Introduction to 2^2 Factorial Design

Dr Austin R Brown

Kennesaw State University

Introduction

- In the last section, we learned about factorial designs.
- Such a design is appropriate when we have multiple treatment factors (each with 2+ levels) and we want to study these main effects and the potential interactions between them.
- But let's consider some other situations where a full factorial design might initially make sense, but some alternatives may prove more effective and useful.

Introduction

- Remember in the Instagram Ads example, we had two treatment factors, Ad Content and Ad Format, each with two levels (funny/informational and image/video).
- This gave us a total of four treatment combinations.
- Suppose we wanted to rerun the experiment, but now we have a third treatment factor, Ad Color Palette, with three levels (blue, red, and green).
- This would give us a total of 12 treatment combinations.

Introduction

- Supposing we would want 10 replications of each treatment combination, this would require a total of 120 experimental units.
- This is an increase of 80 experimental units from the original design (supposing $r = 10$).
- In some cases, the cost of running these additional experiments and/or collecting these additional observations may not be overly high.
- But in other cases, it may be prohibitive.
  - So what do we do?

Introduction

- In experimental situations where we may be interested in screening a large number of factors or where the cost of running all factor combinations is too great in terms of time or money, a 2^k *factorial design* provides and efficient way to explore our main effects and interactions with fewer experiments.

- A 2^k design is used to study the effects of k factors, each at two levels (e.g., high and low), on some outcome/response.
  - It allows researchers to efficiently explore the main effects and interactions of multiple factors with a **manageable number of experimental runs**

Introduction

- The key here is taking a potentially complex design and simplifying it by having every treatment effect simply be a pairwise comparison.
- As we will see, it also simplifies computations quite a bit.
- Before we get into the general case, let's talk about the simplest case, the 2^2 factorial design.

Experimental Example

- A popular hot wings chain called **Buster's Blazing Wings** wants to develop a new mobile app to enhance customer engagement and boost online orders.

- The development team has identified two key factors that may impact **monthly purchases**
  1. Push Notifications (we'll call this factor A) - Enabled (+) or Disabled (-)
  2. Loyalty Program (we'll call this factor B) - Basic (points-based +) or Enhanced (includes exclusive discounts -)
     - Note, we will use $+/-$ notation to denote the levels of the main effects

- For each of the $2^2 = 4$ treatment combinations, suppose we randomly recruit $r = 3$ customers to pilot the app. After a month of using the app, we record their monthly purchases.
  - The data are stored in the `2^2 Wings Example.xlsx`

Considering the Context

- As we've done before, we can work out the roadmap of our experiment:
- We are trying to quantify the effect main effects (Push Notifications and Loyalty Program) as well as their interaction have on our outcome, monthly restaurant sales.
- You can imagine there likely being some lurking variables working with human participants, so hopefully we've done a good job sampling app users/customers who are relatively homogeneous.

Considering the Context

- Our hypotheses in this case would be the same as our hypotheses for the two-factor factorial design.
- We will discuss how the computations simplify in the 2^2 case.

Exploring the Data

- With such a small dataset, we can visually examine the structure in a tabular format:

A	B	Treatment Combo	Rep 1	Rep 2	Rep 3	Total
-	-	Enabled/Enhanced	151.80	179.99	154.92	486.71
-	+	Enabled/Points	136.15	141.49	152.41	430.05
+	-	Disabled/Enhanced	129.30	119.98	131.79	381.07
+	+	Disabled/Points	107.22	44.72	77.85	229.79

Table 1: 2^2 Factorial Design with Treatment Combinations and Replicates

Exploring the Data

- For the sake of notation, let:
  - Capital "A" denote the effect of factor A (Push Notifications in our case)
  - Capital "B" denote the effect of factor B (Loyalty Program in our case)
  - Capital "AB" denotes the interaction between the main effects.

- Additionally, let:
  - a represent the treatment combination of A at the $+$ level and B at the - level
  - b represent the treatment combination of A at the - level and B at the $+$ level
  - ab represents both factors at the $+$ level
  - (1) is used to denote both factors at the - level
  - We can use these symbols to denote the values in the *Total* column of Table 1

Exploring the Data: The Effect of A

- Our goal in a factorial design, 2^2 or otherwise, is to quantify the effect of a given factor or factors on the outcome.

- In a 2^2 factorial design, the average effect of a given factor (say A) is the change in the outcome/response when that factor's level changes ($+$ to - or vice versa), averaged across all levels of the other factor (say B).

Exploring the Data: The Effect of A

- The effect of A at the - level of B is (A at its + level less A at its - level while B is fixed at its - level averaged across the replicates):

$$\frac{a-(1)}{r}$$

- The effect of A at the + level of B is (A at its + level less A at its - level while B is fixed at its + level averaged across the replicates):

$$\frac{ab-b}{r}$$

Exploring the Data: The Effect of A

- Averaging these two quantifies yields the **main effect of A**:

$$A = \frac{1}{2}\left(\frac{a-(1)}{r} + \frac{ab-b}{r}\right)$$

$$A = \frac{1}{2r}(ab + a - b - (1))$$

Exploring the Data: The Effect of A

- So in our case:

$$A = \frac{1}{2(3)}(229.79+381.07-430.05-486.71) = \frac{-305.90}{6} = -50.98$$

- The effect of A is negatively signed suggesting that moving from the - level (Enabled Notifications) to the $+$ level (Disabled Notifications) is associated with a decrease in monthly spend.

Exploring the Data: The Effect of B

- We can do the exact same thing to quantify the effect of B. The effect of B at the - level of A is:

$$\frac{b-(1)}{r}$$

- The effect of B at the $+$ level of A is:

$$\frac{ab-a}{r}$$

Exploring the Data: The Effect of B

- Averaging these two quantifies yields the **main effect of B**:

$$B = \frac{1}{2r}(ab + b - a - (1))$$

- In our case, this works out to be:

$$B = \frac{1}{2(3)}(229.79 + 430.05 - 381.07 - 486.71) = \frac{-207.94}{6} = -34.66$$

Exploring the Data: The Effect of B

- Again, since the effect of B is negatively signed, this suggests that moving from the - level (Enhanced) to the $+$ level (Points) is associated with a decrease in monthly spend.

- Now, what about the interaction effect?

Exploring the Data: The Interaction Effect, AB

- We define the interaction effect, AB, as the average difference between the effect of A at the $+$ level of B and the effect of A at the - level of B:

$$AB = \frac{1}{2r}((ab - b) - (a - (1)))$$

- In our case:

$$AB = 12(3)(229.79 - 430.05 - 381.07 + 486.71) = \frac{-97.62}{6} = -15.77$$

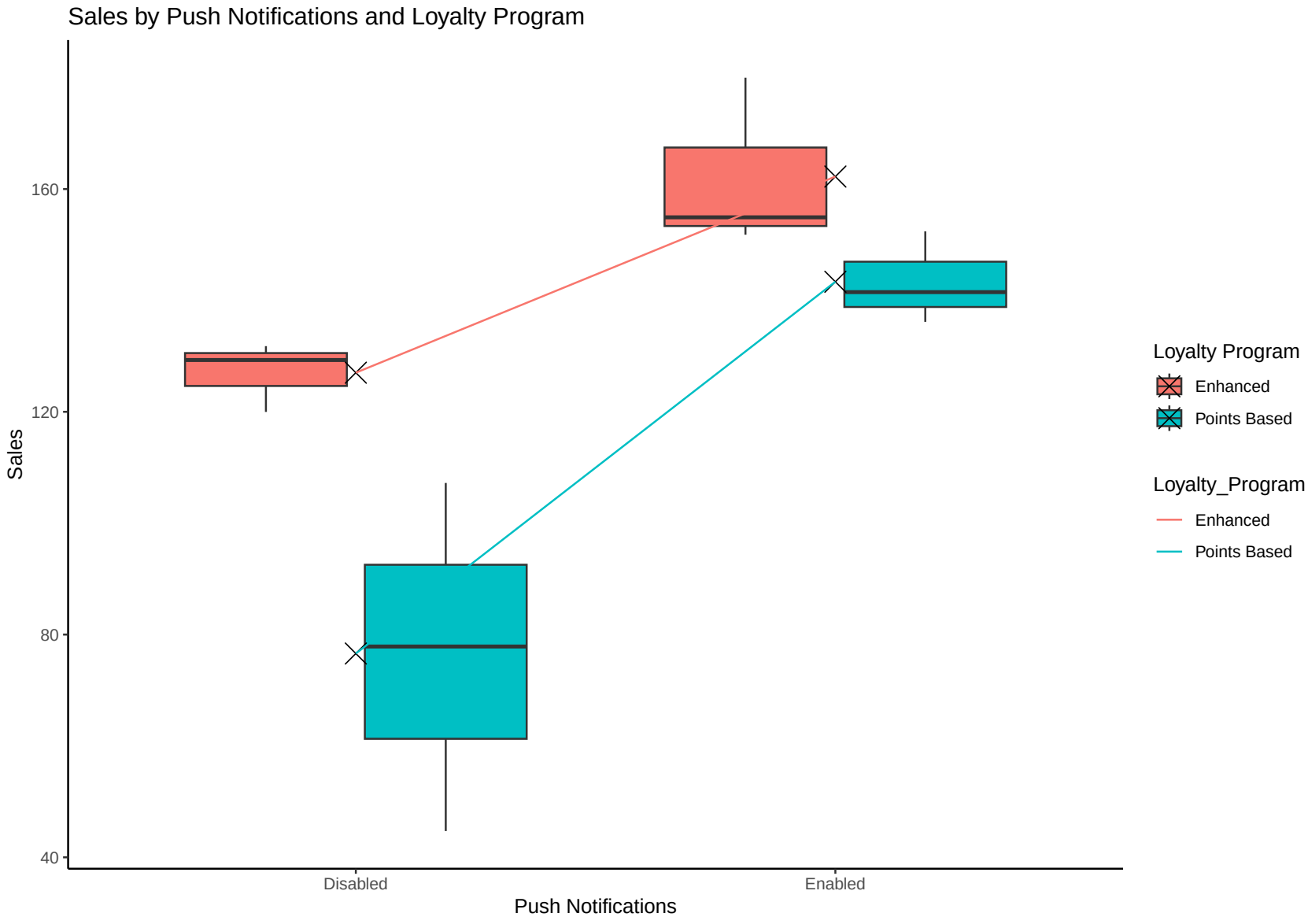
Exploring the Data: The Interaction Effect, AB

- Here, the interpretation is a little more involved.
- Since the sign is negative, this implies that as we move from the - level (Enabled Notifications) of A to the $+$ level of A (Disabled Notifications) while simultaneously moving from the - level (Enhanced) level of B to the $+$ level of B (Points), we would expect to see a decrease in monthly spend.
- This is best observed in our boxplot/interaction combination plot:

Exploring the Data: The Interaction Effect, AB

```
library(tidyverse)
library(readxl)
library(rstatix)
## Read in the Data ##
wings <- read_excel("2^2 Wings Example.xlsx")
## Boxplot/Interaction Combo Plot ##
wings |>
  ggplot(aes(x = Push_Notifications, y = Sales, fill = Loyalty_Program)) +
  geom_boxplot() +
  geom_point(data = wings |>
               group_by(Push_Notifications,Loyalty_Program) |>
               get_summary_stats(Sales,type='mean_sd'),
             aes(x=Push_Notifications,y=mean,group=Loyalty_Program),
             shape=4,size=5) +
  geom_line(data = wings |>
              group_by(Push_Notifications,Loyalty_Program) |>
              get_summary_stats(Sales,type='mean_sd'),
            aes(x=Push_Notifications,y=mean,color=Loyalty_Program,
                group=Loyalty_Program)) +
  labs(title = "Sales by Push Notifications and Loyalty Program",
       x = "Push Notifications",
       y = "Sales",
       fill = "Loyalty Program") +
  theme_classic()
```

Exploring the Data: The Interaction Effect, AB



Evaluating the ANOVA Model

- Contextually, we can see evidence that there may be a significant push notification effect as well as loyalty program effect.

- To evaluate this statistically, we can perform the same tests that we did for regular factorial ANOVA. For our A factor:

$$H_0 : \alpha_{\text{Enabled}} = \alpha_{\text{Disabled}} = 0$$

$$H_1 : \text{At least one } \alpha_i \neq 0$$

Evaluating the ANOVA Model

- For our B factor:

$$H_0 : \beta_{\text{Enhanced}} = \beta_{\text{Points}} = 0$$

$$H_1 : \text{At least one } \beta_j \neq 0$$

Evaluating the ANOVA Model

- And for the interaction, AB:

$$H_0 : (\alpha\beta)_{EE} = (\alpha\beta)_{EP} = (\alpha\beta)_{DE} = (\alpha\beta)_{DP} = 0$$

$$H_1 : \text{At least one } (\alpha\beta)_{ij} \neq 0$$

Evaluating the ANOVA Model

- Now, using R, we can fit the model, check our assumptions (see the code), and interpret the results:

```
## Fit the Model ##
wings_mod <- aov(Sales~Push_Notifications*Loyalty_Program,
                 data=wings)
## Evaluate Results ##
library(broom)
wings_mod |>
  tidy()
```

```
# A tibble: 4 x 6
  term                                df sumsq meansq statistic  p.value
  <chr>                            <dbl> <dbl>  <dbl>     <dbl>    <dbl>
1 Push_Notifications                   1 7798.  7798.      23.6   0.00127
2 Loyalty_Program                      1 3603.  3603.      10.9   0.0109
3 Push_Notifications:Loyalty_Program   1  746.   746.       2.25  0.172
4 Residuals                            8 2648.   331.      NA    NA
```

Evaluating the ANOVA Model

- Here we can see that:
  - For the interaction effect, we have significant evidence in favor of the **null hypothesis** ($F(1, 8) = 2.25$, $p = 0.172$)
  - For the B, Loyalty Program effect, we have significant evidence in favor of the **alternative hypothesis** ($F(1, 8) = 10.9$, $p = 0.0109$)
  - For the A, Push Notifications effect, we also have significant evidence in favor of the **alternative hypothesis** ($F(1, 8) = 23.6$, $p = 0.0012$)

- Since the interaction effect is not significant and since we only have two levels for each of our main effects, a post-hoc test isn't necessary as we know which two levels may be significantly different from each other.

Quantifying Effect Sizes

- The conclusion of our experiment is that we have evidence to suggest that enabling push notifications and using the enhanced loyalty program may lead to increased monthly spend among customers at Buster's Wings.

- But I want to return to our conversation about quantifying effects as this can be very important when our sample size is small or very large.

Quantifying Effect Sizes

- Earlier in the lecture notes, we saw how to calculate the effect of A, B, and AB.
  - We also learned how to generally interpret the sign.
- However, the numeric value of the effects also has meaning (remember how we interpret the correlation coefficient...it's both the sign *and* the magnitude).
- Just like the interpretation of the correlation coefficient, we need to consider the magnitude of the effect in addition to its sign.

Quantifying Effect Sizes

- Recall, the effect of A was -50.98; the effect of B was -34.66, and the effect of AB was -15.77.
- Since our original units were dollars, the units of these effects is also dollars.
- So in an absolute sense, we may find \$51 and \$35 to be large effects whereas \$15 may be a smaller effect.
  - We saw this statistically: the larger effects were statistically significant while the smaller effect was not.

Quantifying Effect Sizes

- However, \$15 is not zero. We saw in our interaction plot lines that were not parallel.
- In other words, we observed an effect, but the effect was not found to be meaningfully different than a 0 effect in a statistical sense.
- So how do we make sense of this result? We observed a non-zero effect but our statistical test is essentially calling it zero?

Quantifying Effect Sizes

- Consider an independent means t-test statistic:

$$t_{\text{Stat}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s^2_{\text{Pooled}} \left(\frac{1}{r_1} + \frac{1}{r_2} \right)}}$$

- where:

$$s^2_{\text{Pooled}} = \frac{(r_1 - 1)s_1^2 + (r_2 - 1)s_2^2}{r_1 + r_2 - 2}$$

Quantifying Effect Sizes

- If we call $N = r_1 + r_2$, then with a little rearranging, we can see:

$$t_{\text{Stat}} = \frac{(\bar{y}_1 - \bar{y}_2)\sqrt{r_1 r_2 (N-2)}}{\sqrt{N[(r_1-1)s_1^2 + (r_2-1)s_2^2]}}$$

- As $N \to \infty$, this suggests for $\bar{y}_1 - \bar{y}_2 \neq 0$ (a non-zero effect), there exists some value of N that will push our test statistic into the critical region of the null distribution.

Quantifying Effect Sizes

- In other words, if we have a negligible difference between our group means, we may still have $p < 0.05$ for a large enough N.
- The converse may also be true. We may have a moderate and contextually meaningful difference between our group means, but we may observe $p > 0.05$ for a relatively small N, like we see in our case.
- This is why considering effect sizes along with the statistical test is important.

Quantifying Effect Sizes

- Here, we've learned how to calculate unstandardized effect sizes, meaning that they still exist in the units of the outcome variable.

- Words like "big" and "small" are tied to the scale of the units, which isn't desirable.
  - This is why we use the correlation coefficient instead of covariance...the former is *standardized* whereas the latter is not.

- So how do we standardize the effects we've already calculated?

Quantifying Effect Sizes: Partial η^2

- A common effect size for factorial ANOVA designs (2^k or otherwise) is called *partial* η^2 (pronounced "eta"), usually denoted η_p^2.

- For a given effect, the calculation of η_p^2 is:

$$\eta_p^2 = \frac{SS_{\text{Effect}}}{SS_{\text{Effect}} + SSE}$$

- We interpret its value as the proportion of variance in the response explained by this specific factor after accounting for all the other factors in the model.

Quantifying Effect Sizes: Partial η^2

- In our 2^2 case, for factor A, we can show that:

$$SSA = A^2 r$$

$$\implies SSA = (-50.98)^2 \times 3 = 7796.881$$

Quantifying Effect Sizes: Partial η^2

- This suggests that:

$$\eta_p^2 = \frac{7796.881}{7796.881 + 2648.0158} = 0.7465$$

- This means that 74.65% of the variance in the response can be explained by the Push Notifications factor, controlling for the other factor in the model.

Quantifying Effect Sizes: Partial η^2

- Fortunately for us, we don't have to do this by hand. We can use the `partial_eta_squared` function as part of the `rstatix` package to help us out:

```
wings_mod |>
  partial_eta_squared()
```

```
                                  Push_Notifications                                  Loya
                                           0.7465023
Push_Notifications:Loyalty_Program
                                           0.2198167
```

Quantifying Effect Sizes: Partial η^2

- In general, values of η_p^2 around 0.01 indicate a "small" effect,
- Values around 0.06 indicate a "medium" effect, and
- Values around 0.14 indicate a "large" effect

Quanitfying Effect Sizes: Partial η^2

- For us, all of our effects are considered large with our main effect being considerably larger than our interaction effect.
- So here, we can see even though our interaction effect is generally considered large, our small sample size ($N = 12$) prevents us from concluding that this large effect is statistically different from 0.
- This shows that a small sample size can be just as impactful on inference as a large sample!