A/B Testing and Beyond

Designed Experiments for Data Scientists

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Week 5

Wednesday October 4th, 2017

Outline

- Recap
- Primer on linear regression
- Experiments with Multiple Conditions
  - Comparing means
  - Comparing proportions
  - The multiple comparison problem
- Experiments with Multiple Factors
  - Factorial vs. One-factor-at-a-time
  - Designing and analyzing factorial experiments

RECAP

Recap

- Experiments with Two Conditions
  - Evaluating Assumptions
    - Welch's t-test
    - Randomization tests
    - χ^2-tests
  - A discussion of "peeking"

LINEAR REGRESSION – A PRIMER

Linear Regression

- This is a form of statistical modeling that is appropriate when interest lies in relating a response variable (Y) to one or more explanatory variables ($x_1, x_2, \ldots, x_p$).

- The idea is that Y is influenced in some manner by $\{x_1, x_2, \ldots, x_p\}$ according to an unknown function:

$$Y = f(x_1, x_2, \ldots, x_p)$$

Linear Regression

- The goal of statistical modeling in general (and linear regression in particular) is to approximate the function $f(\cdot)$

- The linear regression model relates Y to $\{x_1, x_2, \ldots, x_p\}$ via

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

where

  - Y is the response variable
  - The x_j's are explanatory variables we treat as fixed
  - The β's are unknown parameters quantifying the influence of a particular x_j on Y

Linear Regression

- And ϵ is the random error term that accounts for the fact that
  $f(x_1, x_2, \ldots, x_p) \neq \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  and we assume $\epsilon \sim N(0, \sigma^2)$

- This distributional assumption has several consequences. In particular, it implies

$$Y \sim N\left(\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2\right)$$

which means that we expect, for specific values of the x's, the response to be equal to

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Linear Regression

Based on this distributional result

$$E\left[Y|x_1 = x_2 = \cdots = x_p = 0\right] = \beta_0$$

And so β_0 is interpreted as the intercept of the model:

- The expected response when all of the explanatory variables are equal to zero.

Linear Regression

Also notice that

$$
\begin{aligned}
&E\left[Y|x_j = x+1\right] - E\left[Y|x_j = x\right] \\
&= \left(\beta_0 + \beta_1 x_1 + \cdots + \beta_j(x+1) + \cdots + \beta_p x_p\right) \\
&- \left(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \cdots + \beta_p x_p\right) \\
&= \left(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \beta_j + \cdots + \beta_p x_p\right) \\
&- \left(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \cdots + \beta_p x_p\right) \\
&= \beta_j
\end{aligned}
$$

And so β_j is interpreted as the expected change in response associated with a unit increase in x_j, while holding all other explanatory variables fixed

Linear Regression

To actually use the linear regression model we must estimate the β's.

This is typically done with least squares estimation where the goal is to find values of $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimize the model's error, ϵ.

For observed data $\left(y_i, x_{i1}, x_{i2}, \ldots, x_{ip}\right), i = 1,2, \ldots, n$ we wish to minimize

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)\Big)^2$$

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Linear Regression

The linear regression model can be expressed in vector-matrix notation as follows

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear Regression

Using this formulation it can be shown that the least squares estimate of $\boldsymbol{\beta}$ and hence of the individual β's is given by

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

$$= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

Linear Regression

With the regression coefficients estimated we define the fitted values to be

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

which are interpreted as an estimate of the expected response for specific values of the x's

Next we define the residuals to be

$$e_i = y_i - \hat{\mu}_i$$

which represent the difference between observed values of the response and what the model predicts the response to be.

Linear Regression

It can be shown that the least squares estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2}{n-p-1}$$

- This estimate is sometimes referred to as the mean squared error (MSE) of the model
- This is because $\hat{\sigma}$ quantifies the typical distance (error) between an observed response value and the value predicted by the model

Linear Regression

Having estimated $\beta_0, \beta_1, \dots, \beta_p$ and σ^2 the fitted linear regression model can be used for inference and prediction

Of particular importance are hypothesis tests of the form

$$H_0\colon \beta_j = 0 \text{ vs. } H_A\colon \beta_j \neq 0$$

for some $j = 1,2, \dots, p$

And confidence and prediction intervals for predicted values of Y

EXPERIMENTS WITH MULTIPLE CONDITIONS

Comparing Multiple Conditions

- We now consider the design and analysis of an experiment consisting of multiple experimental conditions i.e., an A/B/n Test
- Like an A/B test, the goal is to decide which condition is optimal with respect to some metric of interest – but now we have several conditions

- Given several options, which one is best?

Comparing Multiple Conditions

Designing a multi-condition test:

- Choose your response variable (y)
- Choose a metric θ that summarizes the response
- Choose a design factor and m levels to experiment with
- Choose $n_1, n_2, \ldots, n_m$ – the number of units to assign to each condition

Comparing Multiple Conditions

Data Collection:

- Randomly assign n_j units to condition $j = 1,2, \dots, m$
- Measure the response (y) on each unit and summarize the measurements with the metric of interest θ in each of the conditions and hence obtain

$$\hat{\theta}_1,\ \hat{\theta}_2, \dots,\ \hat{\theta}_m$$

Goal:

- Identify the optimal condition

Comparing Multiple Conditions

In order to identify the optimal condition, we simply need to do a series of pairwise comparisons using two-sample tests

- t-tests, Z-tests, and χ^2-tests may be used for this purpose

However, while identifying the optimal condition is the ultimate goal, it is prudent to first decide whether a difference exists, at all, between the conditions

Comparing Multiple Conditions

To answer this question formally, we may test a hypothesis of the form

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_m \text{ vs. } H_A: \theta_j \neq \theta_k$$

for some $j \neq k$

Next we discuss how to test this hypothesis in the cases that the metric of interest is either a

- Mean, or a
- Proportion (rate)

Comparing Multiple Means

The Linear Regression *F*-test

Here interest lies in testing the hypothesis

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_m \text{ vs. } H_A\colon \mu_j \neq \mu_k$$

for some $j \neq k$.

This may be done with the *F*-test associated with an appropriately defined linear regression model.

Specifically, we adopt the following model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{m-1} x_{i,m-1} + \epsilon_i$$

Comparing Multiple Means

The Linear Regression *F*-test

In this model

- $Y_i \sim N(\mu_j, \sigma^2)$ represents the response observation for unit $i = 1,2, \ldots, N = \sum_{j=1}^{m} n_j$.
- Each x_{ij} is a dummy (indicator) variable taking on the value 1 if unit i is in condition j, and 0 otherwise
- $\epsilon_i \sim N(0, \sigma^2)$ represents the random error term for unit i
- The β's are unknown regression parameters

Comparing Multiple Means

The Linear Regression *F*-test

The parameter β_0 is interpreted as the expected response value when $x_1 = x_2 = \cdots = x_m = 0$

In other words, β_0 is the expected response value in condition m

We can also show that $\beta_0 + \beta_j$ is the expected response value in condition $j = 1,2, \dots, m-1$

Comparing Multiple Means

The Linear Regression *F*-test

As such

$$\begin{aligned}
\mu_1 &= \beta_0 + \beta_1 \\
\mu_2 &= \beta_0 + \beta_2 \\
\mu_3 &= \beta_0 + \beta_3 \\
&\vdots \\
\mu_{m-1} &= \beta_0 + \beta_{m-1} \\
\mu_m &= \beta_0
\end{aligned}$$

and

$$\mu_1 = \mu_2 = \cdots = \mu_m$$

if and only if

$$\beta_1 = \beta_2 = \cdots = \beta_m = 0$$

Comparing Multiple Means

The Linear Regression *F*-test

So testing

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m \text{ vs. } H_A: \mu_j \neq \mu_k$$

for some $j \neq k$

is equivalent to testing

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0 \text{ vs. } H_A: \beta_j \neq 0$$

for some $j = 1,2, \ldots, m$

This latter test corresponds to the *F*-test for overall significance in a linear regression model

Comparing Multiple Means

Example: Candy Crush

Candy Crush is experimenting with three different versions of in-game "boosters":

- The lollipop hammer
- The jelly fish
- The color bomb

Users are randomized to one of these three conditions ($n_1 = 121$, $n_2 = 135$, $n_3 = 117$) and they receive (for free) 5 boosters corresponding to their condition.

Let μ_j represent the average length of game play in condition $j = 1,2,3$.

Comparing Multiple Means

Example: Candy Crush

While interest ultimately lies in finding the booster condition that maximizes user engagement, (i.e., has the largest μ_j) we will first decide whether any difference at all exists between the conditions:

$$H_0: \mu_1 = \mu_2 = \mu_3 \text{ vs. } H_A: \mu_j \neq \mu_k$$

for some $j \neq k$

To do so, we fit an “appropriately defined linear regression model”. The results are shown on the next slide.

Comparing Multiple Means

Example: Candy Crush

```
Call:
lm(formula = time ~ factor(booster), data = candy)
Residuals:
    Min       1Q    Median     3Q      Max
-2.84231 -0.69476 0.02617 0.65326 2.76681
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.01281   0.08664   57.859 <2e-16 ***
factor(booster)2 1.17528   0.11931   9.851  <2e-16 ***
factor(booster)3 4.88279   0.12357   39.515 <2e-16 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 0.953 on 370 degrees of freedom
Multiple R-squared: 0.8216,Adjusted R-squared: 0.8206
F-statistic: 851.9 on 2 and 370 DF, p-value: < 2.2e-16
```

Comparing Multiple Means

Example: Candy Crush

From this output we see that $\hat{\beta}_0 = 5.0128$, $\hat{\beta}_1 = 1.1753$ and $\hat{\beta}_2 = 4.8828$

This means that the average length of game play in each condition is estimated to be

- $\hat{\mu}_1 = 5.0128$ minutes in the lollipop hammer condition
- $\hat{\mu}_2 = 6.1881$ minutes in the jelly fish condition
- $\hat{\mu}_3 = 9{:}8956$ minutes in the color bomb condition

Comparing Multiple Means

Example: Candy Crush

The p-value associated with the *F*-test for overall significance in a linear regression model is less than 2.2×10^{-16} which provides very strong evidence against H_0

Thus we conclude that the average length of game play is not the same for each of the boosters.

To determine which booster is optimal – the one that maximizes game play duration – we must use a series of pairwise t-tests

Comparing Multiple Proportions

χ^2-test of Independence

Here interest lies in testing the hypothesis

$$H_0: \pi_1 = \pi_2 = \cdots = \pi_m \text{ vs. } H_A: \pi_j \neq \pi_k$$

for some $j \neq k$.

This may be done with the same χ^2-test of independence that we discussed in the $m = 2$ case

Yes, it generalizes!

Comparing Multiple Proportions

χ^2-test of Independence

In the case of m conditions we have a $2\times m$ contingency table:

		Condition				
		1	2	$\cdots$	m	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	$\cdots$	$O_{1,m}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	$\cdots$	$O_{0,m}$	O_0
		n_1	n_2	$\cdots$	n_m	$\sum_{j=1}^{m} n_j$

- $O_{1,j}$ and $O_{0,j}$ respectively represent the observed number of conversions and non-conversions in condition $j = 1,2, \dots , m$
- O_1 and O_0 represent the overall number of conversions and non-conversions

Comparing Multiple Proportions

χ^2-test of Independence

If $\pi_1 = \pi_2 = \cdots = \pi_m = \pi$ then we would expect the conversion rate in each condition to be the same

Pooled estimates of $\hat{\pi}$ and $1 - \hat{\pi}$ are given by

$$\hat{\pi} = \frac{O_1}{\sum_{j=1}^{m} n_j} \text{ and } 1 - \hat{\pi} = \frac{O_0}{\sum_{j=1}^{m} n_j}$$

With these we can calculate the expected number of observations in each cell of the contingency table:

$$E_{1,j} = n_j\hat{\pi} \text{ and } E_{0,j} = n_j(1 - \hat{\pi})$$

for$j = 1,2, \ldots , m$

Comparing Multiple Proportions

χ^2-test of Independence

The expected frequencies can also be summarized in a contingency table:

		Condition				
		1	2	$\cdots$	m	
Conversion	Yes	$E_{1,1}$	$E_{1,2}$	$\cdots$	$E_{1,m}$	O_1
	No	$E_{0,1}$	$E_{0,2}$	$\cdots$	$E_{0,m}$	O_0
		n_1	n_2	$\cdots$	n_m	$\sum_{j=1}^{m} n_j$

Note that the margin totals do not change.

As in the 2×2 case, the χ^2-test formally compares what was observed and what is expected under the null hypothesis

Comparing Multiple Proportions

χ^2-test of Independence

The test statistic that compares the observed count in each cell to the corresponding expected count, is defined as

$$T = \sum_{l=0}^{1} \sum_{j=1}^{m} \frac{\left(O_{l,j} - E_{l,j}\right)^2}{E_{l,j}}$$

Assuming H_0 is true, T approximately follows a $\chi^2_{(m-1)}$ distribution

- As a rule of thumb, this approximation may be very poor unless the observed and expected cell frequencies are all greater than 5

Comparing Multiple Proportions

Example: Nike SB

- Suppose that Nike is running an ad campaign for Nike SB, their skateboarding division
- The ad campaign involves $m = 5$ different video ads being shown in Facebook newsfeeds
- In these five video conditions there are $n_1 = 5014$, $n_2 = 4971$, $n_3 = 5030$, $n_4 = 5007$, and $n_5 = 4980$ users, respectively
- The videos in these conditions are viewed 160, 95, 141, 293, and 197 times yielding watch rates:

$$\hat{\pi}_1 = 0.03, \hat{\pi}_2 = 0.02, \hat{\pi}_3 = 0.03,$$

$$\hat{\pi}_4 = 0.06, \hat{\pi}_5 = 0.04$$

Comparing Multiple Proportions

Example: Nike SB

The observed contingency table is

		Condition					
		1	2	3	4	5	
View	Yes	160	95	141	293	197	886
	No	4854	4876	4889	4714	4783	24116
		5014	4971	5030	5007	4980	25002

And the expected contingency table is

		Condition					
		1	2	3	4	5	
View	Yes	177.68	176.16	178.25	177.43	176.48	886
	No	4836.32	4794.84	4851.75	4829.57	4803.52	24116
		5014	4971	5030	5007	4980	25002

Comparing Multiple Proportions

Example: Nike SB

- The observed value of the test statistic for this test is $t = 129.1761$ and the corresponding p-value is 5.84×10^{-27} and so there is strong evidence again H_0

- As such, we conclude that the likelihood that someone "views" a video is not the same for all of the videos

- To determine which video is optimal – the one with the highest likelihood of viewing – we must use a series of pairwise Z-tests or χ^2-tests

The Multiple Comparison Problem

As we saw in the previous two examples, the null hypothesis of overall equality is often rejected

In these cases a family of follow-up pairwise comparisons are necessary to determine which condition(s) is (are) optimal

Statistically we know how to do this

However, when doing multiple comparisons, it is important to recognize that the overall Type I Error rate associated with this family of tests is inflated

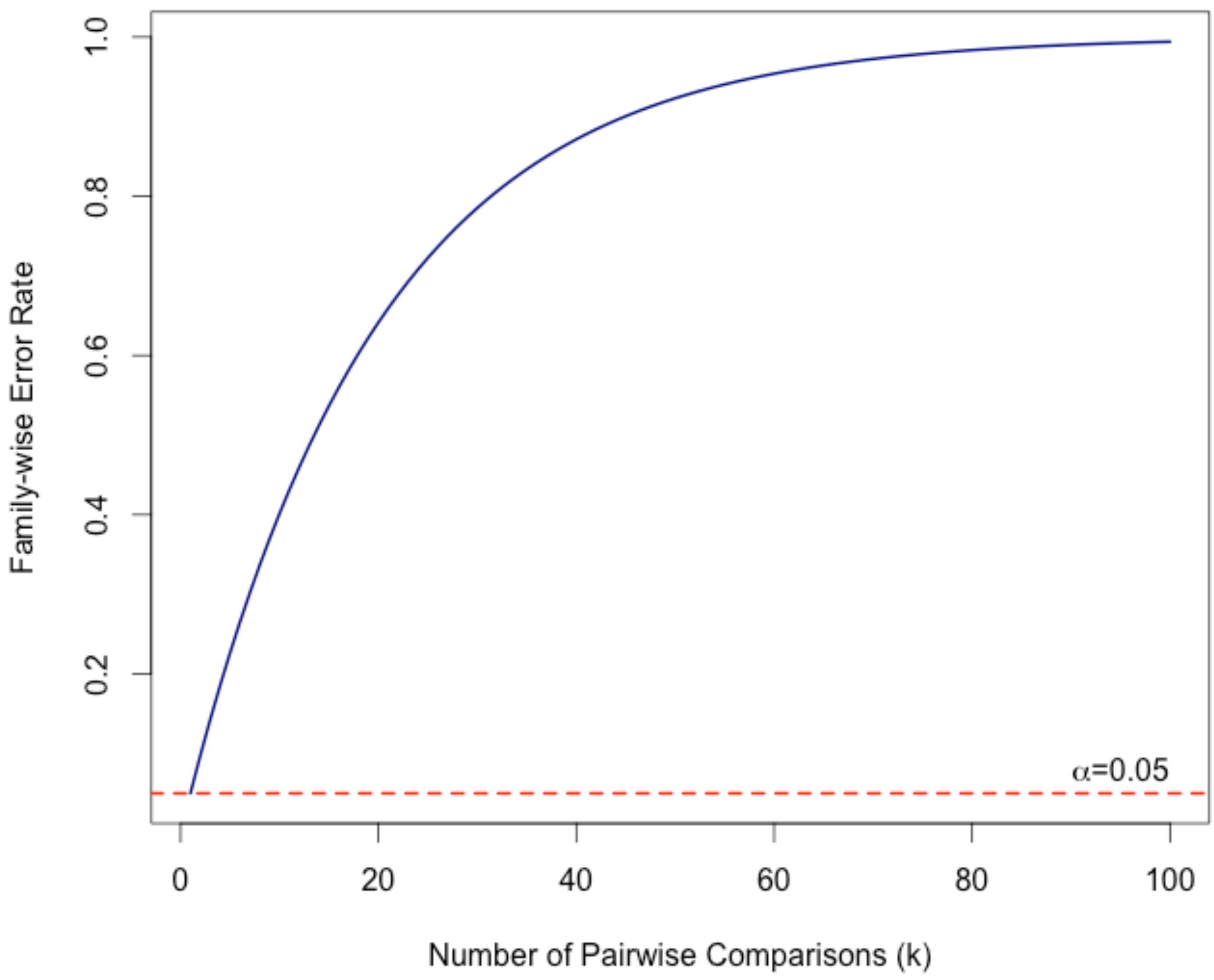
The Multiple Comparison Problem

This problem – where a series of independent hypothesis tests lead to an inflated family-wise error rate – is known as the multiple comparison or multiple testing problem.

It can be shown that for a family of k hypothesis tests, each with significance level α, the family-wise error rate is

$$1 - (1 - \alpha)^k$$

The Multiple Comparison Problem



The Multiple Comparison Problem

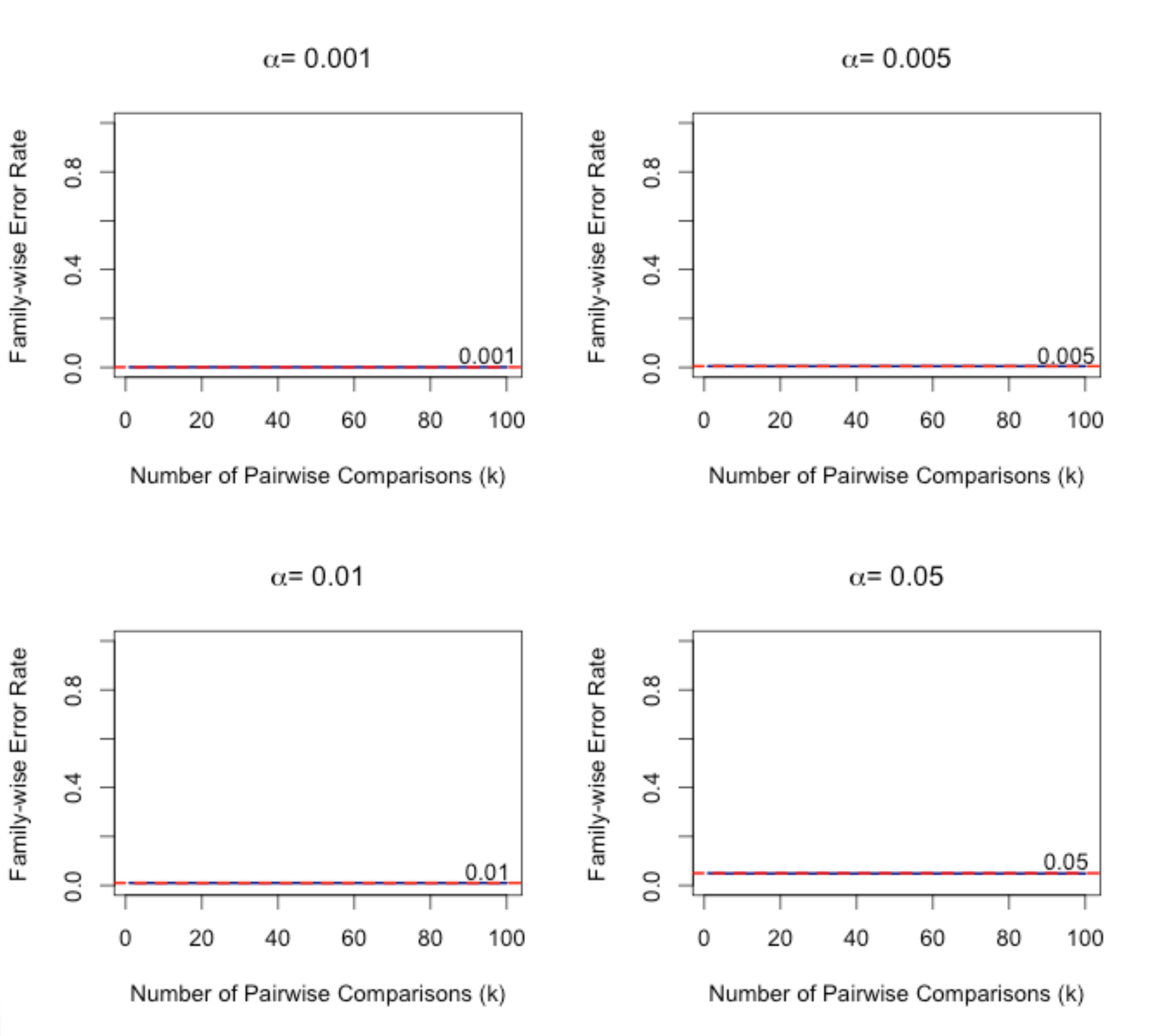
We combat this problem with the Bonferroni correction

- With this correction we test each of the k hypothesis tests at a significance level α/k, if maintaining an error rate of α is of interest
- Doing so yields a family-wise error rate of

$$1 - \left(1 - \frac{\alpha}{k}\right)^k$$

which, for typical values of α is approximately equal to α

The Multiple Comparison Problem



The Multiple Comparison Problem

So what does this mean for sample size calculations and power analyses?

The sample size formulas we derived previously did not account for this multiple comparison problem

In order to do so, when performing a power analysis, use α/k and not α as the significance level in the sample size calculations

EXPERIMENTS WITH MULTIPLE FACTORS

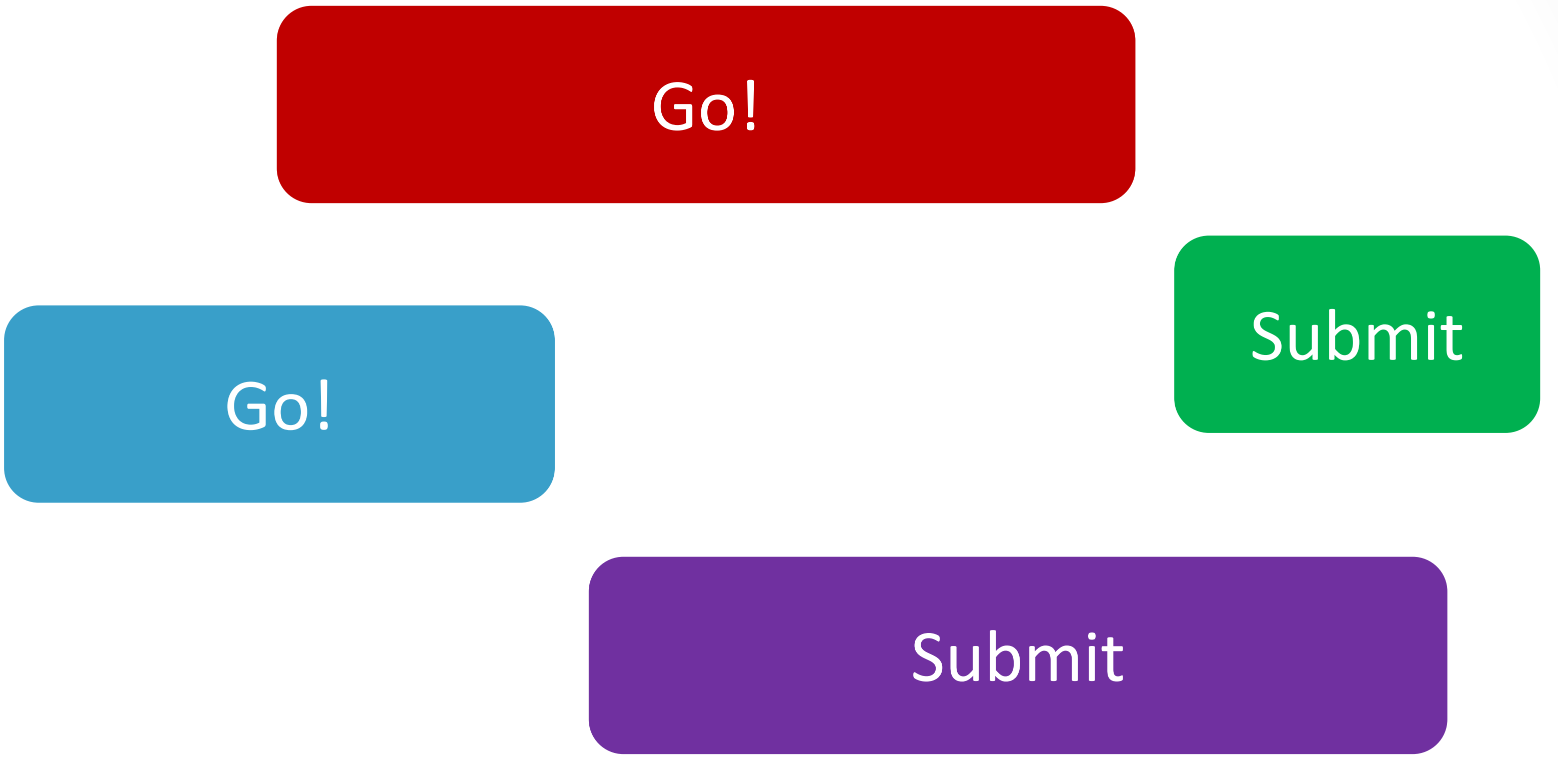
Multivariate Experiments

- So far we have considered experiments with just one design factor
- However, there might be several factors that are expected to impact the response
- We now turn our attention to the so-called "multivariate experiment" in which we manipulate more than one design factor

Multivariate Experiments

- Previously we considered experimenting with the color of a button to determine which color maximized the likelihood that the button is clicked
- But what about the size of the button, the button’s location, or the button’s message?
- All of these things might influence whether the button is clicked
- The goal, then, is to find the combination of factor levels that optimize the response

Multivariate Experiments



- How do we use an experiment to find the optimal combinations?

Multivariate Experiments

The one-factor-at-a-time approach is a simple method for investigating several factors

This approach can be carried out by following these steps:

- Pick a factor to experiment with
- Run an experiment and find that factor's optimal level
- Pick a second factor to experiment with
- Run an experiment with the first factor fixed at its optimal level and then find the optimal level of the second factor

Multivariate Experiments

- Pick a third factor to experiment with
- Run an experiment with the first two factors held fixed at their optimal levels and then find the optimal level of the third factor
- Repeat in this manner until all factors of interest have been investigated

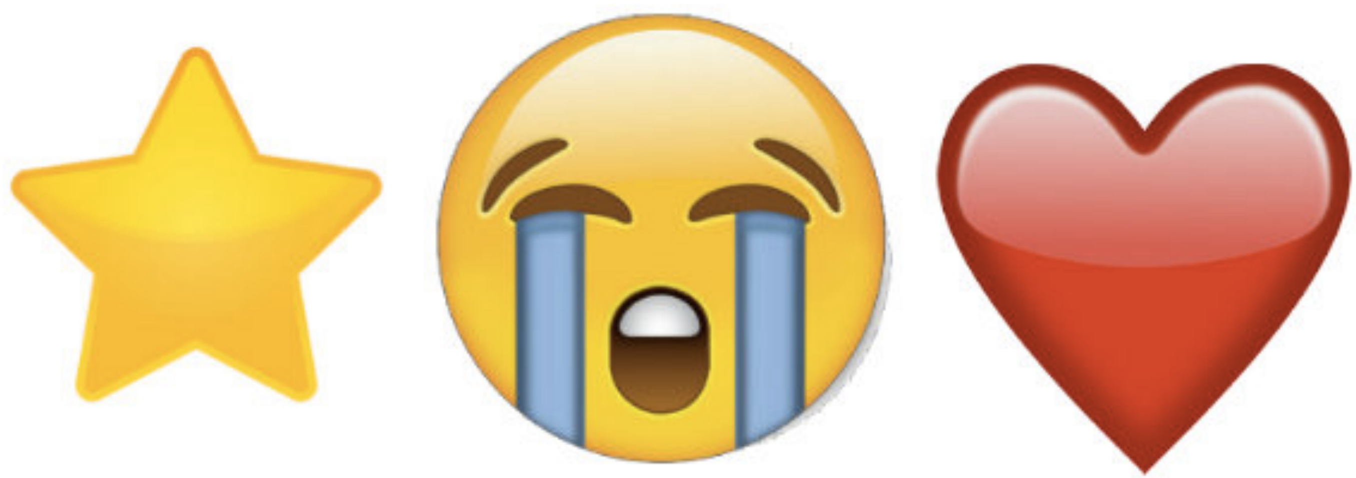
While this approach is simple, it has one major drawback:

- **There may be an optimal combination you did not try**

Multivariate Experiments

Example: Twitter experiment

Twitter changed their star 'favourites' to heart 'likes' and the internet is pissed



Multivariate Experiments

Example: Twitter experiment

The experiment that was run involves two factors each with two levels:

- Icon Shape: 
- Icon Color: 
- Consider investigating these using the the one-factor-at-a-time approach

Multivariate Experiments

Example: Twitter experiment

Test 1:

versus

- **Winner: Heart**

Multivariate Experiments

Example: Twitter experiment

Test 2:

versus

- **Winner: Red Heart**

Multivariate Experiments

Example: Twitter experiment

But what about

- The one-factor-at-a-time approach missed this combination
- What if it's the best?

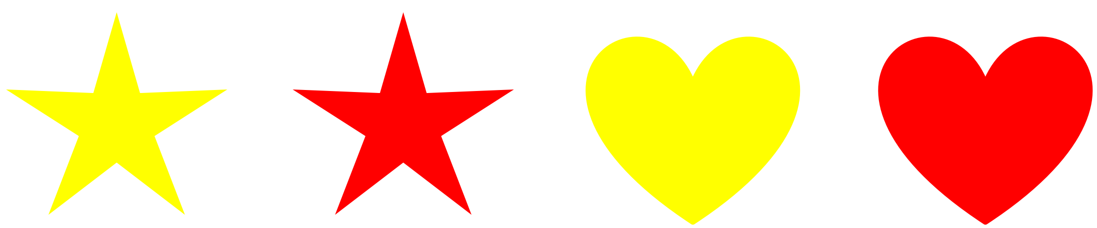
Multivariate Experiments

The Factorial Approach

A factorial approach to multivariate experiments considers **every** combination of factor levels

So it doesn't miss any potentially optimal combinations

In the Twitter example there are 2x2=4 possible combinations:



Multivariate Experiments

The Factorial Approach

A factorial experiment would have investigated all of these combinations – there is no loss of information

In this case, the number of conditions is exactly the same as in the one-factor-at-a-time approach!

But as the number of factors and levels increase, factorial experiments will always have more conditions than a the one-factor-at-a-time approach

Multivariate Experiments

The Factorial Approach

This is the **only drawback** to factorial experiments – they get big, quickly!

However they are still the most efficient way to fully investigate multiple factors

A factorial experiment allows us to investigate

- main effects: the change in response produced by a change in a particular factor
- interaction effects: the difference between the main effect of one factor at different levels of another

Multivariate Experiments

Designing a Factorial Experiment

The design is conceptually simple:

- Pick your design factors
- Pick their levels
- Your experimental conditions are all of the different combinations of these factors' levels

If you have k factors with $m_1, m_2, \dots, m_k$ levels, respectively, the corresponding factorial experiment will have

$$M = m_1 m_2 \cdots m_k$$

experimental conditions

Multivariate Experiments

Designing a Factorial Experiment

However, practically, the design is not simple.

- As the number of factors and levels increase M gets very large
- We need to be careful choosing our factors and levels so as not design an unmanageably large experiment
- Keep it simple!

Multivariate Experiments

Designing a Factorial Experiment

Once the conditions are established experimental units must be randomized to each of them

Like the single-factor multi-level experiments we've discussed previously, factorial experiments consist of multiple conditions

Thus the optimal condition can be found using a series of pairwise comparisons as we have seen

Sample size calculations should be based on two-sample tests that account for the multiple comparison problem

Multivariate Experiments

Designing a Factorial Experiment

Once units have been assigned to each condition, the response variable is measured on all of them

Using the collected data we

(1) Identify which factors are influential, and

(2) Identify which combination of factors is optimal

To do (1) we will apply regression techniques

To do (2) we will use two sample t-, Z- or χ^2-tests

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

We discuss these concepts in the context of the following example:

Suppose, again, Instagram is experimenting with ads to understand their influence on user engagement.

Again we assume the response variable (Y) is session duration (measured in minutes)

But now we assume we have two design factors

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

Factor 1: Ad Frequency

- None (coded as 0)
- 7:1 (coded as 1)
- 4:1 (coded as 2)
- 1:1 (coded as 3)

Factor 2: Ad Type

- Photo (coded as 1)
- Video (coded as 2)

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

Factor 1: Ad Frequency

- None (coded as 0)
- 7:1 (coded as 1)
- 4:1 (coded as 2)
- 1:1 (coded as 3)

This leads to 4x2 = 8 unique conditions

Factor 2: Ad Type

- Photo (coded as 1)
- Video (coded as 2)

Assume we randomize n=1000 units to each and measure Y

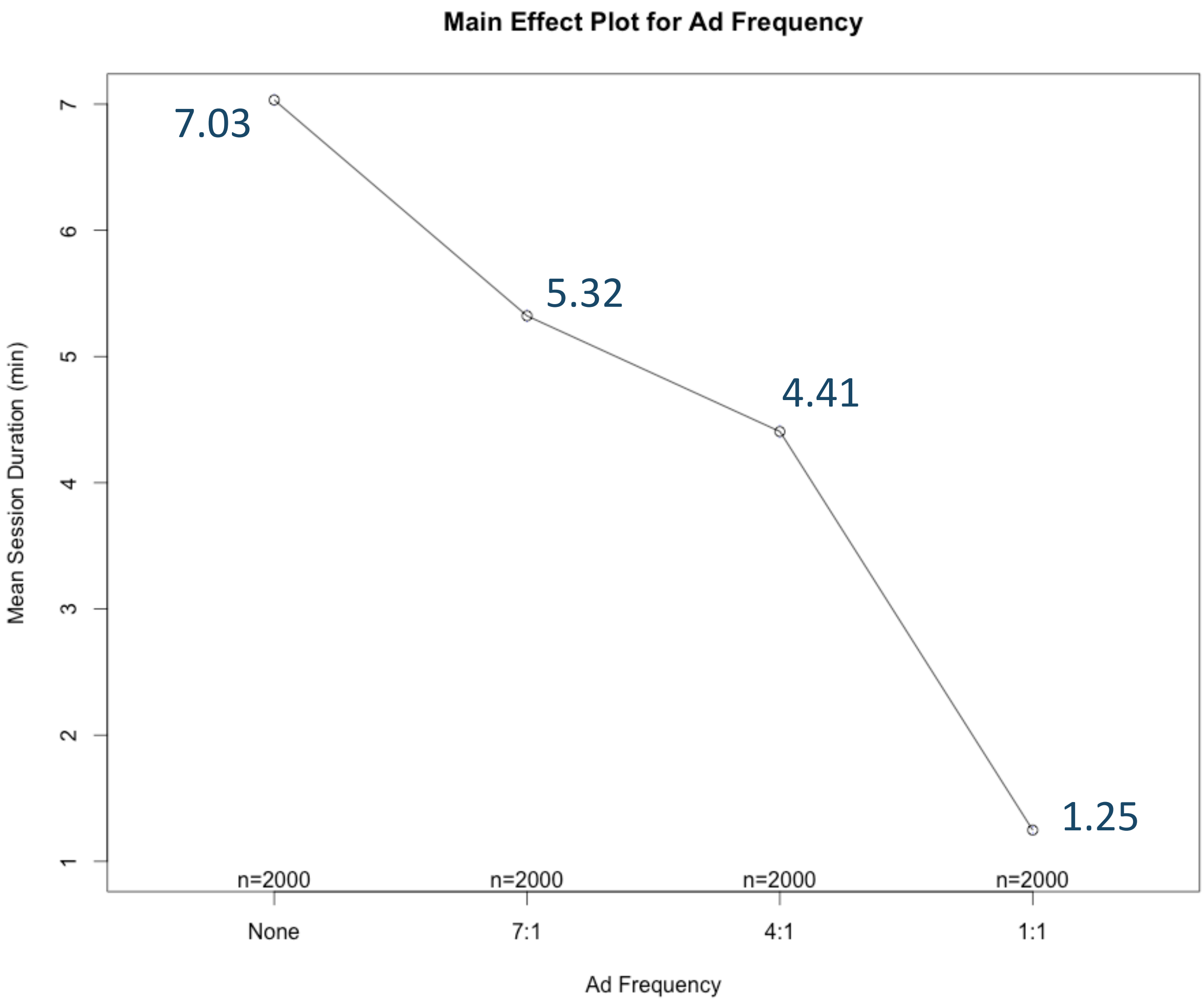
Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

Frequency: None Type: Photo	Frequency: None Type: Video
Frequency: 7:1 Type: Photo	Frequency: 7:1 Type: Video
Frequency: 4:1 Type: Photo	Frequency: 4:1 Type: Video
Frequency: 1:1 Type: Photo	Frequency: 1:1 Type: Video

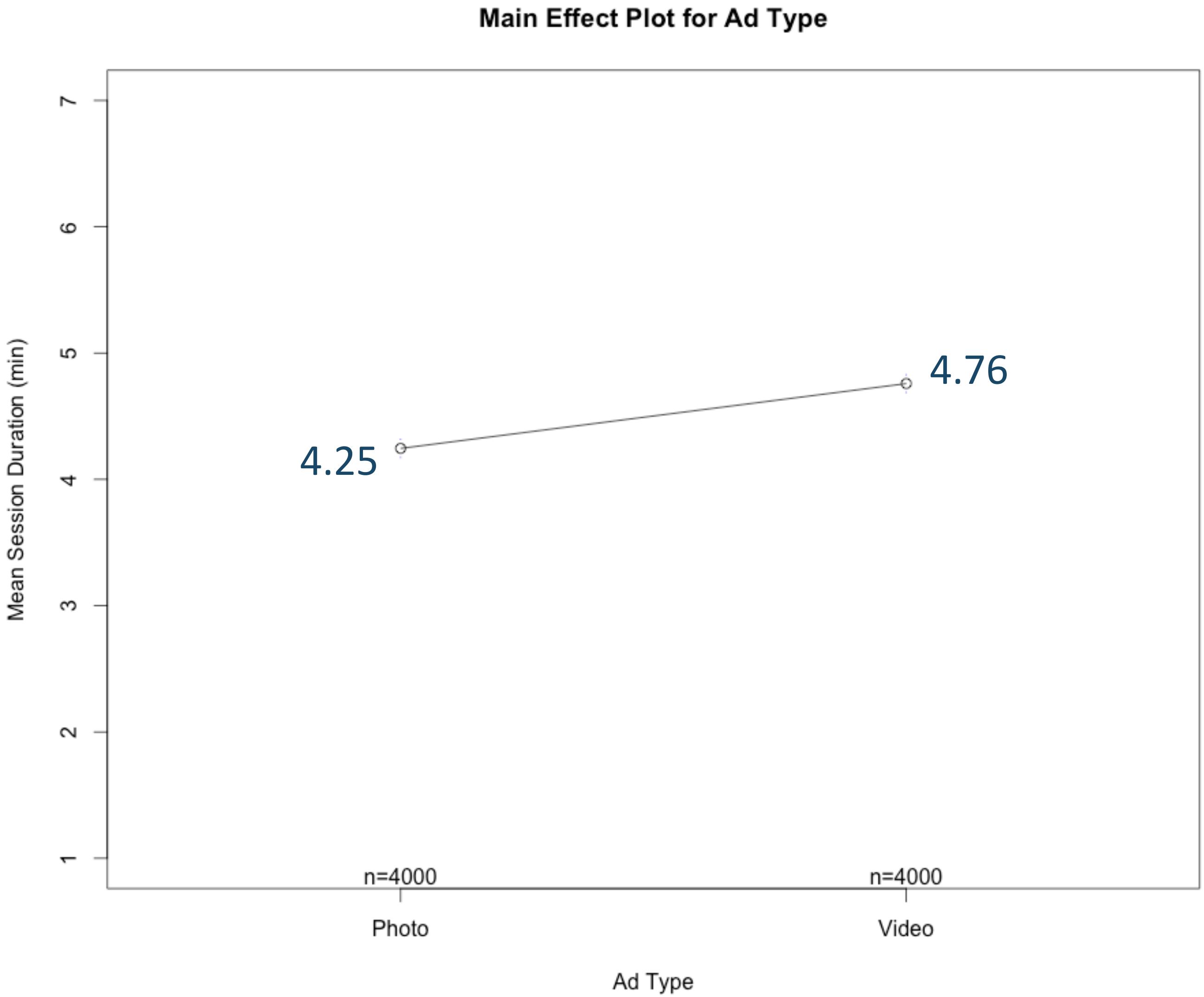
Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y



Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y



Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

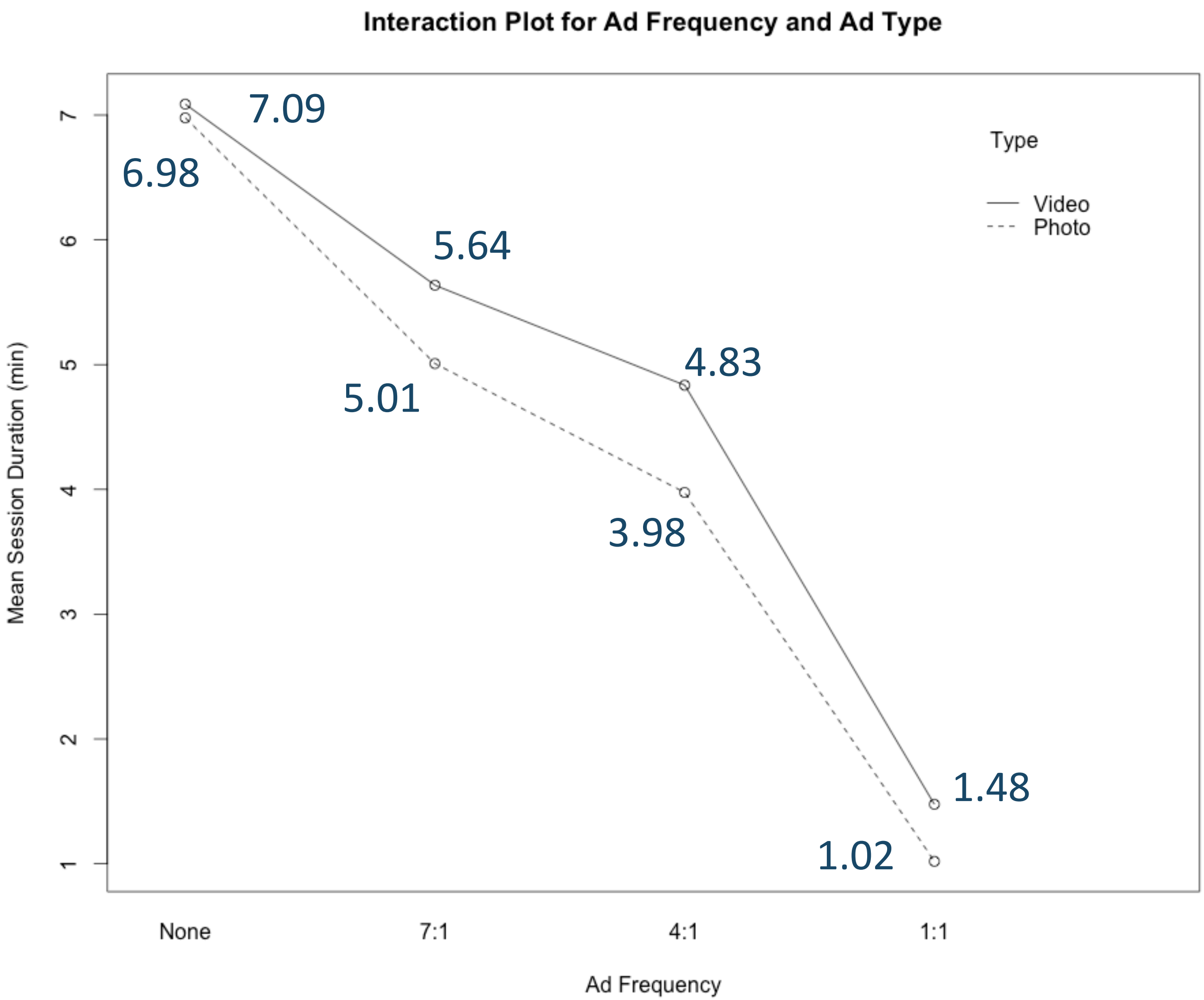
3

The main effect plots tell us:

- Session duration decreases as ad frequency increases
- Session duration is slightly longer for video ads vs. photo ads
- The influence of ad frequency is larger than the influence of ad type

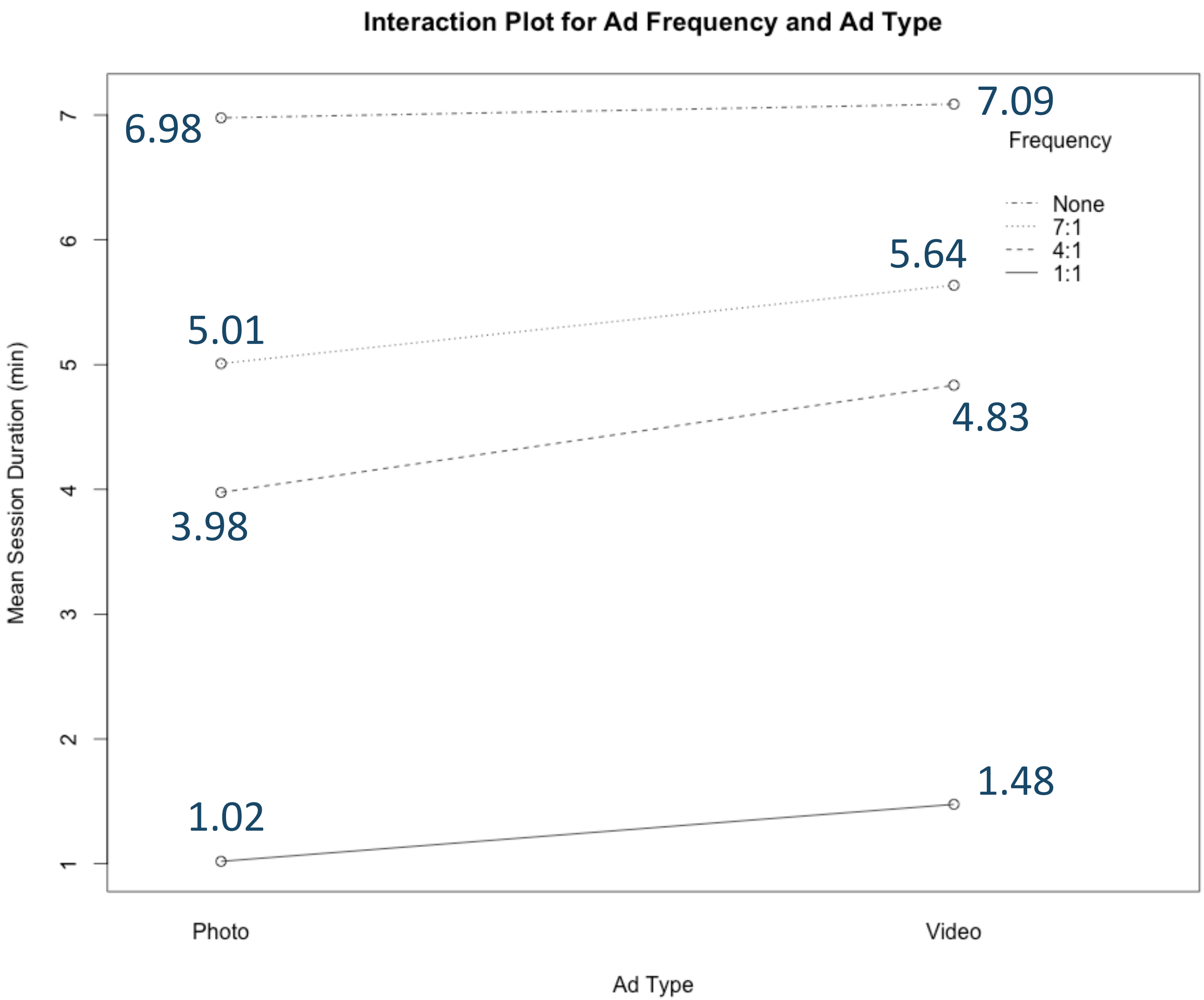
Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y



Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y



Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

The interaction effect plots tell us:

- The effect of ad frequency is not quite the same for both ad types
- The effect of ad type is not quite the same for all ad frequencies
- Thus an interaction is present

To formally decide whether the main and interaction effects are significant, we use linear regression

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

Linear regression models used for this purpose should contain

- Indicator variables for each factor; the number of indicators for a particular factor is equal to the number of levels of that factor, minus 1.
  - This allows us to evaluate main effects

- k-way products of the indicator variables for the k different factors.
  - This allows us to evaluate interaction effects

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

The linear regression model appropriate for the Instagram factorial example is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \\ + \beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4} + \epsilon_i$$

where

- $x_{i1} = 1$ if unit i is in the 7:1 condition
- $x_{i2} = 1$ if unit i is in the 4:1 condition
- $x_{i3} = 1$ if unit i is in the 1:1 condition
- $x_{i4} = 1$ if unit i is in the video condition

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

Main effects become irrelevant in the context of interaction, and so it is common practice to first decide whether the interaction effect is significant

Note that $\beta_5 = \beta_6 = \beta_7 = 0$ removes the interaction terms from the model and so a test of

$$H_0: \beta_5 = \beta_6 = \beta_7 = 0 \text{ vs. } H_A: \beta_j \neq 0$$

formally tests whether the interaction effect is significant for at least one of $j = 5,6,7$

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

If the interaction effect is significant (i.e., we do not reject H_0) we must be careful to only draw conclusions regarding the effect of one factor in the context of the levels of the other factor

However, if the interaction effect is not significant (i.e., we do not reject $\beta_5 = \beta_6 = \beta_7 = 0$) we may use the **reduced** main effects model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

which can be used to evaluate the significance of main effects

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

The expected response, based on this model, in each of the "photo" conditions is shown below.

Frequency	Expected Response
None	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 0] = \beta_0$
7:1	$E[Y_i \mid x_{i1} = 1, x_{i4} = 0] = \beta_0 + \beta_1$
4:1	$E[Y_i \mid x_{i2} = 1, x_{i4} = 0] = \beta_0 + \beta_2$
1:1	$E[Y_i \mid x_{i3} = 1, x_{i4} = 0] = \beta_0 + \beta_3$

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

The expected response, based on this model, in each of the “video” conditions is shown below.

Frequency	Expected Response
None	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 1] = \beta_0 + \beta_4$
7:1	$E[Y_i \mid x_{i1} = 1, x_{i4} = 1] = \beta_0 + \beta_1 + \beta_4$
4:1	$E[Y_i \mid x_{i2} = 1, x_{i4} = 1] = \beta_0 + \beta_2 + \beta_4$
1:1	$E[Y_i \mid x_{i3} = 1, x_{i4} = 1] = \beta_0 + \beta_3 + \beta_4$

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

- Notice that the expectations in each row are identical if $\beta_1 = \beta_2 = \beta_3 = 0$
- Thus, ad frequency does not significantly influence the response if $\beta_1 = \beta_2 = \beta_3 = 0$
- We formally test whether the main effect of ad frequency is significant by testing

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_A: \beta_j \neq 0$$

for at least one of $j = 1,2,3$

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

- Notice that the expected response for photo vs. video ads becomes the same if $\beta_4 = 0$
- Thus, ad type does not significantly influence the response if $\beta_4 = 0$
- We formally test whether the main effect of ad type is significant by testing

$$H_0: \beta_4 = 0 \text{ vs. } H_A: \beta_4 \neq 0$$

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

- All of these hypothesis tests correspond to simultaneously setting a subset of the β's equal to zero
- Thus, each of these tests generates a reduced model with fewer terms than the corresponding full model
- In each case we compare the full and reduced models to decide if they seem significantly different – rejecting H_0 if they do
- This is done formally with a partial *F*-test

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

- The partial F-test compares the mean squared errors between the full and reduced models (similar to the F-test for overall significance in a linear regression)

- The test statistics and p-values associated with this test are provided in standard linear regression ANOVA output like `anova()` in R.

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

```
lm(formula = Time ~ Frequency * Type)
Residuals:
   Min      1Q    Median    3Q      Max
-3.7276 -0.5474 -0.0020 0.5499 4.4332
Coefficients:
                      Estimate Std.Error t value  Pr(>|t|)
(Intercept)            6.97785  0.02824  247.104 < 2e-16 ***
Frequency7:1          -1.96929  0.03994  -49.312 < 2e-16 ***
Frequency4:1          -3.00204  0.03994  -75.173 < 2e-16 ***
Frequency1:1          -5.95856  0.03994 -149.206 < 2e-16 ***
TypeVideo              0.10993  0.03994    2.753 0.00592 **
Frequency7:1:TypeVideo 0.51768  0.05648    9.166 < 2e-16 ***
Frequency4:1:TypeVideo 0.74924  0.05648   13.266 < 2e-16 ***
Frequency1:1:TypeVideo 0.34731  0.05648   6.150 8.14e-10 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
Residual standard error: 0.893 on 7992 degrees of freedom
Multiple R-squared: 0.8497,Adjusted R-squared: 0.8496
F-statistic: 6455 on 7 and 7992 DF, p-value: < 2.2e-16
```

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

```
Analysis of Variance Table
Response: Time
                 Df  Sum Sq  Mean Sq    F value      Pr(>F)
Frequency         3   35353  11784.3  14778.187  < 2.2e-16 ***
Type              1     527    527.3    661.318  < 2.2e-16 ***
Frequency:Type    3     149     49.8     62.398  < 2.2e-16 ***
Residuals      7992    6373      0.8
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Multivariate Experiments

Analyzing a Factorial Experiment – Continuous Y

The p-values in the ANOVA table are sufficiently small so we conclude:

- Ad frequency has a significant main effect
- Ad type has a significant main effect
- The interaction between these factors is also significant

This means that both factors should be considered when trying to optimize session duration.

To determine which condition is optimal we can use a series of pairwise t-tests

Take Home Exercises

Using R or Python, formally do the pairwise comparisons to find the optimal condition in each of the three examples presented here. Be sure to account for the multiple comparison problem.

See you next week!