

2 Way ANOVA

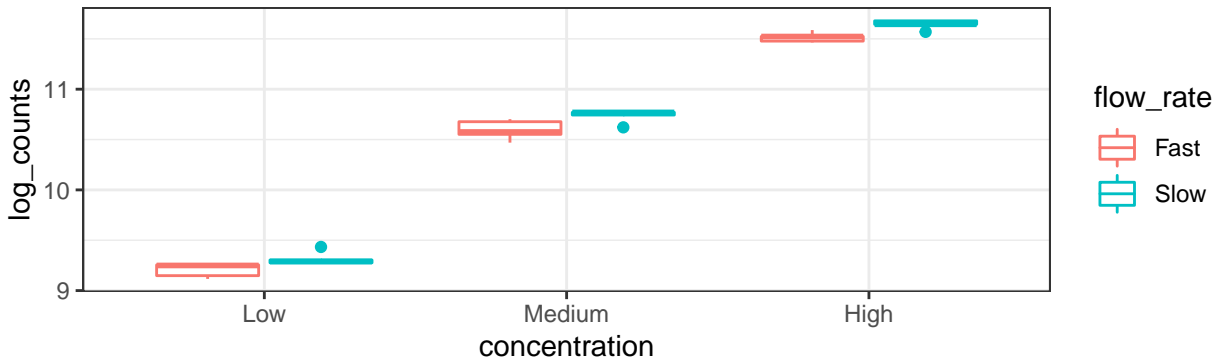
Nov. 4 2019 – Highlights from Sleuth3 Chapter 13

Here are two examples of two-way ANOVA problems (quantitative response, two categorical explanatory variables):

Example 1: a calibration experiment was performed to explore the relationship between:

- the recorded **counts** from a gas chromatograph (response) – we use log counts to stabilize variance across groups
- **concentration** of a compound (Low, Medium, or High) and **flow_rate** through the chromatograph (Slow or Fast)

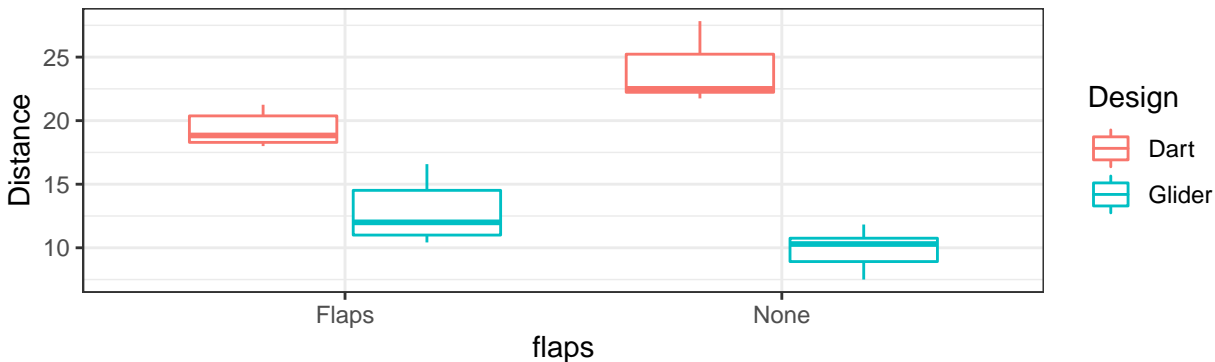
```
ggplot(data = chromatography, mapping = aes(x = concentration, y = log_counts, color = flow_rate)) +  
  geom_boxplot() +  
  theme_bw()
```



Example 2: A motivated paper airplane thrower recorded the following for each of 32 plane tosses:

- The **Distance** travelled (response)
- The **Design** (dart or glider) and whether or not flaps were put on the wings (Flaps or None)

```
ggplot(data = planes, mapping = aes(x = flaps, y = Distance, color = Design)) +  
  geom_boxplot() +  
  theme_bw()
```



1. For each of the examples above would an additive model or an interactions model be more appropriate? How can you tell based on the plots?

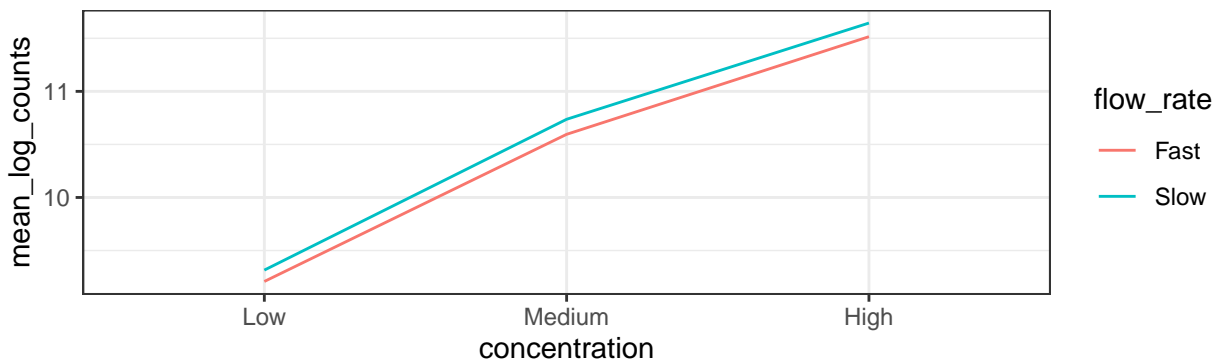
Here is another way to represent the data in the examples above that is sometimes used.

```
chromatography_group_means <- chromatography %>%
  group_by(concentration, flow_rate) %>%
  summarize(
    mean_log_counts = mean(log_counts)
  )
```

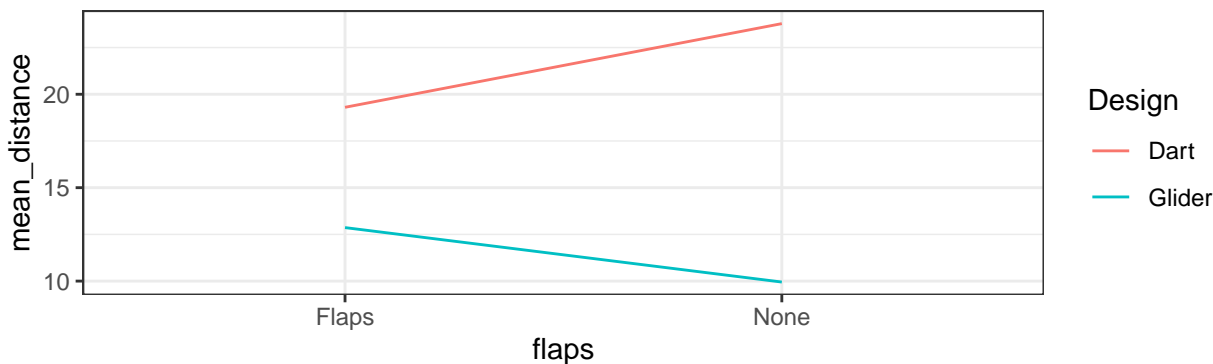
```
chromatography_group_means
```

```
## # A tibble: 6 x 3
## # Groups:   concentration [3]
##   concentration flow_rate mean_log_counts
##   <fct>         <chr>         <dbl>
## 1 Low          Fast           9.21
## 2 Low          Slow           9.32
## 3 Medium       Fast          10.6
## 4 Medium       Slow          10.7
## 5 High         Fast          11.5
## 6 High         Slow          11.6
```

```
ggplot(data = chromatography_group_means,
  mapping = aes(x = concentration, y = mean_log_counts, color = flow_rate, group = flow_rate)) +
  geom_line() +
  theme_bw()
```



Code suppressed, but I did the same thing for the second example:



The line segments will be parallel if an additive model is appropriate, and have different slopes if an interactions model is appropriate.

I'm showing you this because you will see it in other places, but I do not prefer this display for two reasons:

1. It "hides" variability around the mean, which is a critical part of the model!!
2. It uses lines to connect values of a discrete variable along the horizontal axis. Visually, lines imply variable represented on the horizontal axis is continuous. But it's discrete! There is not necessarily a meaningful in-between in the middle of "flaps" and "no flaps" – and even if there was, no guarantee the relationship with the response follows a linear path between those points.

Additive Model: `lm(response ~ explanatory1 + explanatory2, data = data)`

Fit group means on transformed scale

```
additive_fit <- lm(log_counts ~ concentration + flow_rate, data = chromatography)
summary(additive_fit)
```

```
##
## Call:
## lm(formula = log_counts ~ concentration + flow_rate, data = chromatography)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.133752 -0.050017  0.004214  0.048191  0.108745
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.19898    0.02384 385.810 < 2e-16 ***
## concentrationMedium 1.40414    0.02920 48.084 < 2e-16 ***
## concentrationHigh   2.31775    0.02920 79.370 < 2e-16 ***
## flow_rateSlow       0.12576    0.02384  5.274 1.63e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0653 on 26 degrees of freedom
## Multiple R-squared:  0.996, Adjusted R-squared:  0.9955
## F-statistic: 2140 on 3 and 26 DF, p-value: < 2.2e-16
```

2. What model did we just fit? Define all explanatory variables in your model statement.

3. Express the population mean for each combination of levels for the explanatory variables in terms of coefficients from the model in part 2.

concentration = “Low”, flow_rate = “Fast”:

concentration = “Medium”, flow_rate = “Fast”:

concentration = “High”, flow_rate = “Fast”:

concentration = “Low”, flow_rate = “Slow”:

concentration = “Medium”, flow_rate = “Slow”:

concentration = “High”, flow_rate = “Slow”:

4. Find and interpret a confidence interval for β_3 .

```
confint(additive_fit)
```

```
##                2.5 %    97.5 %  
## (Intercept)    9.14996912 9.2479903  
## concentrationMedium 1.34411633 1.4641673  
## concentrationHigh  2.25772204 2.3777730  
## flow_rateSlow     0.07674864 0.1747698
```

5. Conduct a test of the claim that for a given flow rate, the mean log counts is the same at all three concentrations.

```
fit_flow_only <- lm(log_counts ~ flow_rate, data = chromatography)  
anova(fit_flow_only, additive_fit)
```

```
## Analysis of Variance Table  
##  
## Model 1: log_counts ~ flow_rate  
## Model 2: log_counts ~ concentration + flow_rate  
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)  
## 1      28 27.3717  
## 2      26  0.1109  2    27.261 3196.8 < 2.2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions Model: `lm(response ~ explanatory1 * explanatory2, data = data)`

In this case, it's clear from the plot that an additive model is good enough. The following is just to demonstrate the set up and ideas. (I would not do the analysis below in real life!)

Fit group means on transformed scale

```
interactions_fit <- lm(log_counts ~ concentration * flow_rate, data = chromatography)
summary(interactions_fit)
```

```
##
## Call:
## lm(formula = log_counts ~ concentration * flow_rate, data = chromatography)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.125842 -0.040647  0.003796  0.040590  0.118114
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.20835     0.03018  305.068  <2e-16 ***
## concentrationMedium  1.38686     0.04269   32.489  <2e-16 ***
## concentrationHigh    2.30692     0.04269   54.042  <2e-16 ***
## flow_rateSlow        0.10702     0.04269    2.507   0.0193 *
## concentrationMedium:flow_rateSlow  0.03456     0.06037    0.572   0.5723
## concentrationHigh:flow_rateSlow    0.02166     0.06037    0.359   0.7229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06749 on 24 degrees of freedom
## Multiple R-squared:  0.996, Adjusted R-squared:  0.9952
## F-statistic: 1202 on 5 and 24 DF, p-value: < 2.2e-16
```

6. What model did we just fit? Define all explanatory variables in your model statement.

7. Express the population mean for each combination of levels for the explanatory variables in terms of coefficients from the model in part 2.

concentration = "Low", flow_rate = "Fast":

concentration = "Medium", flow_rate = "Fast":

concentration = "High", flow_rate = "Fast":

concentration = "Low", flow_rate = "Slow":

concentration = "Medium", flow_rate = "Slow":

concentration = "High", flow_rate = "Slow":

8. Give the interpretations of the following coefficient estimates in the interactions model:

- $\hat{\beta}_0 = 9.208$:

- $\hat{\beta}_1 = 1.387$:

- $\hat{\beta}_3 = 0.107$:

- $\hat{\beta}_4 = 0.035$:

9. Conduct a test of the claim that the difference in mean log counts between the “Slow” and “Fast” flow rates is the same across the Low and Medium concentrations.

10. Does the result of your hypothesis test in part 9 prove that the difference in mean log counts between the “Slow” and “Fast” flow rates is the same across the Low and Medium concentrations?