Clinical Research Methods

Hypothesis testing (Part I)

ABHAYA INDRAYAN, PIYUSH GUPTA

INTRODUCTION

After the general principles of statistical testing of hypotheses, this article describes some methods of inference from proportions including confidence intervals for relative risk and odds ratio. Part II will deal with tests of hypothesis on means.

NULL HYPOTHESIS AND *P* VALUE

Court judgment

The concepts of null hypothesis and *P* value are best understood with the help of an example of a court decision in a crime case. When a case is presented before a court of law by the prosecution, the judge is supposed to start with the presumption of innocence. It is up to the prosecution to provide sufficient evidence against the innocence of the person and change the initial opinion of the judge. Guilt should be proved beyond reasonable doubt. If the evidence is not sufficient, the person is acquitted, whether the crime was actually committed or not. Sometimes the circumstantial evidence is strong and an innocent person is wrongly pronounced guilty. This is considered to be a serious error. Special caution is exercised to guard against this type of error even at the cost of acquitting some criminals!

Similarly, clinicians look at a lot of evidence including clinical features, laboratory investigations, etc. before reaching a diagnosis. Despite this, *misdiagnosis* (disease is not present but diagnosed as present) is not uncommon. This is a more serious error than *missed diagnosis,* which occurs when the disease is present but missed. Similar errors can occur in statistical decisions.

Null hypothesis

In the case of statistical decisions too, the assumption initially made is that there is no difference between the groups. This is equivalent to the presumption of innocence in the court setting and is called the *null hypothesis.* The notation used for this is H_0. This is the hypothesis under scrutiny and sought to be refuted by conducting a study on a sample of subjects.

The sample observations serve as evidence. Depending upon this evidence, H_0 is either rejected or not rejected. In an empirical set-up, H_0 is never accepted though it could be conceded. The conclusion reached is that the evidence is not enough to reject H_0. This may mean two things: (i) further evidence needs to be collected, and (ii) continue to accept the present knowledge as though this investigation was never done. The 'truth' remains unchanged in this case.

University College of Medical Sciences, Dilshad Garden, Delhi 110095
ABHAYA INDRAYAN Department of Biostatistics and Medical Informatics
PIYUSH GUPTA Department of Paediatrics
Correspondence to ABHAYA INDRAYAN

Alternative hypothesis

When H_0 is rejected, what would be accepted? The alternative hypothesis is the assertion accepted when H_0 is rejected. This is denoted by H_1. The alternative hypothesis could be one-sided where either superiority or inferiority is accepted. In case of a two-sided H_1, only the difference is accepted without any inference as to which is better or higher.

P value

We stated earlier that the values observed in the sample serve as evidence against H_0. The error of rejecting a true null hypothesis is similar to punishing an innocent. This is more serious and is called a *Type I error*; popularly referred to as *P* value. Thus, *P* value is the probability that a true null hypothesis is wrongly rejected. The maximum *P* value allowed in a problem is called the *level of significance* or sometimes the α-level. *P* value is the probability that the conclusion of presence of difference is reached when actually there is no difference. Being serious, *P* value is kept at a low level, mostly less than 5%, or $P<0.05$. When *P* value is this low or lower, it is generally considered safe to conclude that the groups are indeed different. This is also called *statistical significance.* This threshold, 0.05, is the level of significance.

Type II error. The second type of error is failing to reject H_0 when it is false. This corresponds to missed diagnosis or pronouncing a criminal as not guilty. The probability of this error is denoted by β. In a clinical trial, this is equivalent to declaring a drug ineffective when it is effective. This is not so much of an error, albeit it is an error. If the manufacturer believes that the drug is really effective, the company will carry out further trials and collect more evidence. Thus, the only effect of Type II error is that the introduction of the drug is delayed but not denied.

Type II error is calculated after fixing the level of significance α, and for a specific value under the alternative hypothesis.

Power. The complementary of Type II error is called power. Thus, the power of a test is the probability of correctly rejecting an H_0 when it is false. This is the probability of getting a statistically significant result. Power depends on the magnitude of the difference really present in the target population and the level of significance. The power of a test is high if it is able to detect a small difference and thus reject H_0.

Power becomes especially important when the investigator does not want to miss a specified difference. For example, a hypotensive drug may be considered useful if it reduces diastolic blood pressure by an average of at least 5 mmHg after being used for, say, one week. A sufficiently powerful statistical test would be needed to detect this difference with a high probability. Thus $(1-\beta)$ is an important consideration in this set-up. However, the difference (5 mmHg in this case) should be chosen on some objective evidence.

General procedure to obtain a P value

The exact form of test criterion for obtaining *P* depends on:

1. the nature of the data (qualitative or quantitative),
2. the form of distribution if the data are quantitative (Gaussian or non-Gaussian),
3. the number of groups to be compared (two or more than two),
4. the parameter to be compared (can be mean, median, correlation coefficient, etc. in the case of quantitative data),
5. the size of the sample (small or large), and
6. the number of variables considered together (one, two or more).

The exact criteria used to test different statistical hypotheses are described subsequently. The distributional form of these criteria has been obtained and is known. These are used to find *P* values according to the following procedure:

Step 1. Set up a null hypothesis and decide that the alternative is one-sided or two-sided.

Step 2. Identify a criterion suitable for the set-up in hand. This would depend on items 1–6 enumerated above. These criteria are also called tests of statistical significance.

Step 3. Use sample observations to calculate the value of the criterion assuming that the null hypothesis is true.

Step 4. Compare this value with its known distribution, and assess the probability of occurrence of a value of the criterion, which is as, or more extreme towards H_1 than that obtained in Step 3. This is calculated for both the negative and positive side when the alternative is two-sided. Since the comparison is with the distribution under H_0, a high probability indicates that the H_0 is plausible and cannot be rejected. A very low probability indicates that the H_0 is very unlikely to be true and deserves to be rejected. This probability is the *P* value.

Step 5. Decide a threshold α of the *P* value that can be tolerated for the problem at hand. Reject H_0 if *P* is less than this threshold, otherwise not. Generally, $P<0.05$ is considered low enough for H_0 to be rejected. Statistical significance is said to have been achieved when H_0 is rejected. Such a result can be stated in a variety of ways:

- The evidence against the null hypothesis is sufficient to reject it.
- The chance that the null hypothesis is true is very small (or, the null hypothesis is extremely unlikely to be true).
- The alternative hypothesis is accepted.
- The *P* value is less than a predetermined threshold, e.g. 0.05.
- The probability of wrongly rejecting H_0 (Type I error) is very small.
- The result achieves statistical significance.

All these statements are the same. They illustrate how a decision was reached despite the uncertainty. Conceding a null hypothesis can also be similarly stated in a number of different ways.

METHODS OF HYPOTHESIS TESTING: INFERENCE FROM PROPORTIONS

Tests of significance are standard statistical procedures for drawing inferences from the sample about the target population. Sample estimates are never exact, being subject to sampling errors. Tests of significance allow us to decide whether the sample estimates, or the differences between estimates, are within their normal sampling variation.

Inferential methods are different for qualitative and quantitative variables. Variables in medicine and health are predominantly qualitative in nature and lead to proportions, rather than means, which are obtained on quantitative data. The methods of this article are applicable to situations where the interest is in proportions. Procedures to draw inferences from means are presented in a subsequent article.

In this article we shall deal separately with three situations. First, where only one qualitative variable is under consideration and the interest is in finding out whether the population proportion is a specific value, or whether a specified pattern exists in proportions in different categories. The second section is on the association between two dichotomous variables. A useful offshoot of this set-up is a relative risk (RR) and odds ratio (OR). A separate section is devoted to these concepts because of their importance in dealing with uncertainties in the health sciences. The methods for two or more polytomous variables are considered in the last section.

ONE QUALITATIVE VARIABLE

The proportion of obese among hypertensives, the proportion of healthy babies born with different birth weights (<2500 g, 2500–3499 g, ≥3500 g) and the proportion with different grades of head injury (mild, moderate, serious, critical) among those surviving, are examples of proportions based on one qualitative variable. The variables in these examples are obesity, birth weight and head injury, respectively. The variable is dichotomous (obese/non-obese) in the first, and polytomous (multiple categories) in the second and the third cases.

In the case of dichotomous categories, the *P* values under the null hypothesis can be easily obtained by what we call a binomial distribution. This can be approximated by Gaussian form when *n* is large and π is not too small. Gaussian approximation seems to work well when $n\pi \geq 8$, where π is the population proportion.

In the case of polytomous categories, the exact distribution applicable to calculate *P* value is called multinomial. This is complex and a computer is generally needed to compute these probabilities. However, in many situations, when *n* is large, this can be reasonably approximated by Chi-square. In view of the importance of Chi-square in dealing with uncertainties in health sciences, it is explained in detail below.

Polytomous categories (large n)—Goodness of fit χ^2-test

The method is applied to a situation where the variable has several categories. Let the interest be in finding whether the subjects in the target population do or do not follow a pre-specified pattern. Such a problem is known as the problem of 'goodness of fit' because the interest is in finding whether the pattern observed in the sample does or does not fit well into the specified pattern.

Example 1a. **A random sample of 150 patients with acquired immunodeficiency syndrome (AIDS) are investigated to examine the possibility of a preponderance of a particular blood group in AIDS cases. If there is no preponderance, the profile would be the same as in the general population. Suppose this is 6:5:8:1 for blood groups O, A, B and AB, respectively. The sample observations are as follows:**

Blood group	O	A	B	AB	Total
AIDS patients	57	36	51	6	150

What is the chance that this sample has indeed arisen from the population with a blood group pattern in the ratio 6:5:8:1?

Denote the population proportion in the four blood groups by π_1, π_2, π_3 and π_4, respectively. The null hypothesis in this case is

$$H_0: \pi_1=0.30,\ \pi_2=0.25,\ \pi_3=0.40 \text{ and } \pi_4=0.05 \qquad (1)$$

This comes from the blood group ratios in the population. The interest is in testing whether the sample provides enough evidence against this H_0. The alternative hypothesis H_1 is that the pattern is any other than (1).

Before going any further with this example, we need to explain the following.

Chi-square and its explanation. Denote the observed frequencies in the four groups in the sample by O_1, O_2, O_3, and O_4, respectively. That is, O_1=57, O_2=36, O_3=51 and O_4=6. If H_0 is really true then the frequencies expected, denoted by *Es*, would be in the ratio specified in (1), i.e. $E_k = n\pi_k$ (k=1, 2, 3, 4). For n=150, we get E_1=150×0.3=45. Similarly, E_2=37.5, E_3=60 and E_4=7.5. A large difference between *Os* and *Es* would suggest that the observed pattern is different than that stipulated in (1). It would serve as evidence against H_0 and in favour of H_1. Thus, the examination of the differences $(O_k - E_k)$ for different k could be helpful. Since the total of the expected frequencies has to be the same (n=150) as that of the observed frequencies, it is imperative that some of these differences would be negative and some positive. The sum $\Sigma(O_k - E_k)$ would be zero. As in the case of deviations $(x_i - \bar{x})$ for calculating standard deviation (SD), square of these differences helps to get rid of the negative sign. This gives $(O_k - E_k)^2$. The magnitude of these squares is the key to the plausibility of H_0. However, a difference of 1.5 over the expected 7.5 in blood group AB has a different meaning than the same difference over the expected 37.5 in blood group A. The former difference is one-fifth of the corresponding expected frequency while the latter is not even one-twentieth. Thus, the squared differences should be viewed in relation to the expected frequencies. The quantity $[(O_k - E_k)^2/E_k)]$ becomes relatively free of the differentials existing in the expected frequencies in different groups and helps to give nearly equal weight to the groups. Instead of taking the average of these quantities, this time obtain the sum $\Sigma[(O_k - E_k)^2/E_k]$. This criterion is based entirely on frequencies and is thus unit free. This obviates the need to take the square root as is done at the time of calculating SD. As a reminder that the quantity is a square, the sum is called Chi-square (χ^2).

$$\text{Chi-square: } \chi^2 = \Sigma \frac{(O_k - E_k)^2}{E_k} \quad (k=1, 2 \ldots, K) \qquad (2)$$

where K is the total number of cells in the contingency table. In Example 1a, K=4. Note that *Es* are obtained assuming that H_0 is true. Thus the value of χ^2 in (2) is under H_0. When H_0 is true, the difference between O_k and E_k, $(O_k - E_k)$, should be small and thus the value of χ^2 also should be small. In other words, a large value of χ^2 is unlikely if H_0 is true. If the sample gives a large χ^2, it is evidence against H_0.

Now, we need to find the P value. This is the probability of occurrence of the value of the criterion as much or more extreme than that obtained for the sample data. For this, the distribution of the criterion under H_0 is needed. This distribution of χ^2 is known and the critical values are tabulated in standard probability tables for specific values of the level of significance α. However, this requires that four-fifths of the expected frequencies are at least five. The shape of the distribution varies according to the degrees of freedom (df) which in turn depends mostly on the number of cells K. Different distribution of χ^2 for different df is analogous to the different distribution of diastolic blood pressure (DBP) in different age-groups. The df can be explained as follows:

Degrees of freedom (df). In Example 1a, there are four categories of blood group, namely, O, A, B, and AB. However, the frequency in only three of them can be freely chosen, the fourth is automatically determined by the total. If the frequencies chosen for O, A and AB are 70, 20 and 10 then the frequency in B group has to be 50 since the total is 150. If the frequencies chosen for O, A and B are 60, 30 and 20 then the frequency in the AB group has to be 40. Thus, there is a freedom to choose only 3 out of 4 cells. This is called the degrees of freedom (df). For K cells in a one-way contingency table, df=K–1 when the sample values have no restriction other than the fixed total.

Example 1b. **From the data in Example 1a, we get the following:**

	Blood group				
	O	A	B	AB	Total
Observed frequency (O_k)	57	36	51	6	150
Expected frequency under H_0 (E_k)	45	37.5	60	7.5	150
$O_k - E_k$	12.0	−1.5	−9.0	−1.5	0
$(O_k - E_k)^2/E_k$	3.20	0.06	1.35	0.30	4.91

Thus, χ^2=4.91. If a computer software is used, it will automatically compare the calculated value of χ^2 with its known distribution and tell us that P=0.178. Otherwise, from the

- Chi-square does not require the frequency pattern to be Gaussian or any other specific pattern. Thus, Chi-square is a distribution-free procedure.
- Chi-square works fine when the expected frequency in any cell is not less than 5. This is generally met when n is large and no cell probability is very small. Without $E_k \geq 5$ for each k=1, 2, . . . , K, the use of Chi-square is not considered valid. However, when the number of the categories K is large, not more than one-fifth of the categories (i.e. K/5 cells) should have E_k<5 and none should be less than 2.
- Chi-square is calculated on the actual frequencies in the cells and not the percentages.
- Chi-square is basically a two-tail test. Significance in this case implies only presence of some difference and it can seldom be labelled positive or negative. Thus, the alternative hypothesis H_1 is two-sided.
- The χ^2 criterion would be large even if one particular difference (O_kñE_k) is large. Thus, rejecting H_0 only tells us that there is at least one cell where the frequency is substantially different from the expected but fails to specify the cell where it is different. On the other hand, if a large difference is present in only one cell, then this can be masked by the small differences in the other cells. For a more focused inference, partitioning of the table is helpful.
- There is a growing feeling that the use of conventional probability cut-offs such as P<0.05, P<0.01, etc. should be discontinued and the exact P value be used instead. The user then is in a better position to decide how much significance should be to attached to the results.
- Inference from statistical tests is probabilistic rather than definitive. The chance of error is controlled to <5%. This certainly works in the long run but might fail in a particular case.

probability tables, the *P* value for 3 df and $\chi^2 = 4.91$ is more than 0.05. Thus the sample values do not provide sufficient evidence against H_0. The observed frequencies in different blood groups can well arise by chance when the sample is from a population with a pattern shown in (1). Thus, H_0 is plausible and cannot be rejected. Preponderance of any blood group in the cases of AIDS cannot be concluded on the basis of this sample.

Partitioning of table

The data in Example 1 reveal that the observed frequency in blood group O is much higher than that expected from the pattern in the general population. However, the other differences are not as large. To check that this really is so, first check that the pattern in blood groups A, B and AB is nearly the same as expected, and later check the difference in blood group O. The corresponding null hypotheses are

H_{0I}: A, B and AB are in the ratio 5:8:1 (same as before); and
H_{0II}: $\pi_1=0.3$, $\pi_2+\pi_3+\pi_4=0.7$.

The second ratio is also the same (6:14) as in the population. For these two null hypotheses, calculations for χ^2 are as follows:

I. Blood group	A	B	AB	Total
O_k	36	51	6	93
E_k	33.2	53.1	6.6	93
$(O_k-E_k)^2/E_k$	0.24	0.08	0.05	$0.37=\chi^2_{\text{I}}$
II. Blood group	O	Others		Total
O_k	57	93		150
E_k	45.0	105.0		150
$(O_k-E_k)^2/E_k$	3.20	1.37		$4.57=\chi^2_{\text{II}}$

The division of the earlier 4-cell table into two tables as shown above is called partitioning. The first partition gives $\chi^2_{\text{I}}=0.37$. This has 3–1=2 df. The *P* for $\chi^2=0.37$ (df=2) is 0.831. Since it is more than 0.05, H_{0I} cannot be rejected. The evidence is not sufficient to conclude that the pattern of A, B and AB in AIDS cases is not the same as in the general population. Part II has only two cells so that χ^2_{II} has only one df. The *P* in this case is 0.033. This is less than 0.05 and is statistically significant at the 5% level. It can be concluded that blood group O is more common in AIDS cases while nothing can be said about the other three groups.

The conclusion reached after partitioning is different from the one reached earlier when all the cells were considered together. This is because the lack of difference in A, B and AB groups masked the difference in the O group. Partitioning helped to uncover this.

PROPORTIONS IN 2×2 TABLES

We shall now discuss whether the proportion of subjects possessing a particular characteristic is the same in one group as compared to another. The set-up is essentially bivariate. There are two variables, both with two categories each. Such a set-up is also known as a 2×2 table (Table I). Three situations are possible. These are explained for the classical set-up of one variable being antecedent and the other outcome.

Prospective study

The column totals $O_{\cdot 1}$ and $O_{\cdot 2}$ for antecedents are fixed in advance. We can then denote them also by n_1 and n_2, respectively. These are the number of subjects with and without the antecedent and followed up to observe the outcome. The row totals $O_{1\cdot}$ and $O_{2\cdot}$ with and without outcome become known only after the investigation is over. The relevant null hypothesis in this case is

TABLE I. General structure of a 2×2 contingency table

Variable-2 (Outcome)	Variable-1 (Antecedent)		Total
	Present	Absent	
Present	O_{11} (π_{11})	O_{12} (π_{12})	$O_{1\cdot}$ ($\pi_{1\cdot}$)
Absent	O_{21} (π_{21})	O_{22} (π_{22})	$O_{2\cdot}$ ($\pi_{2\cdot}$)
Total	$O_{\cdot 1}$ ($\pi_{\cdot 1}$)	$O_{\cdot 2}$ ($\pi_{\cdot 2}$)	n

The corresponding probabilities are in parentheses

H_0: $\pi_{11}=\pi_{12}$. This states that the incidence rate in the two exposure groups is the same.

Retrospective study

The row totals $O_{1\cdot}$ and $O_{2\cdot}$ for the outcome are fixed in advance and the column totals $O_{\cdot 1}$ and $O_{\cdot 2}$ with and without antecedent are obtained through the study. The fixed row totals can also be denoted by n_1 and n_2. The null hypothesis now is that the rate of presence of antecedent in those with positive outcome is the same as in those with negative outcome, i.e. H_0: $\pi_{11}=\pi_{21}$.

Cross-sectional study

Neither the column totals nor the row totals are fixed in advance and both become known only after the study is over. A sample of *n* subjects is studied and the presence or absence of antecedent and outcome both are simultaneously observed. In this case, $\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22}=1$. As per the law of multiplication of probabilities, the antecedent and outcome are independent only if H_0:$\pi_{rc}=\pi_{r\cdot}*\pi_{\cdot c}$ $(r,c=1,2)$ holds, where $\pi_{r\cdot}=\pi_{r1}+\pi_{r2}$ and $\pi_{\cdot c}=\pi_{1c}+\pi_{2c}$

The H_0 in the first two cases is called *hypothesis of homogeneity* (column homogeneity and row homogeneity, respectively) and the H_0 in the third case is called the *hypothesis of independence*. Such a distinction is required for proper interpretation of results but the method of calculation is the same in all the three situations. We describe the methods for independent samples and matched pairs separately.

Tests for proportions in independent samples

Chi-square test. If the sample size (n) is large, under any of the above three H_0, the test criterion analogous to (2) is given by

$$\text{Chi-square: } \chi^2=\Sigma_{rc}\frac{(O_{rc}-E_{rc})^2}{E_{rc}} \quad (r,c=1,2), \text{ and df=1} \qquad (3)$$

The justification is the same as for (2) and the applicability also requires each expected cell frequency to be at least 5. In a 2×2 table, df=1. There is a freedom to arbitrarily choose frequency in only one cell. Others are automatically decided because the row and the column totals are considered fixed. The test procedure is to calculate χ^2, and find the probability (*P* value) of obtaining this or a higher value. A small value of *P*, as before, is evidence against H_0. If *P* value is sufficiently small, say less than 0.05, then reject H_0, otherwise not.

We would like to point out two things here: first, the popular Yate's correction is advisable in a 2×2 table, particularly if *n* is not very large. This helps in reaching a better approximation of a discrete distribution by a continuous distribution, though this often makes the test very conservative. Second, when frequencies are small, use Fisher's exact test. Both these can be easily taken care of by standard statistical packages.

Tests for proportions in matched pairs

The procedure given above is valid only when the two groups of

TABLE II. Matched pairs with dichotomous antecedent and dichotomous outcome: Prospective study

Partner 2 Antecedent present (Exposed or experiment)	Partner 1 Antecedent not present (Not exposed or control)		Total
	Positive outcome (disease +)	Negative outcome (disease –)	
Positive outcome (disease +)	a	b	$a+b$
Negative outcome (disease –)	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

subjects are independent. Independence is lost when there is some kind of matching or pairing. In such circumstances, simple Chi-square or Fisher's exact test are replaced by McNemar's and exact test for large and small n, respectively. Table II depicts the design of a study where the pairs are matched in a 2×2 set-up. In this table, for example, b is the number of pairs in which the exposed partner develops the disease and the non-exposed partner does not.

McNemar's test for large n. A popular criterion in case of matched pairs is as follows:

$$\text{McNemar's test: } \chi^2_M = \frac{(|b-c|-1)^2}{b+c} \qquad (4)$$

where b and c are as in Table II.

This continues to be referred to Chi-square distribution for obtaining the P value. The restriction of no cell frequency less than 5 applies. Subtraction of 1 in the numerator of (4) represents continuity correction similar to the Yate's correction. Note that the concordant pairs a and d do not contribute at all to the decision. It is solely based on the number of disconcordant pairs of the two types.

Exact test for matched pairs with small n. McNemar's test also ceases to follow Chi-square distribution when n is small. The test is then done with the help of a binomial distribution. The P value under H_0 is

$$P = \sum_{x=b}^{(b+c)} {}^{(b+c)}C_x (1/2)^{(b+c)} \qquad (5)$$

This is the probability of obtaining b or more discordant pairs when H_0 of no association ($\pi=1/2$) is true. The H_1 in this case is that positive outcome is more frequent in those with antecedent present. Thus, b should be large if H_1 is true and c should be small. The probability (5) is of the configurations as much or more extreme favouring H_1 when actually there is no association. If this is small, say less than 0.05, reject H_0 and conclude that an association is present.

Example 2. **To evaluate the role of a therapy in relieving common cold within a week, 15 cases were given therapy and another 15 cases served as controls. The experimental and the control cases were one-to-one matched for age, gender and body mass index so that these do not act as confounders. The results obtained were as follows:**

With therapy (Experimental group)	Without therapy (Control group)		Total
	Relieved in 1 week	Not relieved in 1 week	
Relieved in 1 week	5	3	8
Not relieved in 1 week	1	6	7
Total	6	9	15

In this example, $b=3$ and $c=1$. The probability of a Type I error in this case is the chance that the therapy is effective in at least 3 pairs in the sample when it actually is not effective in the target population. This is

$$P = {}^4C_3(1/2)^4 + {}^4C_4(1/2)^4 = 0.31$$

Since this is high, the observed configuration could well have arisen due to sampling fluctuation even when there is no actual association. Thus the null hypothesis of no association cannot be rejected. It cannot be concluded that the therapy is effective in relieving common cold within a week.

RELATIVE AND ATTRIBUTABLE RISKS, AND ODDS RATIO (large *n*)

A useful application of the comparison of two proportions is studying the RR and attributable risks (AR), and the odds ratio (OR). These terms are differentially used in the context of prospective and retrospective studies, respectively. In case there are more than two groups, the comparison could be made between two groups at a time. However, several RR, AR and OR will have to be computed. The concepts of risk and odds are mostly relevant to a large n only.

Relative risk (RR)

Relative risk in independent samples. Relative risk is the ratio of the risk (incidence) of developing an outcome such as disease (D) in those with antecedent factor (A) compared to those without this factor. This obviously requires a prospective study according to the structure in Table III. In terms of probabilities,

$$\text{RR} = \frac{P(\text{D+/A+})}{P(\text{D+/A–})} = \pi_{11}/\pi_{12}$$

where '+' is for presence and '–' is for absence. Note the condition that $\pi_{11}+\pi_{21}=1$ and $\pi_{12}+\pi_{22}=1$ as statedearlier. This is estimated by

$$\hat{\text{RR}} = \frac{O_{11}/n_1}{O_{12}/n_2}$$

where hat (^) sign is for the estimate. If any O_{rc} $(r,c=1,2)$ is zero then a modified estimate of RR is

$$\hat{\text{RR}}_{\text{mod}} = \frac{(O_{11}+0.5)/n_1}{(O_{12}+0.5)/n_2}$$

Confidence interval. It can be shown that lnRR (natural logarithm of RR) has a Gaussian distribution for large n. Thus, lnRR is used for inferences when n is large. It is known that for a large sample that an approximate estimate of its standard error (SE) is the following:

$$\hat{\text{SE}}(\ln\hat{\text{RR}}) \approx \sqrt{\frac{1}{O_{11}} - \frac{1}{n_1} + \frac{1}{O_{21}} - \frac{1}{n_2}} \qquad (6)$$

TABLE III. Structure of a prospective study

Outcome (D)	Antecedent(A)		Total
	Present	Absent	
Present	O_{11} (π_{11})	O_{12} (π_{12})	$O_{1\cdot}$ ($\pi_{1\cdot}$)
Absent	O_{21} (π_{21})	O_{22} (π_{22})	$O_{2\cdot}$ ($\pi_{2\cdot}$)
Total	n_1 (1)	n_2 (1)	

The corresponding probabilities are in parentheses

The 95% CI for RR can be obtained as $exp\,[\ln\hat{RR} \pm 2\hat{SE}(\ln\hat{RR})]$ following the procedure outlined in our previous article.[1]

Example 3. **Martinez *et al.*[2] reported a prospective study on wheezing lower respiratory tract illness (LRI) during the first year of life of 500 boys and an almost equal number of girls. These were enrolled at birth in Tucson, Arizona, of the USA between 1980 and 1984. Among the objectives was to explore maternal age as a risk factor for wheezing LRI. The data for the boys are given below.**

Maternal age (years)	Lower respiratory tract illness		Total
	Yes	No	
<26	48	117	165
≥26	65	270	335

The estimated RR of wheezing LRI in boys with mother's age <26 years compared to boys with mother's age ≥26 years is

$$RR = \frac{48/165}{65/335} = 1.50$$

Thus, boys born to younger women were 1.5 times more likely to get LRI during their first year of life compared to boys born to older women. Also, ln RR = ln(1.50) = 0.4055, and

$$\hat{SE}(\ln\hat{RR}) = \sqrt{\frac{1}{48} - \frac{1}{165} + \frac{1}{65} - \frac{1}{335}} = 0.1648$$

Therefore, 95% CI for RR is $exp(0.4055 \pm 2 \times 0.1648)$, or $(e^{0.076}, e^{0.735})$, or (1.08, 2.09). With a 'relaxed' meaning of CI, it can be stated with 95% confidence that this interval contains the true RR in the population.

Test of hypothesis. Since the above CI does not include RR=1, the H_0:RR=1 can be safely rejected. It can be concluded that the risk of wheezing LRI in the first year of life was indeed higher in boys born to women of younger age. Thus, one procedure to test H_0 is to calculate

$$Z = \frac{\ln\hat{RR}}{\hat{SE}(\ln\hat{RR})} \qquad (7)$$

and reject it if the corresponding P value is <0.05.

The second procedure to test H_0: RR=1 against H_1: RR≠1 is by Chi-square as explained for 2×2 tables. In this case, χ^2=5.93. The Chi-square value at one df is 3.84 for P=0.05. Since the calculated value is higher, P<0.05. The chance that RR=1 in the target population will give the observed numbers or more in favour of H_1 is exceedingly small. Thus, the null hypothesis is rejected. Conclude that RR≠1. In case there is *a priori* reason to believe that RR would be more than one, then H_1:RR>1. It may then be prudent to use (7) and refer it to the Gaussian distribution table to get the one-sided P value. In this example, Z=2.46. From probability tables, $P(Z \geq 2.46)$=0.007. This P value is not only less than 0.05 but also less than 0.01. There is practically no chance of H_0 being true. The sample provides strong evidence against it. The RR can be said to be highly significantly different from RR=1.

RR in case of matched pairs. A general situation of matched pairs with regard to antecedent and outcome in a prospective study is shown in Table II. In this case, RR is estimated as follows:

$$\text{Relative risk (matched pairs): } \hat{RR}_M = \frac{a+b}{a+c}$$

CI and test of H_0 for RR in case of matched pairs can be done by the method given later for OR.

Attributable risk (AR). Attributable risk is the difference in the risk among the exposed and non-exposed subjects. This is directly estimated as $(O_{11}/n_1 - O_{12}/n_2)$ in case of independent samples. In case of matched pairs,

$$\hat{AR}_M = \frac{b-c}{n}$$

AR measures the expected reduction in risk if the exposure factor is eliminated. Thus, this is of public health importance.

Population attributable risk. The population attributable risk (PAR) estimates the excess rate of disease attributable to the exposure in the total population under study and is calculated as the rate (incidence) of disease in the population including exposed minus the rate in the non-exposed group. This is different from AR since the population comprises both the exposed and non-exposed groups of people.

Both AR and PAR are sometimes calculated as percentage of the risk in the exposed group. In that case, these can be understood as attributable fractions. The PAR fraction can be directly obtained from RR when the proportion of persons with the given risk factor is known.

$$PAR = \frac{p(RR-1)}{p(RR-1)+1}$$

where p is the proportion of persons having the given risk factor.

Odds ratio (OR)

In betting it is often said, that the odd of winning is 1:3. This means that a loss is 3 times more likely than a win. Similarly, in retrospective or case–control studies, an odd is the frequency of presence of the antecedent relative to its absence. It is calculated

- RR=1 implies independence, i.e. the risk in the exposed subjects is the same as in the non-exposed subjects. RR>1 means a higher risk in the exposed and RR<1 means a lower risk. Lower than one RR can be interpreted as a protective effect in place of risk.
- It is preferable to keep the adverse category of antecedent in the first column of the 2x2 contingency table and the adverse outcome in the first row. The interpretation is then easy.
- The term 'risk' literally relates to an adverse outcome. However, statistically, this is not necessarily so.

TABLE IV. Structure of a case–control study

Outcome	Antecedent		Total
	Present	Absent	
Present (cases)	O_{11} (π_{11})	O_{12} (π_{12})	n_1 (1)
Absent (controls)	O_{21} (π_{21})	O_{22} (π_{22})	n_2 (1)
Total	$O_{\cdot 1}$ ($\pi_{\cdot 1}$)	$O_{\cdot 2}$ ($\pi_{\cdot 2}$)	n

The corresponding probabilities are in parentheses

for the cases and controls. The ratio of these two odds is called the OR.

Two independent samples. In the case of case–control studies, Table I takes the form of Table IV. In this case, $\pi_{12}=1-\pi_{11}$ and $\pi_{22}=1-\pi_{21}$. The odds of the antecedent being present among cases is π_{11}/π_{12} and among controls is π_{21}/π_{22}. Thus, the odds ratio is

$$\text{OR} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\times\pi_{22}}{\pi_{12}\times\pi_{21}}$$

This is estimated as

$$\hat{\text{OR}} = \frac{O_{11}\,O_{22}}{O_{12}\,O_{21}}$$

If any of the cell frequencies is zero, then the modified estimate of OR is calculated as:

$$\hat{\text{OR}}_{\text{mod}} = \frac{(O_{11}+0.5)\,(O_{22}+0.5)}{(O_{12}+0.5)\,(O_{21}+0.5)}$$

***Example* 4. Consider the following data on parity status of 210 anaemic and 140 non-anaemic women.**

Anaemia	Parity ≤ 2	Parity ≥ 3	Total
Present	98	112	210
Absent	92	48	140

$$\text{OR} = \frac{98\times 48}{92\times 112} = 0.46$$

The likelihood of parity ≤ 2 among anaemics is less than one-half that among non-anaemics.

CI for OR. OR is a ratio and its natural logarithm (ln) becomes a linear function. The distribution of OR can be shown to be highly skewed but lnOR has nearly Gaussian pattern for large n. It has been established that for large n,

$$\hat{\text{SE}}\,(\ln\hat{\text{OR}}) = \sqrt{1/O_{11}+1/O_{12}+1/O_{21}+1/O_{22}} \qquad (8)$$

where O_{11}, O_{12}, O_{21} and O_{22} are as in Table IV. If any $O_{rc}=0$ then add 1/2 to each O_{rc} in the denominators in (8). For large n,

95% CI for OR: *exp*($\ln\hat{\text{OR}} \pm 2\,\hat{\text{SE}}\,(\ln\hat{\text{OR}})$),

where *exp* is exponent on Neparien base e.

Test of hypothesis on OR. In this case, the H_0 almost invariably is that OR=1. This states that the presence of the antecedent is as common in cases as in controls. A simple statement which takes care of both negative as well as positive directions of relationship is that there is no association under H_0 between the antecedent and the outcome. The alternative could be one-sided H_1:OR<1 or H_1:OR>1, or two-sided H_1:OR≠1. The latter is applicable when there is no *a priori* assurance that the relationship could be one-sided. The hypothesis is tested by the classical Chi-square procedure as given in (3).

TABLE V. Matched pairs in a case–control study with dichotomous antecedent

Cases (Partner 2)	Controls (Partner 1)	
	Antecedent present (exposed)	Antecedent not present (non-exposed)
Antecedent present (exposed)	*A*	*B*
Antecedent not present (non-exposed)	*C*	*D*

OR in matched pairs. Consider Table V on matched pairs. This is similar to Table II. We are now using capital *A*, *B*, *C*, *D* as the notation of cell frequencies in place of lower case *a*, *b*, *c*, *d* to distinguish the case–control set-up from the prospective set-up. Note also how the labelling of the cells has changed.

The total number of pairs is *A*+*B*+*C*+*D*. In this table, *A* is the number of pairs with both case and control subjects exposed, *D* is the number of pairs with both non-exposed. These two together are the concordant pairs. The OR is computed on the basis of the disconcordant pairs: *B* and *C*. In case of an association between exposure and disease, clearly *B* should be more than *C*.

Odds ratio (matched pairs): $\hat{\text{OR}}_{\text{M}} = \dfrac{B}{C}$

Confidence interval: For large *B* and *C*,

$$\hat{\text{SE}}(\ln\hat{\text{OR}}_{\text{M}}) = \sqrt{\frac{1}{B}+\frac{1}{C}} \qquad (9)$$

Thus, 95% CI for log of odds ratio is $\ln\left(\dfrac{B}{C}\right) \pm 2\sqrt{\dfrac{1}{B}+\dfrac{1}{C}}$

Take exponential of the limits and get $\dfrac{B}{C}\,e^{\pm 2\sqrt{\frac{1}{B}+\frac{1}{C}}}$

Test of hypothesis. The relevant null hypothesis in this case is H_0:$\text{OR}_{\text{M}}=1$. To test this against a one-sided alternative H_1:$\text{OR}_{\text{M}}>1$, calculate

$$Z = \frac{B-C}{\sqrt{B+C}} \qquad (10a)$$

For large n, refer it to the usual Gaussian distribution to find whether the *P* value is sufficiently small or not. For a two-tailed test, it may be easier to calculate McNemar's

$$\chi^2 = \frac{(|B-C|-1)^2}{B-C} \qquad (10b)$$

and refer it to Chi-square at one df.

ANALYSIS OF *R*×*C* TABLES (large *n*)

So far, we have discussed analysis of tables with two rows (R=2) and two columns (C=2). This means both the characteristics (or variables) are dichotomous. However, there are a large number of variables that are not dichotomous. Concern now is with a set-up where both the variables are polytomous and qualitative. It is also presumed that n is large and at least 80% cells have a minimum frequency of 5.

The method to find whether or not the two qualitative variables

- The interpretation of OR is similar to that of RR. It can be shown[3] that OR approximates RR fairly well when the outcome of interest is rare, say <5% in the target population. Most outcomes of medical interest are rare. If the outcome is not rare, OR may overestimate RR if it is more than one and underestimate it if less than one.
- The sample OR is always a good estimate of the population OR whether or not the disease is rare in the population.
- In caseñcontrol studies, it could be inappropriate to use the term incidence since the reference is to presence or absence of antecedent characteristic. Thus the term 'odd' is used.

are associated continues to be Chi-square in case of two polytomous variables. If n is large, (3) is used with R and C is now more than two. The df will be $(R-1)(C-1)$. This considers all categories nominal. When the categories are ordinal, the trend in proportions can be investigated by calculating Chi-square for trend. For this see Le.[3]

THREE-WAY TABLES

A three-way contingency table arises when the classification of the subjects is done with respect to three variables. This is a type of multivariate categorical data set-up. Besides row and column, the third dimension is called layer. The number of rows, columns and layers can be denoted by R, C and L, respectively. As in the case of two-way tables, the null hypothesis in a three-way table could be of homogeneity of different types or of independence, depending upon individual variables being factors or responses. The calculation of χ^2 is done as a formula similar to (3).

Under H_0, this now follows a Chi-square distribution with $(R-1)(C-1)(L-1)$ df. Chi-square, if significant, only indicates that an association is present somewhere. To find where exactly this association is, one approach is partitioning. The second and more versatile approach is to use log–linear models.

Log–linear models. The logarithm of expected frequencies in a contingency table can be expressed in terms of additive factors. Thus the name 'log–linear' for these models. These models are useful only when no variable is considered dependent on the other. The dependent variable in these models is the number of subjects in a cell of the contingency table. The objective is to find whether or not the variables categories, individually or jointly, are significantly contributing to determining the cell frequency. These models can be applied to both two-way, three-way and higher dimensional tables. Interested readers may consult Haberman[4] for a detailed discussion on the methodology of log–linear models.

Log–linear models can also be used to evaluate the net association between two variables after removing the effect of the others. Stellman *et al.*[5] used a series of log–linear models in a case–control study on cervical cancer and cigarette smoking. They controlled age and socio-economic status and concluded that net association between cervical cancer and smoking was not statistically significant at 5% level.

Three-way tables are also obtained as a collection of several two-way tables. Suppose an association between smoking and hypertension is investigated for six different professions. Then six contingency tables would be available. The professions can be viewed as strata. While separate Chi-square can be calculated for each profession, a more reliable conclusion can be drawn by combining the data, provided the differences across professions are not significant. This is done by using the Mantel–Haenszel method. For details, consult Le.[3]

TABLE VI. Statistical procedures for inference from proportions

Parameter of interest and set-up	Conditions	Main criterion	Equation number
Proportion			
One dichotomous variable	Independent trials		
	Any n	Binomial	Not given
	Large n ($n\pi \geq 8$)	Gaussian Z	Not given
One polytomous variable	Independent trials		
	Large n	Goodness of fit Chi-square	(2)
	Small n	Multinomial	Not given
Two dichotomous variables (2×2)	Two independent samples		
	Large n	Chi-square	(3)
	Small n	Fisher's exact	Not given
	Matched pairs		
	Large n	McNemar's	(4)
	Small n	Binomial	(5)
*Relative risk and odds ratio**			
Natural logarithm of relative risk	Two independent samples		
	CI	±2SE	Use (6)
	Test of H_0	Gaussian Z	(7)
	Matched pairs		
	CI	±2SE	Use (9)
	Test of H_0	Z or χ^2	(10a, 10b)
Odds ratio	Two independent samples		
	CI	±2SE	Use (8)
	Test of H_0	Chi-square	(3)
	Matched pairs		
	CI	±2SE	Use (9)
	Test of H_0	Z or χ^2	(10a, 10b)
Bigger tables—No matching			
Association	$R \times C$ tables	Chi-square	Similar to
	Test of H_0		(3)
	Three-way tables		
	Test of full independence	Chi-square	Similar to (3)
	Test of other types of independence (log–linear models)	G^2	Not given

* The case of small n is not discussed as the main condition is a large n
CI confidence interval SE standard error

Table VI summarizes the statistical procedures that can be used to derive inferences from proportions, under different conditions. These are mostly restricted to the simple cases that we have discussed. The list is by no means exhaustive. There are several methods that we have not touched upon. If proportion is to be estimated from a set of regressors, logistic regression is used. If proportion relates to survival and duration of survival is of interest, then survival analysis is used. A different set of methods is required for multiple response tables.

REFERENCES

1 Indrayan A, Gupta P. Sampling techniques, confidence intervals and sample size. *Natl Med J India* 2000;**13:**29–36.
2 Martinez FD, Wright AL, Holber CJ, Morgan WJ, Taussig LM. Maternal age as a risk factor for wheezing lower respiratory illness in the first year of life. *Am J Epidemiol* 1992;**136:**1258–68.
3 Le CT. *Applied categorical data analysis.* New York:John Wiley, 1998.
4 Haberman SJ. *Analysis of qualitative data. Vol 1.* New York:Academic Press, 1978.
5 Stellman SD, Austin H, Wynder EL. Cervix cancer and cigarette smoking: A case–control study. *Am J Epidemiol* 1980;**11:**383–8.