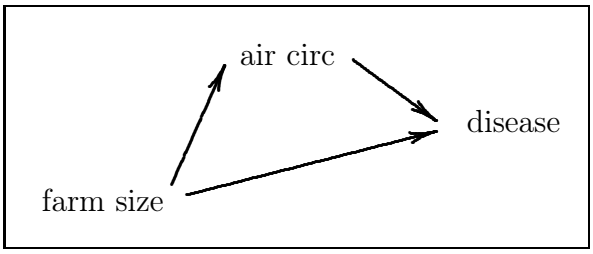
Biostats VHM 811 Course
Fall 2007, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment 3

The data originate from a study in Denmark, reported in Willeberg P (1979), The analysis and interpretation of epidemiological data, Proceedings of the 2nd International Symposium on Veterinary Epidemiology and Economics, Canberra, Australia. The disease is called Swine Enzootic Pneumonia (SEP), and the air circulation systems are referred to as either a fan or no-fan system.

1. Study design and causal diagram

The description of the study indicates that "case" and "control" farms were selected based on a three-year period prior to the study. Although it is not stated explicitly how the controls were selected, this hints at the study being a retrospective case-control study. The study population is not defined, and it is not clear from the text that the control farms included in the study are a sample from a larger population of controls, but is seems likely to be the case. The more plausible causal relation between the explanatory factors is from farm size to air circulation system. That is, larger farms are more likely to have an air circulation system (the association is positive as can be seen directly from the data). The causal relation may be a question of logistics — in a large animal-holding facility there may be a need to install an air circulation system because there is too little natural air flow in the room. It seems unnatural to have the causal relation in the other direction (how can the presence or absence of an air circulation system affect the farm size?). Assuming also a relation of both factors with the disease, we arrive at the following causal diagram.



2. Crude measures of association with disease

For a case-control study, the appropriate measure of association is the odds-ratio (OR). For the binary predictor air circulation, calculation of the odds-ratio with an associated confidence interval and a Pearson chi-square significance test is straightforward. The crude odds-ratio is 2.99, with a 95% confidence interval (CI) of (1.65,3.46). The CI does not contain 1 from which we deduce that there is a significant association. The Pearson chi-square confirms this: $X^2 = 15.3$, $\text{df} = 1$, $P < 0.001$. Presence of an air circulation system seems to increase the risk of disease (at a high level), as measured by the OR, by a factor of about 3.

For the categorical predictor farm size, the Pearson chi-square test still applies but there is no longer a single odds-ratio. The two options are (i) to report odds-ratios for each category relative to a "baseline" category, or to dichotomize the variable at one cut-point to obtain an odds-ratio comparing farms larger and smaller than the cutpoint; this calculation should be preferably be carried out at several cut-points to avoid losing information. For simplicity, the lowest farm size group has been chosen as the baseline category, although it is generally best to choose a category with a fairly large number of observations as the baseline category. See table for results.

	Farm size				
Parameter	0-199	200-299	300-399	400-499	500+
odds-ratio vs baseline	1	2.72	3.91	5.67	25.3
95% CI	n/a	(0.93,8.97)	(1.27,13.3)	(1.34,24.4)	(8.0,86.2)
odds-ratio at cutpoint	n/a	6.30	4.82	6.90	9.23
95% CI	n/a	(2.46,19.0)	(2.70,8.67)	(3.70,13.0)	(4.52,19.7)
test	$X^2 = 57.7,\ \mathrm{df} = 4,\ P < .001$				

The Pearson chi-square test gives strong evidence of a crude association between farm size and disease level. Both sets of odds-ratios show that larger farms have increased risk of disease at a high level because all odds-ratios are >1, and the odds-ratios vs baseline tend to increase with farm size.

3. Confounding

The causal diagram shows that farm size could be a confounder for air circulation. On the other hand, air circulation is on the pathway between farm size and the outcome — an intervening (or intermediate) variable — and can therefore not be a confounder. There are 3 further conditions to check for whether farm size is indeed a confounder:

1. *Substantial effect change.* The Mantel-Haenszel estimate of the effect of air circulation stratified on farm size categories is 1.26 with an associated 95% CI of (0.66,2.42). The effect change is overwhelming: (2.99-1.26)/1.26=137%. Farm size certainly affects the estimate enough to qualify as a confounder. Furthermore, upon stratification the effect of air circulation is not longer significant, as is apparent from the 95% CI and the M-H test: $X^2_{\mathrm{MH}} = 0.48,\ \mathrm{df} = 1,\ P = 0.49$.

2. *Association with outcome.* The crude associations were computed in Question 2, but the assessment should really be done among the exposure-negative farms. The Pearson chi-square statistic is 11.43, which corresponds to $P = 0.022$ in a χ^2-distribution with 4 degrees of freedom. There is a significant association of farm size with the outcome among the exposure negative farms. The odds-ratios at cutpoints range from 3.25 to 11.2, and are thus of a similar magnitude as in the combined dataset including also the exposure-positives. The increased P-value is probably mostly attributable to lower power because of the reduced sample size.

3. *Association with exposure.* We should assess the association with exposure among the controls. The Pearson chi-square statistic is 25.95 which is highly significant in a χ^2-distribution with 4 degrees of freedom. It is evident from the data that the use of air circulation systems is more frequent in larger farms.

We conclude that farm size meets all conditions to be a confounder for the effect of air circulation system on the disease level.

4. Interaction

The Mantel-Haenszel analysis for the effect of air circulation after stratification on farm size included a test for homogeneity of the odds- ratios: $X^2_{\mathrm{hom}} = 1.01$ corresponding to $P = 0.91$ in a χ^2-distribution with 4 degrees of freedom. There is no indication whatsoever of an interaction between farm size and air circulation.

In summary, farm size does act as a confounder for the relation between air circulation and disease, and as air circulation has no substantial or statistically significant effect after controlling for farm

size, it can be said that farm size exerts *complete confounding* on the relation between air circulation and farm size. The M-H odds-ratio for air circulation was given in Question 3. Also, by the lack of (significant) effect of air circulation, the relevant associations between farm size and disease are the crude associations computed in Question 2.

5. Control of confounding by (future) study design

For this question we will consider the confounding of farm size on the association between air circulation and disease level in a case-control study. The two procedures for control of confounding by design are restriction and matching.

Restriction here means restricting the study to involve farms of the same size, or with sizes within a fairly narrow range (e.g. one of the 5 categories in the dataset). This would restrict the scope of the study drastically, and the choice of which range to focus on would probably be driven mostly by considerations on which range was considered of primary interest. One additional consideration is that the chosen range should not represent a too small part of the population. Among the controls, the largest farm size group is 200–299, so this range would be perhaps be the default choice. After the range has been chosen, the data collection proceeds as for an ordinary case-control study within the restricted study population.

Matching on farm size means that the controls are selected to produce the same distribution of farm sizes as among the cases, shown in the table below.

Farm size	0-199	200-299	300-399	400-499	500+	total
Number of case farms	6	23	20	9	58	116
Proportion	0.05	0.20	0.17	0.08	0.50	1

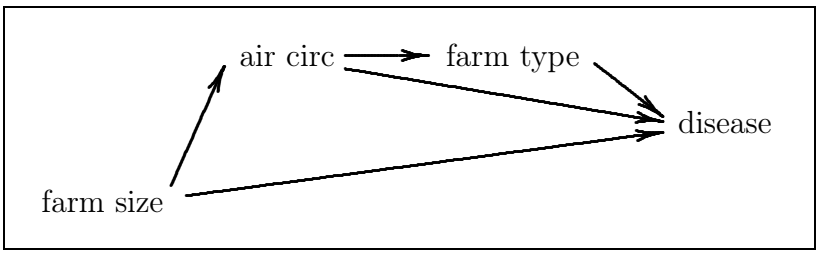
Specifically, when sampling control farms one aims to achieve the size distribution of the table. This would usually involve stratified sampling.

Even if matching the farm size distribution to the table above may be challenging, the matching approach seems far preferable to restriction, simply because restricting the farm size may produce such a narrow scope of the study that its results become of limited interest.

6. Construction of causal structures

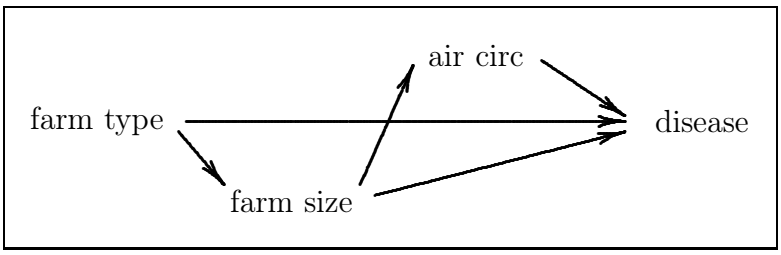
For this question we consider, as before, air circulation as the exposure of primary interest. All statements about the causal role of the added variable, farm type, refer to its impact on the estimation of the association between exposure and disease.

Farm type is not a potential confounder
The simplest example of a relation of this type is perhaps if farm type is an intervening (intermediate) variable between air circulation and disease.



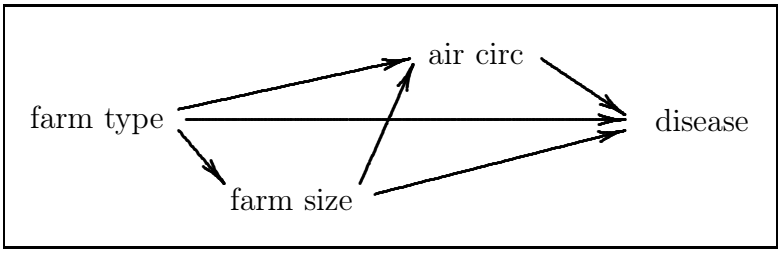
Note that the association between farm size and farm type is only through the other variables (in particular, air circulation system); this is most likely an unrealistic assumption. If a direct relation existed between these two variables, farm type could still be a potential confounder because an alternative pathway through farm type connecting the exposure and outcome would exist.

Farm type is a potential confounder but does not need control
This type of scenario arises for example if the only path from farm type to the exposure is through farm size. In this case, controlling for farm size blocks the path, and no additional control is needed.



This type of relationship may be plausible if farm type does not involve the facilities of the farm (the barns).

Farm type is a potential confounder but does need control
Simply add a direct arrow from farm type to the exposure. As there are now two routes from disease back to exposure, controlling for farm size alone is not sufficient.



This structure may represent the most likely scenario, where the type of air circulation system probably depends directly on the farm type.