M1S : ASSESSED COURSEWORK 2 : SOLUTIONS

This is a problem based on the Hypergeometric formula with $N = 200, n = 20, r = 17$ and R unknown.

(i) The relevant formula for the probability of obtaining $r = 17$ TYPE I (SATISFIED) employees in the sample is

$$\frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} = \frac{\binom{R}{17}\binom{200-R}{3}}{\binom{200}{20}} \tag{1}$$

where this formula can legitimately evaluated for $17 \leq R \leq 197$ as we know that there are at least 17 SATISFIED and at least 3 DISSATISFIED employees.

[4 MARKS]

(ii) Using the general version of Bayes Theorem (as indicated by the hint given in lectures ...), we have

$$\mathrm{P}\left(A_R \mid B_r\right) = \frac{\mathrm{P}\left(B_r \mid A_R\right)\mathrm{P}\left(A_R\right)}{\mathrm{P}\left(B_r\right)} = \frac{\mathrm{P}\left(B_r \mid A_R\right)\mathrm{P}\left(A_R\right)}{\sum_{i=r}^{N-(n-r)} \mathrm{P}\left(B_r \mid A_i\right)\mathrm{P}\left(A_i\right)} \qquad r \leq R \leq N-(n-r) \tag{2}$$

where we regard r as fixed, and consider the probability as R varies, where the term in the numerator is given by (1) as

$$\mathrm{P}\left(B_r \mid A_i\right) = \frac{\binom{i}{r}\binom{N-i}{n-r}}{\binom{N}{n}}$$

$\mathrm{P}(A_R \mid B_r)$ is the posterior probability that the number of satisfied employees is R in light of the information in the sample.

[6 MARKS]

(iii) Substituting in from (1), cancelling terms, and noting that prior knowledge specifies

$$\mathrm{P}\left(A_i\right) = \frac{1}{201} \qquad 0 \leq i \leq 200$$

we have

$$\mathrm{P}\left(A_R \mid B_{17}\right) = \frac{\binom{R}{17}\binom{200-R}{3}}{\sum_{i=17}^{197}\binom{i}{17}\binom{200-i}{3}} \qquad 17 \leq R \leq 197 \tag{3}$$

If $f(R) = \mathrm{P}(A_R \mid B_{17})$, MAPLE output and calculations give a plot (Fig1 in Assol2.mws) of this function, and the following numerical calculations

R	167	168	169	170	171	172	173
$f(R)$	0.02614	0.02644	0.02664	0.02673	0.02672	0.02658	0.026318

so the most likely value in light of the data is (unsuprisingly ?) 170

[6 MARKS]

Note that $f(R)$ is a probability mass function for a discrete random variable, X say, taking values on the range

$$\mathbb{X} = \{r, r+1, ..., N-(n-r)\} = \{17, 18, ..., 197\}$$

Note also that we can **prove** that the denominator in (3) is actually given by

$$\sum_{i=17}^{197} \binom{i}{17}\binom{200-i}{3} = \binom{201}{21}$$

and also that for any value of r in $\{0, 1, 2, ..., 20\}$, we have in the denominator of (2) that

$$\mathrm{P}(B_r) = \sum_{i=r}^{N-(n-r)} \mathrm{P}(B_r \mid A_i)\,\mathrm{P}(A_i) = \frac{1}{201}\frac{1}{\binom{N}{n}} \sum_{i=r}^{N-(n-r)} \binom{i}{r}\binom{N-i}{n-r} = \frac{1}{201}\frac{\binom{N+1}{n+1}}{\binom{N}{n}} = \frac{1}{201}\frac{201}{21} = \frac{1}{21}$$

both results using one of the binomial identities given in the lecture handout; see the verification in the MAPLE output. This point is interesting, but actually not directly relevant to the numerical calculations; we can use MAPLE to perform the calculation of $f(R)$.

Hence, we **can** in fact say that

$$\mathrm{P}(B_r) = \frac{1}{21} \qquad r = 0, 1, ..., 20$$

indicating that, under the model and before the data are collected, each of the possibilities for sample composition $r = 0, 1, ..., 20$ are **equally likely.** However, this must be **proved** using a formal argument, as it is not obvious from the specification given. Note also that we know that the denominator in the function $f(R)$ in (3) is merely a normalization constant which ensures that terms in $f(R)$ sum to 1, but *that does not depend on R.* Therefore we could proceed by calculating only the numerators for all the different values of R in (3), and then taking their sum as the normalization constant - these two methods of calculation give precisely the same numerical answers.

(iv) The probability that the number of SATISFIED employees in the workforce is greater that 95%, that is, greater than 190, is given by MAPLE as

$$\sum_{R=191}^{197} \mathrm{P}(A_R \mid B_r) = 0.01238$$

[4 MARKS]