00	7,500	Selected
31	6,00	5,00	1,00	1,00	7,500	Selected

Statistics

Rank of abs_difference

N	Valid	18
	Missing	0
Variance		58,536
Sum		248,000

Rank of abs_difference

Wilkoxon signed–rank Test: Example (rank sum)

- $S_p = 248$.
- The expected rank sum is given by:

$$E(S_p) = 40(41)/4 = 410$$

- The variance of the rank sum corrected for ties is given by:

$$Var(S_p) = \frac{40(41)(81)}{24} - \frac{(14^2 - 14) + (10^3 - 10) + (7^3 - 7) + (1^3 - 1) + (2^3 - 2) + (2^3 - 2) + (3^3 - 3) + (1^3}{48}$$

$$= 5449.75 \Rightarrow sd(S_p) = \sqrt{Var(S_p)} = \sqrt{5449.75} = 73.82.$$

- Test statistic:

$$Z = \frac{\left| S_p - \frac{n(n+1)}{4} \right| - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{L} \frac{t_i^3 - t_i}{48}}}$$

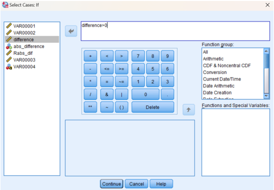
$$= \frac{|248 - 410| - \frac{1}{2}}{73.82} = 2.19.$$

- *p*–value:

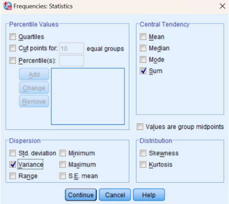
$$p = 2 \cdot [1 - \Phi(2.19)] = 2 \cdot (1 - 0.9857) = 0.029$$

- We therefore can conclude that there is a significant difference between ointments, with ointment A doing better than ointment B because the observed rank sum (248) is smaller than the expected rank sum (410).
- This conclusion differs from the conclusion based on the sign test, where no significant difference between ointments was found. This result indicates that when the information is available, it is worthwhile to consider both magnitude and direction of the difference between treatments, as the signed-rank test does, rather than just the direction of the difference, as the sign test does.

Wilkoxon signed–rank Test: Example (rank sum)



	VAR00001	VAR00002	difference	abs_difference	Rabs_dif	filter_$
16	1,00	8,00	-7,00	7,00	38,000	Not Selected
17	3,00	10,00	-7,00	7,00	38,000	Not Selected
18	,00	8,00	-8,00	8,00	40,000	Not Selected
19	3,00	6,00	-3,00	3,00	28,000	Not Selected
20	4,00	7,00	-3,00	3,00	28,000	Not Selected
21	5,00	7,00	-2,00	2,00	19,500	Not Selected
22	3,00	5,00	-2,00	2,00	19,500	Not Selected
23	1,00	,00	1,00	1,00	7,500	Selected
24	3,00	2,00	1,00	1,00	7,500	Selected
25	2,00	1,00	1,00	1,00	7,500	Selected
26	2,00	1,00	1,00	1,00	7,500	Selected
27	9,00	8,00	1,00	1,00	7,500	Selected
28	8,00	7,00	1,00	1,00	7,500	Selected
29	8,00	7,00	1,00	1,00	7,500	Selected
30	7,00	6,00	1,00	1,00	7,500	Selected
31	6,00	5,00	1,00	1,00	7,500	Selected
32	5,00	4,00	1,00	1,00	7,500	Selected



Statistics

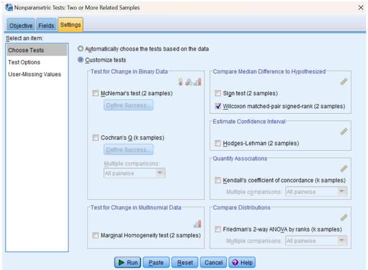
Rank of abs_difference

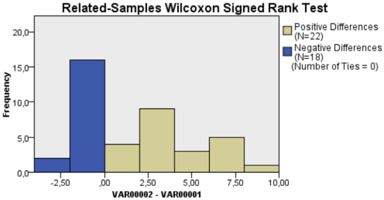
N	Valid	18
	Missing	0
Variance		58,536
Sum		248,000

Rank of abs_difference

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7,500	10	55,6	55,6	55,6
	19,500	6	33,3	33,3	88,9
	28,000	2	11,1	11,1	100,0
	Total	18	100,0	100,0	

Wilkoxon signed–rank Test (SPSS)





Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between VAR00001 and VAR00002 equals 0.	Related-Samples Wilcoxon Signed Rank Test	,028	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Total N	40
Test Statistic	572,000
Standard Error	73,822
Standardized Test Statistic	2,194
Asymptotic Sig. (2-sided test)	,028

Kruskal–Wallis Test

- Nonparametric alternative to the one-way ANOVA. In some instances we want to compare means among more than two samples, but either the underlying distribution is far from being normal or we have ordinal data.

$$H_0 : \delta_{X_1} = \delta_{X_2} = \cdots = \delta_{X_k} \text{ vs.}$$
$$H_1 : \text{ Not } H_0.$$

- Kruskal–Wallis Test test step by step:
  1. Compute the ranks of the observations assuming the k samples as a single sample.
  2. Compute the sums of the ranks R_i, the number of observations n_i in each group and the quantity $T_i = t_i^3 - t_i$ in cases where there are observations with the same rank where t_i is the number of observations with the same rank.
  3. If all sample sizes n_i $(i = 1, \ldots, k)$ are large (i.e., $n_i > 5$), then

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \sim X^2_{k-1},$$

     under H_0.
  4. In the case that they are observations with the same rank, the quantity above is corrected as follows:

$$H' = \frac{H}{1 - \sum_{i=1}^{L} \frac{t_i^2(t_i-1)}{N^2(N-1)}},$$

     where L denotes the number of cases for which we have equal observations and t_i the number of observations with the same rank.
- Remark. This test procedure should be used only if minimum $n_i \geq 5$ (i.e., if the smallest sample size for an individual group is at least 5). Else, combine samples.

Kruskal–Wallis Test: Example

Example.

- A study was conducted to compare the anti-inflammatory effects of four different drugs in albino rabbits after administration of arachidonic acid.
- Six rabbits were studied in each group. Different rabbits were used in each of the four groups. For each animal in a group, one of the four drugs was administered to one eye and a saline solution was administered to the other eye.
- Ten minutes later arachidonic acid (sodium arachidonate) was administered to both eyes. Both eyes were evaluated every 15 minutes thereafter for lid closure.
- At each assessment the lids of both eyes were examined and a lid-closure score from 0 to 3 was determined, where 0 = eye completely open, 3 = eye completely closed, and 1, 2 = intermediate states.
- The measure of effectiveness (x) is the change in lid-closure score (from baseline to follow-up) in the treated eye minus the change in lid-closure score in the saline eye.
- A high value for x is indicative of an effective drug. The results, after 15 minutes of follow-up, are presented in the following Table:

Rabbit Number	Indomethacin		Aspirin		Piroxicam		BW755C	
	Score[a]	Rank	Score	Rank	Score	Rank	Score	Rank
1	+ 2	13.5	+ 1	9.0	+ 3	20.0	+ 1	9.0
2	+ 3	20.0	+ 3	20.0	+ 1	9.0	0	4.0
3	+ 3	20.0	+ 1	9.0	+ 2	13.5	0	4.0
4	+ 3	20.0	+ 2	13.5	+ 1	9.0	0	4.0
5	+ 3	20.0	+ 2	13.5	+ 3	20.0	0	4.0
6	0	4.0	+ 3	20.0	+ 3	20.0	− 1	1.0

[a](Lid-closure score at baseline − lid-closure score at 15 minutes)$_{\text{drug eye}}$ − (lid-closure score at baseline − lid-closure score at 15 minutes)$_{\text{saline eye}}$

- Pool the observations over all samples, thus constructing a combined sample of size $N = \sum_{i=1}^{k} n_i$.
- Assign ranks to the individual observations, using the average rank in the case of tied observations

Lid-closure score	Frequency	Range of ranks	Average rank
−1	1	1	1.0
0	5	2–6	4.0
+1	5	7–11	9.0
+2	4	12–15	13.5
+3	9	16–24	20.0

- Compute the rank sum R_i for each of the k samples:

Rabbit Number	Indomethacin		Aspirin		Piroxicam		BW755C	
	Score[a]	Rank	Score	Rank	Score	Rank	Score	Rank
1	+ 2	13.5	+ 1	9.0	+ 3	20.0	+ 1	9.0
2	+ 3	20.0	+ 3	20.0	+ 1	9.0	0	4.0
3	+ 3	20.0	+ 1	9.0	+ 2	13.5	0	4.0
4	+ 3	20.0	+ 2	13.5	+ 1	9.0	0	4.0
5	+ 3	20.0	+ 2	13.5	+ 3	20.0	0	4.0
6	0	4.0	+ 3	20.0	+ 3	20.0	− 1	1.0

[a](Lid-closure score at baseline − lid-closure score at 15 minutes)$_{\text{drug eye}}$ − (lid-closure score at baseline − lid-closure score at 15 minutes)$_{\text{saline eye}}$

$$R_1 = 13.5 + 4 \cdot 20.0 + 4.0 = 97.5$$
$$R_2 = 2 \cdot 9.0 + 2 \cdot 20.0 + 2 \cdot 13.5 = 85.0$$
$$R_3 = 2 \cdot 9.0 + 4 \cdot 20.0 + 1 \cdot 13.5 = 91.5$$
$$R_4 = 4 \cdot 4.0 + 9.0 + 1.0 = 26.0$$

- Compute the rank sum R_i for each of the k samples:

$$\begin{aligned} R_1 &= 13.5 + 4 \cdot 20.0 + 4.0 = 97.5 \\ R_2 &= 2 \cdot 9.0 + 2 \cdot 20.0 + 2 \cdot 13.5 = 85.0 \\ R_3 &= 2 \cdot 9.0 + 4 \cdot 20.0 + 1 \cdot 13.5 = 91.5 \\ R_4 &= 4 \cdot 4.0 + 9.0 + 1.0 = 26.0 \end{aligned}$$

- Because there are ties, compute the Kruskal-Wallis test statistic H as follows:

$$H = \frac{\frac{12}{24 \times 25} \times \left(\frac{97.5^2}{6} + \frac{85.0^2}{6} + \frac{91.5^2}{6} + \frac{26.0^2}{6} \right) - 3(25)}{1 - \frac{(5^3 - 5) + (5^3 - 5) + (4^3 - 4) + (9^3 - 9)}{24^3 - 24}}$$

$$= \frac{0.020 \times 4296.583 - 75}{1 - \frac{1020}{13{,}800}} = \frac{10.932}{0.926} = 11.804$$

- To assess statistical significance, compare H with a chi-square distribution with $k - 1 = 4 - 1 = 3$ df.
- Since $\chi^2_{3;0.01} = 11.34$, $\chi^2_{3;0.005} = 12.84$. Because $11.34 < H < 12.84$, it follows that $0.005 < p < 0.01$.
- Thus, there is a significant difference in the anti-inflammatory potency of the four drugs.

Comparison of Specific Groups Under the Kruskal-Wallis Test

- Dunn procedure. To compare the i–th and j–th treatment groups under the Kruskal-Wallis test, use the following test statistic:

$$z = \frac{\overline{R}_i - \overline{R}_j}{\sqrt{\frac{N(N+1)}{12} \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim \mathcal{N}(0, 1)$$

  where $\overline{R}_i$ denotes the average rank in the i–th sample.

- For a two–sided level α test, compare test statistic:

$$\text{If } |z| > z_{\alpha^*}, \text{ then reject } H_0,$$
$$\text{If } |z| < z_{\alpha^*}, \text{ then reject } H_0,$$

  where

$$\alpha^* = \frac{\alpha}{k(k-1)}.$$

Comparison of Specific Groups Under the Kruskal-Wallis Test

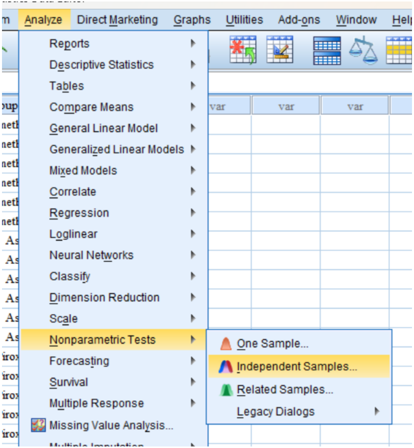
- Application. For the previous Example, we have

$$\overline{R}_1 = \frac{97.5}{6} = 16.25, \quad \overline{R}_2 = \frac{85.0}{6} = 14.17,$$
$$\overline{R}_3 = \frac{91.5}{6} = 15.25, \quad \overline{R}_4 = \frac{26.0}{6} = 4.33.$$

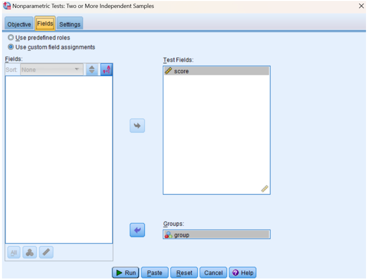
- Therefore, the following test statistics are used to compare each pair of groups:
  - Groups 1 and 2: $z_{12} = \frac{16.25-14.17}{\sqrt{\frac{24\cdot 25}{12}\cdot\left(\frac{1}{6}+\frac{1}{6}\right)}} = \frac{2.08}{4.082} = 0.51,$
  - Groups 1 and 3: $z_{13} = \frac{16.25-15.25}{4.082} = \frac{1.0}{4.082} = 0.24,$
  - Groups 1 and 4: $z_{14} = \frac{16.25-4.33}{4.082} = \frac{11.92}{4.082} = 2.92,$
  - Groups 2 and 3: $z_{23} = \frac{14.17-15.25}{4.082} = \frac{-1.08}{4.082} = -0.27,$
  - Groups 2 and 4: $z_{24} = \frac{14.17-4.33}{4.082} = \frac{9.83}{4.082} = 2.41,$
  - Groups 3 and 4: $z_{34} = \frac{15.25-4.33}{4.082} = \frac{10.92}{4.082} = 2.67.$
- The critical value for $\alpha = 0.05$ is $\alpha^* = \frac{0.05}{4\cdot 3} = 0.0042$, whereby $z_{\alpha^*} = 2.635$.
- Because z_{14} and z_{34} are greater than the critical value, it follows that indomethacin (group 1) and piroxicam (group 3) have significantly better anti-inflammatory properties than BW755C (group 4), whereas the other treatment comparisons are not statistically significant.

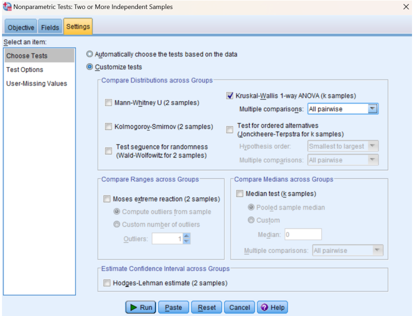
Kruskal-Wallis Test (SPSS)

	score	group
1	2	Indomethacin
2	3	Indomethacin
3	3	Indomethacin
4	3	Indomethacin
5	3	Indomethacin
6	0	Indomethacin
7	1	Aspirin
8	3	Aspirin
9	1	Aspirin
10	2	Aspirin
11	2	Aspirin
12	3	Aspirin
13	3	Piroxicam
14	1	Piroxicam
15	2	Piroxicam
16	1	Piroxicam
17	3	Piroxicam
18	3	Piroxicam
19	1	BW755C
20	0	BW755C



Kruskal-Wallis Test (SPSS), cont'd





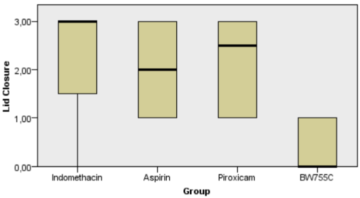
Kruskal-Wallis Test (SPSS), cont'd

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Lid Closure is the same across categories of Group.	Independent-Samples Kruskal-Wallis Test	,015	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

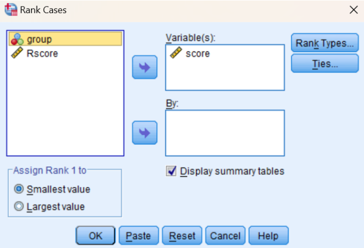
Independent-Samples Kruskal-Wallis Test



Total N	24
Test Statistic	10,510
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	,015

1. The test statistic is adjusted for ties.

Kruskal-Wallis Test (SPSS), cont'd



	score	group	Rscore	var
1	2	Indomethacin	13,500	
2	3	Indomethacin	20,000	
3	3	Indomethacin	20,000	
4	3	Indomethacin	20,000	
5	3	Indomethacin	20,000	
6	0	Indomethacin	3,000	
7	1	Aspirin	8,500	
8	3	Aspirin	20,000	
9	1	Aspirin	8,500	
10	2	Aspirin	13,500	
11	2	Aspirin	13,500	
12	3	Aspirin	20,000	
13	3	Piroxicam	20,000	
14	1	Piroxicam	8,500	
15	2	Piroxicam	13,500	
16	1	Piroxicam	8,500	
17	3	Piroxicam	20,000	
18	3	Piroxicam	20,000	
19	1	BW755C	8,500	
20	0	BW755C	3,000	
21	0	BW755C	3,000	