MATH48082 Design and Analysis of Experiments

Coursework 1

24 March, 2020

Part 1: Tensile Strength of Steel

Introduction

Below is a table of results from an experiment that tested the tensile strength of steel alloy which contained three different percentages of each nickel and aluminium, totaling in 9 combinations. For each batch combination, measurements were taken of four steel bars produced and then recorded in the corresponding group. The focus of this experiment was to try and improve the tensile strength overall by finding the best percentage of each metal to be used.

Steel	0.1% Ni	0.2% Ni	0.3% Ni
0.1% Al	300, 291, 282, 287	342, 331, 338, 339	365, 360, 355, 358
0.2% Al	304, 315, 302, 304	328, 313, 338, 318	346, 357, 352, 361
0.3% Al	305, 292, 307, 307	329, 335, 339, 346	359, 353, 352, 343

In order to analyse this data set, we will firstly state the appropriate model to use given the factors of the experiment. Then, an ANOVA Table will be produced, where the calculations deriving the separate parts will be provided. Finally, statistical conclusions will be drawn from the results found and a suggestion of how to improve the tensile strength of these steel bars will be given using the percentages of Nickel and Aluminium tested.

Analysis

Initially, let us define some notation to be used from now on to correspond to the factors we are dealing with in this question: let A = percentage of Nickel in the steel alloy produced and B = percentage of Aluminium in the steel alloy produced. Then, we begin by proposing a suitable model for this experiment. Since this is a Complete Factorial Design (CFD) with m=2 qualitative factors, with fixed effects, we can consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where $i = 1, 2, 3(= t_A)$, $j = 1, 2, 3(= t_B)$ and $k = 1, 2, 3, 4(= r)$. Thus, there are $3 \times 3 \times 4 = 36$ observations in total, and

- μ is the grand mean,
- α_i is the mean effect of A (at level i),
- β_j is the mean effect of B (at level j),
- γ_{ij} is the interaction term between A (at level i) and B (at level j),
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the residual error.

In order to make some statistical inferences and analysis about the data that has been recorded, an ANOVA Table will be produced. This will help us to answer the main question whether there are any statistically significant effects of A, B or their interaction in the model. In other words, the null hypothesis H_0 is:

$$H_0 : \alpha_i = 0, \quad H_0 : \beta_j = 0, \quad H_0 : \gamma_{ij} = 0$$

The ANOVA Table we want to create looks something like this:

Source	SS	df	MS	F-ratio	P-value
Nickel % (A)	SSA	$t_A - 1$	$MSA = \frac{SSA}{t_A - 1}$	$f_A = \frac{MSA}{MSE}$	$P(F_{(t_A-1),(n-t)} > f_A)$
Aluminium % (B)	SSB	$t_B - 1$	$MSB = \frac{SSB}{t_B - 1}$	$f_B = \frac{MSB}{MSE}$	$P(F_{(t_B-1),(n-t)} > f_B)$
Interaction (AB)	SSAB	$(t_A - 1)(t_B - 1)$	$MSAB = \frac{SSAB}{(t_A-1)(t_B-1)}$	$f_{AB} = \frac{MSAB}{MSE}$	$P(F_{(t_A-1)(t_B-1),(n-t)} > f_{AB})$
Residual (E)	SSE	$n - t$	$MSE = \frac{SSE}{n-t}$	-	-
Total (T)	SST	$n - 1$	-	-	-

Therefore, the sum-of-squares (SS) needs to be calculated for each term in the model, appropriately "corrected" by subtracting the mean, μ, also known as the Correction Factor (CF),

$$CF = \frac{1}{n} y_{...}^2 (= n\bar{y}_{...}^2)$$

Since the total sum-of-squares, (SST), is easiest to work out, this will be worked out first. The formula for SST is

$$SST = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} (y_{ijk} - \bar{y}_{...})^2 = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} y_{ijk}^2 - CF$$

Using the measurements in the table above, this then becomes

$$SST = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{4} y_{ijk}^2 - \frac{1}{36} y_{...}^2 = 3923783 - \frac{1}{36}(11853)^2 = 3923783 - 3902600.25 = 21182.75$$

To calculate the sum-of-squares for A, SSA, we use the formula

$$SSA = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} (\bar{y}_{i..} - \bar{y}_{...})^2 = \frac{1}{r \times t_B} \sum_{i=1}^{t_A} y_{i..}^2 - CF$$

The individual components of this can be worked out using the correct data from the table. In this case, $y_{i..}$ are the corresponding column sums.

$$\begin{aligned} SSA &= \frac{1}{4 \times 3} \sum_{i=1}^{3} y_{i..}^2 - \frac{1}{36}(11853)^2 \\ &= \frac{1}{12}(y_{1..}^2 + y_{2..}^2 + y_{3..}^2) - \frac{1}{36}(11853)^2 \\ &= \frac{1}{12}(3596^2 + 3996^2 + 4261^2) - 3902600.25 \\ &= 3921279.417 - 3902600.25 = 18679.16666\ldots \\ &= 18679.167 \quad \text{(3 decimal places)} \end{aligned}$$

A similar formula for SSB can be used, but instead $y_{.j.}$ are the corresponding row sums.

$$SSB = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} (\bar{y}_{.j.} - \bar{y}_{...})^2 = \frac{1}{r \times t_A} \sum_{j=1}^{t_B} y_{.j.}^2 - CF$$

Again, using the table above,

$$\begin{aligned} SSB &= \frac{1}{4 \times 3} \sum_{j=1}^{3} y_{.j.}^2 - \frac{1}{36}(11853)^2 \\ &= \frac{1}{12}(y_{.1.}^2 + y_{.2.}^2 + y_{.3.}^2) - \frac{1}{36}(11853)^2 \\ &= \frac{1}{12}(3948^2 + 3938^2 + 3967^2) - 3902600.25 \\ &= 3902636.417 - 3902600.25 = 36.1666\ldots \\ &= 36.167 \quad \text{(3 decimal places)} \end{aligned}$$

For the interaction sum-of-squares, $SSAB$ is calculated slightly differently since the values for SSA and SSB must also be subtracted.

$$SSAB = \sum_{i=1}^{t_A}\sum_{j=1}^{t_B}\sum_{k=1}^{r}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 = \frac{1}{r}\sum_{i=1}^{t_A}\sum_{j=1}^{t_B} y_{ij.}^2 - SSA - SSB - CF$$

Within $SSAB$, we need to work out $y_{ij.}$ which corresponds to the sum of the four measurements of the steel bars in each cell for each of the 9 combinations.

$$\begin{aligned}
SSAB &= \frac{1}{4}\sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij.}^2 - 18679.167 - 36.167 - \frac{1}{36}(11853)^2 \\
&= \frac{1}{4}(y_{11.}^2 + y_{12.}^2 + y_{13.}^2 + y_{21.}^2 + y_{22.}^2 + y_{23.}^2 + y_{31.}^2 + y_{32.}^2 + y_{33.}^2) - 18679.167 - 36.167 - \frac{1}{36}(11853)^2 \\
&= \frac{1}{4}(1160^2 + 1225^2 + 1211^2 + 1350^2 + 1297^2 + 1349^2 + 1438^2 + 1416^2 + 1407^2) - 3921315.584 \\
&= 3922451.25 - 3921315.584 = 1135.666 \quad \text{(3 decimal places)}
\end{aligned}$$

Finally, for the Residual sum-of-squares, $SSE = SST - SSA - SSB - SSAB$, so using everything we have just calculated:

$$SSE = 21182.75 - 18679.167 - 36.167 - 1135.666 = 1331.75$$

Now that the sum-of-squares for each of the factors has been calculated, the Mean Squares can be worked out. Like in the table, Mean Squares (MS) is given by the ratio of the sum-of-squares to the degrees of freedom. Hence,

$$\begin{aligned}
MSA &= \frac{SSA}{t_A - 1} = \frac{18679.167}{2} = 9339.584 \\
MSB &= \frac{SSB}{t_B - 1} = \frac{36.167}{2} = 18.084 \\
MSAB &= \frac{SSAB}{(t_A - 1)(t_B - 1)} = \frac{1135.6}{4} = 283.915 \\
MSE &= \frac{SSE}{n - t} = \frac{1331.75}{27} = 49.324
\end{aligned}$$

(The results above are rounded to 3 decimal places.)

The F-ratio for each factor can be worked out by calculating the ratio between the Mean Square of the factor and the Mean Square of the residual:

$$\begin{aligned}
f_A &= \frac{MSA}{MSE} = \frac{9339.584}{49.324} = 189.352 \\
f_B &= \frac{MSB}{MSE} = \frac{18.084}{49.324} = 0.367 \\
f_{AB} &= \frac{MSAB}{MSE} = \frac{283.915}{49.324} = 5.756
\end{aligned}$$

The last thing required to work out the p-values for each of the factors is to find the percentage points $F_{(t-1),(n-t)}$. For this, standard statistical tables were used with a 5% significance level ($q = 0.95$).

$$\begin{aligned}
F_{(t_A-1),(n-t),q} = F_{(t_B-1),(n-t),q} &= F_{2,27,0.95} = 3.35, \\
F_{(t_A-1)(t_B-1),(n-t),q} &= F_{4,27,0.95} = 2.73
\end{aligned}$$

We now have everything we need that can be summarized in the ANOVA Table:

Source	SS	df	MS	F-ratio	P-value
Nickel % (A)	18679.167	2	9339.584	189.352	< 0.05 ($\implies$ Significant)
Aluminium % (B)	36.167	2	18.084	0.367	> 0.05 ($\implies$ Not Significant)
Interaction (AB)	1135.666	4	283.915	5.756	< 0.05 ($\implies$ Significant)
Residual (E)	1331.75	27	49.324	-	-
Total (T)	21182.75	35	-	-	-

From the table above, the p-values for A and the interaction suggest there is strong evidence at the 5% significance level that the percentage of Nickel and the interaction between Nickel and Aluminium affects the tensile strength of the steel bars produced. This also means we would reject the null hypotheses that $H_0 : \alpha_i = 0$ and $H_0 : \gamma_{ij} = 0$ at the 5% significance level.

The least squares problem for this question that needs to be minimized is

$$S = \sum_{i,j,k} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$$

The method to solve this is to take partial derivatives with respect to each of the unknown parameters and set them to zero. These will then be rearranged to find estimates for the parameters of the model. However, in order to solve these resulting equations, we need to set some linear constraints. These constraints are

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_{i,j} \gamma_{ij} = 0$$

If these constraints are imposed, then the least squares estimates are:

- $\hat{\mu} = \bar{y}_{...} = \frac{1}{36} \sum_{i,j,k} y_{...} = \frac{11853}{36} = 329.25$
- $\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad i = 1, 2, 3$
  - $\hat{\alpha}_1 = \bar{y}_{1..} - \bar{y}_{...} = \frac{3596}{12} - 329.25 = -29.583$
  - $\hat{\alpha}_2 = \bar{y}_{2..} - \bar{y}_{...} = \frac{3996}{12} - 329.25 = 3.75$
  - $\hat{\alpha}_3 = \bar{y}_{3..} - \bar{y}_{...} = \frac{4261}{12} - 329.25 = 25.833$
- $\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad j = 1, 2, 3$
  - $\hat{\beta}_1 = \bar{y}_{.1.} - \bar{y}_{...} = \frac{3948}{12} - 329.25 = -0.25$
  - $\hat{\beta}_2 = \bar{y}_{.2.} - \bar{y}_{...} = \frac{3938}{12} - 329.25 = -1.083$
  - $\hat{\beta}_3 = \bar{y}_{.3.} - \bar{y}_{...} = \frac{3967}{12} - 329.25 = 1.333$
- $\hat{\gamma}_{ij} = \bar{y}_{ij.} - \bar{y}_{...} - \hat{\alpha}_i - \hat{\beta}_j, \quad i = 1, 2, 3, j = 1, 2, 3$
  - $\hat{\gamma}_{11} = \bar{y}_{11.} - \bar{y}_{...} - \hat{\alpha}_1 - \hat{\beta}_1 = \frac{1160}{4} - 329.25 - (-29.583) - (-0.25) = -9.417$
  - $\hat{\gamma}_{12} = \bar{y}_{12.} - \bar{y}_{...} - \hat{\alpha}_1 - \hat{\beta}_2 = \frac{1225}{4} - 329.25 - (-29.583) - (-1.083) = 7.666$
  - $\hat{\gamma}_{13} = \bar{y}_{13.} - \bar{y}_{...} - \hat{\alpha}_1 - \hat{\beta}_3 = \frac{1211}{4} - 329.25 - (-29.583) - 1.333 = 1.75$
  - $\hat{\gamma}_{21} = \bar{y}_{21.} - \bar{y}_{...} - \hat{\alpha}_2 - \hat{\beta}_1 = \frac{1350}{4} - 329.25 - 3.75 - (-0.25) = 4.75$
  - $\hat{\gamma}_{22} = \bar{y}_{22.} - \bar{y}_{...} - \hat{\alpha}_2 - \hat{\beta}_2 = \frac{1297}{4} - 329.25 - 3.75 - (-1.083) = -7.667$
  - $\hat{\gamma}_{23} = \bar{y}_{23.} - \bar{y}_{...} - \hat{\alpha}_2 - \hat{\beta}_3 = \frac{1349}{4} - 329.25 - 3.75 - 1.333 = 2.917$
  - $\hat{\gamma}_{31} = \bar{y}_{31.} - \bar{y}_{...} - \hat{\alpha}_3 - \hat{\beta}_1 = \frac{1438}{4} - 329.25 - 25.833 - (-0.25) = 4.667$
  - $\hat{\gamma}_{32} = \bar{y}_{32.} - \bar{y}_{...} - \hat{\alpha}_3 - \hat{\beta}_2 = \frac{1416}{4} - 329.25 - 25.833 - (-1.083) = 0$
  - $\hat{\gamma}_{33} = \bar{y}_{33.} - \bar{y}_{...} - \hat{\alpha}_3 - \hat{\beta}_3 = \frac{1407}{4} - 329.25 - 25.833 - 1.333 = -4.666$

Also, the estimate for σ^2 is just the Mean Squared Error i.e. $\hat{\sigma}^2 = MSE = 49.324 = 7.023^2$ (3 decimal places).

Conclusion

Given the results above and the aim of this experiment, it would be suggested that in order to improve the strength of the steel alloy, 0.3% Nickel and 0.1% Aluminium should be used. We have strong evidence that Nickel is significant in the model since it has a p-value smaller than 0.05 and thus is the dominant factor, so a higher percentage would result in a stronger steel bar. Since there was also strong evidence of an interaction between Nickel and Aluminium, we can evaluate the data that was collected for tensile strength with higher Nickel percentage and see which percentage also contributes to a higher strength. The average tensile strength at 0.1%, 0.2% and 0.3% Aluminium with 0.3% Nickel are 359.5, 354 and 351.75 MPa respectively. This implies a lower percentage of Aluminium produces a stronger steel alloy bar. Therefore the interaction implies a high Nickel and low Aluminium percentage should be used.

In this case, the most desirable combination would be 0.3% of Nickel and 0.1% of Aluminium.

In terms of other statistical analysis that could be carried out, Contrasts, Least Significant Differences (LSD's) and Confidence Intervals could be computed. These would provide us with more information about the treatment estimates we have observed and so be able to make inferences about the smallest difference between estimated treatment effects that are statistically significant. The LSD approach comes with a disadvantage; this method can result in finding too many significant differences and thus the comparison between them becomes meaningless. This could be modified, or instead, Tukey's test or Duncan's test could be used.

In light of new information that manganese increases the tensile strength of steel, a new experiment could be conducted with different levels of all three metals, Nickel, Aluminium and Manganese. A suitable model for this would be

$$y_{ijk\ell} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijk\ell}$$

where

- μ is the overall mean,
- α_i, β_j and γ_k are the main effects,
- $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{jk}$ are the first order interaction terms,
- $(\alpha\beta\gamma)_{ijk}$ is the second order interaction term,
- $\epsilon_{ijk\ell} \sim N(0, \sigma^2)$ is the residual error.

Part 2: Amount of Codeine in a Cough Syrup

Introduction

The objective of this report is to analyse the data provided from an experiment that occurs during a quality control inspection within a factory that produces bottles of cough syrup in batches. The cough syrup contains codeine and the inspection aims to check amount of codeine in each bottle. In total, 7 batches are tested and within each batch, four bottles are randomly selected. Below, the results of two samples from each of the bottles are recorded (mg/100ml).

Batch	Bottle 1	Bottle 2	Bottle 3	Bottle 4
1	0.83 | 0.80	0.77 | 0.80	0.77 | 0.83	0.81 | 0.79
2	0.68 | 0.64	0.62 | 0.63	0.62 | 0.65	0.63 | 0.65
3	0.63 | 0.63	0.67 | 0.65	0.60 | 0.63	0.62 | 0.61
4	0.77 | 0.76	0.77 | 0.74	0.73 | 0.77	0.78 | 0.76
5	0.69 | 0.71	0.74 | 0.75	0.72 | 0.75	0.72 | 0.76
6	0.63 | 0.65	0.61 | 0.61	0.59 | 0.61	0.62 | 0.64
7	0.72 | 0.73	0.74 | 0.73	0.72 | 0.70	0.69 | 0.73

In this report, an ANOVA Table will be produced, resulting in some statistical conclusions being drawn and a 90% confidence interval being produced for the variance of errors, σ^2. Also provided will be estimates of the correlation coefficient between different observations for two scenarios: within the same bottle, and within the same batch but different bottles.

Analysis

To begin with, let A= Batch factor and B(A)= Bottle factor nested within the Batch. This makes notation simpler to write. Since this inspection is set up in a way that has a hierarchical structure, an appropriate model needs to be chosen. For this experiment, we can use a nested model with m=2 factors, where it can be interpreted that the randomly chosen bottle is nested within the batch being tested. Thus, the model is

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk},$$

where $i = 1, 2, \ldots, 7 (= t_A)$, $j = 1, 2, 3, 4 (= t_B)$ and $k = 1, 2 (= r)$. This means there are $7 \times 4 \times 2 = 56$ observations in total, and

- μ is the grand mean,
- $\alpha_i \sim N(0, \sigma_A^2)$, the main effect of A (at level i),
- $\beta_{j(i)} \sim N(0, \sigma_B^2)$, the main effect of B (at level j) within level i of A,
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the residual error, with all random variables independent.

This implies that $y_{ijk} \sim N(\mu, \sigma_A^2 + \sigma_B^2 + \sigma^2)$. We wish to test the following null hypotheses:

$$H_0 : \sigma_A^2 = 0, \quad H_0 : \sigma_B^2 = 0$$

A general ANOVA Table for this type of model is given below to show the different components that will be calculated in this section.

Source	SS	df	MS	E[MS]	F-ratio
Batch (A)	SSA	$t_A - 1$	$MSA = \frac{SSA}{t_A - 1}$	$rt_B\sigma_A^2 + r\sigma_B^2 + \sigma^2$	$f_A = \frac{MSA}{MSB(A)}$
Bottle (B(A))	SSB(A)	$t_A(t_B - 1)$	$MSB(A) = \frac{SSB(A)}{t_A(t_B - 1)}$	$r\sigma_B^2 + \sigma^2$	$f_{B(A)} = \frac{MSB(A)}{MSE}$
Residual (E)	SSE	$n - t$	$MSE = \frac{SSE}{n - t}$	σ^2	-
Total (T)	SST	$n - 1$	-	-	-

The sum-of-squares (SS) will be calculated for each factor of the inspection and appropriately "corrected" by subtracting the Correction Factor (CF),

$$CF = \frac{1}{n} y_{...}^2 (= n\bar{y}_{...}^2)$$

Firstly, in the general case, the formula for SST is

$$SST = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} (y_{ijk} - \bar{y}_{...})^2 = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} y_{ijk}^2 - CF$$

Since we are provided with everything needed in the table above, the calculation for this is simple. By plugging in the values recorded, we get

$$SST = \sum_{i=1}^{7} \sum_{j=1}^{4} \sum_{k=1}^{2} y_{ijk}^2 - \frac{1}{56} y_{...}^2 = 27.699 - \frac{1}{56}(39.2)^2 = 27.699 - 27.44 = 0.259$$

Then, for SSA, the formula is

$$SSA = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} (\bar{y}_{i..} - \bar{y}_{...})^2 = \frac{1}{r \times t_B} \sum_{i=1}^{t_A} y_{i..}^2 - CF$$

This sum corresponds to working out the sum of all the samples taken from each bottle for each batch, $i = 1, \ldots, 7$ (in other words, row sums) and hence

$$\begin{aligned} SSA &= \frac{1}{2 \times 4} \sum_{i=1}^{7} y_{i..}^2 - \frac{1}{56}(39.2)^2 \\ &= \frac{1}{8}(y_{1..}^2 + y_{2..}^2 + y_{3..}^2 + y_{4..}^2 + y_{5..}^2 + y_{6..}^2 + y_{7..}^2) - \frac{1}{56}(39.2)^2 \\ &= \frac{1}{8}(6.4^2 + 5.12^2 + 5.04^2 + 6.08^2 + 5.84^2 + 4.96^2 + 5.76^2) - 27.44 \\ &= 27.6784 - 27.44 = 0.2384 \end{aligned}$$

For $SSB(A)$, we need to adjust this sum according. If this was a Complete Factorial Design (fixed or random), this term would come from pooling together the results of a main effect B and an interaction term AB. In this case, to directly work this out, we can just subtract the value found for SSA:

$$SSB(A) = \sum_{i=1}^{t_A} \sum_{j=1}^{t_B} \sum_{k=1}^{r} (\bar{y}_{ij.} - \bar{y}_{i..})^2 = \frac{1}{r} \sum_{i,j} y_{ij.}^2 - CF - SSA$$

Therefore,

$$\begin{aligned} SSB(A) &= \frac{1}{2} \sum_{i,j} y_{ij.}^2 - \frac{1}{56}(39.2)^2 - SSA \\ &= \frac{1}{2}(y_{11.}^2 + y_{12.}^2 + y_{13.}^2 + \cdots + y_{72.}^2 + y_{73.}^2 + y_{74.}^2) - \frac{1}{56}(39.2)^2 - 0.2384 \\ &= 27.6892 - 27.44 - 0.2384 \\ &= 0.0108 \end{aligned}$$

Finally, for SSE, we have $SSE = SST - SSA - SSB(A)$, so using everything we have just calculated:

$$SSE = 0.259 - 0.2384 - 0.0108 = 0.0098$$

The equations for calculating the Mean Squares are just the ratio of the sum-of-squares and the degrees of freedom for the corresponding factor.

$$MSA = \frac{SSA}{t_A - 1} = \frac{0.2384}{6} = 0.0397333\cdots = 0.03973$$
$$MSB(A) = \frac{SSB(A)}{t_A(t_B - 1)} = \frac{0.0108}{7 \times 3} = 0.00051428\cdots = 0.0005143$$
$$MSE = \frac{SSE}{n - t} = \frac{0.0098}{28} = 0.00035$$

(The results above are rounded to 4 significant figures.)

The equations for calculating the F-ratio for factors in a nested model are slightly different than that of a normal Complete Factorial Design (fixed or random). A general rule is to use the mean square value from the next stratum down as the denominator instead of the standard MSE. Hence here, since the Bottle factor is nested within the Batch, we would use the mean square of $B(A)$ as the denominator for f_A.

$$f_A = \frac{MSA}{MSB(A)} = \frac{0.03973}{0.0005143} = 77.25063193 = 77.25$$
$$f_{B(A)} = \frac{MSB(A)}{MSE} = \frac{0.0005143}{0.00035} = 1.469428571 = 1.469$$

(The results above are rounded to 4 significant figures.)

Lastly, the percentage points of the F-distribution can be found for the Batch and the Bottle factors and then calculate

$$P(F_{(t_A-1),(n-t)} > f_A), \quad P(F_{(t_{B(A)}-1),(n-t)} > f_{B(A)})$$

At the 5% significance level ($q = 0.95$), these are

$$F_{(t_A-1),(n-t),q} = F_{6,28,0.95} = 2.45$$
$$F_{(t_A(t_B-1)),(n-t),q} = F_{21,28,0.95} = 1.90 \text{ (By interpolation)}$$

Now that we have everything we need, we can just plug the results we have worked out above into the ANOVA Table:

Source	SS	df	MS	E[MS]	F-ratio	P-value
Batch (A)	0.2384	6	0.03973	$8\sigma_A^2 + 2\sigma_B^2 + \sigma^2$	77.25	< 0.05 ($\implies$ Significant)
Bottle (B(A))	0.0108	21	0.0005143	$2\sigma_B^2 + \sigma^2$	1.469	> 0.05 ($\implies$ Not significant)
Residual (E)	0.0098	28	0.00035	σ^2	-	-
Total (T)	0.259	55	-	-	-	-

As a result of the calculations above, we can observe that the p-value for the Batch factor is smaller than 0.05 and this suggests there is strong evidence that there is a difference in the amount of codeine between batches, whereas there is little evidence to suggest that the amount of codeine differs between bottles of the same batch.

For further analysis, a 90% Confidence Interval will be calculated for the variance of the errors σ^2. It can be shown that

$$\frac{SSE}{\sigma^2} \sim \chi^2_v$$

where v is the residual degrees of freedom.
Therefore, a suitable $100(1-\alpha)\%$ Confidence Interval for σ^2 is

$$C.I = \left(\frac{SSE}{\chi^2_{v,1-\alpha/2}}, \frac{SSE}{\chi^2_{v,\alpha/2}} \right)$$

If a 90% Confidence Interval is required, this implies that $\alpha = 0.1$. The values of $\chi^2_{v,1-\alpha/2}$ and $\chi^2_{v,\alpha/2}$ can be found in the standard statistical tables. Subbing these values into the formula above, the Confidence Interval becomes

$$C.I = \left(\frac{0.0098}{41.337}, \frac{0.0098}{16.928}\right) = (0.00023707\ldots, 0.00057892\ldots) = (0.0002371, 0.0005789)$$

In the usual way, the estimate for σ^2 is equal to the Mean Squared error,

$$\hat{\sigma}^2 = MSE = 0.00035$$

For a nested model, the estimate for σ^2_A is calculated by using the following formula:

$$\hat{\sigma}^2_A = \frac{1}{r \times t_B}(MSA - MSB(A)) = \frac{1}{2 \times 4}(0.03973 - 0.0005143) = 0.0049019\cdots = 0.004902$$

A similar formula and result is found for the estimate of σ^2_B:

$$\hat{\sigma}^2_B = \frac{1}{r}(MSB(A) - MSE) = \frac{1}{2}(0.0005143 - 0.00035) = 0.00008215$$

Finally, these can then be used to find estimates of the correlation coefficient for two different scenarios; firstly within the same bottle, and then within the same batch but different bottles.

The theory for this is that if we consider the model that is used in this experiment,

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

and then

$$y_{i'j'k'} = \mu + \alpha_{i'} + \beta_{j'(i')} + \epsilon_{i'j'k'},$$

the covariance between these can be computed and used in the formula to find the correlation coefficient. Note, that the covariance between y_{ijk} and $y_{i'j'k'}$ is given by

$$\begin{aligned} cov(y_{ijk}, y_{i'j'k'}) &= cov(\mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}, \mu + \alpha_{i'} + \beta_{j'(i')} + \epsilon_{i'j'k'}) \\ &= cov(\alpha_i, \alpha_{i'}) + cov(\beta_{j(i)}, \beta_{j'(i')}) + cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) \end{aligned}$$

due to independence between the random variables.

By definition,

$$cov(\alpha_i, \alpha_i) = \delta_{ii'}\sigma^2_A, \quad cov(\beta_{j(i)}, \beta_{j'(i')}) = \delta_{ii'}\delta_{jj'}\sigma^2_B, \quad cov(\epsilon_{ijk}, \epsilon_{i'j'k'}) = \delta_{ii'}\delta_{jj'}\delta_{kk'}\sigma^2$$

where

$$\delta_{\ell\ell'} = \begin{cases} 1 & \text{for } \ell = \ell', \\ 0 & \text{for } \ell \neq \ell'. \end{cases}$$

Therefore, we gain the following general expression for the correlation between the different samples which can be easily computed:

$$\begin{aligned} \rho(y_{ijk}, y_{i'j'k'}) &= \frac{cov(y_{ijk}, y_{i'j'k'})}{var(y_{ijk})^{\frac{1}{2}} var(y_{i'j'k'})^{\frac{1}{2}}} \\ &= \frac{\delta_{ii'}\sigma^2_A + \delta_{ii'}\delta_{jj'}\sigma^2_B + \delta_{ii'}\delta_{jj'}\delta_{kk'}\sigma^2}{\sigma^2_A + \sigma^2_B + \sigma^2} \end{aligned}$$

Hence,

$$\rho(y_{ijk}, y_{i'j'k'}) = \begin{cases} 0 & \text{for } i \neq i', \\ \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_B + \sigma^2} & \text{for } i = i', j \neq j', \\ \frac{\sigma^2_A + \sigma^2_B}{\sigma^2_A + \sigma^2_B + \sigma^2} & \text{for } i = i', j = j', k \neq k', \\ 1 & \text{for } i = i', j = j', k = k'. \end{cases}$$

Since we worked out estimates for the variances earlier, $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and $\hat{\sigma}^2$ can just be substituted in:

$$\hat{\rho}(y_{ijk}, y_{i'j'k'}) = \begin{cases} 0 & \text{for } i \neq i', \\ \frac{0.004902}{0.004902+0.00008215+0.00035} = 0.9190 & \text{for } i = i', j \neq j', \\ \frac{0.004902+0.00008215}{0.004902+0.00008215+0.00035} = 0.9344 & \text{for } i = i', j = j', k \neq k', \\ 1 & \text{for } i = i', j = j', k = k'. \end{cases}$$

Using these results, the estimate of the correlation coefficient between different observations within the same bottle occurs when the batch number is the same and the bottle number is the same, corresponding to $i = i', j = j'$ in the cases above. Thus

$$\hat{\rho}(y_{ijk}, y_{i'j'k'}) = 0.9344 \quad \text{(4 significant figures)}$$

Similarly, the estimate of the correlation coefficient between different observations within the same batch but within different bottles is interpreted as when only the batch number is the same so $i = i'$ in the cases above. Hence

$$\hat{\rho}(y_{ijk}, y_{i'j'k'}) = 0.9190 \quad \text{(4 significant figures)}$$

Conclusion

In conclusion, the analysis of the nested model shows that at the 5% significance level, there is strong statistical evidence that there is a difference in the amount of codeine being recorded between the batches taken, but no statistical evidence of a difference between the bottles within the batch itself. This is because we observed a small p-value for the Batch factor (A) (smaller than 0.05) but the p-value for Bottles was not significant. Therefore, we would reject the null hypothesis that $H_0 : \sigma_A^2 = 0$ from this result.

Also, the Confidence Interval obtained for the variance of the error shows that there is some variability in this value, but our estimate sits comfortably in the middle of interval and thus would suggest our estimate is fairly accurate.

Furthermore, the correlation coefficients calculated are both close to 1 meaning that there is some strong positive correlation between the observations in both scenarios, despite the random variables being independent. This correlation is due to the random effects.

If there are many more factories within a region that are producing the same syrup and similar data is available from these factories, then the model suggested here could be extended to include the random variation between factories. In the hierarchical structure, the factory the batches and subsequent bottles the samples are taken from would now be the top layer of the model.
A new model would look something like the following:

$$y_{ijk\ell} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(j(i))} + \epsilon_{ijk\ell}$$

where

- μ is the grand mean,
- $\alpha_i \sim N(0, \sigma_A^2)$, the main effect of A (at level i),
- $\beta_{j(i)} \sim N(0, \sigma_B^2)$, the main effect of B (at level j) within level i of A,
- $\gamma_{k(j(i))} \sim N(0, \sigma_C^2)$, the main effect of C (at level k) within level j of B, within level i of A,
- $\epsilon_{ijk\ell} \sim N(0, \sigma^2)$ is the residual error, with all random variables independent.