1. (a) Describe the distribution of these differences.

Based on the model assumptions of $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, for each subject $j = 1, \cdots, 20$,

$$\begin{aligned} d_j &= y_{1j} - y_{2j} \\ &= \mu_1 + u_j + e_{1j} - (\mu_2 + u_j + e_{2j}) \\ &= (\mu_1 - \mu_2) + e_{1j} - e_{2j} \end{aligned}$$

$E(d_j) = \mu_1 - \mu_2$, $Var(d_j) = Var(e_{1j}) + Var(e_{2j}) = 2\sigma_e^2$. Because a linear combination of independent normal distributions is still normal, we have $d_j \sim N(\mu_1 - \mu_2, 2\sigma_e^2)$.
For any $j \neq j'$, $Cov(d_j, d_{j'}) = Cov(e_{1j} - e_{2j}, e_{1j'} - e_{2j'}) = 0$, so all d_j's are independent. Therefore $d_j \overset{iid}{\sim} N(\mu_1 - \mu_2, 2\sigma_e^2)$, which is a constant mean model. We can write this as a special case of a Gauss-Markov model as follows:

$$\boldsymbol{d} = \mathbf{1}[\mu_1 - \mu_2] + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{d} = (d_1, \ldots, d_{20})' \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, 2\sigma_e^2 \boldsymbol{I}).$$

(b) Provide a formula for a test statistic (as a function of $d_1, \cdots, d_{20}$) to test $H_0 : \mu_1 = \mu_2$.

Given the Gauss-Markov model above, we can find the formula for a test statistic by considering either a t test or and F test of $H_0 : \boldsymbol{C\beta} = \mathbf{0}$. The general formulas for a Gauss-Markov model can be simplified in this case because the "$\boldsymbol{X}$" matrix is just $\mathbf{1}$, the "$\boldsymbol{\beta}$" vector is just the one-element vector with $\mu_1 - \mu_2$ as the only element, and the "$\boldsymbol{C}$" matrix is just the 1×1 matrix with the element 1. Alternatively, can rewrite the model for differences as $d_1, \ldots, d_{20} \overset{iid}{\sim} N(\mu_d, \sigma_d^2)$, where $\mu_d = \mu_1 - \mu_2, \sigma_d^2 = 2\sigma_e^2$. Now the null hypothesis is equivalent to $H_0 : \mu_1 - \mu_2 = \mu_d = 0$. We can now see this as a STAT 101 type of question that asks us to test whether the mean of a normal distribution is zero based on an i.i.d. sample.

Let $\bar{d}. = \dfrac{\sum_{j=1}^{20} d_j}{20}$. Then $\bar{d}. \sim N\left(\mu_d, \dfrac{\sigma_d^2}{20}\right)$, and we can build up a t statistic to test $H_0 = \mu_d = 0$ as follows:

$$\begin{aligned} t &= \frac{\bar{d}. - 0}{\sqrt{\widehat{Var}(\bar{d}.)}} \\ &= \frac{\bar{d}.}{\sqrt{\hat{\sigma}_d^2/20}} \\ &= \frac{\bar{d}.}{\sqrt{\left[\frac{1}{20-1}\sum_{j=1}^{20}(d_j - \bar{d}.)^2\right]/20}} \end{aligned}$$

Or use F test statistic $F = t^2 = \dfrac{380\,\bar{d}.^2}{\sum_{j=1}^{20}(d_j - \bar{d}.)^2}$

(c) Fully state the exact distribution of the test statistic provided in part (b).

$$t \sim t_{19}\left(\frac{\mu_d}{\sqrt{\sigma_d^2/20}}\right) \overset{d}{=} t_{19}\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_e^2/10}}\right)$$

$$F \sim F_{1,19}\left(\frac{5(\mu_1-\mu_2)^2}{\sigma_e^2}\right)$$

(d) Provide a formula for a 95% confidence interval for $\mu_1 - \mu_2$.
Given only the 40 scores of the subjects who received only drink one type, the model for these scores is simplfied to be a Markov model as

$$\boldsymbol{y} = \underbrace{[\boldsymbol{I}_{2\times 2} \otimes \boldsymbol{1}_{20\times 1}]}_{\boldsymbol{X}} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{y} = [a_1, \cdots, a_{20}, b_1, \cdots, b_{20}]'$ and $\boldsymbol{\varepsilon}$ is a vector of random errors $[\varepsilon_{11}, \ldots, \varepsilon_{1,20}, \varepsilon_{21}, \ldots, \varepsilon_{2,20}]'$ where $\varepsilon_{ik} \overset{iid}{\sim} N(0, \sigma_u^2 + \sigma_e^2)$ for $i = 1, 2; k = 1, \cdots, 20$.
So the BLUE for $\mu_1 - \mu_2$ is $\bar{a}. - \bar{b}.$.

$$\begin{aligned} \widehat{Var}(\bar{a}. - \bar{b}.) &= \widehat{Var(\bar{a}.) + Var}(\bar{b}.) \\ &= 2 \times \frac{1}{20}(\widehat{\sigma_u^2 + \sigma_e^2}) \qquad \text{MSE for the Markov model above} \\ &= \frac{1}{10} \cdot \frac{1}{40-2}\left(\sum_{j=1}^{20}(a_j - \bar{a}.)^2 + \sum_{j=1}^{20}(b_j - \bar{b}.)^2\right) \end{aligned}$$

Therefore the 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{a}. - \bar{b}.) + t_{38,\,0.975}\sqrt{\frac{1}{380}\left(\sum_{j=1}^{20}(a_j - \bar{a}.)^2 + \sum_{j=1}^{20}(b_j - \bar{b}.)^2\right)}$$

with $df = n - rank(\boldsymbol{X}) = 38$

(e) Provide formulas for unbiased estimators of σ_u^2 and σ_e^2

From part (b), we have $\hat{\sigma}_d^2 = 2\hat{\sigma}_e^2 = \frac{1}{20-1}\sum_{j=1}^{20}(d_j - \bar{d}.)^2$.
From part (d) we have $\widehat{\sigma_u^2 + \sigma_e^2} = \frac{1}{40-2}\left(\sum_{j=1}^{20}(a_j - \bar{a}.)^2 + \sum_{j=1}^{20}(b_j - \bar{b}.)^2\right)$.
By solving the equations above, we can obtain

$$\begin{cases} \hat{\sigma}_e^2 = \dfrac{\sum_{j=1}^{20}(d_j - \bar{d}.)^2}{38} \\ \hat{\sigma}_u^2 = \dfrac{\left(\sum_{j=1}^{20}(a_j - \bar{a}.)^2 + \sum_{j=1}^{20}(b_j - \bar{b}.)^2\right)}{38} - \dfrac{\sum_{j=1}^{20}(d_j - \bar{d}.)^2}{38} \end{cases}$$

(f) Provide a simplified expression for the best linear unbiased estimator of $\mu_1 - \mu_2$.

Both $\bar{d}.$ and $(\bar{a}. - \bar{b}.)$ are independent unbiased estimators of $\mu_1 - \mu_2$. Thus, the BLUE of $\mu_1 - \mu_2$ is the weighted average of $\bar{d}.$ and $(\bar{a}. - \bar{b}.)$ with weights proportional to the inverse of the variances.

$$\begin{aligned}
\widehat{\mu_1 - \mu_2} &= \frac{Var^{-1}(\bar{d}.)}{Var^{-1}(\bar{d}.) + Var^{-1}(\bar{a}. - \bar{b}.)} \cdot \bar{d}. + \frac{Var^{-1}(\bar{a}. - \bar{b}.)}{Var^{-1}(\bar{d}.) + Var^{-1}(\bar{a}. - \bar{b}.)} \cdot (\bar{a}. - \bar{b}.) \\
&= \frac{\sigma_u^2 + \sigma_e^2}{\sigma_u^2 + 2\sigma_e^2} \cdot \bar{d}. + \frac{\sigma_e^2}{\sigma_u^2 + 2\sigma_e^2} \cdot (\bar{a}. - \bar{b}.)
\end{aligned}$$

2. Suppose the responses in problem 1 were sorted first by subject and then by drink, the response vector $\boldsymbol{y} = [y_{11}, y_{21}, \cdots, y_{1,20}, y_{2,20}, y_{1,21}, \cdots, y_{1,40}, y_{2,41}, \cdots, y_{2,60}]'$.
In model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$,

$$\boldsymbol{X} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \hline 1 & \\ 1 & \\ \vdots & \\ 1 & \\ \hline & 1 \\ & 1 \\ & \vdots \\ & 1 \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\\1\\1\end{array}}\right\} \text{40 rows} \\ \left.\vphantom{\begin{array}{c}1\\1\\1\\1\end{array}}\right\} \text{20 rows} \\ \left.\vphantom{\begin{array}{c}1\\1\\1\\1\end{array}}\right\} \text{20 rows} \end{array} \quad \text{and } \boldsymbol{Z} = \left[\begin{array}{ccccc|ccc|ccc} 1 &&&&&&&&&& \\ 1 &&&&&&&&&& \\ & 1 &&&&&&&&& \\ & 1 &&&&&&&&& \\ && \ddots &&&&&&&& \\ &&& 1 &&&&&&& \\ &&& 1 &&&&&&& \\ &&&&& 1 &&&&& \\ &&&&&& \ddots &&&& \\ &&&&&&& 1 &&& \\ &&&&&&&& 1 && \\ &&&&&&&&& \ddots & \\ &&&&&&&&&& 1 \end{array}\right]$$

the Kronecker product notation for $\boldsymbol{X}$ and $\boldsymbol{Z}$ are

$$\boldsymbol{X}_{80\times 2} = \begin{bmatrix} \mathbf{1}_{20\times 1} \otimes \boldsymbol{I}_{2\times 2} \\ \boldsymbol{I}_{2\times 2} \otimes \mathbf{1}_{20\times 1} \end{bmatrix}$$

$$\boldsymbol{Z}_{80\times 60} = diag(I_{20\times 20} \otimes \mathbf{1}_{2\times 1}, \boldsymbol{I}_{40\times 40})$$

3. By slide 54 of set 12, the BLUE of μ is a weighted average of independent linear unbiased estimators, where the weights are proportional to the inverse variances of the linear unbiased estimators.

   We can divide $\boldsymbol{y}$ into two independent subvectors by considering y_5 separately from $y_1, \ldots, y_4$. By the hint given, the BLUE of μ based only on $y_1, \ldots, y_4$ is $\frac{1}{4}\sum_{i=1}^{4} y_i$. Clearly, the BLUE of μ based on only y_5 is y_5 itself. These two estimators are independent, with variances

$$\begin{aligned}
\text{Var}\left(\frac{1}{4}\sum_{i=1}^{4} y_i\right) &= \text{Var}\left(\frac{1}{4}\mathbf{1}'_{4\times 1}(y_1, y_2, y_3, y_4)'\right) \\
&= \frac{1}{16}\mathbf{1}'_{4\times 1}\,\text{Var}\,(y_1, y_2, y_3, y_4)'\,\mathbf{1}_{4\times 1} \\
&= \frac{1}{16}\mathbf{1}'_{4\times 1}\begin{pmatrix} 5 & 1 & 1 & 1 \\ 1 & 5 & 1 & 1 \\ 1 & 1 & 5 & 1 \\ 1 & 1 & 1 & 5 \end{pmatrix}\mathbf{1}_{4\times 1} \\
&= \frac{32}{16} \\
&= 2,
\end{aligned}$$

$$\text{Var}(y_5) = 4.$$

   Then,

$$\begin{aligned}
\hat{\mu}_{\text{BLUE}} &= \frac{\frac{1}{\text{Var}\left(\frac{1}{4}\sum_{i=1}^{4} y_i\right)}\frac{1}{4}\sum_{i=1}^{4} y_i + \frac{1}{\text{Var}(y_5)}y_5}{\frac{1}{\text{Var}\left(\frac{1}{4}\sum_{i=1}^{4} y_i\right)} + \frac{1}{\text{Var}(y_5)}} \\
&= \frac{\frac{1}{2}\frac{1}{4}\sum_{i=1}^{4} y_i + \frac{1}{4}y_5}{\frac{1}{2} + \frac{1}{4}} \\
&= \frac{2}{3}\left(\frac{1}{4}\sum_{i=1}^{4} y_i\right) + \frac{1}{3}y_5 \\
&= \frac{1}{6}\sum_{i=1}^{4} y_i + \frac{1}{3}y_5.
\end{aligned}$$