Assignment 2 – Experiments in Marketing

The deadline for this assignment is 23.59 on Sunday November 22nd.

For all problems, use a 5% threshold for statistical significance.

Problem 1 (4 points)

Aalto University recently ran a series of A/B tests with the goal of streamlining the application process for prospective students. Since the desired outcome of a prospective student's decision-making journey is application to Aalto's programs, conversion cannot be measured in monetary terms. Moreover, because the application process is long, Aalto University decided to optimize the website against one of the largest 'leaks' in the conversion funnel: in 2016-2017, from the ca. 160,000 users that visited the study options pages of Aalto University's master's programs, only 40 percent moved in the funnel to the "How to Apply" pages. Building on this insight, the goal of the experiments was to see if changes in the website design could increase the likelihood that a user visits the "How to Apply" page within that browsing session. If that happens, the user is considered converted.

The total number of visitors and conversion rates (i.e., percentage of sessions where the user visits the "How to Apply" page) by experimental condition are reported in Table 1. The data is real experimental data pulled from Google Analytics.

Table 1. Results of the A/B tests

	Number of experiment sessions (conversion rate)	
Experiment	Original version	Test version
Experiment 1	2334 (10.88 %)	2065 (9.69 %)
Experiment 2	9500 (12.21 %)	13,623 (15.09 %)
Experiment 3	3087 (14.48 %)	3111 (12.86 %)
Experiment 4	9384 (8.87 %)	9061 (9.50 %)

Analyze the four tests using the Chi-Square test. Which of the tested design changes (Experiments 1-4) should be implemented? Hint: See the script from Lecture 3 for help with building a contingency table to analyze with the chisq.test() function.

To perform the Chi-Square test, we need to first create four contingency tables from the data. The number of conversions is conversion rate*total number of visitors. The rest of the visitors are non-conversions. Once we have calculated the number of conversions and non-conversions per original/test version, we can construct the contingency tables. The easiest way to do this is using either the 'table()' or 'matrix()' functions as in the Introduction to R course, but you can also do this the same way as we did in Lecture 3 by creating vectors and combining them into a table. Once you have a contingency table, run the 'chisq.test()' function to perform the Chi-square test.

For example, in the case of Expriment 1:

```
> click_data <- matrix(c(round(2334-2334*.1088),round(2065-
2065*.0969),round(2334*.1088),round(2065*.0969)),ncol=2,byrow=TRUE)
> colnames(click_data) <-c("Original","Test");
> rownames(click_data) <-c("No click","Click")
> click_data

         Original Test
No click     2080 1865
Click         254  200

> chisq.experiment1<-chisq.test(click_data)
> chisq.experiment1

      Pearson's Chi-squared test with Yates' continuity correction

data:  click_data
X-squared = 1.5703, df = 1, p-value = 0.2102
```

The p-value is above 0.05, so the difference between the original and test version is not statistically significant.

We repeat the same procedure for Experiments 2-4. Out of these, only Experiment 2 yields a statistically significant difference, so we implement that design change and disregard the rest.

Problem 2 (3 points)

An e-commerce company has performed an A/B test to improve the conversion rate of its website. Download the dataset concerning the results of the experiment ('conversion_experiment.csv'). In the experiment, users were randomly shown either the current version of the landing page (test_version = 0), or a newly designed landing page (test_version = 1). The outcome of interest is whether the user makes a purchase (conversion = 1) or not (conversion = 0) during the session.

The first few rows are as look as follows:

```
> head(experiment_data) # this function returns the first few lines of the data
   sessionid geography test_version conversion
1 WCJVZ5046Y        US            0          0
2 MCGDF4182W        US            0          0
3 UXNZH0193Q        US            1          1
4 IQNNJ1719B        US            0          0
5 RDAAM6037N        US            0          0
6 WGUKL6680R        US            0          0
```

Problem 2A. Run a chi-square test to analyze whether the new landing page is associated with a higher conversion rate. Report (i) a contingency table of observed frequencies, (ii) conversion rates for the current and the new version of the landing page, (iii) a table of expected frequencies, and (iv) the results of a Chi-Square (χ^2) test. Hint: before beginning, you may wish to familiarize yourself with the chisq.test() function by running '?chisq.test' in R.

Contingency table:

You can construct a contingency table manually as in the previous example. However, some of you have noticed that you can also import the 'conversion' and 'test_version' variables directly into the 'chisq.test()' function as vectors:

```
> test_results<-
chisq.test(experiment_data$conversion,experiment_data$test_version)
> test_results$observed
```

```
                          experiment_data$test_version
experiment_data$conversion    0    1
                         0 1120 1117
                         1  107  156
```

Conversion rates:

```
> conversion_original <-
test_results$observed[2,1]/colSums(test_results$observed)[1]
> conversion_test <-
test_results$observed[2,2]/colSums(test_results$observed)[2]
> conversion_original
> conversion_test
```

Original ≈ 8.7 %
Test version ≈ 12.3 %

It appears that the test version performs better than the original. However, we need to perform a statistical test to discern how likely this is just a product of coincidence.

Expected frequencies:

```
> test_results$expected
```

```
                          experiment_data$test_version
experiment_data$conversion         0         1
                         0 1097.9196 1139.0804
                         1  129.0804  133.9196
```

This table shows that if the conversion rate was independent of the experimental condition, we should expect about 129 conversions for the original version and about 134 conversions for the new version. These are clearly different from the observed frequencies.

The formal Chi-square test is produced as follows:

```
> test_results
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  experiment_data$conversion and experiment_data$test_version
X-squared = 7.9185, df = 1, p-value = 0.004893
```

The p-value 0.004893 < 1 %, leading us to conclude that there is a significant association between the experimental condition and the conversion rate.

Problem 2B. Randomization is an important principle of successful experiment design. Your users come from United States ('US') and the rest of the world ('Non-US'). If randomization is successful, the probability that users see the current vs. the new version of the landing page should not depend on the geographical location of the user. If randomization is unsuccessful, inferences about the effect of the manipulation (i.e., variable test_version) can become biased. Should we worry about this?

We can test the randomization using the chi-square test. The test can be performed similarly as the test above.

```
> randomization_test <-
chisq.test(experiment_data$geography,experiment_data$test_version)
> randomization_test

        Pearson's Chi-squared test with Yates' continuity correction

data:  experiment_data$geography and experiment_data$test_version
X-squared = 0.0048392, df = 1, p-value = 0.9445
```

The test yields a non-significant p-value (p = 0.9445). Therefore, we do not need to worry about confounding due to unsuccessful randomization.

Problem 3 (3 points)

Problem 3A: An entertainment streaming service has created a new personalization algorithm to help users find relevant content on their platform. Before rolling out the algorithm for all users, the company has first performed an A/B test to see if the algorithm actually improves user experience. The outcome of interest is weekly streaming hours. Download the dataset containing the experimental results ('personalization_experiment.csv'), and perform a t-test to analyze the experimental results. Use 1 % as the threshold of statistical significance.

The script for the t-test is as follows:

```
> test_results <-
t.test(streaming_hours~algorithm,data=personalization_experiment)
> test_results

        Welch Two Sample t-test

data:  streaming_hours by algorithm
t = -8.58, df = 2498, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.2502525 -0.7850856
sample estimates:
mean in group new mean in group old
         9.056679         10.074348
```

The p-value indicates that there is a statistically significant different in the weekly streaming hours, depending on whether the new algorithm is used or not. However, the new algorithm produces a significantly *lower* level of engagement with the service (ca. 9 hours compared to about 10 hours for the old version). Therefore, the new algorithm should not be rolled out!

Problem 3B: Run an ANOVA to test is the effect of the new algorithm is moderated by whether users are classified as predominantly mobile or desktop users. Hint: to help you interpret your results visually, you may recycle the script from Lecture 3, substituting relevant variable names into the ggplot() function where appropriate.

We run an ANOVA to test for moderation, including an interaction term between algorithm and device type:

```
> aov.test<-
aov(streaming_hours~as.factor(algorithm)+as.factor(device)+as.factor(algori
thm)*as.factor(device), data=personalization_experiment)
>tidy(aov.test)
```

```
  term                                    df    sumsq meansq statistic   p.value
  <chr>                                <dbl>    <dbl>  <dbl>     <dbl>     <dbl>
1 as.factor(algorithm)                     1   647.   647.      76.3    4.31e-18
2 as.factor(device)                        1   834.   834.      98.4    8.81e-23
3 as.factor(algorithm):as.factor(device)   1     1.87   1.87     0.220  6.39e- 1
4 Residuals                             2496 21157.     8.48    NA      NA
```

The results indicate the both algorithm and device (lines 1 and 2) have an effect on streaming hours (p-value is well below 0.05). However, the interaction term (line 3) is not statistically significant (p-value 0.693 > 0.05). Therefore, we can conclude that the effect of the algorithm on streaming hours does not depend on (i.e., is not moderated by) device type.