Universitetet i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation: Master of Mathematics and Physics Statistics	Spring semester, 2019 Open/~~Confidential~~
Author: Bogdana Tolokonnikova	Bogdana Tolokonnikova (signature of author)
Programme coordinator: Jan Terje Kvaløy Supervisor(s): Anastasia Ushakova, Ingvild Dalen and Jan Terje Kvaløy (internal)	
Title of master's thesis: Sample size requirements for agreement studies	
Credits: 60	
Keywords: Inter-rater agreement, sample size, reliability, agreement measures, chance agreement, Cohen's kappa, Gwet's AC1, Variance of AC1, planning experiment	Number of pages: 67 + supplemental material/other: …… Stavanger, 14/06/2019 date/year

Sample size requirements for agreement studies

Bogdana Tolokonnikova

Department of Mathematics and Physics
University of Stavanger
Submission Date: June 2019
Supervisors: Anastasia Ushakova, Ingvild Dalen and Jan Terje Kvaløy

Preface

I would like to thank my supervisors Anastasia Ushakova and Ingvild Dalen for the guidance, valuable knowledges that I got from them and right feedbacks during the writing of master thesis this last academic year. I would also like to thank my internal supervisor Jan Terje Kvaløy for estimable advices and help, and for being a very good professor of the statistical subjects during my bachelor and master education, which were basis for writing my master thesis.

Abstract

This thesis examines requirements of subject sample size while planning a medical experiment, describes some known types of measures of inter-rater agreement and discovers some new useful results in this study area. The main goal is to determine how many patients to include in clinical study. There is a particular focus on studying the Gwet's AC1 agreement coefficient by comparing it to the well known Cohen's kappa coefficient and some others, which are popular used in the inter-rater agreement studies. As is shown most agreement coefficients have paradoxical behaviour, which gives unreasonably low estimates of agreement in some situations. Gwet's AC1 has been claimed to be "paradox-free", however, we find its own paradox. That is why the focus holds on studying some useful properties of this measure and analyzing the formula for variance which has the main role in finding the required sample size.

Contents

1 Introduction

Sample size calculation is an important part of any clinical research. There are different methods to estimate the sample size. If we want to provide an effective and safe test of treatment, we need to make sure that needed number of subjects were taken for the trial.

If we talk about clinical research, often two measurement methods need to be compared. Before researches can start the clinical trial, they need to identify sample size to present the contingency table of the agreement between the raters.

Chapters 2 and 3 cover introduction to the main topic, the theory of the power-based and precision-based calculations of sample size is given. An introduction to inter-rater agreement and reliability of diagnostic tests is given in chapter 4, which also contains a brief summary of the theory of some types of measures of inter-rater agreement, gives some simple analytical and practical examples.

In Chapter 5 properties of Gwet's gamma coefficient are studied in detail and important conclusions done. The minimum and maximum for percent chance agreement are established. It is found that for large number of categories Gwet's gamma is no longer a chance-corrected measure.

An upper bound of variance of Gwet's gamma is obtained in chapter 6. Afterwards it is finally shown how to choose the sample size when designing an inter-rater reliability study to attain a pre-specified margin of error for Gwet's gamma.

2 Sample size calculations in medical research planning

Proper research planning is an integral part of evidence-based medicine. Sample size calculations are important parts of planning quantitative studies. During the studies that determine the prevalence of a certain characteristic in a population(for example, the prevalence of asthma in children), the calculation of the sample size is necessary to get the desired degree of accuracy from the obtained estimates.

For example, the prevalence of the disease in 10%, obtained on a sample size of 20 people, will have a 95% confidence interval from 1% to 31%, which can be recognized neither accurate nor informative estimate. On the other hand, the prevalence of the disease in 10%, obtained on a sample size of 400 people, will have a 95% confidence interval from 7% to 13%, which can be considered as fairly accurate result. Pre-calculation/assessment of available precision for different choices of sample size before study start let us avoid the first of these two situations.

In studies aimed to identify precision in estimate of treatment effect (for example, the difference in the effectiveness of two treatment methods, the relative risk of the disease in the presence or absence of a risk factor) estimating the sample size is important to ensure that if a clinically or biologically important effect exists, then it highly probable will be detected, in other words, the analysis will give statistically significant results.

The adequacy of the sample size should also be assessed in accordance with the purpose of the study. For example, if the goal of the study is to demonstrate that a new treatment is better than the existing one, then it is necessary to ensure that the sample size can detect clinically significant differences between the two treatments.

However, sometimes it is required to demonstrate that two treatments are clinically equivalent. This type of research is often called a test of equivalence or "negative" test. The sample size in studies aimed at demonstrating the equivalence of medicine is larger than in studies aimed to identify differences in efficiency. It is important to make sure that the sample size calculations are related to the goals and problems of the study and are based on data about the main outcome variable.

The sample size should also correspond to the methods of analysis used in the study, since both the sample size and the analysis depend on the chosen design of research. There are two approaches to sample size calculations, depending on whether the primary aim is a comparison performed by a statistical test, or an estimate of a population parameter.

2.1 Power-based sample size calculations

Power-based sample size calculations relate to hypothesis testing. Below will be presented some definitions used in power-based sample size calculations. Generally about hypothesis testing: we formulate a null hypothesis, H_0, and an alternative hypothesis, H_1, i.e.

H_0 : current knowledge

H_1 : new knowledge

More precisely, a hypothesis test should be formulated in terms of some population parameter(s). E.g. a general two-sided hypothesis test, for some population parameter θ:

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0.$$

By comparing groups, we initially assume that they do not differ (this is our H_0). If the probability that the identified differences are a random result is very small under the assumption that H_0 is true, then it will be clear to reject the null hypothesis and make a conclusion that there is indeed a difference (H_1 true). The value of p for a particular sample is the probability of obtaining at least the same or even greater differences than observed, provided that the null hypothesis is correct.

	Reality	
Test	H_0 true	H_0 false
Reject H_0	Type I Error (probability = α)	Correct rejection H_0 (probability= $1-\beta$ (=power))
Accept H_0	Correct acceptance of H_0 (probability= $1-\alpha$)	Type II Error (probability= β)

Table 1: Probability of errors Type I and Type II

Type I error (false positive)
Mistakenly rejecting the null hypothesis, i.e. conclude with differences where there are none. The probability of type I error is also called level of significance (alpha):

$$\alpha = P\left(Type \quad I \quad error\right) = level \quad of \quad significance$$

i.e.

$$\alpha = P\left(reject \quad H_0 | H_0 \quad true\right)$$

Type II error (false negative)
It occurs if we accept the null hypothesis when it is not true, in other words, we do not find significant difference. The probability of type II error is denoted by beta.

$$\beta = P\left(Type \quad II \quad error\right)$$

i.e.

$$\beta = P\left(not \qquad reject \quad H_0 | H_1 \quad true\right)$$

Power
The probability of detecting a true difference (i.e. the statistical power of a trial) is equal to $1-\beta$, i.e.

$$power = P\left(reject \quad H_0 | H_1 \quad true\right) = 1 - \beta$$

By using statistical calculations we can compute the p-value, which is later compared to a pre-selected level of significance, often denoted as α. In biomedical research the significance level is usually set at $\alpha = 0.05$ (= 5%). If the significant level was chosen at $\alpha = 0.05$, then all the samples that return $p \leq 0.05$ for the null hypothesis reject this hypothesis, and the samples with $p > 0.05$ do not give grounds for rejecting it.

The point is that in cases there really is no difference, there is only a 5% chance that by random chance an observed difference is large enough to lead to a rejection of the null hypothesis will happen. In other words we set the probability of false rejection of null hypothesis H_0 (standard) in favour of the alternative hypothesis H_1 (studied).

2.1.1 Power calculations: comparison of means

Suppose we have two groups of normally distributed data and we want to compare the mean in both of them to each other (i.e. carry out an unpaired t-test):

$$H_0 : \mu_1 = \mu_2 \qquad\qquad H_1 : \mu_1 \neq \mu_2.$$

Required number, n, in each group for a given level of α and power $1 - \beta$ is given as (R. Cornish, 2006)

$$n = f(\alpha, \beta) \cdot \frac{2s^2}{\delta^2}, \tag{1}$$

where
α - level of significance.
$1 - \beta$ - power of test.
$f(\alpha, \beta)$ - value calculated from α and β (see Table 2).
s - the standard deviation, assumed equal in both groups.
$\delta = \mu_1 - \mu_2$ - the smallest difference in the means which is considered to be clinically meaningful.

	β		
α	0.05	0.1	0.2
0.05	13.0	10.5	7.9
0.01	17.8	14.9	11.7

Table 2: $f(\alpha, \beta)$ for some commonly used choices of α and β

Example

We can see how the formula given above works on concrete example taken from Rosie Cornish (2006)"An introduction to sample size calculations"[5].

Suppose we wish to carry out a trial of a new treatment among men aged between 50 and 60. This treatment is for people who has hypertension (high blood pressure).Also

suppose we want to be 90% sure of detecting a difference in mean blood pressure of 10 mmHg as significant at 5% level (i.e. $\alpha = 0.05$, $\beta = 0.1$, power=0.9, $\delta = 10$). We assume $s = 20mmHg$ (this number is taken from other published papers about blood pressure studies), and calculate the number required in each group

$$n = f(\alpha, \beta) \cdot \frac{2s^2}{\delta^2} = 10.5 \cdot \frac{2 \cdot 20^2}{10^2} = 84,$$

using $f(\alpha, \beta) = 10.5$ from Table 2.

That is we would need 84 subjects in each group, i.e. a total of 168 participants to obtain the desired statistical power.

2.1.2 Power calculations: comparison of probabilities

Suppose we have two groups of size n and we want to compare a binary outcome in these groups. Let
p_1 = probability of events in group 1.
p_2 = probability of events in group 2.

Then the null and alternative hypotheses would be:

$$H_0 : p_1 = p_2 \qquad\qquad H_1 : p_1 \neq p_2.$$

We would like to detect the *smallest* important difference in proportions, $\delta = p_1 - p_2$. General form same as equation (1), however since the variance of a binomial random variable is given by p, this simplifies to the following form, where required number of subjects, n, is given by (R.Cornish, 2006)

$$n = \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2} \cdot f(\alpha, \beta). \tag{2}$$

Example

We look at group of people who have had a heart attack. A new treatment has been developed for such patients. It is known that 10% of people have died within one year after the heart attack. It is clearly that it would be very important clinically if the death percentage were to be reduced from 10% to 5%. We will use $\alpha = 0.05$ and $\beta = 0.1$. It means that our $p_1 = 0.1$ (probability of deaths in placebo group) and $p_2 = 0.05$ (probability of deaths in treatment group). Using the formula above, we get

$$n = \frac{0.1(1-0.1) + 0.05(1-0.05)}{(0.1 - 0.05)^2} \cdot 10.5 = 578$$

So we can conclude that 578 patients is needed in each treatment group so that we are 90% sure of being able to detect a reduction from 10% to 5% as statistically significant at the 5% level.

2.2 Precision-based sample size calculations

Suppose now we want to estimate an unknown parameter with a certain degree of precision. In other words we want our confidence interval to be of a certain width. We have a general formula to find a 95% confidence interval for asymptotic normal case, it is given by

$$Estimate \pm z_{\alpha/2} \times SE,$$

where $z_{\alpha/2} = 1.96$ - quantile of normal distribution, SE is a standard error of what we are estimating. This formula is based on approximation.

The formula for standard error contains number n, which is the sample size. It means we can get the formula that can be solved to find n, by specifying how wide the 95% confidence interval should be. We can use consideration around variance/standard deviation for determining sample size we need. Let us first present the general formulation of the sample size calculation for one parameter.

Assume we have some estimate $\hat{\mu} = \bar{X}$. Then $Var(\bar{X}) = \sigma^2/n$, where $\sigma^2 = Var(X)$, and $SD(\bar{X}) = \sigma/\sqrt{n}$. As we mentioned that our estimate is normal distributed, then there is approximately a $(1-\alpha)100\%$ chance that $\bar{X}$ will fall in: $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, i.e. an interval estimate that with $(1-\alpha)100\%$ probability covers the true/population parameter, which we call confidence interval.

In other words, $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ should not exceed a margin error, e:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq e \Longrightarrow n \geq \frac{z^2_{\alpha/2}\sigma^2}{e^2}.$$

Example 1(Gayatri Vishwakarma [9])

Calculate the sample size needed to achieve plasma lamotrigine (LTG) among patients who have seizures with $\pm 1.0 mg/l$ precision and 95% confidence. In this case the standard deviation of plasma was taken as $2.0 mg/l$, i.e. $\sigma = 2.0$, $e = 1.0 mg/l$, and for 95% significance level $z_{\alpha/2} = 1.96$.

Thus, by inserting given values into the formula for n, we get

$$n = \frac{(1.96)^2 2^2}{1^2} = 15.36.$$

Rounding upwards, it means we need to have a sample size at least 16.

Example 2

We will take the example with a new treatment for people with high blood pressure as we did before(R.Cornish, 2006 [5]). So we select randomly $2n$ subjects, n of them get this new treatment and the other n get the same treatment as before (it means we take both groups of the same size). In addition to testing if the new treatment is better than the old one, we want to estimate a 95% CI for the difference in mean. The

CI should be wider than $10mmHg$. So the question is how many subjects we need to have in this study?

We have the formula for 95% confidence interval for a difference in means. It is given by

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where s_p is pooled estimate of common variance.

We want $1.96 \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq 5mmHg$. Also $n_1 = n_2 = n$, i.e.

$$1.96 \times s_p \sqrt{\frac{2}{n}} \leq 5,$$

We need to know what s_p is, so that we can work out sample size. As we did in similar example above, we took $s_p = 20mmHg$ by using other published papers about this topic. This gives us

$$1.96 \cdot 20 \sqrt{\frac{2}{n}} \leq 5 \Longrightarrow \frac{n}{2} \geq \left(\frac{1.96 \cdot 20}{5}\right)^2 \Longrightarrow n = 122.93.$$

Rounding up, it means we need 123 subjects in each group.

If we wanted, for example, to estimate a 95% CI for the same parameter within $2.5mmHg$ of the estimated, then we would get the following

$$\frac{n}{2} = \left(\frac{1.96 \cdot 20}{2.5}\right)^2 \Longrightarrow n = 491.72,$$

which by rounding gives us 492 subjects.

From this we can conclude that we have to increase our sample size by a factor of 4, if we want to increase our precision by factor of 2, as we showed in example. In general, by increasing the precision by some factor k, the sample size needs to be increased by factor k^2 (as it said in R.Cornish, 2006 [5]).

3 Inter-rater agreement

3.1 Reliability of measurements/classifications

Reliability is the level of constancy of the repetition of the same result in multiple measurements of a trait under identical conditions. Further we want to ask the following question: Can the result be repeated if the measurement is repeated several times?

Let's take a look at the reasons for the low reliability of diagnostics.

- The discrepancy between the results obtained by different raters when measuring the same symptom/trait
- Difference in the results of repeated measurements of a trait provided by the same rater

We want to find out the variation between different researchers. First of all, two raters often do not come to the same result. So knowing the level where two raters agree or disagree with each other is very important in science and in assessing the quality of medical care, regardless of whether a physical examination, laboratory test, or other measurement of human characteristics is performed. Therefore, when performing a research or improving the quality of diagnosis in the practice, it is necessary to know how to represent agreement in quantitative form.

3.2 Inter-rater reliability

Inter-rater agreement is one of many measures of reliability. Inter-rater reliability can be defined as a degree to which two or more individuals referred as raters give an independent data classification with the same set of objects. In other words inter-rater reliability, which measures uniformity, requires to conduct the same form with the same people by two or more raters in order to establish the degree of agreement on the use of this tool by those who use it.

Confidence in accuracy of the clinical study depends on the reliability of gathered data. One of the important factor to get a clinical and medical research with a high quality lies in the importance of raters having a high degree of agreement while testing the samples. Any research project can potentially have a number of errors. And the confidence in the conclusions of the study depends on minimizing those errors by the researcher.

3.3 Types of data and measurement scales

In rating we usually work with data which can be of different type: nominal and ordinal. *Nominal scale*, which is also called a categorical variable scale, is assigned to subjects divided into categories without having any order. Nominal scale is used for labeling variables into distinct classes. By constructing a nominal scale, the following requirement must be met: each item of a set of objects must be assigned only to one class, i.e. none of the objects can be assigned simultaneously to two or more classes. If

we talk about inter-rater agreement, in this case by having a nominal scale, we assume that two individuals or raters agree when their ratings are identical. And they disagree if the ratings are not identical.

In *ordinal scale* the classes of objects are discrete, as in case of nominal scale, and the categories are ordered. However, the numbers can be compared, it is always necessary to remember that quantities in the ordinal scale have only relative and not absolute value. When we have an ordinal measurement, we no longer assume the agreement and disagreement as two distinct definitions. It means that a disagreement can be assumed as some other level of agreement, in other words, quantify level of disagreement.

To look closer at the difference between these types of data, we will take a simple example from Kilem L. Gwet "Handbook of inter-rater reliability"[1].

Suppose a psychiatrist classifies his patients into five categories: Personal Disorder, Schizophrenia, Depression, Neurosis and Other. There is no possible way of meaningful ordering of these categories. That's why we can say that the scale of these five categories is *Nominal*. From the other side,after being tested for Multiple Sclerosis, the patient can be defined as Doubtful, Probable, Possible, Certain, and that is then rated on *Ordinal scale*. In this case it is clear that one of the categories can be closer to another one but not all of them. For example, the category "Certain" is closer to "Probable" than to the "Doubtful" category. By looking at this example we can say that disagreements on nominal and ordinal scales should be considered in different ways. It means that the way to analyse the data depends on which type of rating data we are working with, nominal or ordinal.

Inter-rater reliability can be calculated for all types of data, but we want to focus on categorical data, specifically nominal type of data, where the order of categories does not have any interpretation.

3.4 Establishment of framework

Let's first present the general agreement table for distribution of n subjects rated by two raters into k categories.

	Rater A				
Rater B	1	2	...	k	Total
1	n_{11}	n_{12}	...	n_{1k}	$\mathbf{n_{1.}}$
2	n_{21}	n_{22}	...	n_{2k}	$\mathbf{n_{2.}}$
...	...	...	...	...	...
k	n_{k1}	n_{c1}	...	n_{kk}	$\mathbf{n_{k.}}$
Total	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	...	$\mathbf{n_{.k}}$	$\mathbf{n}$

Table 3: Distribution of n subjects by rater and category.
The categories are nominal

Diagonal elements of Table 3 represent agreement among the raters, while non-diagonal elements represent disagreement. The level of agreement between the raters is defined by different types of measures, agreement coefficients, which will be presented later. Most agreement coefficients (but not all of them) express the value to which the observed agreement exceeds the random agreement and looks like a proportion of maximum possible improvement. We can present a very general formulation of agreement coefficient:

$$\text{Coefficient} = \frac{\begin{array}{c}\text{percent of}\\ \text{observed agreement}\end{array} \; - \; \begin{array}{c}\text{percent of}\\ \text{random agreement}\end{array}}{\begin{array}{c}100\%\end{array} \; - \; \begin{array}{c}\text{percent of}\\ \text{random agreement}\end{array}}$$

4 Types of measures of inter-rater agreement

4.1 Cohen's Kappa coefficient

Cohen's kappa coefficient is a statistical measure of inter-rater agreements for categorical data. Kappa is the most used chance-adjusted agreement coefficient and assumed to be a more reliable measure than the simple calculation of a percent agreement.

Cohen's Kappa measures agreement between two raters and it is calculated as:

$$\kappa = \frac{\frac{1}{n}\left(\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{ii}^{e}\right)}{1 - \frac{1}{n}\sum_{i=1}^{k} n_{ii}^{e}}, \tag{3}$$

where
k - number of categories,
n - the total number of pairs (observations),
n_{ii} - the number of agreed pairs of category i,
n_{ii}^{e} - the expected number of agreed pairs of category i, which is calculated as:

$$n_{ii}^{e} = \frac{1}{n}\sum_{j=1}^{k} n_{ij} \times \sum_{j=1}^{k} n_{ji}.$$

.

Interpretation of Cohen's kappa values is given in Table 4 (Landis & Koch, 1977)[20].

Cohen's Kappa value	Level of agreement
0.00	Poor
< 0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	**Almost Perfect**

Table 4: Level of agreement for Cohen's Kappa value

An agreement table can also be presented in terms of proportions, i.e. where each cell's count is divided by n (the total number of observations), i.e. $p_{ij} = n_{ij}/n$. In this case we can write kappa as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{4}$$

where

$$\begin{cases} p_0 = \sum_{i=1}^{k} p_{ii}, \\ p_e = \sum_{i=1}^{k} p_{i.} p_{.i}, \end{cases} \tag{5}$$

$p_{i.}$ and $p_{.i}$ is the sum of frequencies of each category.

The kappa indicator reaches its maximum value, i.e. 1, if all non-diagonal elements are equal to zero.

Example

We will see how formula (4) works on practice by looking at a simple example. We have a table of agreement (Table 5) with two raters and two categories.

	Rater 2		
Rater 1	Yes	No	Total
Yes	19	2	21
No	3	4	7
Total	22	6	28

Table 5: Distribution of 28 subjects by rater and category

By using the formulas for the observed agreement given for Cohen's kappa, we get:

$$p_0 = (19 + 4)/28 = 0.82,$$

and Cohen's chance agreement:

$$p_e = \frac{21}{28} \times \frac{22}{28} + \frac{7}{28} \times \frac{6}{28} = 0.64.$$

Thus, the Cohen's kappa coefficient (4) will be:

$$\kappa = \frac{0.82 - 0.64}{1 - 0.64} = 0.5.$$

So, we got an agreement coefficient equal to 0.5 which is corresponding to *Moderate* level of agreement using Table 4.

4.2 Scott's Pi coefficient

Cohen's Kappa coefficient can be compared to an other agreement coefficient named *Pi* and invented by William A. Scott, 1955. The general formula is the same as for Cohen's kappa (4), however chance agreement is calculated as

$$p_e = \sum_{i=1}^{k} \pi_i^2$$

with

$$\pi_i = (p_{i.} + p_{.i})/2,$$

which is "the agreement that is expected when the units are statistically unrelated to their descriptions" (Krippendorff, 2007, p. 80).

For situation with two rater and categories, this simplifies to:

$$p_e = \pi_1^2 + (1 - \pi_1)^2.$$

Example

Using the example from Table 5 and the formulas presented above, we can calculate the value for p_e:

$$\pi_1 = \left(\frac{21}{28} + \frac{22}{28}\right)/2 = 0.77,$$

$$p_e = \pi_1^2 + (1 - \pi_1)^2 = 0.77^2 + (1 - 0.77)^2 = 0.65.$$

Insert the values of p_0 and p_e into the formula for Pi and get:

$$Pi = \frac{0.82 - 0.65}{1 - 0.65} = 0.49.$$

So, the Scott's Pi agreement coefficient is equal to 0.49, which in this case is very similar to Cohen's kappa 0.50.

4.3 Krippendorff's Alpha coefficient

Krippendorff's Alpha appeared as a way to determine the validity of data in content analysis (Krippendorff, 2004). It is considered to be a reliable way to describe the level of agreement between the raters. It differs from the previously mentioned coefficients in that it incorporates a small-sample correction to the observed (percent) agreement.

For n subjects categorized into k categories, the Krippendorff's alpha coefficient can be calculated from the following formula:

$$\alpha = \frac{p_0' - p_e}{1 - p_e},$$

where

$$\begin{cases} p_0' = (1-\epsilon_n)p_0 + \epsilon_n, \\ p_0 = \sum_{i=1}^{k} p_{ii}, \\ p_e = \sum_{i=1}^{k} \pi_i^2, \end{cases}$$

and

$$\begin{cases} \epsilon_n = 1/(2n), \\ \pi_i = (p_{i.} + p_{.i})/2. \end{cases}$$

Note: As $n \to \infty$, $\epsilon_n \to 0$; thus for large samples Krippendorff's alpha $\approx$ Scott's Pi. (According to Gwet (2014, p.39 [1]), small samples shows insignificance of ϵ already as $n = 10$.)

Krippendorff's alpha coefficient in 2×2 case simplifies to:

$$\begin{cases} p_e = \pi_1^2 + (1-\pi_1)^2, \\ p_0' = (1-\epsilon_n)(p_{11} + p_{22}) + \epsilon_n. \end{cases}$$

Example

We are going to use the same example as we did for Cohen's kappa and Scott's Pi coefficient. Assume that we have the same dataset from Table 5.

The values for p_0, π_1 and p_e are going to be without change, i.e. $p_0 = 0.82$, $\pi_1 = 0.77$, $p_e = 0.65$.

By taking the formula from above, we can find the values for p_0' and p_e:

$$\epsilon_n = 1/(2n) = 1/56 = 0.018,$$

$$p_0' = (1 - 0.018) \cdot 0.82 + 0.018 = 0.82.$$

Thus, we have

$$\alpha = \frac{0.82 - 0.65}{1 - 0.65} = 0.49.$$

The Krippendorff's alpha agreement coefficient is equal to 0.49, i.e. the same as Scott's Pi.

4.4 Gwet's AC1 coefficient

The AC1 statistics was proposed by K.L. Gwet(2008a) as an alternative, improved agreement coefficient compared to Cohen's kappa. The AC1 coefficient "is defined as the conditional probability that two randomly selected raters agree given that there is no agreement by chance." (Gwet 2001). The main difference from the kappa coefficient is the way we calculate the chance agreement.

The AC1 coefficient for k categories is given by:

$$\gamma = \frac{p_0 - p_e}{1 - p_e}, \tag{6}$$

where

$$p_e = \frac{1}{k-1} \sum_{i=1}^{k} \pi_i (1 - \pi_i), \tag{7}$$

and $\pi_i = (p_{i.} + p_{.i})/2$ has the same form as for Scott's Pi and Krippendorff's alpha coefficients.

For a 2×2 case, the percent chance agreement simplifies to:

$$p_e = 2\pi_1(1 - \pi_1).$$

Notaion for Gwet's AC1 coefficient such as $AC1$ and γ will be used interchangeably following the notation in Gwet's book [1].

Example

Again we take the same dataset from Table 5 as we did before to calculate the agreement coefficient and compare it with the others. The observed percent agreement p_0 and π_1 are calculated the same way as in all other cases. We have $p_0 = 0.82$, $\pi_1 = 0.77$.

To calculate Gwet's chance agreement we insert the value for π_1 into formula for p_e:

$$p_e = 2 \cdot 0.77 \cdot (1 - 0.77) = 0.35.$$

Thus, we can find the value for gamma:

$$\gamma = \frac{0.82 - 0.35}{1 - 0.35} = 0.72.$$

Gwet's AC1 agreement coefficient is equal to 0.72. This value of agreement is the highest among all the agreement coefficients presented in this chapter, due to lowest chance agreement.

4.5 Paradoxes of agreement measures

When we talk about kappa coefficient, one can notice sometimes unexpected values kappa gives us. This value can be quite low if we compare it to the percent agreement. Feinstein and Cicchetti (1990) shows us issues of kappa statistics which is called two kappa paradoxes. As it is shown in Gwet(2014, p. 58)[1], these two paradoxes were described by Feinstein and Cicchetti like this:

- *"The first paradox of κ is that if p_e is large, the correction process can convert a relatively high value of p_0 into a relatively low value of κ" (Feinstein & Cicchetti (1990, p. 544)*
- *"The second paradox occurs when unbalanced marginal totals produce higher values of κ than more balanced totals."(Feinstein & Cicchetti (1990, p. 545)*

4.5.1 The first paradox

This paradox is a function of a high unbalanced prevalence in the sample. Let's take a look at the agreement between two raters, where they agree almost perfectly about rating some number of subjects into two category.

	Rater 2		
Rater 1	Yes	No	Total
Yes	25	5	30
No	0	0	0
Total	25	5	30

Table 6: Distribution of 30 subjects by rater and category

By using the data from the Table 6, the kappa coefficient will be:

$$p_0 = (25 + 0)/30 = 0.83,$$

$$p_e = \frac{30}{30} \times \frac{25}{30} + \frac{0}{30} \times \frac{5}{30} = 0.83.$$

Thus,

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0.83 - 0.83}{1 - 0.83} = 0.$$

So, we can see that kappa calculations gives us the result equal to 0, which one can read as no agreement between the raters. Yet, by looking at the data in the

contingency table, we have almost prefect agreement between the raters. If we look at the observed marginals, we can see that rater 1 put all the subjects into the category "Yes", which gives us probability equal to 1. At the same time the rater 2 put 83% to same category. Since we used all subject to compute p_0 and p_e, it is logically that $p_0 - p_e$ will have no meaning. It means that maybe it is not smart to use these observed marginals(concentrated in one category), which is leading to this paradox.

Let us check if this paradox holds for the other agreement coefficients that have been presented in this chapter before.

As we said earlier, the observed agreement $p_0 = 0.83$ will be the same for all coefficients ($p_0' = p_0$ with $n = 30$ for Krippendorff's alpha). We need to find only the value for percent chance agreement. First we need to find π_1:

$$\pi_1 = (\frac{30}{30} + \frac{25}{30})/2 \approx 0.91,$$

then p_e will be the same both for Scott's Pi and Krippendorff's Alpha coefficients:

$$p_e = 0.92^2 + (1 - 0.92)^2 \approx 0.83.$$

Thus,

$$Pi = \alpha = \frac{0.83 - 0.83}{1 - 0.83} = 0.$$

p_e and γ for AC1 will be

$$p_e = 2 \cdot 0.91(1 - 0.91) = 0.16,$$

$$\gamma = \frac{0.83 - 0.16}{1 - 0.16} = 0.8.$$

As we see this paradox is not unique to Cohen's kappa. Of the studied coefficients, Gwet's AC1 is the only one giving a reasonable estimate of agreement in this situation.

4.5.2 The second paradox

The second paradox includes such issue as symmetry of observations in the disagreement categories. It means that higher kappa may be produced by raters who disagree more on the marginal counts. We can illustrate this paradox by going through the example taken from Feinstein and Cicchetti (1990)):

	Rater 2		
Rater 1	Yes	No	Total
Yes	45	15	60
No	25	15	40
Total	70	30	100

Table 7: Distribution of 100 subjects by rater and category

For the situation in Table 7 the kappa calculation will be:

$$p_0 = (45 + 15)/100 = 0.6,$$

$$p_e = \frac{60}{100} \times \frac{70}{100} + \frac{40}{100} \times \frac{30}{100} = 0.54,$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0.6 - 0.54}{1 - 0.54} = 0.13.$$

	Rater 2		
Rater 1	Yes	No	Total
Yes	25	35	60
No	5	35	40
Total	30	70	100

Table 8: Distribution of 100 subjects by rater and category

For the situation in Table 8, the observed agreement has the same value as in Table 7, i.e $p_0 = 0.6$. Chance agreement and kappa coefficient will be:

$$p_e = \frac{60}{100} \times \frac{30}{100} + \frac{40}{100} \times \frac{70}{100} = 0.46,$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0.6 - 0.46}{1 - 0.46} = 0.26.$$

So we got that kappa value associated with second case is two times the kappa in the first one. Same as in first paradox, the problem lies in the calculation of the chance

agreement. It shows us that chance agreement is dependent of the marginal counts, hence kappa reacts sensitively and its value strongly depend on that.

Now we want to do the same as we did with first paradox and check if the second one holds to any of other presented agreement coefficients.

First, we find all the values for Table 7:

$$\pi_1 = (\frac{60}{100} + \frac{70}{100})/2 = 0.65,$$

then Scott's Pi and Krippendorff's Alpha will be:

$$p_e = 0.65^2 + (1 - 0.65)^2 \approx 0.55,$$

$$Pi = \alpha = \frac{0.6 - 0.55}{1 - 0.55} = 0.11.$$

For AC1 the result will be:

$$p_e = 2 \cdot 0.65(1 - 0.65) = 0.46,$$

$$\gamma = \frac{0.6 - 0.46}{1 - 0.46} = 0.26.$$

Then for the values from Table 8:

$$\pi_1 = (\frac{60}{100} + \frac{30}{100})/2 = 0.45,$$

Scott's Pi and Krippendorff's Alpha will be:

$$p_e = 0.45^2 + (1 - 0.45)^2 \approx 0.5,$$

$$Pi = \alpha = \frac{0.6 - 0.5}{1 - 0.5} = 0.2.$$

For AC1 the result will be:

$$p_e = 2 \cdot 0.45(1 - 0.45) = 0.5,$$

$$\gamma = \frac{0.6 - 0.5}{1 - 0.5} = 0.2.$$

To sum up: corresponding values for Scott's Pi, Krippendorff's Alpha and AC1 are 0.11, 0.11, 0.26 for the situation in Table 7, and 0.2, 0.2, 0.2 for Table 8 respectively.

Hence, we note the similar behaviour for Scott's and Krippendorff's coefficients as in case with kappa. The value got two times bigger after changing the marginals counts. While the value for Gwet's AC1 stayed almost unchangeable in both situations (we note actually a little decreasing in last one). This shows us that second paradox of kappa holds for Pi and Alpha as well, but does not hold the same way for AC1.

4.6 Illustration of measures of inter-rater agreement

We here introduce some definitions which were used to make some illustration of each type of measurement of inter-rater agreement for the 2×2 case (Qingshu Xie [11]).

Definitions:

- Observed agreement

$$p_0 = \frac{p_{11} + p_{22}}{n}.$$

- Prevalence index (PI). As it said in reference paper, PI was defined by Byrt, Bishop and Carlin (1993) as a difference in probabilities of given categories. It is estimated by the difference between $(p_{1.} + p_{.1})/2$ and $(p_{2.} + p_{.2})/2$.

$$PI = \frac{p_{1.} + p_{.1}}{2} - \frac{p_{2.} + p_{.2}}{2} =$$
$$\frac{\frac{p_{11}+p_{21}}{n} + \frac{p_{11}+p_{21}}{n}}{2} - \frac{\frac{p_{21}+p_{22}}{n} + \frac{p_{12}+p_{22}}{n}}{2} =$$
$$\frac{2p_{11} + p_{12} + p_{21}}{2n} - \frac{2p_{22} + p_{12} + p_{21}}{2n} = \frac{2p_{11} - 2p_{22}}{2n} = \frac{p_{11} - p_{22}}{n}.$$

  So, the prevalence index is

$$PI = \frac{p_{11} - p_{22}}{n}.$$

- Bias index (BI) is the difference in probabilities in the agreement category (category "Yes" in our case). It means difference between $p_{1.}$ and $p_{.1}$.

$$BI = p_{1.} - p_{.1} = \frac{p_{11} + p_{12}}{n} - \frac{p_{11} + p_{21}}{n} = \frac{p_{12} - p_{21}}{n}.$$

  So, the bias index is

$$BI = \frac{p_{12} - p_{21}}{n}.$$

4.6.1 Comparison of presented agreement coefficients

We take some basic example to compare all the agreement coefficients. Assume we have 100 subjects, and we want to look at the effect of prevalence on the coefficients. In the following illustration we used the indices introduced in above and such data: $n = 100$, $p_0 = 0.9$, $BI = 0.1$.

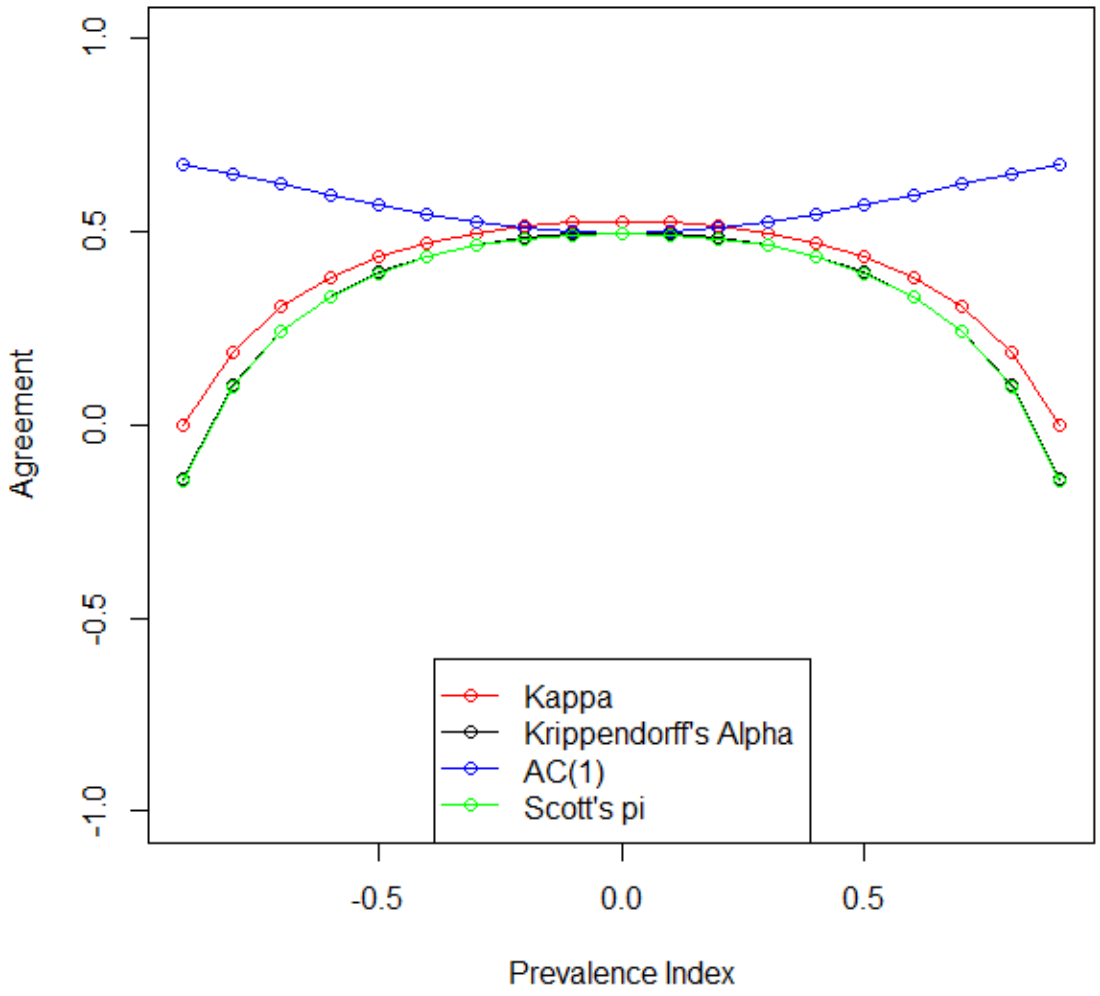

Figure 1: Comparison of all agreement coefficients

The figure 1 shows the behaviour of agreement coefficients as a function of prevalence index (PI). As we took a fixed rate at 90% observed agreement (p_0) and 10% a bias (BI), then we can see how the agreement coefficients are influenced by the effect of the

prevalence index.

We can clearly see that Cohen's kappa, Scott's Pi and Krippendorff's alpha coefficients decreases, when the absolute value of the prevalence index increases. Also we can see the equality of the Scott's Pi and Krippendoff's alpha coefficients, same as we showed it analytically before.

If we look at the figure, we note that all four coefficients illustrate an agreement with a very similar magnitude in "balanced data"($PI = 0$). When the absolute value of prevalence index increases, we can see the difference in the behaviour only for AC1, while three others behave as described above. The behaviour of AC1 diverges from the other three coefficients; while AC1 increases somewhat in magnitude, the others decrease drastically as $PI = |1|$.

5 Properties of AC1

Let us assume the example of an experiment where we have two raters (A and B) and k categories. The agreement table represented in terms of frequencies is given below.

	Rater A				
Rater B	1	2	...	k	Total
1	p_{11}	p_{12}	...	p_{1k}	$\mathbf{p_{1.}}$
2	p_{21}	p_{22}	...	p_{2k}	$\mathbf{p_{2.}}$
...	...	...	...	...	...
k	p_{k1}	p_{k2}	...	p_{kk}	$\mathbf{p_{k.}}$
Total	$\mathbf{p_{.1}}$	$\mathbf{p_{.2}}$	...	$\mathbf{p_{.k}}$	**1**

Table 9: Distribution of frequencies by rater and category

where the agreement matrix P is

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}$$

And remind from section 4.4 that AC1 coefficient is defined as

$$\gamma = \frac{p_0 - p_e}{1 - p_e},$$

where

$$p_0 = \sum_{i=1}^{k} \pi_{ii},$$

$$p_e = \frac{1}{k-1} \sum_{i=1}^{k} \pi_i (1 - \pi_i),$$

$$\pi_i = (p_{i.} + p_{.i})/2.$$

In the remaining, p_e is taken implicitly to mean the chance agreement pertaining to Gwet's AC1.

We would like to mention some interesting properties of π_i:

- π_i is a function of agreement matrix P, and it can be regarded as a prevalence of category i.
- π_i is a probability mass function. Namely $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{k} \pi_i = 1$.
- If P is diagonal, then $\pi_i = p_{ii}$, for $i = 1, \ldots, k$.

We need to remember that raters A and B "put" subjects in some category i and that each $p_{i.}$ and $p_{.i}$ define the number of those subjects. If both raters classify subjects in the same category i, then the number of those will be defines as p_{ii} not related but assigned. So this gives us that π_i will define the probability of some random subjects being related to category i by some random rater.

Gwet's theory

Gwet bases his theory about chance agreement measure on the notion that subjects can be Hard or Easy to classify:

- **Easy (for the rater) subjects are** *always classified correctly*.
- **If for a rater the subject is hard $\Longrightarrow$ it is assigned randomly** *uniformly* **into any of the available categories.**

5.1 Properties of chance agreement p_e

We want to introduce three main properties of p_e:

1) p_e is a degree of uniformity.

"Subjects distributed more uniformly across categories are more likely to contain H-subjects" (Gwet 2014, p.116)[1].

2) $0 \leq p_e \leq \frac{1}{k}$.

Let us remind the formulation of chance agreement:

$$p_e = \frac{1}{k-1} \sum_{i=1}^{k} \pi_i(1 - \pi_i). \tag{8}$$

And let us find the minimum and maximum of p_e. The minimum value of equation (8) is 0, it is easy to show. If we, for example, take the case with following agreement table

Rater B	Rater A 1	2	Total
1	1	0	1
2	0	0	0
Total	1	0	1

Table 10: Distribution of frequencies by rater and category

then from Table 10 it is obviously that $\pi_1 = 1$, hence $p_e = 2\pi_1(1 - \pi_1) = 0$. So, as we said, minimum of $p_e = 0$.

Lemma 1. *Maximal p_e is attained, if $\pi_i = \frac{1}{k}$, for $\forall i = 1, 2, ..., k$.*

Proof. If we take the equation (8) and open the parentheses, we get:

$$p_e = \frac{1}{k-1}\left(\sum_{i=1}^{k}\pi_i - \sum_{i=1}^{k}\pi_i^2\right).$$

Note that

$$\sum_{i=1}^{k}\pi_i = 1,$$

therefore

$$p_e = \frac{1}{k-1}\left(1 - \sum_{i=1}^{k}\pi_i^2\right).$$

Note that $\sum\limits_{i=1}^{k}\frac{1}{k} = 1$.

Consider

$$\left(\frac{1}{k} + a_1\right)^2 + \left(\frac{1}{k} + a_2\right)^2 + ... + \left(\frac{1}{k} + a_k\right)^2, \tag{9}$$

such that $(\frac{1}{k} + a_1) + (\frac{1}{k} + a_2) + ... + (\frac{1}{k} + a_k) = 1$, i.e. $\sum\limits_{i=1}^{k} a_k = 0$.

Then (9) is equivalent to

$$\frac{1}{k^2} + \frac{2a_1}{k} + a_1^2 + \frac{1}{k^2} + \frac{2a_2}{k} + a_2^2 + ... + \frac{1}{k^2} + \frac{2a_k}{k} + a_k^2 =$$

$$\sum_{i=1}^{k}\frac{1}{k^2} + \frac{2}{k}\sum_{i=1}^{k}a_k + \sum_{i=1}^{k}a_k^2.$$

We know that $\sum_{i=1}^{k} a_k^2 \geq 0$. On the other hand $\sum_{i=1}^{k} a_k = 0$, thus $\frac{2}{k}\sum_{i=1}^{k} a_k = 0$.

Therefore we can conclude that $\sum_{i=1}^{k} \frac{1}{k^2} + \frac{2}{k}\sum_{i=1}^{k} a_k + \sum_{i=1}^{k} a_k^2 \geq 0 \implies$ expression (9) is always bigger than or equal to 0. It means that for $\sum_{i=1}^{k} \pi_i^2$ to be minimal, the π_i has to be equal to $\frac{1}{k} \implies \sum_{i=1}^{k} \pi_i^2 = \sum_{i=1}^{k} \frac{1}{k^2}$.

Hence,

$$p_e \leq \frac{1}{1-k}\left(1 - \sum_{i=1}^{k} \frac{1}{k^2}\right) = \frac{1}{k-1}\left(1 - \frac{1}{k}\right) = \frac{1}{k-1} \cdot \frac{k-1}{k} = \frac{1}{k}.$$

□

Corollary 1. *p_e tends to 0 as k grows. Therefore $\gamma \approx p_0$, if k is big enough.*

When $\gamma \approx p_0$, AC1 is no longer a chance-adjusted agreement coefficient.

Corollary 2. (AC1 paradox) *Value of AC1 grows, when we add an "empty" category.*

Assume we have an agreement matrix:

$$P_2 = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix} \implies \gamma(P_2) = 0.6.$$

Now take the same example matrix and add one "empty" category. We expect to get a similar value for AC1 as for P_2, but:

$$P_3 = \begin{pmatrix} 0.4 & 0.1 & 0 \\ 0.1 & 0.4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \implies \gamma(P_3) = 0.73.$$

We see that the value became higher. We continue with couple more examples to see the tendency of behaviour of AC1:

$$P_4 = \begin{pmatrix} 0.4 & 0.1 & 0 & 0 \\ 0.1 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \implies \gamma(P_4) = 0.76.$$

For number of categories $k = 8$:

$$P_8 = \begin{pmatrix} 0.4 & 0.1 & 0 & \dots & 0 \\ 0.1 & 0.4 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \implies \gamma(P_8) = 0.78.$$

So, we note that the more "empty" categories we add, the higher values of AC1 we get.

3) For a given p_0, maximum p_e is attained if the agreement matrix P has the following form: all diagonal elements are equal, i.e $p_{ii} = p_0/k$, and all non-diagonal elements are equal to each other as well, i.e. $p_{ij} = p_{ji} = (1 - p_0)/(k(k-1))$.

Proof. We need to show that to attain maximal p_e, the matrix P have to be in the following form:

$$P^* = \begin{pmatrix} \frac{p_0}{k} & \frac{1-p_0}{k(k-1)} & \cdots & \frac{1-p_0}{k(k-1)} \\ \frac{1-p_0}{k(k-1)} & \frac{p_0}{k} & \cdots & \frac{1-p_0}{k(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-p_0}{k(k-1)} & \frac{1-p_0}{k(k-1)} & \cdots & \frac{p_0}{k} \end{pmatrix}$$

We want to prove that maximum chance agreement p_e for the given agreement matrix above will be $\frac{1}{k}$ as we showed in property 2).

Let us take a general form of agreement matrix and divide it into two matrices:

- P_d - includes only diagonal elements while all others are equal to 0,
- P_{nd} - includes all non-diagonal elements while diagonal elements are equal to 0.

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix} = P_d + P_{nd} =$$

$$= \begin{pmatrix} p_{11} & 0 & \ldots & 0 \\ 0 & p_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_{kk} \end{pmatrix} + \begin{pmatrix} 0 & p_{12} & \ldots & p_{1k} \\ p_{21} & 0 & \ldots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \ldots & 0 \end{pmatrix}$$

Following the same logic as in proof of Lemma 1, it is clear that maximum values of P_d is attained, if all the summation terms (which in case of diagonal matrix will be the diagonal elements) are equal. Same can be said about non-diagonal elements. They also should be all equal to attain the maximum of P_{nd}.

From $p_0 = \sum_{i=1}^{k} p_{ii}$ follows that $p_{kk} = \frac{p_0}{k}$.

Sum of all elements in P is given by:

$k \cdot$ (elements of $P_d + (k-1) \cdot$ elements of $P_{nd}) = 1 \Longrightarrow$
$(k-1) \cdot$ elements of $P_{nd} = \frac{1}{k} - \frac{p_0}{k} \Longrightarrow$
elements of $P_{nd} = \frac{1-p_0}{k(k-1)}$.
Thus, we have a new form for matrices P_d and P_{nd}:

$$P_d + P_{nd} = \begin{pmatrix} \frac{p_0}{k} & 0 & \dots & 0 \\ 0 & \frac{p_0}{k} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p_0}{k} \end{pmatrix} + \begin{pmatrix} 0 & \frac{1-p_0}{k(k-1)} & \dots & \frac{1-p_0}{k(k-1)} \\ \frac{1-p_0}{k(k-1)} & 0 & \dots & \frac{1-p_0}{k(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-p_0}{k(k-1)} & \frac{1-p_0}{k(k-1)} & \dots & 0 \end{pmatrix}$$

And now we got a new agreement matrix in the form

$$P^* = \begin{pmatrix} \frac{p_0}{k} & \frac{1-p_0}{k(k-1)} & \dots & \frac{1-p_0}{k(k-1)} \\ \frac{1-p_0}{k(k-1)} & \frac{p_0}{k} & \dots & \frac{1-p_0}{k(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-p_0}{k(k-1)} & \frac{1-p_0}{k(k-1)} & \dots & \frac{p_0}{k} \end{pmatrix}$$

Let us find the chance agreement for P^*. We can first find the value for π_i:

$$\pi_i = \frac{p_{i.} + p_{.i}}{2} = \frac{2 \cdot \left(\frac{p_0}{k} + (k-1) \cdot \frac{1-p_0}{k(k-1)} \right)}{2} = \frac{p_0}{k} + \frac{1-p_0}{k} = \frac{1}{k}.$$

So we got $\pi_i = \frac{1}{k} \Longrightarrow p_e = \frac{1}{k}$, which is maximum value for p_e.

□

6 Variance of Gwet's AC1

Let γ define the Gwet's AC1 coefficient. Remind from section 4.4 that

$$\gamma = \frac{p_0 - p_e}{1 - p_e},$$

which is the same formulation as for kappa with p_0 and p_e observed and percent chance agreement respectively.

The variance of the AC1 coefficient for two-rater reliability experiments was presented by Gwet(2008a) and it is given by:

$$Var(\gamma) = \frac{1}{n(1-p_e)^2}\left[p_0(1-p_0) - 4(1-\gamma)\left(\frac{1}{k-1}\sum_{i=1}^{k} p_{ii}(1-\pi_i) - p_0 p_e\right) + \right.$$
$$\left. +4(1-\gamma)^2\left(\frac{1}{(k-1)^2}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}[1-(\pi_i+\pi_j)/2]^2 - p_e^2\right)\right], \qquad (10)$$

where
n - number of observations,
k - number of categories,
$p_{i.}$ and $p_{.i}$ - marginal frequencies from agreement table, $\pi_i = \frac{p_{i.}+p_{.i}}{2}$.

Asymptotic normality

Gwet mentions asymptotic normality of AC(1) in his article (2008)[7] "It can be shown that $v(\hat{\gamma_k}|Agreement)$ captures all terms of magnitude order up to n^{-1}, is consistent for estimating the true population variance and provides valid normality-based confidence intervals when the number of participants is reasonably large."

Example

We want to see how the standard error and the confidence interval for gamma coefficient behave in case of small and big number of observations. We take a look at basic example where we simulate some data: $n = 40$ and $n = 300$ to see the difference, $p_0 = 0.9$, $BI = 0.1$.

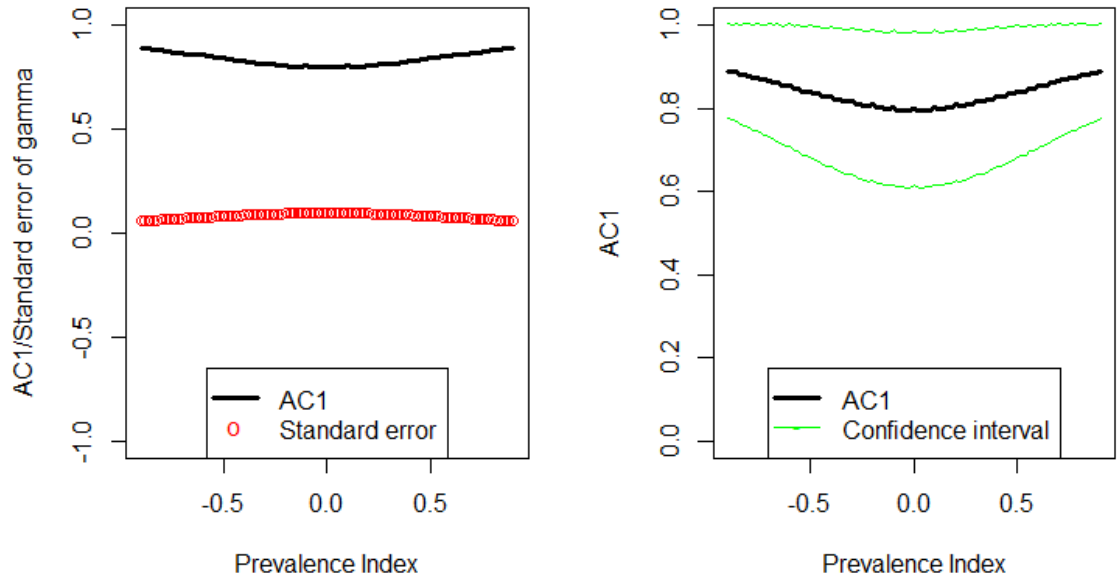

Figure 2: Illustration of standard error of AC1 and 95% confidence interval for gamma with $n = 40$

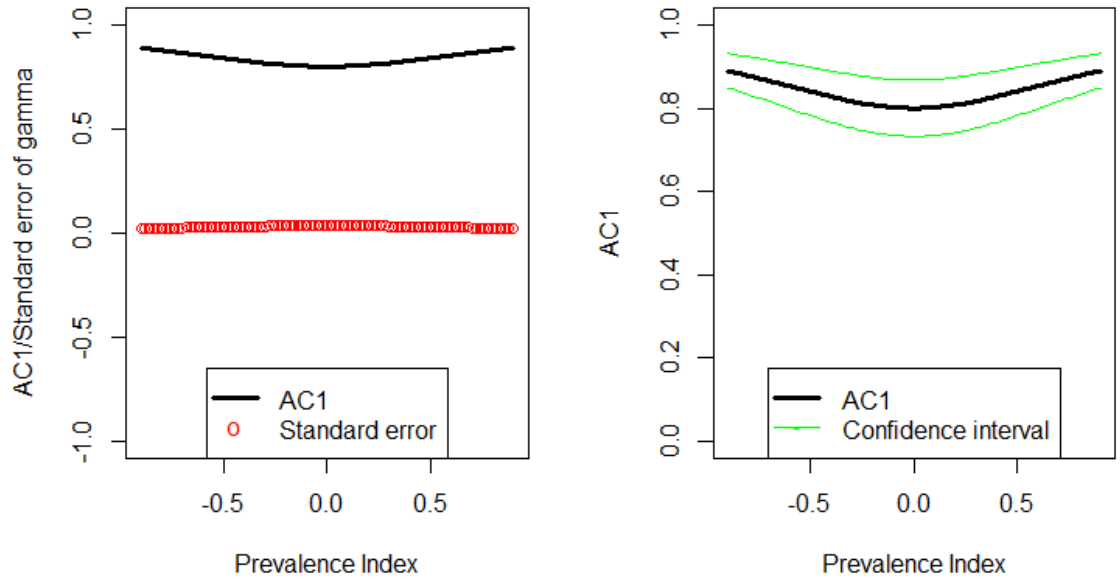

Figure 3: Illustration of standard error of AC1 and 95% confidence interval for gamma with $n = 300$

We note that standard error is dependent on sample size as expected. There is also a dependence of the effect of the prevalence index. The larger number of observations we have, the smaller we get the standard error. Same happens with the confidence interval. If we calculate the kappa for small number of observations, we get CI quite wide.

So we can conclude that the bigger the sample size, the smaller confidence interval it can produce, which gives us that the estimate of agreement is very accurate.

6.1 Variance of Kappa

We want to compare the behaviour of variance for gamma with the kappa agreement coefficient. Let us remind the Cohen's kappa coefficient:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}.$$

The formulation of variance was presented by Gwet (2008a) and it is equivalent to the formulation of Fleiss, Cohen and Everitt(1969). The last ones represented the formula for estimated standard error of κ. As we know from basic $Sd(\kappa) = \sqrt{Var(\kappa)}$.

$$se(\kappa) = \frac{\sqrt{A + B - C}}{(1 - p_e)\sqrt{n}}. \tag{11}$$

This formulation was taken from Fleiss, Levin and Paik[2] (2003), where

$$A = \sum_{i=1}^{k} p_{ii} \left[1 - (p_{i.} + p_{.i})(1 - \kappa)\right]^2,$$

$$B = (1 - \kappa)^2 \sum \sum_{i \neq j} p_{ij} (p_{.i} + p_{j.})^2,$$

$$C = [\kappa - p_e(1 - \kappa)]^2.$$

Approximate $100(1 - \alpha)\%$ confidence interval for kappa is

$$\kappa - z_{\alpha/2} se(\kappa) \leq \kappa \leq \kappa + z_{\alpha/2} se(\kappa).$$

D. Altman [3] proposed the following approximation of the variance:

$$Var(\kappa) \approx \frac{p_0(1 - p_0)}{n(1 - p_e)^2}. \tag{12}$$

This estimate was first presented in Cohen, 1960. According to Fleiss, Cohen and Everitt (1969), formula (11) is more accurate and is based on a better theoretical foundation. However, according to Altman [3], formula (12) is often used during design planning due to its simplicity. Unfortunately, (12) is not an upper bound and it can underestimate the real variance when there is disbalance in prevalences of categories. A following example demonstrate it.

Example

We do some simulations to run the example by using formulas given by Fleiss, Cohen and Everit from above for the same data as in previous example:

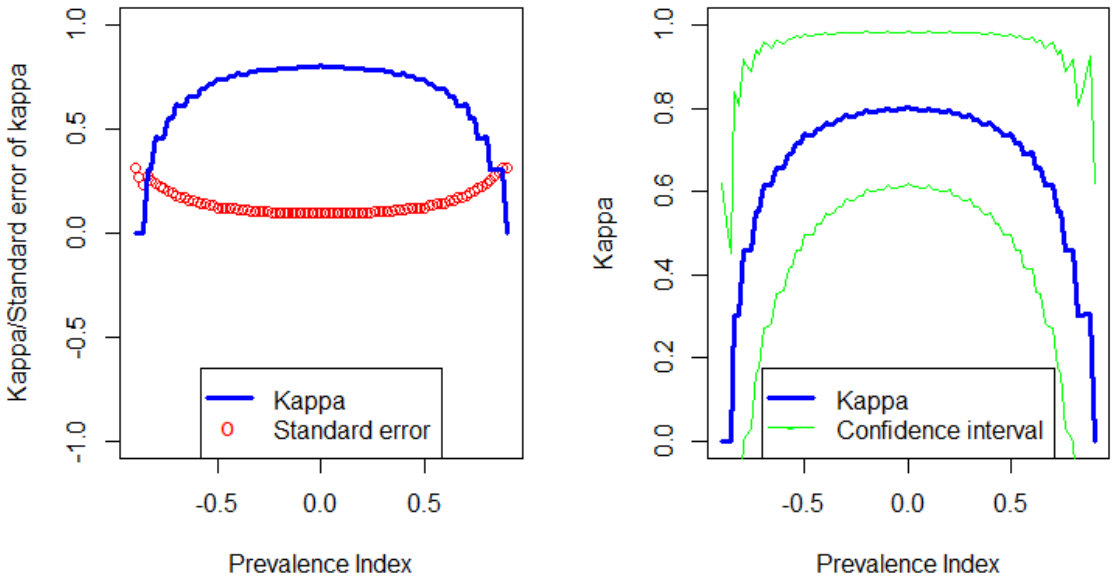

Figure 4: Illustration of standard error of kappa and 95% confidence interval for kappa with $n = 40$

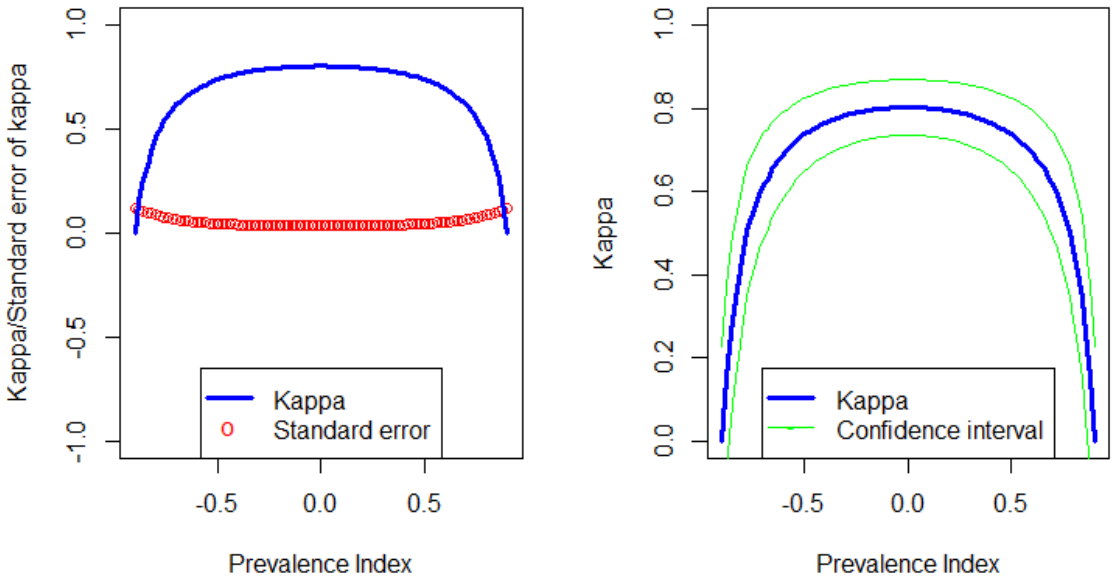

Figure 5: Illustration of standard error of kappa and 95% confidence interval for kappa with $n = 300$

We note the same behaviour of kappa coefficient as we did in AC1 ones, which depends on sample size and the prevalence index. The larger number of observation we use, the less standard error we get and smaller CI will be produced. Though, we notice that kappa has a larger standard error and wider confidence interval than AC1 coefficient.

6.2 Upper bound for $Var(\gamma)$

Now we will obtain an upper bound for variance of AC1. Variance for Gwet's coefficient was presented in (10) and, using definition of γ in (6), it can be rewritten as:

$$Var(\gamma) = \frac{1}{n(1-p_e)^2}\left\{p_0(1-p_0) - 4\,\frac{1-p_0}{1-p_e}\left(\frac{1}{k-1}\sum_{i=1}^{k} p_{ii}(1-\pi_i) - p_0 p_e\right)\right.$$
$$\left. +4\,\frac{(1-p_0)^2}{(1-p_e)^2}\left(\frac{1}{(k-1)^2}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}[1-(\pi_i+\pi_j)/2]^2 - p_e^2\right)\right\}. \tag{13}$$

Expression in the parentheses (13) is a sum of three terms:

$$T_1 = p_0(1-p_0), \tag{14}$$

$$T_2 = -\frac{4(1-p_0)}{(1-p_e)}\left(\frac{1}{k-1}\sum_{i=1}^{k} p_{ii}(1-\pi_i) - p_0 p_e\right), \tag{15}$$

$$T_3 = \frac{4(1-p_0)^2}{(1-p_e)^2}\left(\frac{1}{(k-1)^2}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}[1-(\pi_i+\pi_j)/2]^2 - p_e^2)\right). \tag{16}$$

We will search for upper bounds for T_1, T_2, T_3 separately.

Upper bound of T_1

T_1 takes its maximum value at $p_0 = 1/2$ which is equal to $1/4$.

Upper bound of T_2

Substitute the definition of p_e into the expression under the parentheses of T_2 and multiply by $(k-1)$:

$$\sum_{i=1}^{k} p_{ii}(1-\pi_i) - p_0\pi_i(1-\pi_i).$$

Assume without loss of generality that $p_0 \neq 0$. Then p_0 can be also taken out of parentheses:

$$p_0\left\{\sum_{i=1}^{k}\frac{p_{ii}(1-\pi_i)}{p_0} - \sum_{i=1}^{k}\pi_i(1-\pi_i)\right\}. \tag{17}$$

Recall that $p_0 = \sum_{i=1}^{k} p_{ii}$ and all $p_{ii} \geq 0$. Denote $\tilde{p_i} = \frac{p_{ii}}{p_0}$.

Upper bound of T_2 can be obtained by finding the maximum of the expression

$$\sum_{i=1}^{k} \tilde{p_i}\pi_i - \sum_{i=1}^{k} \pi_i^2, \tag{18}$$

where $\tilde{p_i}$ and π_i are probability mass functions of distributions that are not related to each other.

Lemma 2. *The maximum of* $\sum_{i=1}^{k} \tilde{p_i}\pi_i - \sum_{i=1}^{k} \pi_i^2$ *is equal to* $\frac{1}{4} - \frac{1}{4k}$.

Proof. Assume without loss of generality $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_k$, then expression (18) less or equal to

$$p_1\pi_1 - \pi_1^2 - \sum_{i>1} \pi_i^2. \tag{19}$$

Using the result of Lemma 1 (chapter 5.1), we conclude that minimum of the last term of equation (19), $\sum_{i>1} \pi_i^2$, is attained, if all π_i are equal (for $i > 1$), i.e. $\pi_i = \frac{1-\pi_1}{k-1}$.

Maximum of (19) is attained for $\tilde{p_1} = 1$ and, by inserting value for π_i, we get

$$\pi_1 - \pi_1^2 - (k-1)\frac{(1-\pi_1)^2}{(k-1)^2} = \pi_1 - \pi_1^2 - \frac{(1-\pi_1)^2}{k-1} =$$
$$\frac{k\pi_1 - \pi_1 - k\pi_1^2 + \pi_1^2 - 1 + 2\pi_1 - \pi_1^2}{k-1} =$$
$$\frac{1}{k-1}\left(-\pi_1^2 k + \pi_1(k+1) - 1\right). \tag{20}$$

Expression (20) is a quadratic function of π_1 which is a parabola pointing downwards. Thus, the maxim of (19) is attained at π_1 maximal, which is $\pi_1 = \frac{k+1}{2k}$, and is equal to

$$\frac{1}{k-1}\left(-k\big(\frac{k+1}{2k}\big)^2 + (k+1)\frac{k+1}{2k} - 1\right) =$$
$$\frac{1}{k-1}\left(-\frac{(k+1)^2}{4k} + \frac{(k+1)^2}{2k} - 1\right) = \frac{1}{k-1} \cdot \frac{(k+1)^2 - 4k}{4k} =$$
$$\frac{1}{k-1} \cdot \frac{k^2 - 2k + 1}{4k} = \frac{k-1}{4k} = \frac{1}{4} - \frac{1}{4k}.$$

□

Finally

$$T_2 \leq \frac{p_0(1-p_0)}{1-p_e}\big(1 - \frac{1}{k}\big)\big(\frac{1}{k-1}\big). \tag{21}$$

Upper bound of T_3

Theorem 1.

$$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[1-(\pi_i+\pi_j)/2\right]^2 - p_e^2 \leq \sum_{i=1}^{k}\pi_i^3 - \Big(\sum_{i=1}^{k}\pi_i^2\Big)^2.$$

Proof. We start with analysis of T_3 that is expression under the parentheses:

$$\frac{1}{(k-1)^2}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}[1-(\pi_i+\pi_j)/2]^2 - p_e^2. \tag{22}$$

Lemma 3.

$$\frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}[1-(\pi_i+\pi_j)/2] = p_e.$$

Proof. (of Lemma 3)

The left hand side of the equation we want to prove is

$$\sum_{i=1}^{k}\sum_{j=1}^{k}[p_{ij}-p_{ij}(\pi_i+\pi_j)/2] = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}(\pi_i+\pi_j)/2,$$

while the right hand side is equal to

$$\sum_{i=1}^{k}[\pi_i-\pi_i^2] = \sum_{i=1}^{k}\pi_i - \sum_{i=1}^{k}\pi_i^2.$$

Since $\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{k} p_{ij} = 1$ and $\sum\limits_{i=1}^{k}\pi_i = 1$, it is enough to show that $\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{k} p_{ij}(\pi_i+\pi_j)/2$ is equal to $\sum\limits_{i=1}^{k}\pi_i^2$.

First we simplify the left side of the equation by using the formula (7) for π and opening the sums:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\frac{\pi_i+\pi_j}{2} = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\frac{p_{i.}+p_{.i}+p_{j.}+p_{.j}}{2\cdot 2} =$$

$$\frac{1}{4}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[\sum_{m=1}^{k} p_{im} + \sum_{l=1}^{k} p_{li} + \sum_{g=1}^{k} p_{jg} + \sum_{s=1}^{k} p_{sj}\right] =$$

$$\frac{1}{4}\left[\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{m=1}^{k} p_{ij}p_{im} + \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{k} p_{ij}p_{li} + \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{g=1}^{k} p_{ij}p_{jg} + \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{s=1}^{k} p_{ij}p_{sj}\right].$$

And then do the same work with the right hand side:

$$\sum_{i=1}^{k} \pi_i^2 = \sum_{i=1}^{k} \left(\frac{p_{i.} + p_{.i}}{2} \right)^2 = \sum_{i=1}^{k} \left[\frac{1}{2} (\sum_{j=1}^{k} p_{ij} + \sum_{l=1}^{k} p_{li}) \right]^2 =$$

$$\frac{1}{4} \sum_{i=1}^{k} p_i \left[\left(\sum_{j=1}^{k} p_{ij} \right)^2 + 2 \left(\sum_{j=1}^{k} p_{ij} \right) \left(\sum_{l=1}^{k} p_{li} \right) + \left(\sum_{l=1}^{k} p_{li} \right)^2 \right] =$$

by using the expression $\left(\sum_{i=1}^{k} a_i \right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j$, we can now open the parentheses

$$= \frac{1}{4} \left[\sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{m=1}^{k} p_{ij} p_{im} + 2 \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{l=1}^{k} p_{ij} p_{li} + \sum_{i=1}^{k} \sum_{l=1}^{k} \sum_{s=1}^{k} p_{li} p_{si} \right].$$

□

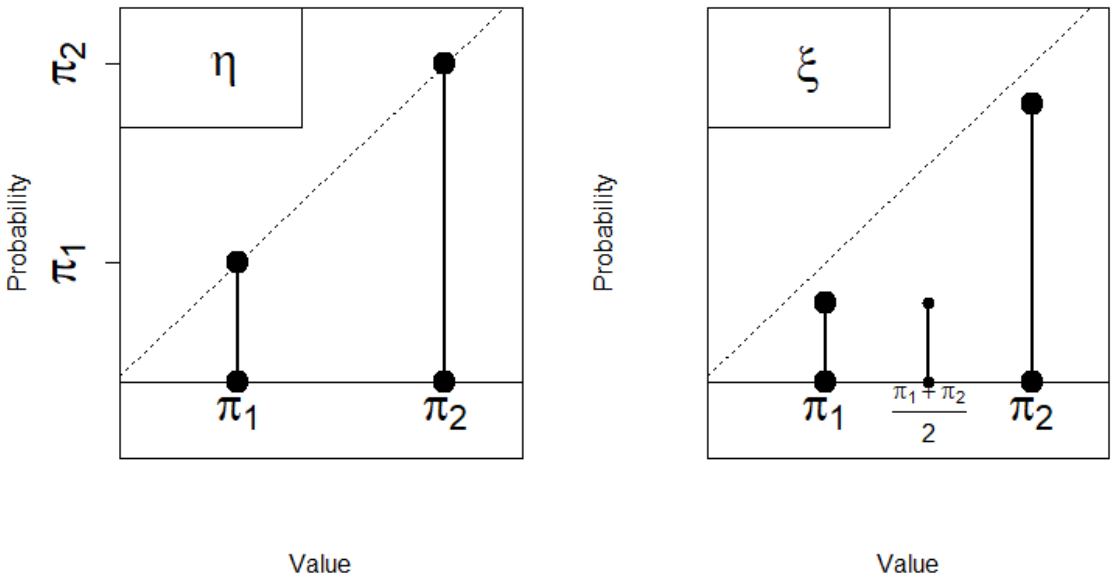

Figure 6: Illustration of value and probability of variables η and ξ for particular case $k = 2$.

Let us get back to the expression (22). Suppose that all $(\pi_i + \pi_j)/2$ are all not equal (for $i < j$). Let us introduce a random variable ξ

$$\xi: \quad \begin{array}{l|l|l|l} \text{Value} & (\pi_i + \pi_j)/2 & \text{for } i < j & \pi_i \text{ for } i = 1, \ldots, k \\ \hline \text{Probability} & p_{ij} + p_{ji} & \text{for } i < j & p_{ii} \text{ for } i = 1, \ldots, k \end{array}$$

We can show that $\sum_{i=1}^{k}\sum_{j=i+1}^{k}\{p_{ij}+p_{ji}+p_{ii}\}=1$. Indeed, since $i<j$, then p_{ij} are the elements that lie below the diagonal, p_{ji} are the elements that lie above the diagonal and p_{ii} are the diagonal elements in the agreement matrix (and the sum of all elements of an agreement matrix must be equal to 1).

Since $Var(\xi)=E\xi^2-(E\xi)^2$ and $Var(1-\xi)=Var(\xi)$, Lemma 3 proves that expression (22) is $Var(\xi)\cdot\frac{1}{(k-1)^2}$.

Suppose that all π_i are not equal and let us introduce another discrete random variable η defined as

η:

Value	π_i for $i=1,\ldots,k$
Probability	π_i

From Lemma 3 it follows that $E\xi=E\eta$.

□

Lemma 4.

$$Var(\xi)\leq Var(\eta)$$

Proof. (of Lemma 4)

For an agreement matrix P, the corresponding ξ will take a value π_i with probability equal to the diagonal element p_{ii}, and $(\pi_i+\pi_j)/2$ with probability $p_{ij}+p_{ji}$ (for $i>j$).

In the limiting case of the diagonal agreement matrix P, definitions of ξ and η coincide, therefore $Var(\xi)=Var(\eta)$.

Let us now consider a stepwise process that transforms diagonal agreement matrix to a general form agreement matrix with the same π_i, $\forall i=1,\ldots,k$. We will show that on each step variance of corresponding ξ is less or equal to variance of corresponding η.

The process is such that

- Start from the diagonal form matrix
- On each step insert elements p_{ij} and p_{ji} so that all other elements in the matrix except from the diagonal ones p_{ii} and p_{jj} remain the same.
- All π_l, $l=1,\ldots,k$ remain unchanged.

Let us consider what the conditions of the stepwise process imply. For $i<j$ we have that

$$\pi_i=p_{ii}+\frac{p_{ij}+p_{ji}}{2}+\text{other elements},$$
$$\pi_j=p_{jj}+\frac{p_{ij}+p_{ji}}{2}+\text{other elements}.$$

If $p_{ij}8{+}p_{ji}$ changes from 0 to δ and π_i remains unchanged (as well as other elements), then p_{ii} has to be decreased by $\delta/2$, the same is true for p_{jj}.

Suppose that $\mu = E\eta$. Suppose further that d_i and d_j are distances between μ and π_i (μ and π_j) respectively. Than it is possible to show that the distance between μ and $(\pi_i + \pi_j)/2$ equals to $|d_j - d_i|/2$.

For a discrete random variable, the variance is equal the sum of squared deviations from its expected value weighted by the corresponding probabilities. On each step of our stepwise process the corresponding items in the sum change from

$$p_{ii} \cdot d_i^2 + p_{jj} \cdot d_j^2$$

to

$$(p_{ii} - \frac{\delta}{2}) \cdot d_i^2 + (p_{jj} - \frac{\delta}{2}) \cdot d_j^2 + \delta((d_j - d_i)/2)^2 =$$

$$p_{ii} \cdot d_i^2 - \frac{\delta d_i^2}{2} + p_{jj} \cdot d_j^2 - \frac{\delta d_j^2}{2} + \frac{\delta d_j^2}{4} - \frac{\delta d_j d_i}{2} + \frac{\delta d_i^2}{4} =$$

$$p_{ii} \cdot d_i^2 + p_{jj} \cdot d_j^2 - \frac{\delta d_i^2}{4} - \frac{\delta d_j^2}{4} - \frac{\delta d_j d_i}{2} \leq p_{ii} \cdot d_i^2 + p_{jj} \cdot d_j^2$$

Thus, $Var(\xi) \leq Var(\eta)$.

Remark 1

For the ease of proof, we made an assumption that all $(\pi_i + \pi_j)/2$ are not equal (for $i < j$). But it is easy to see that the main result still holds if this requirement is relaxed.

Indeed, if it happens that $\frac{\pi_{i1}+\pi_{j1}}{2} = \frac{\pi_{i2}+\pi_{j2}}{2}$ (where $i_1 \neq i_2$ and either $j_1 = j_2$ or $j_1 \neq j_2$), then the impact of adding $\frac{\pi_{i1}+\pi_{j1}}{2}$ into the set of values for random variable ξ will be considered twice (and on each step it will not increase the variance).

Remark 2

The assumption that all π_i-s are not equal can also be relaxed. Thus, if $\pi_i = \pi_j$, the in the definition of η the value π_{i_1} has probability $2\pi_{i_1}$, and it is still true that $E\eta = \sum_{i=1}^{k} \pi_i^2$.

The stepwise process described above is still applicable, and in the case of $\pi_i = \pi_j$, the variance of ξ on the corresponding step will not change the variance.

The expectation of variable η will be

$$E(\eta) = \sum_{i=1}^{k} \pi_i^2,$$

and

$$E(\eta^2) = \sum_{i=1}^{k} \pi_i^2 \pi_i.$$

Thus, we can write down the expression for variance for variable η:

$$Var(\eta) = E(\eta^2) - (E(\eta))^2 = \sum_{i=1}^{k} \pi_i^3 - \left(\sum_{i=1}^{k} \pi_i^2 \right)^2 . \tag{23}$$

□

Now we want to check how the maximum of expression $\frac{Var(\eta)}{(k-1)^2}$ behaves for different values of k. We run a numerical simulation as follows: we consider all possible combinations of π_i such that they take values on a grid from 0 to 1 with a step 0.01 and the sum is equal to 1. Find the maximum values for $k = 2, \ldots, 5$.

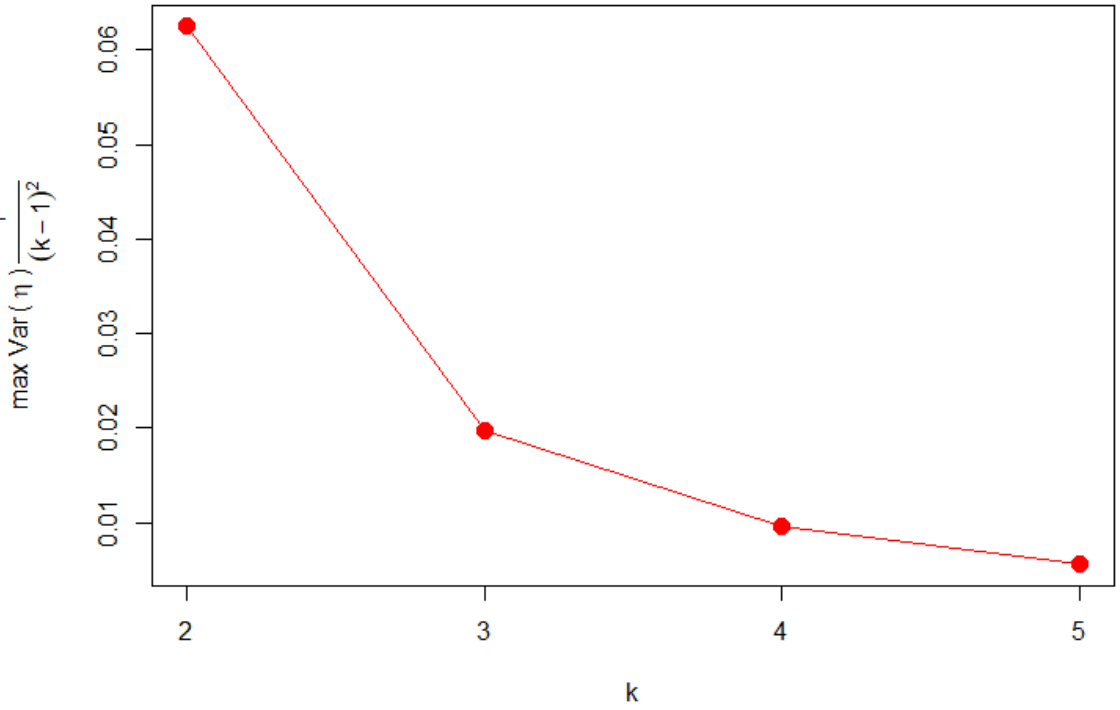

Figure 7: Maximum of $Var(\eta)\frac{1}{(k-1)^2}$ for 2, 3, 4 and 5 categories

k	2	3	4	5
$\max Var(\eta)\frac{1}{(k-1)^2}$	0.062475	0.015619	0.006942	0.003905

We can see that the more categories we have, the less value this maximum takes. Apparently $max(\frac{Var(\eta)}{(k-1)^2})$ converges to 0, when the number of categories k grows. It means that expression (23) takes so small values that we can eventually just neglect it.

6.2.1 Conservative upper bound

Now we can write down the main formula for upper bound of variance of gamma by taking into account all simplifications we made above for each term (T_2, T_2, T_3). Also we have to remember to take the maximum value for chance agreement, p_e, which was found in Lemma 1 (chapter 5.1). Thus,

$$Var(\gamma) < \frac{k^2(1-p_0)}{n(k-1)^2}\left\{p_0\big(1+\frac{1}{k-1}\big)+(1-p_0)\cdot\tilde{C}_k\right\}, \tag{24}$$

where $\tilde{C}_k$ takes values:

k	2	3	4	5
$\tilde{C}_k$	0.999600	0.176976	0.067925	0.035125

For a larger number of categories, the values for $\tilde{C}_k$ will be small that we can neglect it and obtain the following upper bound:

$$Var(\gamma) \leq \frac{k^2(1-p_0)}{n(k-1)^2}\left\{p_0\big(1+\frac{1}{k-1}\big)\right\}. \tag{25}$$

Generally, for a 2×2 agreement formula (24) simplifies to

$$Var(\gamma) < \frac{4(1-p_0)}{n}\left\{2p_0+0.9996\cdot(1-p_0)\right\}. \tag{26}$$

Let us compare the main result (24) for upper bound for $Var(\gamma)$ with the full expression (10). We take the same example as we did in chapter 6 with number of observations equal to 300.

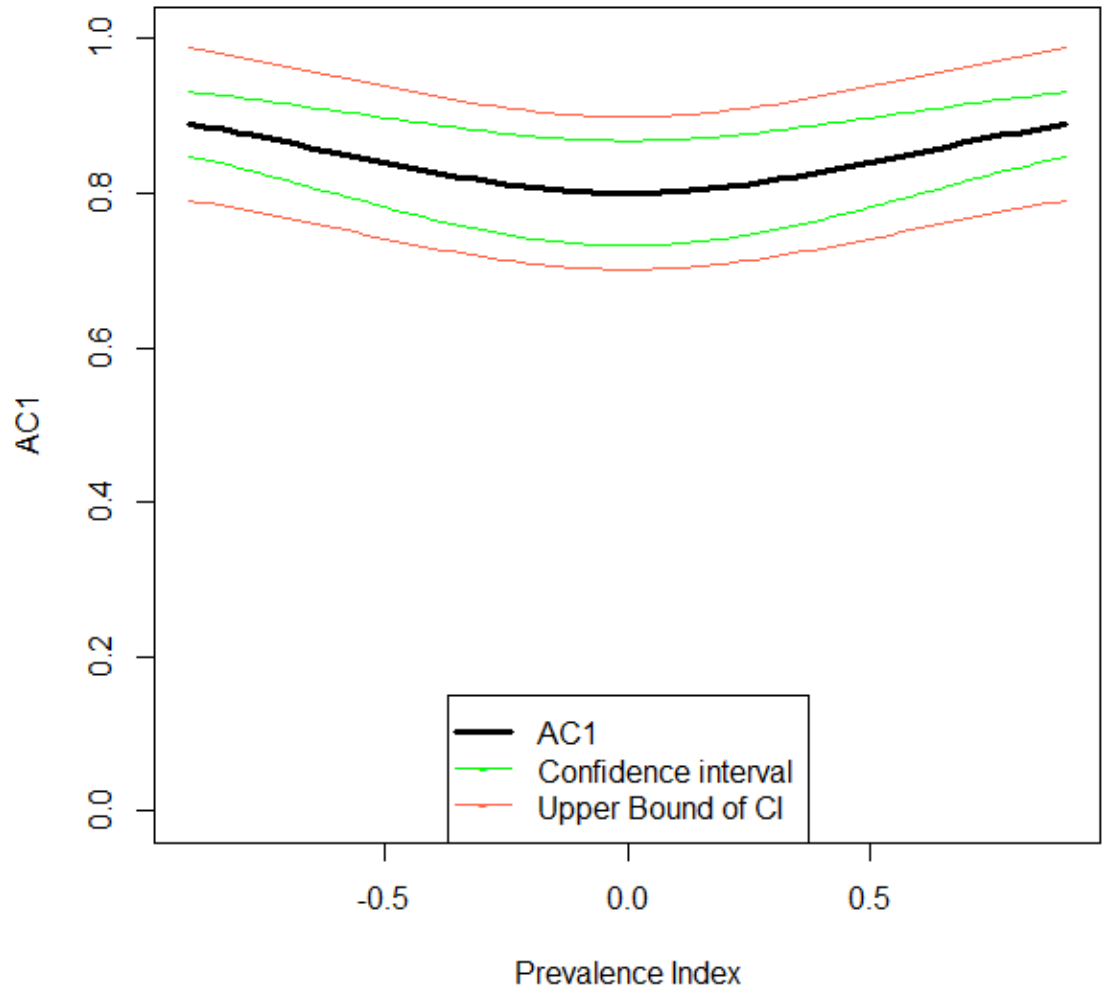

Figure 8: Upper bound for 95% confidence interval for gamma with $n = 300$

6.2.2 Improvement of upper bound

We found a formal strict upper bound of variance and considered the three sources of variation that contribute to its magnitude. Thus, T_1 is the part of variation that is due to p_0 only, T_3 is the part of variation that is due to the prevalence of categories. As for T_2, it incorporates the percent agreement and prevalences, and there is a dependency between T_2 and T_3. In addition, the upper bounds for T_2 and T_3 are strict, since they are based on mutually exclusive special cases. This motivates us to search for an improved upper bound.

Proposition

For $k \leq 5$ the obtained upper bound for T_3 can be excluded from the expression of the conservative upper bound (24).

We show this by numerical simulation. For $k = 2, \ldots, 5$ we generate an agreement matrix using $Uniform(0, 1)$ distribution and divide each element of matrix by the sum of all elements, so that $\sum_{i=1}^{k} p_{ij} = 1$. This is done 10000 times for each k and $T_2 + T_3$ is compared with the upper bound of T_2. The result is that $T_2 + T_3$ is always less than $p_0(1 - p_0)/(k - 1)$. The result is presented graphically in Figure 9.

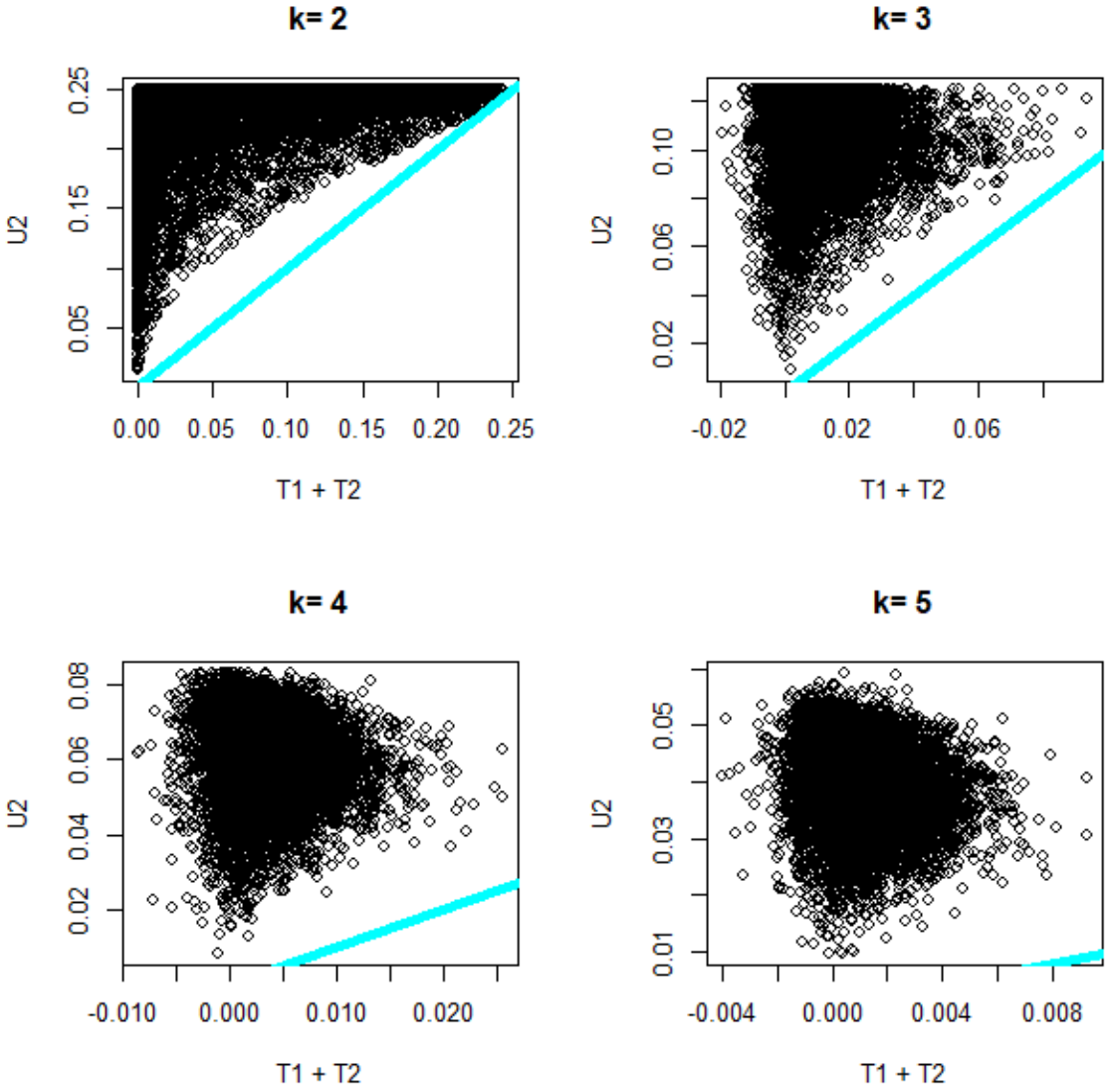

Figure 9: Numerical simulation: comparison of $T_2 + T_3$ with obtained conservative upper bound of T_2

The improved upper bound is therefore

$$Var(\gamma) \leq \frac{k^2(1-p_0)}{n(k-1)^2} \left\{ p_0\big(1 + \frac{1}{k-1}\big) \right\}. \tag{27}$$

7 Planning experiment

If we have an upper bound of variance of gamma and use asymptotic normality of gamma, it is possible to select sample size to achieve predefined margin of error of gamma. Let us demonstrate this in the following example.

Example 1:

The introduction of a new image modality at the mammography center at hospital X has prompted the question: what is the inter-rater reliability for this new modality? A study is planned to estimate the inter-rater agreement in radiologists' evaluation of images when it comes to the classification of breast tissue as malignant vs. benign. (In real life the radiologists will be using a richer spectrum of categories, however we simplify for the purpose of this illustration.) How many patients should be assessed to estimate the inter-rater agreement within limits of error ± 0.05?

Solution:

In this case the number of categories $k = 2$ and the desired error margin $e = 0.05$. Then by using the formula (27), we can find the required sample size for this study. $Var(\gamma)$ can be found from the following equation:

$$1.96 \cdot SE = 0.05, \qquad then \qquad Var(\gamma) = (SE)^2.$$

Thus, sample size requirements are presented below:

Value of observed agreement, p_0	Required sample size of subjects
0.5	3074
0.6	2951
0.7	2582
0.8	1967
0.9	1107

Table 11: Required sample size of subjects for different values of observed agreement

Example 2:

The reliability of classification of fractures from x-ray images is important for the planning of treatment and surgeries at the orthopedic department at hospital Y. A study is planned to assess the inter-rater agreement in the classification of growth plate fractures in children. According to the Salter-Harris classification scheme (reference [19]), these fractures are classified into one of five categories, i.e. types I-V. There

are vast differences in the expected proportions of the different categories: Type I-V fractures have reported incidences of 6, 75, 8, 10 and 1%, respectively. How many patients should be included to estimate the inter-rater agreement in the Salter-Harris classification within limits of error ± 0.05?

Solution

We have $k = 5$ - number of categories, $e = 0.05$ - error margin. We use formula (27), where variance of gamma can be found the same way as in Example 1. The required sample size for different values of observed agreement is presented in the Table 12.

Value of observed agreement, p_0	Required sample size of subjects
0.5	751
0.6	721
0.7	631
0.8	481
0.9	271

Table 12: Required sample size of subjects for different values of observed agreement

8 Conclusion

In this thesis the main goal was to find the requirements of sample size in planning a clinical study. The decision of how many patients, for example, to include is usually based on considerations of statistical power or precision. The cases for inter-rater agreement, where agreement is between two raters on k categories, has been studied. Usually such models include studies when the researcher needs to compare a new diagnostic tool with an existing one. Thus, the aim of the thesis was to study about different types of measures of inter-rater agreement by comparing them to each other.

Four agreement coefficients were presented: Cohen's kappa, Scott's Pi, Krippendorff's alpha and Gwet's AC1. All of them were analyzed and compared in case of two raters and two categories. The conclusion that Scott's Pi and Krippendorff's alpha are identically was made. The Cohen's kappa very similar result to the mentioned before two coefficients was observed, while only AC1 coefficient gave the highest value of agreement. That is why the main focus stayed on studying AC1 agreement coefficient in detail.

Some very useful properties of chance agreement of AC1 were studied, such as minimum and maximum values. Two important corollaries were found, one of them is the paradox of AC1, while the author of this coefficient (K.L.Gwet [1]) presented it as a paradox free coefficient. The formula for variance of AC1 was analyzed, and the goal was to find a simplification so it can be easier to use it during the clinical studies. This goal was reached by finding an upper bound of variance of AC1, which was shown in practice as a very useful one to estimate a sample size in planning experiment.

8.1 Further work

This thesis was mainly focused on studying Gwet's AC1 agreement coefficient and its use in finding the sample size requirements. The way the percent chance agreement was chosen for this coefficient is still unclear. After comparing agreement measurements, we conclude that AC1 gives the highest level of agreement among the raters, but still have not been convinced that this measure is good enough to be used in clinical studies.

Among other we have considered raters as a "fixed effect", i.e. the inter-rater agreement is assessed for the two raters that we are interested in. In real-life this is seldom the case, usually the classifications will be performed by any from a pool of raters, which may also vary with regard to educational background and training. If interest lies in the generalized inter-rater agreement between two random raters, the experiment would also have to plan for an appropriate number of raters.

References

[1] Kilem L. Gwet Ph.D., *Handbook of Inter-Rater Reliability*, 4th Edition, 2014 Gaithersburg, MD 20886-2696, pp. 1-168.

[2] Joseph L. Fleiss, Bruce Levin, Myunghee Cho Paik, *Statistical Methods for Rates and Proportions*, Third Edition, 2003, pp. 598-627.

[3] Douglas G. Altman, David Machin, Trevor N. Bryant, Martin J. Gadner, *Statistics with confidence. Confidence intervals and statistical guidelines*, Second Edition, 2000, pp. 116-177.

[4] Helena Chmura Kraemer1, Vyjeyanthi S. Periyakoil and Art Noda, *TUTORIAL IN BIOSTATISTICS. Kappa coefficients in medical research*, Statist. Med. 2002, pp. 2109–2129.

[5] Rosie Cornish, *An introduction to sample size calculations*, Mathematics Learning Support Centre, 2006, pp. 1-5.

[6] John S. Uebersax, *Diversity of Decision-Making Models and the Measurement of Interrater Agreement*, Center for Health Policy Research and Education, Duke University, 1987, Vol. 101, No. 1, pp. 140-146.

[7] Kilem Li Gwet, *Computing inter-rater reliability and its variance in the presence of high agreement*, STATAXIS Consulting, Gaithersburg, USA, British Journal of Mathematical and Statistical Psychology, 2008, pp. 29-48.

[8] Anthony J. Viera, MD; Joanne M. Garrett, PhD, *Understanding Interobserver Agreement: The Kappa Statistic*, Fam Med 2005, pp. 360-363.

[9] Gayatri Vishwakarma, PhD, *Sample Size and Power Calculation*, Clinical Development Services Agency (CDSA), Faridabad, May 2017, pp. 1-21.

[10] Olga Krasko, *Statistical analysis of data in medical research*, United Institute of Informatics Problems, January 2017, pp. 47-56.

[11] Qingshu Xie, *Agree or Disagree? A Demonstration of An Alternative Statistic to Cohen's Kappa for Measuring the Extent and Reliability of Agreement between Observers*, MacroSys, LLC, 11, pp. 1-12.

[12] Mary L. McHugh, *Interrater reliability: the kappa statistic*, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/`, Biochem Med (Zagreb). 2012 Oct; 22(3): 276–282.

[13] Laura Flight and Steven. A. Julious, *The Disagreeable Behaviour of the Kappa Statistic*, `https://www.sheffield.ac.uk/polopoly_fs/1.404095!/file/RSS_Poster_Laura_Flight_Final.pdf`, The University of Sheffield, School of Health and Related Research.

[14] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding and Kilem L Gwet, *A comparison of Cohen's Kappa and Gwet's AC1when calculating inter-rater reliability coefficients: a study conducted with personality disordersamples*, https://www.researchgate.net/publication/257872016, Wongpakaran et al. BMC Medical Research Methodology 2013, 13:61, pp: 1-7.

[15] Sahar Zafar, M.Brandon Westover, Nicolas Gaspard, Emily Gilmore, Brandon Foreman, Kathryn O'Connor and Eric S. Rosenthal, *Inter-rater agreement for consensus definitions of delayed ischemic events following aneurysmal subarachnoid hemorrhage*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4894325/,J Clin Neurophysiol. 2016 Jun; 33(3): 235–240.

[16] Adi Bhat, *Nominal, Ordinal, Interval, Ratio sales with examples*, https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/, Global VP - Sales and Marketing at QuestionPro.

[17] Mohamad Amin Pourhoseingholi, Mohsen Vahedi, Mitra Rahimzadeh, *Sample size calculation in medical studies*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017493/, Gastroenterol Hepatol Bed Bench. 2013 Winter; 6(1): 14–17.

[18] Ragna Elise Støre Govatsmark, Sylvi Sneeggen, Hanne Karlsaune, Stig Arild Slørdahl and Kaare Harald Bønaa, *Interrater reliability of a national acute myocardial infarction register*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4993256/, Clin Epidemiol. 2016; 8: 305–312.

[19] WikipediA, *Salter–Harris fracture*, https://en.wikipedia.org/wiki/Salter-Harris_fracture.

[20] Landis JR, Koch GG, *The measurement of observer agreement for categorical data.*, https://www.ncbi.nlm.nih.gov/pubmed/843571, Biometrics, 1977 Mar; 33(1): 159-74.

A R code

A.1 Comparison of agreement coefficients

```
#Use package 'rel' which gives us functions
#gac(), ckap(), spi(), kra().
#These functions estimate the agreement between meauserements.

library(rel)

# transforms agreement table into the input form for gac(), ckap()
#and others which were mentioned above
matrix.function <- function(tt) {
  N = sum (tt)

  M <- matrix(NA, nrow=N, ncol=2)
  if (tt[1,1]>0) M[1:tt[1,1], c(1,2)] <-1
  if (tt[2,2]>0) M[(tt[1,1]+1):(tt[1,1]+tt[2,2]), c(1,2)] <- 2
  if (tt[1,2]>0) {
    M[(tt[1,1]+tt[2,2]+1):(tt[1,1]+tt[2,2]+tt[1,2]), 1] <- 1
    M[(tt[1,1]+tt[2,2]+1):(tt[1,1]+tt[2,2]+tt[1,2]), 2] <- 2
  }
  if (tt[2,1]>0 ) {
    M[(tt[1,1]+tt[2,2]+tt[1,2]+1):N, 1] <- 2
    M[(tt[1,1]+tt[2,2]+tt[1,2]+1):N, 2] <- 1
  }
  M
}

# Creates agreement table with specified values of
# Prevalence index, Observed agreement

table.function<- function(Pi = 0.5, Pa = 0.9, Bi = 0.1,  N = 100){

  a <- round( 0.5*N*(Pa+Pi) )
  d <- round( 0.5*N*(Pa-Pi) )
  b <- round( 0.5*N*(Bi+1-Pa) )
  if ((N - a - b - d)<0) {
    c <- 0
    i <- which(c(a,b,d)==max(c(a,b,d)))[1]
    if (i==1) {
      a <- a - 1
    } else if (i==2) { b <- b - 1
    } else {  d <- d - 1
    }
```

```
  } else {
    c <- N - a - b - d
  }
  matrix(c(a,c,b,d), nrow=2)
}

#compare Kappa and AC1
#table.function(Pi = 0.5, Pa = 0.9, Bi = 0.1,  N = 100)

G <- K <- Pr <- NULL

for (alpha in seq(-0.9, 0.9, by=0.1)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 100)
  # contigency table  ->  table in required form
  M <- matrix.function(tt)

  G <- c(G, gac(M)$est)
  K <- c(K, ckap(M)$est)
  Pr <- c(Pr, alpha)
}

plot(Pr, G, type='o', col=("blue"), ylim=c(-1, 1),
     xlab="Prevalence Index", ylab="Agreement")
points(Pr, K, type='o', col=("red"))
legend("bottom", c("Kappa", "AC(1)"), pch=c(1,1),
       lty=c(1,1), col=c("red", "blue"))

#####################################################################

#####################################################################

#Compare Kappa and Krippendorff's Alpha

K <- Kripp <- Pr <- NULL

for (alpha in seq(-0.9, 0.9, by=0.1)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 100)
  # contigency table  ->  table in required form
  M <- matrix.function(tt)

  K <- c(K, ckap(M)$est)
  Kripp <- c(Kripp, kra(M)$est)
  Pr <- c(Pr, alpha)
}
```

```
plot(Pr, K, type='o', ylim=c(-1, 1),
     xlab="Prevalence Index", ylab="Agreement")
points(Pr, Kripp, type='o', col=2)
legend("bottom", c("Krippendorff's Alpha", "Kappa"),
       pch=c(1,1), lty=c(1,1), col=c(2, 1))

#####################################################################

#####################################################################

#Compare Krippendorff's Alpha and AC1

G <- Kripp <- Pr <- NULL

for (alpha in seq(-0.9, 0.9, by=0.1)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 100)
  # contigency table  ->  table in required form
  M <- matrix.function(tt)

  G <- c(G, gac(M)$est)
  Kripp <- c(Kripp, kra(M)$est)
  Pr <- c(Pr, alpha)
}

plot(Pr, G, type='o', ylim=c(-1, 1),
     xlab="Prevalence Index", ylab="Agreement")
points(Pr, Kripp, type='o', col=2)
legend("bottom", c("Krippendorff's Alpha", "AC(1)"),
       pch=c(1,1), lty=c(1,1), col=c(2, 1))


#####################################################################

#####################################################################

#Compare Kappa and Scott's Pi

Sc <- K <- Pr <- NULL

for (alpha in seq(-0.9, 0.9, by=0.1)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 100)
  # contigency table  ->  table in required form
  M <- matrix.function(tt)
```

```
  Sc <- c(Sc, spi(M)$est)
  K <- c(K, ckap(M)$est)
  Pr <- c(Pr, alpha)
}

plot(Pr, Sc, type='o', ylim=c(-1, 1),
     xlab="Prevalence Index", ylab="Agreement")
points(Pr, K, type='o', col=2)
legend("bottom", c("Kappa", "Scott's Pi"),
       pch=c(1,1), lty=c(1,1), col=c(2, 1))

#####################################################################

#####################################################################

#Compare AC1 and Scott's Pi

G <- Sc <- Pr <- NULL

for (alpha in seq(-0.9, 0.9, by=0.1)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 100)
  # contigency table  ->  table in required form
  M <- matrix.function(tt)

  G <- c(G, gac(M)$est)
  Sc <- c(Sc, spi(M)$est)
  Pr <- c(Pr, alpha)
}

plot(Pr, G, type='o', ylim=c(-1, 1), xlab="Prevalence Index",
     ylab="Agreement")
points(Pr, Sc, type='o', col=2)
legend("bottom", c("Scott's Pi", "AC(1)"), pch=c(1,1), lty=c(1,1),
       col=c(2, 1))

#####################################################################

#####################################################################

#Compare Scott's Pi and Krippendorff's Alpha

Sc<- Kripp <- Pr <- NULL
```

```
for (alpha in seq(-0.9, 0.9, by=0.1)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 100)
  # contigency table  ->  table in required form
  M <- matrix.function(tt)

  Sc <- c(Sc, spi(M)$est)
  Kripp <- c(Kripp, kra(M)$est)
  Pr <- c(Pr, alpha)
}

plot(Pr, Sc, type='o', col=("green"), ylim=c(-1, 1),
     xlab="Prevalence Index", ylab="Agreement")
points(Pr, Kripp, type='o', col=("black"))
legend("bottom", c("Krippendorff's Alpha", "Scott's Pi"),
       pch=c(1,1), lty=c(1,1), col=c("black", "green"))


#####################################################################

#####################################################################

#Compare all agreement coefficients
Kripp <- K <- Pr <- G <- Sc <- NULL

for (alpha in seq(-0.9, 0.9, by=0.1)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 100)

  if (min(tt)>=0) {
    # contigency table  ->  table in required form
    M <- matrix.function(tt)

    Kripp <- c(Kripp, kra(M)$est)
    K <- c(K, ckap(M)$est)
    G <- c(G, gac(M)$est)
    Sc <- c(Sc, spi(M)$est)
    Pr <- c(Pr, alpha)
  }
}

plot(Pr, Kripp, type='o',col=("black"), ylim=c(-1, 1),
     xlab="Prevalence Index", ylab="Agreement")
points(Pr, K, type='o', col=("red"))
points(Pr, G, type='o', col=("blue"))
points(Pr, Sc, type='o', col=("green"))
```

```
legend("bottom", c("Kappa", "Krippendorff's Alpha", "AC(1)", "Scott's pi"),
       pch=c(1,1, 1, 1), lty=c(1,1,1,1),
       col=c("red", "black", "blue", "green"))
```

A.2 Examples of Cohen's kappa and Gwet's AC1

```
#Simulate some examples for Cohen's kappa and Gwet's AC1
#We want to see the bahaviour of standard error and
#confidence interval with dependence on prevalence
#index for different sample size

#We use same library and functions (matrix.function and
#table.function) as we did before.

#table.function(Pi = 0.5, Pa = 0.9, Bi = 0.1,  N = 40)

#For Kappa with N=40

K <- Pr <- Sd <- NULL

for (alpha in seq(-0.9, 0.9, by=0.02)) {

  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 40)

  # contigency table  ->  table in required form

  M <- matrix.function(tt)
  # agreement table => frequency agreement table
  n <- sum(tt)
  p <- tt/n

  #Parts of formula varince of kappa

  A <- sum(diag(p) * (1 - (apply(p, 1, sum) + apply(p, 2, sum))*
                        (1-ckap(M)$est))^2)

  B <- (1-ckap(M)$est)^2 *(p[1,2] *(sum(p[1,]) + sum(p[,2]))^2 +
                             p[2,1]*(sum(p[2,]) + sum(p[,1]))^2)

  pe <- sum(p[1,])*sum(p[,1]) + sum(p[2,])*sum(p[,2])

  C <- (ckap(M)$est - pe*(1-ckap(M)$est))^2
```

```
  SE <- sqrt(A+B-C)/((1-pe)*sqrt(n))

  K <- c(K, ckap(M)$est)
  Pr <- c(Pr, alpha)
  Sd <- c(Sd, SE)
}

#find the upper and lower bond for 95% CI where quantile=1.96
upK=K+Sd*1.96
lowK=K-Sd*1.96

#plot kappa and its standard error with dependence on PI
par(mfrow=c(1,2))
plot(Pr,Sd,ylim=c(-1,1), xlab="Prevalence Index",
     ylab="Kappa/Standard error of kappa", col="red")
lines(Pr,K, type = "l", lwd=3,col="blue")
legend("bottom", c("Kappa", "Standard error"),pch = c("-","o"),
       lty=c(1,0) , lwd = c(3,0),col=c("blue","red"))

#plot kappa and its 95% CI
plot(Pr, K,type = "l",lwd=3,ylim=c(0,1), xlab="Prevalence Index",
     ylab="Kappa", col="blue")
lines(Pr,upK, col="green")
lines(Pr,lowK, col="green")
legend("bottom", c("Kappa", "Confidence interval"), pch=c("-","-"),
       lty=c(1,1), lwd = c(3,0), col=c("blue","green"))

#####################################################################

#for Kappa with N=300 we do the same simulations without changes,
#only for bigger N

K <- Pr <- Sd <- NULL
for (alpha in seq(-0.9, 0.9, by=0.02)) {

  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 300)
  M <- matrix.function(tt)
  n <- sum(tt)
  p <- tt/n

  A <- sum(diag(p) * (1 - (apply(p, 1, sum) + apply(p, 2, sum)) *
                        (1-ckap(M)$est))^2)

  B <- (1-ckap(M)$est)^2 *(p[1,2] *(sum(p[1,]) + sum(p[,2]))^2 +
```

```
                              p[2,1]*(sum(p[2,]) + sum(p[,1]))^2)

  pe <- sum(p[1,])*sum(p[,1]) + sum(p[2,])*sum(p[,2])

  C <- (ckap(M)$est - pe*(1-ckap(M)$est))^2

  SE <- sqrt(A+B-C)/((1-pe)*sqrt(n))

  K <- c(K, ckap(M)$est)
  Pr <- c(Pr, alpha)
  Sd <- c(Sd, SE)
}

upK=K+Sd*1.96
lowK=K-Sd*1.96

par(mfrow=c(1,2))
plot(Pr, Sd,ylim=c(-1,1),xlab="Prevalence Index",
     ylab="Kappa/Standard error of kappa", col="red")
lines(Pr, K, type = "l", lwd=3, ylim=c(0,1), col="blue")
legend("bottom", c("Kappa", "Standard error"),pch = c("-","o"),
       lwd=c(3,0), lty=c(1,0), col=c("blue","red"))

plot(Pr, K, type = "l", lwd=3, ylim=c(0,1),xlab="Prevalence Index",
     ylab="Kappa", col="blue")
lines(Pr,upK, col="green")
lines(Pr,lowK, col="green")
legend("bottom", c("Kappa", "Confidence interval"), pch=c("-","-"),
       lwd=c(3,0), lty=c(1,1), col=c("blue","green"))

######################################################################

#Simulations for AC1 where N=40, to compare it to the result
#we got in Kappa
#First we write down all the components of formula for
#variance of AC1

#Pi function
pi = function(i, p) {
  (sum(p[i,]) + sum(p[,i]))/2
}

#pe-chance agreement
pe = function(p) {
```

```
  S = 0
  for (i in 1:nrow(p)) {
    pe1=pi(i,p)*(1-pi(i,p))
    S = S + pe1
  }
  k=nrow(p)-1
  S = S/k
}

#po-observed agreement
po = function(p) { sum(diag(p))}

#gamma coefficient formula
gamma = function(p)  {(po(p) - pe(p))/(1-pe(p))}

#part 1 of formula Var(AC1)
part1=function(p){
  po(p)*(1-po(p))
}

#part 2 of formula
part2=function(p){
  S=0
  for(i in 1:nrow(p)) {
    S=S+p[i,i]*(1-pi(i,p))}

  4*(1-gamma(p))*((1/(nrow(p)-1))*S-po(p)*pe(p))
}

#part 3 of formula
part3=function(p){
  S=0
  for (i in 1:nrow(p)){
    for (j in 1:nrow(p)){
      S=S+p[i,j]*(1-(pi(i,p)+pi(j,p))/2)^2
    }
  }
  4*(1-gamma(p))^2 *((1/(nrow(p)-1)^2)*S - pe(p)^2)
}

#########################################################################

#Numerical simulation
```

```
G <- Sd <- Pr<- NULL

for (alpha in seq(-0.9, 0.9, by=0.02)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 40)
  M <- matrix.function(tt)
  n <- sum(tt)
  p <- tt/n

  #set all the components into the formula
  VarG= (1/(n*(1-pe(p))^2))*(part1(p) - part2(p) + part3(p))

  SE <- sqrt(VarG)

  G <- c(G, gac(M)$est)
  Pr <- c(Pr, alpha)
  Sd <- c(Sd, SE)
}
#Upper and lower bound for 95% CI
upG=G+Sd*1.96
lowG=G-Sd*1.96

#AC1 and standard error
par(mfrow=c(1,2))
plot(Pr, Sd,ylim=c(-1,1),xlab="Prevalence Index",
     ylab="AC1/Standard error of gamma",  col="red")
lines(Pr,G, type = "l", lwd=3, ylim=c(0,1), col="black")
legend("bottom", c("AC1", "Standard error"), pch = c("-","o"),
       lty=c(1,0),lwd=c(3,0), col=c("black","red"))

#AC1 and 95% CI
plot(Pr, G, type = "l", lwd=3, ylim=c(0,1),xlab="Prevalence Index",
     ylab="AC1", col="black")
lines(Pr,upG, col="green")
lines(Pr,lowG, col="green")
legend("bottom", c("AC1", "Confidence interval"),  pch=c("-","-"),
       lty=c(1,1), lwd=c(3,0), col=c("black","green"))


######################################################################
#Same simulations for AC1 with N=300

G <- Sd <-Pr<- NULL

for (alpha in seq(-0.9, 0.9, by=0.02)) {
```

```
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 300)
  M <- matrix.function(tt)
  n <- sum(tt)
  p <- tt/n

  VarG= (1/(n*(1-pe(p))^2))*(part1(p) - part2(p) + part3(p))
  SE <- sqrt(VarG)

  G <- c(G, gac(M)$est)
  Pr <- c(Pr, alpha)
  Sd <- c(Sd, SE)
}
upG=G+Sd*1.96
lowG=G-Sd*1.96

par(mfrow=c(1,2))
plot(Pr, Sd,ylim=c(-1,1), xlab="Prevalence Index",
     ylab="AC1/Standard error of gamma", col="red")
lines(Pr,G, type = "l", lwd=3, ylim=c(0,1), col="black")
legend("bottom", c("AC1", "Standard error"), pch = c("-","o"),
       lty=c(1,0),lwd=c(3,0), col=c("black","red"))


plot(Pr, G, type = "l", lwd=3, ylim=c(0,1),xlab="Prevalence Index",
     ylab="AC1", col="black")
lines(Pr,upG, col="green")
lines(Pr,lowG, col="green")
legend("bottom", c("AC1", "Confidence interval"),  pch=c("-","-"),
       lty=c(1,1),lwd=c(3,0), col=c("black","green"))
```

A.3 Upper bound

Term 2

```
#To establish for what set of values of expression under the parantheses
#in Term 2 reaches its maximum, we preformed numerical simulations.
#We considered all possible combinations of p and pi taking the
#values from the grid (0 to 1) with a step 0.01 (for k=2) and 0.025
#(for k=3) and such that sum of all p and sum of all pi equals to 1.

#We show the example where k=3

k=3
x <- seq(0, 1, by=0.025)
```

```
D <- matrix(rep(x, k-1), nrow=k-1)
ZZ <- t(expand.grid(data.frame(t(D))))
ZZ <- matrix(ZZ[, apply(ZZ, 2, sum)<=1], nrow=k-1)
CC <- matrix(nrow(ZZ), k)
CC <- rbind(ZZ, 1-apply(ZZ,2,sum))
ZZ <- NULL

res <- rep(NA, 1000000)
#ncol(CC)*ncol(CC) #   rep(NA, 1000000) #

pi.max <- NULL
p.max <- NULL

max_ <- -10
i=1

for (j1 in 1:ncol(CC)) {
  for (j2 in 1:ncol(CC)) {

    if (i %% 1000000) {i=1}

    res[i] <- sum((CC[,j2] - CC[,j1])*CC[,j1])

    if (res[i] >= max_) {

      pi.max <- CC[,j1]
      p.max <- CC[,j2]
      }
     max_ <- max(max_, res[i])
    i = i + 1
  }
}

pi.max
p.max
```

Term 3

```
#We want to check how the maximum of expression
#(1/(k-1)^2)*Var(eta) behaves for different values of k.
#We run a numerical simulation as follows: we consider all
#possible combinations of pi such that their sum is equal to 1.
#On a grid (0 to 1) with a step 0.01, find the maximum
```

```
#values for k=2,...,5.

x <- seq(0, 1, by=0.01)
D <- matrix(rep(x, k-1), nrow=k-1)
ZZ <- t(expand.grid(data.frame(t(D))))
ZZ <- matrix(ZZ[, apply(ZZ, 2, sum)<=1], nrow=k-1)
CC <- matrix(nrow(ZZ), k)
CC <- rbind(ZZ, 1-apply(ZZ,2,sum))
ZZ <- NULL

result <- rep(NA, 1000000)
#ncol(CC)*ncol(CC) # rep(NA, 1000000) #

max.glob <- NULL

result <- rep(NA, ncol(CC))
i=1
for (k in 2:5){
for (j in 1:ncol(CC)) {
  result[i]=sum(CC[,j]^3)- sum(CC[,j]^2)^2
  i=i+1
}
max.glob  <- c(max.glob , max(result)/(k-1)^2)
}
max.glob

plot(x=2:5, y=max.glob, type="o",pch=16, cex=1.5, xlab="k",
     ylab=bquote("max Var" ~"("~ eta~ ")" ~frac(1, (k-1)^2)),
     xaxt="n", col="red")
axis(1, at=2:10,labels=2:10)
```

Example of using the upper bound formula

```
#We run the same example as we did for Gwet's AC1 with N=300
#to illustrate both confidence interval and new formula
#for upper bound of corresponding variance.

#We use absolutely identical simulation as we did before.

#table.function(Pi = 0.5, Pa = 0.9, Bi = 0.1,  N = 300)

G <- Sd <-Pr<-UB <- NULL
```

```
for (alpha in seq(-0.9, 0.9, by=0.02)) {
  tt <- table.function(Pi=alpha, Pa = 0.9, Bi = 0.1,  N = 300)
  M <- matrix.function(tt)
  n <- sum(tt)
  p <- tt/n

  VarG= (1/(n*(1-pe(p))^2))*(part1(p) - part2(p) + part3(p))
  SE <- sqrt(VarG)
  #formula for upper bound for k=2
    u <- sqrt((2^2*(1-po(p))/n)*
              ( po(p)*(1 + 1) + (1-po(p))*0.062475*4*4 ))

  G <- c(G, gac(M)$est)
  Pr <- c(Pr, alpha)
  Sd <- c(Sd, SE)
  UB <- c(UB, u)
}
#95% CI
upG=G+Sd*1.96
lowG=G-Sd*1.96

plot(Pr, G, type = "l", lwd=3, ylim=c(0,1),xlab="Prevalence Index",
     ylab="AC1", col="black")
lines(Pr,upG, col="green")
lines(Pr, G + 1.96* UB, col="tomato")
lines(Pr, G - 1.96* UB, col="tomato")
lines(Pr,lowG, col="green")
legend("bottom", c("AC1", "Confidence interval", "Upper Bound of CI"),
       pch=c("-","-","-"), lty=c(1,1,1),lwd=c(3,0,0),
       col=c("black","green","tomato"))
```

Improving upper bound

```
par(mfrow=c(2,2))
for (k in 2:5) {
  T1=T2=U2=NULL
  for (i in 1:10000){
    P=matrix(runif(k*k), nrow=k)
    P=P/sum(P)

    T1=c(T1, -part2(P))
```

```
    T2=c(T2, part3(P))
    U2=c(U2, po(P)*(1-po(P))*(1/(k-1)))
  }
  plot(T1+T2, U2, main=paste("k=", k))
  abline(0,1,col=5, lwd=5)
}
```

A.4 Planning experiment

```
#Example 1

#error +/- 0.025
e=0.05

#Standard error for a normal distributed estimate 1.96*SE=0.025
SE=e/1.96

#Variance gamma
var=SE^2

#two categories
k=2

#observed agreement
p0=0.5

#By using the main result for upper bound we can find samle size
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.6
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.7
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.8
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.9
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
```

```
n

#####################################################################

#Example 2

#error +/- 0.05
e=0.05

#Standard error for a normal distributed estimate 1.96*SE=0.05
SE=e/1.96

#Variance gamma
var=SE^2

#five categories
k=5

#observed agreement
p0=0.5

#Sample size
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.6
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.7
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.8
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n

p0=0.9
n=(k^2*(1-p0)*p0*(1+(1/(k-1))))/((k-1)^2 * var)
n
```