M3S12 BIOSTATISTICS - EXERCISES 2
SOLUTIONS

1. In the usual notation, the *risk difference* is given by

$$RD = P(F|E) - P\left(F|E'\right) = \pi_1 - \pi_0$$

and let $\theta = P(E)$. Now, by an elementary result

$$\begin{aligned} P(F) &= P(F|E)\,P(E) + P\left(F|E'\right)P\left(E'\right) \\ &= P(F|E)\,P(E) + P\left(F|E'\right)(1 - P(E)) \\ &= P(E)\left(P(F|E) - P\left(F|E'\right)\right) + P\left(F|E'\right) \\ &= \theta(\pi_1 - \pi_0) + \pi_0 \end{aligned}$$

Now, $P(F)$ is the incidence rate in the population, and the primary factor that influences decisions about healthcare provision; if $P(F)$ is high, then the need for, say, General Practice or hospital provision is also high. Clearly, from equation above $P(F)$ increases linearly with RD. Now, typically, $P(F|E')$ may be thought to be small, and so a good approximation may be

$$P(F) \approx \theta(\pi_1 - \pi_0).$$

2. Using the following notation for the *observed* counts

	EXPOSURE		
OUTCOME	E	E'	TOTAL
F	n_{11}	n_{12}	$n_{1.}$
F'	n_{21}	n_{22}	$n_{2.}$
TOTAL	$n_{.1}$	$n_{.2}$	$n_{..}$

with notation $\{N_{11}, N_{12}, N_{21}, N_{22}\}$ and $\{N_{.1}, N_{.2}, N_{1.}, N_{2.}, N_{..}\}$ for the corresponding *random variables*, the count data in a cohort study can be thought of as being derived either from independent **binomial** distributions in the columns (assuming *fixed column totals*)

$$N_{11}|N_{.1} = n_{.1} \sim Binomial(n_{.1}, \pi_1) \qquad N_{12}|N_{.2} = n_{.2} \sim Binomial(n_{.2}, \pi_0)$$

or independent **binomial** distributions in the rows (assuming *fixed row totals*)

$$N_{11}|N_{1.} = n_{1.} \sim Binomial(n_{1.}, \gamma_1) \qquad N_{21}|N_{2.} = n_{2.} \sim Binomial(n_{2.}, \gamma_0)$$

or even a **multinomial** distribution across the whole table (assuming *grand total* $N_{..} = n_{..}$). Similarly, the row and column totals have independent binomial distributions conditional on the grand total

$$N_{.1}|N_{..} = n.. \sim Binomial(n_{..}, \theta) \qquad N_{1.}|N_{..} = n.. \sim Binomial(n_{..}, \phi).$$

Hence maximum likelihood estimation for a binomial problem is key. Suppose, in general

$$X \sim Binomial(n, p).$$

Then the likelihood function after observing $X = x$ is given by

$$L(p) = f_X(x;p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad 0 < p < 1$$

and hence

$$\log L(p) = c(x,n) + x \log p + (n-x)\log(1-p) \qquad \text{where } c(x,n) = \log \binom{n}{x}.$$

Thus

$$\frac{d}{dp}\{\log L(p)\} = \frac{x}{p} - \frac{n-x}{1-p} \qquad \therefore \frac{d}{dp}\{\log L(p)\} = 0 \Longrightarrow \widehat{p} = \frac{x}{n}$$

and

$$\frac{d^2}{dp^2}\{\log L(p)\} = -\left[\frac{x}{p^2} + \frac{n-x}{(1-p)^2}\right] < 0 \qquad \text{for all } p$$

verifies the maximum. Hence the results follow;

$$\widehat{\pi}_1 = \frac{n_{11}}{n_{.1}} \qquad \widehat{\pi}_0 = \frac{n_{12}}{n_{.2}} \qquad \widehat{\theta} = \frac{n_{1.}}{n_{..}} \qquad \phi = \frac{n_{.1}}{n_{..}}$$

3. Using the following formulae, all the quantities can be evaluated;

Parameter	Notation	Estimate
Risk Difference	$\pi_1 - \pi_0$	$\frac{n_{11}}{n_{.1}} - \frac{n_{12}}{n_{.2}}$
Relative Risk/Risk Ratio	$\frac{\pi_1}{\pi_0}$	$\frac{n_{11}/n_{.1}}{n_{12}/n_{.2}}$
Odds Ratio	$\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$	$\frac{n_{11}n_{22}}{n_{12}n_{21}}$

(1)

For the numerical calculations, see the attached SPLUS output sheet.

4. To compare the results for the two levels of the risk factor E, and the risk factor (or potential *confounder* X), we can inspect either the relative risk or the odds ratio as given in (1) across different levels of either variable. For example: let

$$RR^{(x)} = \frac{\pi_1^{(x)}}{\pi_0^{(x)}}$$

denote the relative risk for $X = x \in \{0,1\}$, and let then consider

$$\frac{\pi_1^{(0)}}{\pi_1^{(1)}} \qquad \text{or} \qquad \frac{\pi_0^{(0)}}{\pi_0^{(1)}}$$

to compare the different levels of risk factor X for the same exposure status. To uncover the relationship between X and E, it is legitimate to pool across the outcome variable F in the two tables. For example

(i) In the exposed group E

$$X = 0 : \widehat{\pi}_1^{(0)} = \frac{70}{70+30} = \frac{70}{100} = 0.7$$

$$X = 1 : \widehat{\pi}_1^{(1)} = \frac{160}{160+40} = \frac{160}{200} = 0.8$$

so that

$$\frac{\widehat{\pi}_1^{(0)}}{\widehat{\pi}_1^{(1)}} = \frac{0.7}{0.8} = 0.875$$

which indicates that $X = 1$ increases incidence in the exposed group, and hence is a risk factor. However, for the unexposed group E'

$$X = 0 : \widehat{\pi}_0^{(0)} = \frac{80}{80+120} = \frac{80}{200} = 0.4$$

$$X = 1 : \widehat{\pi}_0^{(1)} = \frac{40}{40+60} = \frac{40}{100} = 0.4$$

and

$$\frac{\widehat{\pi}_0^{(0)}}{\widehat{\pi}_0^{(1)}} = \frac{0.4}{0.4} = 1$$

which indicates that there is no difference in risk for the two levels of X in the unexposed group. Thus the conclusion is different for E and E', so it appears that there is an interaction between X and E in terms of effect size, that is, there is *effect moderation.*

(ii) In the exposed group E

$$X = 0 : \widehat{\pi}_1^{(0)} = \frac{90}{90+10} = \frac{90}{100} = 0.9$$

$$X = 1 : \widehat{\pi}_1^{(1)} = \frac{80}{80+120} = \frac{80}{200} = 0.4$$

so that

$$\frac{\widehat{\pi}_1^{(0)}}{\widehat{\pi}_1^{(1)}} = \frac{0.9}{0.4} = 2.25$$

which indicates that $X = 1$ decreases incidence in the exposed group, and hence is a risk factor. For the unexposed group E'

$$X = 0 : \widehat{\pi}_0^{(0)} = \frac{60}{40+60} = \frac{60}{100} = 0.6$$

$$X = 1 : \widehat{\pi}_0^{(1)} = \frac{20}{20+180} = \frac{20}{200} = 0.1$$

and

$$\frac{\widehat{\pi}_0^{(0)}}{\widehat{\pi}_0^{(1)}} = \frac{0.6}{0.1} = 6$$

which indicates that $X = 1$ decreases incidence in the unexposed group, and hence is a risk factor. Thus although it appears that there is an interaction between X and E in terms of effect size, the effect is in the same direction in both cases. Hence we can conclude that $X = 1$ decreases incidence.

To test whether X and E are related, can look at the odds ratio in pooled tables

(i)
	E	E'
$X=0$	100	200
$X=1$	200	100

$$\Longrightarrow \widehat{\psi} = \frac{100 \times 100}{200 \times 200} = \frac{1}{4} \Longrightarrow X, E \text{ are related}$$

(ii)
	E	E'
$X=0$	100	100
$X=1$	200	200

$$\Longrightarrow \widehat{\psi} = \frac{100 \times 200}{100 \times 200} = 1 \Longrightarrow X, E \text{ are not related}$$

See the attached SPLUS sheet for full computational details.

5. From notes ML estimate of

$$\frac{P(E|F)}{P(E'|F)} = \frac{\gamma_1}{1-\gamma_1}$$

is n_{11}/n_{12} Now, by Bayes Theorem

$$\frac{P(E|F)}{P(E'|F)} = \frac{P(F|E)}{P(F|E')}\frac{P(E)}{P(E')} \Longrightarrow \frac{P(F|E)}{P(F|E')} = \frac{P(E|F)}{P(E'|F)}\frac{P(E')}{P(E)} = \frac{\gamma_1}{1-\gamma_1}\frac{1-\theta}{\theta}$$

but it is known that $\theta = 0.4$, so

$$\frac{\pi_1}{\pi_0} = \frac{P(F|E)}{P(F|E')} = \frac{\gamma_1}{1-\gamma_1}\frac{0.6}{0.4} = \frac{3}{2}\frac{\gamma_1}{1-\gamma_1}$$

and by similar calculations

$$\frac{1-\pi_1}{1-\pi_0} = \frac{P(F'|E)}{P(F'|E')} = \frac{P(E|F')}{P(E'|F')}\frac{P(E')}{P(E)} = \frac{\gamma_0}{1-\gamma_0}\frac{0.6}{0.4} = \frac{3}{2}\frac{\gamma_0}{1-\gamma_0}.$$

(i) In the case control study, we have

$$\widehat{\gamma}_1 = \frac{n_{11}}{n_{1.}} \qquad \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} = \frac{n_{11}/n_{1.}}{n_{12}/n_{1.}} = \frac{n_{11}}{n_{12}}$$

and similarly

$$\widehat{\gamma}_0 = \frac{n_{21}}{n_{2.}} \qquad \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} = \frac{n_{21}/n_{2.}}{n_{22}/n_{2.}} = \frac{n_{21}}{n_{22}}.$$

The usual arguments say that the relative risk, π_1/π_0 cannot be estimated from a case control study. However, in this case, the estimation is possible, as the exposure rate θ is **known** From above, we have two equations

$$\frac{\pi_1}{\pi_0} = \frac{\gamma_1}{1-\gamma_1} \times \frac{3}{2} = k_1 \Longrightarrow \pi_1 = k_1\pi_0 \tag{2}$$

$$\frac{1-\pi_1}{1-\pi_0} = \frac{\gamma_0}{1-\gamma_0} \times \frac{3}{2} = k_0 \Longrightarrow 1-\pi_1 = k_1(1-\pi_0) \tag{3}$$

say, for two constants k_0, k_1 that are estimable from the data. Rearranging these formulae, noting that

$$\pi_1 + (1-\pi_1) = 1$$

we have that $k_1\pi_0 + k_1(1-\pi_0) = 1$, so that

$$\pi_0 = \frac{1-k_0}{k_1-k_0} \qquad \pi_1 = \frac{k_1(1-k_0)}{k_1-k_0} \tag{4}$$

Form the data we have that

$$k_1 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1}\frac{3}{2} = \frac{96}{104}\times\frac{3}{2} = 1.385$$

$$k_0 = \frac{\widehat{\gamma}_0}{1-\widehat{\gamma}_0}\frac{3}{2} = \frac{109}{666}\times\frac{3}{2} = 0.245$$

and hence, from (4),

$$\widehat{\pi}_1 = 0.917 \qquad \widehat{\pi}_0 = 0.662.$$

The standard errors are in theory available from (4), as the estimators of

$$\frac{\gamma_1}{1-\gamma_1} \qquad \frac{\gamma_0}{1-\gamma_0}$$

are merely functions of the data, but are complicated. Simulation-based methods can be used (in particular, the bootstrap; we will touch on this approach during the course). The standard error for the relative risk π_1/π_0 is also complex, but on the log scale the formulae from notes can be used, that is for $\log\dfrac{\pi_1}{\pi_0}$, estimate is

$$\log\left(\frac{n_{11}}{n_{12}}\times\frac{3}{2}\right) = \log\left(\frac{96}{104}\times\frac{3}{2}\right) = 0.325$$

and standard error is computed in a manner identical for the standard error for the odds ratio. Let

$$g(t) = \log\left(\frac{t}{1-t}\right) + \log\left(\frac{3}{2}\right) \Longrightarrow g'(t) = \frac{1}{t\,(1-t)}$$

and thus

$$\log\left(\frac{\gamma_1}{1-\gamma_1}\times\frac{3}{2}\right) \text{ has standard error } \sqrt{\frac{1}{n_{1.}\left(\frac{n_{11}}{n_{1.}}\right)\left(\frac{n_{12}}{n_{1.}}\right)}} = \sqrt{\frac{n_{1.}}{n_{11}n_{12}}} = \sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}}$$

thus $\log\dfrac{\pi_1}{\pi_0}$ has estimated standard error

$$\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}} = \sqrt{\frac{1}{96}+\frac{1}{104}} = 0.142.$$

(ii) From above

$$\widehat{\delta} = \widehat{\pi}_1 - \widehat{\pi}_0 = 0.255$$

Again, the standard error is not readily available, but on the log scale, the standard error of difference can be estimated

(iii) For the odds ratio, formulae from notes say $\log\psi$ has estimate and standard error

$$\log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) = 1.730 \qquad \sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} = 0.175$$

so that an approximate 95% confidence interval for $\log\psi$ is

$$1.730 \pm 2\times 0.175 = (1.380, 2.080)$$

that is, an interval that does not contain zero, so the log odds ratio is significantly different from zero, and the direction of effect indicates that the odds on disease is greater in the exposed group. We can exponentiate this interval to get a 95% confidence interval for ψ, that is

$$(3.974, 8.00)$$