NON-PARAMETRIC TESTS WITH EFFECT SIZES

H.S. Steyn
Statistical Consultation Services
North-West University
2020

Preface

In my work as statistical consultant, I began to realise that there also exists a need for the following up of non-parametric statistical tests with effect sizes. In my *Manual for the determination of effect size indices and practical significance* (Steyn, 2012), I tried to provide each statistical test with at least one effect-size index, but the focus was mainly on parametric statistical methods. To apply one of these effect sizes (meant to supplement parametric statistics) in non-parametric statistics, would not make sense and be incorrect.

When contributing in writing an article in the field of physiotherapy (Pautz et. al., 2018), I became more aware of researchers' need to be guided in using correct effect sizes after applying non-parametric methods in statistical analyses.

I thus decided to write this manual for the clients of the Statistical Consultation Services. To use the manual without referring back to my previous one, I have included some topics that had already been discussed in the latter in order to make this one more comprehensive. These topics are dealt with in subsections 2.1 and 2.2; they appear in Chapter 5 (subsections 5.4 and 5.5) of my previous manual, but have been shortened and slightly altered.

The calculations of non-parametric tests' statistics can all be done by using the statistical packages SAS (SAS Institute Inc. 2020), SPSS (IBM SPSS Statistics, 2020) and STATISTICA (TIBCO STATISTICA, 2020). As the calculations can also be done manually and certain principles can thus be illustrated, I always provide the methods of those calculations before giving attention to the calculation of accompanying effect sizes.

In the determination of confidence intervals, calculations are not always possible without the use of the SAS program *CI_w* (see Steyn, 2012) and the Excel *Nonparametric Effect Size and CI Calculator* spreadsheet. This Excel spreadsheet is an adapted and extended version of Pautz et.al.'s (2018) *Supplementary Calculator*. Both the program and spreadsheet are enclosed in this manual.

1. Introduction

When working with measurements on the nominal scale (i.e., categorical measurements), non-parametric statistics is directly applicable. These tests and methods are dealt with in Section 2, together with the relevant effect-size indices.

For ordinal measurements (e.g., on a five-point Likert scale), discrete measurements (e.g., the number of people living in a house), interval scale measurements (e.g., weights and heights of people) as well as when groups are small (e.g., smaller than 30), many parametric methods cannot be used anymore, as the assumption of normality is necessary. An example is when 12 persons are randomly drawn from each of the populations of healthy and ill persons, but the measurements from the populations are not necessarily normally distributed. Suppose the blood pressure of persons is measured and they are asked to declare on a seven-point scale how regularly they come down with a cold. In this case, it is appropriate to use non-parametric methods by employing ranks based on such measurements. These cases with effect-size indices receive attention in Sections 4 and 5.

Relationships between two or more variables measured on ordinal, discrete and interval scales are not necessarily linear or the underlying joint distribution is not always multivariate normal. Therefore, it is not possible to test statistically for significant correlations. Once again, non-parametric measures of relationships (based on the ranks of these measurements) can then be used as effect-size indices. This receives attention in Section 3.

2. Categorical relationships

With nominal or categorical measurements, relationships between two such variables can be determined by using two-way frequency tables. If the variables are dichotomous (i.e., comprising only two categories, e.g., pass/fail or positive/negative), a two-by-two (2 x 2) frequency table (also called a four-fold table) is formed. If the two dichotomous variables are **independent** (e.g., gender and pass/fail), a considerable number of measures or effect sizes can be calculated from the table and this is the topic of subsection 2.1. In subsection 2.2, the relationships between two independent categorical variables are viewed where at least one of them has **more than two categories** (e.g., four language categories vs. nine provinces).

The case of two dichotomous **dependent** variables (e.g., pass/fail of first test vs. pass/fail of second test of the same subject) is subsequently dealt with in subsection 2.3. Lastly, the case of m**ore than two dependent** dichotomous variables (e.g., when four examiners assess the same students in an oral examination to judge whether they pass or fail) is the topic of subsection 2.4.

2.1 Effect sizes in two-by-two (2 x 2) frequency tables

When population or sample elements can be classified simultaneously according to two dichotomous categories, these data can be represented in a 2 x 2 frequency or contingency table (also called a four-fold table), as in Table 1 (see Steyn, 2002 and Kline, 2004a: 146):

Table 1

The 2 x 2 frequency table of *x* and *y*

	y: Category 1	*y: Category 2*	*Total*
x: Category 1	*a*	*b*	*a + b*
x: Category 2	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n*

Here, *a, b, c* and *d* are the frequencies in the four combinations of the categories of *x* and *y* and $n = a + b + c + d$ is the population or sample size.

2.1.1 Relationship between *x* and *y*

The Pearson correlation coefficient between *x* and *y* (where each one takes on two values, e.g., 1 and 2) is in terms of the frequencies in Table 1:

$$\varphi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \qquad (1)$$

the ***phi coefficient***. Therefore, this coefficient has the same characteristics as ρ_{xy} and r_{xy} in Chapter 5, subsection 5.1 (see Steyn, 2012) and can as such be used as an effect-size index. As in the case of ρ_{xy} and r_{xy}, φ can also be negative, which is the case when $bc > ad$. Because categories 1 and 2 are usually in an arbitrary sequence (e.g., the first category of *x* is men and the second is women), the frequency table could usually be

constructed in such a way that the greater frequencies appear in category 1 of both *x* and *y* and category 2 of both *x* and *y*. As a result, such a construction has a positive φ.

Remark: The chi-square statistic for a two-by-two frequency table is:

$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = n\varphi^2$.

It thus holds true that $\varphi = \sqrt{\frac{X^2}{n}}$.

With reference to the guideline values for ρ_{xy}, Cohen (1969, 1977, 1988) proposes the same values for φ, namely:

- small effect: $\varphi = 0{,}1$;
- medium effect: $\varphi = 0{,}3$;
- large effect: $\varphi = 0{,}5$.

2.1.2 Binomial effect-size display (Rosenthal et.al., 2000: 17)

To interpret the φ-coefficient in terms of a 2 x 2 frequency table, the so-called BESD (binomial effect-size display) is used. As an example, consider an experimental and a control group, each consisting of size 100. Suppose that of the 200 persons, 100 improved after a certain treatment and 100 did not improve; the 2 x 2 table then represents the following:

	Improved	Not improved	Total
Experiment	66	34	100
Control	34	66	100
Total	100	100	200

Thus, $\varphi = \frac{66\times66-34\times34}{\sqrt{100\times100\times100\times100}} = 0{,}32$.

In general, the content of the 2 x 2 table is:

100(0,5 + r/2) 100(0,5 - r/2)

100(0,5 - r/2) 100(0,5 + r/2),

with $r = \varphi$. In the example, $r = 0{,}32$ and thus 66% - 34% = 32%. The value of *r* therefore gives the difference in improvement rates (66% vs 34%) when one half of the population (belonging to the experimental group) receives treatment and the other half (belonging to the control group) receives no treatment. The following table represents the improvement rates for values of $r = \varphi$:

$r = \varphi$	**Improvement: from**	**to**	**Effect (Cohen,1988)**
0,0	0,50	0,50	
0,1	0,45	0,55	small
0,2	0,40	0,60	
0,3	0,35	0,65	medium
0,4	0,30	0,70	
0,5	0,25	0,75	large
0,6	0,20	0,80	
0,7	0,15	0,85	
0,8	0,10	0,90	
0,9	0.05	0,95	
1,0	0.00	1,00	

2.1.3 Interpretation of φ:

To get a feel for φ-values in terms of four-fold tables, Steyn (2002) provides the following examples in Table 2:

Table 2

Examples of 2 x 2 tables

(a) $\varphi = 0$: if frequencies in two rows (or columns) are equal, e.g.,

	$y = 1$	$y = 2$	Total
$x = 1$	50	50	100
$x = 2$	25	25	50
Total	75	75	150

(b) $\varphi = 0{,}1$ (small effect):

	$y = 1$	$y = 2$	Total
$x = 1$	45	55	100
$x = 2$	55	45	100
Total	100	100	200

(c) $\varphi = 0{,}3$ (medium effect):

	y = 1	*y* = 2	Total
x = 1	65	35	100
x = 2	35	65	100
Total	100	100	200

(d) $\varphi = 0{,}5$ (large effect):

	y = 1	*y* = 2	Total
x = 1	75	25	100
x = 2	25	75	100
Total	100	100	200

(e) $\varphi = 1$: if frequencies are 0 in any diagonal of the table, e.g.,

	y = 1	*y* = 2	Total
x = 1	100	0	100
x = 2	0	100	100
Total	100	100	200

Table 2(e) is an example of a *strictly perfect relationship* between *x* and *y* (Smithson, 2000: 324). This means that *x* determines *y* completely as well as the other way round. If a person has a 1 for *x*, it will be a 1 for *y* too, whereas all persons with a 2 for *x* will also receive a 2 for *y*.

Consider the following table, though:

	y = 1	*y* = 2	Total
x = 1	100	0	100
x = 2	75	25	100
Total	175	25	200

Here, we find a *weak perfect relationship* (Smithson, 2000: 324) in the sense that *y* can only be fully determined for category 1 of *x*, but not for category 2; *x* is also fully determined when *y* = 2. Here, $\varphi = 0{,}38$, which is a significant decrease from $\varphi = 1$. It indicates that φ *is not a suitable measure in the measurement of weak perfect relationships.* Later on, we will show that the relative odds ratio (OR) is more suitable for this purpose.

Example 1:

In Example C, Chapter 3 of Steyn (2012), the last three categories of smoking are combined so that it becomes a two-by-two table:

	Coronary heart disease: Yes	Coronary heart disease: No	Total
Smoke: Yes	78	59	137
Smoke: No	42	61	103
Total	120	120	240

In order to determine the relationship between coronary heart disease and smoking, φ is calculated as

$$\varphi = \frac{78\times61-59\times42}{\sqrt{137\times103\times120\times120}} = 0{,}16,$$

which indicates a small effect.

Suppose the 240 employees have been selected randomly from all workers at the company. Then, φ could be estimated with the value of 0,16.

In general, the sample value of φ, $\hat{\varphi}$ can be used as an estimator of the population value of φ. This estimator is asymptotically unbiased, but overestimate φ for small samples with approximately $1/\sqrt{n}$ (Johnson et.al., 1995: 447).

Remark:

On the basis of an example, Fleiss (1994) points out the following problem with φ as effect-size index. Consider two studies whose relative frequencies of *y* for given *x* are the same, but whose relative fequencies of *x* differ:

Study		*y* = 1	*y* = 0	Total
1	*x* = 1	45	5	50
	x = 0	120	30	150
	Total	165	35	200
2	*x* = 1	90	10	100
	x = 0	80	20	100
	Total	170	30	200

In both studies, the relative frequencies for *x* = 1, as well as for *x* = 0, are 45 / 50 = 90 / 100 and 5 / 50 = 10 / 100. However, the totals of the relative frequencies at *x* = 1, as well as at *x* = 0, are 50 / 200 and 100 / 200, which are thus different.

The φ-coefficients, though, are 0,11 and 0,14 for the two studies.

This means that the φ-coefficient is influenced by the degree to which the categories of *x* are represented in the data. The same holds true for the y-categories.

For this reason, $\hat{\varphi}$ is **not** a *valid estimator* if it is based on something else than a *random sample.* With randomness, the marginal totals of the 2 x 2 frequency table should be in the same relationships as those of the population. Consider the following fictitious frequency table obtained from Example 1, but where a random sample was drawn from the company instead of a stratified sample with equal numbers of employees with or without heart diseases:

	Coronary heart disease: Yes	Coronary heart disease: No	Total
Smoke: Yes	26	98	124
Smoke: No	14	102	116
Total	40	200	240

The table was obtained by distributing the 240 employees who have heart diseases into 40 instead of 120 and taking the number of smokers from them as one third of the original number of 78. In the same way, the number 98 was approximated to the nearest integer, $(59/120) \times 200$. This table should be a realisation of a random sample if one-sixth (i.e., 40 / 240) of the employees have heart diseases. The value $\hat{\varphi} = 0{,}119$ will give a valid estimate of the population φ-coefficient, whereas on the basis of a stratified sample, Example 1's value of $\hat{\varphi} = 0{,}16$ does not serve as a valid estimator.

2.1.4 Confidence interval (*CI*) for φ

For large samples, Fleiss (1994) gives the variance of $\hat{\varphi}$ approximated as:

$$Var(\hat{\varphi}) = \frac{1}{n}\left[1 - \hat{\varphi}^2 + \hat{\varphi}\left(1 + \frac{\hat{\varphi}^2}{2}\right)C_1 - \frac{3}{4}\hat{\varphi}^2 C_2\right], \qquad (2)$$

where

$$C_1 = \frac{(a+b-c-d)(a+c-b-d)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

and

$$C_2 = \frac{(a+b-c-d)^2}{(a+b)(c+d)} + \frac{(a+c-b-d)^2}{(a+c)(b+d)}.$$

The approximated $100(1-\alpha)\%$ confidence interval (*CI*) for φ has boundaries:

$$\varphi_O = \hat{\varphi} - z_{\alpha/2}\sqrt{Var(\hat{\varphi})}$$

and (3)

$$\varphi_B = \hat{\varphi} + z_{\alpha/2}\sqrt{Var(\hat{\varphi})}.$$

As alternative, the approximated *CI* can be determined by using the SAS program *CI_w* (see Steyn, 2012) for φ as a special case of *w* in subsection 2.2. Use $X2 = n\hat{\varphi}^2$, *n* and df = 1 as inputs. Excel's *Nonparametric Effect Size and CI Calculator* can also be used as an alternative for SAS.

Example 1 (continued):

For Example 1, $C_1 = \frac{(78+59-42-61)(78+42-59-61)}{\sqrt{137\times103\times120\times120}} = 0$, $C_2 = \frac{(137-103)^2}{137\times103} + 0 = \frac{1156}{14111} = 0{,}082$

$$Var(\hat{\varphi}) = \frac{1}{240}\left[1 - 0{,}16^2 + 0{,}16\left(1 + \frac{0{,}16^2}{2}\right) \times 0 \quad -\frac{3}{4}0{,}16^2 \times 0{,}082\right]$$

$$= \frac{0{,}9728}{2} = 0{,}00405.$$

Then, a *95% CI's* boundaries are:

$$\varphi_O = 0{,}16 - 1{,}96\sqrt{0{,}00405} = 0{,}16 - 0{,}125 = 0{,}035,$$

$$\varphi_B = 0{,}16 + 0{,}125 = 0{,}285.$$

For the approximate *CI* (by means of SAS or Excel), the inputs are:

X2 = $240(0{,}16)^2 = 6{,}144$, n = 240 and vg = 1.

This produces the 95% *CI* of 0,032 and 0,287, which is very close to the approximate *CI*.

Thus, even with a large sample such as 240, the *95% CI* is moderately wide and the value of φ varies in such a way that it is a small to medium effect.

2.1.5 Probability measures from 2 x 2 frequency tables

Suppose the proportions of population elements of populations 1 and 2 are *p* and *q,* respectively. Suppose the response on *y* can be positive or negative (e.g., 'agree' versus 'differ'; in case-control studies in epidemiology, 'exposed' versus 'not exposed'; in intervention studies, 'improve' versus 'not improve'). Take the probabilities (proportions) for positive π_1 and π_1 in the two populations. The 2 x 2 frequency table thus represents the following:

Table 3

General 2 x 2 table

	y: positive	*y: negative*	Total
x: population 1	$pN\pi_1$	$pN(1-\pi_1)$	*pN*
x: population 2	$qN\pi_2$	$qN(1-\pi_2)$	*qN*
Total	$N\pi$	$N(1-\pi)$	*N*

Here, $\pi = p\pi_1 + q\pi_2$ is the probability of a positive response by both populations, whereas *N* is the total number of elements in both populations.

By using the table above, we now discuss the following three *comparative risk or rate measures:*

- Difference in proportion of positive responses $\pi_1 - \pi_2$
- Ratio of proportion of positive responses π_1/π_2, the rate or risk ratio
- Relative odds ratio

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}\,.$$

Rate measures is a more general term, because only when 'positive' means something undesirable, for example 'exposed', 'identified', 'ill' or 'dead', we can use the term *risk measures*.

2.1.6 Difference in proportions

As in the case of averages, two kinds of effect-size indices are here under discussion: firstly, standardised differences in proportions and secondly, the relationship between the response *y* and the population distribution *x*.

(a) Standardised differences in proportions:

Take y_i = 1 if population i is positive and y_i = 0 if population i is negative.

Then, the population average of y_i is $\mu_i = \pi_i$ and its population variance is ${\sigma_i}^2 = \pi_i(1-\pi_i)$, so that, from subsection 4.21 in Chapter 4 of the manual (Steyn, 2012), with weights $W_1 = p$ and $W_2 = q$, it follows that:

$$\delta_g = \frac{\pi_1 - \pi_2}{\sqrt{p\pi_1(1-\pi_1)+q\pi_2(1-\pi_2)}}. \quad (4)$$

If equal population variances are accepted, each variance can be replaced by $\pi(1-\pi)$ (remember that $\pi = p\pi_1 + q\pi_2$ and $p + q = 1$) and this produces the proportion analogue of δ:

$$\delta = \frac{\pi_1 - \pi_2}{\sqrt{\pi(1-\pi)}}. \quad (5)$$

With population 1 as reference point (i.e., the control), the effect-size index becomes:

$$\Delta_1 = \frac{\pi_1 - \pi_2}{\sqrt{\pi_1(1-\pi_1)}}. \quad (6)$$

The estimators δ_g, δ and Δ_1 can be obtained by replacing the proportions π_1 and π_2 with p_1 and p_2, which are the sample proportions for the two populations.

However, the problem with all three indices mentioned above is that the standard deviation by which they are divided depends on π_1 and π_2.

Cohen (1969, 1977, 1988) therefore proposes the following *effect-size index:*

$$\psi = 2[bgsin(\sqrt{\pi_1}) - bgsin(\sqrt{\pi_2})]. \quad (7)$$

Take note that $bgsin(x)$ is the angle in radians a of which $sin(a) = x$.

Remarks:

- On pocket calculators, the function $bgsin(x)$ is also indicated by $arcsin(x)$ or $sin^{-1}(x)$.
- Radians are obtained in degrees from an angle by $a = \frac{\theta}{360} \times 6{,}283$, where θ is the angle in degrees.
- If π_1 or $\pi_2 = 0$, use $bgsin\left(\sqrt{1/(4n)}\right)$ instead of $bgsin(0)$.
- If π_1 or $\pi_2 = 1$, use 1,571 - $bgsin\left(\sqrt{1/(4n)}\right)$ instead of $bgsin(1)$.
- The standard deviation of ψ is independent of π_1 and π_2 so that, as in comparison of averages, the scale remains constant. For example, for $\pi_1 = 0{,}65$ and $\pi_2 = 0{,}35$, $\psi = 0{,}61$, whereas for $\pi_1 = 0{,}5$ and $\pi_2 = 0{,}2$, $\psi = 0{,}64$. This means that a difference of 0,3 in proportions produces more or less a difference of 0,6 on the ψ-scale. With the index δ_g, the concordant values would be 0,63 and 0,50 if $p = q = 0{,}5$ is assumed.

- Take note that in a BESD two-by-two table (see previous section), all the marginal totals are 100 and $\pi_1 - \pi_2$ is transformed to φ. Therefore, $r = \pi_1 - \pi_2$ can be taken and the BESD determined from the difference in proportions.

If random samples of sizes n_1 and n_2 from the populations produce p_1 en p_2 as proportions, the estimator

$$\hat{\psi} = 2\left[bgsin\left(\sqrt{p_1}\right) - bgsin\left(\sqrt{p_2}\right)\right] \quad (8)$$

can be used. For large samples, $\hat{\psi}$ is normally distributed with average ψ and variance $\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{n_1+n_2}{n_1 n_2}$, so that the *100(1-α)% CI* is given by the boundaries: $\psi_O = \hat{\psi} - z_{\alpha/2}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

and (9)

$\psi_B = \hat{\psi} + z_{\alpha/2}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$.

Example 1 (continued):

Consider Example 1 and take the coronary heart disease patients as population 1 and those without the disease as population 2. Now, $\pi_1 = \frac{78}{120} = 0{,}65$ and $\pi_2 = 0{,}49$, $p = \frac{120}{240} = 0{,}5$.

$$\delta_g = \frac{0{,}65-0{,}49}{\sqrt{0.5\times0{,}65\times0{,}35+0{,}5\times0{,}49\times0{,}51}} = \frac{0{,}17}{0{,}489} = 0{,}348.$$

To determine δ, we calculate π = 137 / 249 = 0,57, so that

δ = 0,17 / $\sqrt{0{,}57 \times 0{,}43}$ = 0,17 / 0,495 = 0,343, which is for all practical purposes the same as δ_g. $\psi = 2\left[bgsin\left(\sqrt{0{,}65}\right) - bgsin(\sqrt{0{,}49}\right] = 2(0{,}9377 - 0{,}7754) = 0{,}325$, which gives more or less the same effect size.

If the assumption is made that there are two random samples drawn from populations 1 and 2, the approximated 95% *CI* for ψ's boundaries are:

$$\psi_O = 0{,}325 - 1{,}96\sqrt{\frac{120+120}{120\times120}} = 0{,}325 - 1{,}96 \times 0{,}129 = 0{,}072$$

$$\psi_B = 0{,}325 + 1{,}96 \times 0{,}129 = 0{,}578.$$

The value ψ in the population can thus be as low as $0{,}072$, but also as high as $0{,}578$ (with a probability of 95%).

2.1.7 Guideline values for differences in proportions

From Example 1 (continued), it seems as if all three the effect-size indices δ_g, δ and ψ produce more or less the same values. In practice, this holds true for all the combinations of $0{,}1 \leq \pi_1, \pi_2 \leq 0{,}9$ and $0{,}25 \leq p \leq 0{,}5$. With reference to the guideline values of δ, based on averages, Cohen (1969, 1977, 1988) proposes the same guidelines:

Small effect: δ_g, δ and $\psi = 0{,}2$: This value is obtained when (π_1, π_2) form, for example, the following pairs: (0,005; 0,1), (0,2; 0,29), (0,4; 0,5), (0,6; 0,7), (0,8; 0,87) and (0,9; 0,95).

Medium effect: δ_g, δ and $\psi = 0{,}5$: Here, (π_1, π_2)-values are, for example, the pairs: (0,05; 0,21), (0,2; 0,43), (0,4; 0,65), (0,6; 0,82), and (0,8; 0,96).

Large effect: δ_g, δ and $\psi = 0{,}8$: Here, (π_1, π_2)-values are, for example, the pairs:

(0,05; 0,34), (0,2; 0,58), (0,4; 0,78), (0,6; 0,92), and (0,8; 0,996).

Burnand et. al. (1990) propose guidelines that were determined empirically from a survey of 392 articles in the medical literature:

- Significant: δ = 0,28
- Substantially significant: δ = 0,35
- Highly significant: δ = 0,65

2.1.8 Rate or risk ratio

The rate ratio is the ratio of the probabilities π_1 and π_2, as defined in subsection 2.1.5. If population 1 is the control population and population 2 the treatment population, π_1/π_2 is the ratio of the proportion of positive responses of the control persons relative to the treated persons. If 'positive' means something such as illness or death, one refers to a risk ratio. If $\pi_1/\pi_2 > 1$, it means the risk is greater in the control than in the treatment group, which means that the treatment has been advantageous. The calculation of π_1/π_2 in terms of the cell frequencies of a two-by-two frequency table (Table 3) is the following:

$$\frac{\pi_1}{\pi_2} = \frac{a/(a+b)}{c/(c+d)}\,. \qquad (10)$$

If working with random samples of sizes n_1 and n_2 from the populations, the estimated rate ratio is p_1/p_2, where p_1 and p_2 are the sample proportions of samples from the two populations.

The disadvantage of the rate ratio is that it can become very large if π_2 becomes very small relative to π_1. It therefore does not serve as an effect-size index such as, for example, φ or η^2 that lies between 0 and 1; it should be evaluated by how far it lies from 1, as $\pi_1/\pi_2 = 1$ means there is no difference in rate or risk. The natural logarithm of π_1/π_2, namely $ln\left(\frac{\pi_1}{\pi_2}\right) = ln\pi_1 - ln\pi_2$, also serves as an effect-size index. It can take on any value, with the zero as no difference in rate. According to Fleiss (1994) and Kline (2004a), $ln\left(\frac{p_1}{p_2}\right)$ is approximately normally distributed if the samples are large. Further,

$$Var\left[\ln\left(\frac{p_1}{p_2}\right)\right] = \frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}\,, \qquad (11)$$

so that the *100(1-α)% CI* for the boundaries (*L; U*) of $ln(\frac{\pi_1}{\pi_2})$ is obtained from:

$$ln(\frac{p_1}{p_2}) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}\,. \qquad (12)$$

Then, the *CI* for π_1/π_2 has boundaries:

$$(\frac{\pi_1}{\pi_2})_L = e^L \text{ and } (\frac{\pi_1}{\pi_2})_U = e^U, \qquad (13)$$

where (*L, U*) are the *CI* with boundaries in (12) and (13).

Example 1 (continued): From Example 1, take population 1 as persons with coronary heart disease and population 2 as those without the disease; then, $\pi_1 = \frac{78}{120} = 0{,}65$ and $\pi_2 = \frac{59}{120} = 0{,}49$ are the proportions of probabilities that persons from these populations are smokers. $\frac{\pi_1}{\pi_2} = \frac{0{,}65}{0{,}49} = 1{,}327$, which means that persons with coronary heart disease are 1,3 times more inclined to smoke than those without the disease. Smoking can therefore be a risk factor for this disease. If the 120 per group are viewed as two random samples, $p_1 = 0{,}65$ and $p_1 = 0{,}49$, so that $\frac{p_1}{p_2} = \frac{0{,}65}{0{,}49} = 1{,}327$ is the estimate of the rate ratio, whereas

$$\ln\left(\frac{p_1}{p_2}\right) = \ln(1{,}327) = 0{,}283,$$

$$Var\left[\ln(\frac{p_1}{p_2})\right] = \frac{1-0{,}65}{120\times 0{,}65} + \frac{1-0{,}49}{120\times 0{,}49} = 0{,}00449 + 0{,}00864 = 0{,}0131.$$

100(1-α)%CI for $\ln(\frac{\pi_1}{\pi_2})$: $0{,}283 \pm 1{,}96\sqrt{0{,}0131} = 0{,}283 \pm 0{,}225 = (0{,}058; 0{,}508)$, so that $(\frac{\pi_1}{\pi_2})_O = e^{0{,}058} = 1{,}06$, $(\frac{\pi_1}{\pi_2})_B = e^{0{,}508} = 1{,}661$.

This means that with a 95% probability, $\frac{\pi_1}{\pi_2}$ can be as low as 1,06 but also as high as 1,661. There is thus an indication of a risk.

2.1.9 Odds ratio

First, it is necessary to define *odds*. In terms of Table 3, the odds for population 1 is $\pi_1/(1-\pi_1)$ and for population 2, $\pi_2/(1-\pi_2)$. It thus gives the ratio of the probability of *y* being positive with reference to the probability of *y* being negative.

Example 1 (continued): In Example 1, the odds ratios for persons with coronary heart disease is $\frac{78}{42} = 1{,}857$, whereas it is $\frac{59}{61} = 0{,}967$ for persons without the disease. Thus, with heart disease patients, around 1,9 of them smoke for each one who does not smoke, whereas it is nearly 1 with persons who do not have heart disease.

If the two populations' odds are to be compared, it can be done by determining the ratio thereof.

This ratio is called the *odds ratio* (*OR*):

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} = \frac{ad}{bc}. \qquad (14)$$

To calculate, it is easiest to use $\frac{ad}{bc}$ from Table 1. The value of *OR* can vary between 0 and infinity, with the value of 1 when the two odds are the same. The values 0 and infinity are obtained if any of the frequencies in the 2 x 2 table has a value of 0. It is precisely the case in subsection 2.1.3 if a weak perfect relationship exists between *x* and*y*.

Example 2 (Smithson, 2000: 324):

While researching snake phobia, a clinical psychologist obtained the following frequency table:

	Snakes: persons who like them	Snakes: persons who dislike them	Total
Snakes: persons who fear them	49 (*a*)	5 (*b*)	54
Snakes: persons who don't fear them	159 (*c*)	49 (*d*)	208
Total	208	54	262

The odds ratio for people who fear snakes = 49 / 5 = 9,8.

The odds ratio for people who do not fear snakes = 159 / 49 = 3,24.

The odds ratio is:

$$OR = 9{,}8 / 3{,}24 = 3{,}02.$$

Take note that *OR* could also be obtained from:

$$OR = (ad) / (cd) = (49 \times 49) / (5 \times 159) = 3{,}02.$$

(Here, *a* is taken as the frequency of the 'Yes/Yes' category etc.)

This means that the odds ratio for people who fear snakes is three times higher than for those who do not fear them. An *OR* of 1 / 3,02 = 0,331 would have the same meaning if the odds ratio of people who do not fear snakes would be compared to those who do fear snakes.

Smithson (2000: 326) names *two advantages* of *OR* as measure of ratio over those of the φ-coefficient:

a) It also serves as a measure of weak perfect relationships.
b) It stays the same, even when a row or column of the 2 x 2 table is multiplied by a factor.

If the random sample is drawn from a population, the population *OR* (ω) is estimated with $\hat{\omega}$ where *a, b, c* and *d* are the sample frequencies. If *b* or *c* (or both) are zero, $\hat{\omega}$ is undefined. The estimator introduced by Jewell (see Shoukri & Chaudhary, 2007) can then be used, namely

$$\widehat{\omega}_J = \frac{ad}{(b+1)(c+1)}.$$

From Monte-Carlo simulations for *n* = 25, it appears that $\widehat{\omega}_J$ has a smaller bias and average squared error than other estimators such as $\widehat{\omega}$.

Example 2 gives a near weak perfect relationship (the frequency of 5, which is near to zero). Here, *OR* was 3,02 and if the frequencies of the first row and column were 3 and 51, it would have changed to 5,552 and become infinitely large if the frequencies were 0 and 54. The φ-coefficient for Example 2 is 0,143 and increases to 0,26 if the first cell frequency 0 is taken. It only indicates that the relationship is far from perfect, which illustrates advantage a).

Regarding advantage b), we refer to the remark in subsection 2.1.3, where two studies with different relative frequencies for the two categories of *x* produce different φ-values, namely 0,11 and 0,17. However, for these two studies, the

OR values are the same: Study 1: (45 x 30) / (120 x 5) = 2,25

Study 2: (90 x 20) / (80 x 10) = 2,25

In the same way as the rate ratio π_1/π_2, *OR* is evaluated with regard to the distance from 1.

Therefore, the natural logarithm of *OR* is sometimes easier to use, because the distance from 1 is then transformed to a distance from 0.

If working with a random sample from a population, the population's *OR*(ω) is estimated with $\widehat{\omega}$, where *a, b, c* and *d* are the sample frequencies. For large samples, it further holds true that $ln(\widehat{\omega})$ is approximately normally distributed with an average of $ln(\omega)$ and variance (Fleiss, 1994):

$$Var[ln(\widehat{\omega})] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}. \qquad (15)$$

Thus, a *100(1-α)% CI* for $ln(\omega)$ (with boundaries (*L, U*)) is

$$ln\left(\frac{ad}{bc}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}, \qquad (16)$$

so that the *CI* boundaries of ω are

$$\omega_L = e^L \quad \text{and} \quad \omega_U = e^U. \qquad (17)$$

Additional applications of odds ratios are discussed in Fleiss (1994):

a) When other variables (covariants) influence the response variable *y* (on top of grouping variable *x*), a logistic regression analysis can be done from which the *OR* value can be obtained directly.
b) The Mantel-Haenszel estimator is another method to combine the $ln(\omega)$ values when the covariants in a) is categorical (and the data are thus divided into strata).

Newcombe (2006) gives the following reasons why *OR* is the measure most used in 2 x 2 frequency tables:

- It has a natural role in logistic regression.
- It is the only significant measure when the sampling is non-random, but retrospective case-control study designs are dealt with, which happens many times in epidemiological studies.
- When the occurrence of an event (e.g., a disease) is rare, the value of *OR* is much the same as that of the risk ratio (*RR*), seeing that for values of *a* and *c* small in comparison to those of *b* and *d*, it follows that $a / b \approx a / (a+b)$ and $c / d \approx c / (c+d)$.

However, Newcombe warns that the *OR* value always lies further from 1 than the *RR* value and the risk is thus exaggerated. Further, the *OR* is used as if it is the same as *RR,* which is only true in rare occurrences.

Example 2 (continued)

If the frequency table in Example 2 renders the results of a random sample, $\widehat{\omega}$ = 3,02 and the 90% *CI* for $ln(\omega)$:

$$ln(3{,}02) \pm 1{,}645\sqrt{\frac{1}{5}+\frac{1}{49}+\frac{1}{49}+\frac{1}{159}} = 1{,}105 \pm 1{,}645 \times 0{,}497 = (0{,}287; 1{,}923).$$

Thus, 90% *CI* for ω is (1,333; 6,840), so that the population *OR* can be as small as $1{,}33$ and as large as 6,84 with a 90% probability.

For Example 1, if two random samples from the populations of persons with heart disease and those without this disease are drawn, the 95% *CI* for the *OR* is calculated as:

$$ln\left(\frac{78 \times 61}{42 \times 59}\right) \pm 1{,}96\sqrt{\frac{1}{78}+\frac{1}{59}+\frac{1}{42}+\frac{1}{61}} = ln(1{,}92) \pm 1{,}96 \times 0{,}07 = 0{,}652 \pm 0{,}137$$

$= (0{,}515; 0{,}789)$.

Thus, the 95% *CI* for ω has boundaries:

$$\omega_L = e^{0{,}515} = 1{,}674;\ \omega_U = e^{0{,}789} = 2{,}201.$$

Here again, ω of the population varies from a small to a medium effect (see next subsection).

2.1.10 Interpretation of *OR* as effect size

According to Kline (2004a: 147) and Chinn (2000), *OR* can be transformed to a standardised difference analogous to δ. Because $ln[p_1/(1-p_1)]$ and $ln[p_2/(1-p_2)]$ both have a logistical distribution that is approximately normal with standard deviation $\pi/\sqrt{3}$ =1,81, the standardised difference becomes

$$\hat{\delta}_{OR} = \frac{ln[p_1/(1-p_1)] - ln[p_2/(1-p_2)]}{1{,}81} = \frac{logit(p_1) - logit(p_1)}{1{,}81}. \qquad (18)$$

Note that $logit(p) = ln[p/(1-p)]$. The standardised difference δ_{OR} can thus be identified similar to δ, and the same guideline values can be used, so that

small effect: $\delta_{OR} = 0{,}2$;

medium effect: $\delta_{OR} = 0{,}5$;

large effect: $\delta_{OR} = 0{,}8$.

Because it follows from (18) that $\omega = e^{1{,}81\delta_{OR}}$, it holds true that:

small effect: $\omega = 1{,}44$, take as 1,5;

medium effect: $\omega = 2{,}48$, take as 2,5;

large effect: $\omega = 4{,}27$, take as 4,25.

Although an *OR* larger than 1 indicates that the odds ratio of the one population is larger than the other one, it cannot necessarily be concluded that there is an important difference in odds ratios. As 0,5 and 0,8 per guideline values bring about medium and large effects, only the guideline values of 2,5 and 4,25 would suggest medium and large effects at *OR* values.

On the basis of a survey in medical magazines in which 392 articles were relevant, Burnand et. al. (1990) propose the following guideline values for *OR*:

- Significant: *OR* = 2,2
- Substantially significant: *OR* = 2,5
- Highly significant: *OR* = 4,0

The last two guideline values concur with 'medium' and 'large' effects.

An additional interpretation of *OR* is the following:

Tritchler (1995) is of opinion that with two normal populations (Pop. 1 and Pop. 2) with averages μ_1 and μ_2 and the same standard deviation σ, it holds true that

$E = P$ (classify x in Pop. 1 $\mid x$ is from Pop. 2)

$= P$ (classify x in Pop. 2 $\mid x$ is from Pop. 1)

$= \Phi\left(-\frac{\delta}{2}\right)$, (19)

with $\Phi(\mathrm{t})$ the cumulative distribution function of a standard normal distribution, and

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

the standardised absolute difference in averages, as defined in Steyn (2012, Chapter 4).

The special case in univariate populations of the linear classification rule in discriminant analysis (see Steyn, 2012, Chapter 8) transforms to the following:

Classify x in Pop.1 if $x > \frac{\mu_1+\mu_2}{2}$, if $\mu_1 > \mu_2$.

Tritchler (1995) then proposes the joint probabilities of the two dichotomisations, $x > c$, $x \leq c$, against Pop. 1, Pop. 2, as follows:

	Pop. 1	*Pop. 2*
x > c	*(1-E).P(1)*	*E.P(2)*
x ≤ c	*E.P(1)*	*(1-E).P(2)*

where *P*(i) = *P*(*x* is from Pop. i).

The odds ratio of this 2 x 2 table is therefore:

$$\omega = \frac{\left(\frac{1-E}{E}\right)}{\left(\frac{E}{1-E}\right)} = \left(\frac{1-E}{E}\right)^2, \qquad (20)$$

so that it follows from (19) that

$$\omega = \left[\frac{1-\Phi\left(-\frac{\delta}{2}\right)}{\Phi\left(-\frac{\delta}{2}\right)}\right]^2. \qquad (21)$$

When we use the guideline values of Cohen (1988), we obtain the following effects from Steyn's (2012) equation (5.63) in Chapter 5:

small effect: δ = 0,2: ω = 1,38;

medium effect: δ = 0,5: ω = 2,25;

large effect: δ = 0,8: ω = 3,64.

These values of ω concur to a certain degree with those obtained from δ_{OR} and those proposed by Burnand et.al. (1990).

Here, the same warnings as in Steyn (2012, subsection 4.5.4) are applicable; the proposed guideline values must thus be handled with circumspection.

2.2 Effect size of relationship between two nominal variables

A significant measure of the degree to which the cell frequencies in a two-way frequency table deviate from the expected frequencies if no relation is assumed, is (Cohen, 1969, 1977, 1988):

$$w = \sqrt{\frac{X^2}{N}} = \sqrt{\sum_{i-1}^{m} \frac{(f_i - v_i)^2}{N v_i}}, \qquad (22)$$

where f_i is the *i-th* cell's frequency, v_i the expected frequency of cell *i* if no relationship (i.e the null hypothesis) is assumed and *m* = *IJ*, with *I* the number of rows and *J* the

number of columns in the frequency table. *N* is the total of the cell-frequencies (i.e. the total number of observations).

X^2 is also the chi-square test statistic when a test is done (on the basis of a random sample) for a statistically significant relationship.

The expected frequency of a cell (if there is no relationship) is:

(row total) x (column total) / *N* (totals of row and column in which the cell falls), where *N* = sum of row totals = sum of column totals = total frequency.

Example 3:

In Example B, Chapter 3 (Steyn, 2012) is the table (in which the column with the lecturers' frequencies is ommitted) with frequencies and expected frequencies (in brackets):

	Male students	Female students	Total
Temperament SJ	57(64,79)	79(71,21)	136
Temperament SP	29(24,77)	23(27,23)	52
Temperament NT	23(20,01)	19(21,99)	42
Temperament NF	12(11,43)	12(12,57)	24
Total	121	133	254

$$X^2 = \frac{(57-64,79)^2}{64,79} + \frac{(79-71,21)^2}{71,21} + \frac{(29-24,77)^2}{24,77} + \frac{(23-27,23)^2}{27,23} + \frac{(23-20,01)^2}{20,01} + \frac{(19-21,99)^2}{21,99} + \frac{(12-11,43)^2}{11,43} + \frac{(12-12,57)^2}{12,57} = 4,074,$$

therefore $w = \sqrt{\frac{4,074}{254}} = 0,127.$

The measure w can serve as an effect-size index to measure the relationship between two nominal variables (temperament type and gender of students in Example 3). It is clear that the more f_i differs from v_i, the larger $(f_i - v_i)^2/v_i$ becomes and if there are large differences in most of the cells, X^2 should be large. Because the size of X^2 is also influenced by *N*, $\frac{X^2}{N}$ is a more significant measure. In the special case of 2 x 2 tables,

$$\varphi^2 = \frac{X^2}{N} = w^2, \qquad (23)$$

which is also a reason why $\sqrt{\frac{X^2}{N}}$ is used as an effect-size index to indicate a relationship.

Smithson (2000: 313) points out that, except for the fact that *N* influences the size of X^2, the number of cells also plays a role in the sense that the more cells there are, the larger X^2 becomes (the number of terms in the sum becomes larger). To compensate for this, *Cramer's V* (see also Cohen, 1969, 1977, 1988) can be used:

$$V = \sqrt{\frac{X^2}{N(k-1)}} = \frac{w}{\sqrt{k-1}}, \qquad (24)$$

where *k* = *min(I, J)*.

In Example 3, *k* = 2, because *I* = 4 and *J* = 2, so that *V* has the same value as w.

Remark:

For smaller tables, *V and* w are almost the same, but where w can be interpreted like a correlation because it lies between 0 and 1, the same cannot be said of *V* in larger tables. For *k* > 2, the maximum value of *V* becomes smaller than 1, so that the size of the table has an influence on the value of *V*.

2.2.1 Estimation of *w*

When a random sample is selected from a population, the effect-size index w can be estimated with $\widehat{w}$ by using the sample's frequencies.

For smaller samples, *w* is *overestimated* and the bias of w^2 is approximately $\frac{(I-1)(J-1)}{n}$ where *n* is the sample size (see Steyn, 2002).

Therefore, *w* can rather be estimated by:

$$\widetilde{w} = \sqrt{\widehat{w}^2 - \frac{(I-1)(J-1)}{n}}, \qquad (25)$$

which is approximately unbiased for w.

Example 4 (Smithson, 2000):

By using the Crosspatch program of Smithson, the following frequencies were obtained in a random sample in which the preferences of 10 to 40-year-old persons in three age groups were asked for four kinds of shoes:

Shoe

Age	Kind 1	Kind 2	Kind 3	Kind 4	Total
10 - 19	86(44,0)	5(12,7)	38(54,6)	14(31,7)	143
20 - 29	4(18,8)	14(5,4)	4(23,3)	39(13,5)	61
30 - 39	14(41,2)	11(11,9)	87(51,2)	22(29,7)	134
Total	104	30	129	75	338

$X^2 = 194{,}01$ (p $<$ 0,0001), $\widehat{w} = \sqrt{194{,}01/338} = 0{,}758$

$\widehat{V} = \frac{\widehat{w}}{\sqrt{2}} = 0{,}536$ (because *k* = 3).

There is a statistically significant relationship (*p* < 0,0001). The estimation of $\widehat{w}$ of 0,758 can be used to obtain the effect of relationship between the kind of shoe and age in the population and is practically unbiased, because the bias of $\widehat{w}^2$ is approximately (2 x 3) / 338 = 0,018, so that $\widetilde{w} = \sqrt{0{,}758^2 - 0{,}018}$ = 0,746.

2.2.2 Confidence interval for *w*

According to Johnson et.al. (1995: 467), the chi-square statistic X^2 has approximately a non-central chi-square distribution with *(I – 1)(J – 1)* degrees of freedom and non-centrality parameter *nsp* = nw^2. As in subsection 2.1.4 for *φ*, it is now possible to determine a *100(1 - α)% CI* for *nsp* by means of a computer program, and to obtain from that an approximate *CI* for *w* and *V* with, say, lower and upper bounds (*L, U*). Then, the *CI* for *w* has boundaries $(\sqrt{L/n}\,, \sqrt{U/n})$ and those of *V* is $(\sqrt{L/n(k-1)}\,, \sqrt{U/n(k-1)})$. The calculation is done by the SAS program *CI_w,* but Excel can also be used as an alternative for SAS by using the *Nonparametric Effect Size and CI Calculator.xls.*

Example 4 (continued): **I**n Example 4, the 95% *CI* for *w* is (0,640; 0,855), which means that the unknown population *w* can vary between 0,64 and 0,86 with a probability of 0,95. For *V*, the *CI* is (0,45; 0,61).

2.2.3 Guideline values for *w*

Cohen (1969, 1977, 1988) links guideline values for *w* to a table in which *w* and Cramer's *V* are given for different values of *k*. Table 4 provides an excerpt from it and uses the relationship in (24).

Table 4

Values of w and concordant *V*

k	2	3	4	5	6
w = 0,1	0,1	0,071	0,058	0,05	0,045
w = 0,3	0,3	0,212	0,173	0,15	0,134
w = 0,5	0,5	0,354	0,289	0,25	0,224

Take note that when *k* = *min(I, J)* = 2, *w* = *V*. When *I* = *J* = *2*, it also holds true that *w* = φ.

The guideline values for *w* could thus be chosen on the basis of those for φ:

- small effect: *w* = 0,1;
- medium effect: *w* = 0,3;
- large effect: *w* = 0,5.

However, Cohen warns that perhaps these guidelines are not realistic for larger tables. Because Cramer's *V* is an adjusted index for larger tables, Table 4 could then be used. For example, if *I* = 6 and *J* = 10, *k* = 6 and *V*-values of 0,224, 0,134 and 0,045 can already be viewed as large, medium and small effects.

In Example 3, *w* = 0,127 and because *k* = 2 in Table 4, it is a small effect. For a larger table with the same *w*-value, it could be viewed as a medium effect if *k* was, for example, larger than 4. In Example 4, even the lower bound of the 95% *CI* (i.e., 0,45) gives us the right to classify it as a large effect, because at *k* = 3, a large effect is 0,354.

2.3 Effect sizes in 2 x 2 frequency tables with dependent pairs

Suppose a diagnostic test is applied to persons in order to classify them as positive (they have, e.g., contracted a disease), or negative (they are healthy). If both groups are treated thereafter, the test can also be done on the same persons after the treatment. A 2 x 2 frequency table of the results can now be drawn up. Take note how it differs from Table 1 in subsection 2.1:

Table 5

The 2 x 2 frequency table of the same people before and after treatment

	After treatment: positive	After treatment: negative	Total
Before treatment: positive	*a*	*b*	*a + b*
Before treatment: negative	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n*

In Table 5, *a* represents the number of persons who were positive before and after treatment, whereas *d* represents the number of persons who stayed negative after treatment. The treatment had therefore no effect on these persons. Usually, there is no interest in the latter, but rather in those who have changed from before to after their treatment. In the table, *b* represents ill persons (thus positive) who healed (negative) after treatment, whereas *c* represents healthy persons who became ill after treatment.

To test the null hypothesis that the probability of persons being positive before and after the treatment is the same, must

$$p_a + p_b = p_a + p_c, \qquad (26)$$

where p_a is the probability that a person is before and after the treatment positive et cetera.

For the null hypothesis that the probability of persons being negative before and after treatment is the same, must

$$p_b + p_d = p_c + p_d. \qquad (27)$$

Both (26) and (27) imply that

$H_0: p_b = p_c$ against the alternative $H_1: p_b \neq p_c$.

In the case of a complete population (of size *N*), $p_b = \frac{b}{N}$ and $p_c = \frac{c}{N}$, whereas in a random sample of size *n* from this population, the probabilities can be estimated as $\hat{p}_b = \frac{b}{n}$ and $\hat{p}_c = \frac{c}{n}$.

In the case of a sample, the test statistic of the **McNemar test** is:

$$X^2 = \frac{(|b-c|-1)^2}{b+c}, \qquad (28)$$

with X^2 under H_0 approximately chi-square distributed with 1 degree of freedom, provided that *b* or *c* is not too small (usually *b* + *c* > 25).

In cases where *b* and *c* are small, the **binomial test** can be used to test H_0, where only the number of persons who have changed (i.e., the *b* and *c* frequencies) are viewed.

Let $p_b^* = \frac{b}{b+c}$ be the probability that some persons who have changed, were positive before treatment and became negative afterwards. Similarly, let $p_c^* = \frac{c}{b+c}$, so that

$H_0: p_b = p_c = p_b^* = p_c^*$ = 0,5, which can be tested for one proportion with the binomial test.

2.3.1 Effect sizes

(a) From the McNemar test statistic follows that:

$$w_M = \sqrt{\frac{X^2}{b+c}}. \qquad (29)$$

The effect size w_M is then interpreted in the same way as *w* previously was (see subsection 2.2).

(b) From the probability (proportion) of change from positive to negative (Cohen, 1988; Steyn, 2012: Chapter 5):

g = $p_b^* - 0{,}5$ if *b* ≥ *c*, otherwise *g* = $p_c^* - 0{,}5$.

According to Cohen (1988), the guideline values are: small effect when *g* = 0.05; medium effect when *g* = 0,15; and large effect when *g* = 0,25.

(c) Odds ratio of change from before to after treatment:

$$OR_M = \frac{b}{c}. \qquad (30)$$

Because $OR_M = \frac{b/(b+c)}{c/(b+c)}$, it gives the ratio of probability of change from positive to negative to that of change from negative to positive (only for persons whose condition changed). If

positive means 'ill', a large OR_M would mean a small risk, whereas a small OR_M would bring about a large risk.

OR is also called the Mantel-Haenszel odds ratio (see Olivier et. al., 2017) and is indicated as OR_{MH}. The relationship between the effect size *g* and OR_{MH} is then

$g = ½\frac{OR_{MH}-1}{OR_{MH}+1}$, so that $OR_{MH} = \frac{1+2g}{1-2g}$.

Cohen's guideline values (based on *g*) then produce the following for OR_{MH} (and thus for *OR*): small effect - $OR_{MH} = 1{,}22;$ medium effect - OR_{MH} = 1,86; and large effect - OR_{MH} = 3,00.

Take note:

- If working with complete populations, the ratios mentioned above are in reality the same as in the populations, whereas in the case of random samples, they become estimated ratios.
- $OR_M = \frac{b}{c} = \frac{b/N}{c/N}$, so that in the interpretation of OR_M, the ratios could be meant to come from either (*b* + *c)* or *N* (where *N* could be the population or sample size).

2.3.2 Confidence intervals in samples

(a) Effect size w_M: If *n* is large, X^2 has approximately a non-central chi-square distribution with 1 degree of freedom and non-centrality $nw_M{}^2$. See subsection 2.2.2 for computer programs that determine an approximate *100(1- α)% CI* for the population value of w_M. For SAS's *CI_w,* the inputs are X2 = X^2, df = 1 to first determine the *100(1- α)% CI* of non-centrality as *(nc_lower, nc_upper)*. The *CI* of w_M is then *(nc_lower/(b+c), (nc_upper/(b+c).* It can also be calculated in Excel's *Nonparametric Effect Size and CI Calculator.xls*.

(b) Effect size *g*: For *n* large, $Z = \frac{(p_b^*-0{,}5)}{\sqrt{p_b^*(1-p_b^*)/(b+c)}}$ standard normally distributed, so that the *100(1- α)% CI* for *g* is given by:

$g \pm z_{\alpha/2}\sqrt{p_b^*(1-p_b^*)/(b+c)}$. (31)

(c) Effect size for OR_M: Similar to rate or risk ratios in subsection 2.1.8, it holds true for *n* large that $ln\left(\frac{b}{c}\right)$ is approximately normally distributed with $Var\left[ln\left(\frac{b}{c}\right)\right] = \frac{1}{b} + \frac{1}{c}$. Thus, the *100(1- α)% CI* for $ln\left(\frac{b}{c}\right)$ is:

$$ln\left(\frac{b}{c}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{b} + \frac{1}{c}}, \qquad (32)$$

from which (e^L, e^U) forms *CI* for $OR_M = \frac{b}{c}$, with *L* and *U* the lower and upper confidence limits of (32).

Example 5

In a sample of 164 persons, it was first determined whether they had contracted a certain disease before treating them. After the treatment was completed, the persons were tested for the disease again. The following 2 x 2 frequency table renders the results:

	After treatment: positive (ill)	After treatment: negative (healthy)	Total
Before treatment: positive (ill)	42	44 (b)	86
Before treatment: negative (healthy)	14 (c)	64	78
Total	56	108	164

Only the frequencies 44 (was ill before but is healthy after treatment) and 14 (was healthy before but is ill after treatment) are viewed further.

(a) Effect size w_M: $X^2 = \frac{(|44-14|-1)^2}{44+14}$ = 14,5, $w_M = \sqrt{\frac{14,5}{58}}$ = 0,5.

The computer programs in SAS and the Excel spreadsheet (see subsection 2.1) give the approximate 95% *CI* for w_M as: (0,24; 0,76); there is thus a large effect that can also be medium with a 95% probability (if looking at the guideline values for *w* in, e.g., subsection 2.2.3).

(b) Effect size *g*: $p_b^* = \frac{44}{58}$ = 0,76, so that *g* = 0,76 – 0,5 = 0,26 – also a large effect (see subsection 2.3.1 [b] above).

The approximate 95% *CI*: $0,26 \pm 1,96\ \sqrt{0,76(1-0,76)/58}$ = (0,15; 0,37), so that the population value of *g* can also have a medium effect with a probability of 95%.

(c) Effect size of OR_M: OR_M = 44/14 = 3,1, which indicates a large effect according to the guideline values above. The approximate 95% *CI* for $ln(OR_M)$: $ln(3{,}1) \pm 1.96\sqrt{\frac{1}{44} + \frac{1}{14}}$ = (0,82; 1,73), so that die 95% *CI* for approximate OR_M is: $(e^{0{,}82}, e^{1{,}73}) = (2{,}3; 5{,}6)$; the population effect thus tends towards medium with a 95% probability.

2.4 Dependent sets of dichotomous measurements

Where dependent pairs of dichotomous measurements led up to McNemar's test with accompanying effect sizes, we now view the case of sets of dependent dichotomous measurements of three or more on a unit (block, person, object etc.).

Example 6:

Suppose there are 10 persons who have to try and solve three logical problems (A, B and C) and then obtain the correct (indicated by 1) or wrong (indicated by 0) solution. The data are rendered in the first four columns of the following table:

Person	A	B	C	L_i	L_i^2
1	0	0	1	1	1
2	1	1	1	3	9
3	1	1	0	2	4
4	0	0	1	1	1
5	1	1	1	3	9
6	1	0	1	2	4
7	1	1	1	3	9
8	0	1	0	1	1
9	0	0	0	0	0
10	0	0	1	1	1
Total	5	5	7	17	39

The Cochran Q-test is used to test the hypothesis of equal proportions of correct solutions for the logical problems A, B and C (see Siegel, 1956).

Let G_j be the *j-th* number of successes over *n* units (in the example, the number of correct solutions over the 10 persons) of measurement *j,* and *k* the number of measurements (in the example, there are three logical problems); let L_i be the total number of successes for unit *i.* The test statistic is:

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} G_j^2 - \left(\sum_{j=1}^{k} G_j\right)^2\right]}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2}. \qquad (33)$$

To test the null hypothesis of equal proportions for *k* measurements, the fact is used that under the null hypothesis, *Q* is approximately chi-square distributed with *k-1* degrees of freedom for *n* large.

In the example, $Q = \frac{(3-1)\left[3(25+25+49)-(5+5+7)^2\right]}{3\,(17)-39}$ = 16/12 = 1,33 (p = 0,51).

Thus, the hypothesis of equal proportions is not rejected.

2.4.1 Effect size

With *Q* approximately chi-square distributed, the following effect size is proposed (the same as in subsections 2.1, 2.2 and 2.3):

$$w_Q = \sqrt{\frac{Q}{N}}, \qquad (34)$$

where *N* is the population or sample size, depending whether the data are based on a complete population or a random sample drawn from it.

Here, the guideline values of Cohen (1988) can also be applicable, namely

- small effect: $w_Q = 0{,}1$;
- medium effect: $w_Q = 0{,}3$;
- large effect: $w_Q = 0{,}5$.

In Example 6, $w_Q = \sqrt{\frac{1{,}33}{10}}$ = 0,36, which is a medium effect.

2.4.2 Confidence interval for population w_Q

Similar to the previous sections, an approximate *CI* can be calculated by using SAS or Excel (see subsection 2.1). However, in Example 6, it becomes clear that *n* = 10 is too small to use this method. We thus give another example:

Example 7:
In a sample of 405 respondents, three different markers, indicating risk of stress-related illnesses, were compared. All three markers were applied to each respondent. It was

determined at each marker whether respondents were at risk. The result of Cochran's Q-test was: Q = 35,02 (p < 0,001).

With n = 405 and k = 3, it then holds true that: $w_Q = \sqrt{\frac{35,02}{405}}$ = 0,29 and 95% *CI*: (0,19; 0,39).

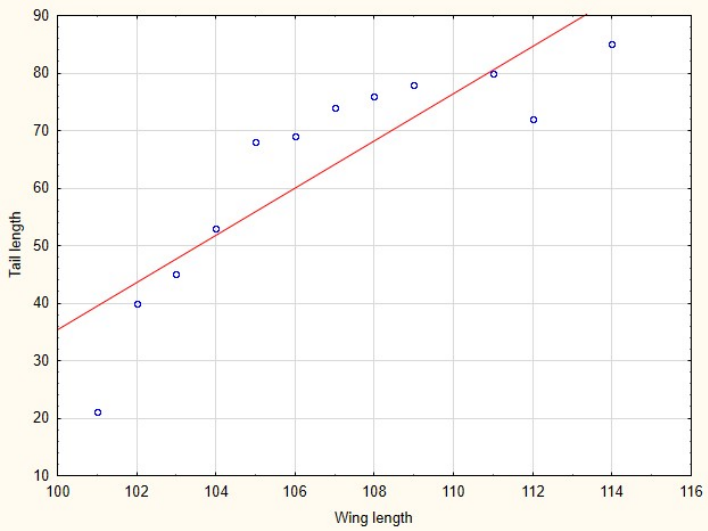
Therefore, although the difference of the three markers' proportions of risks is highly significant (p < 0,001), it is only a medium effect with effect size w_Q = 0,29. The real effect of the population lies with a 95% probability between 0,19 and 0,39, which could even point to a small effect.

3. Relationships between dependent variables

In Chapter 5 (Steyn, 2012), the author viewed linear relationships between two continuous variables whose degree of relationship is measured by the Pearson correlation coefficient. The underlying assumption was that these variables have a bivariate normal distribution, with correlation coefficient ρ. In cases where normality does not necessarily holds true and/or the variables are discrete or ordinal and the relation non-linear, the Pearson correlation is not always a good measure of relationship. However, there exist various measures based on ranks, of which the rank correlation coefficients of Spearman (r_s) and Kendall's tau (τ) are going to be discussed here; they can also be used as effect sizes. When more than two variables are under discussion, we discuss Kendall's coefficient of concordance (W) (which is also based on ranks) as measure of relationship.

3.1 Spearman's rank correlation

This is the correlation between the ranks of two continuous (or interval-scaled) variables and is a measure of monotone association. The relationship between, for example, tail and wing lengths of birds is very clear if one views the following scatter plot:



Although the least squares straight line indicates otherwise on the diagram, there is more probably a monotone ascending relation with the Spearman rank correlation (r_s = 0,93) as measure. The Pearson correlation of r = 0,87 would provide the measure of linear relationship if this would be assumed.

Suppose X and Y are continuous variables, with R_X and R_Y as their ranks. The Pearson correlation between R_X and R_Y gives the Spearman rank correlation r_s. A simple formula that can be derived from this, is:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad (35)$$

where $d_i = R_{X_i} - R_{Y_i}$.

Example 8 (Wikipedia):

Ten children's IQ and hours/week in front of the TV are determined:

IQ (X)	Hours TV (Y)	R_X	R_Y	d	d^2
86	2	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16

110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36
				$\sum d^2$:	194

Now, $r_s = 1 - \frac{6\times194}{10(10^2-1)}$ = 1 – 1,176 = -0,176.

Because of the unexpected values of the first child (with an IQ of 86 and Hours of TV of 2), there is almost no linear relationship (*r* = -0.07), whereas there exists a good negative relationship (*r* = -0,62), as well as a good monotone descending relationship (r_s = -0,62), without that child.

3.1.1 Effect size

Because r_s measures monotone associations, which are a generalisation of linear relationships, r_s can be viewed as an effect size of monotone associations, similar to that of *r* for linear relationships.Therefore, the guideline values of Cohen (1988) for correlation *r* of 0,1 (small effect), 0,3 (medium effect) and 0,5 (large effect) are proposed here too.

In the example of the birds' tail length versus their wing lengths, r_s = 0,93 was a large effect and in Example 10, r_s = -0,18 was a small effect.

3.1.2 Confidence interval for population ρ_s

With the Pearson correlation coefficient *r,* the following transformation was used (see Chapter 5, Steyn, 2012):

$$F(r) = \frac{1}{2} ln \frac{1+r}{1-r}\,, \qquad (36)$$

which means that for larger values *n, F(r)* is approximately normal with a mean of *F(ρ)* and variance of *1/(n - 3)*, and with *ρ* the population correlation coefficient.

However, with Spearman's rank correlation r_s *,* it holds true (see Fieller et.al., 1957) that for *n* large,

$\sqrt{\frac{n-3}{1,06}} F(r_s)$ has an approximate *N(0; 1)* distribution. Then, the approximate *100(1- α)% CI* for $F(\rho_s)$ is

$$F(r_s) \pm z_{\alpha/2} \sqrt{\frac{1,06}{n-3}}\,. \qquad (37)$$

Take $F(r_{sL})$ and $F(r_{sU})$ as the lower and upper bounds of the CI in (37); then, the *100(1- α)% CI* for ρ_s becomes $(r_{sL}; r_{sU})$,

with $r_{sL} = \frac{e^{2F(r_{sL})}-1}{e^{2F(r_{sL})}+1}$ and $r_{sU} = \frac{e^{2F(r_{sU})}-1}{e^{2F(r_{sU})}+1}$. (38)

Example 8 (continued):

In Example 8, r_s = -0,176 and n = 10, so that the approximate 95% *CI* for ρ_s is obtained as follows: $F(-0{,}176) = \frac{1}{2} ln \frac{1-0{,}176}{1+1{,}176}$ = -0,178,

so that $F(r_{sL}) = -0{,}178 - 1{,}96\sqrt{\frac{1{,}06}{7}}$ = -0,940 and similarly, $F(r_{sU})$ = 0,585. Then,

$r_{sL} = \frac{e^{2(-0{,}940)}-1}{e^{2(-0{,}940)}+1}$ = -0,74 and $r_{sU} = \frac{e^{2(0{,}585)}-1}{e^{2(0{,}585)}+1}$ = 0,53.

As a result of the very small sample, the 95% *CI* for ρ_s is very wide. With a sample size of 100, the interval would have narrowed to: (-0,36; 0,03).

3.2 Kendall's tau

This is another measure of relationship and is (similar to Spearman) based on ranks. The original measurements can also be continuous again, but not necessarily normally distributed. The requirement is (similar to Spearman) that the measurements must be at least ordinal to obtain ranks from them. Siegel (1956) explains in the following way how to proceed when calculating this measure:

Suppose two judges A and B had to arrange the same four articles from 1 to 4. The ranks were the following:

Judge	Article 1	Article 2	Article 3	Article 4
A	3	4	2	1
B	3	1	4	2

Next, arrange them according to Adjudicator A:

Judge	Article 4	Article 3	Article 1	Article 2
A	1	2	3	4
B	2	4	3	1

Now, begin with the first rank of B, namely 2, and count how many articles to the right of it have larger ranks (2 larger: 4 and 3) and how many have smaller ranks (1 smaller: 1). Do the same with the second rank to the right of 2, namely 4. There are no ranks larger than 4, but there is 1 smaller (1). With the following rank, namely 3, there is only 1 smaller (1).

Total larger = 2 + 0 + 0 = 2.

Total smaller = 1 + 2 + 1 = 4.

The difference of larger totals and smaller totals = -2. Indicate it with *S*.

The maximum value that *S* can assume is when A and B award the same ranks, that is, B also awards 1, 2, 3 and 4. There are 3 counts (2, 3 and 4) larger than 1; 2 counts (3 and 4) larger than 2; and there is only 1 count (4) larger than 3. Counts smaller than 1, 2 and 3 are 0, so that *S* = 3 + 2 + 1 – (0 + 0 + 0) = 6.

The ratio *S* / 6 now provides a measure of relationship between the ranks awarded by A and B, namely -2 / 6 = -0,33, known as Kendall's tau (τ).

In general, with *n* entities (e.g., persons, or articles as in the example) awarded ranks with regard to two variables (e.g., aspects, judges as in the example, or tests), Kendall's τ becomes:

$$\tau = \frac{Total\ larger - Total\ smaller}{n(n-1)/2} = \frac{2S}{n(n-1)}. \qquad (39)$$

Take note that *S's* maximum value with *n* ranks can be obtained as follows: In the row of ranks *1, 2, 3, ..., n,* there are *(n-1)* ranks larger than *1, (n-2)* ranks larger than *2, ...,* and *1* larger than *(n-1)*, which thus gives a total of

(n-1) + (n-2) +.... + 1 = n(n-1)/2.

Example 9 (Siegel, 1956)

Suppose 12 persons' counts of 'Strive for Status' and 'Authoritarianism' are known. The relationship between the two aspects can be determined from the ranks awarded to each of the aspects. Without indicating the original count for each person, the following table already provides the ranks of 'Strive for Status' (X) 1 to 12, together with the 'Authoritarianism' (Y) ranks:

X	1	2	3	4	5	6	7	8	9	10	11	12
Y	1	5	2	6	7	3	4	10	11	8	9	12

Total larger = 11 + 7 + 9 + 6 + 5 + 6 + 5 + 2 + 1 + 2 + 1 = 55.

Total smaller = 0 + 3 + 0 + 2 + 2 + 0 + 0 + 2 + 2 + 0 + 0 = 11.

S = 55 – 11 = 44 and *n* = 12, so that $\tau = \frac{2\times44}{12\times11}$ = 0,67.

Because τ is a different measure of relationship from Spearman's rank correlation r_s, the values will differ. For Example 12, r_s = 0,82. Because r_s is the Pearson correlation of two rank sets, its value could be interpreted similarly to the Pearson *r*. The question, though, is: How is the value of τ interpreted? In the same way as with other correlation measures, it holds true that $-1 \leq \tau \leq 1$ and near 0-values indicate no relationship, near -1 a good negative relationship, and near 1 a good positive relationship. According to Daniels (1950), it holds true that:

$(2r_s - 1)/3 \leq \tau \leq (2r_s + 1)/3.$

It is thus difficult to interpret τ in terms of r_s.

3.2.1 Effect size

For *n* > 10, it holds approximately true (Siegel, 1956) that for τ the population value and $\hat{\tau}$ the sample value:

$\hat{\tau}$ has approximately an $N(\tau, \sigma_\tau^2)$ distribution with

$$\sigma_\tau^2 = \frac{2(2n+5)}{9n(n-1)}. \qquad (40)$$

Now, the effect size is similar to that of *w* in subsections 2.1 to 2.3, which of the form $w = \sqrt{\frac{X^2}{n}}$, where X^2 chi-square is distributed with 1 degree of freedom. Thus, $X = Z$ is standard normally distributed, so that it is the same as $r = \frac{Z}{\sqrt{n}}$. Now, take $Z = \frac{\hat{\tau}}{\sigma_\tau^2}$, then the effect size

$$r_\tau = \frac{\hat{\tau}}{\sqrt{n\sigma_\tau^2}} = \frac{3\,\hat{\tau}\sqrt{n-1}}{\sqrt{2(2n+5)}}. \qquad (41)$$

The interpretation of r_τ is similar to that of *w*, namely r_τ = 0,1 – small effect; r_τ = 0,3 – medium effect; and r_τ = 0,5 – large effect.

Example 9 (continued):

Here, $\hat{\tau} = 0{,}67$ and *n* = 12, so that $r_\tau = \frac{3 \times 0{,}67\sqrt{11}}{\sqrt{2(2\times 12+5)}}$ = 0,88, which is a large effect.

3.2.2 Confidence interval for population r_τ

From the fact that for larger *n*, $\hat{\tau}$ had approximately an $N(\tau, \sigma_\tau^2)$ distribution, follows that the approximate *100(1- α)% CI* is given for the population r_τ by:

$r_\tau \pm z_{\alpha/2} / \sqrt{n}$. (42)

Example 9 (continued):

The 95% CI for population r_τ is $0{,}88 \pm 1{,}96/\sqrt{12}$ = (0,31; 1,45), which is very wide as a result of the small sample and entails that the effect size can, with a 95% probability, even be of medium effect.

3.3 Kendall's coefficient of concordance

In the previous two sections, measures of relationship between pairs of interval scale or ordinal measures were viewed by using ranks. However, if three or more repeated measurements per block or unit (i.e., a person or object) are determined, Kendall's coefficient of concordance gives a measure of relationship between the measurements.

Example 10

Twelve patients receive three treatments each. In the following table, ranks are awarded for the counts of the 12 patients at each treatment. As before, averages of ranks that compete at equal values were awarded (e.g., with Treatment 1, the average rank of the two values 178 was 3,5). The last column (Totals) gives the sum of the ranks in the *jth* row (i.e., the R_j*s*).

Patient	Treatment 1	Ranks 1	Treatment 2	Ranks 2	Treatment 3	Ranks 3	Totals
1	209	5	88	1,5	109	7	13,5
2	412	12	388	11	142	10,5	33,5
3	315	9	451	12	155	12	33
4	389	11	325	9	121	8,5	28,5
5	210	6	126	5	75	5	16

6	136	2	118	4	49	2,5	8,5
7	178	3,5	227	8	101	6	17,5
8	228	7	98	3	49	2,5	12,5
9	240	8	205	7	142	10,5	25,5
10	113	1	88	1,5	45	1	3,5
11	178	3,5	194	6	55	4	13,5
12	321	10	349	12	121	8,5	30,5
						Total	236

Let *S* be the sum of squares of the R_j*'s* deviances from the average of the R_j *'s*. Here, R_j is the sum of the ranks of the *k* repeated measurements allocated to each block (in Example 10 it was patients).

$$S = \sum_{j=1}^{n}(R_j - \frac{\sum_{j=1}^{n} R_j}{n})^2 = (n - 1)\ Var\ (R_i)\ . \qquad (43)$$

The coefficient of concordance is then:

$$W = \frac{S}{k^2(n^3-n)/12}\ , \qquad (44)$$

which takes on a value between 0 en 1, because the denominator of (44) is the maximum value that *S* can attain if there is complete concordance in the ranks of all *k* the repeated ranks.

Example 10 (continued)

The average of the R_j *'s* is 236 / 12 = 19,67, so that
$S = (13,5 - 19,67)^2 + (33,5 - 19,67)^2 + ... + (30,5 - 19,67)^2 = 1130,17.$

$W = \frac{1130,17}{3^2(12^3-12)/12}$ = 1130,17 / 2717 = 0,88.

3.3.1 Interpretation of *W*

Suppose the average of the Spearman rank correlations of all possible pairs of repeated measurements is indicated by r_{sav}. Then, according to Siegel (1956), there exists the following linear relation between *W* and r_{sav} :

$$r_{sav} = \frac{kW-1}{k-1},$$

so that $W = \frac{(k-1)r_{sav}+1}{k}$. (45)

Guideline values for Spearman rank correlations (r_s) are given in subsection 3.1, which could also be used for r_{sav}, seeing that it is based on r_s. Because W is according to (45) a function of r_{sav} and k, guideline values for W can be obtained from Table 6.

Table 6: W as function of r_{sav} and number of repeated measurements k

r_{sav} \ k	2	3	4	5	7	10
0.00	0.50	0.33	0.25	0.20	0.14	0.10
0.10	0.55	0.40	0.33	0.28	0.23	0.19
0.20	0.60	0.47	0.40	0.36	0.31	0.28
0.30	0.65	0.53	0.48	0.44	0.40	0.37
0.40	0.70	0.60	0.55	0.52	0.49	0.46
0.50	0.75	0.67	0.63	0.60	0.57	0.55
0.60	0.80	0.73	0.70	0.68	0.66	0.64
0.70	0.85	0.80	0.78	0.76	0.74	0.73
0.80	0.90	0.87	0.85	0.84	0.83	0.82
0.90	0.95	0.93	0.93	0.92	0.91	0.91
1.00	1.00	1.00	1.00	1.00	1.00	1.00

For increasing values of k and larger values of r_{sav}, the values of W become closer and closer to those of r_{sav} and can thus be interpreted in the same way as r_{sav} or r_s. For $k = 2$, $r_{sav} = r_s$, whose values differ a lot from W for smaller values of r_s. In the table, the guideline values of small, medium and large effect for W are highlighted in grey. In Example 13, W = 0,88 and according to Table 6, this can be viewed as a large effect (that is more or less in concordance with r_{sav} = 0,8).

Take note that with $0 \leq W \leq 1$, the relationship (45) is only valid for $r_{sav} > -1 / (k - 1)$; what it amounts to in practice is that for larger values of k, one can only view positive and smaller negative values of r_{sav}.

3.3.2 Confidence interval for population W

According to Siegel (1956), $k(n-1)\widehat{W}$ (for $n > 7$) has a non-central chi-square distribution with $n - 1$ degrees of freedom and non-centrality parameter $nsp = k(n - 1)W$, with $\widehat{W}$ the estimator of W from a random sample. As in subsection 2.2.2 for w, it is now possible to first determine (by means of a computer program) an approximate $100(1 - \alpha)\%$ CI with lower and upper bounds (L, U) for nsp; then, W's CI is given by $(\frac{L}{k(n-1)}, \frac{U}{k(n-1)})$. Once again, it can be calculated by means of SAS's *CI_w* or with the Excel spreadsheet (as in subsection 2.2).

Example 10 (continued):

Here, $\widehat{W}$ = 0,88, k = 3 and n = 12; it follows that the degrees of freedom = 11, so that the 95% CI for nsp: (3,94; 42,33), from which the CI for W follows as (0,12; 1,28). The upper bound can be taken as 1,0, which still gives a very wide interval. With a 95% probability, W can be as low as 0,12 – thus a small effect.

4. Two groups compared with interval and ordinal scale measurements

In Steyn (2012: Chapter 4), we have already viewed the parametric effect sizes that are based on standardised differences between two means. However, there are also effect sizes that can be used after non-parametric tests have been applied. Usually, non-parametric tests are applied to interval scale measurements when normality of data cannot be assumed and samples are small. With ordinal measurements, it is also more appropriate to use non-parametric methods when dealing with small samples. First, the case of two independent groups is viewed and thereafter, that of dependent groups.

4.1 Two independent groups – the Mann-Whitney test

This test is based on the statistic U: the number of observations from the first group (A) (population or a sample thereof), which is smaller than the number of observations from the second group (B). The value of U is usually obtained by the execution of packages such as SPSS, Statistica or SAS. The Mann-Whitney (or Wilcoxon two-sample) statistic is then:

$$Z = \frac{U - m_U}{s_U}, \qquad (46)$$

where $m_U = \frac{n_A n_B}{2}$, n_A and n_B are the group sizes, and

$$s_U = \sqrt{\frac{n_A n_B (n_A + n_B + 1)}{12}}. \qquad (47)$$

If the smallest of n_A and $n_B > 20$, it holds true that under the null hypothesis of populations A and B having the same distribution, *Z* is approximately standard normally distributed.

4.1.1 Effect sizes

(a) $p_{A,B} = \frac{U}{n_A n_B}$, (48)

that is, the probability that a randomly chosen observation from population B will be larger than a randomly chosen observation from population A. If working with samples, $p_{A,B}$ becomes the proportion $\hat{p}_{A,B}$ and this serves as an estimator for the probability (see Pautz et.al., 2018). In Steyn (2012: subsection 8.5), $p_{A,B}$ is given as *AUC*, the area under the ROC curve.

(b) According to Pautz et.al. (2018),

$r = \frac{|Z|}{\sqrt{n_A+n_B}}$. (49)

The interpretation of *r* is the same as *w* (see subsection 2.2), because Z^2, based on $n_A + n_B$ observations, has a chi-square distribution with 1 degree of freedom, so that $r = \sqrt{\frac{Z^2}{n_A+n_B}}$.

Thus, as with *w,* we take *r* = 0,1 (small effect), *r* = 0,3 (medium effect), and *r* = 0,5 (large effect) as guidelines again.

4.1.2 Confidence intervals in samples

(a) According to Pautz et.al. (2018), an approximate *100(1- α)% CI* for $p_{A,B}$ is given by:

$\hat{p}_{A,B} \pm z_{\alpha/2}\sqrt{V(\hat{p}_{A,B})}$, (50)

where

$V(\hat{p}_{A,B}) = \hat{p}_{A,B}(1-\hat{p}_{A,B})\left[1 + \frac{n_A^*(1-\hat{p}_{A,B})}{2-\hat{p}_{A,B}} + \frac{n_B^*\hat{p}_{A,B}}{1+\hat{p}_{A,B}}\right]/n_A n_B,$ (51)

with $n_A^* = n_A - 1$ and $n_B^* = n_B - 1$.

(b) An approximate *100(1- α)% CI* for population *r* :

$r \pm \frac{z_{\alpha/2}}{\sqrt{n_A+n_B}}$. (52)

The Excel spreadsheet can also be used to calculate these effect sizes as well as their confidence intervals.

Example 11

Two random samples with a size of 12 persons each were taken from populations A and B; their counts for a test were then determined. The table underneath provides the counts:

A: 40	30	25	29	37	43	25	27	30	35	39	42
B: 44	41	34	35	40	44	39	39	45	44	46	32

To determine *U,* each value in group B is compared with each one in group A (or the other way round); the values that are smaller are then counted. In this way, for example, the first value of Group B, namely 44, is not smaller than any other counts in group B and the second value, 41, is smaller than 42 and 43. All the 12 x 12 combinations are examined similarly. This produces *U* = 29, so that

$\hat{p}_{A,B} = \frac{29}{12 \times 12}$ = 0,2, whereas

$V(\hat{p}_{A,B}) = (0{,}2 \times 0{,}8)[1 + 11 \times 0{,}8\ /1{,}8 + 11 \times 0{,}2\ /\ 1{,}2\ ]/144$ = 0,0086.

Thus, the approximate 95% *CI* for $p_{A,B}$: $0{,}2 \pm 1{,}96\sqrt{0{,}0086} = (0{,}02; 0{,}38)$.

The approximate probability that the counts of population B are smaller than those of population A is 0,2, but can vary between more or less 0 and 0,4 with a 95% probability.

(b) $Z = \frac{|29-72|}{\sqrt{12 \times 12 \times 25/12}} = \frac{43}{17{,}3} = 2{,}49,$

with $r = \frac{2{,}49}{\sqrt{24}} = 0{,}51,$

and 95% *CI* for population value of *r* : $0{,}51 \pm \frac{1{,}96}{\sqrt{24}} = (0{,}11; 0{,}91)$.

The population *r* is estimated at 0,51 (large effect), but as a result of small samples, it can vary between 0,11 and 0,91 with a 95% probability.

4.2 Dependent groups – The paired samples Wilcoxon test

Here, interval or ordinal scale observations are obtained on dependent pairs, for example on persons before and after a certain intervention. Consider the following example from the physiotherapy field:

Example 12 (Pautz et.al., 2018):

In a sample of 29 persons, measurements were taken on each one's L4[right] and L4[left] muscles. The first three columns of the following table render the measurements:

Person	L4 [right]	L4 [left]	Differences	Positive sign (1) / Negative sign (0)	Rank positives
1	601	592	9	1	7
2	915	984	-69	0	
3	651	670	-19	0	
4	626	718	-92	0	
5	754	743	11	1	8
6	673	654	19	1	10.5
7	678	679	-1	0	
8	769	776	-7	0	
9	885	736	149	1	28
10	659	805	-146	0	
11	694	585	109	1	22
12	860	750	110	1	23
13	793	801	-8	0	
14	796	800	-4	0	
15	918	917	1	1	2
16	642	641	1	1	2
17	979	1090	-111	0	
18	963	935	28	1	13
19	738	821	-83	0	
20	780	605	175	1	29
21	740	835	-95	0	
22	829	948	-119	0	
23	324	373	-49	0	
24	868	988	-120	0	
25	690	648	42	1	14
26	564	661	-97	0	
27	587	602	-15	0	
28	461	439	22	1	12
29	860	787	73	1	17

The differences between the two measurements on each person are indicated in the fourth column, whereas the sign of the difference is indicated in column 5 by 1 (positive) and 0

(negative). The absolute differences (thus without the sign) are arranged from small to large and ranks 1 to 29 are awarded to them (at equal values or ties, the average of ranks competing for them is used, e.g., the values 19 that compete for ranks 10 and 11 both receive 10,5 as rank). Column 6 only gives the ranks of positive differences.

The number of persons with positive differences (i.e., the total of column 5) is 13. To test the null hypothesis that in the population, half of the persons' L4[right] values are larger than their L4[left] values, the sign test with *n* = 29 and number of successes = 13 can be used. On the example's data, the null hypothesis cannot be rejected against the one-sided alternative that the proportion is smaller than 0,5 (p = 0,13).

The test statistic of the **paired samples Wilcoxon test** is:

$$Z = \frac{S_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}, \qquad (53)$$

with S_+ = sum of the positive ranks.

Example 12 (continued):

Now, S_+ =187,5 (total of column 6 in the table above, so that

$Z = \frac{187{,}5 - 29 \times 30/4}{\sqrt{29 \times 30 \times 59/24}}$ = -0,65 (*p* = 0,26, one-sided).

4.2.1 Effect sizes

(a) According to Pautz et.al. (2018), the probability that the first measurement is larger than the second one becomes

$$PS_{dep} = \frac{n_+}{n_*}, \qquad (54)$$

where n_+ gives the number of positive differences between the first and second measurements, whereas n_* gives the number of pairs with differences (0-differences are thus excluded).

(The notation is due to Grissom & Kim, 2012, who use PS_{dep} to indicate "**P**robability of **S**uperiority of **dep**endent pairs".)

Example 12 (continued):

Because it is a sample drawn from a population, the estimated probability that a person's L4[right] value is larger than the L4[left] value is the following:

$\widehat{PS}_{dep} = \frac{13}{29}$ = 0,45.

(b) As with the Mann-Whitney test, the normal approximation of the Wilcoxon test (usually when *n* > 25) can be used to give the following effect size:

$$r = \frac{|Z|}{\sqrt{n}}. \qquad (55)$$

Example 12 (continued):

Now, *Z* = -0,65, so that $r = \frac{0,65}{\sqrt{29}}$ = 0,12, which gives a small effect. This compares well with the interpretation of the estimated probability $\widehat{PS}_{dep}$ = 0,45, which does not differ much from 0,5.

4.2.2 Confidence intervals in samples

(a) With *n* large, the usual normal approximation of proportions holds true and the

100(1- α)% CI for PS_{dep} is:

$$\widehat{PS}_{dep} \pm z_{\alpha/2}\sqrt{\frac{\widehat{PS}_{dep}\,(1-\widehat{PS}_{dep})}{n_*}}. \qquad (56)$$

The Excel spreadsheet can also be used to calculate these effect sizes as well as their confidence intervals.

Example 12 (continued):

The 95% for PS_{dep}, thus $0{,}45 \pm 1{,}96\sqrt{\frac{0,45(1-0,45)}{29}}$ = 0,45 ± 0,18 = (0,27; 0,63). Therefore, no finding can be made concerning the population probability (i.e., that the L4[right] value is larger than the L4[left] value) since it lies between 0,27 and 0,63 with a 95% probability (here, lying around 0,5).

(b) The *100(1- α)% CI* for the population value of *r* is (as in subsection 4.1) the following for *n* large:

$$r \pm \frac{z_{\alpha/2}}{\sqrt{n_*}}. \qquad (57)$$

Example 12 (continued):

The 95% CI for the population value of *r* :

$0{,}12 \pm \frac{1{,}96}{\sqrt{29}}$ = (-0,24; 0,48); take it as (0; 0,48), because negative values of *r* does not make sense. The effect size of the population value of *r* can thus vary with a 95% probability between 0 and 0,48, which is meaningless and does not indicate any effect.

5. More than two groups compared

As with two groups, non-parametric tests can be applied to compare three or more groups. This is usually done when groups are small and interval scale (continuous) measurements are not necessarily normally distributed. Such tests are also applied to small groups with ordinal or discrete measurements. As in Section 4, we distinguish between independent and dependent groups again.

5.1 Independent groups – the Kruskal-Wallis test

Example 13 (Siegel,1956)

Three groups (types) of teachers are measured on an authoritarianism scale (*A*-scale). The following table gives the *A*-scale values, as well as the ranks, of each group. Take note that ranks are awarded to the pooled data.

Educationally oriented (Group 1)	Ranks of Group 1	Administratively oriented (Group 2)	Ranks of Group 2	Administrative teachers (Group 3)	Ranks of Group 3
96	4	82	2	115	7
128	9	124	8	149	13
83	3	132	10	166	14
61	1	135	11	147	12
101	5	109	6		
	$R_1 = 22$		$R_2 = 37$		$R_3 = 46$

Where a one-way analysis of variance (ANOVA) tests whether the population from which random samples are drawn has the same means (see Steyn, 2012: Chapter 6), the null hypothesis will now be tested to see whether the probability is 50% that a randomly chosen teacher from Population 1 has a larger value (on the *A* -scale) than one who was chosen from Population 2; the same applies to Population 2 versus 3 and 1 versus 3.

Here, the **Kruskal-Wallis test** is used with test statistic

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1), \qquad (58)$$

where n_i is the *i-th* group from *k* groups' sample size and $N = \sum_{i=1}^{k} n_i$, whereas R_i is the sum of ranks in the *i-th* group.

Example 13 (continued):

$n_1 = n_2 = 5$ and $n_3 = 4$, so that $N = 14$. From the table, it follows that: $R_1 = 22$, $R_2 = 37$ and $R_3 = 46$, so that

$$H = \frac{12}{14 \times 15}\left[\frac{22^2}{5} + \frac{37^2}{5} + \frac{46^2}{4}\right] - 3 \times 15 = 6{,}4.$$

For large samples under the null hypothesis, *H* is approximately chi-square distributed with $k - 1$ degrees of freedom, so that $P = 0{,}04$; consequently, the null hypothesis is rejected on a significance level of 5%.

5.1.1 Effect sizes

(a) The following is represented similarly to the effect sizes of Section 2:

$$w_H = \sqrt{\frac{H}{N}}\ , \qquad (59)$$

with interpretation as before, namely

- small effect: $w_H = 0{,}1$;
- medium effect: $w_H = 0{,}3$;
- large effect: $w_H = 0{,}5$.

Example 13 (continued): With $H = 6{,}4$ and $N = 14$, the effect size becomes

$w_H = \sqrt{\frac{6{,}4}{14}} = 0{,}68$, which can be considered to be a large effect.

(b) As with the one-way ANOVA (see Steyn, 2012: Chapter 6), there also exists an eta-square here (η^2), the proportion of the total variance of the dependent variable that can be ascribed to the independent variable. In Example 13, it is the proportion of the variance of the measurements (in the *A*-scale) that can be ascribed to the three groups of teachers. According to Tomczak and Tomczak (2014), this effect size is

$$\eta_H^2 = \frac{H-k+1}{N-k}\ . \qquad (60)$$

Steyn (2012: Chapter 6) states that then, the guideline values for η_H^2 is:

- small effect: $\eta_H^2 = 0{,}01$;
- medium effect: $\eta_H^2 = 0{,}06$;
- large effect: $\eta_H^2 = 0{,}14$.

Example 13 (continued): $\eta_H^2 = \frac{6{,}4-3+1}{14-3} = 0{,}4$, a large effect – 40% of the total variance of the *A*-scale measurements can be ascribed to the three groups.

5.1.2 Confidence interval for population w_H

For large samples, the approximation of the chi-square distribution with *k – 1* degrees of freedom for *H* is good. As before (see, e.g., subsection 2.2), a *100(1- α)% CI* can be determined by using the non-central chi-square distribution with the aid of the Excel program.

In Example 13, it is not possible to obtain a lower boundary for the confidence interval, as the samples are so small. Consider therefore another example:

Example 14:

For three samples of 12 persons each, the totals of ranks with regard to a certain aspect were determined as $R_1 = 139$, $R_2 = 200$ and $R_3 = 327$; $N = 12 \times 3 = 36$, so that

$$H = \frac{12}{36 \times 37}\left[\frac{139^2}{12} + \frac{200^2}{12} + \frac{327^2}{12}\right] - 3 \times 37 = 13{,}8 \ (P = 0{,}001).$$

$w_H = \sqrt{\frac{13{,}8}{36}} = 0{,}62$, whereas the 95% *CI* proves to be (0,26; 0,93) from calculations by the Excel program. This means there are highly significant differences between the groups with a large effect; however, with a 95% probability, they can also be as low as 0,26 – thus, a medium effect.

5.2 Dependent groups – repeated measurements: the Friedman test

As with the paired samples test of Wilcoxon (in subsection 4.2) with pairs of observations on blocks (e.g., persons or objects), the case of three or more repeated measurements on each block is now considered. When the number of blocks (or persons) are few and normality of interval scale or ordinal measurements does not necessarily hold true, ranks are determined at each block and the Friedman test is applied to test the null hypothesis that each of the repeated measurements on the blocks comes from the same population.

Example 15

In Example 10, subsection 3.3, 12 patients were subjected to three treatments (repeated measurements) and the ranks thereof on each of the 12 patients (blocks) are given in the following table:

Patient	Treatment 1	Ranks 1	Treatment 2	Ranks 2	Treatment 3	Ranks 3
1	209	3	88	1	109	2
2	412	3	388	2	142	1
3	315	2	451	3	155	1
4	389	3	325	2	121	1
5	210	3	126	2	75	1
6	136	3	118	2	49	1
7	178	2	227	3	101	1
8	228	3	98	2	49	1
9	240	3	205	2	142	1
10	113	3	88	2	45	1
11	178	2	194	3	55	1
12	321	2	349	3	121	1
Totals	-	32	-	27	-	13

The **Friedman test statistic** for testing the null hypothesis is:

$$F = \frac{12}{nk(k+1)} \sum_{i=1}^{k} K_i^2 - 3n(k+1), \qquad (61)$$

where k is the number of repeated measurements and n the number of blocks, whereas K_i is the *i-th* (repeated) measurement's total of ranks (i.e., the sum of ranks in the *i-th* treatment in Example 15).

For n large, it holds true that under the null hypothesis F is approximately distributed according to a chi-square distribution with $k - 1$ degrees of freedom.

Example 15 (continued):

The three totals of the treatments' ranks are $K_1 = 32$, $K_2 = 27$; $K_3 = 13$, $k = 3$; and $n = 12$, so that

$$F = \frac{12}{12 \times 3 \times 4}(32^2 + 27^2 + 13^2) - 3 \times 12 \times 4 = 16{,}17.$$

The null hypothesis is rejected ($P < 0{,}001$), so that the population distributions of the three treatments are different.

5.2.1 Effect size

The following effect size is used, similar to previous sections where the test statistics under the null hypotheses are approximately chi-square distributed:

$$w_F = \sqrt{\frac{F}{n}}\,. \qquad (62)$$

Then, the guideline values for the interpretation of w_F is $w_F = 0{,}1$ – small effect; $w_F = 0{,}3$ – medium effect; and $w_F = 0{,}5$ – large effect.

Example 15 (continued): $w_F = \sqrt{\frac{16{,}17}{12}} = 1{,}17$, a very large effect.

5.2.2 Confidence interval for population w_F

The chi-square statistic *F* has approximately a non-central chi-square distribution with *k - 1* degrees of freedom and non-centrality parameter of $nsp = nw_F$. As in subsection 2.1.4 for *φ*, it is now possible to determine (by means of a computer program) an approximate

100(1 – α)% CI with lower and upper bounds (*L, U*) for *nsp;* then, the *CI* for the bounds of w_F is $(\sqrt{L/n}\,, \sqrt{U/n})$. The Excel spreadsheet can be used.

Example 15 (continued): With *F* = 16,17, degrees of freedom = 2 and *n* = 12, *CI_w* gives the 95% *CI* for *nsp* as (3,53; 34,55), from which follows the *CI* for w_F as (0,54; 1,70). With a 95% probability, even the lower bound of 0,54 is a large effect.

References:

Burnand, B., Kernan, W.N. & Feinstein, A.R. (1990). Indexes and boundaries for quantative significance in statistical decisions. *Journal of Clinical Epidemiology*, 43, 1273 - 1284.

Cohen, J. (1969). *Statistical Power Analysis*. Academic Press, Inc., Orlando.

Cohen, J. (1977). *Statistical Power Analysis. Revised Edition*. Academic Press, Inc., Orlando.

Cohen, J. (1988). *Statistical Power Analysis. Second Edition*. Academic Press, Inc., New York.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19, 3127 - 3131.

Daniels, H.E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society (B)*, 12(2), 171 – 191.

Fieller, E.C., Hartley, H.O. & Pearson, E.S. (1957).Tests for rank correlation coefficients I. *Biometrika* 4(3-4), 407 – 481.

Fleiss, J.L. (1994). Measures of effect size for categorical data. In: *The handbook of research synthesis* (eds. Cooper, H & Hedges L.V.), p. 245-260. Russel Sage Foundation, New York.

IBM SPSS Statistics (2020), Version 26, Copyright© IBM Corporation and its licensors. http://www-01.ibm.com/software/analytics/spss/

Grissom, R.J. & Kim, J.J. (2005). *Effect sizes for research: a broad practical approach.* Lawrence Erlbaum Associates, New York

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). *Continuous univariate distributions, Volume 2 (second edition)*. John Wiley, New York.

Kline, R.B. (2004a). *Beyond Significance Testing: Reforming data analysis methods in behavioural research.* American Psychological Association. Washington, DC.

Newcombe, R.G. (2006a). Confidence intervals for an effect size measure based on the Mann-Whitney statistic I. *Statistics in Medicine*, 25:543 - 557.

Newcombe, R.G. (2006b). Confidence intervals for an effect size measure based on the Mann-Whitney statistic II. *Statistics in Medicine*, 25:559 - 573.

Olivier, J, May, W.L. & Bell, M.L. (2017). Relative effect sizes for measures of risk. *Communications in Statistics – Theory and Methods*, 46(14), 6774 – 6781.

Pautz Nikolas, Olivier Benita, Steyn Faans. 2018. The use of nonparametric effect sizes in single study musculoskeletal physiotherapy research: A practical primer. *Physical Therapy in Sport*, 33, 117 – 124. www.elsevier.com/ptsp

Rosenthal, R, Rosnow, R.L & Rubin, D.B. (2000). *Contrast and effect sizes in behavioural research: a correlational approach.* New York, Cambridge University Press.

SAS Institute Inc. (2020). The SAS System for Windows Release 9.4 TS Level 1M3 X64_8PRO platform Copyright© by SAS Institute Inc., Cary, NC, USA

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York.

Smithson, M. (2000). *Statistics with confidence*. Sage Publications, London.

Steyn, H.S. jr. (2002). Practically significant relationship between two variables. *S.A. Tydskrif vir Bedryfsielkunde*, 28(3), 10-15.

Steyn, H.S. 2012. *Effect sizes and practical significance.* http://natural-sciences.nwu.ac.za/scs/effect, North West University, Potchefstroom [date when downloaded].

TIBCO STATISTICA (data analysis software system) 2020. Version 13 TIBCO Software Inc. All rights

Tomczak, M and Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*. 1(21), 19-25.

Tritchler, D. (1995). Interpreting the standardized difference. *Biometrics.* 51, 351 - 353