**MSDS 6371 Fall 2018 Midterm**

**Analysis Questions**

**Browse Time**



A retail company wants to analyze the length of time that consumers spend on their main webpage. The company has gathered the time (in seconds) that all customers during a one hour period spent on the page. The marketers at the retailer wanted to see if the position (left or right) of a promotional banner affected how long browsers spent on the page. So, they designed an experiment where half of the browsers randomly saw the banner positioned on the left and half randomly saw the banner positioned on the right.

* **Use significance level  unless otherwise stated.**
* **Include your code and relevant output at the end of each response (or within, if appropriate). You may use SAS or R in your analyses.**
* **Please save your work in a Word document.**

1. (20 points) Test the claim that the browsing time is significantly different for the left and right banner positions.

The data are recorded in BannerPositionOnly.csv.

Please provide a **complete analysis** for thequestion of interest. Remember that a **complete analysis** includes doing the following:

* 1. state the problem
  2. **address the assumptions** and clearly state which test you feel is most appropriate
  3. conduct the appropriate test, including 6 steps (you do not need to provide a shaded graph, and the last step should include the conclusion and provide a scope of inference). Be sure to include confidence intervals where appropriate.

Problem statement: We are testing the claim that the browsing times of customers who see the left banner are different from those who see the right banner.

Assumptions: A two-sample t-test would be great to use, but we must test assumptions first:

Normally distributed sample means:

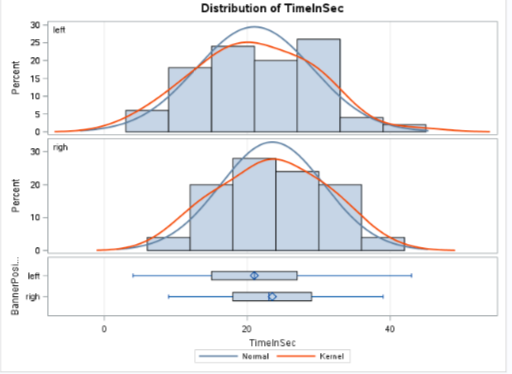
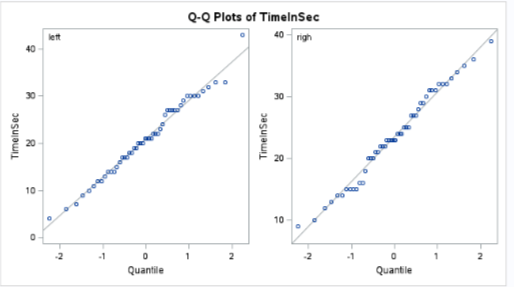
In SAS, most results can be obtained from the following:

proc ttest data = banneronly;

class BannerPosition;

var TimeInSec;

run;



Both the q-q plots and histograms show little evidence against normally distributed data, which means the sample means should be normally distributed.

Equal standard deviations: The histograms and box plots show little evidence against equal standard deviations, and the equal sample sizes should overcome any slight differences.

Independence between and within groups. Based on how the data was chosen (a census in a randomized experiment), we will proceed with the assumption of independence both between and within group.

**Two sample t-test:**

Step 1 - Hypotheses:

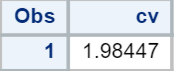
Step 2 - Identification of Critical Value:

At alpha = 0.05,

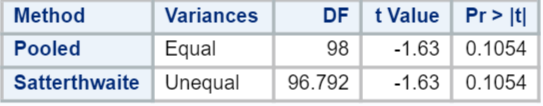
data mycritval;

cv=quantile("t", 0.975, 100-2);

run;



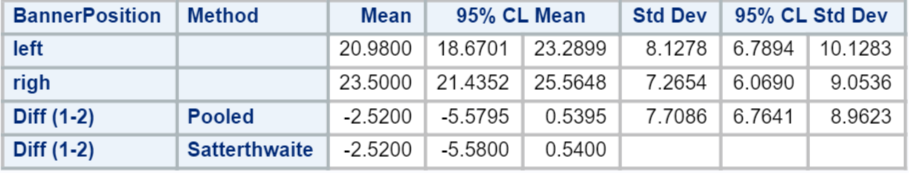
Step 3 - Value of Test Statistic:



Step 4 - Give p-value:

Step 5 - Decision: Fail to Reject at significance level alpha = 0.05.

Step 6 - Conclusion: On the basis of this test, there is not enough evidence to suggest that the mean browsing times of customers who saw the left and right banner positions are different ( from a two-sample t-test). A 95% confidence interval for this difference is seconds, in favor of the right bannered group. Since the subjects in this sample constituted a census of the customers over one hour, these results cannot be generalized to all customers. Even though this was a randomized experiment, the results are not significant, so there is no difference of mean browse times to discuss here.



1. (20 points) The data scientists then notice that those who saw the left banner were directed to the site from two different web pages that they visited just before this retailer (prior pages 1 & 2), while those who saw the right banner were directed to the site from two entirely different web pages (prior pages 3 & 4). How odd! Now the data scientists are interested in seeing if the prior web pages (and only the prior web pages) had any effect on the browsing time of the consumers. Perform an analysis that answers this question. **After listing which assumptions might be required, you may assume those assumptions are met.** (You do not need to provide evidence supporting these assumptions.) This data is found in Browsetime.csv. This data set contains the same data as BannerPositionOnly.csv along with the additional variable of PriorPage.

Problem statement: We would like to test the claim that any mean browsing times are different across prior webpages.

Assumptions: We would like to perform pure ANOVA, so we should check for normally distributed sample means, equal standard deviations, and independence between and within groups.

Hypothesis Test:

**Step 1 - Hypotheses:**

**All mean browsing times are the same across prior web pages.**  
 **At least one pair of mean times are different between prior web pages.**

**Step 2 - Identification of Critical Value: You may skip step 2 (critical value) in ANOVA settings, although one could be found (and the comparison to the F statistic should match the p-value’s comparison to alpha).**

**Step 3 - Value of Test Statistic:**

The following SAS code answers parts 2, 3, and the bonus.

proc sort data = browsetime;

by PriorPage;

run;

proc glm data = browsetime;

class PriorPage;

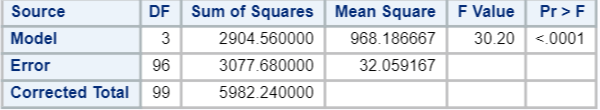
model TimeInSec = PriorPage;

means Priorpage/ HOVTest = BF bon cldiff;

lsmeans PriorPage/ pdiff tdiff adjust = bon cl;

estimate 'average PriorPage 1 & 2 versus 3 & 4' PriorPage 1 1 -1 -1 /divisor = 2;

run;



**Step 4 - Give p-value:**

**Step 5 - Decision: Reject**

**Step 6 - Conclusion: There is strong evidence to suggest that at least one of the mean browsing times for a particular prior web page is different from the others ( from a pure ANOVA). Because the prior web pages are not randomly assigned, we cannot infer that the prior web page caused the difference.** Since the subjects in this sample constituted a census of the customers over one hour, these results cannot be generalized to all customers.

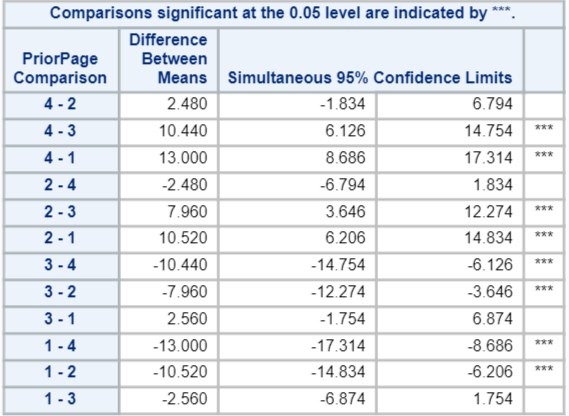
1. (20 points) The data scientists are now looking to see which specific prior sites differ in browsing time. **After listing which assumptions might be required, you may assume those assumptions are met.** (You do not need to provide evidence supporting these assumptions.) Conduct an analysis using Browsetime.csv that answers this question. Provide confidence intervals where appropriate.

Problem statement: We would like to test the claim that for each pairing of prior web pages, the mean browsing times are different.

Hypotheses: For i, j, where i is not equal to j:

Assumptions: We would like to perform multiple comparisons, which typically have assumptions similar to pure ANOVA; so, we should check for normally distributed sample means, equal standard deviations, and independence between and within groups.

All pairwise comparisons are significantly different except groups 2 and 4 and groups 1 and 3 according to Bonferroni adjusted t-tests. Confidence intervals for the differences in the means are shown below.



P-values could also have been provided. Bonferroni is not the only possible choice of multiple comparison tests.

The scope of inference is identical to question 2.

1. (Bonus 10 points, up to 100 maximum exam points) Reconsider the question of interest in problem 1. Considering that more information and data were provided in questions 2 and 3, would you go about answering the question of interest in question 1 the same way? If so, explain why the additional data is irrelevant. If not, state and perform what you would do instead.

A contrast to test the following hypothesis would be even better, as it reduces the variance, which increases the power of the test.

Problem statement: Is the average of the means of the first and second group different from the average of the means of the third and fourth group.

Assumptions: Assumptions of pure ANOVA should be checked, such as normally distributed sample means, equal standard deviations across groups, and independence both between and within groups.

Hypothesis Test:

Step 1 - Hypotheses:

Step 2 - Identification of Critical Value: Can skip if looking at F-statistic.

Critical t-value =

data thecritval;

critv=quantile("t", 0.975, 100-4);

run;



Step 3 - Value of Test Statistic:



Step 4 - Give p-value:

Step 5 - Decision: Reject at significance level alpha = 0.05.

Step 6 - Conclusion: On the basis of this test, there IS sufficient evidence to suggest that the average of the mean browsing times of customers who came from pages 1 and 2 (left banners), respectively, are different ( from a contrast) from the average of the means who came from pages 3 and 4 (right banners). A 95% confidence interval for this contrast is seconds in favor of the mean of groups 3 and 4, the right banners. Since the subjects in this sample constituted a census of the customers over one hour, these results cannot be generalized to all customers. Here, because there is an odd alignment of randomly assigned banner positions to those coming from two distinct sets of web pages, we should be careful about inferring causality here.

\*95% Confidence interval for the contrast :

Notice that this 95% confidence interval has the same point estimate as the confidence interval in question 1 (95% CI for : [-0.54, 5.58]), but the width is much narrower here. This is nearly entirely due to being able to reduce the estimate of the pure variance by subdividing the data into more groups. (The reduced degrees of freedom of the contrast make the confidence interval slightly wider, but negligibly so, and not nearly enough to counteract the reduced variance.)

**To upload your file : Please make sure this is a Word doc**

**1. Select Add**

**2. Select Upload a file**

**3. Select Choose file**

**4. Locate file you want to upload**