SUTD ESD

Statistics 2018

Assignment 4 Solutions

Refer to the spreadsheet for calculations.

Question 1.

(a) Using the table below,

d_i	-4	0	12	18	-4	-12	8	16
Sign(d_i)	$-$		$+$	$+$	$-$	$-$	$+$	$+$

$s_+ = 4$ and $s_- = 3$. Then the exact P-value is

$$P = 2P(S \leq s_{\min}) = 2P(S \leq 3) = 2 \times 0.5 = 1.0$$

Since $P > \alpha = 0.05$, do not reject H_0 and conclude that glaucoma does not affect corneal thickness.

(b) Using the table below,

d_i	-4	0	12	18	-4	-12	8	16
r_i	1.5		4.5	7	1.5	4.5	3	6

$$w_+ = 4.5 + 7 + 3 + 6 = 20.5$$

As covered in class w_+ is normally distributed with $\mu = \frac{n(n+1)}{4} = \frac{7\cdot 8}{4} = 14$ and $\sigma^2 = \frac{n(n+1)(2n+1)}{24} = \frac{7\cdot 8\cdot 15}{24} = 35$. Computing the z-score (using continuity correction)

$$z = \frac{w_+ - 0.5 - 14}{\sqrt{35}} = 1.014.$$

The p-value is $2p(z > 1.014) = 2 \cdot 0.155 = 0.310$. Fail to reject the null. See Excel for details.
In this case sign test and signed rank test give the same result. But in general the results might differ, since the signed rank test is using the magnitudes of the differences, as opposed to just the sign of the difference for the sign test. This could happen for example when the distribution is not symmetric.

Question 2. If MST = MSA + MSE, then we would have

$$\frac{\text{SST}}{N-1} = \frac{\text{SSA}}{k-1} + \frac{\text{SSE}}{N-k} = \frac{\text{SSA} + \text{SSE}}{(k-1)+(N-k)}.$$

Therefore, after renaming the variables, we are asking if it is possible to have

$$\frac{a}{c} + \frac{b}{d} = \frac{a+b}{c+d}.$$

After multiplying both sides of this equation by $c+d$, we get

$$a \times \frac{d}{c} + b \times \frac{c}{d} = 0,$$

which is *impossible* since all the quantities involved are positive. (The only exception is when $a = b = 0$, i. e. SSA = SSE = 0, which does not occur in practice.)

Question 3. (a) F is much larger than the critical value – in fact the p-value is 4×10^{-11}; hence this is very strong evidence that the mean salinity at the three sites are different.

(b) We first compute $\bar{\bar{y}} = (1118.66 + X)/30$, $\bar{y}_1 = (408.91 + X)/12$, $\bar{y}_2 = 40.10375$, and $\bar{y}_3 = 38.892$.

Now SSA = 38.8008825 from the table; on the other hand, SSA $= 12(\bar{y}_1 - \bar{\bar{y}})^2 + 8(\bar{y}_2 - \bar{\bar{y}})^2 + 10(\bar{y}_3 - \bar{\bar{y}})^2$ by the formula. After making the substitutions, we obtain a quadratic in X which has two solutions, 38.85 and 89.66.

We can check that the larger solution is invalid (e. g. by constructing an ANOVA table based on that solution), hence we must have $X = 38.85$.

It is also possible to solve for X using SSE or SST. You can use *Excel*'s Solver or Goal Seek function (see spreadsheet for Goal Seek using SSE); however, *Excel* may return the wrong root of the quadratic.

Question 4. See spreadsheet. Here $\alpha = 0.05$, $m = \binom{3}{2} = 3$, $s^2 = 3.996$, and the cut-off difference, computed using the formula, is 1.559.

We conclude that shelf 2 is significantly different from shelf 1, and also from shelf 3. However, there is not enough evidence to show that shelf 1 and shelf 3 are different.

Question 5. (a) See *Excel* for the interaction plot. Since the two lines are almost parallel, there does not appear to be any interaction.

(b) An ANOVA table is produced using *Excel*'s Anova: Two-Factor With Replication option, with 8 as the 'Rows per sample'. The p-value for F_{AB} is 0.93, which is greater than any α we would pick, thus confirming our answer to part (a) (that we cannot reject the hypothesis of no interaction).

The p-value for F_A is 0.0098, and the p-value for F_B is 0.00059. Thus we can conclude that IQ seems to depend on both the biological and the adoptive parents' socioeconomic status, if α is set at a reasonable value, such as 0.05 or 0.01.

Question 6.

Since

$$f(x|p) = P(X = x|p) = p(1-p)^{x-1},$$

then the log likelihood is

$$\ln L(p) = \ln p + (x-1)\ln(1-p).$$

To find the MLE, set

$$\frac{d \ln L(p)}{dp} = \frac{1}{p} - \frac{x-1}{1-p} = 0,$$

and solve for p, which yields

$$\hat{p} = \frac{1}{x}.$$

Question 7.

(a) If the prior distribution is

$$\pi(\theta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta} \text{ for } \theta > 0,$$

then the posterior distribution is

$$\pi^*(\theta|x) = \frac{f(x_1,\ldots,x_n|\theta)\pi(\theta)}{\int_0^1 f(x_1,\ldots,x_n|\theta)\pi(\theta)\,d\theta}.$$

The numerator is

$$\begin{aligned} f(x_1,\ldots,x_n|\theta)\pi(\theta) &= \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod x_i!}\frac{1}{\Gamma(\alpha)}\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta} \\ &= \theta^{\sum x_i+\alpha-1}e^{-\theta(n+\beta)}\frac{\beta^{\alpha}}{\prod x_i!\Gamma(\alpha)} \\ &= \theta^{\alpha'-1}e^{-\theta\beta'}\frac{\beta^{\alpha}}{\prod x_i!\Gamma(\alpha)}, \end{aligned}$$

where $\alpha' = \alpha + \sum x_i$ and $\beta' = n + \beta$.

Therefore the posterior distribution is proportionate to $\theta^{\alpha'-1}e^{-\beta'\theta}$. I.e. posterior is Gamma$(\alpha + \sum_i x_i, \beta + n)$.

(b) The posterior mean of θ primarily depends on the prior information if the sample size is small. As more and more data are collected, $E(\theta|x) \Rightarrow \bar{x}$, so that the prior distribution and posterior mean are overwhelmed by the data.