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

Solution to home assignment III

The data were used by the famous statistician Sir Ronald A. Fisher (considered as one of the founders of modern statistics) in discussions about the evidence for smoking to cause lung cancer. In the late 1950s, the British medical authorities leaned towards accepting the hypothesis of a causal relationship between smoking and cancer. Fisher, himself a passionate pipe smoker, opposed this conclusion vehemently and argued publicly that further evidence was needed before such a conclusion could be drawn. The core of the issue was (and still is) that inferring causation from observational studies is difficult, if not impossible. Fisher was however fighting a lost cause, and although he apparently never conceded defeat on this issue, nowadays the detrimental health effects caused by smoking are no longer questioned. The media page for VHM 801 links to a collection of writings by Fisher on this issue, including a letter to *Nature* in which these data are shown and discussed.

This solution is, for pedagogical reasons, substantially more detailed than expected for a 100% mark.

1. Association between likeness of smoking habits and type of twin pair

The data constitute a 2×3 table of counts of twin pairs in the categories corresponding to the six combinations of the two variables (likeness of smoking habits (*smoking*), type of twin pair (*zygosity*)). From the description of how the counts were obtained it is reasonable to assume both variables to be response variables. The temporal relation between the two variables is such that only zygosity could affect smoking, not the other way around. As the question further asks for an assessment of association (or dependence) between the two variables, it seems most natural to assume a model for a single population ("Model 2"). Specifically, the counts (N_{ij}) are assumed to follow a multinomial distribution $(N, (p_{ij}))$, where $N = 71$ is total number of twin pairs. The null hypothesis of independence can be expressed as,

$$\mathrm{H}_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j},$$

where $p_{i\cdot}$ and $p_{\cdot j}$ are the marginal probabilities for rows (smoking) and columns (zygosity), respectively, and the alternative hypothesis H_a is two-sided, expressing that H_0 does not hold. The calculations for the Pearson X^2 statistic are summarized in the table below.

count[#]	Zygosity		
Smoking	dizygotic	monozyg., separ.	monozyg., joint
same	9 (13.4/1.46)	21 (19.4/0.13)	23 (20.2/0.40)
different	9 (4.56/4.31)	5 (6.59/0.38)	4 (6.85/1.18)

[#] (expected count/X^2-contribution)

The Pearson test statistic is: $X^2 = 7.88$, $\mathrm{df} = 2$, $P = \mathrm{P}(\chi^2(2) > 7.88) = 0.019$ (all values obtained from Minitab). Because only one out of six ($\sim 17\%$) of the expected values is < 5 (and not < 1), the conditions for use of the test are met. The result gives evidence to reject the hypothesis of independence, i.e. there is (moderate) evidence of an association (or dependence) between smoking and zygosity. The largest contribution to the test value comes from the cell (different, dizygotic), where the observed count under H_0 is larger than the expected count. That is, the data show too many dizygotic twin pairs with different smoking habits (and too few dizygotic twin pairs with same smoking habits) when compared to the distribution among the monozygotic twin pairs. This difference can also be seen directly in the data table.

2. Smoking among monozygotic and dizygotic pairs

The focus in this question is on the distribution on the two smoking categories for the two types of twin pairs. We are therefore considering only smoking as a response variable; formally this analysis can be thought of as conditional on twin pair type from the multinomial model in **1)**. For each of the twin pair types we will consider the data as corresponding to a binomial setting, say $X \sim \mathrm{B}(n, p)$, where X is the count of twin pairs with the same smoking habits, and p is the probability of the twins having the same smoking habits. The monozygotic and dizygotic twin pairs constitute independent samples. Estimates and confidence intervals are listed in the table below. For reference purposes all three types of confidence intervals are included, although only one interval is needed. The choice between intervals is discussed below.

	count	total	estimate	95% confidence interval		
Twin pair type	X	n	$\hat{p} = X/n$	"classical"	"plus four"	"exact binomial"
dizygotic	9	18	0.500	$(0.269, 0.731)$	$(0.291, 0.709)$	$(0.260, 0.740)$
monozygotic	44	53	0.830	$(0.729, 0.931)$	$(0.705, 0.909)$	$(0.702, 0.919)$

The values in the table have been computed using software but the classical and plus four intervals could also be obtained from the formulae, using $z^* = 1.96$:

$$\begin{aligned} \text{classical} &: \hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}, \\ \text{plus four} &: \tilde{p} \pm z^* \sqrt{\tilde{p}(1-\tilde{p})/(n+4)}, \quad \tilde{p} = (X+2)/(n+4). \end{aligned}$$

Neither of the twin pair groups meet the condition for the classical CI, that both the number of cases and non-cases exceed 15. Therefore it seems most natural to use the plus four interval (for which the conditions are met) in both groups. The "exact binomial" intervals are valid as well, but are seen to be the widest, as expected from its generally too large coverage (i.e., above 95%). The first impression is that the probability of same smoking habits is higher in monozygotic pairs, i.e. the pairs with the stronger genetic similarity. We continue our analysis by testing the hypothesis of equal probabilities in the two groups against the one-sided alternative (as specified in the question),

$$\mathrm{H}_0 : p_m = p_d \quad \text{versus} \quad \mathrm{H}_a : p_m > p_d.$$

Our options for the test is the classical, normal approximation z-test, or the equivalent X^2-test (though with a two-sided alternative), and Fisher's exact test. Because the conditions for use of the z-test and X^2-test are the same and are probably easiest assessed in the two-way table, we lay out our calculations in that format.

count (expected)	Zygosity	
Smoking	dizygotic	monozygotic
same	9 (13.44)	44 (39.56)
different	9 (4.56)	9 (13.44)

The test statistics are: $X^2 = 7.74 = 2.78^2 = Z^2$, and $P = \mathrm{P}(Z > 2.78) = 0.003$. Minitab gives $P = 0.005$ for the X^2-test, but as it has a two-sided alternative, the P-value for the one-sided H_a is half of that. The tests do however not meet their condition for use because one of the four expected counts is (just) below 5. In the two-way table setup, Minitab gives $P = 0.010$ for Fisher's exact test, also against a two-sided alternative. We would need another way of getting the P-value for the one-sided alternative, but we can at least conclude that $P \leq 0.010$. It turns out that the "2 Proportions" menu gives the one-sided $P = 0.008$ for Fisher's exact test. In any case, the test(s)

give evidence against H_0, and in support of H_a. We conclude that monozygotic twins have larger probability of having the same smoking habits than dizygotic twins. This could possibly indicate a genetic component to smoking habits.

3. Environmental effect on smoking habits

As the preceding analysis showed stronger similarity of smoking habits in monozygotic than dizygotic twins, it is of further interest to explore whether this could be attributable to a shared environment (i.e. upbringing). We only have such data for the monozygotic twin pairs. With our focus on the probability of twin pairs having the same smoking habits, the statistical model also here consists of two independent binomial distributions. A potential environmental effect could be described by the difference in the probabilities between twin pairs with joint and separate upbringing, i.e. $D = p_j - p_s$. The effect would manifest itself by greater likeness in smoking habits among twin pairs with joint upbringing, i.e. $D > 0$. We use a similar tabular layout for the calculations as above, again with multiple confidence intervals (when only one is needed) and the choice between them discussed below the table.

	count	total	estimate	95% confidence interval for D	
Upbringing	X	n	$\hat{p} = X/n$	"classical"	"plus four"
separated	23	27	0.852	$(-0.246, 0.158)$	$(-0.247, 0.163)$
joint	21	26	0.808		

The values in the table have been computed using software but could also be obtained from the formulae, using $z^* = 1.96$:

$$\text{classical} \quad : \quad \hat{p}_j - \hat{p}_s \pm z^* \sqrt{\hat{p}_j(1-\hat{p}_j)/n_j + \hat{p}_s(1-\hat{p}_s)/n_s},$$

$$\text{plus four} \quad : \quad \tilde{p}_j - \tilde{p}_s \pm z^* \sqrt{\tilde{p}_j(1-\tilde{p}_j)/(n_j+2) + \tilde{p}_s(1-\tilde{p}_s)/(n_s+2)}, \quad \tilde{p} = (X+1)/(n+2).$$

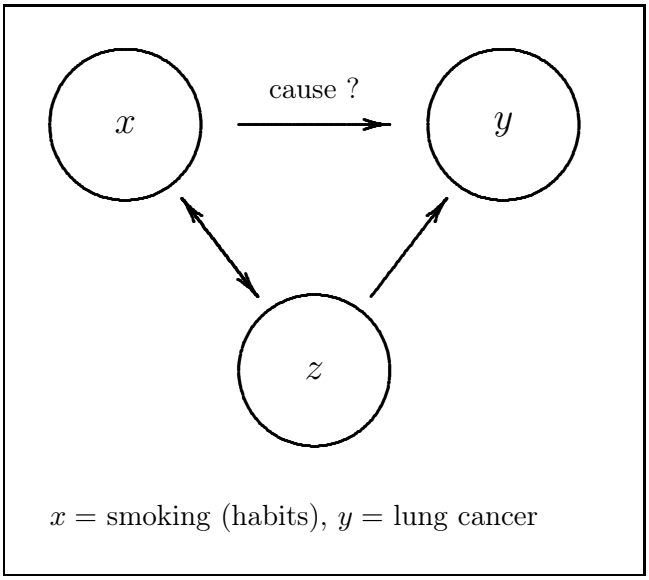
The condition for the classical CI for D, that both the number of cases and non-cases exceed 10 in both groups, is far from met. So only the plus four interval is appropriate (it easily meets its condition of sample sizes ≥ 5 in both groups). The first impression is that the probability of same smoking habits is very similar in pairs with separated and joint upbringing. The CI easily includes zero, and with $\hat{D} = \hat{p}_j - \hat{p}_s < 0$ the observed difference is not even in the direction of interest. We continue our analysis by testing the hypothesis of equal probabilities in the two groups against a two-sided alternative (a one-sided $H_a : p_j > p_s$ could also be argued),

$$H_0 : p_s = p_j \quad \text{versus} \quad H_a : p_s \neq p_j.$$

Because of the low total number of twin pairs with different smoking habits (i.e. $4 + 5 = 9$ such twin pairs), it is clear that we will not be able to get expected counts above 5 in all cells. Therefore we turn to Fisher's exact test, and it gives $P = 0.73$ against a two-sided alternative. It is clearly non-significant, and the same conclusion would have resulted from $Z = 0.40$ or $X^2 = 0.18$. We conclude that there is no evidence of a different likeness in smoking habits among monozygotic pairs with separated and joint upbringing. With so similar estimated probabilities in the two groups, we may view this result as an indication of no relevant effect (the CI for the difference quantifies how large an effect could be supported by the data) rather than simply lack of evidence against the null hypothesis.

4. Argument against causal hypothesis

We may illustrate the question of causation between smoking and detrimental health effects (in particular lung cancer) in the form of the diagrams from the IPS and PSLS textbooks used to discuss causation.



The lurking (or confounding) variable z of interest is genetical factors. It is not unreasonable to assume that genetical factors are directly linked to disease, i.e. the required association between z and y. The findings from **2)** hint at a possible genetical component to smoking habits, i.e. an association between z and x. Because the genetics temporally predates both of the other variables, the arrows can be thought of as going from z to x and y, respectively. This would the make causal structure correspond to that of z being a common cause of x and y. Our analysis in **3)** plays the role of confirming that the demonstrated association between zygosity and smoking habits cannot be explained by environment rather genetics. It would seem as a bold overinterpretation of the result from **3)** to claim that smoking habits in general have no environmental component, but for the argument related to the diagram (i.e. the specific twin pairs) that is not necessary. One of the weaknesses of the argument is perhaps that the *same* genetic components (or at least genetical components that are strongly linked) must be responsible for predispositions to both smoking and disease. In order to establish that one would for example need to have disease information for the individuals for which we already have the genetical and smoking habit information.