Andre Guimaraes Duarte
MSAN 631-01
Final Exam
June 28, 2017

1. Discuss the merits and drawbacks of an experiment.

> **Solution:** A well planned and conducted experiment has the tangible merit of establishing and quantifying cause-and-effect relationships between factors and a response. Since the experiment is carefully laid out before the data collection, it is usually fast and efficient, seeing as we know how much and what type of data we need to collect.
>
> However, an experiment may be risky, since an active modification of a factor level can lead to a decrease in activity/return. The experiment must be supervised closely so that the potential negative cost on the product is minimized. In addition, experiments may cross the border of ethics, since we purposefully give some users a different treatment than others to assess how their behavior changes.

2. Discuss the importance of replication, randomization, and blocking in the context of an experiment.

> **Solution:** Replication allows for estimation of experimental error, and it allows us to be more sure of our conclusions. By repeating each observation many times and taking (for example) the average of the responses, we can make better statistical inferences on the results.
>
> Randomization allows for balancing out extraneous factors not under study. Indeed, the experimental units should be homogeneous across conditions. Also, a random sample also facilitates valid statistical inference.
>
> Blocking allows for controlling of nuisance factors, i.e., factors that we expect/know to influence the response but that we do not care about. The goal is to choose blocks such that the response variation within a block is smaller than the response variation between blocks.

3. Briefly explain the philosophy and mechanics of an "A/B" test.

> **Solution:** In an "A/B" test, we are interested in comparing the effect of two different treatments on a response variable. For this single-factor two-level experimental design, we usually want to compare two means or proportions (or variances). Therefore, we want to test the following hypothesis:
>
> $$H_0 : \mu_1 = \mu_2 \text{ or } \pi_1 = \pi_2 \text{ vs } H_A : \mu_1 \neq \mu_2 \text{ or } \pi_1 \neq \pi_2$$
>
> We test the hypothesis by calculating a test statistic using our available samples. Under the null hypothesis, this statistic follows a known null distribution. Therefore, we evaluate how extreme this value is given the null distribution using either p-value or rejection regions. If the observed value of the test statistic is sufficiently extreme, then we reject the null hypothesis.

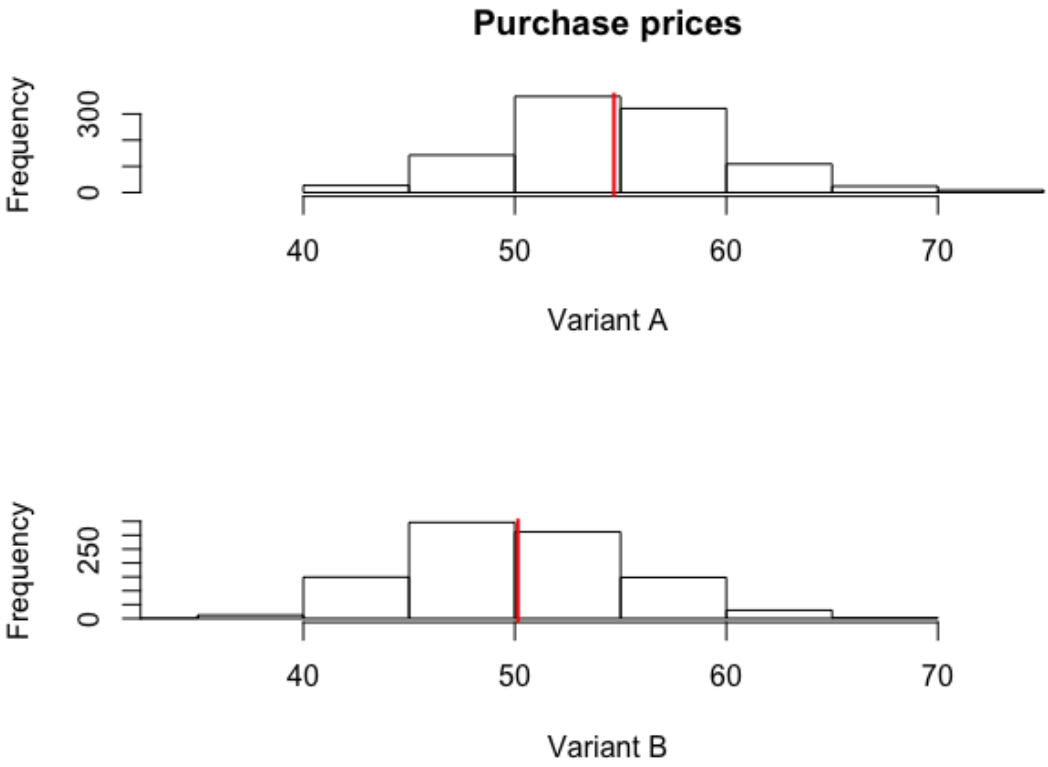
4. Suppose that an online retailer is interested in determining whether the average purchase price is larger among customers in "condition A" than among customers in "condition B". Formally:

$$H_0 : \mu_A \leq \mu_B \text{ versus } H_A : \mu_A > \mu_B$$

To test this hypothesis the retailer randomizes $n = 1000$ customers to each condition and records their purchase prices. Data for this experiment can be found in the file `retail.txt`. Use R to analyze this data and hence conclude whether condition "A" results, on average, in significantly larger purchase prices than condition "B", at a 5% level of significance. Be sure to assess the validity of any assumptions your analysis makes.

Assuming an effect size of $\delta = \mu_A - \mu_B = 1$, and a common standard deviation of 5, calculate the power of this hypothesis test, and determine the sample size that would be sufficient to achieve 90% power.

Solution: See code in `retail.R` file.



We have $\bar{y_A} = 54.68681$ and $\bar{y_B} = 50.14506$. Additionally, the two sample variances are very similar, so we assume that $\sigma_A = \sigma_B = \sigma$. Therefore, we calculate the test statistic

$$t = \frac{(\bar{y_A} - \bar{y_B}) - (\mu_A - \mu_B)}{s_p \sqrt{\frac{1}{n_A^2} + \frac{1}{n_B^2}}} \sim t_{(n_A + n_B - 2)}$$

Using the function `t.test(x = A, y = B, alternative = "greater", mu = 0, paired = F, var.equal = T, conf.level = 0.95)`, we obtain a p-value $< 2.2 \times 10^{-16} < 5\%$. Therefore, we reject the null hypothesis that $\mu_1 \leq \mu_2$.

Using the function `power.t.test(n = 1000, delta = 1, sd = 5, sig.level = 0.05, alternative = "one.sided")`, we obtain a power of 99.76% of this hypothesis test.

To achieve 90% power, a sample size of 429 (428.87 rounded up) per group would be sufficient.

5. Suppose your colleague wishes to compare two groups but they are unwilling to use parametric hypothesis tests such as t-tests for means or z-tests for proportions. Instead they wish to use a non-parametric randomization test, and they ask you for your help. Unfortunately you are busy, so you only have time to write pseudocode for such a test. Provide that pseudocode here.

Solution: Let's assume that we are testing the hypothesis

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

Initialize t∗ = $\bar{y_1} - \bar{y_2}$
$s \leftarrow y_1 + y_2$ # join all observations

Repeat B times:
$s_1 \leftarrow$ sample(all observations, n_1) # randomly assign n_1 observations to a new sample_1
$s_2 \leftarrow s - s_1$ # assign the other n_2 observations to a new sample_2
$t_b \leftarrow$ mean(s_1) − mean(s_2) # calculate the test statistic on these new samples

p−value $\leftarrow \frac{\text{count } t_b \text{ outside } \bar{t} \pm |t* - \bar{t}|}{B}$
Reject H_0 if p−value $< \alpha$

This code is also applicable for comparing proportions and easily adaptable for one-sided hypothesis testing as well.

6. Suppose we are interested in testing the following hypothesis:

$$H_0 : \pi_1 = \pi_2 \text{ versus } H_A : \pi_1 \neq \pi_2$$

where π_i is the proportion of individuals that conform to some condition in population $i = 1, 2$. Derive a formula for the sample size n that each sample would need to have to be able to this hypothesis a $100 \times \alpha\%$ significance level and a power of $1 - \beta$.

Using this formula, calculate the sample size necessary for testing this hypothesis assuming Type I and II error rates of 5%, and assuming $p_1 = 0.78$ and $p_2 = 0.84$. Repeat this calculation using the `power.prop.test()` function in R.

Solution:

$$\begin{aligned}
\text{Power} &= Prob(\text{Reject } H_0 | H_0 \text{ is false}) \\
&= P(t \in \text{ Rejection region} | \pi_1 \neq \pi_2) \\
&= P(\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \geq Z_{1-\alpha/2} \text{ or } \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq -Z_{1-\alpha/2} | \pi_1 - \pi_2 = \delta) \\
&= P(\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \geq Z_{1-\alpha/2} | \pi_1 - \pi_2 = \delta) + P(\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq -Z_{1-\alpha/2} | \pi_1 - \pi_2 = \delta) \\
&= P(\frac{(p_1 - p_2) - \delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \geq Z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} | \pi_1 - \pi_2 = \delta) + \\
&\quad P(\frac{(p_1 - p_2) - \delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq -Z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} | \pi_1 - \pi_2 = \delta)
\end{aligned}$$

Without loss of generality, we have Power $= 1 - \beta = P(Z \geq Z_\beta)$ where $Z_\beta = Z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$.

Therefore, we can write:

$$
\begin{aligned}
& Z_\beta && = Z_{1-\alpha/2} - \frac{\delta}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \\
\Leftrightarrow & Z_\beta \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} && = Z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} - \delta \\
\Leftrightarrow & \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} && = \frac{\delta}{Z_{1-\alpha/2} - Z_\beta} \\
\Leftrightarrow & \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} && = \left(\frac{\delta}{Z_{1-\alpha/2} - Z_\beta} \right)^2 \\
\Leftrightarrow & \frac{p_1(1-p_1) + k \cdot p_2(1-p_2)}{k \cdot n_2} && = \left(\frac{\delta}{Z_{1-\alpha/2} - Z_\beta} \right)^2 && (n_1 = k \cdot n_2) \\
\Leftrightarrow & n_2 && = \left(\frac{p_1(1-p_1)}{k} + p_2(1-p_2) \right) \left(\frac{Z_{1-\alpha/2} - Z_\beta}{\delta} \right)^2
\end{aligned}
$$

Using this definition, with $k = 1$ and $\delta = p_1 - p_2$, we get $n_1 = n_2 = 1104.55$.

Using the function `power.prop.test(p1 = 0.78, p2 = 0.84, sig.level = 0.05, power = 0.95, alternative = "two.sided")`, we obtain $n_1 = n_2 = 1108.08$.

We can see that the two values obtained are not exactly equal, but they are fairly close.

7. Online retailers often use banner ads to drive traffic to their websites. Nike is trying to determine which of five online banner ads is related to the largest click-through-rate. To investigate which is best, the experimentation team plans to display each of the 5 ads to various US Facebook users in their newsfeeds. Because ad impressions are expected to differ by geographic region, each of the 5 ads is displayed 5000 times in Facebook newsfeeds for IP addresses originating in each of the four major US geographic regions: Northeast, Midwest, South, and West.

   The data may be described as follows:

$$
y_{ijk} = \begin{cases} 1, & \text{if ad i was clicked by person j in region k} \\ 0, & \text{otherwise} \end{cases}
$$

   where $i = 1, \ldots, 5$ indexes ads, $j = 1, \ldots, 5000$ indexes subjects and $k = 1, 2, 3, 4$ corresponds to geographic regions Northeast, Midwest, South and West. The data for this experiment can be found in the file `nike.txt`.

   Let π_{ik} denote the click-through-rate for ad i in region (block) k. Use R to conduct a χ^2 test of the hypothesis

$$
H_0 : \pi_{1k} = \ldots = \pi_{5k} \text{ versus } H_A : \pi_{ik} \neq \pi_{lk} \text{ for some } i \neq l
$$

   in each of the four geographic regions. Use a series of pairwise tests in each region to decide which ads are associated with the highest and lowest click-through-rates per region. For purposes of this question, you may ignore the "multiple testing" problem.

   Discuss the relevance of blocking here, and why pooling data across geographic regions may be misleading.

Solution: R solution available in the file `nike.R`.

Within each region, we reject the null hypothesis that the click-through-rate (CTR) is equal for all 5 ads with $\alpha = 0.05$. Next, we want to verify within each region which ad has the highest and lowest CTR. By performing a series of pairwise tests, we obtain the following results:

Region	Highest CTR (value)	Lowest CTR (value)
NE	1 (0.2964)	3 (0.1028)
MW	2 (0.3172)	3 (0.0976)
S	5 (0.3020)	3 (0.1044)
W	4 (0.3016)	3 (0.1052)

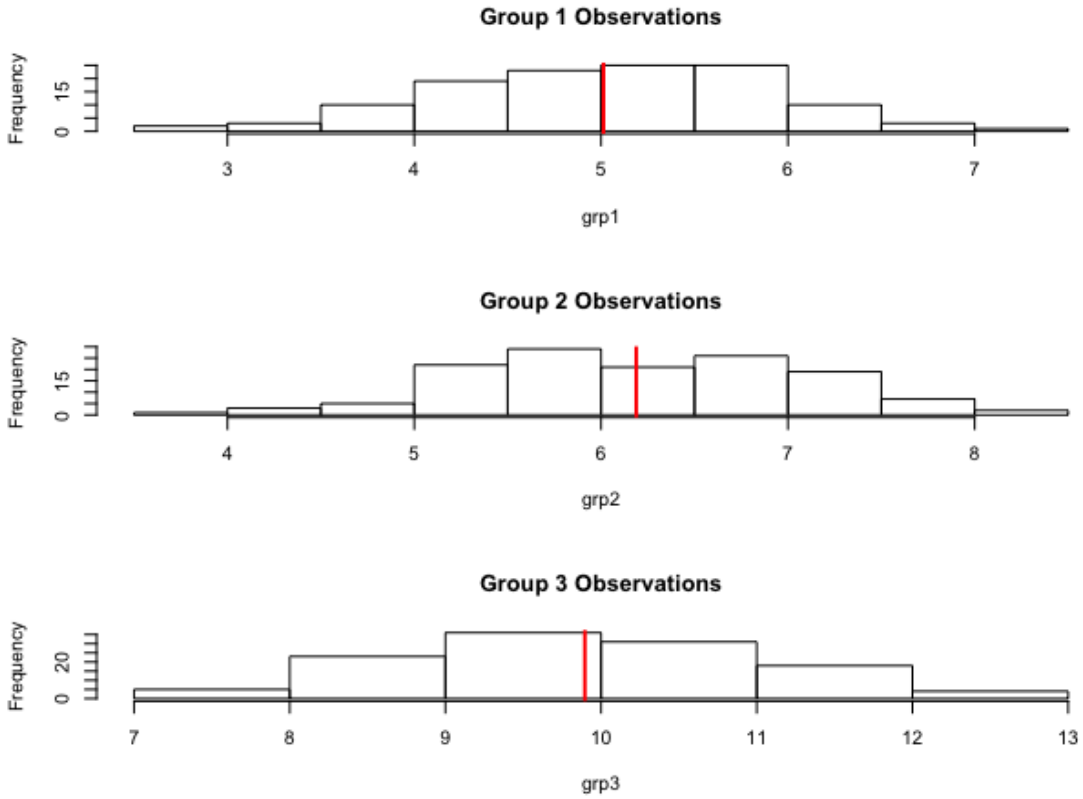
We can see that ad 3 consistently has the lowest CTR, but the ad with highest CTR varies by region. These results show that blocking by region allows us to discern specific ads that work best for each region, and we would not find the optimal ad type by region had we pooled the data across geographic regions. The results would then have been misleading. Indeed, with all regions combined, ad 4 has the highest CTR of 0.24995 (which is lower than each of the highest CTR per region).

8. Candy Crush is experimenting with three different versions of in-game "boosters": the lollipop hammer, the jelly fish, and the color bomb. Users in each condition receive (for free) 5 boosters corresponding to their condition and interest lies in evaluating the effect of these different boosters on the length of time a user plays the game. Let μ_i represent the average length of game play (in minutes) associated with booster $i = 1, 2, 3$. Use R and the data found in `candycrush.txt` to conduct an F-test of the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ versus } H_A : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Use a series of pairwise t-tests to decide which booster is associated with the largest average length of game play. For each of the pairwise tests use a significance level in accordance with the 'Bonferroni correction' to help mitigate the "multiple testing" problem.

Solution: R solution available in the file `candycrush.R`.



By looking at the ANOVA table for this model, the F-test produces a p-value $< 2.2 \cdot 10^{-16}$, so we reject the null hypothesis that the means are all equal.

Using the Bonferroni correction, the significance level for each pairwise t-test is now $\frac{0.05}{k=3} \approx 0.017$. We obtain the results that $\mu_1 < \mu_2 < \mu_3$. A histogram of each group helps visualize this result very easily.

9. (a) Show that when k hypothesis tests are being conducted, each at significance level α, the experimentwise Type I error rate is given by:

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

(b) Calculate the experimentwise Type I error rate for large k (i.e., as $k \to \infty$) when using the Bonferroni correction, and justify the efficacy of this correction.

Solution: (a)

$$P(\text{Type I error}) = 1 - P(\text{no Type I error in any test})$$

Since we are making k independent individual tests, each at significance level α, we get

$$P(\text{no Type I error in any test}) = P(\text{no Type I error in one test})^k = (1 - \alpha)^k$$

Therefore, we get

$$P(\text{Type I error}) = 1 - (1 - \alpha)^k$$

(b) Using the Bonferroni correction, we take the significance level of each of the k individual tests to be $\frac{\alpha}{k}$. So we get $P(\text{Type I error}) = 1 - (1 - \frac{\alpha}{k})^k$.

In addition, we know that $e^x = \lim_{n\to\infty}(1 + \frac{x}{n})^n$.

Therefore, as $k \to \infty$, $P(\text{Type I error}) = 1 - e^{-\alpha}$. In particular, this is useful when α is small (which is the usual scenario, i.e., $\alpha < 0.2$), because then we have $1 - e^{-\alpha} \approx \alpha$.

Using the Bonferroni correction, as $k \to \infty$, we get

$$P(\text{Type I error}) \approx \alpha$$

10. In the world of Big Data, there is typically little emphasis placed on obtaining an adequate sample size as it is often believed that exceedingly large samples are easily attainable. However, when performing inference via usual hypothesis tests and/or confidence intervals, a sample that is too large can be problematic. Consider a simple one-sample t-test of the hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_A : \mu \neq \mu_0$$

where the test statistic is given by

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Notice that as n gets very large, t also becomes very large and hence the associated p-value gets infinitesimally small. In other words, if your sample is large enough, you will always find sufficient evidence to reject the null-hypothesis.

Do some light reading on the topic of "equivalence tests". Provide a brief description of this method of hypothesis testing, and discuss some of the advantages it enjoys relative to the standard hypothesis-testing framework.

Solution: With a very large number of observations, it becomes increasingly easy to reject the null hypothesis of a t-test. So in this case, instead of wanting to show a difference between groups (as is the case in simple hypothesis testing), we want to show an equivalence between groups (i.e., $H_0 : \mu \neq \mu_0$ versus $H_A : \mu = \mu_0$). In this case, we assume that there is a difference, and we want to show, through testing, that there isn't.

The null hypothesis for equivalence testing can be rewritten with respect to a practical equivalence margin δ. This margin corresponds to the range of differences in mean such that we would consider the groups to be equivalent. The null hypothesis can then be written as $H_0 : |\mu - \mu_0| > \delta$ or even $H_0 : \mu - \mu_0 < -\delta$ or $\mu - \mu_0 > \delta$.

This last version of H_0 can then be interpreted as two one-sided t-tests. In particular, we can conclude equivalence if, and only if, the two tests are rejected. Then the difference in means is within the equivalence margin and we can claim that the means are equivalent at the specified significance level.

Equivalence testing is useful when we want to show equivalence. With traditional statistical testing, the null hypothesis that the group means are equal is never truly accepted, as we only fail to reject it. Indeed, if we want to show that the null is true, we can't begin by assuming it is true! So the idea

behind equivalence testing is to "switch" the hypotheses, so that we suppose there is a difference (null) and we will use evidence (data) to conclude that there isn't (alternative) by rejecting the initial hypothesis.

11. Consider an experiment with a continuous response and one factor with a levels. The *effects model* that describes data arising from a such a design is given by:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where $i = 1, \ldots, a$ indexes the levels of the factor, $j = 1, \ldots, n$ indexes replication and we assume the ϵ_{ij} are i.i.d. $N(0, \sigma^2)$.

In class we showed that the ANOVA F-test is an appropriate test of $H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$, and that in the context of this effects model it is equivalent to the test of $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$. The appropriateness of the ANOVA F-test relied on the fact that $E(MS_{Cond}) = E(MS_E)$ if H_0 is true. Here you are required to derive these two expected mean squares. In particular, prove that the following expected-mean-square equivalences are true.

$$E(MS_{Cond}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$
$$E(MS_E) = \sigma^2$$

Solution:

$$\begin{aligned}
E(MS_E) &= E(\tfrac{SS_E}{a(n-1)}) \\
&= \tfrac{1}{a(n-1)} E\left[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a}(y_{ij} - \bar{y_{\cdot j}})^2\right] \\
&= \tfrac{1}{a(n-1)} E\left[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a}(y_{ij}^2 - 2y_{ij}\bar{y}_{\cdot j} + \bar{y_{\cdot j}}^2)\right] \\
&= \tfrac{1}{a(n-1)} E\left[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a} y_{ij}^2 - 2\sum_{i=0}^{n}\sum_{j=0}^{a}\bar{y}_{\cdot j}y_{ij} + \sum_{i=0}^{n}\sum_{j=0}^{a}\bar{y_{\cdot j}}^2\right] \\
&= \tfrac{1}{a(n-1)} E\left[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a} y_{ij}^2 - 2n\sum_{j=0}^{a}\bar{y}_{\cdot j}^2 + n\sum_{j=0}^{a}\bar{y_{\cdot j}}^2\right] \\
&= \tfrac{1}{a(n-1)} E\left[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a} y_{ij}^2 - n\sum_{j=0}^{a}\bar{y_{\cdot j}}^2\right] \\
&= \tfrac{1}{a(n-1)} E\left[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a} y_{ij}^2 - n\frac{1}{n^2}\sum_{j=0}^{a}\left(\sum_{i=1}^{n} y_{ij}\right)^2\right] \\
&= \tfrac{1}{a(n-1)} E\left[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a}(\mu + \tau_j + \epsilon_{ij})^2 - \frac{1}{n}\sum_{j=0}^{a}\left(\sum_{i=1}^{n}\{\mu + \tau_j + \epsilon_{ij}\}\right)^2\right] \\
&= \tfrac{1}{a(n-1)} E\Big[\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a}\mu^2 + \sum_{i=0}^{n}\sum_{j=0}^{a}\tau_j^2 + \sum_{i=0}^{n}\sum_{j=0}^{a}\epsilon_{ij}^2 \\
&\quad \cancel{+2\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a}\mu\tau_j} + \cancel{2\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a}\mu\epsilon_{ij}} + \cancel{2\textstyle\sum_{i=0}^{n}\sum_{j=0}^{a}\tau_j\epsilon_{ij}} \\
&\quad -\tfrac{1}{n}\textstyle\sum_{j=1}^{a}\left(\{\sum_{i=1}^{n}\mu\}^2 + \{\sum_{i=1}^{n}\tau_j\}^2 + \{\sum_{i=1}^{n}\epsilon_{ij}\}^2\right)\Big] \\
&= \tfrac{1}{a(n-1)} E\left[an\mu^2 + n\textstyle\sum_{j=1}^{a}\tau_j^2 + an\sigma^2 - \frac{1}{n}\left(\sum_{j=1}^{a}\{n\mu\}^2 + \sum_{j=1}^{a}\{n\tau_j\}^2 + \sum_{j=1}^{a}\{\epsilon_{\cdot j}\}^2\right)\right] \\
&= \tfrac{1}{a(n-1)} E\left[\cancel{an\mu^2} + \cancel{n\textstyle\sum_{j=1}^{a}\tau_j^2} + an\sigma^2 - \cancel{\tfrac{1}{n}an^2\mu^2} - \cancel{\tfrac{1}{n}n^2\textstyle\sum_{j=1}^{a}\tau_j^2} - \cancel{\tfrac{1}{n}an\sigma^2}\right] \\
&= \cancel{\tfrac{1}{a(n-1)}}\cancel{a(n-1)}\sigma^2 \\
&= \sigma^2
\end{aligned}$$

$$
\begin{aligned}
E(MS_{Cond}) &= E(\tfrac{SS_{Cond}}{a-1}) \\
&= \tfrac{n}{a-1} E\left[\textstyle\sum_{j=0}^{a} (\bar{y_{\cdot j}} - \bar{y_{\cdot\cdot}})^2\right] \\
&= \tfrac{n}{a-1} E\left[\textstyle\sum_{j=0}^{a} (\bar{y_{\cdot j}}^2 - 2\bar{y}_{\cdot j}\bar{y}_{\cdot\cdot} + \bar{y_{\cdot\cdot}}^2)\right] \\
&= \tfrac{n}{a-1} E\left[\textstyle\sum_{j=0}^{a} \bar{y_{\cdot j}}^2 - 2\sum_{j=1}^{a} \bar{y}_{\cdot j}\bar{y}_{\cdot\cdot} + \sum_{j=1}^{a} \bar{y_{\cdot\cdot}}^2\right] \\
&= \tfrac{n}{a-1} E\left[\textstyle\sum_{j=0}^{a} \bar{y_{\cdot j}}^2 - 2a\bar{y}_{\cdot\cdot}^2 + a\bar{y}_{\cdot\cdot}^2\right] \\
&= \tfrac{n}{a-1} E\left[\textstyle\sum_{j=0}^{a} \bar{y_{\cdot j}}^2 - a\bar{y}_{\cdot\cdot}^2\right] \\
&= \tfrac{n}{a-1} E\left[\tfrac{1}{n^2}\textstyle\sum_{j=0}^{a} \left(\sum_{i=1}^{n} y_{ij}\right)^2 - \tfrac{1}{an^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{a} y_{ij}\right)^2\right] \\
&= \tfrac{1}{n(a-1)} E\left[\textstyle\sum_{j=0}^{a} \left(\sum_{i=1}^{n} y_{ij}\right)^2\right] - \tfrac{1}{an(a-1)} E\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{a} y_{ij}\right)^2\right] \\
&= \tfrac{1}{n(a-1)} E\left[\textstyle\sum_{j=0}^{a} \left(\sum_{i=1}^{n} (\mu + \tau_j + \epsilon_{ij})\right)^2\right] - \tfrac{1}{an(a-1)} E\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{a} (\mu + \tau_j + \epsilon_{ij})\right)^2\right] \\
&= \tfrac{1}{n(a-1)} E\left[\textstyle\sum_{j=0}^{a} \left(\sum_{i=1}^{n} \mu\right)^2 + \sum_{j=1}^{a} \left(\sum_{i=1}^{n} \tau_j\right)^2 + \sum_{j=1}^{a} \left(\sum_{i=1}^{n} \epsilon_{ij}\right)^2\right] \\
&\quad - \tfrac{1}{an(a-1)} E\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{a} \mu\right)^2 + \left(\sum_{i=1}^{n}\sum_{j=1}^{a} \tau_j\right)^2 + \left(\sum_{i=1}^{n}\sum_{j=1}^{a} \epsilon_{ij}\right)^2\right] \\
&= \tfrac{1}{n(a-1)} E\left[an^2\mu^2 + n^2\textstyle\sum_{j=1}^{a} \tau_j^2 + an\sigma^2\right] \\
&\quad - \tfrac{1}{an(a-1)} E\left[a^2n^2\mu^2 + \cancel{n^2\left(\textstyle\sum_{j=1}^{a} \tau_j^2\right)^2} + an\sigma^2\right] \\
&= \cancel{\tfrac{an^2}{n(a-1)}\mu^2} + \tfrac{n^2}{n(a-1)}\textstyle\sum_{j=1}^{a} \tau_j^2 + \cancel{\tfrac{na}{n(a-1)}}\sigma^2 - \cancel{\tfrac{a^2n^2}{an(a-1)}\mu^2} - \cancel{\tfrac{an}{an(a-1)}}\sigma^2 \\
&= \sigma^2 + \tfrac{n}{a-1}\textstyle\sum_{j=1}^{a} \tau_j^2
\end{aligned}
$$

12. Consider the Instagram factorial experiment from class, in which the influence of two factors on engagement time was investigated. Use R to answer the following questions. The data are available in the `instagram-factorial.txt` file.

    (a) Construct a main effects plot for the factor "prevalence" and explain whether and why you do/do not believe it is a significant factor. Comment on which level is most detrimental to the user experience.

    (b) Construct a main effects plot for the factor "type" and explain whether and why you do/do not believe it is a significant factor. Comment on which level is most detrimental to the user experience.

    (c) Using the factors "prevalence" and "type" as an example, explain what is meant by interaction. Construct an interaction plot for these factors explain whether and why you do/do not believe this effect is significant.

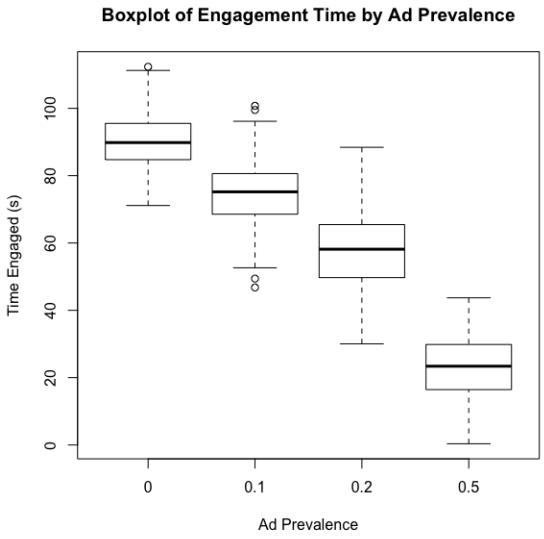
    (d) Fit the full model with all main effects and interaction terms and formally evaluate the statistical significance of each effect. What do you conclude? Do these formal conclusions agree with the informal ones drawn in parts (a)-(c)?

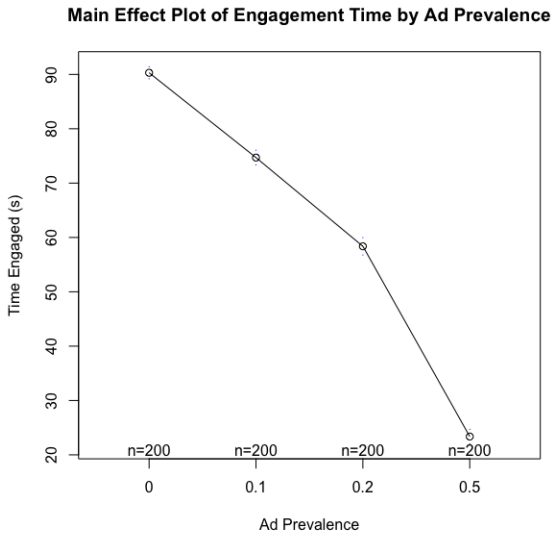
    (e) Based on this analysis, which combination of factor levels appears to maximize engagement? Discuss the practical implications of this choice.

    (f) Assess the residuals of the model in part (d). Do the typical ordinary least squares assumptions appear to be met?

Solution: R solution available in the file `instagram-factorial.R`.

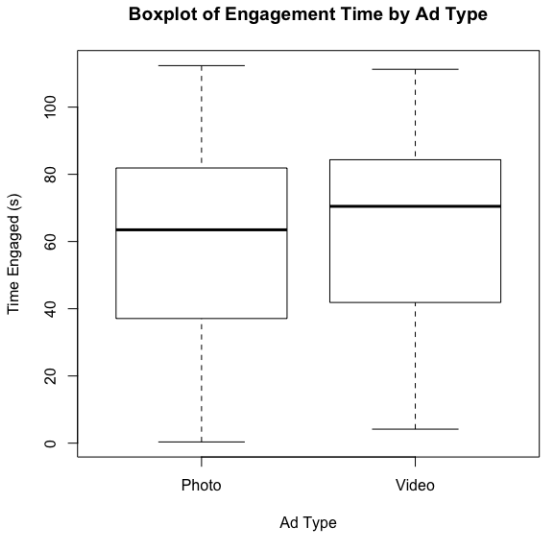
(a)

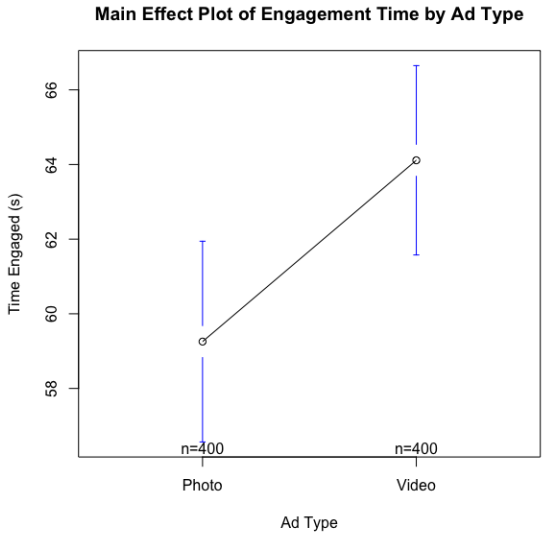




Looking at the main effects plot for the factor "prevalence", I am led to believe that it is a significant factor, since the time engaged decreases apparently significantly as the ad prevalence increases. A prevalence level of 0.5 is the most detrimental within the levels we have available, with an average engagement of just above 20 seconds (compared to around 90 seconds when ad prevalence is 0).

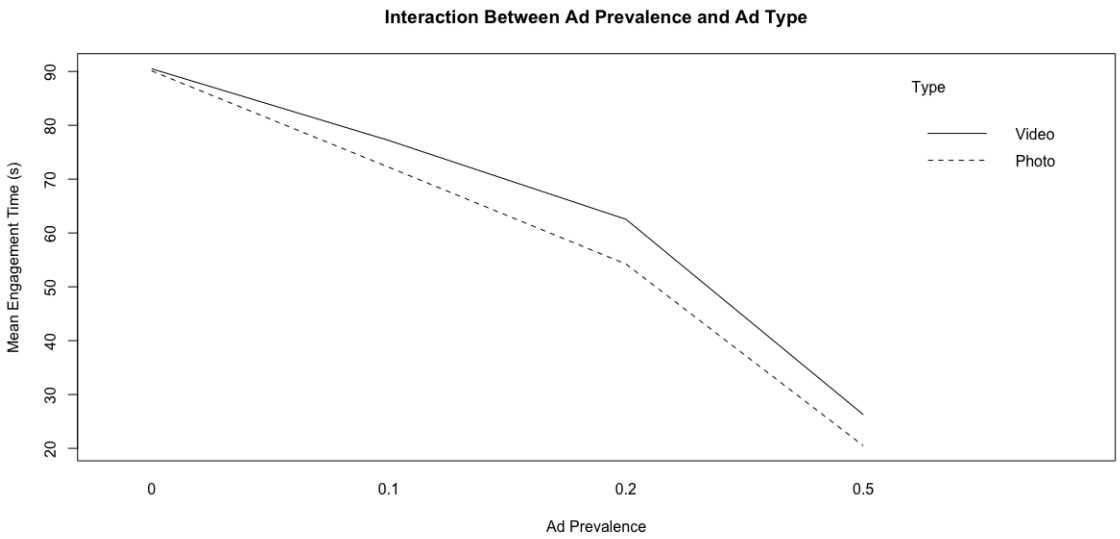
(b)





The main effects plot for the factor "type" shows that average engagement time is around 59 seconds when the ad type is "Photo", while it is around 64 seconds when the ad type is "Video". So the type "Photo" is more detrimental to the user experience. It is difficult to say only by looking that this plot whether the effect is statistically significant.

(c)

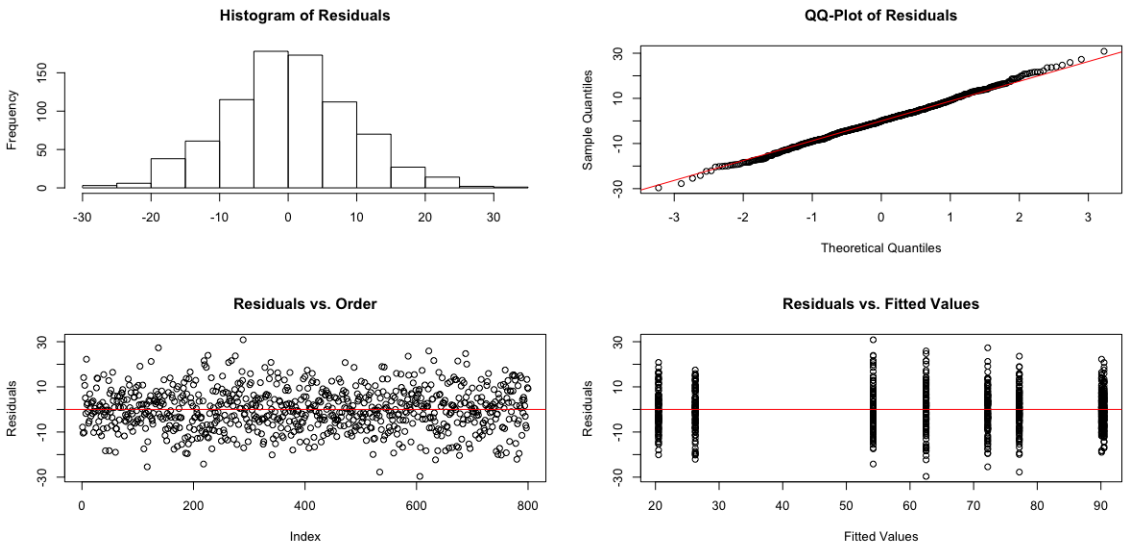


If the engagement time is not the same at all levels of other factors, we say there is an interaction present. Here, in order to have interaction, we would want to see that the response (engagement time) varies at different rates for Photo and Video ads as we increase (or decrease) the ad prevalence. In the interaction plot, this would be represented by intersecting (within or outside the plot boundaries) lines. Looking at the interaction plot between these two factors, it seems like there is interaction between "prevalence" and "type", but it is minimal. The main effects seem to drive the variation in the response (engagement time).

(d) We fit the full model by using the formula `lm(Time ~ Prevalence * Type)`. Looking at the `anova`, we can see that our intuition was correct: both main effects are significant, with "prevalence" explaining most of the variation in engagement, and the interaction, while minimal, is significant.

(e) Looking at the `summary` of this model, we can conclude that the combination of factor levels that maximizes engagement are a prevalence of 0 and an ad type set to "Video". This choice is certainly not wise, since a prevalence of 0 means that no ads will show (and hence even the choice of an ad type for this factor level does not make sense).

(f)



The residuals seem to be normally distributed, independent, and homoskedastic. There are no outliers, and the residuals vs fitted plot doesn't show any non-linearity in the residuals. Hence, the assumptions of ordinary least squares are met.

13. In the context of a 2^k factorial design, we typically model factor levels in coded units (± 1) rather than natural units. A model for a single replicate of a 2^2 factorial design can be written as follows:

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

where $\mathbf{y} = (y_1, y_2, y_3, y_4)^T$, $\beta = (\beta_0, \beta_1, \beta_2, \beta_{12})^T$, $\epsilon = (\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_{12})^T$, $\epsilon_i \sim N(0, \sigma^2)$ and

$$X = \begin{bmatrix} \mathbf{1} & \mathbf{x_1} & \mathbf{x_2} & \mathbf{x_1 \cdot x_2} \end{bmatrix} = \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \end{bmatrix}$$

One consequence of using these coded units is that the columns of the design matrix in such models are orthogonal. There are several convenient by-products of this.

(a) Show that the least squares estimates $\hat{\beta}_0, \hat{\beta}_1,$ and $\hat{\beta}_2$ are unaffected by the presence/absence of the interaction term.

(b) Show that the standard errors of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2,$ and $\hat{\beta}_{12}$ are all the same.

(c) Explain why $\hat{\beta}_1, \hat{\beta}_2,$ and $\hat{\beta}_{12}$ are each interpreted as one-half of the corresponding factor's effect estimate.

Solution: (a) With interaction term, we have

$$\mathbf{X^T X} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \text{ and therefore } (\mathbf{X^T X})^{-1} = \begin{bmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}.$$

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

Therefore, we get $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_{12} \end{bmatrix} = \begin{bmatrix} \frac{y_1+y_2+y_2+y_4}{4} \\ \frac{-y_1+y_2-y_3+y_4}{4} \\ \frac{-y_1-y_2+y_3+y_4}{4} \\ \frac{y_1-y_2-y_3+y_4}{4} \end{bmatrix}$.

Without interaction term, we have

$\mathbf{X^T X} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$, and therefore $(\mathbf{X^T X})^{-1} = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}$.

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

Therefore, we get $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \frac{y_1+y_2+y_2+y_4}{4} \\ \frac{-y_1+y_2-y_3+y_4}{4} \\ \frac{-y_1-y_2+y_3+y_4}{4} \end{bmatrix}$.

We can see that the estimates $\hat{\beta}_0, \hat{\beta}_1,$ and $\hat{\beta}_2$ are unaffected by the presence/absence of the interaction term.

(b) We have $\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}$.

Here, $\mathbf{X^T X} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$, and therefore $(\mathbf{X^T X})^{-1} = \begin{bmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}$.

We know that the variance of any model regression coefficient is
$V(\hat{\beta}) = \sigma^2 \times$ (corresponding diagonal element of $(\mathbf{X^T X})^{-1}$).

Therefore, we get $V(\hat{\beta}) = \frac{\sigma^2}{4}$ and all model regression coefficients have the same variance.

(c) $\hat{\beta}_1, \hat{\beta}_2,$ and $\hat{\beta}_{12}$ are each interpreted as one-half of the corresponding factor's effect estimate because of the way we code our factor levels (-1 and 1). The regression coefficients measure the effect of a one-unit change of the variable on the response. However, in case of a factorial design, we want the effect of a two-unit change of the variable on the response (i.e., from -1 to 1).

14. Fractional factorial designs are an efficient approach to investigating the effects of a large number of factors.

(a) Briefly explain the philosophy and mechanics of a 2^{k-p} fractional factorial design.

(b) Briefly explain *aliasing* and *resolution*, and why a resolution-IV design is to be preferred over a resolution-III design.

Solution: (a) The number of conditions in a full 2^k factorial design can become very large very quickly. In addition, the principle of effect sparsity suggests that 3 and higher order interactions are unlikely to be significant. Therefore, by using a full factorial design, we are effectively "wasting" degrees of freedom to estimate these factors. 2^{k-p} fractional factorial designs allow us to investigate

the same number of factors in fewer conditions or investigate a larger number of factors in the same number of conditions. We effectively investigate k factors in $(\frac{1}{2})^p 2^k$ conditions. In order to do this, we pick p design generators, which means we choose an interaction column to alias each of the p additional factors.

(b) When we alias a main effect with an interaction, we are given a prescription dictating when to run the effect at its low and high levels. By doing this, the main effect and the interaction are confounded and we can't separately identify their individual effects.

A design is of resolution R if main effects are aliased with interaction effects involving at least $R-1$ factors. With higher order resolution, main effects are aliased with higher order interactions, which are less likely to be significant. This means that main effects would not be confounded with more significant two-level interactions.