Chapter 811

Kappa Test for Agreement Between Two Raters

Introduction

This module computes power and sample size for the test of agreement between two raters using the kappa statistic. The power calculations are based on the results in Flack, Afifi, Lachenbruch, and Schouten (1988). Calculations are based on ratings for k categories from two raters or judges. You are able to vary category frequencies on a single run of the procedure to analyze a wide range of scenarios all at once. For further information about kappa analysis, see chapter 18 of Fleiss, Levin, and Paik (2003).

Technical Details

Suppose that N subjects are each assigned independently to one of k categories by two separate judges or raters. The results are placed in a $k \times k$ contingency table. Each p_{ij} represents the proportion of subjects that Rater A classified in category i, but Rater B classified in category j, with $i, j = 1, 2, \ldots, k$. The proportions $p_{i.}$ and $p_{.j}$ are the frequencies or marginal probabilities of assignment into categories i and j for Rater A and Rater B, respectively. For each rater, the category frequencies sum to one.

	Rater B				
Rater A	1	2	…	k	Total
1	p_{11}	p_{12}	…	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	…	p_{2k}	$p_{2.}$
⋮	⋮	⋮	⋱	⋮	⋮
k	p_{k1}	p_{k2}	…	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$	…	$p_{.k}$	1

The proportions on the diagonal, p_{ii}, represent the proportion of subjects in each category for which the two raters agreed on the assignment. The overall proportion of observed agreement is

$$p_o = \sum_{i=1}^{k} p_{ii} ,$$

and the overall proportion of agreement expected by chance is

$$p_e = \sum_{i=1}^{k} p_{i.} p_{.i} \ .$$

The overall value of kappa, which measures the degree of rater agreement, is then

$$\kappa = \frac{p_o - p_e}{1 - p_e} .$$

A kappa value of 1 represents perfect agreement between the two raters. A kappa value of 0 indicates no more rater agreement than that expected by chance. A kappa value of -1 would indicate perfect disagreement between the raters.

The true value of kappa can be estimated by replacing the observed and expected proportions by their sample estimates

$$\hat{\kappa} = \frac{\hat{p}_o - \hat{p}_e}{1 - \hat{p}_e} ,$$

where

$$\hat{p}_o = \sum_{i=1}^{k} \hat{p}_{ii}$$

$$\hat{p}_e = \sum_{i=1}^{k} \hat{p}_{i.} \hat{p}_{.i} \ .$$

The minimum possible value of $\hat{\kappa}$ depends on the marginal proportions. If the marginal proportions are such that $\hat{p}_e = 0.5$, then the minimum value is -1. Otherwise, the minimum value is between -1 and 0.

The standard error of $\hat{\kappa}$ is

$$s.e.(\hat{\kappa}) = \frac{\tau(\hat{\kappa})}{\sqrt{N}},$$

where

$$\tau(\hat{\kappa}) = \frac{1}{(1-p_e)^2} \left\{ p_o (1-p_e)^2 + (1-p_o)^2 \sum_{i=1}^{k} \sum_{j=1,\, j \neq i}^{k} p_{ij} \left(p_{j.} + p_{.i} \right)^2 \right.$$
$$\left. - 2(1-p_o)(1-p_e) \sum_{i=1}^{k} p_{ii} \left(p_{i.} + p_{.i} \right)^2 - \left(p_o p_e - 2 p_e + p_o \right)^2 \right\}^{1/2} .$$

Again, an estimate of the standard error can be obtained by replacing the unknown values p_{ij} by their sample estimates $\hat{p}_{ij}$.

Hypothesis Tests

One- and two-sided hypothesis tests can be conducted using the test statistic

$$z = \frac{\hat{\kappa} - \kappa_0}{s.e.(\hat{\kappa})},$$

where κ_0 is the null hypothesized value of kappa, and the denominator is the estimated standard error. For a one-sided alternative, the test rejects H_0 if $|z| \geq z_\alpha$, where z_α is the value that leaves α in the upper tail of the standard normal distribution. For a two-sided alternative, the test rejects H_0 if $|z| \geq z_{\alpha/2}$.

Power Calculation

The standard error for the kappa statistic is based on values p_{ij}, which are unknown prior to conducting a study. Therefore, the power is computed at the maximum standard error based on given category frequencies or marginal probabilities. The following steps are taken to compute the power of the test.

1. Determine the category assignment frequencies for both raters. In practice, the category frequencies may not be equivalent, but the standard error maximization method of Flack, Afifi, Lachenbruch, and Schouten (1988) assumes that the category frequencies are equal for both raters. Therefore, only one set of frequencies is needed.
2. Determine the maximum standard error under the null and alternative hypotheses for the given marginal frequencies. This is equivalent to finding the maximum $\tau(\hat{\kappa})$ under the null and alternative hypotheses.
3. Find the critical value using the standard normal distribution. The critical value, $z_{critical}$, is that value of z that leaves exactly the target value of alpha (or alpha/2) in the upper tail of the standard normal distribution. For example, for an upper-tailed test with a target alpha of 0.05, the critical value is 1.645.
4. Without loss of generality, for a one-sided test of the alternative hypothesis that $\kappa > \kappa_0$, compute the power at an alternative value of kappa, κ_1, as

$$\begin{aligned} 1-\beta &= \Pr\left(z \geq z_{critical} \mid H_1\right) \\ &= 1-\Phi(u) \end{aligned},$$

where $\Phi()$ is the cumulative standard normal distribution and

$$u = \frac{\sqrt{N}\left(\kappa_0 - \kappa_1\right) + z_{critical} \max \tau\left(\hat{\kappa} \mid \kappa = \kappa_0\right)}{\max \tau\left(\hat{\kappa} \mid \kappa = \kappa_1\right)}.$$

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains most of the parameters and options that you will be concerned with.

Solve For

Solve For

This option specifies the parameter to be solved for from the other parameters. The parameters that may be selected are *Power*, *Sample Size*, or *κ1*.

Test

Alternative Hypothesis (H1)

Specify the alternative hypothesis of the test. Since the null hypothesis is the opposite, specifying the alternative is all that is needed. The alternative hypothesis determines how the alternative value(s) of Kappa ($\kappa 1$) should be entered. Usually, the two-sided option is selected.

For a one-sided alternative hypothesis test of $\kappa 1 > \kappa 0$, all values for $\kappa 1$ should be greater than $\kappa 0$.

For a one-sided alternative hypothesis test of $\kappa 1 < \kappa 0$, all values for $\kappa 1$ should be less than $\kappa 0$.

For a two-sided test, the values for $\kappa 1$ can be either greater than or less than $\kappa 0$.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of equal means when in fact the means are different.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. In this procedure, a type-I error occurs when you reject the null hypothesis of equal means when in fact the means are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

You may enter a range of values such as *0.01 0.05 0.10* or *0.01 to 0.10 by 0.01*.

Sample Size

N (Sample Size)

Enter a value for the sample size (N). This is the number of subjects rated by the two judges in the study. You may enter a range such as *10 to 100 by 10* or a list of values separated by commas or blanks such as *50 60 70*.

Kappa and Categories – Kappa

κ1 (Kappa|H1)

This is the value of Kappa under the alternative hypothesis, H1. Values must be between -1 and 1. If a one-sided hypothesis is used, the alternative value(s) for Kappa in relation to the null value(s), κ0, should match the direction of the alternative hypothesis. You can enter a list of values separated by blanks or commas such as *0.6 0.7 0.8* or *0.6 to 0.8 by 0.05*.

κ0 (Kappa|H0)

This is the value of Kappa under the null hypothesis, H0. Values must be between -1 and 1. If a one-sided hypothesis is used, the alternative value(s) for Kappa in relation to the null value(s), κ0, should match the direction of the alternative hypothesis. You can enter a list of values separated by blanks or commas such as *0.2 0.3 0.4* or *0.2 to 0.4 by 0.05*.

Kappa and Categories – Classification Frequencies

Specify using

Select the method used to specify the marginal classification frequencies.

List Input

Specify a single set of marginal frequencies as a list. For example, with three categories you might enter "2 3 5."

Spreadsheet Column Input

Specify more than one set of marginal frequencies (proportions) using the spreadsheet. Each spreadsheet column becomes a set of marginal frequencies. For example, if you have three sets of frequencies in the three columns C1, C2, and C3, you would enter "=C1 C2 C3".

P (Frequencies)

Specify two or more category frequencies, proportions, or marginal proportions. These are the proportions of subjects assigned to each category by the two raters or judges. In practice the proportions may not be exactly equal for the two raters, but the power calculations assume that the category frequencies are equal for the two raters. Each proportion set must sum to 1.

List Input

Specify a single set of proportions as a list. For example, with three groups you might enter *0.2 0.3 0.5*.

Spreadsheet Column Input

Specify more than one set of proportions using the column input syntax

=[column 1] [column 2] etc.

For example, if you have three proportion sets stored in the spreadsheet in columns C1, C2, and C3, you would enter *=C1 C2 C3* in the P (Category Frequencies) box.

Each column in the spreadsheet corresponds to a single set of proportions. The columns may contain different numbers of frequencies, but in all cases, the values in each column must sum to 1.

Example 1 – Finding the Power

Suppose a study is being planned to measure the degree of inter-rater agreement for two psychiatrists. The two psychiatrists will independently classify each of a series of patients into one of three diagnostic categories: personality disorder, neurosis, or psychosis. The study will then determine how well the psychiatrists "agree" with a hypothesis test using the kappa statistic.

Before the data are collected, the organizers would like to study the relationship between sample size and power. From previous experience, they have determined to use frequencies of 0.4, 0.5, and 0.1 for the personality disorder, neurosis, and psychosis diagnoses, respectively. They would like to determine the power for detecting alternative kappa values of 0.5, 0.6, and 0.7 when the null value is 0.4. A two-sided hypothesis test will be conducted at alpha = 0.05. What will be the power for a wide range of sample sizes?

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Kappa Test for Agreement Between Two Raters** procedure window by expanding **Correlation**, then clicking on **Kappa**, and then clicking on **Kappa Test for Agreement Between Two Raters**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Alternative Hypothesis (H1)	**Two-Sided**
Alpha	**0.05**
N (Sample Size)	**30 to 200 by 10**
κ1	**0.5 0.6 0.7**
κ0	**0.4**
Specify Using	**List Input**
P (Frequencies)	**0.4 0.5 0.1**

Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Test Type = Two-sided Z test.
H0: Kappa = κ0 vs. H1: Kappa ≠ κ0.

Power	Sample Size N	Kappa|H0 κ0	Kappa|H1 κ1	Alpha	Beta	- Rating Categories - k	Frequencies
0.07748	30	0.40	0.50	0.05000	0.92252	3	0.40, 0.50, 0.10
0.19421	30	0.40	0.60	0.05000	0.80579	3	0.40, 0.50, 0.10
0.43345	30	0.40	0.70	0.05000	0.56655	3	0.40, 0.50, 0.10
0.09199	40	0.40	0.50	0.05000	0.90801	3	0.40, 0.50, 0.10
0.26055	40	0.40	0.60	0.05000	0.73945	3	0.40, 0.50, 0.10
0.58208	40	0.40	0.70	0.05000	0.41792	3	0.40, 0.50, 0.10
0.10677	50	0.40	0.50	0.05000	0.89323	3	0.40, 0.50, 0.10
0.32746	50	0.40	0.60	0.05000	0.67254	3	0.40, 0.50, 0.10
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

(Report Continues)

References

Flack, V.F., Afifi, A.A., Lachenbruch, P.A., and Schouten, H.J.A. 1988. 'Sample Size Determinations for the Two Rater Kappa Statistic'. Psychometrika 53, No. 3, 321-325.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the total sample size.
κ0 is the value of Kappa under the null hypothesis, H0.
κ1 is the value of Kappa under the alternative hypothesis, H1.
Alpha is the probability of rejecting a true null hypothesis. It should be small.
Beta is the probability of accepting a false null hypothesis. It should be small.
k is the number of rating categories.
Frequencies lists the rating category frequencies. The number of frequencies is equal to k.

Summary Statements

In a test for agreement between two raters using the Kappa statistic, a sample size of 30 subjects achieves 8% power to detect a true Kappa value of 0.50 in a test of H0: Kappa = 0.40 vs. H1: Kappa ≠ 0.40 when there are 3 categories with frequencies equal to 0.40, 0.50, and 0.10. This power calculation is based on a significance level of 0.05000.

This report shows the numeric results of this power study. Following are the definitions of the columns of the report.

Power

The probability of rejecting a false null hypothesis.

N

The total sample size for the study.

κ0

The value of kappa under the null hypothesis.

κ1

The value of kappa under the alternative hypothesis.

Alpha

The probability of rejecting a true null hypothesis. This is often called the significance level.

Beta

The probability of accepting a false null hypothesis.

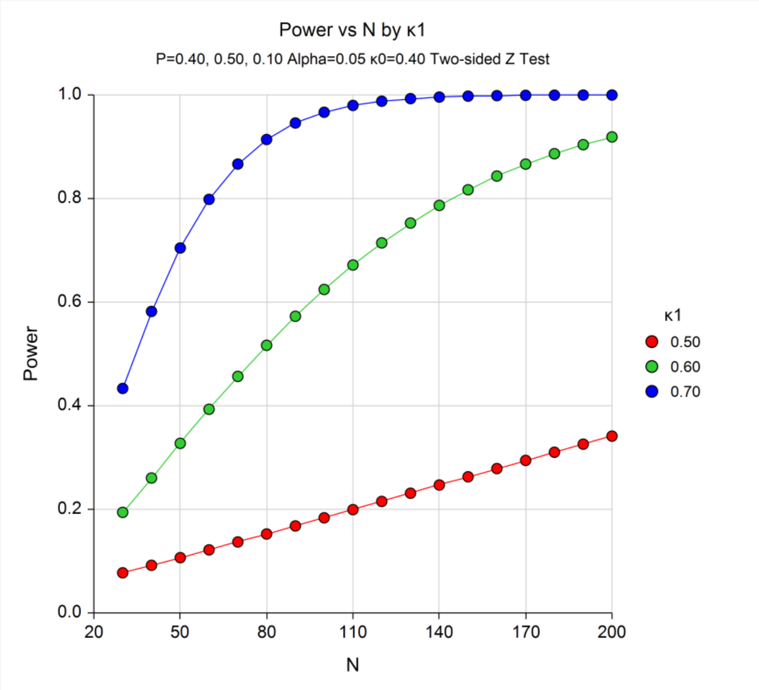
k

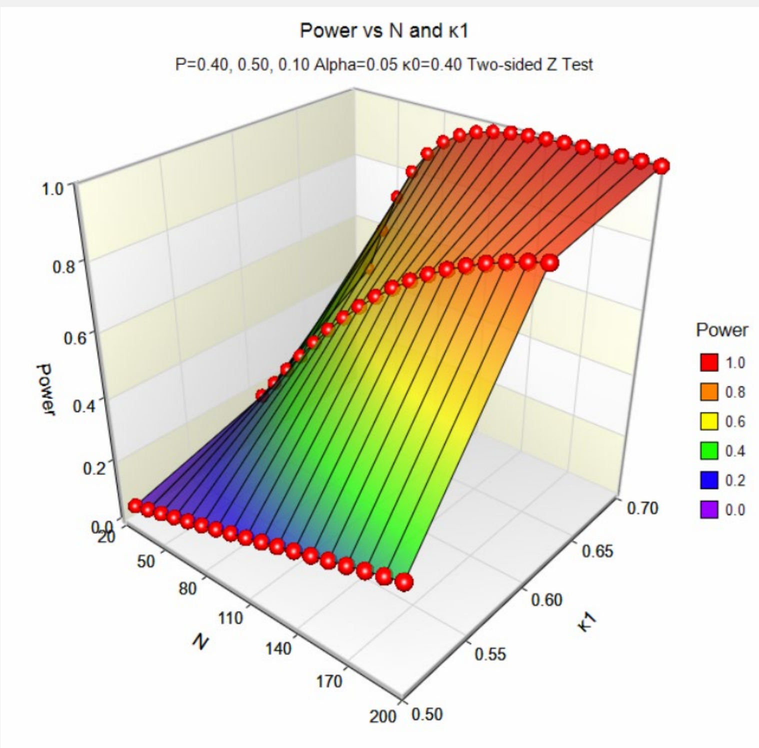
The number of rating categories.

Frequencies

The rating category frequencies used.

Plots Section





These plots give a visual presentation to the results in the Numeric Report. We can quickly see the impact on the power of increasing the sample size for the different values of κ1.

When you create these plots, it is important to use trial and error to find an appropriate range for the horizontal variable so that you have results with both low and high power.

Example 2 – Finding the Sample Size

Continuing with the last example, we will determine how large the sample size would need to be for the three values of κ1 to have the power at least 0.95 with an alpha level of 0.05.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Kappa Test for Agreement Between Two Raters** procedure window by expanding **Correlation**, then clicking on **Kappa**, and then clicking on **Kappa Test for Agreement Between Two Raters**. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Alternative Hypothesis (H1)	**Two-Sided**
Power	**0.95**
Alpha	**0.05**
κ1	**0.5 0.6 0.7**
κ0	**0.4**
Specify Using	**List Input**
P (Frequencies)	**0.4 0.5 0.1**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Test Type = Two-sided Z test.
H0: Kappa = κ0 vs. H1: Kappa ≠ κ0.

Power	Sample Size N	Kappa|H0 κ0	Kappa|H1 κ1	Alpha	Beta	- Rating Categories - k	Frequencies
0.95003	983	0.40	0.50	0.05000	0.04997	3	0.40, 0.50, 0.10
0.95031	228	0.40	0.60	0.05000	0.04969	3	0.40, 0.50, 0.10
0.95078	92	0.40	0.70	0.05000	0.04922	3	0.40, 0.50, 0.10

The required sample sizes are 983, 228, and 92 for alternative kappa values of 0.5, 0.6, and 0.7, respectively.

Example 3 – Finding the Minimum Detectable Kappa

Continuing with the last example, we will now determine what is the minimum value of kappa that can be detected with 100 subjects and power of 0.95.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Kappa Test for Agreement Between Two Raters** procedure window by expanding **Correlation**, then clicking on **Kappa**, and then clicking on **Kappa Test for Agreement Between Two Raters**. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**κ1 (κ1 > κ0)**
Alternative Hypothesis (H1)	**Two-Sided**
Power	**0.95**
Alpha	**0.05**
N (Sample Size)	**200**
κ0	**0.4**
Specify Using	**List Input**
P (Frequencies)	**0.4 0.5 0.1**
Reports Tab	
Decimal Places - Kappa	**4**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Test Type = Two-sided Z test.
H0: Kappa = κ0 vs. H1: Kappa ≠ κ0.

Power	Sample Size N	Kappa|H0 κ0	Kappa|H1 κ1	Alpha	Beta	- Rating Categories - k	Frequencies
0.95000	200	0.4000	0.6122	0.05000	0.05000	3	0.40, 0.50, 0.10

The test detects a kappa value of 0.6122 with 95% power.

Example 4 – Validation using Flack, Afifi, Lachenbruch, and Schouten (1988)

Flack, Afifi, Lachenbruch, and Schouten (1988) page 324 presents a table (Table 2) of calculated sample sizes required for 80% power in a one-sided test of H1: Kappa > 0.4 vs. H1: Kappa = 0.4 computed at $\kappa 1$ = 0.6 and alpha = 0.05. The sample sizes are computed for various sets of category frequencies.

Table 2

	Frequencies		Sample Size
PD	N	PS	for 80% Power
0.50	0.26	0.24	93
0.50	0.30	0.20	99
0.55	0.30	0.15	109
0.60	0.30	0.10	119
0.60	0.21	0.19	107

This example will replicate these results.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Kappa Test for Agreement Between Two Raters** procedure window by expanding **Correlation**, then clicking on **Kappa**, and then clicking on **Kappa Test for Agreement Between Two Raters**. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**. You can see that the values have been loaded into the spreadsheet by clicking on the spreadsheet button.

Option	Value
Design Tab	
Solve For	**Sample Size**
Alternative Hypothesis (H1)	**One-Sided (H1: κ1 > κ0)**
Power	**0.80**
Alpha	**0.05**
κ1	**0.6**
κ0	**0.4**
Specify Using	**Spreadsheet Column Input**
P (Frequencies)	**=C1-C5**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results
Test Type = One-sided Z test.
H0: Kappa ≤ κ0 vs. H1: Kappa > κ0.

Power	Sample Size N	Kappa|H0 κ0	Kappa|H1 κ1	Alpha	Beta	- Rating Categories - k	Frequencies
0.80218	93	0.40	0.60	0.05000	0.19782	3	0.50, 0.26, 0.24
0.80143	99	0.40	0.60	0.05000	0.19857	3	0.50, 0.30, 0.20
0.80253	109	0.40	0.60	0.05000	0.19747	3	0.55, 0.30, 0.15
0.80286	120	0.40	0.60	0.05000	0.19714	3	0.60, 0.30, 0.10
0.80259	106	0.40	0.60	0.05000	0.19741	3	0.60, 0.21, 0.19

The sample sizes computed by PASS match those in Flack, Afifi, Lachenbruch, and Schouten (1988). Slight differences are due to rounding.