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* × *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, …, *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

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

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

$$1 - \beta = \Pr\left(z \geq z_{critical} | H_1\right)$$

$$= 1 - \Phi(u)$$

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

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

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Output

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

Numeric Reports

Numeric Results

Solve For: Power
Number of Rating Categories (k): 3
Test Type: Two-sided Z-test
Hypotheses: H0: Kappa = κ0 vs. H1: Kappa ≠ κ0

		Degree of Agreement Between Raters			
Power	**Sample Size N**	**Kappa | H0 κ0**	**Kappa | H1 κ1**	**Alpha**	**Classification Frequencies P**
0.07748	30	0.4	0.5	0.05	0.4, 0.5, 0.1
0.19421	30	0.4	0.6	0.05	0.4, 0.5, 0.1
0.43345	30	0.4	0.7	0.05	0.4, 0.5, 0.1
0.09199	40	0.4	0.5	0.05	0.4, 0.5, 0.1
0.26055	40	0.4	0.6	0.05	0.4, 0.5, 0.1
0.58208	40	0.4	0.7	0.05	0.4, 0.5, 0.1
0.10677	50	0.4	0.5	0.05	0.4, 0.5, 0.1
0.32746	50	0.4	0.6	0.05	0.4, 0.5, 0.1
0.70452	50	0.4	0.7	0.05	0.4, 0.5, 0.1
0.12180	60	0.4	0.5	0.05	0.4, 0.5, 0.1
0.39325	60	0.4	0.6	0.05	0.4, 0.5, 0.1
0.79842	60	0.4	0.7	0.05	0.4, 0.5, 0.1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Power The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N The total sample size.
κ0 The value of Kappa under the null hypothesis, H0.
κ1 The value of Kappa under the alternative hypothesis, H1.
Alpha The probability of rejecting a true null hypothesis.
P The classification frequencies (proportions) in each of the k categories as assigned by the two raters (judges).

Summary Statements

An agreement between two raters design (with 3 categories and frequencies of 0.4, 0.5, and 0.1) will be used to test whether the rater agreement (Kappa) is different from 0.4 (H0: Kappa = 0.4 versus H1: Kappa ≠ 0.4). The comparison will be made using a two-sided Kappa statistic Z-test, with a Type I error rate (α) of 0.05. To detect a Kappa of 0.5 with a sample size of 30 subjects, the power is 0.07748.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size N	Dropout-Inflated Enrollment Sample Size N'	Expected Number of Dropouts D
20%	30	38	8
20%	40	50	10
20%	50	63	13
20%	60	75	15
20%	70	88	18
20%	80	100	20
20%	90	113	23
20%	100	125	25
20%	110	138	28
20%	120	150	30
20%	130	163	33
20%	140	175	35
20%	150	188	38
20%	160	200	40
20%	170	213	43
20%	180	225	45
20%	190	238	48
20%	200	250	50

Dropout Rate The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.

N The evaluable sample size at which power is computed (as entered by the user). If N subjects are evaluated out of the N' subjects that are enrolled in the study, the design will achieve the stated power.

N' The total number of subjects that should be enrolled in the study in order to obtain N evaluable subjects, based on the assumed dropout rate. N' is calculated by inflating N using the formula N' = N / (1 - DR), with N' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)

D The expected number of dropouts. D = N' - N.

Dropout Summary Statements

Anticipating a 20% dropout rate, 38 subjects should be enrolled to obtain a final sample size of 30 subjects.

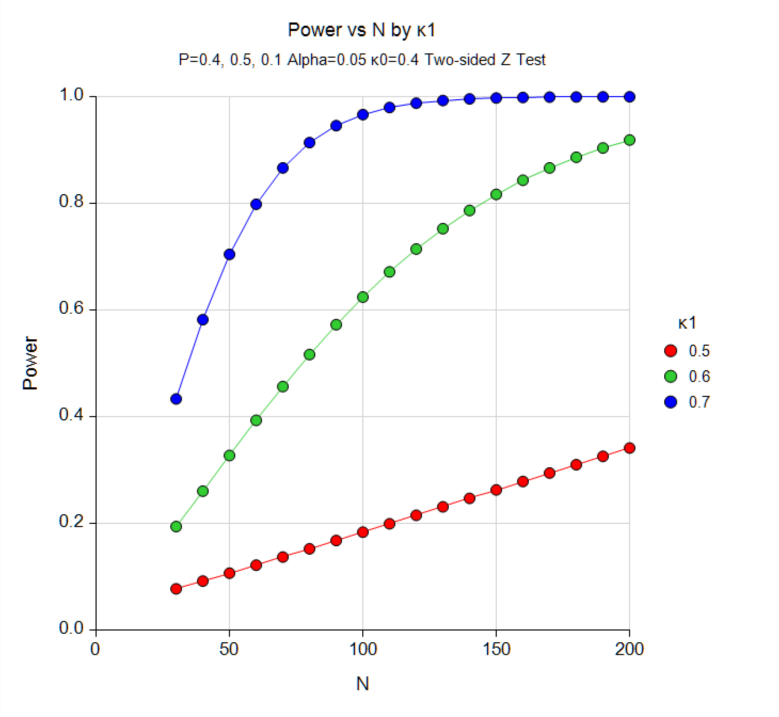
References

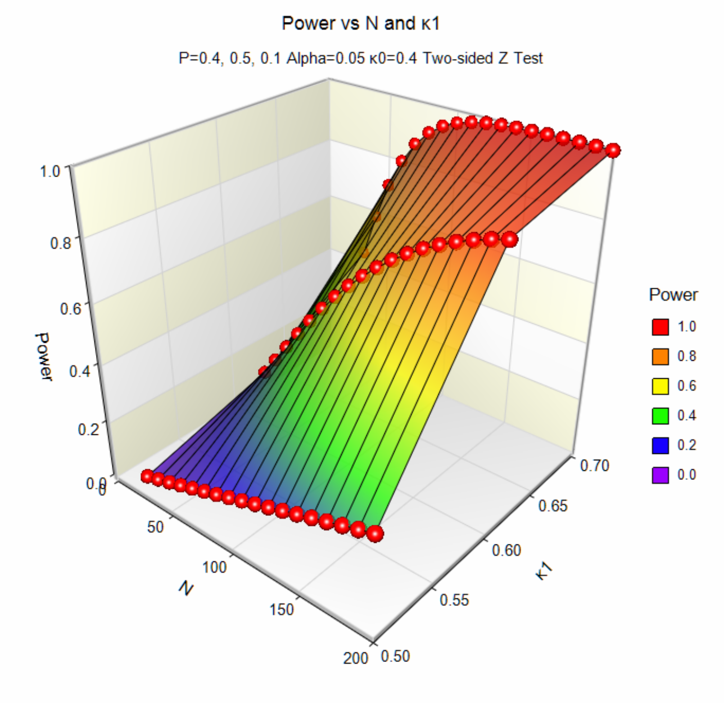
Flack, V.F., Afifi, A.A., Lachenbruch, P.A., and Schouten, H.J.A. 1988. 'Sample Size Determinations for the Two Rater Kappa Statistic'. Psychometrika 53, No. 3, 321-325.

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

Plots Section

Plots





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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

```
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
─────────────────────────────────────────────────────────────────────────────
Solve For:                          Sample Size
Number of Rating Categories (k):    3
Test Type:                          Two-sided Z-test
Hypotheses:                         H0: Kappa = κ0   vs.   H1: Kappa ≠ κ0
─────────────────────────────────────────────────────────────────────────────
                                Degree of Agreement
                                   Between Raters
                   Sample     ──────────────────────             Classification
                     Size     Kappa | H0    Kappa | H1           Frequencies
Power                   N             κ0            κ1    Alpha  P
─────────────────────────────────────────────────────────────────────
0.95003               983            0.4           0.5     0.05  0.4, 0.5, 0.1
0.95031               228            0.4           0.6     0.05  0.4, 0.5, 0.1
0.95078                92            0.4           0.7     0.05  0.4, 0.5, 0.1
─────────────────────────────────────────────────────────────────────────────
```

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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

```
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
──────────────────────────────────────────────────────────────────────────────
Solve For:                          κ1 (κ1 > κ0)
Number of Rating Categories (k):    3
Test Type:                          Two-sided Z-test
Hypotheses:                         H0: Kappa = κ0   vs.   H1: Kappa ≠ κ0
──────────────────────────────────────────────────────────────────────────────
                              Degree of Agreement
                                Between Raters
           Sample       ─────────────────────────             Classification
             Size       Kappa | H0      Kappa | H1             Frequencies
Power           N               κ0              κ1     Alpha    P
───────────────────────────────────────────────────────────────────────────
0.95          200              0.4          0.6122      0.05    0.4, 0.5, 0.1
──────────────────────────────────────────────────────────────────────────────
```

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 κ1 = 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

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

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**

Input Spreadsheet Data

Row	C1	C2	C3	C4	C5
1	0.50	0.5	0.55	0.6	0.60
2	0.26	0.3	0.30	0.3	0.21
3	0.24	0.2	0.15	0.1	0.19

Output

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

```
Numeric Results
──────────────────────────────────────────────────────────────────────────────────
Solve For:                          Sample Size
Number of Rating Categories (k):    3
Test Type:                          One-sided Z-test
Hypotheses:                         H0: Kappa ≤ κ0   vs.   H1: Kappa > κ0
──────────────────────────────────────────────────────────────────────────────────
                            Degree of Agreement
                              Between Raters
              Sample     ─────────────────────────             Classification
                Size     Kappa | H0     Kappa | H1             Frequencies
Power              N             κ0             κ1     Alpha   P
───────────────────────────────────────────────────────────────────────────
0.80218           93            0.4            0.6      0.05   0.5, 0.26, 0.24
0.80143           99            0.4            0.6      0.05   0.5, 0.3, 0.2
0.80253          109            0.4            0.6      0.05   0.55, 0.3, 0.15
0.80286          120            0.4            0.6      0.05   0.6, 0.3, 0.1
0.80259          106            0.4            0.6      0.05   0.6, 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.