Biostats VHM 801 Course
Winter 2018, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment III

The solution is more detailed and verbose than required for a 100% mark, and includes additional remarks on outlier tests.

The study is observational, and its purpose is to compare the two groups of twins, monozygotic and dizygotic. Even though no treatments were imposed, the status of each twin pair as either monozygotic or dizygotic was established and known prior to the interviews that determined the state of alcoholism for each person. It is therefore most natural to think of the design as a comparison of two populations.

a. Probability of alcoholism of both twins

The population of interest here is the monozygotic twin pairs with at least one alcoholic twin in the pair. With the narrow criteria there are $n = 45 + 8 = 53$ such pairs. To estimate the probability p that both twins in such a pair are alcoholics we use a binomial model: $X \sim \mathrm{B}(n, p)$, where X is the number of pairs with both twin pairs alcoholics. We estimate by the sample proportion $\hat{p} = X/n$, and compute a 95% confidence interval by the normal approximation or the "plus four" method if the data don't contain at least 15 positives and 15 negatives. The results are shown in the table below for all criteria, with the preferred CI in bold. The "exact" (Clopper-Pearson) intervals are valid in every situation, but less transparent in their calculation.

				95% confidence interval		
Criteria	n	X	Estimate	normal approx.	plus four	"exact"
narrow	53	8	0.151	$(0.055, 0.247)$	$\mathbf{(0.077, 0.274)}$	$(0.067, 0.276)$
intermediate	80	15	0.188	$\mathbf{(0.102, 0.273)}$	$\mathbf{(0.116, 0.288)}$	$(0.109, 0.290)$
broad	147	45	0.306	$\mathbf{(0.232, 0.380)}$	$(0.237, 0.385)$	$(0.233, 0.387)$

The values in the table show how the probabilities of both twins in a pair being alcoholics increase when the criteria get broadened. This is what one would expect when more persons are being classified as alcoholics in general.

b. Probability of alcoholism with an alcoholic twin

The population representing a monozygotic twin classified as an alcoholic in the data consists of all the *persons* classified as alcoholics, not the twin pairs. For the narrow criteria, there are $m = 45+2\cdot 8 = 61$ such persons. We are looking for the proportion of these where the other person in the pair ("you") would also be classified as an alcoholic. Among the $m = 61$ (other) persons, 45 of them would not get this designation and $Y = 16$ would get it, giving a proportion of $Y/m = 16/61 = 0.262$. This probability is known as the probandwise concordance rate, and it is often used in twin studies (for further references and discussion, see e.g. Lee et al., 2003, *European Journal of Epidemiology* **18**, 1047–1050). The main difference to the probability calculated in **a)** is the shift from pairs to individuals.

For the intermediate and broad criteria we similarly get,

$$\begin{aligned} \text{intermediate:} \quad & m = 95,\ Y = 30 \quad \text{probandwise concord. rate} = 30/95 = 0.316, \\ \text{broad:} \quad & m = 192,\ Y = 90 \quad \text{probandwise concord. rate} = 90/192 = 0.469. \end{aligned}$$

These statistics also seem to increase when the condition becomes more common. Moreover, for all three criteria the value is larger than the one computed in **a)**.

As an added note, let us see how the probability can be computed as a conditional probability. For use of the formula: $\mathrm{P}(B|A) = P(A \text{ and } B)/P(A)$, the events A and B need to be defined appropriately. We let $A =$ the event that a co-twin is classified with alcoholism; among the 1180 women, there are 61 alcoholics (for the narrow criteria) and hence 61 alcoholic co-twins, so $P(A) = 61/1180$. Next we let $B =$ the event that a women is classified as an alcoholic, whereby (A and B) is the event that both the woman herself and her twin are alcoholics; this happened for 16 out of the 1180 women, therefore $P(A \text{ and } B) = 16/1180$. Using the formula, the desired probability becomes $16/61$, as before.

3. Comparison of monozygotic and dizygotic twins

The data for the narrow criteria constitute a 3×2 table with the number of alcoholic twins in a pair $\sim$ rows as a response variable, and the genetic type of the pair $\sim$ columns as an explanatory variable (as discussed above in the introduction). Therefore, the model is two independent multinomials, one for each column. In detail, for the counts in the first column we assume $(N_{11}, N_{21}, N_{31}) \sim$ Multinomial distribution $(n_1, (p_{11}, p_{21}, p_{31}))$, and for the second column we assume $(N_{12}, N_{22}, N_{32}) \sim$ Multinomial distribution $(n_2, (p_{12}, p_{22}, p_{32}))$. The column totals are $n_1 = 590$ and $n_2 = 440$. The model corresponds to assuming a multinomial setting for each type of twin pairs. In textbook terminology this is a model for comparing independent populations (model I).

We estimate the p_{ij}'s by the respective sample proportions N_{ij}/n_j. Our interest is in the null hypothesis of same distribution in the two genetic groups, that is,

$$\mathrm{H}_0 : p_{11} = p_{12}, \text{ and } p_{21} = p_{22}, \text{ and } p_{31} = p_{32}.$$

The alternative hypothesis is that some differences between these probabilities exist. We test the null hypothesis by the (Pearson chi-square) X^2-test; the table below gives expected values under the null hypothesis, as well as the X^2-statistic and P-value. The X^2-statistic has $(3-1)\cdot(2-1) = 2$ degrees of freedom, so the critical value is $\chi^2_{.95}(2) = 5.99$. For the intermediate and broad criteria, the statistical model and analysis are entirely similar, and the results are shown in the same table below. The guidelines for use of the X^2-statistic are met; in the narrow criteria data, there is one expected value just above 5 but even if it had been less than 5, we would still have had 80% of cells >5.

Count	Narrow criterion		Intermediate criteria		Broad criteria	
(Exp. value)	Monozygotic	Dizygotic	Monozygotic	Dizygotic	Monozygotic	Dizygotic
neither	537 (524)	377 (390)	510 (499)	361 (372)	443 (426)	301 (318)
one	45 (60)	59 (44)	65 (76)	68 (57)	102 (123)	113 (92)
both	8 (6.9)	4 (5.1)	15 (15)	11 (11)	45 (41)	26 (30)
X^2 test (P)	9.59 (0.008)		4.42 (0.110)		11.14 (0.004)	

Two of the test statistics, for the narrow and broad criteria, are clearly significant with $P < 0.01$; there is evidence against the same distribution of the number of alcoholics in a twin pair for monozygotic and dizygotic twins. The data for the intermediate criteria fails to achieve significance. The pattern in the deviations between observed and expected values is the same for all criteria: too few observed monozygotic pairs with one twin only classified as alcoholic, and too many observed dizygotic pairs in this middle category. The lower (higher) observed in the middle category is then counterbalanced by too high (low) observed counts in the "neither" category and (mostly) the "both" category. Thus,

in a monozygotic pair the twins tend to more similar (either both non-alcoholic or both alcoholic) than in a dizygotic pair; the similarity refers not only to the event that both twins are alcoholics but also to both of them being non-alcoholics. This would seem in agreement with a possible genetic factor behind alcoholism. The middle category becomes the one with the largest discrepancies, and by far the largest X^2-contributions, because it needs to compensate for the two other categories. By the observational nature of the study, however, there is ample room for confounding; the obvious candidate is a shared environment which may not have been equally common among the two types of twin pairs.

4. Assessing (in)dependence between the two twins in a pair

The questions take us through the computation of a chi-square test of a different type than those described in the textbook. All calculations are done under the hypothesis (H_0) of independence between the two siblings in each pair.

Estimate probability p of alcoholism. Among the 880 persons in the 440 dizygotic twin pairs, a total of $113+2\cdot 26 = 165$ were in a state of alcoholism (by the broad criteria), therefore $\hat{p} = 165/880 = 0.1875$.

Estimate probabilities of 0, 1, and 2 twins in a state of alcoholism. Under the assumption of independence, the probability of both twins being in a state of alcoholism is $p \cdot p = p^2$, the probability of none them being in that state is $(1-p)\cdot(1-p) = (1-p)^2$, and the probability of exactly one of them in that state is $2p(1-p)$. Effectively, these are probabilities in a binomial distribution $(2, p)$. The estimated probabilities are given in the table below.

Expected number of twin pairs under H_0. The total number of twin pairs, $n = 440$, is multiplied onto the estimated probabilities from the previous step.

Computation of chi-square statistic. The calculations are summarized in the table below.

Twin pair	Observed	Estimated prob.	Expected value	Chi-square contrib
neither	301	0.66016	290.47	0.38
one	113	0.30469	134.06	3.31
both	26	0.03516	15.47	7.17
Total	440	1.00001	440.00	$X^2 = 10.86$

The table gives a value of $X^2 = 10.86$, which is highly significant in a $\chi^2(1)$-distribution. Statistical tables give $\chi^2_{.999}(1) = 10.83$, therefore $P < 0.001$. We conclude that there is overwhelming (strong) evidence against the null hypothesis of independence between alcoholism in the two twins within a pair. The data show too few pairs with just one twin in a state of alcoholism, and too many pairs where the twins are in the same state. Thus, the two twins are more similar (with respect to alcoholism) than would be expected if they were independent. The same findings exist in the data for monozygotic pairs and different alcoholism criteria (not shown). Whether this can be attributed to the genetic similiarity between the twins, would be subject to the same confounding factors as in the previous question.