Lesson 12: Pistachios (3.2)

LTC James K. Starling

Review (-2) What did we talk about in the previous lesson? Can you write out a mathematical representation of the model?

(-1) What is the null and alternative hypothesis with a multi-factor experiment?

In this chapter we're going to talk about *interactions* in multi-factor designs.

Background: Pistachios imported from the Middle East are cleaned and bleached with peroxide to turn them white. Some governments have banned bleaching over concerns of health risks; others have explored the impact on the health benefits of pistachios. Researchers (Gazor & Minaei, Dry. Technol., 2005) wanted to investigate the effects of the air velocity of the fan and the drying temperature of the oven on the amount of peroxide which remains (as a percentage) on the pistachios after the bleach process. Two values of air velocity (1.5 and 2.5 mph) and two values of drying temperature (60° and 90°F) were investigated. A full factorial design was conducted such that five batches, each consisting of 24 ounces of nuts, were randomly assigned to each treatment. We would like to determine the optimal settings of temperature and air velocity to minimize the percentage of peroxide remaining.

(1) Create a sources of variability diagram. What were the response variables and the explanatory variable(s)? How are these variables classified? How many treatments do we have?

(2) Are any of these explanatory variables blocking variables? Why or why not.

(3) We are treating temperature and air velocity as factors and not as continuous random variables. What is a downside of this?

(4) Conduct a two-variable ANOVA with a two-factor model (use effect coding). Calculate the percent of variation explained by each explanatory variable and the overall model.

```
# pist.dat2 <- pist.dat
# contrasts(pist.dat2$Temperature) <- contr.sum
# contrasts(pist.dat2$AirVelocity) <- contr.sum
#
# pist.lm <- lm(Peroxide~ xxx + xxx, data=pist.dat2)
# pist.anv <- anova(pist.lm);
# pist.anv
#
# SST <- xxx
# rsq.temp <- xxx; rsq.temp
# rsq.av <- xxx/SST; rsq.av
# rsq.mod <- xxx + xxx; rsq.mod
```

(5) What is the fitted, two-variable, statistical model?
Which effects are statistically significant? According to the model, which combination is the best? (lower `Peroxide` level is better)

```
# summary(pist.lm)
```

(6) According to the model, which combination is the best? (lower `Peroxide` level is better)

```
# pred.means <- data.frame(Temperature = c(60,   60,  90, 90),
#                          AirVelocity = c(1.5, 2.5, 1.5, 2.5),
#                          mn = c(NA, NA, NA, NA))
# pred.means[1,'mn'] <- predict(pist.lm, data.frame(Temperature=xxx,AirVelocity=xxx))
# pred.means[2,'mn'] <- xxx
# pred.means[3,'mn'] <- xxx
# pred.means[4,'mn'] <- xxx
# pred.means
```

(7) Calculate the means for each combination of Temperature and AirVelocity using the `group_by()` and `summarise()` commands. Is this contrary to our model predictions above?

```
# pist.means <- pist.dat %>% group_by(xxx, xxx) %>%
#   summarise(mn=mean(xxx)) %>% as.data.frame()
# pist.means
```

(8) This is evidence of a **Statistical Interaction**. An interaction means that the effect is modified by the presence of another variable. In our experiment, if the temperature is 60, then it'd be more advantageous to have an air velocity of 2.5, but if the temperature is 90, then it'd be more advantageous to have an air velocity of 1.5.

Note that we are not saying that Air Velocity and Temperature are confounders. If we know the air velocity do we gain any knowledge of the temperature?

(9) How do we update our sources of variation diagram now?
What is our updated statistical model? What is our null and alternative hypotheses?

(10) Use the `pist.means` dataframe and use the pipe command and `ggplot()` + `geom_line()` + `geom_point()` + (other stuff that makes sense) to create an interaction plot.

If there were no interaction effects, what would we expect the lines to look like?

```
# p <- pist.means %>% ggplot(aes(x=xxx, y=xxx, group= xxx, color=xxx)) +
#   geom_line() +
#   geom_point() +
#   labs(title="Interaction Plot", y="mean peroxide %")
# p
```

(11) Using the previous figure, add horizontal lines accounting for the effect of the air velocity and (separately) the effects of temperature. Use `geom_hline` to plot. What do you notice about dashed lines compared to the means?

```
# mod.mean <- coef(pist.lm)[xxx];
# temp.adj <- coef(pist.lm)[xxx];
# av.adj <- coef(pist.lm)[xxx];
#
# p + geom_hline(yintercept = mod.mean, linetype = "dashed") +
#   geom_hline(yintercept = mod.mean - av.adj, linetype = "dashed") +
#   geom_hline(yintercept = mod.mean + av.adj, linetype = "dashed") +
#   labs(title="Interaction Plot with effects of AirVelocity")
#
# p + geom_hline(yintercept = mod.mean, linetype = "dashed") +
#   geom_hline(yintercept = mod.mean - temp.adj, linetype = "dashed") +
#   geom_hline(yintercept = mod.mean + temp.adj, linetype = "dashed") +
#   labs(title="Interaction Plot with effects of Temperature")
```

(12) Using the `pist.means` table, calculate the `difference in the differences`. How do you interpret this answer?

```
# pist.means
#
# # diff of diffs degrees
# dd.deg <- (xxx - xxx) - (xxx - xxx)
# dd.deg
#
# # diff of diffs AirVelocity
# dd.av <- (xxx - xxx) - (xxx - xxx)
# dd.av
```

(12) To see if this statistic is significant, we can simulate the distribution under H_0 ($\alpha = \beta = 0$). How rare is it to observe something greater than 3.24?

```
# set.seed(376)
# options(dplyr.summarise.inform = FALSE)
#
# M <- 1000
# stats <- rep(NA, M)
# for(i in 1:M){
#   pist.dat.shuff <- pist.dat
#   pist.dat.shuff$temp.shuf <- sample(pist.dat.shuff$Temperature)
#   pist.dat.shuff$av.shuff <- sample(pist.dat.shuff$AirVelocity)
#   p.shuff <- pist.dat.shuff %>% group_by(temp.shuf, av.shuff) %>%
#     summarise(mean.perox=mean(Peroxide))
#   stats[i] <- (p.shuff$mean.perox[1] - p.shuff$mean.perox[2]) -
#               (p.shuff$mean.perox[3] - p.shuff$mean.perox[4])
# }
# as.data.frame(stats) %>% ggplot(aes(x=stats)) + geom_histogram(bins = 50) + geom_vline(xintercept = d
# sum(stats > dd.deg) / M
```

(13) Verify the numbers of the interaction effects by using a linear model. Where do the interaction effects show up here?

```
# pist.int.lm <- pist.dat2 %>% lm(xxx~xxx * xxx, data=.)
# summary(pist.int.lm)
# anova(pist.int.lm)
```

(14) How does the ANOVA output change from our first model `pist.lm` to the model with interactions `pist.int.lm`? What happens to the residuals (SSE, MSE)?

```
# anova(pist.lm)
# anova(pist.int.lm)
```

(15) We can now use the post-hoc test to see which of the interactions is significant.

```
# TukeyHSD(aov(xxx ~ xxx * xxx, data = pist.dat2))
```