STA 100 Practice Final Solution

Yidong Zhou

1. Let K and G denote the events that a randomly selected person develops kidney cancer and has the gene, separately. It follows that

$$P(G) = 0.5, \quad P(K|G) = 0.3, \quad P(K|G^C) = 0.015.$$

(a)

$$\begin{aligned} P(K) &= P(K \cap G) + P(K \cap G^C) \\ &= P(K|G)P(G) + P(K|G^C)P(G^C) \\ &= 0.3 \times 0.5 + 0.01 \times (1 - 0.5) \\ &= 0.15 + 0.005 \\ &= 0.155. \end{aligned}$$

(b)

$$\begin{aligned} P(G|K) &= \frac{P(G \cap K)}{P(K)} \\ &= \frac{P(K|G)P(G)}{P(K)} \\ &= \frac{0.3 \times 0.5}{0.155} \\ &= \frac{0.15}{0.155} \\ &= 0.9677. \end{aligned}$$

(c)

$$\begin{aligned} P(G \cup K) &= P(G) + P(K) - P(G \cap K) \\ &= P(G) + P(K) - P(K|G)P(G) \\ &= 0.5 + 0.155 - 0.3 \times 0.5 \\ &= 0.5 + 0.155 - 0.15 \\ &= 0.505. \end{aligned}$$

(d)

$$\begin{aligned} P(G|K^C) &= \frac{P(G \cap K^C)}{P(K^C)} \\ &= \frac{P(G) - P(G \cap K)}{1 - P(K)} \\ &= \frac{P(G) - P(K|G)P(G)}{1 - P(K)} \\ &= \frac{0.5 - 0.3 \times 0.5}{1 - 0.155} \\ &= 0.4142. \end{aligned}$$

(e) $Y \sim B(100, 0.5)$ can be approximated by $N(50, 25)$. Using continuity correction, one has

$$\begin{aligned} P(Y \geq 50) &= P(Y > 49.5) \\ &= P(Z > \frac{49.5 - 50}{\sqrt{25}}) \\ &= P(Z > -0.1) \\ &= 1 - P(Z \leq -0.1) \\ &= 1 - 0.4602 \\ &= 0.5398. \end{aligned}$$

2. (a) $H_0 : \mu_1 - \mu_2 \geq 0$ v.s. $H_A : \mu_1 - \mu_2 < 0$.

(b) Here we have $n_1 = 50$, $n_2 = 70$, $\nu = 100$, $\bar{Y}_1 = 490$, $\bar{Y}_2 = 500$, $s_1 = 32$, and $s_2 = 48$. The test statistic is thus

$$\begin{aligned} T &= \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\text{SE}_{\bar{Y}_1 - \bar{Y}_2}} \\ &= \frac{(490 - 500) - 0}{\sqrt{32^2/50 + 48^2/70}} \\ &= -1.3685. \end{aligned}$$

(c) From t table with df $= 100$, we find that $P(t_{100} > 1.290) = 0.10$ and $P(t_{100} > 1.660) = 0.05$. The range of p-value is thus $(0.05, 0.10)$.

(d) Since p-value $> \alpha = 0.05$, we fail to reject the null at the 0.05 level of significance.

(e) Yes, the upper one-sided 95% confidence interval for the difference in costs would include zero since we fail to reject the null for the left-sided t test.

3. (a) H_0 : The incidence of CHD and smoking are independent.
H_A: The incidence of CHD and smoking are dependent.

(b) See the following table for the row, column, and grand totals.

	CHD	No CHD	Total
Smoked	84	87	171
Did not smoke	296	491	787
Total	380	578	958

The expected frequencies are

$$e_1 = \frac{171 \times 380}{958} = 67.83, \quad e_2 = \frac{171 \times 578}{958} = 103.17,$$

$$e_3 = \frac{787 \times 380}{958} = 312.17, \quad e_4 = \frac{787 \times 578}{958} = 474.83.$$

Therefore, the test statistic is

$$\begin{aligned} T &= \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{(84 - 67.83)^2}{67.83} + \frac{(87 - 103.17)^2}{103.17} + \frac{(296 - 312.17)^2}{312.17} + \frac{(491 - 474.83)^2}{474.83} \\ &= 7.78. \end{aligned}$$

(c) The null distribution for the test statistic is χ^2_1. The critical value for $\alpha = 0.05$ is thus $\chi^2_1(0.05) = 3.84$. Since the test statistic is greater than the critical value, we reject the null at the 0.05 level of significance.

(d) The Wilson-adjusted sample proportions are

$$\tilde{p}_1 = \frac{84+1}{171+2} = 0.4913, \quad \tilde{p}_2 = \frac{296+1}{787+2} = 0.3764.$$

The standard error for $\tilde{p}_1 - \tilde{p}_2$ is

$$\text{SE}_{\tilde{p}_1-\tilde{p}_2} = \sqrt{\frac{0.4913 \times (1-0.4913)}{171+2} + \frac{0.3764 \times (1-0.3764)}{787+2}} = 0.04174.$$

The 95% confidence interval for $p_1 - p_2$ is thus

$$(0.4913 - 0.3764) \pm 1.96 \times 0.04174$$

or $(0.0331, 0.1967)$.

(e) Yes, the confidence interval is consistent with the conclusion in (c) since it does not include zero.

4. (a) Here we have $I = 3, n = \sum_{i=1}^{3} n_i = 63$, and

$$\begin{aligned}\bar{Y} &= \frac{\sum_{i=1}^{3} n_i \bar{Y}_i}{n} \\ &= \frac{21 \times 8 + 21 \times 8 + 21 \times 5}{63} \\ &= \frac{441}{63} \\ &= 7.\end{aligned}$$

It follows that

$$\begin{aligned}\text{SSB} &= \sum_{i=1}^{3} n_i (\bar{Y}_i - \bar{Y})^2 \\ &= 21 \times (8-7)^2 + 21 \times (8-7)^2 + 21 \times (5-7)^2 \\ &= 126.\end{aligned}$$

and

$$\begin{aligned}\text{SSW} &= \sum_{i=1}^{3} (n_i - 1) s_i^2 \\ &= (21-1) \times 3^2 + (21-1) \times 3^2 + (21-1) \times 3^2 \\ &= 540.\end{aligned}$$

Therefore,

$$\text{MSB} = \frac{\text{SSB}}{3-1} = 63, \quad \text{MSW} = \frac{\text{SSW}}{63-3} = 9.$$

(b) $H_0 : \mu_1 = \mu_2 = \mu_3$ v.s. H_A : The μ_i's are not all equal.

(c) The test statistic is

$$T = \frac{\text{MSB}}{\text{MSW}} = \frac{63}{9} = 7.$$

(d) The null distribution of the test statistic is $F_{2,60}$. From F table with numerator df = 2 and denominator df = 60, we find that $P(F_{2,60} > 4.98) = 0.01$ and $P(F_{2,60} > 7.77) = 0.001$. The range of p-value is thus $(0.001, 0.01)$.

(e) Since the p-value $< \alpha = 0.01$, we reject the null at the 0.01 level of significance.

5. Here the response variable is the tooth length and the explanatory variable is the amount of vitamin C. It follows that $n = 60, \bar{X} = 1.17, \bar{Y} = 18.81, s_X = 0.63, s_Y = 7.65$.

   (a) The slope is
   $$b_1 = r\frac{s_Y}{s_X} = 0.803 \times \frac{7.65}{0.63} = 9.7507.$$
   The intercept is
   $$b_0 = \bar{Y} - b_1\bar{X} = 18.81 - 9.7507 \times 1.17 = 7.4017.$$
   The fitted regression line is thus
   $$\hat{Y} = 7.4017 + 9.7507X.$$

   (b) The prediction based on the fitted regression line is
   $$\hat{Y} = 7.4017 + 9.7507 \times 2 = 26.9031.$$
   The prediction error is thus
   $$e = \hat{Y} - Y = 29.4 - 26.9031 = 2.4969.$$
   It follows that we under estimate the actual value.

   (c) The residual standard deviation is
   $$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{1227.905}{60-2}} = 4.6012.$$
   The typical error when using the fitted regression line to predict tooth length is 4.6012 mm.

   (d) The coefficient of determination is
   $$r^2 = 0.803^2 = 0.6448.$$
   The proportion of the variance in tooth length that is explained by the linear relationship between tooth length and the amount of vitamin C is 0.6448.

   (e) The standard error for b_1 is
   $$\text{SE}_{b_1} = \frac{s_e}{s_X\sqrt{n-1}} = \frac{4.6012}{0.63 \times \sqrt{60-1}} = 0.9508.$$
   The t multiplier is $t_{60-2}(0.05/2) \approx t_{50}(0.025) = 2.009$. The 95% confidence interval for the slope is thus
   $$9.7507 \pm 2.009 \times 0.9508$$
   or $(7.8405, 11.6609)$. We are 95% confident that when the dosage of vitamin C increases by 1 mg, we expect the tooth length to increase by between 7.8405 mm and 11.6609 mm on average.