Study Design and Confounder Control in Observational Studies: Two Cases

John Andrew Harraway
The University of Otago, Department of Mathematics & Statistics
Science III building, Cumberland Street
Dunedin, New Zealand
email: jharraway@maths.otago.ac.nz

1. Introduction

In New Zealand students complete high school education at ages 17 to 18 and most of those who proceed to university do so at ages 18 to 19. The final year at high school is called Year 13. In 2006 the New Zealand Qualifications Authority introduced a subject at Year 13 called Statistics and Modelling which replaced the earlier Mathematics with Statistics (New Zealand Qualifications Authority Report, 2005).

Expectations of the new subject placed pressure on high school teachers to develop new material. This has been aggravated by the need to award passes at different levels called achieve, merit and excellence, the latter requiring considerable insight into the statistical concepts and methodology involved in project work; there is an extensive project component associated with the subject in the final year at school as well as an examination.

In order to assist teachers and provide ideas for the classroom which would motivate students, an education day was organized at the New Zealand Statistical Association Conference in July 2005 at the University of Otago. Seven researchers from the University in the areas of human nutrition, epidemiology, marketing, psychology, zoology, sociology and statistics were invited to discuss one of their current research projects. The presentations were recorded and subsequently re-recorded and edited in the studio. This produced a DVD of 145 minutes. At the same time the relevant data sets were placed on a CD. Subsequently, Statistics New Zealand added two presentations along with a link to their website giving access to further data. The final package including DVD and CD was made available to teachers in February 2007 (Harraway and van der Vyver, 2007)

As well as providing a source of data for project work, the research topics were selected to facilitate the interest of students in statistics by showing current projects relevant to this age group. Study design principles, sampling procedures, observational data and confounding effects were covered. The two case studies presented in this paper discuss these ideas and are supported by two movie clips, segments of which will be shown during this presentation (Dickson, 2007; Ferguson, 2007). The first study can be viewed at http://www.maths.otago.ac.nz/~jharraway/Presentation1.avi and the second study can be viewed at http://www.maths.otago.ac.nz/~jharraway/Presentation2.avi. Both are observational and adjustment is made for confounder effects, either by stratification and regression modelling or by ignoring a number of the cases affected by some values of a confounding variable.

2. A longitudinal study: long term effects of circumcision

The first case study is based on the Dunedin Multidisciplinary Health and Development Study (Silva and Stanton, 1996). Information has been compiled from a birth cohort of 1037 children born in Dunedin between April 1, 1972 and March 31, 1973. The members of the cohort have been assessed at ages three, five, seven, nine, 11, 13, 15, 18, 21, 26 and 32. The investigation here reports on some observational data at age 26 and aims to determine if circumcision by age three years affects the risk of acquiring herpes (Dickson et.al., 2005).

For ethical reasons data from each individual cohort member cannot be published but equivalent data involving 1890 cases have been generated by simulation for analysis. The simulated data give the same results as those obtained with the actual data. The simulated data are reported in the CD. The response variable is “Presence of herpes at age 26” (Yes, No). The predictors are “Circumcised by age 3” (Yes, No), “Number of sexual partners” (< 10, ≥ 10), “Socioeconomic status at age 26” (High, Medium, Low) and “Education level” (High School, Postsecondary, University). Several other variables are in the CD file but not used in the analysis reported here or in the DVD.

2.1 Preliminary Results

The first analysis of the data calculates the crude effect of circumcision by age three on the presence of herpes at age 26. Table 1 summarizes the 1890 simulated case results. The prevalence of herpes in the circumcised group is 6.23% and in the uncircumcised group is 8.75%. The risk ratio of herpes in the circumcised group compared with the uncircumcised group is 0.71 (95% CI 0.51-0.99). The 29% reduction in risk resulting from circumcision is not therefore likely to occur by chance and it is tentatively concluded that circumcision by age three years is protective of herpes.

Table 1 Crude relationship between herpes and circumcision by age 3.

	Herpes		Total	Prevalence
	Yes	No		
Circumcised by age 3	48	722	770	6.23%
Not circumcised by age 3	98	1022	1120	8.75%
Total	146	1744	1890	

But this procedure measures only a crude relationship between herpes and circumcision without adjustment for potential confounding variables such as, for example, sexual behaviour which could alter the conclusion about the effect of circumcision on the acquisition of herpes.

2.2 Stratification and Confounder Control

Circumcision is not a random event and the lower prevalence of herpes in the circumcised group could reflect the different sexual behaviour among those circumcised. To adjust for such a confounder a stratified analysis is carried out with results summarized in Table 2 and Table 3. Here a variable measuring sexual behaviour during the years of adult life is categorical with the two levels, "< 10 sexual partners" and "10 or more sexual partners" by age 26.

Table 2 Relationship between herpes and circumcision if < 10 sex partners.

	Herpes		Total	Prevalence
	Yes	No		
Circumcised by age 3	24	536	560	4.29%
Not circumcised by age 3	24	470	494	4.86%
Total	48	1006	1054	

Table 3 Relationship between herpes and circumcision if 10 or more sex partners.

	Herpes		Total	Prevalence
	Yes	No		
Circumcised by age 3	24	186	210	11.43%
Not circumcised by age 3	74	552	626	11.82%
Total	98	738	836	

The number of sexual partners is clearly related to circumcision. In particular, 560 of the 1054 (53.1%) with less than 10 partners were circumcised versus 210 of the 836 (25.1%) with 10 or more partners. Also 48 of 1054 (4.6%) with less than 10 partners developed herpes versus 98 of 836 (11.7%) with 10 or more partners. The outcome of herpes is therefore related to the number of sexual partners as well.

To further investigate the relationship between herpes and circumcision the separate risk ratios are calculated from Table 2 and Table 3. The risk ratio in the group with less than ten partners is 0.88 (95% CI 0.51-1.53) and the risk ratio in the group with 10 or more partners is 0.85 (95% CI 0.55-1.31). The risk ratios increase towards one in both cases, the confidence intervals include one and it can be concluded that there is no evidence that circumcision affects the acquisition of herpes once this measure of sexual behaviour is taken into account.

At this stage of the investigation the ideas should be within the understanding of year 13 students at school. But students at university can be taken further. The variable defining the number of sexual partners is said to be a negative confounder since it is negatively related to the exposure variable, circumcision, and positively related to the outcome of herpes. The shift towards one for the risk ratios in Tables 2 and 3 is a consequence of this negative confounding.

The odds ratio can also be introduced and its relationship to the risk ratio for rare diseases discussed. The Mantel-Haenszel test as a method of inference for stratified categorical data can also be developed from the two strata defined in Tables 2 and 3. This can be extended to a discussion of effect modification or interaction if the strength of the association between circumcision and herpes is not the same in Tables 2 and 3. These ideas are reasonably straightforward and can be presented to students who may not be familiar with the logistic regression model.

2.3 Logistic Regression and Confounder Control

The second analysis uses logistic regression to model the odds of herpes for those circumcised compared with those not circumcised. Table 4 shows a crude odds ratio of 0.69 for the presence of herpes for those circumcised compared with those not circumcised by age three years. The confidence interval excludes 1 showing evidence of reduced odds of herpes among those circumcised (p-value = 0.045). This is in agreement with the result obtained with the earlier crude risk ratio.

Table 4 Effect of circumcision on odds of herpes (1 = present, 0 = absent)

Factor	Parameter	Odds Ratio (95% CI)
Circumcision (Ref: not circumcised)	0.366	0.69(0.49,0.99)
Constant	-2.345	

Table 5 shows the effect of circumcision on herpes presence after adjusting for sexual behaviour defined by the categories “10 or more sexual partners” (the reference category) and “less than 10 sexual partners”. There is evidence that the odds of herpes decreases if there are fewer than 10 sexual partners but there is no evidence that circumcision by age three years has an effect on herpes presence after adjusting for this effect. This agrees with the shift identified by stratification in Section 2.2.

Table 5 Factors associated with presence of herpes (1 = present, 0 = absent)

Factor	Parameter	Odds Ratio(95% CI)
Circumcision (Ref: not circumcised)	-0.077	0.93(0.64,1.35)
Sexual partners (Ref: 10 or more)	-1.002	0.37(0.25,0.53)
Constant	-2.000	

Table 6 shows the results of adding further factors to the model. There is evidence that the number of sexual partners is the only factor with an effect on herpes. The simulated data available in the accompanying CD involving 1890 cases reports the values of more variables which students can explore in their own time. A modification resulting from the interaction between circumcision status and sexual partner category can also be assessed by introducing a new variable which is the product of the values of these two variables.

Table 6 Factors associated with presence of herpes (1 = present, 0 = absent)

Factor	Parameter	Odds Ratio (95% CI)
Circumcision (Ref: not circumcised)	-0.072	0.93(0.64, 1.36)
Sexual partners (Ref: 10 or more)	-0.992	0.37(0.26, 0.54)
Socioeconomic (Ref: High)		
Medium	-0.103	0.90(0.58, 1.42)
Low	-0.225	0.80(0.49, 1.31)
Education (Ref: High School only)		
Postsecondary	0.113	0.89(0.61, 1.30)
University	0.270	0.76(0.47, 1.25)
Constant	1.786	

2.4 Discussion

This example introduces students early in their education to study design ideas including cohort analysis, the limitations of observational data and confounder control as well as ethical issues present when dealing with the human population. With further reading students will discover that a second large cohort study (Ferguson et.al., 2006) produces results from observational data in conflict with the Dunedin results described here. But is this a serious problem? An as yet unpublished communication shows reasons for this discrepancy.

Observational data may not lead to conclusions about causation but such data can suggest hypotheses which can then be investigated properly by designed randomized experiments and clinical trials. With these randomized experiments confounder control will not be the issue it is in observational studies. In Africa clinical trials have been established to assess the effect of circumcision on the incidence of HIV/AIDS (Alanis et.al., 2004; Auvert, et.al., 2005). In a Kenyan study (Bailey, et.al., 2007), 1391 circumcised men were compared to 1393 who were not circumcised. In Uganda (Gray, et.al., 2007) 2474 circumcised men were compared to 2522 who were not. After tracking the men for two years, it was found that circumcised men were 51% to 60% less likely to contract HIV than their uncircumcised counterparts. The results were confirmed by further tests carried out by the United States' National Institutes of Health. (Sawires, et.al., 2007).

This example has been used in a large first year class to motivate students. The data have also been used for project work in a second year class on regression and modelling, and this has facilitated student interest. The context has been helped by recent reports in the media about the relationship between circumcision and sexually transmitted diseases. One current study based on the analysis of the Dunedin cohort at age 32 shows evidence that the risk of herpes acquisition increases over early adulthood. (Dickson et.al., 2007).

3. A cross sectional study: iron levels in New Zealand children

Iron deficiency is a concern for the health and development of young children. This issue has been extensively researched (Soh et.al., 2004). The study described in the second DVD clip (Ferguson, 2007) estimates the prevalence of iron deficiency in New Zealand children and explores factors associated with low body iron store as measured by serum ferritin levels.

The study is cross sectional with 323 participating families randomly identified and interviewed in three New Zealand cities; convenience sampling of hospital records involving small numbers of cases was not used. But the data are again observational with information being collected on iron status, health status, blood composition, various socioeconomic factors and past feeding practices of babies such as extent of breast feeding, formula feeding, cows milk feeding and dietary iron intake. C-reactive protein (CRP) was also measured. This is an index of infection which can affect a child’s blood indices and give an inaccurate assessment of body iron store. Children with infection should therefore be excluded from the analysis or alternatively a method found for adjusting for this confounding effect. In the first case conclusions can only be drawn for children unaffected by infection. No attempt was made to control infection in the collection of the data.

3.1 Preliminary Results

The children were first classified according to whether they had normal body iron store levels or body iron store deficiency at increasing levels defined as Stage 1, Stage 2 and Stage 3 (Ferguson, 2007). Those with infection, a confounding factor defined by CRP > 10 μg/L, were excluded from this initial analysis to remove the effect of infection on the results. The prevalence of body iron store deficiency over all 233 children remaining in the analysis turned out to be 20.2% (95% CI 15.2% to 25.9%).

The 233 children were then divided into a group of 72 infants aged 6 to 11 months and a group of 161 toddlers aged 12 to 24 months. Chi square tests on the infants and toddlers with the numbers in the three stages of body iron store deficiency and overall deficiency produced respectively p-values 0.014, 0.350, 0.234 and 0.051. Counts in the various categories suggest higher prevalence of body iron store deficiency among toddlers at the stage 1 level but no evidence of differences at stage 2 and stage 3 levels (Ferguson, 2007). Overall the evidence of higher levels among the toddlers is borderline significant indicating attention should be focussed on the toddlers rather than the infants in New Zealand if wanting to improve iron levels in children.

The effects of breast feeding, formula feeding and cows milk feeding on body iron store (ferritin) levels are summarized in Table 7. For example 20% of those breast fed are in the first (low) tertile of ferritin levels, 18% in the second (average) tertile of ferritin levels and 31% in the third (high) tertile of ferritin levels. Those with infection measured through elevated CRP level are again excluded. If there were to be no effects of these dietary variables on iron levels we would expect the percentages to be similar across each row. The data show that formula fed children have higher ferritin levels but cows milk fed children have lower ferritin levels with no trend apparent for those breast fed. The p-values for the chi square tests show evidence that the type of milk has an effect on iron stores and in particular cows milk consumption results in low iron stores.

Table 7 Categorical dietary variables: % per ferritin tertile

	1st tertile (%)	2nd tertile (%)	3rd tertile (%)	p-value
Breast fed	20	18	31	0.098
Formula fed	15	36	46	< 0.001
Milk fed (> 500ml	53	34	15	< 0.001

3.2 Regression and Confounder Control

So far infection has been controlled by excluding those children with elevated CRP from the analysis. A regression model is now used to adjust for the effect of this confounder and also to incorporate the effects of influential non dietary factors. To normalize the distributions of the ferritin levels and the diet iron levels the logs of these values are taken and the new variables are used when fitting the multiple regression. The results are shown in Table 8. Continuous variables for age and log of diet iron level are predictors in the model fitted. The level of dietary iron intake remains positively associated with higher body iron stores even after adjusting for the confounding factor (CRP) and influential non-dietary factors (age, sex and prematurity). Age is negatively associated with iron stores as expected from the results for the infants and toddlers. Girls have slightly higher iron stores and premature birth results in slightly lower iron stores.

Table 8 Factors associated with iron levels in children (Response: log serum ferritin level)

Factor	Parameter	Standard error	t	p-value
Log(diet iron level)	0.148	0.068	2.186	0.030
Age	-0.030	0.008	-4.032	<0.001
Sex: (Ref male)	0.156	0.080	1.941	0.053
Premature: (Ref normal)	-0.206	0.113	-1.830	0.068
Infection (Ref: CRP $\Leftarrow$ 10mg/L)	0.794	0.165	4.800	< 0.001
Constant	3.104			

3.3 Discussion

Although the contingency tables with infected children deleted result in simple methods of analysis and the more sophisticated regression modelling produces similar results, this may not always be the case and it is therefore essential to control for any anticipated confounding effects. It is wasteful to ignore a large block of data by removing infected children. This will lead to results only relevant for children without infection.

The data file associated with this study recorded on the CD has other variables not mentioned here and therefore is a rich source of data for further project work. Not only modelling and confounder control but other multivariate procedures can be used.

4. Future Research

The DVD has been made available to all high schools in New Zealand and to tertiary training institutes, universities, polytechnics and colleges of education. Interest has also been expressed by teachers of Year 13 biology as well as Year 13 statistics and copies of the DVD/CD package have been ordered already by some biology teachers. The Higher Education Development Unit at the University of Otago viewed this as a community project and therefore provided staff to compile the DVD free of charge. Feedback will be monitored carefully to determine which aspects of the DVD and CD are the most useful for teaching and learning. This will be achieved through an appropriate survey of teachers and students and focus group discussion at an education session at the next conference of the New Zealand Statistical Association. These assessments are particularly important. Many people produce DVD's for teaching but it is essential to know if they are actually used and are beneficial.

Once the useful sections of the DVD and CD are established it will be possible to see how best we can develop further teaching aids. Several projects currently being suggested for future inclusion are injury prevention, the profile of women who may consume alcohol during pregnancy (Parackal et.al., 2007) and lifestyle and health survey data.

REFERENCES

Alanis, M.C. and Lucidi, R.S. Neonatal Circumcision: a Review of the World's Oldest and Most Controversial Operation. *Obstetrical and Gynecological Survey* (2004), 59, 379-396.

Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R. and Puren, A. Randomized, Controlled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk. The ANRS 1265 Trial. *PLoS Medicine* (2005), 2, 11.

Bailey, R.C., Moses, S., Parker, C.B., Agot, K., McLean, A., Krieger, J.N., Williams, C.F.M., Campbell, R.T. and Ndinya-Achola, J.O. Male Circumcision for HIV Prevention in Young Men in Kisumu, Kenya: a Randomised Controlled Trial *Lancet* (2007), 369, 643-656.

Dickson, N., van Roode, T. and Paul, C. Herpes Simplex Virus Type 2 Status at Age 26 is not Related to Early Circumcision in a Birth Cohort. *Sexually Transmitted Diseases* (2005), 32, 8, 517-519.

Dickson, N., van Roode, T., Herbison, P., Taylor, J., Cunningham, A. and Paul, C. Risk of Herpes simplex Virus Type 2 acquisition increases over early adulthood: evidence from a cohort study. *Sexually Transmitted Infections* (2007), 83: 87-90.

Dickson, N. A Ten Minute Course in Epidemiology. In: *Statistics in Research: a Visual Teaching Resource*. Editors Harraway, J.A and van der Vyver, R. Department of Mathematics and Statistics, University of Otago, Dunedin (2007). [available for viewing at http://www.maths.otago.ac.nz/~jharraway/Presentation1.avi]

Ferguson, D.M. ,Boden, J.M. and Horwood, L.J. Circumcision Status and Risk of Sexually Transmitted Infection in Young Adult Males: An Analysis of a Longitudinal Birth Cohort. *Pediatrics* (2006), 118,1971-1977.

Ferguson, E. Is Iron Deficiency Common Among New Zealand Infants and Toddlers? In: *Statistics in Research: a Visual Teaching Resource*. Editors Harraway, J.A. and van der Vyver, R. Department of Mathematics and Statistics, University of Otago, Dunedin (2007). [available for viewing at http://www.maths.otago.ac.nz/~jharraway/Presentation2.avi]

Gray, R.H., Kigozi, G., Serwadda, D., et.al. Male Circumcision for HIV Prevention in Men in Rakai, Uganda: a Randomised Trial. *Lancet* (2007), 369, 657-666.

Harraway, J.A. and van der Vyver, R. Editors. *Statistics in Research: a Visual Teaching Resource.* Department of Mathematics and Statistics, University of Otago, Dunedin (2007).

New Zealand Qualifications Authority. Combined report for the Review of Mathematics and Statistics and Probability and the Review of Mathematics (Mathematics with Calculus) Level 3 achievement standards. http://www.nzqa.govt.nz/nqfdocs/summaries/2005/nov05/revsumnov05-04.pdf. 14 February, 2007.

Parackal, S., Ferguson, E. and Harraway, J.A. Alcohol and tobacco consumption among 6-24-months post-partum New Zealand women. *Maternal and Child Nutrition* (2007) 3(1), 40-51.

Sawires, S.R., Dworkin, S.L., Fiamma, A.F., Peacock, D., Szekeres, G. and Coates, T.J. Male Circumcision and HIV/AIDS: Challenges and Opportunities. *Lancet*. (2007), 369, 708-713.

Silva, P.A. and Stanton, W.R. *From Child to Adult: the Dunedin Multidisciplinary Health and Development Study*. Oxford University Press (1996).

Soh, P., Ferguson, E.L., McKenzie, J.E., Homs, M.Y.V. and Gibson, R.S. Iron deficiency and risk factors for lower iron stores in 6-24-month-old New Zealanders. *European Journal of Clinical Nutrition* (2004) 58,71-79