Project 4 Report

Weijiang Hou, Nathaniel Morgan, Christian Otto, & Jianxiong Shen

04/22/2022

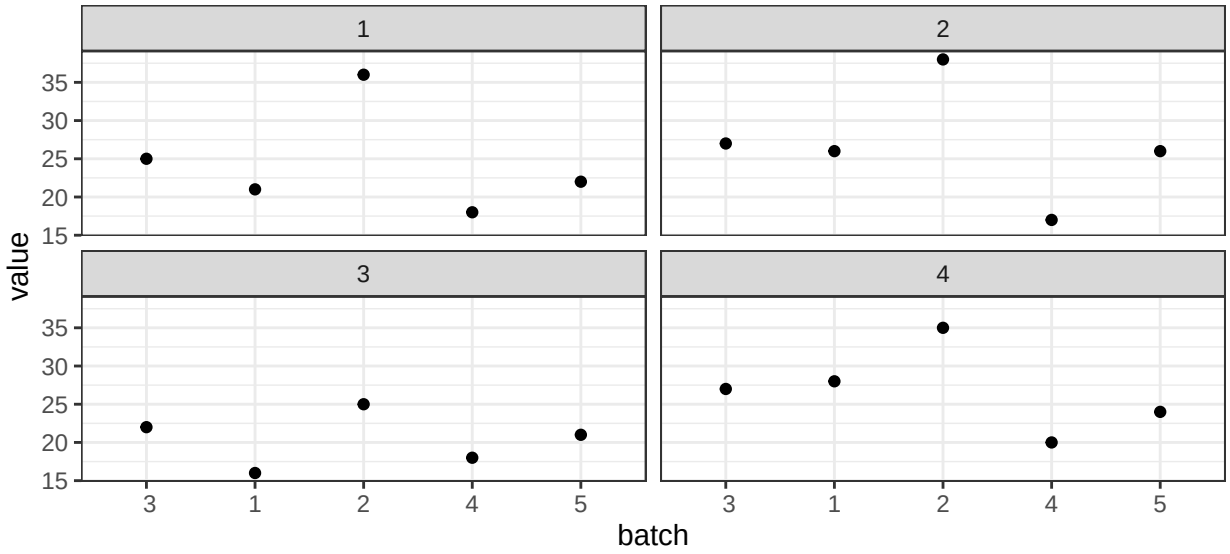
Introduction

In the early stages of processing, natural fibers (such as cotton and wool) require cleaning. A downside of cleaning is that the process results in a loss of wool. For wool sellers, losing wool weight is detrimental for their business because it causes their overall wool production to decrease, which causes them to miss out on possible profit. Therefore, it is important to know which wool cleaning process results in the least wool lost. To achieve this goal, a textile specialist investigated four cleaning processes. Wool from five different ranchers, suppliers, etc, were received for the study. After removing foreign debris, the wool from each batch was thoroughly mixed and an equal amount was assigned to each of the four cleaning processes. For our analysis, we will utilize the Randomized Complete Block Design to determine if there are differences in the wool weight loss for each process.

The data for this study is provided below. The sample includes 20 observations, with a single observation from each combination of Process and Batch. All the analysis was performed using the R programming language.

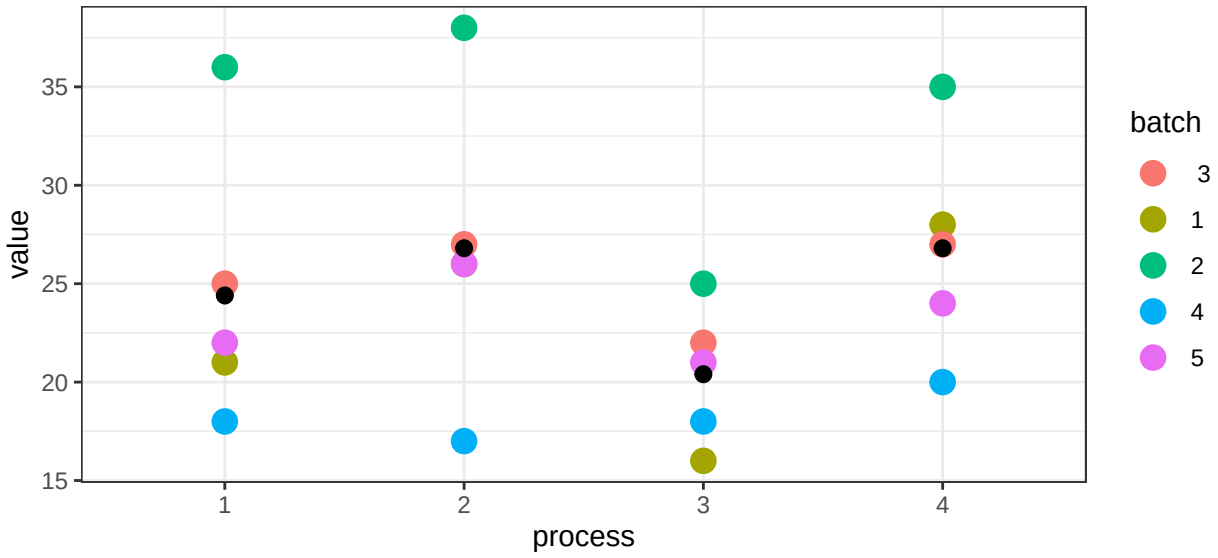
	Batch 1	2	3	4	5
Process 1	21	36	25	18	22
2	26	38	27	17	26
3	16	25	22	18	21
4	28	35	27	20	24

Exploration



From the plot above, we can see that the batch variable seems to follow a specific pattern regardless of the process.

We now plot each of the batches to the corresponding process to look at the spread of data. The black dot represents the mean for a specific process.



Since the process and batch are fixed, We are assuming we have a restricted model.

Often the actual differences in the treatment means are hidden because the large amount of variations within the treatment groups or cells. Often this variation can be explained by considering another variable (called a block or blocking effect) which partitions the variation in the treatment means into smaller homogeneous parts.

We first fit a one-way ANOVA with the variable "process" to see if, ignoring the "batch" variable, there is a significant difference between processes.

Table 2: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
process	3	136.8	45.6	1.241	0.3277
Residuals	16	588	36.75	NA	NA

If we use process as the treatment only and ignore the batch variable, the process variable is not significant.

To get a better estimate of process, we should conduct a Randomized Complete Block Design and treat batch as a block variable.

The Randomized Complete Block Design is also known as the two-way ANOVA without interaction. A key assumption in the analysis is that the effect of each level of the treatment factor is the same for each level of the blocking factor. In RCBD, there is one observation for each combination of levels of the treatment and block factors.

The model for an RCBD (or two-way ANOVA without interactions) is:

$y_{ijk} = \mu + \tau_i + \beta_j + e_{ijk}$, for $i = 1, 2, ..., a, j = 1, 2, ..., b$ and $k = 1, 2, ..., n, N = nab$.

Since it is a restricted model, the expected value for the mean square terms are,

$$E(MS_{block}) = \sigma^2 + a\frac{\sum_{j=1}^{b} \beta_j^2}{b-1}$$

$$E(MS_{treatment}) = \sigma^2 + b\frac{\sum_{i=1}^{a} \tau_i^2}{a-1}$$

$$E(MS_{error}) = \sigma^2$$

There are two hypothesis tests in an RCBD, and they are always the same:

H_0: The means of all treatments are equal versus

H_1: At least one of the treatments has a different mean

and

H_0: The means of all blocks are equal versus

H_1: At least one of the blocks has a different mean

We will now conduct the Randomized Complete Block Design

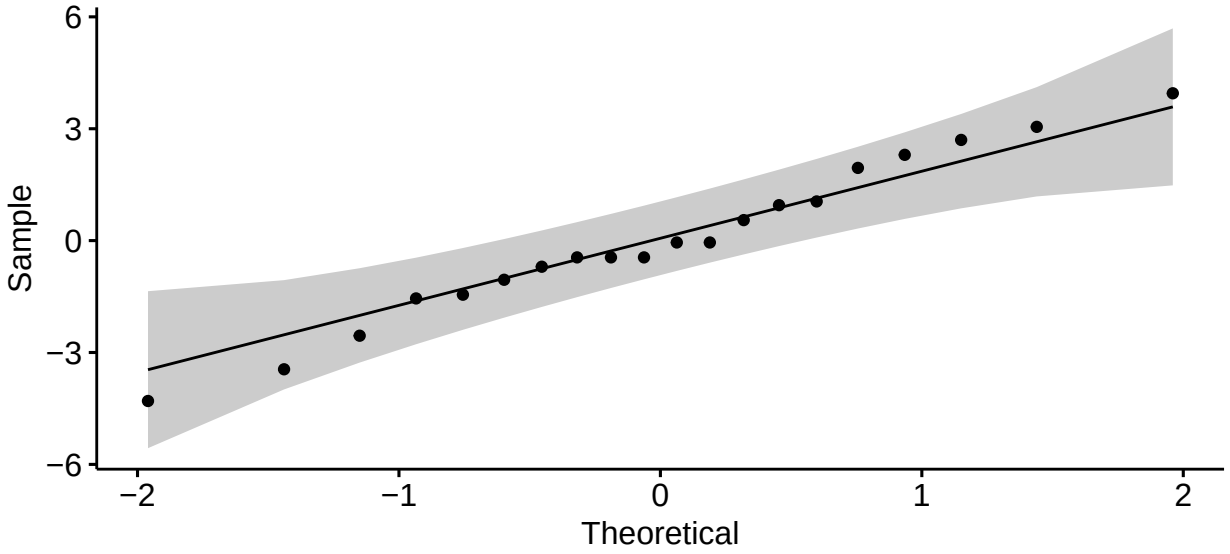
Table 3: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
process	3	136.8	45.6	6.275	0.008326
batch	4	500.8	125.2	17.23	6.5e-05
Residuals	12	87.2	7.267	NA	NA

Now that we use the "batch" variable as a block variable, the process variable is significant. Therefore, we can reject the null hypothesis and state that there is sufficient evidence to conclude that at least one of the processes has a different mean.

Assumption checking

The first assumption to check is normality in the residuals:



According to the qqplot, all of the values are close to the straight line and all of of them fall within the confidence interval. We can assume normality in the residuals.

To further support this, we conduct a shapiro-wilk test, where the hypotheses are as follows

h_0 : The data is normally distributed

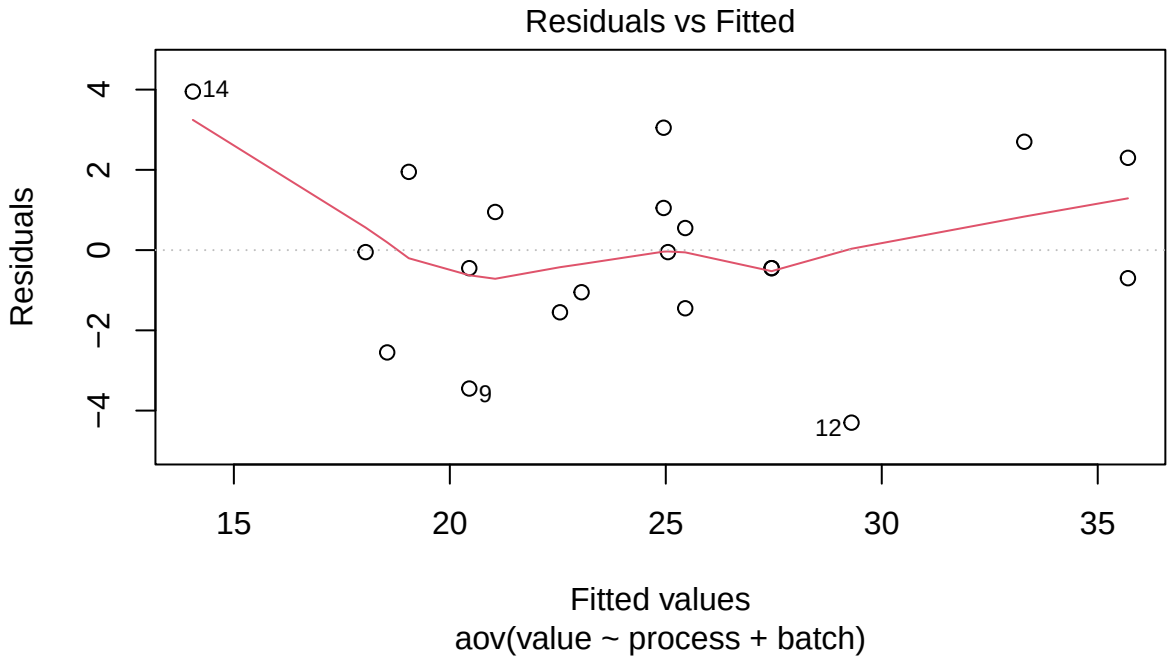
h_a : The data is not normally distributed

Table 4: Shapiro-Wilk normality test: `model_2way$residuals`

Test statistic	P value
0.9825	0.962

The p-value for Shapiro-Wilk test is 0.962. We **do not** have sufficient evidence to suggest the data is **not** normally distributed.

The next assumption we look at is the assumption of homoscedasticity:



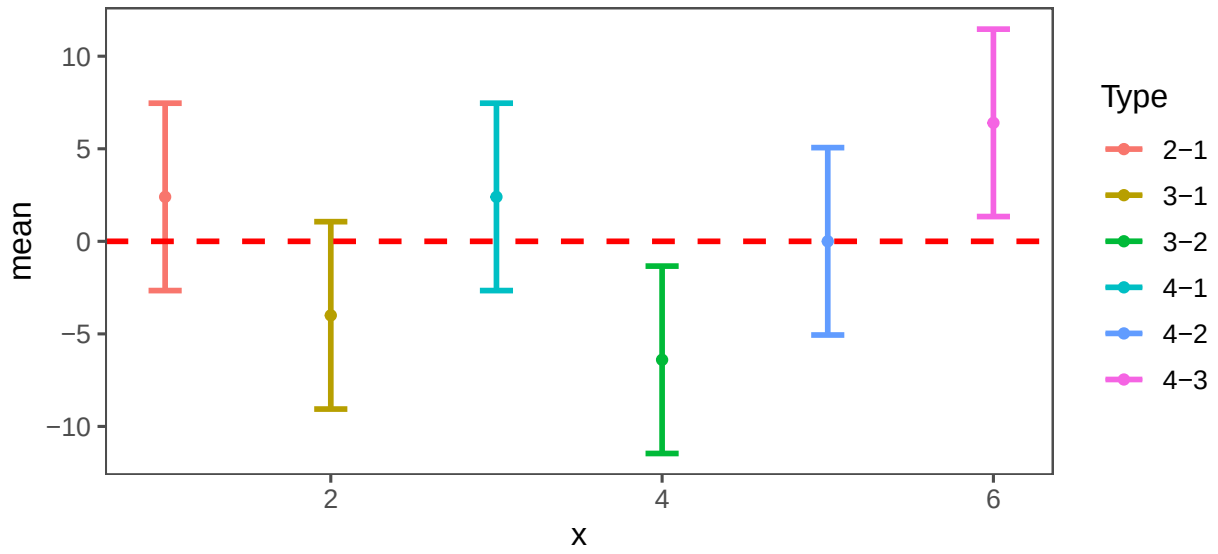
We found that the residuals are randomly distributed around 0. We can assume equal variance.

Multiple Comparisons

Now that we have concluded there is a difference in at least one of the means, we will use Tukey's Honesty Significant Difference method to determine where this difference takes place

	diff	lwr	upr	p adj
2-1	2.4	-2.662	7.462	0.5183
3-1	-4	-9.062	1.062	0.1417
4-1	2.4	-2.662	7.462	0.5183
3-2	-6.4	-11.46	-1.338	0.01269
4-2	0	-5.062	5.062	1
4-3	6.4	1.338	11.46	0.01269

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



We are 95% confident

• that the mean response losses in weight from process 2 is between -2.661664 and 7.461664 milligrams more than that of process 1, there is no statistical difference.

• that the mean response losses in weight from process 3 is between -9.061664 and 1.061664 milligrams more than that of process 1, there is no statistical difference.

• that the mean response losses in weight from process 4 is between -2.661664 and 7.461664 milligrams more than that of process 1, there is no statistical difference.

• that the mean response losses in weight from process 3 is between -11.461664 and -1.338336 milligrams more than that of process 2, there is a statistical difference.

• that the mean response losses in weight from process 4 is between -5.061664 and 5.061664 milligrams more than that of process 2, there is no statistical difference.

• that the mean response losses in weight from process 4 is between 1.338336 and 11.461664 milligrams more than that of process 3, there is a statistical difference.

Conclusion

The Randomized Complete Block Design is the model we should use. The batch variable does explain some of the variance for the processes. It indicates there is a difference in the means for the processes. After conducting multiple comparisons, we are confident that process 3 is the best due to the lowest loss in weight.