MATH 523: MATCHED PAIRS ANALYSIS

Consider $2n$ data that collected as matched pairs, that is n pairs of data (y_{i1}, y_{i2}) for $i = 1, \ldots, n$. Suppose that the outcomes are binary, so that the data can be collated into a 2×2 table:

		$j=2$	
		$Y=1$	$Y=0$
$j=1$	$Y=1$	n_{11}	n_{10}
	$Y=0$	n_{01}	n_{00}

where n_{11} counts the number of pairs in which $y_{i1} = 1$, $y_{i2} = 1$ etc. Consider a multinomial model for these data, $(N_{00}, N_{01}, N_{10}, N_{11}) \sim Multinomial(n; \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$. The ML estimates for these parameters are

$$(\widehat{\pi}_{00}, \widehat{\pi}_{01}, \widehat{\pi}_{10}, \widehat{\pi}_{11}) = \frac{1}{n}(n_{00}, n_{01}, n_{10}, n_{11})$$

Denote by π_{1*} and π_{*1} the probabilities of response $Y = 1$ for $j = 1$ and $j = 2$ respectively, that is

$$\pi_{1*} = \pi_{10} + \pi_{11} \qquad \pi_{*1} = \pi_{01} + \pi_{11}$$

(a) We may estimate (π_{1*}, π_{*1}) from the observed data, but the estimators will be **dependent** due to pairing; in fact, by properties of the multinomial distribution,

$$\text{Cov}[\widehat{\pi}_{1*}, \widehat{\pi}_{*1}] = \text{Cov}[\widehat{\pi}_{10} + \widehat{\pi}_{11}, \widehat{\pi}_{01} + \widehat{\pi}_{11}] = \frac{(\pi_{00}\pi_{11} - \pi_{01}\pi_{10})}{n}$$

(b) If $\pi_{1*} = \pi_{*1}$, so that $\pi_{0*} = \pi_{*0}$ also, then the marginal distributions are identical (so we have *marginal homogeneity*). Furthermore,

$$\delta = \pi_{1*} - \pi_{*1} = \pi_{10} - \pi_{01}$$

so if $\delta = 0$, $\pi_{1*} = \pi_{*1}$, then $\pi_{10} = \pi_{01}$, and this corresponds to *symmetry* in the 2×2 table.

(c) We have that $\widehat{\delta} = \widehat{\pi}_{1*} - \widehat{\pi}_{*1}$, and, by the multivariate Delta method,

$$\sqrt{n}(\widehat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, V(\pi))$$

where

$$V(\pi) = \pi_{1*}(1 - \pi_{1*}) + \pi_{*1}(1 - \pi_{*1}) - 2(\pi_{00}\pi_{11} - \pi_{01}\pi_{10})$$

(d) For large n, $\widehat{\delta}$ is approximately distributed as $\mathcal{N}(\delta, V(\widehat{\pi})/n)$, where

$$V(\widehat{\pi}) = (\widehat{\pi}_{10} + \widehat{\pi}_{01}) - (\widehat{\pi}_{10} - \widehat{\pi}_{01})^2$$

(e) Under the null hypothesis of $H_0 : \delta = 0$, or equivalently $H_0 : \pi_{1*} = \pi_{*1}$ or $H_0 : \pi_{10} = \pi_{01}$ (corresponding to marginal homogeneity), the Wald test statistic W_n takes the form

$$W_n = \frac{\sqrt{n}\,\widehat{\delta}}{\sqrt{V(\widehat{\pi})}}$$

(f) Alternatively, a different (simpler) estimator of the asymptotic variance is $V_0(\widehat{\pi}) = (\widehat{\pi}_{10} + \widehat{\pi}_{01})$. Then

$$W_n = \frac{\sqrt{n}\,\widehat{\delta}}{\sqrt{V_0(\widehat{\pi})}} = \frac{n_{10} - n_{01}}{\sqrt{n_{01} + n_{10}}}$$

The squared Wald statistic, W_n^2 may be used; under H_0, $W_n^2 \sim \chi_1^2$.

Example:- Prime Minister Approval 1600 UK voters were asked on two occasions separated by six months whether they approved or disapproved of the performance of the Prime Minister.

		Second Survey	
		Approve	Disapprove
First Survey	Approve	794	150
	Disapprove	86	570

Here $\widehat{\delta} = \widehat{\pi}_{1*} - \widehat{\pi}_{*1} = (150 + 794)/1600 - (86 + 794)/1600 = 0.04$, with estimated standard error

$$\frac{1}{\sqrt{n}}\sqrt{(\widehat{\pi}_{10} + \widehat{\pi}_{01}) - (\widehat{\pi}_{10} - \widehat{\pi}_{01})^2} = \frac{1}{\sqrt{1600}}\sqrt{\left(\frac{150}{1600} + \frac{86}{1600}\right) - \left(\frac{150}{1600} - \frac{86}{1600}\right)^2} = 0.009549215$$

Thus the Wald statistic is

$$W_n = \frac{0.04}{0.00955} = 4.188$$

which is a highly significant result (after comparison with the standard Normal distribution); the probability of approval has decreased significantly. The second Wald statistic uses $V_0(\widehat{\pi})$

$$W_n = \frac{n_{10} - n_{01}}{\sqrt{n_{01} + n_{10}}} = \frac{150 - 86}{\sqrt{150 + 86}} = 4.166$$

Matched pairs can also be analyzed using *conditional logistic regression*; a pair or *stratum* indicator is used to identify paired observations, and then the comparisons made between surveys which are treated as a levels of a binary factor. In R, conditional logistic regression can be carried out using the `clogit` function in the `survival` library. In this analysis X is the survey number (1-first survey, 0-second survey), and Y is approval indicator (1-approve, 0-disapprove):

```
1  library(survival)
2  Pair<-rep(c(1:1600),2)
3  X<-rep(c(1,0),each=1600)
4  Y<-rep(0,3200)
5  for(i in 1:794){Y[Pair == i]<-1}
6  for(i in 795:944){Y[Pair == i]<-c(1,0)}
7  for(i in 945:1030){Y[Pair == i]<-c(0,1)}
8  for(i in 1030:1600){Y[Pair == i]<-0}
9  Prime<-data.frame(X,Y,Pair)
10 clogit(Y~X+strata(Pair),data=Prime)
11
12
13 > clogit(Y~X+strata(Pair),data=Prime)
14 Call: clogit(Y ~ X + strata(Pair), data = Prime)
15
16    coef exp(coef) se(coef)    z       p
17 X 0.568      1.76    0.136 4.18 2.9e-05
18
19 Likelihood ratio test=18.2  on 1 df, p=1.97e-05  n= 3200, number of events= 1823
```

The `clogit` function on line 10 fits the conditional logistic regression. The results on line 17 indicate a significant association; the estimate 0.568 is the subject-specific log odds ratio, $\widehat{\zeta}$, which is significantly less than zero (p-value is 2.90e-5). This confirms the result that approval rating decreased between surveys.