Design of Experiments: Project

Anthony Anderson

April 18, 2019

a) Does the "location" variable affect the duration of seal lion calls?

```
seal.lions <- read.csv("sealion_bark.csv")
seal.lions$location<- as.factor(seal.lions$location)
seal.fit <-aov(duration~location, data=seal.lions)
summary(seal.fit)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## location       6  98545   16424   22.47 <2e-16 ***
## Residuals   1233 901080     731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$; $i = 1...7$, $j = 1...n_i$
Where μ is the overall mean, each α_i is the main effect of each treatment corresponding to the various locations in the location factor, and the last term is the irreducible error, which is assumed to be distributed iid standard normal.
Null hypothesis: H_0: $\alpha_i = 0$
The extremely small p-value of the F-statistic suggests we would reject the null hypothesis, so location does have a significant effect on prediction the duration of seal lion barks.

b) Is there significant interaction between temperature and pressure to predict the foam index? Which of the two explains more variability, and thus is "more important"?

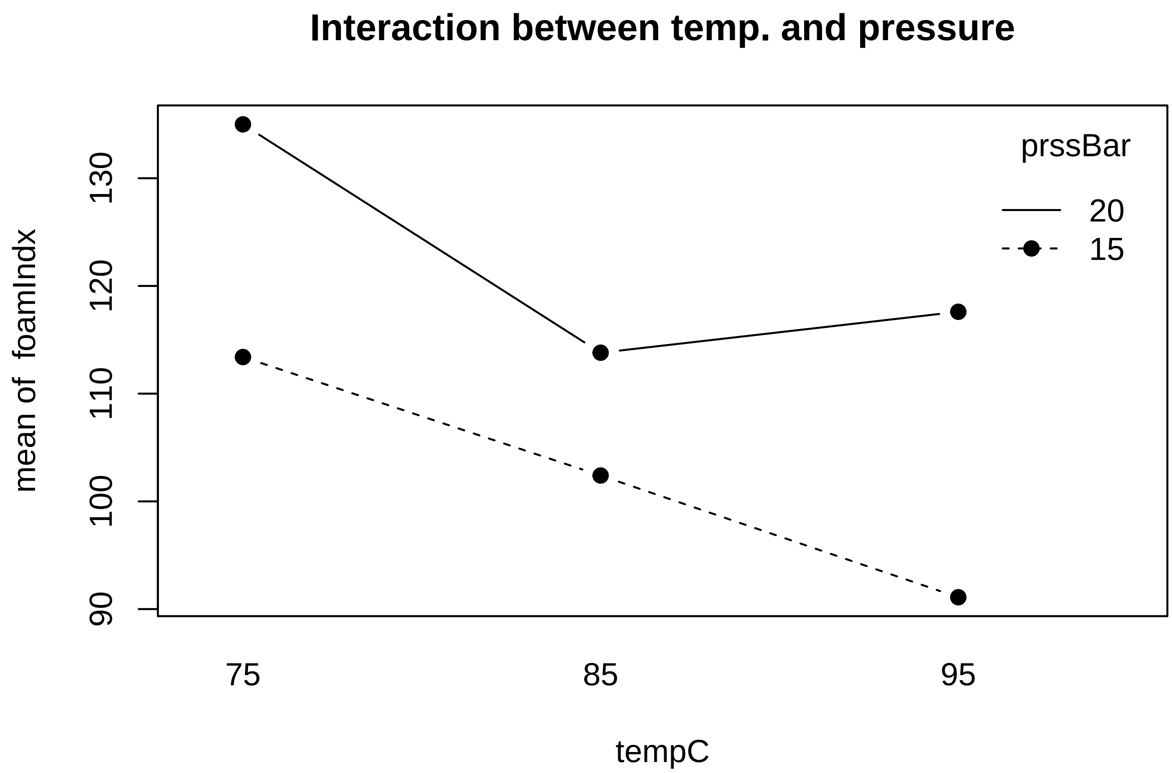
```
espresso.data <- read.csv("espresso2.csv")
espresso.data$trt_id<- as.factor(espresso.data$trt_id)
espresso.data$tempC<- as.factor(espresso.data$tempC)
espresso.data$prssBar<- as.factor(espresso.data$prssBar)
table(espresso.data$tempC,espresso.data$prssBar)
```

```
##
##      15 20
##   75  9  9
##   85  9  9
##   95  9  9
```

```
espresso.fit <- aov(foamIndx~tempC*prssBar,data=espresso.data)
summary(espresso.fit)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## tempC          2   4004    2002   5.491 0.007123 **
## prssBar        1   5310    5310  14.564 0.000388 ***
## tempC:prssBar  2    534     267   0.732 0.486075
## Residuals     48  17501     365
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
attach(espresso.data)
interaction.plot(tempC,prssBar,foamIndx,type="b",
                 main="Interaction between temp. and pressure", pch=19)
```




```r
detach(espresso.data)
```

Model: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$; $i = 1...3$, $j = 1...2$, $k = 1...9$
Where μ is the overall mean, each α_i is the main effect of each level of temperature, each β_j is the main effect of each level of pressure, then their interaction term, and the last term is the irreducible error, which is assumed to be distributed iid standard normal. We also assume that the variance for each effect is unique, so $\text{Var}(\alpha_i) = \sigma^2_\alpha$, $\text{Var}(\beta_j) = \sigma^2_\beta$, $\text{Var}((\alpha\beta)_{ij}) = \sigma^2_{\alpha\beta}$
Variance components:
$\sigma^2_\alpha = \frac{MS_A - MS_{AB}}{bn} = (2002 - 267)/2(9) = 96.389$
$\sigma^2_\beta = \frac{MS_B - MS_{AB}}{an} = (5310 - 267)/3(9) = 186.778$
$\sigma^2_{\alpha\beta} = \frac{MS_{AB} - MS_E}{n} = (267 - 365)/9 = -10.89$
$\sigma^2_E = MS_E = 365$
Null hypothesis: H_0: $(\alpha\beta)_{ij} = 0$
Recalculated F-stats, where A signifies temperature, B is pressure, AB their interaction, and E the error:
$F_A = \frac{MS_A}{MS_{AB}} = 2002/267 = 7.5$
$F_B = \frac{MS_B}{MS_{AB}} = 5310/267 = 19.9$
$F_{AB} = \frac{MS_{AB}}{MS_E} = 267/365 = .73$
Comparatively, the F-stat under the null hypothesis is $F_{.05,2,50} = 3.18$
The recalculated F-stats suggest that the interaction effect is small and non significant, since it's F-value is much smaller than the F-stat under the null hypothesis. Moreover, the F-statistics and anova output suggests that pressure has a greater effect on the foam index of the espresso than temperature, but both are very significant.

c) Does the randomness of the horse/rider variable affect the main effect of seat position?

```r
horse.data <- read.csv("horse.csv")
horse.data$position<-as.factor(horse.data$position)
horse.data$combo<-as.factor(horse.data$combo)
horse.fit<-aov(stride~position*combo,data=horse.data)
summary(horse.fit)
```

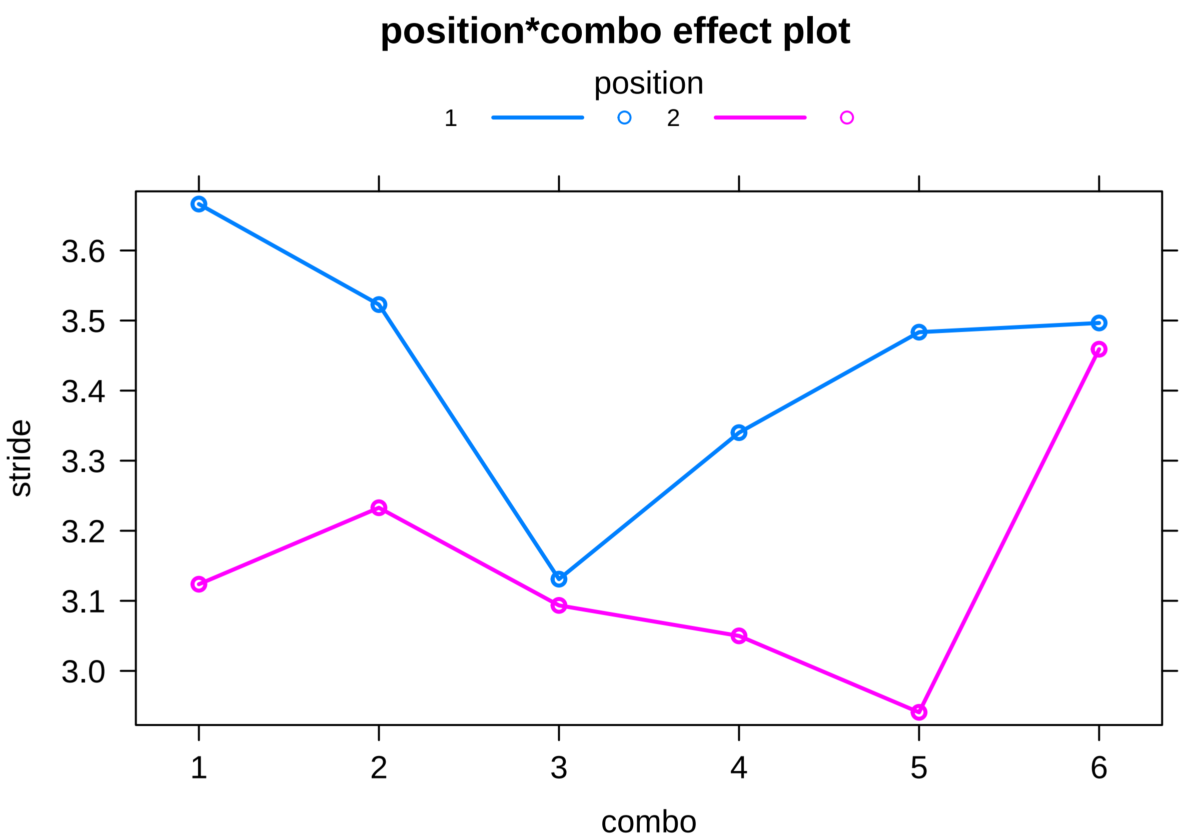
```
##                 Df Sum Sq Mean Sq F value   Pr(>F)    
## position         1 0.7570  0.7570  60.636 5.07e-08 ***
## combo            5 0.6035  0.1207   9.668 3.73e-05 ***
## position:combo   5 0.3827  0.0765   6.131 0.000855 ***
## Residuals       24 0.2996  0.0125                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
library(effects)
```

```
## Loading required package: carData

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
plot(effect("position:combo",horse.fit,,list(combo=c(seq(1,6)))),multiline=TRUE)
```



Model: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$; $i = 1, 2$, $j = 1...6$, $k = 1, 2, 3$
Where μ is the overall mean, each α_i is the main effect of each level of seat position (fixed), each β_j is the main effect of each level of rider/horse combination (random), then their interaction term, and the last term is the irreducible error, which is assumed to be distributed iid standard normal. We also assume that the variance for each random effect is unique, so $\mathrm{Var}(\beta_j) = \sigma^2_\beta$, and $\mathrm{Var}(\alpha\beta_{ij}) = \frac{a-1}{a}\sigma^2_{\alpha\beta}$, so the interaction term

is not independent since it depends on the level of the position variable.
Variance components:
$\sigma^2_\beta = \frac{MS_B - MS_E}{an} = (.1207 - .0125)/2(3) = .018$
$\sigma^2_{\alpha\beta} = \frac{MS_{AB} - MS_E}{n} = (.0765 - .0125)/3 = .021$
$\sigma^2_E = MS_E = .0125$
Null hypothesis: H_0: $(\alpha\beta)_{ij} = 0$
Recalculated F-stats, where A signifies temperature, B is pressure, AB their interaction, and E the error:
$F_A = \frac{MS_A}{MS_{AB}} = .757/.0765 = 9.9$
$F_B = \frac{MS_B}{MS_E} = .1207/.0125 = 9.67$
$F_{AB} = \frac{MS_{AB}}{MS_E} = .0765/.0125 = 6.12$
F-stat under null hypothesis: $F_{.05,5,24} = 2.62$

The interaction plot shows different slopes, so it's clear that interaction between seat position and combo is present. Also, the interaction term's F-stat is larger than the F-stat under the null, so we would reject the null hypothesis. The combination of horse/rider will have a significant effect on seat position and vice-versa, they interact.

d) Is poker mostly skill based, or luck of the draw? (Is the hand you have going to affect your final take home more than being skilled at poker?)

```
poker.data<- read.csv("poker.csv")
poker.data$skill<- as.factor(poker.data$skill)
poker.data$hand<- as.factor(poker.data$hand)
poker.data$limit<- as.factor(poker.data$limit)
poker.fit<- aov(final~skill*hand*limit, data=poker.data)
summary(poker.fit)
```

```
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## skill               1     49    49.2   2.839 0.09308 .
## hand                2   2647  1323.3  76.412 < 2e-16 ***
## limit               1     32    31.7   1.829 0.17726
## skill:hand          2    219   109.5   6.324 0.00205 **
## skill:limit         1    119   119.1   6.878 0.00919 **
## hand:limit          2     97    48.6   2.809 0.06192 .
## skill:hand:limit    2     42    21.2   1.224 0.29565
## Residuals         288   4987    17.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model: $y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$; $i = 1, 2,\ j = 1, 2, 3,\ k = 1, 2$
Where μ is the overall mean, each α_i is the main effect of each level of skill, each β_j is the main effect of each level of hand, each γ_k is the main effect of each level of limit, then all their interaction terms, and the last term is the irreducible error, which is assumed to be distributed iid standard normal.
Null hypothesis: H_0: $\alpha_i = 0$
The above output shows us that the main effect of "skill" is not very significant, and doesn't explain much variability in the "final" response. Instead, the hand predictor explains a very large amount of variation and is found to be significant, with a p-value of essentially 0. There is significant interaction between skill and hand, as well as skill and limit, although again they do not explain nearly as much variability as the main effect of hand.

e) Caffeine is usually avoided for physical activity because it's a diuretic, it supposedly dehydrates you and so you will tend to not perform as well. Are higher levels of caffeine associated with lower endurance times?

```
caffeine.data<- read.csv("caffeine.csv")
caffeine.data$subject<-as.factor(caffeine.data$subject)
```

```r
caffeine.data$dose<-as.factor(caffeine.data$dose)
caffeine.fit<- aov(time~dose+subject, data=caffeine.data)
summary(caffeine.fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## dose         3    933   311.0   5.917  0.00359 ** 
## subject      8   5558   694.7  13.216 4.17e-07 ***
## Residuals   24   1262    52.6                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(caffeine.fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = time ~ dose + subject, data = caffeine.data)
## 
## $dose
##             diff       lwr       upr     p adj
## 5-0   11.2366667  1.808030 20.665303 0.0153292
## 9-0   12.2411111  2.812474 21.669748 0.0076616
## 13-0  11.7088889  2.280252 21.137526 0.0110929
## 9-5    1.0044444 -8.424192 10.433081 0.9909369
## 13-5   0.4722222 -8.956414  9.900859 0.9990313
## 13-9  -0.5322222 -9.960859  8.896414 0.9986162
## 
## $subject
##        diff        lwr        upr     p adj
## 2-1  23.5875   6.161425  41.013575 0.0030722
## 3-1  26.1000   8.673925  43.526075 0.0009323
## 4-1  13.1150  -4.311075  30.541075 0.2549156
## 5-1  15.4675  -1.958575  32.893575 0.1103037
## 6-1  29.5050  12.078925  46.931075 0.0001850
## 7-1   1.1325 -16.293575  18.558575 0.9999997
## 8-1  19.8925   2.466425  37.318575 0.0170160
## 9-1  -8.7025 -26.128575   8.723575 0.7423114
## 3-2   2.5125 -14.913575  19.938575 0.9998747
## 4-2 -10.4725 -27.898575   6.953575 0.5311020
## 5-2  -8.1200 -25.546075   9.306075 0.8040672
## 6-2   5.9175 -11.508575  23.343575 0.9586954
## 7-2 -22.4550 -39.881075  -5.028925 0.0052322
## 8-2  -3.6950 -21.121075  13.731075 0.9979401
## 9-2 -32.2900 -49.716075 -14.863925 0.0000501
## 4-3 -12.9850 -30.411075   4.441075 0.2657958
## 5-3 -10.6325 -28.058575   6.793575 0.5118647
## 6-3   3.4050 -14.021075  20.831075 0.9988407
## 7-3 -24.9675 -42.393575  -7.541425 0.0015977
## 8-3  -6.2075 -23.633575  11.218575 0.9462013
## 9-3 -34.8025 -52.228575 -17.376425 0.0000158
## 5-4   2.3525 -15.073575  19.778575 0.9999236
## 6-4  16.3900  -1.036075  33.816075 0.0766574
## 7-4 -11.9825 -29.408575   5.443575 0.3601341
## 8-4   6.7775 -10.648575  24.203575 0.9147625
```

```
## 9-4 -21.8175 -39.243575  -4.391425 0.0070435
## 6-5  14.0375  -3.388575  31.463575 0.1867695
## 7-5 -14.3350 -31.761075   3.091075 0.1680887
## 8-5   4.4250 -13.001075  21.851075 0.9930235
## 9-5 -24.1700 -41.596075  -6.743925 0.0023325
## 7-6 -28.3725 -45.798575 -10.946425 0.0003163
## 8-6  -9.6125 -27.038575   7.813575 0.6357055
## 9-6 -38.2075 -55.633575 -20.781425 0.0000034
## 8-7  18.7600   1.333925  36.186075 0.0281653
## 9-7  -9.8350 -27.261075   7.591075 0.6086662
## 9-8 -28.5950 -46.021075 -11.168925 0.0002846
```

Model: $y_{ij} = \mu + \alpha_i + \beta_j + +\epsilon_{ij}$; $i = 1, 2,\ j = 1...9$ Where μ is the overall mean, each α_i is the main effect of each level of dosage , each β_j is the main effect of each level of subject (each person, and in this case the blocking variable), and the last term is the irreducible error, which is assumed to be distributed iid standard normal.
Contrasts from the output suggest that the presence of caffeine does have an effect on endurance, the first 3 contrasts have p-values less than .05 suggesting they are significant. However, it seems that as the amount of caffeine increases, the effect lessens. That is to say, regardless if a subject was given 5, 9, or 13 mg of caffeine, there was not much difference between them in terms of endurance time. The presence of caffeine has an effect, but increasing the amount does not seem to have an increasing linear relationship with endurance time.

f) Wine is very particular, and the most avid fans will insist that the grapes and area the wine comes from has massive effects on the taste. In fact, "Champagne" is specifically certified as Champagne only if it grows in the Champagne region of France. In today's globalized world, is the country of origin still important to consumers?

```r
wine<- read.csv("wine.csv")
wine$weeks<-as.factor(wine$weeks)
wine$label<-as.factor(wine$label)
wine$country<-as.factor(wine$country)
wine.fit<-aov(score~country+weeks+label, data=wine)
summary(wine.fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## country      3 1937.7   645.9   9.740 0.0101 *
## weeks        3  729.2   243.1   3.665 0.0824 .
## label        3  414.7   138.2   2.085 0.2037
## Residuals    6  397.9    66.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$; $i, j, k = 1...4$
Where μ is the overall mean, each α_i is the main effect of each country, each β_j is the main effect of each week, each γ_k is the main effect of each label, and the last term is the irreducible error, which is assumed to be distributed iid standard normal.

Country explains the most variability in consumer's scores, suggesting that the country of origin is still an important factor in the taste of wines. The differences in consumers by week is the next largest explanation, and the label the wine has is the least. In fact, both weeks and label have a p-value larger than .05, suggesting their effects on consumer scores are not significant.