Maltese Medical Journal, 2000; 12(1,2): 46

INVITED ARTICLE
Measuring reliability and consistency in contingency tables

Anton Buhagiar

ABSTRACT: The association between two categorical variables is very often assessed by making a cross-tabulation and calculating the χ^2 statistic for that table. However there are many other related parameters which can be used to assess subtle patterns in the table. In this article we will discuss parameters which can be fruitfully used in situations such as :

- **the test-retest method for the reliability of questions in a pilot questionnaire,**
- **the measurement of the change of people's attitude with time,**
- **the comparison of two medical diagnoses of a given patient, and**
- **the prediction of heart disease status using an independent risk scale.**

Correspondence: Prof. Anton Buhagiar, Department of Mathematics, University of Malta, Msida.

Keywords: Cross-tabulations, contingency tables, proportions, association between categorical variables, Pearson's χ^2 statistic, test-retest, change of attitude with time, comparison of medical diagnoses, efficacy of risk scale, Kappa measure of reliability, correlation, McNemar's test of symmetry, test of marginal homogeneity, test for linear trend.

Introduction

Everything dealing with the collection, processing, analysis, and interpretation of numerical data belongs to the realm of statistics. In medicine, this could include such diversified tasks as calculating the average duration of a patient in hospital, collecting and presenting data on the numbers of persons afflicted with a given disease in a given year, evaluating the effectiveness of a given analgesic, predicting the reliability of an X-ray machine, or describing the cardiogram of a patient. One can never deny the importance of descriptive statistics in almost every major phase of human activity.

In recent years, however, there has been a shift in emphasis from descriptive statistics to statistical inference. Who has not heard of the t-test or analysis of variance to compare the means of variables in two or more random samples, or the paired t-test or repeated measures anova to describe how the mean of a variable changes with time for one or more groups of individuals? We will discuss these in future articles of this journal.

In the present article, we will concern ourselves with inferences about proportions rather than means of variables. As seen further on in this article, the standard method of comparing proportions is by cross tabulation of the categorical variables concerned. We will also discuss various important parameters associated with these cross-tabulations, and the practical situations where they are relevant.

Contingency tables and Pearson's χ^2 Statistic

It is very common in statistical work to examine the relation between two categorical variables. For example a sociologist might pose the question whether males in Malta are more in favour of the EU than females. That is, does the attitude towards the EU depend on gender? In statistical jargon, one would be testing whether the proportion of males in favour of the EU is equal to the corresponding proportion of females. In another application, a medical doctor might require to evaluate the efficacy of an anti-influenza injection. The proportion of subjects who subsequently contracted influenza even though they had the injection is compared to the corresponding proportion of subjects who did not have the injection. If the two proportions are similar, it would mean that the injection is not effective in the prevention of influenza.

The standard way to tackle such problems is to set up a contingency table or a cross-tabulation of the two variables and perform the χ^2 test[1] to determine whether an association exists between the two variables. If we do this for the first application given above, we will end up with a cross-tabulation such as the following, all data being fictitious:

Gender	Males	Females	**Total**
Attitude			
In favour	82	79	161
Against EU	75	63	138
Total	157	142	299

In this table 82 out of 157 males (that is 52% of the sampled males) were in favour of the EU, as opposed to 79 out of 142 for females (that is 56% of the sampled females). When the χ^2 statistic is worked out for this table we get:

Statistic	Value	Degrees of freedom	Probabilities
Pearson χ^2	0.348	1	0.5554

It can be seen that the χ^2 statistic is quite small and is not significant since its probability value of 0.5554 exceeds the critical p-value of 0.05. The attitude towards the EU is therefore unrelated to gender. The proportion of males in favour of the EU is practically equal to that of females.

Similarly to measure the efficacy or otherwise of an influenza injection, it was noted that from a total of 443 people who were administered the injection, only 12 subsequently contracted influenza, that is a proportion of 2.7%. Conversely in a control group who were not administered this injection, it was noted that 38 out of a total of 525 contracted influenza, a proportion of 7.2%. These figures (again fictitious) are summarised in the following table.

Group	Without Injection	With Inj.	Total
Influenza			
Caught Influenza	38	12	50
Did not catch flu	487	431	918
Total	525	443	968

Statistic	Value	Degrees of freedom	Probabilities
Pearson χ^2	10.062	1	0.0015

It can be noted in this case that the percentage of subjects who are affected by influenza decreases from 7.2% in the untreated group to 2.7% for the sample who were administered the injection. This is quite a sizeable decrease, as is evidenced by the high χ^2 value of 10.062. The probability of this χ^2 value is in fact 0.0015, which is highly significant, being less than the critical value of 0.05. One can thus deduce that the injection is indeed effective in protecting subjects from influenza.

Other parameters associated with contingency tables

As shown briefly above, the χ^2-statistic is the most popular parameter used by researchers in cross-tabulations. However, in a contingency table, numerous other statistical parameters and their significance can be estimated. Different parameters could be relevant, depending on the hypothesis one sets out to prove, on the number of rows and columns of the table, and on the types of variables featuring in the table.

For measuring reliability and consistency, the most popular parameters are the kappa coefficient of reliability, the related McNemar statistic and the coefficient of marginal homogeneity[1,2,4]. These are valid for all square tables, including those with unordered categories. If both categorical variables are of the ordinal type, the correlation coefficients of Pearson and Spearman are most widely used to measure the extent of association between the two[1,2,4]. The correlation coefficients can also be worked out for rectangular tables. The test of linear trend[1,2] can also be used to see whether a proportion is increasing over the ordered categories of a second variable. This would be relevant for rectangular tables where one variable has two categories, and the other variable has three or more ordered categories.

To illustrate the correct use of these parameters we will discuss some practical examples including:

i) the test-retest method for the reliability of questions in a pilot questionnaire,
ii) measurement of the change of people's attitude with time,
iii) comparison of two medical diagnoses of a given patient, and
iv) prediction of heart disease status using an independent risk scale.

The test-retest method for the reliability of questions in a questionnaire

A pilot questionnaire is sometimes administered on two different occasions to the same respondents to determine how clear the questions are to them. Unambiguous and clear questions are given identical answers in both occasions. Conversely, questions which are not well understood by the respondents tend to be answered differently in separate instances. Subsequently, such questions should be either clarified by suitable rewording or should be omitted outright from the main questionnaire.

To illustrate this imagine that in a pilot survey three questions Q1, Q2 and Q3 could all have *no* and *yes* as possible answers, 0 and 1 being their respective codes. We can assume that the survey was administered twice to the same respondents, and we note the responses on the two occasions. The cross-tabulation for question Q1 and its statistical analysis are as follows:

Q1: First Time	No	Yes	Total
Second Time			
No	34	0	34
Yes	0	42	42
Total	34	42	76

Statistic	Value	ASE1	T-Value	P-Value
Kappa, Meas. Reliability	1.000	0.000	8.718	0.0000
Assuming ordinal categories:				
Correlation Coefficient	1.000	0.000	41.179	0.0000
Spearman Rank Corr.	1.000	0.000	41.179	0.0000

We note that 34 respondents answered NO to Q1 on both occasions, and 42 answered YES on both occasions. Not one single respondent answered differently on different occasions, and they did not seem to find the

question Q1 ambiguous in any way. This is reflected in a high value for the coefficient Kappa, which measures the reliabilty of the given question. In fact the value 1.00 for Kappa indicates perfect agreement for Q1 in the test-retest experiment. If we interpret the categories as ordinal, one can also use the correlation coefficients of Pearson and Spearman, which in this case also attain their highest value of 1.00 for this case of perfect agreement.

Assume now, that on repeating the same exercise for question Q2, we get the following table:

Q2:	First Time	No	Yes	Total
	Second Time			
	No	30	6	36
	Yes	4	36	40
	Total	34	42	76

Statistic	Value	ASE1	T-Value	P-Value
Kappa, Meas. Reliability	0.735	0.078	6.420	0.0000
Assuming ordinal categories:				
Correlation Coefficient	0.736	0.077	9.323	0.0000
Spearman Rank Corr.	0.736	0.077	9.323	0.0000

In this case, 30 respondents said NO on both occasions, while 36 answered YES. However, in this case, 6 subjects answered YES the first time and NO in the retest, whilst another 4 answered NO at first and YES subsequently. In this table again there is a preponderance of frequencies down the main diagonal NO-NO to YES-YES, but the off-diagonal terms are now no longer zero, as they were in Q1. In fact, the kappa coefficient of reliability and the coefficient of correlation have decreased from the maximum possible value of 1.0 in Q1 to the smaller value of 0.7 in Q2 indicating the lesser degree of agreement in the test-retest for Q2.

Again assume that the test retest for Q3 is given as follows.

Q3	First Time	No	Yes	Total
	Second Time			
	No	28	18	46
	Yes	6	24	40
	Total	34	42	76

Statistic	Value	ASE1	T-Value	P-Value
Kappa, Meas. Reliability	0.382	0.100	3.503	0.0004
Assuming ordinal categories:				
Correlation Coefficient	0.402	0.101	3.911	0.0001
Spearman Rank Corr.	0.402	0.101	3.911	0.0001

We note that the number of discrepancies between the test and retest is now 18 YES/NO and 6 NO/YES. The number of discrepancies has increased further from Q2, so we will expect that kappa and the correlation will be even smaller for Q3 than for Q2. As expected, these parameters work out to about 0.4 as shown above. Although this value is significantly different from zero ($p<0.0005$) for all three parameters, 0.4 is small in absolute value, and the reliability is poor in this case. It is important to note that here, it is the magnitude of the parameter, which is even more important than its significance. The kappa measure of reliability is usually very similar to the value of the correlation (as shown in the above three cases), and both can vary between 1 and –1. Kappa can be used in all square tables where either or both categories are not necessarily ordered. Conversely, the correlation coefficient can be used for any shape of table, provided both variables have ordered categories.

Measuring change of attitude with time

In a similar vein to the above, using appropriate statistical parameters, one can measure to what extent the attitude of people changes with time. Assume that in Q4, people were asked whether they agreed with a certain war. Each person was asked this question just before the war started, and then again two months after it ended. We would like to determine to what extent people changed their minds between the two interviews. In the following table, we give the observed frequencies, followed by the relevant statistical output.

Q4:	First Time	Agree	Disagree	Unsure	Total
	Second Time				
	Agree	47	28	26	101
	Disagree	56	61	47	164
	Unsure	38	31	10	79
	Total	141	120	83	344

Statistic	Value	Degrees of freedom	Probabilities
Pearson χ^2	11.584	4	0.0207
McNemar Test of Symmetry	14.865	3	0.0019
Marginal Homogeneity	14.778	2	0.0006

Statistic	Value	ASE1	T-Value	P-Value
Kappa, Meas. Reliability	0.001	0.036	0.039	0.97

One can immediately notice in the cross-tabulation that there is no preponderance of frequencies along the main diagonal. In fact only 47, 61 and 10 people gave the same response on the two occasions (118 out of 344, or about one-third of all respondents). In fact this leads to a kappa value of zero as shown in the last line of the output. One can also notice from the above frequencies that 56 people out of 344 (that is 16% of all respondents) changed from AGREE before to DISAGREE after the war, whereas only 28 out of 344 people (that is 8%) changed their mind from DISAGREE before the war to AGREE afterwards. This lack of symmetry in the table is reflected in the large value of McNemar's statistic (14.9) which is highly significant ($p=0.002$).

McNemar's statistic will only be small when the square table is nearly symmetrical about the main diagonal. Another parameter is that of marginal homogeneity, which compares the marginal proportions for rows and columns (before and after) in a square table. The ratio Agree: Disagree: Unsure is 141: 120: 83 before the war, and 101: 164: 79 after the war. Again the parameter is large (14.8) and highly significant (p=0.0006), thus showing that the two ratios have changed considerably over that time period. In fact the proportion of people disagreeing increased from 35% before the war to 48% afterwards. It is important to note that the parameters kappa and McNemar and the statistic for marginal homogeneity are all relevant for a square table with categories, which are not necessarily ordered.

Agreement of two diagnoses

One can also use kappa to measure how well two doctors agree in diagnosing the same 77 psychiatric patients. Each patient was diagnosed by the two doctors separately. The result is given in the cross-tabulation below.

One would like to ask how concordant are the two diagnoses? Again the kappa coefficient can be used for this purpose. In this case kappa is equal to 0.6 and is highly significant (p=0.0000). We see that there is good agreement between the two doctors, because the counts along the diagonal differ significantly from those expected by chance. According to Fleiss[1], values of kappa below 0.40 reflect poor agreement, whereas values between 0.40 and 0.75 indicate agreement which is fair to good. Values of kappa above 0.75 imply that there is strong agreement between the two diagnoses.

If the above categories are assumed to be ordinal, the correlation coefficient can also be quoted. In fact its value of 0.65 is also highly significant and is very similar to the value of kappa.

McNemar's statistic (P=0.18) and that of marginal homogeneity (P=0.10) are not significant in this case. This shows that the table is essentially symmetrical about the main diagonal, and that the marginal ratio schizo : manic : other for Doctor A is equal to that of Doctor B.

Diagnoses: Doctor A	Schizo	Manic	Other	**Total**
Doctor B				
Schizo	24	5	3	32
Manic	2	16	1	19
Other	3	6	17	26
Total	29	27	21	77

Statistic	Value	Degrees of freedom	Probabilities
Pearson χ^2	58.854	4	0.0000
McNemar Test of Symmetry	4.857	3	0.1826
Marginal Homogeneity	4.677	2	0.0965

Statistic	Value	ASE1	T-Value	P-Value
Kappa, Meas. Reliability	0.609	0.074	7.657	0.0000
Assuming categories were ordered:				
Correlation Coefficient	0.646	0.089	6.159	0.0000

The efficacy of a risk scale in predicting true heart disease status

One would sometimes like to measure the efficacy of a risk scale in predicting the presence or otherwise of disease. We now give an illustration of this.

Two hundred subscribers to a company health plan were surveyed on lifestyle factors and then rated on their risk for heart disease on a Likert scale varying from 1, meaning very low risk, to 5, meaning very high risk. True heart disease status, diseased or healthy, was independently determined for each subscriber. How successful is the five-point risk scale in predicting whether a subscriber is diseased or healthy?

For this purpose one can arrange the data as a cross-tabulation as follows:

Likert Scale for Risk	1	2	3	4	5	Total
Actual						
Diseased	3	1	3	4	12	23
Healthy	67	41	37	21	11	177
Total	70	42	40	25	23	200

Statistic	Value	Degrees of freedom	Probabilities
Pearson χ^2	45.524	4	0.0000
Test for Linear Trend	29.682	1	0.0000

One can note from the table of observed frequencies that as the risk increases from 1 to 5, the proportion of diseased subjects increases gradually from 3/70 for risk level 1, through 1/42 for risk level 2, 3/40, 4/25 until it reaches 12/23 for those with the highest risk level of 5. The corresponding percentages for the five levels of risk are 4%, 2%, 8%, 16%, and 52% respectively. Except for risk level 2, there is a consistent increase in the percentage of diseased persons as the risk level increases.

The statistic of linear trend is the parameter which best measures the increase in the percentage of diseased persons over the ordered categories of risk. As shown above, the test for linear trend is highly significant with P=0.0000. This means that the percentage of diseased persons increases significantly with increasing level of risk.

This statistic can be defined on any table with 2 rows and n columns, or 2 columns and n rows, where n in each case is larger than or equal to three. The variable corresponding to the n categories has to be of the ordinal type, as in a Likert scale.

To further investigate the success of the risk scale in predicting disease one can perform a logistic regression of the binary variable (disease status) on the risk level.

The area of 0.81 under the resulting Receiver Operating Characteristic (ROC) curve shows that the 5-point risk scale correctly predicts diseased and healthy healthplan subscribers with a probability of 81%.

Statistical Analyses

The above statistical analyses were performed with BMDP, the Bio-Medical Data Package[2,3]. In particular we used the program 4F for contingency tables[2], and the program LR for logistic regression[3]. The above analyses can also be easily performed with other programs such as SPSS[4,5].

Suggestions for further reading

The subject of contingency tables and the ubiquitous χ^2 test is treated in many elementary textbooks on statistics. In particular, two books[6,7] are very readable and should be readily intelligible to most readers.

Three other books[8,9,10] are excellent biostatistical texts and all contain a good section on contingency tables. The book by Fleiss[1] is an important book dedicated solely to contingency tables in the medical field. Although specialised, it is very readable and discusses all the statistical parameters described above. One must also recall the reference manuals of statistical software packages like BMDP[2,3] or SPSS[4] which contain many interesting examples, not only on contingency tables but also on the many other techniques normally encountered in statistics.

References

1. Fleiss J. L. Statistical Methods for Rates and Proportions. 2nd Edition. New York, Wiley. 1981.
2. Dixon W.J. BMDP Statistical Software Manual, Volume 1. Berkeley, University of California Press. 1992. See in particular program 4F, mostly pages 282 to 287.
3. Dixon W.J. BMDP Statistical Software Manual, Volume 2. Berkeley, University of California Press. 1992. See in particular program LR, mostly pages 1105 to 1109, and pages 1126 to 1128.
4. Norusis M.J. SPSS Base System Use's Guide. Chicago, SPSS Inc. 1990.
5. Bryman A. & Cramer D. Quantitative Data Analysis with SPSS for Windows. London, Routledge. 1997.
6. Shennan S. Quantifying archaeology. Edinburgh, Edinburgh University Press. 1988.
7. Freund J.E. & Simon G.A. Modern elementary statistics. 9th Edition. New York, Prentice-Hall. 1997.
8. Sokal R.R and Rohlf F.J. Introduction to biostatistics. 2nd Edition. New York, W.H. Freeman and Company. 1987.
9. Zar J.H. Biostatistical Analysis. 2nd Edition. New Jersey, Prentice-Hall. 1984.
10. Sokal R.R. and Rohlf F.J. Biometry. 3rd Edition. New York, W.H. Freeman and Company. 1995.