STA 2101F Notes on 2×2 tables, and case-control studies. October, 2022

Following Cox & Donnelly (Table 3.6), we outline three 2×2 tables, indicating the sampling distribution of the pair (t, y), both binary; y is the response of interest and t is the explanatory variable. The first represents the joint distribution in the population. In the second, we follow individuals forward in time, recording the value of t at the outset, and observe the value of y. These are called prospective, or cohort, studies. In the third, we we start with the response y, usually finding a group with $y = 1$, (a case) and then for each item with $y = 1$ we find one or more with $y = 0$ (a control) and observe the value of t. These are retrospective, also called case-control, studies. Sometimes the control is matched to each case, and we refer to a matched-pairs study. Note that the row margins in (b) are 1, whereas the column margins in (c) are 1.

Table 1: (a) is the population, (b) is prospective, and (c) is retrospective. The response is y, and the covariate is t (treatment), although CD use z.

(a) Population

	$y = 0$	$y = 1$
$t = 0$	π_{00}	π_{01}
$t = 1$	π_{10}	π_{11}

(b) Prospective study

	$y = 0$	$y = 1$
$t = 0$	$\pi_{00}/(\pi_{00} + \pi_{01})$	$\pi_{01}/(\pi_{00} + \pi_{01})$
$t = 1$	$\pi_{10}/(\pi_{10} + \pi_{11})$	$\pi_{11}/(\pi_{10} + \pi_{11})$

(c) Retrospective study

	$y = 0$	$y = 1$
$t = 0$	$\pi_{00}/(\pi_{00} + \pi_{10})$	$\pi_{01}/(\pi_{01} + \pi_{11})$
$t = 1$	$\pi_{10}/(\pi_{00} + \pi_{10})$	$\pi_{11}/(\pi_{01} + \pi_{11})$

In (b) we collect samples from $t = 0$ ("untreated") and $t = 1$ ("treated") and y is measured.
In (c) we collect samples from $y = 0$ ("control") and $y = 1$ ("case"); t is measured.

Generic notation often used for the data that has been collected is

	$y = 0$	$y = 1$	
$t = 0$	n_{00}	n_{01}	n_{0+}
$t = 1$	n_{10}	n_{11}	n_{1+}
	n_{+0}	n_{+1}	

although this notation doesn't tell us though what the sampling scheme was, because it doesn't clearly indicate which margin is fixed.

Davison (SM, Table 10.10) uses slightly different notation and different words:

	$y = 0$ ("failure")	$y = 1$ ("success")	
$t = 0$ ("control")	$m_0 - R_0$	R_0	m_0
$t = 1$ ("case")	$m_1 - R_1$	R_1	m_1
	$m_1 + m_0 - (R_0 + R_1)$	$R_0 + R_1$	

His use of capital letters for random variables indicates that the row margins are fixed, so the sampling is prospective. At the bottom of p.493 he notes that if the column ("horizontal") margins are fixed then the study is retrospective. Both Examples 10.19 and 10.20 are prospective studies. The probability model he uses is $R_1 \sim \text{Binom}(m_1, \pi_1)$, and $R_0 \sim \text{Binom}(m_0, \pi_0)$.[1]

The following famous case-control study, first published by Doll & Hill (1950), compares lung cancer patients ($y = 1$) with matched controls ($y = 0$) on their smoking status. This data is reported in Agresti (2002, p.42).

	Lung cancer	
	1	0
	cases	controls
smoke = 1 (yes)	688	650
smoke = 0 (no)	21	59
	709	709

The cases were all lung cancer patients in a set of twenty hospitals in London, in the year preceding the study. For each case, a non-cancer patient in the same hospital, of the same gender and in the same 5-year age group was chosen as a control. This type of case-control study is called a matched pairs study.

We would like to make inference on $\Pr(\textit{cancer} \mid \textit{smoking})$. But the design provides only information on $\Pr(\textit{smoking} \mid \textit{cancer})$. We could compute the probability of interest if we knew the population prevalence from lung cancer, but we don't have any information on that for this rather special population. (The rates in the general population would not suffice; this haphazard selection of cases and controls is not a random sample from the general population.)

However, the odds ratio, that is the ratio of the odds of cancer among smokers to the odds of cancer among non-smokers can be computed from retrospective data. An application of Bayes' theorem confirms that

$$OR = \frac{\Pr(c = 1 \mid s = 1)/\Pr(c = 0 \mid s = 1)}{\Pr(c = 1 \mid s = 0)/\Pr(c = 0 \mid s = 0)} = \frac{\Pr(s = 1 \mid c = 1)/\Pr(s = 0 \mid c = 1)}{\Pr(s = 1 \mid c = 0)/\Pr(s = 0 \mid c = 0)}$$

[1]Note that Davison (SM, Table 10.10) calls groups with $t = 0$ controls, and $t = 1$ cases, which is a bit confusing when dealing with retrospective studies, which are often case-control studies. It would be clearer if he called them "untreated" and "treated".

and we have the data to estimate the second expression. For the table above it is estimated by

$$\frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 3.0.$$

It can be shown[2] that an estimate of the standard error of $\log(OR)$ is

$$v = \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}\right)^{1/2};$$

the estimated log-odds ratio is more likely to be distributed symmetrically about 0 than is the odds ratio about 1. Thus an approximate 95% confidence interval for $\log(OR)$ for our data is $\log(3.0) \pm 1.96v = 1.10 \pm 1.96 \times 0.26 = (0.589, 1.61)$ leading to a confidence interval for the OR of $(1.8, 5.0)$.

Writing π_1 for $\Pr(cancer \mid smoke = 1)$, and π_0 for $\Pr(cancer \mid smoke = 0)$, we have

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}.$$

The relative risk is defined to be π_1/π_0, i.e. the risk of cancer among smokers relative to the risk of cancer among non-smokers. The RR cannot be estimated from case-control data, but, since

$$\frac{\pi_1}{\pi_0} = OR \times \frac{1-\pi_1}{1-\pi_0}$$

we see that the RR is approximated by the OR if both π_1 and π_0 are small, i.e. if lung cancer is very rare in both groups. Relative risk is easier to interpret than the odds ratio, but the odds ratio is the only quantity estimable from retrospective data.

This can all be framed in the notation of logistic regression, where we might have, in addition to exposure, some additional covariates. We would like to know about $\Pr(Y = 1 \mid \underline{x})$, say, but our case-control data does not provide direct information on this. Let $Z = 1$ if a subject is sampled. Using Bayes' theorem again, we have

$$\Pr(Y = 1 \mid Z = 1, \underline{x}) = \frac{\Pr(Z = 1 \mid Y = 1, \underline{x})\Pr(Y = 1 \mid \underline{x})}{\Pr(Z = 1 \mid Y = 1, \underline{x})\Pr(Y = 1 \mid \underline{x}) + \Pr(Z = 1 \mid Y = 0, \underline{x})\Pr(Y = 0 \mid \underline{x})}.$$

Further, assume $\Pr(Z = 1 \mid Y = 1, \underline{x}) = \Pr(Z = 1 \mid Y = 1) = p_1$, and similarly $\Pr(Z = 1 \mid Y = 0, \underline{x}) = \Pr(Z = 1 \mid Y = 0) = p_0$, i.e. that sampling is not related to the covariates. Then under a logistic regression model for the $\Pr(Y = 1 \mid \underline{x})$ we have

$$\begin{aligned}
\Pr(Y = 1 \mid Z = 1, \underline{x}) &= \frac{p_1 \exp(\underline{x}^T\beta)/\{1 + \exp(\underline{x}^T\beta)\}}{p_1 \exp(\underline{x}^T\beta)/\{1 + \exp(\underline{x}^T\beta)\} + p_0/\{1 + \exp(\underline{x}^T\beta)\}} \\
&= \frac{p_1 \exp(\underline{x}^T\beta)}{p_1 \exp(\underline{x}^T\beta) + p_0} = \frac{\frac{p_1}{p_0}\exp(\underline{x}^T\beta)}{1 + \frac{p_1}{p_0}\exp(\underline{x}^T\beta)} \\
&= \frac{\exp(\underline{x}^T\beta + \alpha^*)}{1 + \exp(\underline{x}^T\beta + \alpha^*)}.
\end{aligned}$$

[2] see end

If the first entry of $\underline{x}$ is a 1, as would usually be the case, then we see that its coefficient cannot be estimated from the data, it is confounded with α^*. However the remaining components of β can be estimated from retrospective data. This is one of the reasons that the logit transform for probabilities is often used with binary or binomial data.

In the smoking example above, there is a single x which is either 1 or 0, so we have

$$\Pr(Y = 1 \mid Z = 1, x = 1) = \frac{e^{\alpha^*+\beta}}{1 + e^{\alpha^*+\beta}}, \quad \Pr(Y = 1 \mid Z = 1, x = 0) = \frac{e^{\alpha^*}}{1 + e^{\alpha^*}};$$

the odds ratio is simply e^{β}.

Logistic regression is applied in Example 10.19 (data in Table 6.8) where the association between smoking and survival is stratified according to age groups. This data is based on a survey of 1314 women, entered into a prospective study in 1972-1974, and followed for twenty years. The age group is the age at entry into the study; twenty years later all the women in the 75+ group had died, as would be expected. In the overall table, 139/582 of the smokers had died, and 230/732 of the non-smokers, suggesting a (misleading) protective effect of smoking. Fitting the model

$$\Pr(Y = 1(survive) \mid Smoke = 1) = \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}}, \quad \Pr(Y = 1 \mid Smoke = 0) = \frac{e^{\alpha}}{1 + e^{\alpha}}$$

gives estimates $\hat{\alpha} = 0.78, \hat{\beta} = 0.38$, with estimated standard errors of 0.08 and 0.13, respectively; note that $\exp(\hat{\beta}) > 1$. The model that uses age information as a factor variable to index groups is

$$\Pr(Y = 1 \mid Smoke = 1, Age = a) = \frac{e^{\alpha_a+\beta}}{1 + e^{\alpha_a+\beta}}; \quad \Pr(Y = 1 \mid Smoke = 0, Age = a) = \frac{e^{\alpha_a}}{1 + e^{\alpha_a}};$$

and under this model the estimate of $\hat{\beta}$ is $-0.43(0.18)$ indicating that after age is accounted for, smoking is associated with increased risk of death.

The same data is analysed in §4.4 of Faraway (ELM-1) (§6.5 of ELM-2), as a set of 2×2 tables. An overall test for association is the Cochran-Mantel-Haenzsel test, and there is an exact calculation of the p-value available using `mantelhaen.test`. This gives a 95% confidence interval for the odds ratio of $(1.0689, 2.2034)$ and point estimate 1.5303. In SM Example 10.19 the approximate 95% CI is $-0.43 \pm 1.96 \times 0.18$; on the odds ratio scale this is $(0.457, 0.926)$, but SM is analysing "alive" rather than "dead", so to compare to ELM we invert these values to get $(1.08, 2.19)$; ELM has $(1.07, 2.20)$, essentially the same.

The CMH parameter of interest is the effect of smoking on survival, aggregated across the 2×2 tables for each age group. The logistic regression test is the same, the effect of smoking on survival, adjusted for age. The only difference is that SM uses likelihood theory, which uses the normal approximation, and the CMH test is based on Fisher's exact test, which uses the hypergeometric distribution.

Approximate variance of the log-odds ratio

There are two ways to calculate this, using the expression for $\log(OR)$ on p.1, or via the logistic regression model version. First, write the generic 2×2 table as

	1 cases	0 controls
exposure = 1 (yes)	n_{11}	n_{10}
exposure = 0 (no)	n_{01}	n_{00}
	n_{+1}	n_{+0}

with associated probabilities $\pi_1 = \pi_{11}$, $\pi_0 = \pi_{10}$, $1 - \pi_1 = \pi_{01}$, $1 - \pi_0 = \pi_{00}$. The odds ratio is $\pi_1(1-\pi_0)/\pi_0(1-\pi_1)$, and the estimate of the log-odds ratio is

$$\hat{\psi} = \log n_{11} - \log n_{01} - \log n_{10} + \log n_{00} = \log \hat{\pi}_1 - \log(1-\hat{\pi}_1) - \log \hat{\pi}_0 + \log(1-\hat{\pi}_0).$$

Using the delta method (SM, Example 2.5), we have that $\mathrm{var}(\log X) \simeq (1/E(X))^2 \mathrm{var}(X)$, so

$$\begin{aligned}
\mathrm{var}(\hat{\psi}) &= \mathrm{var}(\log \hat{\pi}_1) + \mathrm{var}(\log(1-\hat{\pi}_1)) - 2\mathrm{cov}(\log \hat{\pi}_1, \log(1-\hat{\pi}_1)) \\
&\quad + \mathrm{var}(\log \hat{\pi}_0) + \mathrm{var}(\log(1-\hat{\pi}_0)) - 2\mathrm{cov}(\log \hat{\pi}_0, \log(1-\hat{\pi}_0)) \\
&\dot{\sim} \frac{\pi_1(1-\pi_1)}{n_{+1}\pi_1^2} + \frac{(1-\pi_1)\pi_1}{n_{+1}(1-\pi_1)^2} + 2\frac{\pi_1(1-\pi_1)/n_{+1}}{\pi_1(1-\pi_1)} \\
&\quad + \frac{\pi_0(1-\pi_0)}{n_{+0}\pi_0^2} + \frac{(1-\pi_0)\pi_0}{n_{+0}(1-\pi_0)^2} + 2\frac{\pi_0(1-\pi_0)/n_{+0}}{\pi_0(1-\pi_0)} \\
&= \frac{1}{\pi_1(1-\pi_1)n_{+1}} + \frac{1}{\pi_0(1-\pi_0)n_{+0}}; \quad (*)
\end{aligned}$$

on estimating π_1 and π_0 by n_{11}/n_{+1} and n_{00}/n_{+0}, respectively, we have

$$\widehat{\mathrm{var}}(\log \psi) \doteq \frac{n_{+1}}{n_{11}(n_{+1}-n_{11})} + \frac{n_{+0}}{n_{10}(n_{+0}-n_{10})} = \frac{1}{n_{11}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{00}},$$

as claimed. Note that we have assumed the column totals n_{+1} and n_{+0} are the sample sizes for the two binomials, as this is how the lung cancer data was collected. However, the argument above is the same if the row totals are fixed instead, or indeed if only n is fixed and the table is a 4-category multinomial.

Using the logistic representation, we have

$$\frac{\pi_1}{1-\pi_1} = \frac{\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)}, \quad \frac{\pi_0}{1-\pi_0} = \frac{\exp(\alpha)}{1+\exp(\alpha)},$$

giving β as the log-odds ratio; $e^\beta = \psi$ above. The log-likelihood function for (α, β) is a special case of that for logistic regression:

$$\ell(\alpha,\beta) = n_{11}(\alpha+\beta) - n_{+1}\log(1+\exp(\alpha+\beta)) + n_{00}\alpha - n_{+0}\log(1+\exp(\alpha)).$$

The asymptotic variance of $\hat{\beta}$ is given by the $(2,2)$ element of the inverse of the 2×2 observed information matrix $j(\hat{\alpha}, \hat{\beta})$. This entry is $j_{\alpha\alpha}/(j_{\alpha\alpha}j_{\beta\beta} - j_{\alpha\beta}^2)$, where

$$\begin{aligned}
j_{\alpha\alpha} &= -\ell_{\alpha\alpha} = n_{+1}\frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} + n_{+0}\frac{e^{\alpha}}{(1+e^{\alpha})^2} = n_{+1}\pi_1(1-\pi_1) + n_{+0}\pi_0(1-\pi_0), \\
j_{\alpha\beta} &= -\ell_{\alpha\beta} = n_{+1}\frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} = n_{+1}\pi_1(1-\pi_1), \\
j_{\beta\beta} &= -\ell_{\beta\beta} = n_{+1}\frac{e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} = n_{+1}\pi_1(1-\pi_1),
\end{aligned}$$

so

$$\begin{aligned}
(j^{-1})_{2,2} &= \frac{n_{+1}\pi_1(1-\pi_1) + n_{+0}\pi_0(1-\pi_0)}{\{n_{+11}\pi_1(1-\pi_1) + n_{+0}\pi_0(1-\pi_0)\}n_{+1}\pi_1(1-\pi_1) - \{n_{+1}\pi_1(1-\pi_1)\}^2} \\
&= \frac{n_{+1}\pi_1(1-\pi_1) + n_{+0}\pi_0(1-\pi_0)}{n_{+1}\pi_1(1-\pi_1)n_{+0}\pi_0(1-\pi_0)} \\
&= \frac{1}{n_{+1}\pi_1(1-\pi_1)} + \frac{1}{n_{+0}\pi_0(1-\pi_0)},
\end{aligned}$$

which gets us back to the same calculation as at (*).

Reference: Agresti, A. (2002). *Categorical Data Analysis.* John Wiley & Sons, New York.