Problem 1:

Assume missing data in this problem z_1, z_2

z_1 : number of male who plays Minecraft in survey one

$x-z_1$: number of female who plays MC in survey one.

z_2: number of male who doesn't play MC in survey one.

$n-x-z_2$: female ⋯ doesn't play ⋯

So we look at complete data:

survey #1	z_1+z_2 male	z_1 YES	z_2 NO
	$n-z_1-z_2$ female	$x-z_1$ YES	$n-x-z_2$ NO
survey #2	25 male	20 YES	5
	25 female	6 YES	19

E-step

So the likelihood: $\binom{z_1+z_2}{z_1} p_1^{z_1}(1-p_1)^{z_2} \binom{n-z_1-z_2}{x-z_1} p_2^{x-z_1}(1-p_2)^{n-x-z_2}$

$$\cdot \binom{m_1}{x_1} p_1^{x_1}(1-p_1)^{m_1-x_1} \binom{m_2}{x_2} p_2^{x_2}(1-p_2)^{m_2-x_2}$$

$$\text{Log-likelihood} = \text{Constant} + z_1 \log p_1 + z_2 \log(1-p_2) + (x-z_1)\log p_2 + (n-x-z_2)\log(1-p_2)$$

$$+ x_1 \log p_1 + (m_1-x_1)\log(1-p_1) + x_2 \log p_2 - (m_2-x_2)\log(1-p_2)$$

Note that, $E(z_1) = E\left(x \cdot \frac{p_1}{p_1+p_2}\right) = x \cdot \frac{p_1}{p_1+p_2}$

$$E(z_2) = \frac{(n-x)(1-p_1)}{2-p_1-p_2}$$

$$\text{So } E(\text{log-likelihood}) = \underbrace{\frac{x p_1}{p_1+p_2}}_{A} \log p_1^{(t)} + \underbrace{\frac{(n-x)(1-p_1)}{2-p_1-p_2}}_{B} \log(1-p_1^{(t)}) + \underbrace{\left(x - \frac{x p_1}{p_1+p_2}\right)}_{C} \log p_2^{(t)}$$

$$+ \underbrace{\left(n-x-\frac{(n-x)(1-p_1)}{2-p_1-p_2}\right)}_{D} \log(1-p_2^{(t)}) + x_1 \log p_1^{(t)} + (m_1-x_1)\log(1-p_1^{(t)})$$

$$+ x_2 \log p_2^{(t)} + (m_2-x_2)\log(1-p_2^{(t)})$$

M-Step

$$\frac{\partial E(\text{log-lik})}{\partial p_1^{(t)}} = A \cdot \frac{1}{p_1^{(t)}} - \frac{B}{1-p_1^{(t)}} + \frac{x_1}{p_1^{(t)}} - \frac{(m_1-x_1)}{1-p_1^{(t)}} = 0$$

$$\text{so } \hat{p}_1^{(t)} = \frac{A+x_1}{B+(m_1-x_1)+A+x_1} = \frac{A+x_1}{A+B+m_1}$$

$$\frac{\partial E(\text{log-lik})}{\partial p_2^{(t)}} = \frac{C}{p_2^{(t)}} - \frac{D}{1-p_2^{(t)}} + \frac{x_2}{p_2^{(t)}} - \frac{(m_2-x_2)}{1-p_2^{(t)}} = 0$$

$$\text{so } \hat{p}_2^{(t)} = \frac{C+x_2}{D+m_2+C}$$

where A, B, C, D are defined above

Problem 2 :

sp_i : species of beetle i , g_i be genus

m_i be mass, r_i be ratio, w_i be swamp value.

E-step :

Sps we have $I_{sp_i} = 1$ iff $sp_i = s$ where $s \in \{1, 2, \cdots, 10\}$

$$E(I_{sp_i=s}) = P(sp_i = s \mid \text{given data})$$
$$= P(sp_i = s \mid m_i, r_i, w_i, g_i)$$
$$= \frac{P(m_i, r_i, w_i, g_i \mid sp_i = s) P(sp_i = s)}{P(m_i, r_i, w_i, g_i)}$$
$$= \frac{P(m_i \mid sp_i = s) P(r_i \mid sp_i = s) P(w_i \mid sp_i = s) P(g_i \mid sp_i = s) P(sp_i = s)}{P(m_i, r_i, w_i, g_i)}$$
$$= \frac{P(g_i \mid sp_i = s) \exp(-(\log m_i - \mu_s)^2/(2\times 0.08^2)) \exp(-(\log r_i - \nu_s)^2/(2\times 0.1^2)) \varphi_s^{w_i} (1-\varphi_s)^{1-w_i} \cdot \alpha_s}{\sum_{s=1}^{10} P(g_i \mid sp_i = s) \exp(-(\log m_i - \mu_s)^2/(2\times 0.08^2)) \exp(-(\log r_i - \nu_s)^2/(2\times 0.1^2)) \varphi_s^{w_i} (1-\varphi_s)^{1-w_i} \alpha_s}$$

For $P(g_i \mid sp_i = s)$, we know it's determined by genus vs. species.

M-step Likelihood $= \prod_{n=1}^{500} \prod_{s=1}^{10} (N(\mu_s, m_i, 0.08^2) \times N(\nu_s, r_i, 0.1^2) \times Bern(\varphi_s, w_i) \times \alpha_s)^{I_{sp_i=s}}$

Log-likelihood $:= \sum_{n=1}^{500} \sum_{s=1}^{10} I_{sp_i=s} \{ -\log(\sqrt{2\pi}\, 0.08) - \log(\sqrt{2\pi}\, 0.1) - \frac{(\log m_i - \mu_s)^2}{2\times 0.08^2} - \frac{(\log r_i - \nu_s)^2}{2\times 0.1^2}$
$+ w_i \log \varphi_s + (1-w_i)\log(1-\varphi_s) + \log \alpha_s)$

$$\frac{\partial E(ll)}{\partial \mu_s} = \frac{2\log m_i}{2\times 0.08^2} - \frac{\mu_s}{0.08^2} \Rightarrow \sum_{i=1}^{500} P(sp_i = s \mid data) \frac{\log m_i}{0.08^2} - \frac{\mu_s}{0.08^2} = 0$$

so $\sum P \log m_i = \sum P \mu_s \Rightarrow \hat{\mu}_s = \frac{\sum P \log m_i}{\sum P}$

similarly, for $\frac{\partial E(ll)}{\partial \nu_s} \Rightarrow \hat{\nu}_s = \frac{\sum P \log r_i}{\sum P}$

$$\frac{\partial E(ll)}{\partial \varphi_s} = \sum\sum P(sp_i = s \mid data)\left(\frac{w_i}{\varphi_s} - \frac{1-w_i}{1-\varphi_s}\right) = 0 \Rightarrow \varphi_s = \frac{\sum P w_i}{\sum P}$$

$\frac{\partial E(ll)}{\partial \alpha_s} \Rightarrow$ Lagrange function $L = \sum_{s=1}^{10} (\sum P \cdot \log \alpha_s) + \lambda(\sum_{s=1}^{10} \alpha_s - 1)$

$$\frac{\partial L}{\partial \alpha_s} = \sum \frac{P}{\alpha_s} = -\lambda$$
$$\frac{\partial L}{\partial \lambda} = \sum_{s=1}^{10} \alpha_s = 1$$

$$\sum P = -\alpha_s \lambda$$
$$\sum_s^{10} \sum P = \sum_s^{10} \alpha_s \lambda \ (-1)$$
$$\Rightarrow N = -\lambda$$

so $\alpha_s = \frac{\sum P}{N}$

Note that all $\sum P$ are shorthand of $\sum_{i=1}^{N} P(sp_i = s \mid data)$

Problem 2

Let S_i be species of ith beetle, → we just write as S

m_i be mass, r_i be ratio, w_i be swamp value,

E-Step

Likelihood $L(\mu, \nu, \varphi, \alpha) = \prod_{i=1}^{500} P(m_i, r_i, w_i | s) P(S)$

$$= \prod_{i=1}^{500} \frac{1}{0.08\sqrt{2\pi}} \exp\left(-\frac{(\log m_i - \mu_s)^2}{2 \times 0.08^2}\right) \cdot \frac{1}{0.1\sqrt{2\pi}} \exp\left(-\frac{(\log r_i - \nu_s)^2}{2 \times 0.1^2}\right) \cdot \varphi_s^{w_i} (1-\varphi_s)^{1-w_i} \cdot \alpha_s$$

Log-likelihood: $ll(\mu, \nu, \varphi, \alpha) = \sum_{i=1}^{500} \left\{ \text{constant} - \frac{(\log m_i - \mu_s)^2}{2 \times 0.08^2} - \frac{(\log r_i - \nu_s)^2}{2 \times 0.1^2} + w_i \log(\varphi_s) + (1-w_i)\log(1-\varphi_s) + \log(\alpha_s) \right\}$

Now we want to discuss about S, we divide full data into 3 parts, where:

① A beetle has genus & species
② A beetle has only genus.
③ A beetle misses two such data.

By some manipulation: we know such 3 groups have 71, 102, 327 beetles respectively

First we claim the missing data (only one!) is Species of a beetle Z_i.

Sps for ②, it has genus(i), then for species j must be in the range of species that can only belong to genus i.

So conditional probability $P_{ij} = P\{Z_i = j \mid m_i, r_i, w_i, \theta^{(t)}\} = \frac{p(m_i, r_i, w_i | Z_i = j, \theta) P(Z_i = j | \theta)}{P(m_i, r_i, w_i | \theta)}$

$$= \frac{p(m_i, r_i, w_i | Z_i = j, \theta) \cdot P(Z_i = j | \theta)}{\sum_{k=1}^{10} p(m_i, r_i, w_i | Z_i = k, \theta) p(Z_i = k | \theta)}$$

actually k can be replaced by j since no matter what j is, $\sum$ is required.

For ③:

$$q_{ij} = \frac{p(m_i, r_i, w_i | Z_i = j, \theta) p(Z_i = j | \theta)}{\sum_{k=1}^{10} p(m_i, r_i, w_i | Z_i = j, \theta) p(Z_i = k | \theta)}$$

Take expectation about Z_i

For ①: $\sum_{n=1}^{71} E_{Z_i}(\text{loglike} | \theta^{(t)}) = ll(\mu_s, \nu_s, \varphi_s, \alpha_s)$

② $\sum_{n=1}^{102} E_{Z_i}(\text{log-like} | \theta^{(t)}) = \sum_{n=1}^{102} ll(\mu_j, \nu_j, \varphi_j, \alpha_j) \cdot P_{ij}$ where j is species, must belong to genus i

③ $\sum_{n=1}^{327} E_{Z_i}(\text{loglike} | \theta^{(t)}) = \sum_{n=1}^{102} \sum_{j=1}^{10} ll(\mu_j, \nu_j, \varphi_j, \alpha_j) \cdot q_{ij}$

M-Step

As derivative of summantion = summation of derivative.

$\frac{\partial E(ll①)}{\partial \mu_s} = \sum_{i \in ①} \frac{\log m_i - \mu_s}{0.08^2}$ for ① as $s \in [0, \cdots, 10]$

$\frac{\partial E(ll②)}{\partial \mu_s} = \sum_{i \in ②} P_{is} \cdot \frac{\log m_i - \mu_s}{0.08^2}$ for ②

$$\frac{\partial E(\mathcal{U}③)}{\partial \mu_s} = \sum_{i\in③} q_{is} \cdot \frac{\log m_i - \mu_s}{0.08^2}$$

so add together: $$\frac{\partial \mathcal{U}}{\partial \mu_s} = \sum_{i\in①} \log m_i - \sum_{i\in①} \mu_s + \sum_{i\in②} P_{is} \log m_i - \sum_{i\in②} P_{is}\, \mu_s + \sum_{i\in③} q_{is} \log m_i - \sum_{i\in③} q_{is}\, \mu_s = 0$$

~~therefore~~:

$$\mu_s = \frac{\sum_{i\in①} \log m_i + \sum_{i\in②} \log m_i \cdot P_{is} + \sum_{i\in③} q_{is} \log m_i}{\sum_{i\in①} 1 + \sum_{i\in②} P_{is} + \sum_{i\in③} q_{is}}$$

Similarly, $$\nu_s = \frac{\sum_{i\in①} \log r_i + \sum_{i\in②} \log r_i \cdot P_{is} + \sum_{i\in③} q_{is} \log r_i}{\sum_{i\in①} 1 + \sum_{i\in②} P_{is} + \sum_{i\in③} q_{is}}$$

For φ_s:

3 parts:

$$\frac{\partial E(\mathcal{U}①)}{\partial \varphi_s} = \sum_{i\in①} \left(\frac{w_i}{\varphi_s} - \frac{1-w_i}{1-\varphi_s}\right)$$

$$\frac{\partial E(\mathcal{U}②)}{\partial \varphi_s} = \sum_{i\in②} P_{is} \left(\frac{w_i}{\varphi_s} - \frac{1-w_i}{1-\varphi_s}\right)$$

$$\frac{\partial E(\mathcal{U}③)}{\partial \varphi_s} = \sum_{i\in③} q_{is} \left(\frac{w_i}{\varphi_s} - \frac{1-w_i}{1-\varphi_s}\right)$$

Add together:

$$\frac{\partial E(\mathcal{U})}{\partial \varphi_s} = \sum(\quad) + \sum(\quad) + \sum(\quad) = 0$$

$$\Rightarrow \varphi_s = \frac{\sum_{i\in①} w_i + \sum_{i\in②} P_{is} \cdot w_i + \sum_{i\in③} q_{is} w_i}{\sum_{i\in①} 1 + \sum_{i\in②} P_{is} + \sum_{i\in③} q_{is}}$$

$\hookrightarrow$ pretty similar as before, eh?

For α_s

Use Lagrange function $L = \sum_{s=1}^{10} \left(\sum_{i\in①} \log(\alpha_s) + \sum_{i\in②} \log(\alpha_s) \cdot P_{is} + \sum_{i\in③} \log(\alpha_s)\right) + \lambda \cdot \left(\sum_{s=1}^{10} \alpha_s - 1\right)$

$$\frac{\partial L}{\partial \alpha_s} = \sum_{i\in①} \frac{1}{\alpha_s} + \sum_{i\in②} \frac{P_{is}}{\alpha_s} + \sum_{i\in③} \frac{q_{is}}{\alpha_s} = -\lambda$$

$$\frac{\partial L}{\partial \lambda} = \sum_{s=1}^{10} \alpha_s = 1$$

$$\text{so } \alpha_s = -\frac{\sum_{①'} 1 + \sum_{②'} P_{is} + \sum_{③} q_{is}}{\lambda}$$

$$\text{then } \lambda = -\left(\sum_{①'} 1 + \sum_{②'} P_{is} + \sum_{③} q_{is}\right)$$

Here we use stricter condition for ① & ②

for i genus, ①': where $S = sp(i)$

②': $S \subseteq$ {possible species of i}

$$\text{So } \alpha_s = \frac{\sum_{①'} 1 + \sum_{②'} P_{is} + \sum_{③} q_{is}}{n}$$

As for here, $n=500$