Solutions theoretical exercises for STK4900/9900.

Exercise 7

a) With $n_M = 7180$ the number of men interviewed and $X_M = 1630$ classified as binge drinkers we estimate the proportion of binge drinkers among men p_M to be $\hat{p}_M = X_M/n_M = 1630/7180 = 0.227$. Similarly the estimated proportion female binge drinkers equals $\hat{p}_F = X_F/n_F = 1684/9916 = 0.170$.

The 95% confidence intervals for the true proportions p_M and p_F becomes

$$\hat{p}_M \pm 1.96\sqrt{\tfrac{\hat{p}_M(1-\hat{p}_M)}{n_M}} = (0.217, 0.237)$$
$$\hat{p}_F \pm 1.96\sqrt{\tfrac{\hat{p}_F(1-\hat{p}_F)}{n_F}} = (0.162, 0.177)$$

b) The risk difference $p_M - p_F$ is estimated as $\hat{p}_M - \hat{p}_F = 0.227 - 0.170 = 0.057$. The 95% confidence interval for the risk difference is given as

$$\hat{p}_M - \hat{p}_F \pm 1.96\sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{n_M} + \frac{\hat{p}_F(1-\hat{p}_F)}{n_F}} = (0.045, 0.069)$$

Since this interval does not contain the value zero we can conclude that the proportions among men and women are significantly different,

c) More formally we test the null hypothesis $\mathrm{H}_0 : p_M = p_F$ vs. alternative $\mathrm{H}_0 : p_M \neq p_F$ by the test statistic

$$Z = \frac{\hat{p}_M - \hat{p}_F}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_M} + \frac{\hat{p}(1-\hat{p})}{n_F}}}$$

where $\hat{p} = (X_M + X_F)/(n_M + n_F) = 0.199$ is the estimate of the common proportion under the null hypothesis. This test statistic will be approximately standard normally distributed under the null.

Plugging in the data we observe $Z = 9.34$, this corresponds to a very small p-value (from R 10^{-20}).

d) The full 2x2 table over men/women and binge drinking becomes

	Freq. binge drinkers	Not freq. binge dr.	Total
Males	1630	5550	7180
Women	1684	8232	9916
Total	3314	13782	17096

e) With $T_{i\bullet}$ the total in row i, $T_{\bullet j}$ the total in column j and $T_{\bullet\bullet} = 17096$ the overall total of the 2x2 table we get the expected values in cell (i, j) as $E_{ij} = T_{\bullet j}T_{i\bullet}/T_{\bullet\bullet}$.

Perhaps simpler we get $E_{11} = \hat{p}n_M$, $E_{12} = (1-\hat{p})n_M$, $E_{21} = \hat{p}n_F$ and $E_{22} = (1-\hat{p})n_F$. Doing the calculation the 2x2 matrix of expected values becomes

	Freq. binge drinkers	Not freq. binge dr.	Total
Males	1391.8	5788.2	7180
Women	1922.2	7993.8	9916
Total	3314	13782	17096

e) The (Pearson) chi-square statistic is given as

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} are the numbers in the 2x2 table of the observations and the sum is taken over all cells in the 2x2 tables.

Under the null hypothesis $\mathrm{H}_0 : p_M = p_F$ this statistic follows a chi-square distribution with 1 degree of freedom (since 2x2 table). We reject with large values of X^2.

Here we get $X^2 = 87.17$ which correspond to a tiny p-value. In fact it becomes 10^{-20} (from R) just as the p-value in question c). This correspond to the fact that $87.17 = 9.34^2$ where 9.34 was test statistic from question c).

We actually have the algebraic identity $X^2 = Z^2$ where X^2 is the chi-square statistic and Z is the standard normal statistic (for all such 2x2 tables).