Final Exam
Design of Experiments

Mikaela Hoffman-Stapleton

June 2, 2017

1

There are two main advantages to conducting an experiment (as opposed to an observational study): (1) causal inference is easy, and (2) it's fast and efficient. The first refers to the fact that experimental conditions differ only by a difference in factor levels, so any change observed in the response variable is said to be caused by the changes made in the experiment. This can be extremely useful in identifying which factor levels should be used in production in order to optimize the response. The second refers to the fact that in designing an experiment and purposefully collecting data, you have much more control over everything, which often translates into a faster, more efficient approach to answer the question at hand.

There are also two main disadvantages to conducting an experiment: (1) it can be risky or costly, and (2) possibly unethical. The first refers to the idea that some of the treatments may negatively affect the response variable in a significant way. For example, one version of a website may deter users so much that the website loses money as a result of that part of the experiment. The second refers to the idea that some experimental designs can result in questionable, or sometimes obviously unethical, situations. For example, charging users different amounts of money for the same product or service is not good practice (and possibly illegal).

2

Randomization refers to the manner in which experimental units are selected for inclusion in the study, and how they are assigned to experimental conditions. This is a key requirement for valid statistical inference in the sense that improper randomization can lead to conclusions that are based on extraneous factors not under study. For example, if one treatment accidentally includes an overwhelming majority of females in a website experiment, a significant change in the response variable might be attributed to the gender imbalance as opposed to the version of the website.

Replication refers to a repeated run of each experimental condition. This allows us to estimate the experimental error and be more confident in any conclusions that we may draw. As we increase the number of trials, we can better account for random variation inherent to the experiment, and obtain estimates that are more representative of the true population. We use power analyses to systematically determine an appropriate number of replications for a given experiment.

Blocking is the mechanism by which nuisance factors are controlled for. This entails carrying out each experimental condition at each level of the nuisance factor. For example, if it is expected that day of week is a nuisance factor for website traffic, then it would be necessary to carry out separate versions of the same experiment for each day of the week. We expect that the variation in website traffic within each day is smaller than the variation between days, which allows us to make separate, but valid, conclusions for each day.

3

An A/B test is a specific type of experiment that evaluates the effect of two different treatments on a response variable. It has a single-factor two-level design where we are interested in comparing two means, proportions, or variances. Depending on the nature of the problem, you would designate an appropriate response variable and two levels for the factor of interest, and then choose an experimental design that falls within these considerations and any assumptions. Randomization, replication, and blocking should be implemented as necessary. You would then conduct the experiment, collecting (an adequate amount of) data for both (randomized) groups, and compare the results using the hypothesis test that was chosen as part of the experimental design. If significant results are found, then these can be applied in production; if insignificant results are found, then continued experimentation may be necessary or desired.

4

We assume $y_{ij} \sim N(\mu_j, \sigma_j^2)$ where y_{ij} denotes the response measurement for unit $i = 1, 2, ..., 1000$ in treatment $j = 1, 2$. We also assume that observations in the different treatments are independent. In order to check the first assumption, we plot histograms for the response in each treatment:

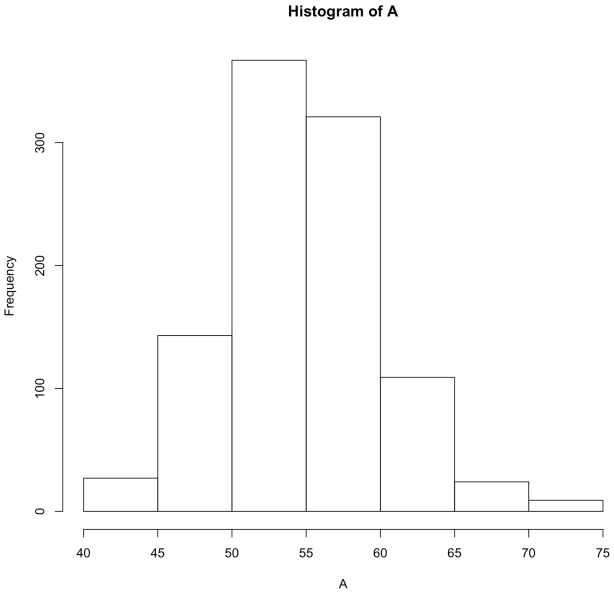

Figure 1: Histogram for the response in treatment A.

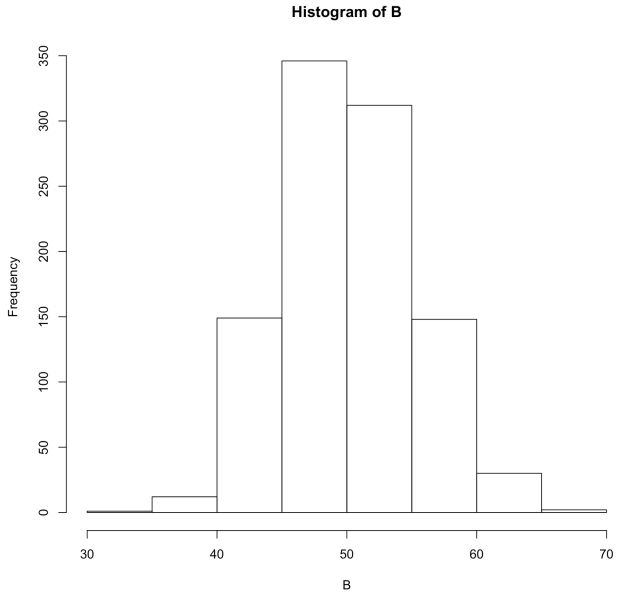

Figure 2: Histogram for the response in treatment B.

These look normal, but we use the `shapiro.test` in R to make sure. In each test, the null hypothesis is that the data are normally distributed, while the alternative hypothesis is that they are not normally distributed. We obtain $p = 9.246 \times 10^{-6}$ for treatment A, and $p = 0.0218$ for treatment B. Both p-values are significant, indicating that we should reject the null hypothesis and conclude that the data are not normally distributed. However, t-tests are relatively robust to non-normality, and since the data generally looks normal from Figures 1 and 2, we will go ahead and assume that this assumption is valid. The second assumption can only be verified through proper randomization in the experimental design phase. Since we don't have access to this information, we will go ahead and assume independence.

We also need to check whether the variances in each treatment are equal or not, i.e. whether our initial assumption of $y_{ij} \sim N(\mu_j, \sigma_j^2)$ should actually be $y_{ij} \sim N(\mu_j, \sigma^2)$. The variance for treatment A is $\sigma = 26.3821$, and the variance for treatment B is $\sigma = 26.8142$. In order to check whether these are significantly different, we use the `var.test` in R. The null hypothesis is that the variances are equal, and the alternative hypothesis is that they are not equal. We obtain $p = 0.7974$, which is insignificant, indicating that we should not reject the null hypothesis and conclude that the variances are equal. Hence, we assume $y_{ij} \sim N(\mu_j, \sigma^2)$.

Using `t.test` in R, we test our initial hypothesis: $H_0 : \mu_A \leq \mu_B$ vs. $H_A : \mu_A > \mu_B$. We obtain $p < 2.2 \times 10^{-16}$, which is significant, indicating that we should reject the null hypothesis and conclude that the mean of treatment A is larger than the mean of treatment B. Hence, implementing treatment A will, on average, result in larger purchase prices.

Using `power.t.test` in R, we can calculate the power of this hypothesis test. We obtain power $= 0.9976$. Assuming a slightly different situation where we did not have n and wanted to achieve 90% power, we can use the same function to calculate the sample size. We obtain $n = 428.8705 \approx 429$, which is the number of units required for each treatment.

5

```
info = two-sample hypothesis test using data1 and data2     # an example would be t.test in R
t_star = extract test statistic from info
data = c(data1, data2)
stat_list = vector()
for (i in 1:B):                                             # B = 1000, 2000, etc.
   randomized = sample(data, length(data), replace = FALSE)
   new_data1 = randomized[1:length(data1)]
   new_data2 = randomized[length(data1)+1:length(data)]
   new_info = two-sample hypothesis test using new_data1 and new_data2     # same as above
   t = extract test statistic from new_info
   stat_list = c(stat_list, t)
final_info = one-sample hypothesis test using stat_list and t_star
p = extract p-value from final_info
if (p <= significance level):
   reject null hypothesis
else:
   do not reject null hypothesis
```

6

Since this is a test of proportions, we use the test statistic (derived in class)

$$t = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \sim \mathrm{N}(0,1)$$

Power is equivalent to

$$\begin{aligned}
1 - \beta &= \mathrm{P}(\text{reject } H_0 | H_0 \text{ false}) \\
&= \mathrm{P}(\text{test statistic is in rejection region} | \pi_1 \neq \pi_2) \\
&= \mathrm{P}\left(\frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \geq z_{1-\alpha/2} \text{ or } \frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \leq z_{\alpha/2} | \pi_1 - \pi_2 = \delta \right) \\
&= \mathrm{P}\left(\frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \geq z_{1-\alpha/2} | \pi_1 - \pi_2 = \delta \right) + \mathrm{P}\left(\frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \leq z_{\alpha/2} | \pi_1 - \pi_2 = \delta \right) \\
&\quad \mathrm{P}\left(\frac{(p_1 - p_2) - \delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \geq z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \right) + \\
&\quad \mathrm{P}\left(\frac{(p_1 - p_2) - \delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \leq -z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \right) \\
&= \mathrm{P}\left(Z \geq z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \right) + \mathrm{P}\left(Z \leq -z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \right)
\end{aligned}$$

For any $\delta \neq 0$, only one term dominates, so hereafter we can just consider

$$1 - \beta = \mathrm{P}\left(Z \geq z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}} \right)$$

We can define

$$z_\beta = z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}}$$

Solving for n, we get

$$n = \frac{(p_1(1-p_1) + p_2(1-p_2))(z_{1-\alpha/2} - z_\beta)^2}{\delta^2}$$
$$= \frac{(p_1(1-p_1) + p_2(1-p_2))(z_{1-\alpha/2} - z_\beta)^2}{(p_1 - p_2)^2}$$

where we have assigned $\delta = p_1 - p_2$ as the effect size.

If our Type-I and Type-II error rates are 5%, then α = significance level = P(Type-I error) = 0.05, and β = P(Type-II error) = 0.05 so power = $1 - \beta = 0.95$. In our sample size formula, $z_{1-\alpha/2}$ is the critical value of the normal distribution at $1 - \alpha/2$ (e.g. for $\alpha = 0.05$ the critical value is $z_{0.975} = -1.9600$) and z_β is the critical value of the normal distribution at β (e.g. for $\beta = 0.05$ the critical value is $z_{0.05} = 1.6449$). Plugging these in as well as our sample proportions $p_1 = 0.78$ and $p_2 = 0.84$, we get

$$n = \frac{(0.78(1-0.78) + 0.84(1-0.84))(-1.9600 - 1.6449)^2}{(0.78 - 0.84)^2} = 1104.6008 \approx 1105$$

Using `power.prop.test` in R, we obtain $n = 1108.0810 \approx 1108$.

7

Region	Ad1	Ad2	Ad3	Ad4	Ad5
NE	0.2964	0.2024	0.1028	0.2938	0.2044
MW	0.2084	0.3172	0.0976	0.1984	0.1978
S	0.1478	0.1984	0.1044	0.2060	0.3020
W	0.2970	0.2090	0.1052	0.3016	0.1988

Region	χ^2	p	reject?
NE	745.21	$< 2.2 \times 10^{-16}$	yes
MW	746.22	$< 2.2 \times 10^{-16}$	yes
S	708.68	$< 2.2 \times 10^{-16}$	yes
W	760.88	$< 2.2 \times 10^{-16}$	yes

In all cases, we reject the null hypothesis and conclude that at least one of the proportions is different. Using the first table to identify the ads with the largest and smallest proportion in each region, we set up the following tests:

Region NE

H_0	H_A	t	p	reject?
$\pi_1 \leq \pi_2$	$\pi_1 > \pi_2$	10.926	$< 2.2 \times 10^{-16}$	yes
$\pi_1 \leq \pi_3$	$\pi_1 > \pi_3$	24.959	$< 2.2 \times 10^{-16}$	yes
$\pi_1 \leq \pi_4$	$\pi_1 > \pi_4$	0.28501	0.3878	no
$\pi_1 \leq \pi_5$	$\pi_1 > \pi_5$	10.677	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_2$	$\pi_3 < \pi_2$	13.982	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_4$	$\pi_3 < \pi_4$	24.667	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_5$	$\pi_3 < \pi_5$	14.23	$< 2.2 \times 10^{-16}$	yes

Region MW

H_0	H_A	t	p	reject?
$\pi_2 \leq \pi_1$	$\pi_1 > \pi_2$	12.454	$< 2.2 \times 10^{-16}$	yes
$\pi_2 \leq \pi_3$	$\pi_1 > \pi_3$	28.13	$< 2.2 \times 10^{-16}$	yes
$\pi_2 \leq \pi_4$	$\pi_1 > \pi_4$	13.705	$< 2.2 \times 10^{-16}$	yes
$\pi_2 \leq \pi_5$	$\pi_1 > \pi_5$	13.781	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_1$	$\pi_3 < \pi_1$	15.573	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_4$	$\pi_3 < \pi_4$	14.337	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_5$	$\pi_3 < \pi_5$	14.262	$< 2.2 \times 10^{-16}$	yes

Region S

H_0	H_A	t	p	reject?
$\pi_5 \leq \pi_1$	$\pi_1 > \pi_2$	18.788	$< 2.2 \times 10^{-16}$	yes
$\pi_5 \leq \pi_2$	$\pi_1 > \pi_3$	12.045	$< 2.2 \times 10^{-16}$	yes
$\pi_5 \leq \pi_3$	$\pi_1 > \pi_4$	25.327	$< 2.2 \times 10^{-16}$	yes
$\pi_5 \leq \pi_4$	$\pi_1 > \pi_5$	11.093	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_1$	$\pi_3 < \pi_1$	6.5502	3.017×10^{-11}	yes
$\pi_3 \geq \pi_2$	$\pi_3 < \pi_2$	13.225	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_4$	$\pi_3 < \pi_4$	14.168	$< 2.2 \times 10^{-16}$	yes

Region W

H_0	H_A	t	p	reject?
$\pi_4 \leq \pi_1$	$\pi_1 > \pi_2$	0.50219	0.3078	no
$\pi_4 \leq \pi_2$	$\pi_1 > \pi_3$	10.678	$< 2.2 \times 10^{-16}$	yes
$\pi_4 \leq \pi_3$	$\pi_1 > \pi_4$	25.153	$< 2.2 \times 10^{-16}$	yes
$\pi_4 \leq \pi_5$	$\pi_1 > \pi_5$	11.95	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_1$	$\pi_3 < \pi_1$	24.639	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_2$	$\pi_3 < \pi_2$	14.408	$< 2.2 \times 10^{-16}$	yes
$\pi_3 \geq \pi_5$	$\pi_3 < \pi_4$	13.146	$< 2.2 \times 10^{-16}$	yes

Our results indicate that for region NE, the best ads were 1 and 4 (no significant difference between them), and the worst ad was 3. For region MW, the best ad was 2, and the worst ad was 3. For

region S, the best ad was 5, and the worst ad was 3. For region W, the best ads were 1 and 4 (no significant difference between them), and the worst ad was 3. While the worst ad was 3 across all regions, we see that the best ads differ among regions. These results confirm that geographic region was indeed a factor that affects the outcome of the tests. By implementing blocking, the experimenters were able to control for this nuisance factor, and figure out which ad was best within each block. This allows Nike to show the best ad selected for each region in order to maximize click-through-rates in all regions. Had the data not been pooled in this way, then a single best ad would have been selected, which might have only maximized click-through-rates for a subset of the regions.

8

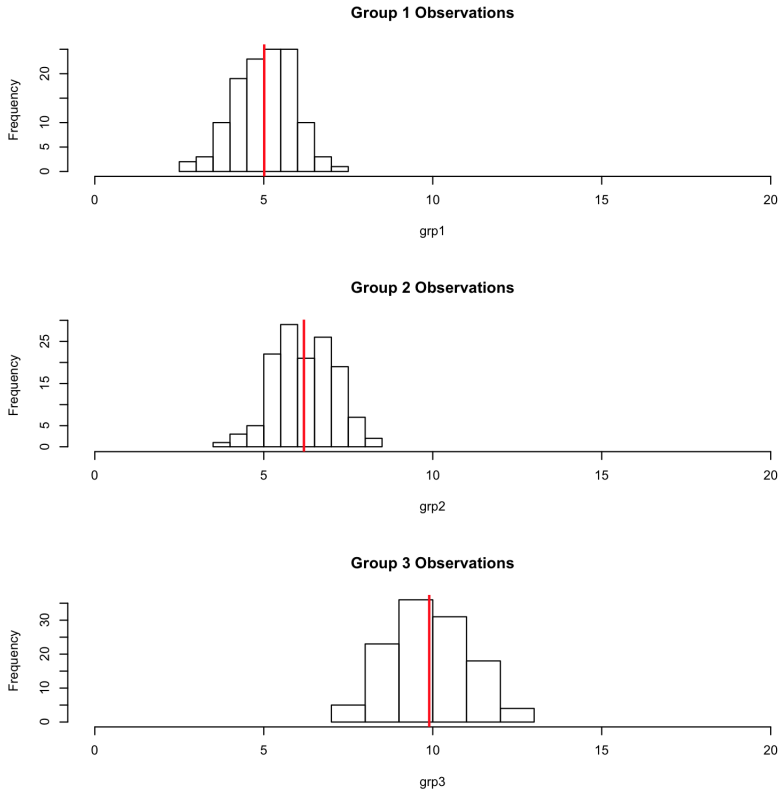

Figure 3: Histogram of each group. We see that the first two groups are relatively similar, but the third group is quite different. We expect the F-test to reject the null hypothesis.

Fitting a linear regression in R and using the `summary` function, we obtain $F = 851.9$ and $p < 2.2 \times 10^{-16}$. Hence, we reject the null hypothesis and conclude that at least one of the means is different. We desire $\alpha = 0.05$, so using the Bonferroni correction, we will substitute in $\frac{\alpha}{k} = \frac{0.05}{2} = 0.025$. Here we use $k = 2$ because we will be conducting the two tests specified in the table below. These were chosen because group 3's sample mean was the largest.

H_0	H_A	t	p	reject?
$\mu_3 \leq \mu_1$	$\mu_3 > \mu_1$	37.456	$< 2.2 \times 10^{-16}$	yes
$\mu_3 \leq \mu_2$	$\mu_3 > \mu_2$	28.835	$< 2.2 \times 10^{-16}$	yes

In both cases, we reject the null hypothesis and conclude that group 3's mean is larger. This is sufficient evidence to support the claim that group 3 is associated with the largest average length of game play.

9

(a)

$$\begin{aligned}
&\mathrm{P}_{\text{experimentwise}}(\text{type-I error}) = 1 - \mathrm{P}_{\text{experimentwise}}(\text{no type-I error}) \\
&= 1 - \mathrm{P}_1(\text{no type-I error})\mathrm{P}_2(\text{no type-I error}) \cdots \mathrm{P}_k(\text{no type-I error}) \\
&= 1 - (1-\alpha)(1-\alpha)\cdots(1-\alpha) \\
&= 1 - (1-\alpha)^k
\end{aligned}$$

(b)

$$\mathrm{P}_{\text{experimentwise}}(\text{type-I error}) = 1 - \left(1 - \frac{\alpha}{k}\right)^k$$

Taking part of this and letting $k \to \infty$, we get

$$\begin{aligned}
\lim_{k\to\infty} \left(1 - \frac{\alpha}{k}\right)^k &= \lim_{k\to\infty} e^{\log\left(1-\frac{\alpha}{k}\right)^k} \\
&= \lim_{k\to\infty} e^{k\log\left(1-\frac{\alpha}{k}\right)} \\
&= e^{\lim_{k\to\infty} \frac{\log\left(1-\frac{\alpha}{k}\right)}{\frac{1}{k}}}
\end{aligned}$$

Using L'Hospital's rule, we get

$$\begin{aligned}
&e^{\lim_{k\to\infty} \frac{\frac{1}{1-\frac{\alpha}{k}}\frac{\alpha}{k^2}}{-\frac{1}{k^2}}} \\
&= e^{\lim_{k\to\infty} -\frac{\alpha}{1-\frac{\alpha}{k}}} \\
&= e^{-\alpha}
\end{aligned}$$

Putting it all together, we have

$$P_{\text{experimentwise}}(\text{type-I error}) = 1 - e^{-\alpha}$$

as desired.

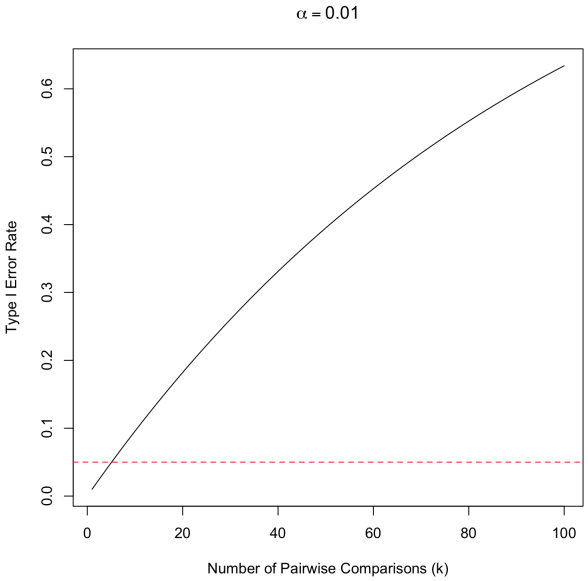

Figure 4: Demonstration of how Type-I error increases as k increases.

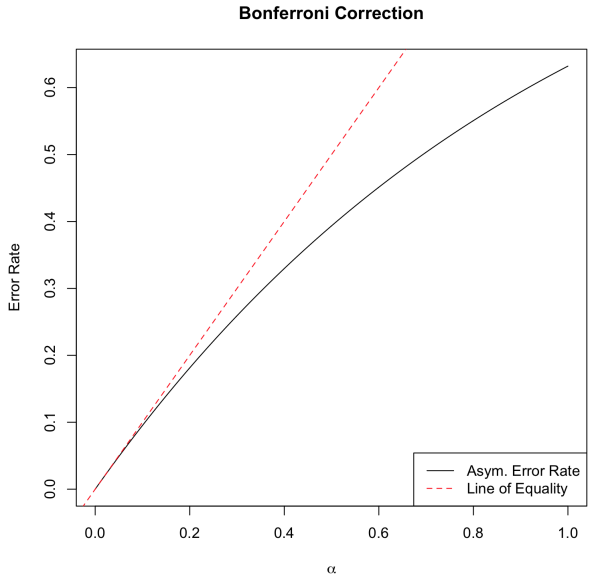

Figure 5: Demonstration of how the Bonferroni correction keeps the desired Type-I error for reasonable α.

As shown above (done in class), the Type-I error quickly gets much larger than our desired α as k increases. However, we also see that $1-e^{-\alpha}$ versus α has approximately a slope of 1 for $\alpha \in [0, 0.2]$. This means that we can say $1 - e^{-\alpha} \approx \alpha$ for $\alpha \in [0, 0.2]$. Since this range of α is generally what we use in statistics, we can see that the Bonferroni correction gets us back to our original desired significance level.

10

Equivalence testing is when you switch around the null and alternative hypotheses associated with standard hypothesis testing. It is used when we are interested in figuring out whether two groups are similar. We can't use standard hypothesis testing because this requires us to assume that the groups are the same, which is exactly what we are trying to figure out. Naturally, we switch around the null and alternative hypotheses as follows:

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

This brings up the question of how exactly to define "the same." Since we are working with data in the real world, it is highly unlikely (if not impossible) that two groups will be exactly the same. So we specify some practical value (context-specific), and re-state the equivalence test as follows:

$$H_0 : |\mu_1 - \mu_2| > \Delta \text{ vs. } H_A : |\mu_1 - \mu_2| < \Delta$$

Manipulating the null hypothesis, we get:

$$H_0 : \mu_1 - \mu_2 < -\Delta \text{ or } \mu_1 - \mu_2 > \Delta$$

From this, we can derive the most basic form of equivalence testing (TOST), which uses the following test statistics:

$$\frac{(\bar{y}_1 - \bar{y}_2) + \Delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{1-\alpha} \text{ or } \frac{(\bar{y}_1 - \bar{y}_2) - \Delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -z_{1-\alpha}$$

The interpretation is similar to standard hypothesis testing, except we require both tests to be significant in order to reject the null hypothesis and conclude that the two groups are the same within some Δ.

One advantage to equivalence testing – beyond the fact that it is a more rigorous way to show that two groups are the same – is that the specification of Δ gives us more control over what amount of difference is actually significant. Given that this amount is context-specific, we can easily adjust the test for many different types of situations. Additionally, this means that we don't run into the same problem with standard hypothesis testing, where, given a large enough sample size, we can always find sufficient evidence to reject the null hypothesis. Equivalence testing allows for scientifically irrelevant differences while still concluding that the groups are the same. Lastly, it penalizes poor precision and small n values.

11

$$\begin{aligned}
\mathrm{E}[\mathrm{MS_E}] &= \mathrm{E}\left[\frac{\mathrm{SS_E}}{N-a}\right] \\
&= \frac{1}{N-a}\mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2\right] \\
&= \frac{1}{N-a}\mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2 - 2y_{ij}\bar{y}_{i.} + \bar{y}_{i.}^2)\right] \\
&= \frac{1}{N-a}\mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - 2n\sum_{i=1}^{a}\bar{y}_{i.}^2 + n\sum_{i=1}^{a}\bar{y}_{i.}^2\right] \\
&= \frac{1}{N-a}\mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a}y_{i.}^2\right]
\end{aligned}$$

$$= \frac{1}{N-a}\mathrm{E}\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}\left(n\mu+n\tau_i+\epsilon_{i.}\right)^2\right]$$

Given that $\mathrm{E}[\epsilon_{ij}] = \mathrm{E}[\epsilon_{i.}] = \mathrm{E}[\epsilon_{..}] = 0, \mathrm{E}[\epsilon_{ij}^2] = \sigma^2, \mathrm{E}[\epsilon_{i.}^2] = n\sigma^2,$ and $\mathrm{E}[\epsilon_{..}^2] = an\sigma^2$, after squaring and taking the expectation, we obtain:

$$= \frac{1}{N-a}\left(N\mu^2 + n\sum_{i=1}^{a}\tau_i^2 + N\sigma^2 - N\mu^2 - n\sum_{i=1}^{a}\tau_i^2 - a\sigma^2\right)$$
$$= \sigma^2$$

The next one is derived similarly:

$$\mathrm{E}[\mathrm{MS}_{\mathrm{Cond}}] = \mathrm{E}\left[\frac{\mathrm{SS}_{\mathrm{Cond}}}{a-1}\right]$$
$$\mathrm{SS}_{\mathrm{Cond}} = \frac{1}{n}\sum_{i=1}^{a}y_{i.}^2 - \frac{1}{an}y_{..}^2$$
$$\mathrm{E}[SS_{\mathrm{Cond}}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{a}y_{i.}^2\right] - \mathrm{E}\left[\frac{1}{an}y_{..}^2\right]$$

The first term is:

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{a}y_{i.}^2\right] = \frac{1}{n}\sum_{i=1}^{a}\mathrm{E}[(n\mu+n\tau_i+\epsilon_{i.})^2]$$

Again, given that $\mathrm{E}[\epsilon_{ij}] = \mathrm{E}[\epsilon_{i.}] = \mathrm{E}[\epsilon_{..}] = 0, \mathrm{E}[\epsilon_{ij}^2] = \sigma^2, \mathrm{E}[\epsilon_{i.}^2] = n\sigma^2,$ and $\mathrm{E}[\epsilon_{..}^2] = an\sigma^2$, after squaring and taking the expectation, we obtain:

$$= \frac{1}{n}\left[a(n\mu)^2 + n^2\sum_{i=1}^{a}\tau_i^2 + an\sigma^2\right]$$
$$= an\mu^2 + n\sum_{i=1}^{a}\tau_i^2 + a\sigma^2$$

The second term is:

$$\mathrm{E}\left[\frac{1}{an}y_{..}^2\right] = \frac{1}{an}\mathrm{E}\left[\left(an\mu + n\sum_{i=1}^{a}\tau_i + \epsilon_{..}\right)^2\right]$$
$$= \frac{1}{an}\mathrm{E}\left[(an\mu+\epsilon_{..})^2\right]$$

since $\sum_{i=1}^{a} \tau_i = 0$. Again, considering the above, squaring, and taking the expectation, we obtain:

$$= \frac{1}{an}\left[(an\mu)^2 + an\sigma^2\right]$$
$$= an\mu^2 + \sigma^2$$

All together, we have:

$$\mathrm{E}[SS_{\mathrm{Cond}}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{a} y_{i\cdot}^2\right] - \mathrm{E}\left[\frac{1}{an}y_{\cdot\cdot}^2\right]$$
$$= an\mu^2 + n\sum_{i=1}^{a}\tau_i^2 + a\sigma^2 - (an\mu^2 + \sigma^2)$$
$$= \sigma^2(a-1) + n\sum_{i=1}^{a}\tau_i^2$$

or

$$\mathrm{E}[\mathrm{MS}_{\mathrm{Cond}}] = \mathrm{E}\left[\frac{\mathrm{SS}_{\mathrm{Cond}}}{a-1}\right]$$
$$= \frac{\sigma^2(a-1) + n\sum_{i=1}^{a}\tau_i^2}{a-1}$$
$$= \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

12

(a)

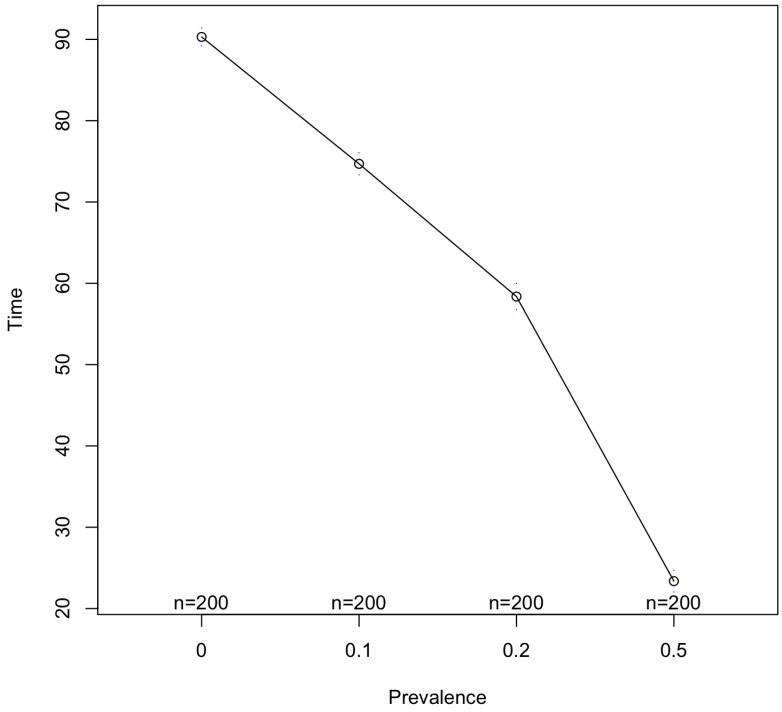

Figure 6: Main effects plot for prevalence factor.

Based on Figure 6, I believe prevalence is a significant factor. This is because the difference in time goes from about 90 to 25, which seems to be a significant loss in terms of user time on site. As expected, the plot shows that a prevalence of 0.5 is the most detrimental to user experience. This makes sense since a higher ad rate will inevitably deter users from the site.

(b)

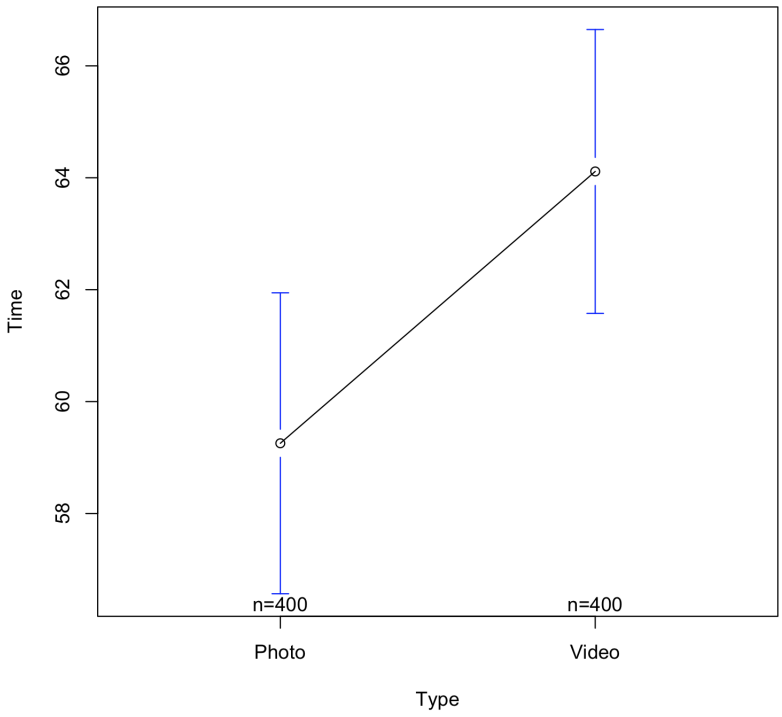

Figure 7: Main effects plot for type factor.

Based on Figure 7, I believe type is a significant factor, although not nearly as significant as prevalence. This is because the difference in time goes from about 64 to 59, which could still be a significant loss – albeit obviously a smaller one – in terms of user time on site. The plot shows that the type corresponding to photo is the most detrimental to user experience. This could be because videos require a certain amount of time, and so we might expect to see longer times on site associated with this factor level.

(c)

Given that we have two different factors each with their own set of levels, when we talk about interaction, we are talking about how all the different combinations of levels for each factor affect the response. Holding prevalence constant at 0.1 and switching types, we can compare the difference in response. Holding prevalence constant again, but this time at 0.2, and switching types, we can again compare the difference in response. We say there is an interaction between prevalence and type when these two differences are not equal. This makes intuitive sense since we would expect the response to change by a similar amount – unless it has some other dependence that might affect it, which is exactly what is meant by interaction.

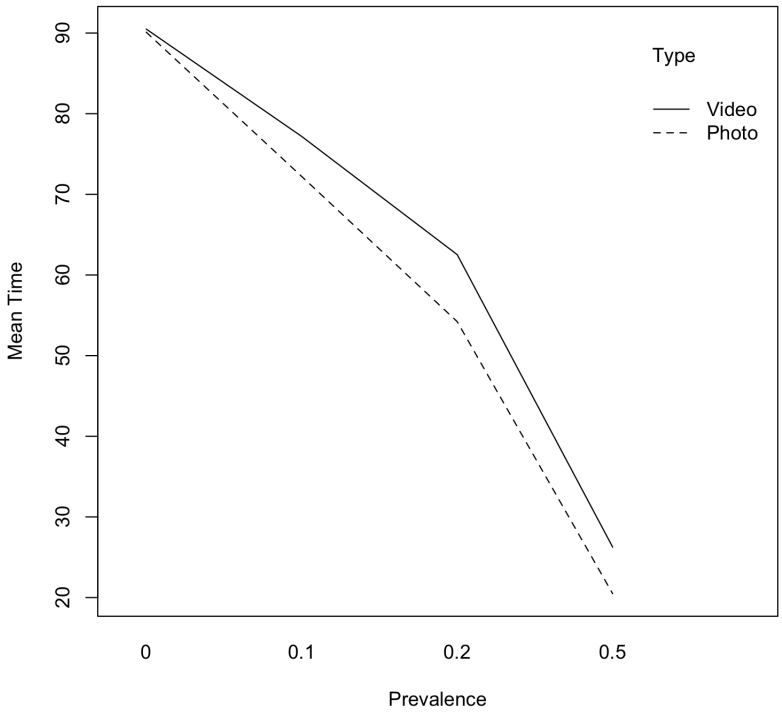

Figure 8: Interaction plot for prevalence and type factors.

Based on Figure 8, I believe the interaction effect is significant. In order for there to be no interaction, or insignificant interaction, we would require the plot to be piecewise parallel (the solid and dotted lines between each set of points should be nearly parallel). This is obviously not the case, so I expect the interaction effect to be significant.

(d)

```
Call:
lm(formula = y ~ as.factor(prev) * as.factor(type))

Residuals:
    Min      1Q  Median      3Q     Max
-29.657  -5.925  -0.015   5.921  30.850

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            90.1075     0.9341  96.462  < 2e-16 ***
as.factor(prev)0.1                    -17.8942     1.3210 -13.545  < 2e-16 ***
as.factor(prev)0.2                    -35.8900     1.3210 -27.168  < 2e-16 ***
as.factor(prev)0.5                    -69.6244     1.3210 -52.704  < 2e-16 ***
as.factor(type)Video                    0.3868     1.3210   0.293  0.76974
as.factor(prev)0.1:as.factor(type)Video   4.5785     1.8682   2.451  0.01447 *
as.factor(prev)0.2:as.factor(type)Video   7.9217     1.8682   4.240  2.5e-05 ***
as.factor(prev)0.5:as.factor(type)Video   5.3829     1.8682   2.881  0.00407 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.341 on 792 degrees of freedom
Multiple R-squared:  0.8785,    Adjusted R-squared:  0.8775
F-statistic: 818.3 on 7 and 792 DF,  p-value: < 2.2e-16
```

Figure 9: R output for summary statistics on the full model.

From Figure 9, we can see that all effects are significant except for the main effect of type. This is mostly consistent with the above results, since I predicted that the main effect of prevalence as well as the interaction effects would be significant. I also predicted that the main effect of type would be significant, and while this was not actually the case, I did mention this would be much less significant as compared to prevalence.

(e)

Based on the above results, the combination that appears to maximize engagement is using video ads with a prevalence of 0. Practically, this doesn't make sense since you can't have any ads if your prevalence is going to be 0. Additionally, this result could have essentially been inferred through common sense. While it is useful to have a prevalence of 0 to use as a control, in real life we would exclude this level as an option and choose a different level that maximized user engagement within realistic bounds for ad rates.

(f)

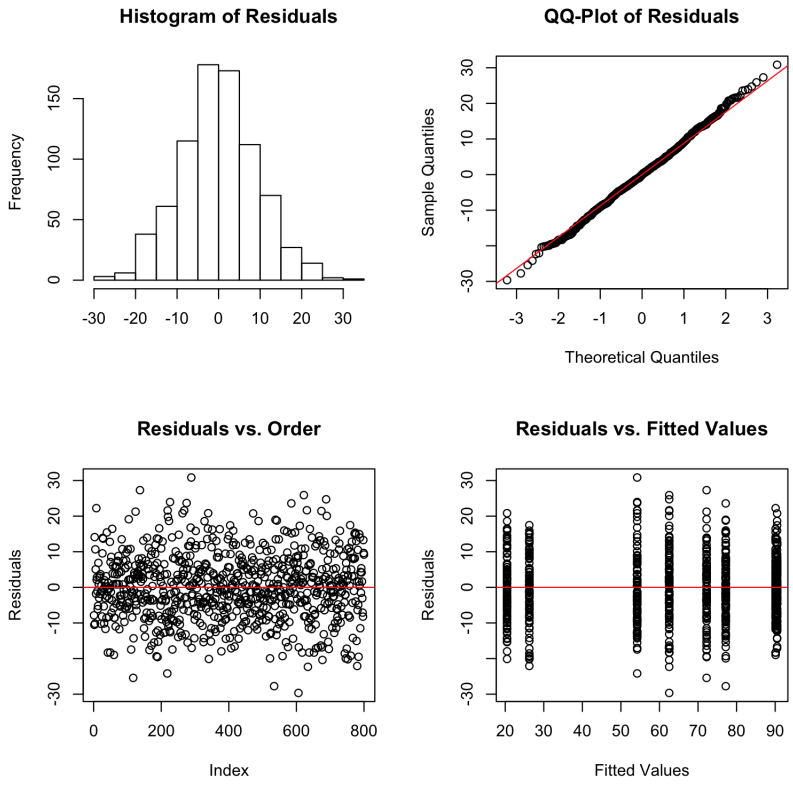

Figure 10: Residual diagnostics for the full model.

We see that the residuals look to be normal since they follow a bell curve in the histogram and lay on the diagonal line in the QQ-plot. From the remaining two plots, we see that the residuals look to be uncorrelated since there are no obvious patterns, have mean zero since they are randomly

scattered about zero, and have constant variance since their spread is at the same level throughout. Hence, the typical ordinary least squares assumptions appear to be met.

13

(a)

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

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} y_1 + y_2 + y_3 + y_4 \\ -y_1 + y_2 - y_3 + y_4 \\ -y_1 - y_2 + y_3 + y_4 \\ y_1 - y_2 - y_3 + y_4 \end{bmatrix}$$

All together, we have:

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

$$= \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} y_1 + y_2 + y_3 + y_4 \\ -y_1 + y_2 - y_3 + y_4 \\ -y_1 - y_2 + y_3 + y_4 \\ y_1 - y_2 - y_3 + y_4 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{y_1+y_2+y_3+y_4}{4} \\ \frac{-y_1+y_2-y_3+y_4}{4} \\ \frac{-y_1-y_2+y_3+y_4}{4} \\ \frac{y_1-y_2-y_3+y_4}{4} \end{bmatrix}$$

Hence, the least squares estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are $\frac{y_1+y_2+y_3+y_4}{4}$, $\frac{-y_1+y_2-y_3+y_4}{4}$, and $\frac{-y_1-y_2+y_3+y_4}{4}$, respectively. Now, if we take out the interaction term, we have:

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} y_1 + y_2 + y_3 + y_4 \\ -y_1 + y_2 - y_3 + y_4 \\ -y_1 - y_2 + y_3 + y_4 \end{bmatrix}$$

All together, we have:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ &= \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} y_1 + y_2 + y_3 + y_4 \\ -y_1 + y_2 - y_3 + y_4 \\ -y_1 - y_2 + y_3 + y_4 \end{bmatrix} \\ &= \begin{bmatrix} \frac{y_1+y_2+y_3+y_4}{4} \\ \frac{-y_1+y_2-y_3+y_4}{4} \\ \frac{-y_1-y_2+y_3+y_4}{4} \end{bmatrix} \end{aligned}$$

Hence, the least squares estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are again $\frac{y_1+y_2+y_3+y_4}{4}$, $\frac{-y_1+y_2-y_3+y_4}{4}$, and $\frac{-y_1-y_2+y_3+y_4}{4}$, respectively. We have therefore shown that the presence or absence of the interaction term does not affect the least squares estimates.

(b)

The variance-covariance matrix is given by:

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \sigma^2 (X^T X)^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix} \end{aligned}$$

using our results from above. Hence, the variance of each $\hat{\beta}$ is $\sigma^2/4$, and the standard error of each is $\sqrt{\sigma^2/4} = \sigma/2$. We have therefore shown that the standard errors of the least squares estimates are all the same.

(c)

The intuition behind why the least squares estimates are only half of the corresponding factor's effect estimate lies in the notation. Here we define an effect as the difference in the means between the high and the low levels of a factor, whereas the least squares coefficient in the model is the difference between the marginal mean and the overall mean. The latter will only get you halfway to the former, so this explains why the least squares estimates are interpreted in this way.

14

(a)

In a 2^{k-p} fractional factorial design, we are investigating k factors in $(\frac{1}{2})^p 2^k$ conditions. For example, if we originally have $k = 6$ factors and are interested in only using 2^4 conditions, then we can do this with a 2^{6-2} fractional factorial design. The key idea is that instead of using a full factorial experiment, which in our above case would require 64 distinct conditions, we only use a fraction of these conditions. This allows us to either (1) investigate the same number of factors in fewer conditions, or (2) investigate a larger number of factors in the same number of conditions. This works by aliasing (see below) p of the interactions – what you are giving up – with each of the p additional factors – what you gain. You conduct the experiment in the same way as before, but the interpretations become slightly more complicated since main effects are now confounded with interactions (i.e. we can't separately identify their individual effects). This is fine as long as your resolution (see below) is adequately high.

(b)

Aliasing is essentially when you exchange one factor for another. This commonly takes the form of exchanging a high-order interaction – which you believe does not have a significant effect on the response – with an additional factor – whose main effect you do believe has a significant effect on the response. When you alias the main effect of your additional factor with the interaction, you are given a prescription dictating when to run that factor at its low and high levels. In other words, the design matrix stays the same even though you have changed what each condition entails. As mentioned above, this complicates the interpretation since main effects are now confounded with interactions. In an effort to quantify this idea, we have resolution.

Resolution is a metric that quantifies our ability to separately estimate effects. Specifically, a design is of resolution R if main effects are aliased with interaction effects involving at least $R - 1$ factors. Hence, higher resolutions are preferable since this corresponds to confounding main effects with higher-order interactions that are typically less significant. For example, a resolution-XXX design entails confounding a main effect with a 2-factor interaction, while a resolution-IV design entails confounding a main effect with a 3-factor interaction. The latter is much better since it is highly likely that the 3-factor interaction is less significant than the 2-factor interaction. Our interpretation will be much more straightforward since we are more confident that the effect is due to the additional factor, and not the interaction.