b8 STATISTICS **HT 2002**

Problem Sheet 10

1. In 1986 the British Medical Journal published the results of a study by Weindling *et al.* comparing the health of juvenile delinquent boys and a non-delinquent control group. The following data relate to a subset of the boys who failed a vision test and show the numbers who did and did not wear spectacles.

		Juvenile delinquents	
		Yes	No
Spectacle wearers	Yes	1	5
	No	8	2

The data are insufficient for a χ^2-test but Fisher's exact test is possible. Are delinquents with poor eyesight more or less likely to wear glasses than non-delinquents with poor eyesight?

2. Consider a 2×2 contingency table from a prospective study in which people in a particular industry who were or were not exposed to a pollutant are followed up and, after several years, categorised according to the presence or absence of a disease. The table below gives the probabilities for each cell.

	Exposed	
	Yes	No
Diseased	p_1	p_2
Not diseased	$1 - p_1$	$1 - p_2$

Define the odds ratio ψ of the disease for the *exposed* and *not exposed* groups.

For the simple logistic model

$$p_i = \frac{e^{\beta_i}}{1 + e^{\beta_i}}$$

show that $\psi = 1$ corresponds to no difference between the exposed and unexposed groups.

Now consider n such 2×2 tables, one for each level x_j of a factor, with $j = 1, \ldots, n$. For the logistic model

$$p_{ij} = \frac{\exp(\alpha_i + \beta_i x_j)}{1 + \exp(\alpha_i + \beta_i x_j)}, \qquad i = 1, 2, \quad j = 1, \ldots, n,$$

show that $\log \psi$ is constant over all tables if $\beta_1 = \beta_2$.

3. Suppose Y_1 and Y_2 are independent binomial random variables with $Y_i \sim B(n_i, p_i)$. Show that $\widehat{p}_i = Y_i/n_i$.

Let φ denote the log odds ratio

$$\varphi = \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) = h(p_2) - h(p_1)$$

where $h(p) = \log(p^{-1} - 1)$. Using the delta method, show that the approximate variance of $\widehat{\varphi}$ is

$$\begin{aligned}&\left(h'(p_1)\right)^2 \frac{p_1(1-p_1)}{n_1} + \left(h'(p_2)\right)^2 \frac{p_2(1-p_2)}{n_2} \\ &= \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}.\end{aligned}$$

If $n_1 = 50$, $n_2 = 100$, $y_1 = 30$, $y_2 = 70$, find an approximate 95% confidence interval for the log odds ratio φ, and hence an approximate 95% confidence interval for the odds ratio itself.

4. Let the random variable Y have a binomial distribution $B(n,p)$ and consider the transformation $\psi(P)$, for some function ψ, where $P = (Y+a)/(n+b)$, and a and b are constants. Use the Taylor expansion of $\psi(P)$ about p along with the approximation

$$\frac{Y+a}{n+b} - p = \frac{1}{n}\left[(Y-np) + (a-bp)\right]\left[1 - \frac{b}{n} + \left(\frac{b}{n}\right)^2 - \cdots\right]$$

to show that

$$E\left[\psi\left(\frac{Y+a}{n+b}\right)\right] = \psi(p) + \frac{\psi'(p)(a-bp)}{n} + \frac{\psi''(p)p(1-p)}{2n} + o\left(\frac{1}{n}\right),$$

$$\operatorname{var}\left[\psi\left(\frac{Y+a}{n+b}\right)\right] = \frac{\left[\psi'(p)\right]^2 p(1-p)}{n} + o\left(\frac{1}{n}\right).$$

For $\psi(t) = \log[t/(1-t)]$ show that, by choosing $a = \frac{1}{2}$ and $b = 1$, the so-called empirical logistic transform $\log[(Y+\frac{1}{2})/(n-Y+\frac{1}{2})]$ is less biased than $\log[Y/(n-Y)]$.