ST520: Statistical Principles of Clinical Trials HW 1 Solutions

Problem 1

Throughout this problem, we will define D as the indicator of esophageal cancer (1 = has cancer, 0 = does not have cancer), and E as the indicator of heavy drinking (1 = heavy drinker, 0 = light drinker).

(a) No, we cannot estimate the proportion of heavy drinkers in the French population. Because we want $P(E=1) = P(E=1|D=0)P(D=0) + P(E=1|D=1)P(D=1)$ and we cannot estimate $P(D=0)$ and $P(D=1)$ from the data, by the nature of the study, we cannot estimate $P(E=1)$.

(b) No, we cannot estimate the overall prevalence of the disease. Because this is a case-control study, we can't estimate the prevalence of the disease in France.

(c) No, we cannot. Because this is a case-control study, we can't estimate $P(D=1|E=1)$, which is the prevalence of disease among heavy drinkers.

(d) No, we cannot. Because this is a case-control study, we can't estimate $P(D=1|E=0)$, which is the prevalence of disease among light drinkers.

(e) No, we cannot. The probabilities $P(D=1|E=1)$ and $P(D=1|E=0)$ cannot be estimated in this case, so their ratio also cannot be estimated.

(f) Yes, we can. The estimates of odds of having cancer can obtained from the estimates of conditional probabilities of exposure, $P(E=1|D=1)$ and $P(E=1|D=0)$, and we can take their ratio. The estimate is given by $\frac{\frac{96/200}{1-96/200}}{\frac{109/775}{1-109/775}} = \frac{(96)(666)}{(109)(104)} = \boxed{5.640}$. This means the odds of having esophageal cancer are 5.64 times higher for heavy drinkers than for light drinkers. Assuming that the disease is rare, we can estimate that the relative risk is also about 5.6.

(g)

$\hat{Var}(\hat{\theta}) = \hat{\theta}^2(1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}) = 5.640^2 * (1/96 + 1/109 + 1/104 + 1/666)$

=0.977

CI: $\hat{\theta} \pm z_{\alpha/2}[\hat{Var}(\hat{\theta})]^{1/2} = 5.640 \pm 1.96 * 0.977$= $[3.686, 7.594]$

From the confidence interval, we can conclude that the odds of having the esophageal cancer for heavy drinkers is significantly higher than the odds for light drinkers.

Problem 2

Throughout this problem, we will define D as the indicator of Heart Attach, and E as the indicator of Coffee Drinker.

(a) Relative risk: $\psi = \dfrac{P(D|E)}{P(D|\bar{E})} = \dfrac{0.006/(0.006+0.495)}{0.004/(0.004+0.495)} = 1.494$

Odds-ratio: $\theta = \dfrac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))} = \dfrac{0.006/0.495}{0.004/0.496} = 1.5$

(b) $P(E|D) = \dfrac{0.006}{0.006+0.004} = 0.6$

$P(E|\bar{D}) = \dfrac{0.495}{0.495+0.495} = 0.5$

Odds-ratio: $\theta' = \dfrac{P(E|D)/(1-P(E|D))}{P(E|\bar{D})/(1-P(E|\bar{D}))} = \dfrac{0.6/0.4}{0.5/0.5} = 1.5$

Yes. The odds ratio is the same

(c) $n_{+1} = n_{+2} = 400/(1+4) = 80$. She can have 80 cases and controls in her study given her budget constraint.

$n_{11} = n_{+1} * P(E|D) = 80 * 0.6 = 48$

$n_{12} = n_{+2} * P(E|\bar{D}) = 80 * 0.5 = 40$

	Heart arrack	No hear attack
Coffee Drinker	48	40
None coffee drinker	32	40

$v\hat{a}r(log(\hat{\theta})) = 1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}$=0.102

(d) Let the case sample size=x, control sample size=400-4x.

$n_{11} = 0.6x, n_{12} = 0.4x, n_{21} = 0.5(400-4x) = 200-2x, n_{22} = 200-2x$

$v\hat{a}r(\hat{\theta}) = 1/0.6x + 1/0.4x + 2/(200 - 2x) = 25/(6x) + 1/(100 - x) = f(x)$

$f'(x) = -25/6 * x^{-2} + 1/(100 - x)^2$

Set $f'(x) = 0$, which leads to x=67.12

$f''(x) = 50/6 * x^{-3} + 2/(100 - x)^3 > 0$ for 0<x<100.

So when x=67.12, f(x) takes the minimum. However, x can only take integers, consider 67 and 68.

$f(67) = 0.09249209, f(68) = 0.09252451$

The optimal design is the case sample size=67, and the control sample size=132. The variance for $log(\hat{\theta})$ is 0.0925, which is smaller than that in (c).

Problem 3

(a) The sample sizes of the exposure and non-exposure groups are equal to n, so the expected sizes of the four groups are given by

$$\begin{aligned}
E(n_{11}) = \quad & P(D=1|E=1)n = \quad & \frac{0.05}{0.05+0.35}n = \quad & 0.125n \\
E(n_{12}) = \quad & P(D=0|E=1)n = \quad & \frac{0.35}{0.05+0.35}n = \quad & 0.875n \\
E(n_{21}) = \quad & P(D=1|E=0)n = \quad & \frac{0.02}{0.02+0.58}n = \quad & 0.033n \\
E(n_{22}) = \quad & P(D=0|E=0)n = \quad & \frac{0.58}{0.02+0.58}n = \quad & 0.967n
\end{aligned}$$

$$\begin{aligned}
\hat{var}\{\log(\hat{\theta})\} &= \frac{1}{0.125n} + \frac{1}{0.875n} + \frac{1}{0.033n} + \frac{1}{0.967n} \\
&= \frac{40.177}{n}
\end{aligned}$$

(b) The sample sizes of the case and control groups are equal to n, so the expected sizes of the four groups are given by

$$\begin{aligned}
E(n_{11}) = \quad & P(E=1|D=1)n = \quad & \frac{0.05}{0.05+0.02}n \\
E(n_{21}) = \quad & P(E=0|D=1)n = \quad & \frac{0.02}{0.02+0.05}n \\
E(n_{12}) = \quad & P(E=1|D=0)n = \quad & \frac{0.35}{0.35+0.58}n \\
E(n_{22}) = \quad & P(E=0|D=0)n = \quad & \frac{0.58}{0.35+0.58}n
\end{aligned}$$

The variance estimate of $\log(\hat{\theta})$ is then

$$\hat{var}\{log(\hat{\theta})\} = 9.16/n$$

(c)The ratio of variances is 40.177/9.16=4.386. This means that the case-control study is more efficient, because its expected variance is lower.