mathematics

MDPI

Article

Stability of Dependencies of Contingent Subgroups with Merged Groups: Vaccination Case Study

Tomas Macak

Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamycka 129, 16500 Prague, Czech Republic; macak@pef.czu.cz; Tel.: +420-224-382-029

Abstract: The answers to extreme phenomena both in nature and in business sectors are the constructions of the distribution of random variables with extreme values. Another area in which appropriate theoretical research is conducted regarding the influence of suppressor (third) variables in categorical data. When examining dependencies in PivotTables, we often find it necessary to merge data into larger sets (e.g., due to a greater number of theoretical frequencies lower than their critical value). A phenomenon many exist wherein the partial relation is stronger than the zero relation. For example, in such a combination, instability may occur, which indicates contingent subgroups with the merged group. The dependence of dependencies is practically manifested because the data of contingent subgroups indicate inconsistent (inverted) conclusions compared to the associated group. For this reason, this paper aimed to find the critical ratios of partial probabilities in the contingency table of subgroups of the original variables, and to determine the conditions of result consistency and contingency stability, including the proof. For practical use and for the ease of repeating the proposed procedure, the solution is based on a case study that compares the effectiveness of vaccination.

Keywords: partial probability; suppressive variable; contingency; causality; consistency

Citation: Macak, T. Stability of Dependencies of Contingent Subgroups with Merged Groups: Vaccination Case Study. *Mathematics* **2021**, *9*, 2917. https://doi.org/10.3390/math9222917

Academic Editors: Tatjana von Rosen and Carlos Agra Coelho

Received: 1 October 2021
Accepted: 12 November 2021
Published: 16 November 2021

1. Introduction

Regarding the portability of statistical testing (parametric/nonparametric) and the search for categorical data dependencies through correlations to the causality factors of corporate governance, the current state of the knowledge of the professional community is that experts focus on some criticism, which states "correlation does not imply causality". In addition to this, another critique mentions the absence of a proposed solution to unambiguously verify which correlation is causal, and uncertainty in how one can determine the direction of causality of factors. The direction of the management factors is critical in terms of effective business management.

Theoretical and applied scientists regularly aim for strict, unbiased approximation when making cogent presumptions regarding scientific problems (A). The prevailing, standard approach has been formulated in terms of two opposing statistical hypotheses: one representing no difference between two populations (i.e., the null hypothesis (Ho)) and the other representing either unidirectional or bidirectional options (i.e., the alternative hypothesis (Ha)). These hypotheses primarily correspond to different models. For example, when comparing two samples of populations, the presumption is that they are from the same primary data set, so the difference between their correct means is equal to 0.

A statistical test and a multinomial regression model are usually calculated from sample data, and are equated to the hypothesized null distribution to explore the conformity of the data with the null hypothesis. More extreme values of a statistical test indicate that the sample data are not consistent with the null hypothesis. A mainly random level (a) is often present to serve as a cut-off point (i.e., the unambiguous background for a verdict) for statistically relevant versus negligible events. This approach is known by different names, e.g., null hypothesis testing or null hypothesis significance testing. This method is

a modification of Fisher's (1928) significance testing [1], and Neyman and Pearson's (1933) hypothesis testing [2–5]. There are many problems that surround the application of the null hypothesis testing method, especially if we consider the test result or the parameters of the regression model as an indication of a causal relationship. Thus, in the case of hypothesis testing, we believe these problems are the result of a binary expression of causality. Some of these problems are mentioned in [6–9]. Although uncertainties among statisticians concerning the utility of null hypothesis testing are hardly new [10–12], the prevalence of criticism has increased in the scientific literature in the last five years. More than 200 references now exist in the academic literature that point out the limitations of regression models and statistical hypothesis testing in the sense that the statistical correlation test is not guaranteed to find the causality of the studied phenomena, but finding the causality of phenomena/processes is essential for effective business management.

A specific area in which conventional statistical approaches fail is the area of association and contingency dependencies. The question of how to transform an association dependency into a causal relation is an issue. A second, little-known problem is the inconsistency of subgroups of categorical data with their associated group. The initial description of these issues, using conditional probability, was expressed by Judea Pearl in [13]. This basic framework was then used in [14–17].

The problem of causality direction can be described using the phenomenon whereby an event (C) increases the probability (E) in a given population (p) and, at the same time, decreases the probability (E) in every sub-population of (p). In other words, if F and ¬F (a negation of (F)) are two complementary properties describing two subpopulations, we might well encounter inequalities (expressed by conditional probability and the negation of phenomena), as expressed by Pearl [13]:

$$P(E|C) > P(E|\neg C) \tag{1}$$

$$P(E|C,F) < P(E|\neg C,F) \tag{2}$$

$$P(E|C,\neg F) < P(E|\neg C,\neg F) \tag{3}$$

Although such a reversal order is not surprising from the perspective of the theory of probability, it is paradoxical by causal interpretation. For example, if C is associated (implying cause) with taking a certain financial recovery of the company (for example, through state subsidies), E (implying effect) with recovery, and F with being a company producing services, then—under the causal interpretation of (2) and (3)—financial recovery seems to be harmful to both manufacturing companies and companies producing services, but yet beneficial to the whole population of companies (Equation (1); Pearl [13]).

In a case study that represents the numerical interpretation of the paradoxical case study, we can, for example, assume that overall, the recovery rate for a company in financial crisis receiving financial recovery (C) at 50% exceeds that of the control (¬C) at 40%, and so the state subsidy treatment is apparently preferred. However, when we inspect the separate data regarding manufacturing companies and companies producing services, the recovery rate for "financially untreated" companies is 10% higher than for the treated ones (for both manufacturing companies and companies producing services).

The explanation for this paradox can be clear from an exact viewpoint because it has taken appropriate care to distinguish "seeing from doing". The conditional operator in probability calculus represents the causal dependent "given that we do". In contrast, the do operator was devised to represent the causal conditional "given then we do" [13]. According to the previous statement, the inequalities are as follows:

$$P(E|C) > P(E|\neg C) \tag{4}$$

$$P(E|do(C)) > P(E|do(\neg C)) \tag{5}$$

The C can be positive evidence for E, which may be due to spurious confounding factors that cause both C and E. In this case study, financial recovery appears beneficial

overall because manufacturing companies are more often in a financial crisis (regardless of the state subsidies) than companies producing services and are more likely to use financial recovery. Indeed, finding a financial recovery-using company C of unknown company type (making services versus products) would benefit from inferring that the company is more likely to be a manufacturing company and, hence, more likely to recover. This statement agrees with Formulas (1)–(3).

Thus, from a theoretical point of view, it is appropriate to supplement the current state of knowledge with an analytical point of view, which will make it possible to unambiguously determine whether the data in the contingency table show inconsistencies of the sorted subgroups with the merged group. For subsequent practical application, this aspect should use relationships at associated frequencies and distinguish whether the association indicates causality. For this purpose, it is appropriate to discard the analytical form of the critical ratio of marginal probabilities of the pool. Furthermore, for practical purposes, it is reasonable to create a graph of stability, which relates the real ratios of marginal probabilities with the theoretical values of marginal probabilities determined using combined frequencies. This diagram then makes it possible to identify a consistent and inconsistent case unambiguously. In the area of categorical data, there is a gap in the design of solutions.

Because the human population is still facing a worldwide coronavirus pandemic and vaccination appears to be the most effective interim solution to date, the theoretical solution is illustrated in a vaccination case study.

When evaluating the efficacy of a given type of vaccine, the stratified population is usually vaccinated with the vaccine, and the same control population is administered the substance without affecting infectious resistance. After that, the control population is monitored, and after infecting a certain proportion of the population, the amounts infected are compared between the vaccinated and control groups. Thus, if 100 infected individuals from the control group were expected and the number of those infected in the vaccinated group was 10, then the difference in 100% = 90 would lead to a 90% vaccine efficacy. This indirect method of determining vaccine efficacy replaces an unethical method of direct experimentation that would directly infect the vaccinated population and measure the level of antibody resistance (proportion of infections manifested).

The reliability of a method for indirectly determining vaccine efficacy is usually examined in terms of stratification and randomization experimental and control populations and in terms of a sample size of the population. It is assumed that a larger experimental and control population automatically indicates greater reliability under the condition of randomization and stratification [18]. It is here that a paradoxical phenomenon can be found, where one vaccine dominates in terms of the total population, and the other vaccine dominates in sorted groups according to the third criterion (this is the number of vaccine doses, i.e., one or two vaccine doses). Therefore, it is interesting to examine the consistency of the results after one dose of vaccine, after two doses of vaccine, and after merging these two groups when comparing the two types of vaccine.

2. Materials and Methods

Thus, we will first label the variables to meet the objectives described at the end of Section 1. Next, we derive a critical ratio of marginal probabilities. We start from a special case where the theoretical equality of the associated frequencies $n_{.12} = n_{.22}$ which indicates the same aggregated efficacy of the vaccines. We also determine the theoretical values of the ratio of marginal probabilities concerning real marginal probabilities for different cases. Using these cases, we then derive the rules of consistency. We then summarize these rules in the combinational consistency of data subgroups and a merged groups table. Then, we create a stability diagram that visualizes this table.

First, we introduce the labeling of variables:

Let $i \in (1,2)$, $j \in (1,2)$, $k \in (1,2)$, be:

n_{ijk} is the number of individuals who were in the i-th state had the j-th treatment, and the action ended in the k-th result;

n_{ij} is number of individuals who were in the i-th state and had the j-th treatment;

$n_{i.k}$ is the number of individuals who were in the i-th state and the action ended with the k-th result;

$n_{.jk}$ is the number of individuals who had the j-th treatment and the action ended in a k-th result;

$n_{i..}$ is the number of individuals who were in the i-th state;

$n_{.j.}$ is the number of individuals who had j-th treatment;

$n_{..k}$ is the number of individuals for whom the action ended in a k-th result;

$n_{...}$ is the number of all individuals.

These variables apply to the following:

$$n_{ij.} = \sum_{k=1}^{2} n_{ijk}$$

$$n_{i.k} = \sum_{j=1}^{2} n_{ijk}$$

$$n_{.jk} = \sum_{i=1}^{2} n_{ijk}$$

$$n_{i..} = \sum_{j=1}^{2} \sum_{k=1}^{2} n_{ijk}$$

$$n_{.j.} = \sum_{i=1}^{2} \sum_{k=1}^{2} n_{ijk}$$

$$n_{..k} = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ijk}$$

$$n_{...} = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} n_{ijk}.$$

These variables can be applied to the case of vaccination where the dependence is reversed, e.g., in the association/contingency table (or the determination of the sample size for the interval estimation error will not work), where the examined subcategories show a different conclusion than when the whole population is merged (populations are the same size). The effect of a partial relationship (partial correlation) is probably stronger than the primary relationship between variables (zero relationships).

The relationship between two variables, X and Y, may not always express the relationship that actually exists. According to A, the relationship between X and Y is called a zero-order relationship. After introducing the third variable, called the test variable labeled Z, a first-order relationship is established. To illustrate, consider the association table sorted by income (low–high) and by gender (female–male). A man has a 1.5 higher frequency of a high income than a woman. It would probably confirm the association dependence that the income is gender-dependent. By introducing the third variable, Z, the number of hours worked, we men find that men work at a three times higher frequency. Thus, the partial correlation (association) will probably be stronger than the zero-order association. Thus, more hours worked for the average man than for the average woman will explain the average higher income for men than for women. Therefore, there will probably be no discrimination against gender in income. This designation of correlation has been used, for example, in scholarly articles [19–23].

3. Results

Table 1 shows a case where the total population $n_{...} = 5000$ is divided according to the criterion of the type of vaccine (A and B) and according to the criterion (variation) of the number of vaccinations (one or two doses of vaccine). The subpopulation vaccinated with vaccine A is the same size as the subpopulation vaccinated with vaccine B ($n_{.1.} = n_{.2.} = 2500$).

The data in Table 1 are intended to provide essential information on which vaccine is preferred in terms of vaccine efficacy. Vaccine efficacy is expressed here as the combined value of the number without infection to the total number ($n_{.11}/n_{.1.} = 1 - p_{.1.} = 0.8$) for vaccine A. Furthermore, for vaccine B, for the combined value of the number without infection to the total number ($n_{.21}/n_{.2.} = 1 - p_{.2.} = 0.9$. In the pooled value, therefore, vaccine B is more effective than vaccine A. However, if we sort the pooled group according to the number of vaccinations, we come to the opposite conclusion. In this case, the efficacy of vaccine A for one dose of vaccination ($n_{111}/n_{11.} = 1 - p_{11.} = 0.525$) is greater than the efficacy of vaccine B ($n_{121}/n_{12.} = 1 - p_{12.} = 0.300$). The efficacy of vaccine A for two doses of vaccination ($n_{211}/n_{21.} = 1 - p_{21.} = 0.984$) is again greater than the efficacy of vaccine B ($n_{221}/n_{22.} = 1 - p_{22.} = 0.967$). For the instability of the conclusions of the sorted and combined set, the criterion of the critical ratio of marginal probabilities $p_{11.}/p_{12}$ and $p_{21.}/p_{22.}$ is derived in the following text using the combined frequencies indicated in Table 1. Furthermore, the rules between the critical probability ratio and the real values of these ratios are derived (see Table 2).

Table 1. Frequencies for determining the unreliability of the effect of vaccination (probability of infection after vaccination in the observed period).

Total Population (Combined)				
	No Infection (Tested Negative) N	**With Infection (Tested Positive) P**	**Total N + P**	**Probability of Infection**
A-type vaccine	2000 $n_{.11}$	500 $n_{.12}$	2500 $n_{.1.}$	0.2 $p_{.1.}$
B-type vaccine	2250 $n_{.21}$	250 $n_{.22}$	2500 $n_{.2.}$	0.1 $p_{.2.}$
Total	4250 $n_{..1}$	750 $n_{..2}$	5000 $n_{\ldots}$	B-type is better than A-type vaccine
Vaccinated Once				
	No Infection (Tested Negative) N	**With infection (Tested Positive) P**	**Total N + P**	**Probability of Infection**
A-type vaccine	525 n_{111}	475 n_{112}	1000 $n_{11.}$	0.475 $p_{11.}$
B-type vaccine	75 n_{121}	175 n_{122}	250 $n_{12.}$	0.700 $p_{12.}$
Total	600 $n_{1.1}$	650 $n_{1.2}$	1250 $n_{1..}$	A-type is better than the B-type vaccine
Vaccinated Twice				
	No Infection (Tested Negative) N	**With Infection (Tested Positive) P**	**Total N + P**	**Probability of Infection**
A-type vaccine	1475 n_{211}	25 n_{212}	1500 $n_{21.}$	0.016 $p_{21.}$
B-type vaccine	2175 n_{221}	75 n_{222}	2250 $n_{22.}$	0.033 $p_{22.}$
Total	3650 $n_{2.1}$	100 $n_{2.2}$	3750 $n_{2..}$	A-type is better than the B-type vaccine

Suppose we have equally large total populations that we can compare with each other. In our case, these populations are the number of people vaccinated with vaccine A and vaccine B. Thus, in this case, $n_{.1.} = n_{.2.} = 2500$. Then, we can start from the theoretical equality of the associated frequencies $n_{.12} = n_{.22}$, from which we calculate the probability of the investigated phenomenon (here, the likelihood of infection). This theoretical equality of combined frequencies allows us to determine the cut-off point (or indifferent limit) at which vaccine A is as effective as vaccine B. This indifferent ratio means that if it uses a blunt sign comparing the partial efficacy of vaccines A and B, then:

$$\frac{p_{11.}}{p_{12.}} \leq 1 \wedge \frac{p_{11.}}{p_{12.}} \geq 1; \frac{p_{21.}}{p_{22.}} \leq 1 \wedge \frac{p_{21.}}{p_{22.}} \geq 1$$

Table 2. Combinational consistency of data subgroups and merged groups.

Situation	Real Ratios of Marginal Likelihoods		Real Ratios of Combined Likelihoods	Theoretical Ratios of Marginal Likelihoods		Occupancy Quadrants of the Stability Graph	Consistency
	$\frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.real}}{p_{22.real}}$	$\frac{p_{.1.}}{p_{.2.}}$	$\frac{p_{11.}}{p_{12.}}$	$\frac{p_{21.}}{p_{22.}}$	**Q1, Q2, Q3, Q4, Q3c, Q4c**	**Yes, No**
1.	$(0;1)$	$(0;1)$	$(0;1)$	$\frac{p_{11.}}{p_{12.}} > \frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.}}{p_{22.}} < \frac{p_{21.real}}{p_{22.real}}$	Q2 ∨ Q3	Yes
2.	$(0;1)$	$(0;1)$	$(0;1)$	$\frac{p_{11.}}{p_{12.}} < \frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.}}{p_{22.}} > \frac{p_{21.real}}{p_{22.real}}$	Q1 ∨ Q4	Yes
3.	$(0;1)$	$(0;1)$	$(1;\infty)$	$\frac{p_{11.}}{p_{12.}} < \frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.}}{p_{22.}} < \frac{p_{21.real}}{p_{22.real}}$	Q3 ∧ Q3c	No
4.	$(1;\infty)$	$(1;\infty)$	$(0;1)$	$\frac{p_{11.}}{p_{12.}} > \frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.}}{p_{22.}} > \frac{p_{21.real}}{p_{22.real}}$	Q1 ∧ Q2	No
5.	$(1;\infty)$	$(0;1)$	$(0;1)$	$\frac{p_{11.}}{p_{12.}} < \frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.}}{p_{22.}} > \frac{p_{21.real}}{p_{22.real}}$	Q2 ∨ Q4	Yes
6.	$(1;\infty)$	$(0;1)$	$(1;\infty)$	$\frac{p_{11.}}{p_{12.}} > \frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.}}{p_{22.}} < \frac{p_{21.real}}{p_{22.real}}$	Q2 ∨ Q4	Yes
7.	1	1	1	$\frac{p_{11.}}{p_{12.}} = \frac{p_{11.real}}{p_{12.real}}$	$\frac{p_{21.}}{p_{22.}} = \frac{p_{21.real}}{p_{22.real}}$	$\{1; 1\}$	Yes

In this way, consistency will be maintained between the associated group and the subgroups of PivotTable values. To find this indifferent ratio of probabilities in terms of associations of frequencies, we start from the theoretical equality of the associated frequencies:

$$n_{.12} = n_{.22} \tag{6}$$

After substituting for $n_{.22}$:

$$n_{.12} = n_{112} + n_{212} = n_{122} + n_{222}. \tag{7}$$

Thus:

$$n_{122} = n_{.12} - n_{222}. \tag{8}$$

Then, instead of the real ratio $\frac{n_{112}}{n_{11.}}$, the theoretical marginal probability $p_{11.}$ is expressed by the theoretical ratio:

$$p_{11.} = \frac{n_{122}}{n_{11.}} = \frac{n_{.12} - n_{222}}{n_{11.}} \tag{9}$$

After substituting the values for the combined frequencies, n_{122}, n_{222} and $n_{11.}$, we obtain the following:

$$p_{11.} = \frac{500 - 75}{1000} = 0.425 \tag{10}$$

The $p_{11.}$ = 0.425 value is the maximum value of the marginal probability (probability of infection with one dose of vaccine A) for the consistency of the pooled data with the data sorted by the number of vaccine applications (one dose or two doses). This value of marginal probability will vary not only depending on the associated frequencies, but will also be implicitly affected by the value of the likelihood of infection with a single dose of vaccine B. Therefore, it is appropriate to determine the probability of infection with a single dose of vaccine A concerning the likelihood of infection with a single dose of vaccine B.

Thus:

$$\frac{p_{11.}}{p_{12.}} = \frac{\frac{n_{.12} - n_{222}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} \tag{11}$$

After substituting values for the combined frequencies of n_{112}, n_{212}, $n_{11.}$, and $n_{12.}$ to the previous relationship, we obtain:

$$\frac{p_{11.}}{p_{12.}} = \frac{\frac{500-75}{1000}}{\frac{175}{250}} = 0.607 \tag{12}$$

The ratio $p_{11.}/p_{12.}$ = 0.607 is the theoretically critical (maximum) value that cannot be exceeded by the actual ratio of marginal probabilities $p_{11.\text{real}}/p_{12.\text{real}}$ to maintain the consistency of the results of the total (combined) file with sorted files, in this case, according to the method of treatment (vaccine used) in a single dose. In our case, the actual ratio of marginal probabilities $p_{11.\text{real}}/p_{12.\text{real}}$ is greater than the critical (maximum) value of the consistent ratio, i.e.,

$$\frac{p_{11.real}}{p_{12.real}} = \frac{0.475}{0.700} = 0.679 > 0.607 = \frac{p_{11.}}{p_{12.}}. \tag{13}$$

This relationship indicates a reversal of the correlation, where exceeding the critical value of the ratio of marginal probabilities leads to inconsistencies in the sub-files and the associated file. The primary cause of inconsistency is due to excessive unevenness of the associated frequencies in the classification by a number of vaccinations, which acts as the third factor (mediator factor) of causality, in addition to the number of vaccinations and the type of vaccine. We proceed similarly for two batch applications. We start again from the equality of combined frequencies for the positive tested for different vaccinations:

$$n_{.12} = n_{.22}. \tag{14}$$

After substituting for $n_{.22}$:

$$n_{.22} = n_{122} + n_{222} = n_{112} + n_{212}. \tag{15}$$

Thus:

$$n_{212} = n_{.22} - n_{212}. \tag{16}$$

Here, the theoretical marginal probability $p_{21.}$ is expressed by the theoretical ratio instead of the real ratio, $\frac{n_{212}}{n_{21.}}$:

$$p_{21.} = \frac{n_{212}}{n_{21.}} = \frac{n_{.22} - n_{112}}{n_{21.}}. \tag{17}$$

After substituting the values for the combined frequencies of n_{112}, $n_{21.}$, and $n_{22.}$, we obtain:

$$p_{21.} = \frac{250 - 475}{1500} = -0.150 \tag{18}$$

Thus, the marginal probability $p_{21.}$ (the association with both possible vaccination results (infection/without infection) is less than 0, i.e., outside its domain) that the sorted subsets indicate the same tendency of vaccination efficiency as their combined set is $n_{.12} = n_{.22}$. In reality, however:

$$n_{.12} > n_{.22}. \tag{19}$$

A more complex explanation is based on the extension of the range of probability values respective to the complex probability in the analogy of complex numbers. If we want to determine the likelihood of infection after the second dose of vaccine A in relation (in proportion) to the likelihood of infection with two doses of vaccine B, we can express this ratio as:

$$\frac{p_{21.}}{p_{22.}} = \frac{\frac{n_{.22} - n_{112}}{n_{21.}}}{\frac{n_{222}}{n_{22.}}}. \tag{20}$$

After substituting values for combined frequencies of n_{112}, n_{212}, $n_{11.}$, and $n_{12.}$, we obtain:

$$\frac{p_{21.}}{p_{22.}} = \frac{\frac{250-475}{1500}}{\frac{n_{222}}{n_{22.}}} \tag{21}$$

$$\frac{p_{21.}}{p_{22.}} = \frac{\frac{250-475}{1500}}{\frac{75}{2250}} = -4.5 \tag{22}$$

If we omit that the ratio of marginal probabilities is outside the range of values, then even in this case, the critical value of the ratio $p_{21.}/p_{22.} = -4.5$ is exceeded by the actual ratio of marginal probabilities $p_{21.\text{real}}/p_{12.\text{real}}$. In this case, the necessary condition to maintain the consistency of the total (combined) group with sorted groups according to the method of treatment (vaccine used) in one-dose administration is not fulfilled. In this case, the actual ratio of marginal probabilities $p_{21.\text{real}}/p_{22.\text{real}}$ is greater than the critical (maximum allowable) value of the consistent ratio.

Table 2 shows the rules for combinations of values of real marginal probabilities $p_{11.\text{real}}/p_{11.\text{real}}$ and $p_{21.\text{real}}/p_{22.\text{real}}$, and real associated probabilities $p_{.1.}/p_{.2.}$ with their respective theoretical ratios. Theoretical ratios are obtained by substituting the derived relations into (11) and (20). Thus, this table represents a small expert system to decide whether the data in a particular pooled table are consistent with its subgroup classifications. That is, whether we can trust the conclusions of the aggregated data. This expert system is complemented by the evidence represented by Formulas (23)–(39).

Proof for selected rows of Table 2

The seventh row of Table 2

We start from the simplest seventh situation Thus:

$$\frac{\frac{n_{.12}-n_{222}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} = \frac{\frac{n_{112}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} \tag{23}$$

Here, the denominators of the upper fractions of the equation are equal, as are the lower fractions of the equation. Therefore:

$$n_{.12} - n_{222} = n_{122}. \tag{24}$$

Which is applied for equations: $p_{.1.} = p_{.2.} \vee n_{112} = n_{122}$. Additionally, when: $n_{11.} = n_{12.}$, it must be applied again.

The first row of Table 2

Furthermore, to prove the validity of the first line of the situation, both real ratios are less than one, and it is assumed that the following applies:

$$\frac{p_{11.}}{p_{12.}} > \frac{p_{11.real}}{p_{12.real}} \vee \frac{p_{21.}}{p_{22.}} < \frac{p_{21.real}}{p_{22.real}}, \tag{25}$$

$$\frac{\frac{n_{.12}-n_{222}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} = \frac{\frac{n_{112}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}}. \tag{26}$$

To achieve Equality (26), we add 1 to the combined frequency n_{222}, and to maintain the total frequencies, we add this 1 to n_{212}, so that aby $n_{.1.} = n_{.2.} = const$. Let us mark the combined frequencies adjusted in this way with the index "*":

$$n^*_{222} = (n_{222} + 1) \vee n^*_{212} = (n_{212} + 1). \tag{27}$$

Then, Relation (26) is adjusted to the form:

$$\frac{\frac{n_{.12}-n^*_{222}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} \neq \frac{\frac{n_{112}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}}. \tag{28}$$

Because we reduce the value of the left side of the previous equation and left the right side unchanged, the inequality must apply:

$$\frac{\frac{n_{.12}-n_{222}+1}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} < \frac{\frac{n_{112}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} \tag{29}$$

Thus:

$$\frac{p_{11.}}{p_{12.}} < \frac{p_{11.real}}{p_{12.real}} \tag{30}$$

The following inequality applies to the second share of theoretical probabilities:

$$\frac{p_{21.}}{p_{22.}} \neq \frac{p_{21.real}}{p_{22.real}} \tag{31}$$

Expressed as a ratio of the associated frequencies, we obtain:

$$\frac{p_{21.}}{p_{22.}} \neq \frac{\frac{n_{.22}-n_{112}}{n_{21.}}}{\frac{n_{222}}{n_{22.}}}. \tag{32}$$

Under the achievement of associated frequencies instead of marginal frequencies, we obtain the equation:

$$\frac{\frac{n_{212}}{n_{21.}}}{\frac{n_{222}}{n_{22.}}} \neq \frac{\frac{n_{.22}-n_{112}}{n_{21.}}}{\frac{n_{222}}{n_{22.}}}. \tag{33}$$

Because we introduce star frequencies on the left side, increase the frequency n_{212} by 1, and increase the numerator $n_{222}+1$ (i.e., decreased the left side of the inequality) at the same time, the following must apply:

$$\frac{\frac{n_{212}}{n_{21.}}}{\frac{n_{222}}{n_{22.}}} > \frac{\frac{n_{.22}-n_{112}}{n_{21.}}}{\frac{n_{222}}{n_{22.}}}. \tag{34}$$

The third row of Table 2

Assume that, unlike the situation represented by line 7, the probability of p_{11} is much greater than the probability of p_{21}, and also, that of p_{12} is much greater than the probability of p_{22}:

$$p_{11.} >> p_{21.} \vee p_{12.} >> p_{22.}. \tag{35}$$

In our case, this difference is because the effectiveness of two doses of a vaccine is 20–30 times more than that of one dose. Next, suppose that the combined frequency of n_{112} is much greater than the frequency of n_{212}, and also that n_{112} is much greater than the frequency of n_{112}:

$$n_{112} >> n_{212} \tag{36}$$

Thus, the efficacy of two doses of vaccine is significantly higher, but concerns, for example, a significantly smaller population in one case. The marginal frequency $n_{21.}$ and $n_{22.}$ do not differ by order. Let us make these frequencies equal to x:

$$n_{21.} \cong n_{22.} \cong x$$

$$\frac{\frac{n_{112}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} \neq \frac{\frac{n_{.12}-n_{222}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} \tag{37}$$

Because the lower fractions in the previous equation are equal to:

$$\frac{n_{122}}{n_{12.}} = \frac{n_{122}}{n_{12.}}. \tag{38}$$

Likewise, the denominators of the upper fractions are equal: $n_{11.} = n_{11.}$ At the same time, the following applies: $n_{112} > n_{.12} - n_{222}$. Then, the following must apply:

$$\frac{\frac{n_{112}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}} > \frac{\frac{n_{.12}-n_{222}}{n_{11.}}}{\frac{n_{122}}{n_{12.}}}. \tag{39}$$

The left side of the equation represents the ratio of real marginal probabilities, and the right side of the equation represents the ratio of the respective theoretical probabilities.

Similarly, it is possible to make evidence for the remaining rows of the table.

4. Discussion

The theoretical goal was to create an analytical solution for determining the critical ratio of partial probabilities in terms of the consistency of conclusions with a merged group of data. This task was solved through part 2.

For practical purposes, the subsequent goal was formulated in the form of a stability diagram. Therefore, a stability diagram is created in order to process control and for use in the visual assessment of the consistency between causality and contingency (see Figure 1). The stability diagram is divided into four real quadrants (Q1, Q2, Q3, and Q4), supplemented by two complex quadrants (Q3c and Q4c) and three forbidden areas (i.e., more precisely than the quadrates; the areas should be called ninths of the graph). Here, the Cartesian system is shifted to point [1,1]. This shift is because real–theoretical likelihood ratios are applied in the system, where the horizontal axis is determined for real likelihood ratios, and the vertical axis is determined for theoretical probability ratios. A ratio value equal to one indicates the identity of the marginal frequencies for the data subsets for real ratios. A value different from one then tends to dominate in a given classification of one set.

Because we always calculate two ratios (from real and also from theoretical probabilities (in our case, from two states, and from two treatments), we always have a combination of two resulting values (each value has its real and theoretical coordinates). A totally stable solution (data consistent with both subsets) is only possible when placing just one theoretical ratio in the interval $(0;1)$ and the other theoretical ratio in the interval $(1;\infty)$; therefore, the point [1,1] is selected as the primary (central) point of the diagram of stability and consistency of data subgroups with their merged group. In other words, for total consistency, it is sufficient if one point (given by the coordinates of theoretical and real marginal ratios) lies in a conditionally unstable region (Q_1 or Q_3). The other point (provided by the second coordinates of theoretical and real marginal ratios) lies in a stable region (Q_2 or Q_4). This finding has a surprising practical impact on finding a stably consistent solution in which all data subgroups are consistent (e.g., in terms of the magnitude of the effect of the vaccine) with their associated data group. A sufficient condition to ensure the consistency of all data subgroups with their associated data group is based on finding exactly one point (given by the coordinates of theoretical and real marginal ratios) in the stable region and exactly one point in the conditionally unstable region. If we connect these two points with a line, this line intersects the boundary between the stable and conditionally unstable areas. Point [1,1] is then a special case, where all associations of theoretical and real marginal probabilities are equal.

The point [1,1] geometrically intersects the boundary of a conditionally unstable and stable region in just one place. Therefore, in this situation, the solution of data consistency is also totally stable. Conversely, a sufficient condition for finding an unstable implementation of data subgroups with their merged group is if at least one point (given by the coordinates of theoretical and real marginal ratios) lies in an unstable region (Q3c or Q47c). This is just a case of a data paradox. Data subgroups indicate the exact opposite conclusion to their combined data group (e.g., in terms of vaccine effect, one subgroup shows better efficiency in both applications; after merging the data, the second subgroup appears to be better). Another possibility is that both points (given by the coordinates of the theoretical and real marginal ratios) lie in a stable region, but each point lies in a different quadrant (Q2 and Q4). In this case, it is a stable solution (or particularly stable), which is realized by just one case of consistency of the data subgroup with the merged data group. The last possibility is the description of three forbidden areas, which express the range of values of real ratios of marginal probabilities less than zero. Such a realization is impossible even in the field of complex values of probabilities. Therefore, they are marked as forbidden.

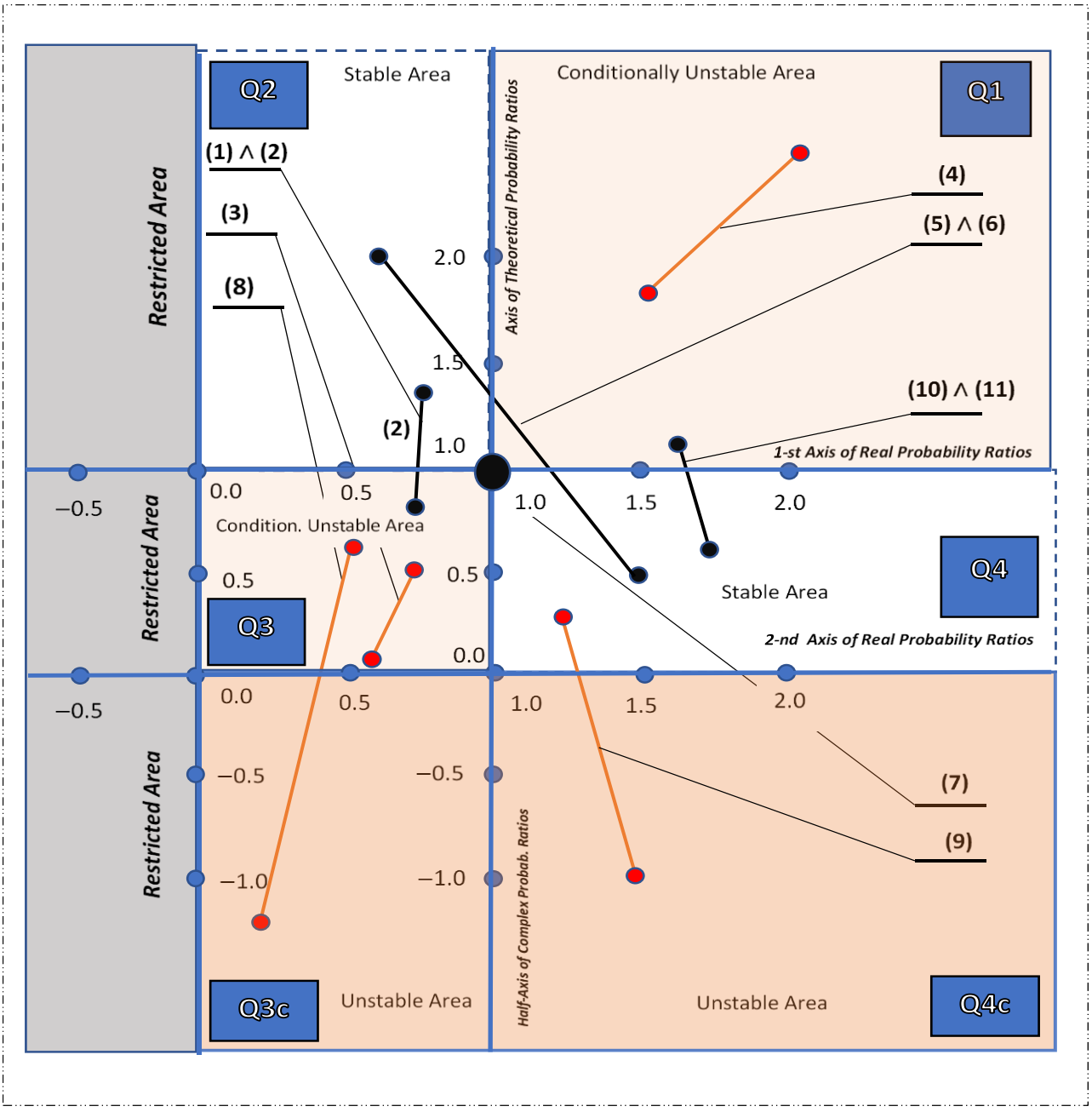

Figure 1. Stability diagram of contingent subgroups with an associated group.

5. Conclusions

Undoubtedly, in statistics, the larger the amount of data, the more reliable the results. There is a case where partial relations are significantly stronger than zero-order relationships (this association is shown in the original, aggregated table). Still, this weak zero-order relationship is significantly amplified when a third variable is introduced, called (in this case) a suppressor variable, to the point where the zero relationships are completely reversed. A paradoxical phenomenon occurs, where the correlation (with respect to contingency or association of data) implies the opposite causality (e.g., consequences precede their causes, or contingency subgroups indicate opposite conclusions than aggregated groups). For practical use, this is supplemented by a diagram of the stability of contingent subgroups with an associated group, which allows for the easy identification of cases of the data paradox.

Subsequent research on this topic will be based on solving cases where the ratio of marginal probabilities is outside the range of values. For this purpose, a theory of complex probability will be introduced, which will use the direction vector of the square of a certain phenomenon. This phenomenon will even make it possible to formally solve situations where the ratio of marginal probabilities is outside the range of values. Furthermore, this use of the direction vector of the square of a certain phenomenon will allow for consistent and inconsistent cases of association to be differentiated in a different way and for correlations to be made regarding the relation of causality.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available within the article.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Fisher, R.A. *Statistical Methods for Research Workers*, 2nd ed.; Oliver and Boyd: London, UK, 1928.
2. Neyman, J.; Pearson, E.S. IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math Phys. Character* **1933**, *231*, 289–337. [CrossRef]
3. Reeves, J.H.; Royall, R.M. Statistical Evidence: A Likelihood Paradigm. *J. Am. Stat. Assoc.* **1998**, *93*, 1235. [CrossRef]
4. Setchi, R.; Anuar, F.M. Multi-faceted assessment of trademark similarity. *Expert Syst. Appl.* **2016**, *65*, 16–27. [CrossRef]
5. Hesamian, G. One-way ANOVA based on interval information. *Int. J. Syst. Sci.* **2016**, *47*, 2682–2690. [CrossRef]
6. Cohen, J. The earth is round ($p < 0.05$). *Am. Psychol.* **1994**, *49*, 997–1003. [CrossRef]
7. Nester, M.R. An Applied Statistician's Creed. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **1996**, *45*, 401. [CrossRef]
8. Braeken, J.; Mulder, J.; Wood, S. Relative Effects at Work. *J. Manag.* **2014**, *41*, 544–573. [CrossRef]
9. Kruschke, J.K.; Aguinis, H.; Joo, H. The Time Has Come. *Organ. Res. Methods* **2012**, *15*, 722–752. [CrossRef]
10. Sirvanci, M.B.; Durmaz, M. Variation Reduction by the Use of Designed Experiments. *Qual. Eng.* **1993**, *5*, 611–618. [CrossRef]
11. Cherry, S. Statistical Tests in Publications of the Wildlife Society. *Wildl. Soc. Bull.* **1998**, *26*, 947953.
12. Cox, D.R. Some Problems Connected with Statistical Inference. *Ann. Math. Stat.* **1958**, *29*, 357–372. [CrossRef]
13. Pearl, J. Comment: Understanding Simpson's Paradox. *Am. Stat.* **2014**, *68*, 8–13. [CrossRef]
14. Witmer, J. Simpson's Paradox, Visual Displays, and Causal Diagrams. *Am. Math. Mon.* **2021**, *128*, 598–610. [CrossRef]
15. Spanos, A. Yule–Simpson's paradox: The probabilistic versus the empirical conundrum. *J. Ital. Stat. Soc.* **2021**, *30*, 605–635. [CrossRef]
16. Liebl, D. Nonparametric testing for differences in electricity prices: The case of the Fukushima nuclear accident. *Ann. Appl. Stat.* **2019**, *13*, 1128–1146. [CrossRef]
17. Ma, Y.Z. Simpson's paradox in GDP and per capita GDP growths. *Empir. Econ.* **2015**, *49*, 1301–1315. [CrossRef]
18. Carletti, M.; Pancrazi, R. Geographic Negative Correlation of Estimated Incidence between First and Second Waves of Coronavirus Disease 2019 (COVID-19) in Italy. *Mathematics* **2021**, *9*, 133. [CrossRef]
19. Eubank, R.; Hsing, T. Canonical correlation for stochastic processes. *Stoch. Process. Their Appl.* **2008**, *118*, 1634–1661. [CrossRef]
20. Kupresanin, A.; Shin, H.; King, D.; Eubank, R. An RKHS framework for functional data analysis. *J. Stat. Plan. Inference* **2010**, *140*, 3627–3637. [CrossRef]
21. Huang, Q.; Renaut, R. Functional partial canonical correlation. *Bernoulli* **2015**, *21*, 1047–1066. [CrossRef]
22. Martínez-Galicia, D.; Guerra-Hernández, A.; Cruz-Ramírez, N.; Limón, X.; Grimaldo, F. Windowing as a Sub-Sampling Method for Distributed Data Mining. *Math. Comput. Appl.* **2020**, *25*, 39. [CrossRef]
23. Jiang, W.; Song, S.; Hou, L.; Zhao, H. A set of efficient methods to generate high-dimensional binary data with specified correlation structures. *Am. Stat.* **2020**, *75*, 1–37. [CrossRef]