Search for Cause I

1 Question

The objective of this investigation is to determine the process step that is home to the dominant cause of variation in y300. We will compare y200 to y300 and y100 to y200 to see which part of the process the variation occurs.

2 Plan

Based on our baseline investigation, we expect to see the full extent of variation in the output in 15 total shifts. We sampled 30 random parts over the 15 shifts/5 days (2 parts per shift) by using random & systematic sampling. We measured the corresponding output characteristics in y_{100}, y_{200} and y_{300}. The total cost of this investigation was $750.

3 Data

Below is a view of our dataset. This table shows the first 6 values.

daycount	shift	partnum	y100	y200	y300
11	1	14670	-3.0	-3.0	-0.2
11	1	14698	3.2	3.4	6.0
11	2	15230	-5.4	-2.8	-1.8
11	2	15326	-1.6	0.4	-1.4
11	3	15452	1.0	-0.8	1.2
11	3	15614	-1.6	-2.4	-2.0

Table 1: Data

4 Analysis

4.1 Scatterplots

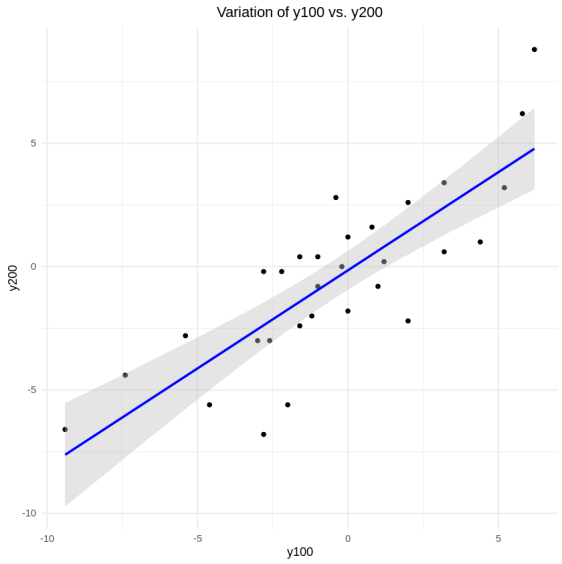

Figure 1: y_{100} vs y_{200}

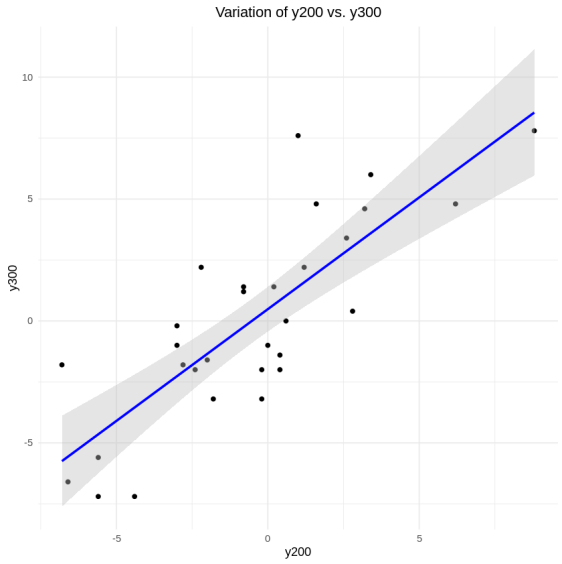

Figure 2: y_{200} vs y_{300}

The two scatter plots provided depict the relationships between pairs of variables, namely y100 vs. y200 and y200 vs. y300. Both graphs show a positive linear correlation, indicating that as one variable increases, so does the other. The regression lines, represented in blue, are accompanied by shaded areas that denote the confidence intervals, suggesting the degree of certainty in the predictions of the linear relationship. The narrower the shaded area, the more confident we can be about the relationship; and in both cases, these intervals are quite tight, which implies strong linear associations between the pairs of variables. Furthermore, the data points are consistently distributed around the regression line without any significant outliers, reinforcing the strength of the linear models. The range of values for y100, y200, and y300 across the graphs varies, with y300 displaying a wider spread than y100 and y200. This variation could reflect differing degrees of variability in the underlying data. Overall, the strong correlations and lack of significant outliers suggest that the linear models provide a good fit for the data, and could potentially be used for predictive purposes.

4.2 Model Summaries

```
Call:
lm(formula = y200 ~ y100, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4231 -1.3281  0.2274  1.5638  4.0198

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1503     0.3854  -0.390      0.7
y100          0.7952     0.1065   7.464 3.95e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.093 on 28 degrees of freedom
Multiple R-squared:  0.6655,    Adjusted R-squared:  0.6536
F-statistic: 55.71 on 1 and 28 DF,  p-value: 3.954e-08
```

Figure 3: Linear Model Between y_{100} and y_{200}

```
Call:
lm(formula = y300 ~ y200, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6500 -1.8954 -0.2663  1.4046  6.2008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4827     0.4479   1.078     0.29
y200          0.9165     0.1267   7.234 7.11e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.426 on 28 degrees of freedom
Multiple R-squared:  0.6515,    Adjusted R-squared:  0.639
F-statistic: 52.33 on 1 and 28 DF,  p-value: 7.108e-08
```

Figure 4: Linear Model Between y_{200} and y_{300}

Summary of the y200 ~ y100 Model: The first model predicts ‘y200‘ with ‘y100‘ as the independent variable. The estimated intercept is -0.1503, which is not statistically significant with a p-value of 0.7, implying that

the intercept is not different from zero in a statistically meaningful sense. The coefficient for 'y100' is 0.7952, with a p-value of 3.95×10^{-8}, which is highly significant, denoting a strong positive relationship; a unit increase in 'y100' is associated with a 0.7952 unit increase in 'y200'.

Residuals range from -4.4231 to 4.0198, and the interquartile range indicates a symmetrical distribution around the median. The model's R^2 is 0.6655, explaining 66.55% of the variance in 'y200'. The adjusted R^2 is 0.6536, which is slightly lower but still indicative of a good fit. The F-statistic is 55.71 with a p-value of 3.954×10^{-8}, confirming the model's overall significance.

Summary of the y300 $\sim$ y200 Model: In the second model, 'y300' is predicted by 'y200'. The intercept is 0.4827 and is not statistically significant (p-value = 0.29), indicating the intercept's contribution to the model at 'y200 = 0' is not significant. The 'y200' coefficient is 0.9165 with a p-value of 7.11×10^{-8}, signifying a very strong positive linear relationship.

The residuals have a range from -3.6500 to 6.2088, with quartiles that suggest a symmetrical distribution around the median. The R^2 for this model is 0.6515, indicating that 65.15% of the variance in 'y300' can be accounted for by 'y200'. The adjusted R^2 is 0.639, reflecting a good fit after accounting for the number of predictors. The F-statistic is 52.33 with a p-value of 7.108×10^{-8}, further confirming the model's statistical significance.

Both models exhibit strong and significant relationships between their respective variables. The high R^2 values in both models suggest a significant proportion of variability in the dependent variables is explained by the models. The non-significant intercepts indicate they do not play a significant role in these models. The substantial F-statistics and their corresponding p-values strongly suggest that the observed relationships are not due to random variation.

4.3 ANOVAs

Table 2: ANOVA Table for Response y_{200}

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
y_{100}	1	243.98	243.980	55.713	3.954e-08
Residuals	28	122.62	4.379		

Table 3: ANOVA Table for Response y_{300}

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
y_{200}	1	307.95	307.955	52.334	7.108e-08
Residuals	28	164.76	5.884		

In Table 2, the ANOVA for the response y200 shows that the regression model with y100 explains a significant portion of the variance in y200. The Sum of Squares for y100 is 243.98, which is the variation explained by the model, and the residual Sum of Squares is 122.62, indicating the unexplained variation. The Mean Square, calculated as the Sum of Squares divided by the degrees of freedom, yields 243.98 for y100 and 4.379 for the residuals. The F value, a ratio of these Mean Squares, is 55.713, suggesting that the model with y100 significantly improves the prediction of y200 compared to the mean of y200. This is further supported by the extremely low p-value of 3.95e-08, indicating that the probability of seeing such an F value if y100 had no effect is virtually zero.

Table 3 presents the ANOVA for the response y300, where y200 serves as the predictor. The model's explained variance for y200 is substantial, with a Sum of Squares of 307.95 compared to the unexplained variance with a residual Sum of Squares of 164.76. The Mean Square for y200 is 307.955, representing the model's variance, and the residual Mean Square is 5.884, representing the variance of the errors. The F value is 52.334, meaning the model substantially improves the prediction of y300 over the mean. The associated p-value of 7.108e-08 again indicates a highly significant effect of y200 on y300.

Both ANOVA tables confirm the statistical significance of the respective models. The F values are high and the corresponding p-values are extremely low, well below the conventional threshold of 0.05 for statistical significance. This implies that the relationships between the predictor variables (y100 and y200) and the response variables (y200 and y300) are unlikely to be due to random chance, affirming the predictors' relevance in explaining the variability of the responses.

4.4 Computed Ratios

We use the following ratio detect if a subprocess is home to a dominant cause

$$\text{Ratio} = \frac{\sqrt{\text{var}(y_i) - \sigma^2_{\text{model}}}}{\text{sd}(y_i)}$$

4.4.1 y200 vs y100

$$\text{sd}(y_{200}) = 3.555$$

$$\sigma^2_{\text{model}} = 4.38$$

$$\text{Ratio} = \frac{\sqrt{3.55^2 - 4.38}}{3.555} = 0.8084$$

$$\frac{\sigma_{\text{y100}}}{\sigma_{\text{y200}}} = 1.02585$$

4.4.2 y200 vs y300

$$\text{sd}(y_{300}) = 4.0374$$

$$\sigma^2_{\text{model}} = 5.884$$

$$\text{Ratio} = \frac{\sqrt{4.0374^2 - 5.884}}{4.0374} = 0.7994$$

$$\frac{\sigma_{\text{y200}}}{\sigma_{\text{y300}}} = 0.8806$$

5 Conclusion

Because both ratios are greater than 0.5, we conclude that both sub-processes are not home to a dominant cause.

The computed ratio for y_{200} vs y_{100} is 0.8084, and the ratio of standard deviations $\frac{\sigma_{y_{100}}}{\sigma_{y_{200}}}$ is 1.02585. These metrics suggest that while there is a significant relationship between y_{100} and y_{200}, as indicated by the robust linear model and its associated statistics (high R^2, low p-value), the process step from y_{100} to y_{200} does not exhibit a dominant cause of variation. The ratio being less than 1 but close to it indicates that the process variation in y_{200} relative to y_{100} is present but not overwhelmingly dominant.

For y_{300} vs y_{200}, the computed ratio is 0.7994, and the ratio of standard deviations $\frac{\sigma_{y_{200}}}{\sigma_{y_{300}}}$ is 0.8806. Similar to the y_{200} vs y_{100} comparison, the relationship between y_{200} and y_{300} is significant, which is also supported by a strong linear relationship, high R^2 value, and a very low p-value. The ratio being less than 1 suggests that there is an increase in process variation in the step from y_{200} to y_{300}; however, it is not so large as to indicate a dominant cause of variation.

Neither comparison (y_{200} vs y_{100} or y_{300} vs y_{200}) shows a ratio significantly greater than 1, which would indicate a dominant source of variation in the process. Both steps show increases in variability, with the step from y_{200} to y_{300} showing a slightly higher ratio, suggesting a bit more increase in variation at this step, but not overwhelmingly so.