

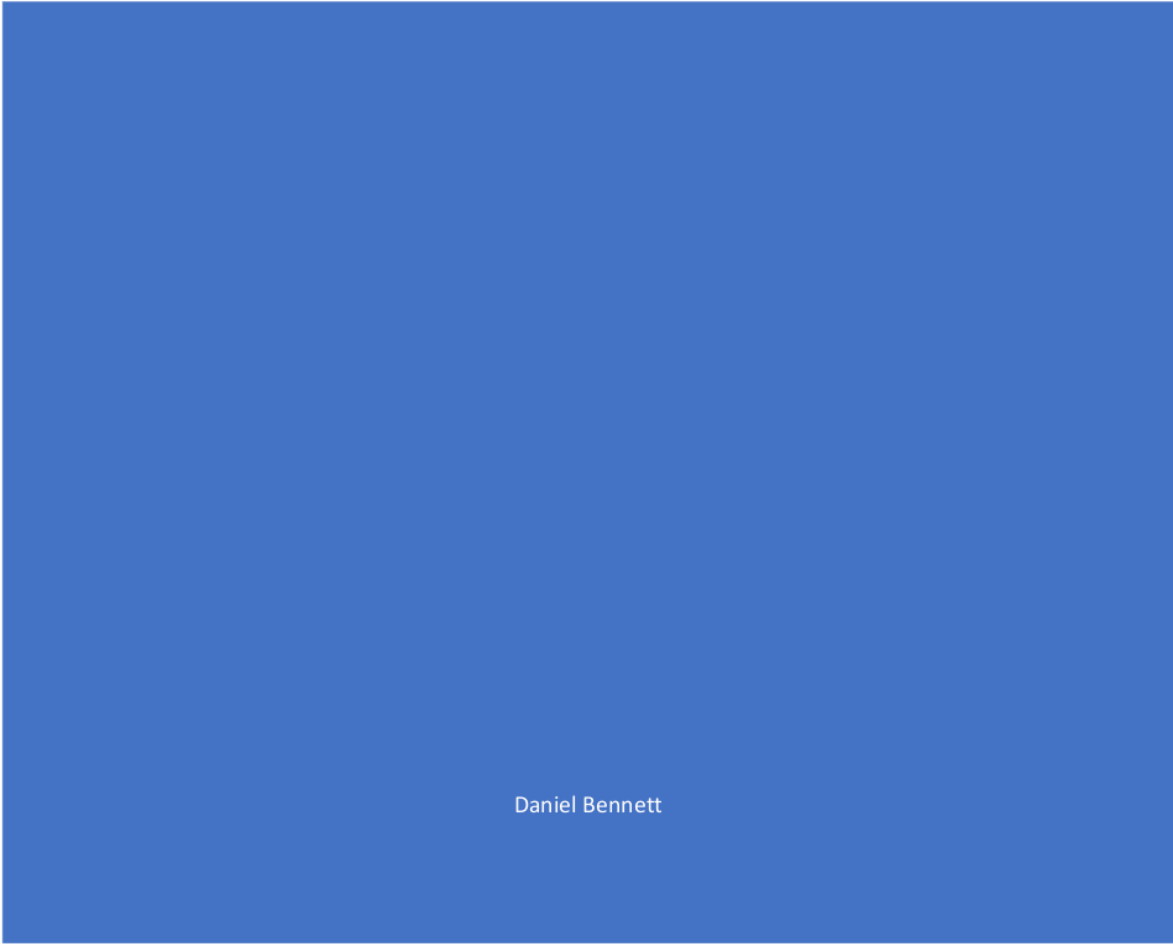
Data Analysis and Visulaisation - Assingment 2

by Daniel Bennett

FILE	14023_DANIEL_BENNETT_DATA_ANALYSIS_AND_VISULAISATION_- _ASSINGMENT_2_192216_1975674232.DOCX (120.42K)		
TIME SUBMITTED	29-MAY-2020 11:51PM (UTC+0100)	WORD COUNT	1789
SUBMISSION ID	129591503	CHARACTER COUNT	8177



DATA ANALYSIS AND VISUALISATION - ASSIGNMENT 2



Daniel Bennett

Contents

Introduction	2
Abstract	2
Initial Data	2
Hypotheses	4
Identifying Suitable Tests	4
Testing	4
Test 1 (0.10 significance level):	5
Test 2 (0.05 significance level):	5
Test 3 (0.01 significance level):	6
Analysis	6
Considering Outliers	6
Test 1 (0.10 significance level):	7
Test 2 (0.05 significance level):	7
Test 3 (0.01 significance level):	8
Analysis	8
Conclusion	8
References	9

Introduction

This document will be an analysis of the data provided by an online retail company regarding sales. The data has been obtained via A/B testing, each user that accesses the company's website was either presented their old design or new design. It was then tracked if any sales were made. For 40 days, this data was tallied and presented to the author, to conduct an analysis and find which version was better at producing sales.

Abstract

There are many ways to determine if sales were better. The objective of this document was to take the data provided and show which website had better sales through sophisticated tests.

In summary, to use the data provided to find which site had better sales.

To achieve this a thorough analysis of the data was conducted. To determine the results of the analysis and hypothesis and null hypothesis was used. The data was also analysed and had a T-Test conducted on it at various significance levels.

This document proves that the new website is a better website and the sales generated are statistically significant.

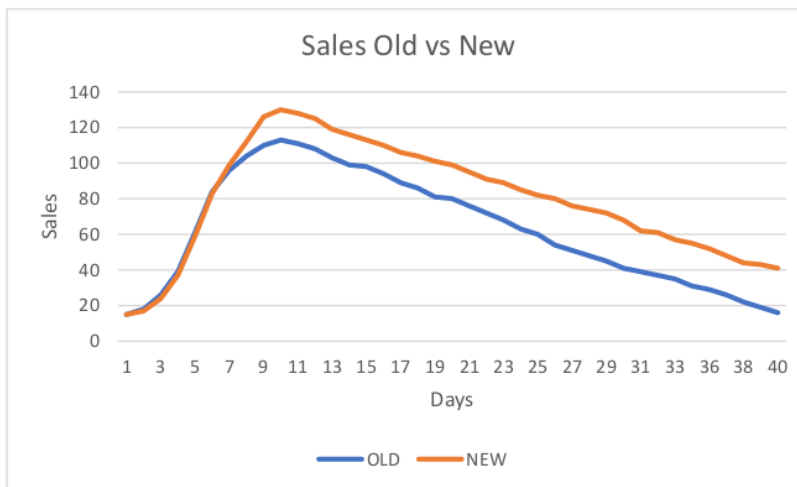
Initial Data

The initial data provided was in the form of a table. This allowed the data to be easily inspected for any obvious trends and data. The data was complete and showed the sales for every of the 40 days on both the old and the new website. Below the base data in a table:

DAY	OLD	NEW	DAY	OLD	NEW
1	15	15	21	76	95
2	18	17	22	72	91
3	26	24	23	68	89
4	39	37	24	63	85
5	61	59	25	60	82
6	84	83	26	54	80
7	96	99	27	51	76
8	104	112	28	48	74
9	110	126	29	45	72
10	113	130	30	41	68
11	111	128	31	39	62
12	108	125	32	37	61
13	103	119	33	35	57
14	99	116	34	31	55
15	98	113	35	76	103
16	94	110	36	29	52
17	89	106	37	26	48
18	86	104	38	22	44

19	81	101	39	19	43
20	80	99	40	16	41

This data was collected through the use of A/B testing over 40 days. With the old column meaning the old website and the new column meaning the new website. From initial observation a simple trend can be seen of numbers rising quite high to begin with a then diminishing over time. However, to view this in a better format, a simple line graph is good.



Label

This graph helps show this trend very well. Both the old and the new sites gain more sales from day 1 peaking at day 10. They then start to diminish at a similar rate every day with slightly less sales every day from the old site. This graph also helps show an outlier on day 36, this will be discussed in more detail later.

Another way for us to look at this data is by using a very useful feature of Microsoft Excel called Data Analysis, we can get what's called "Descriptive Statistics". Below is the descriptive statistics of the old website and the new website.

<i>Old Website</i>		<i>New Website</i>	
Mean	63.075	Mean	80.025
Median	62	Median	82.5
Mode	26	Mode	99
Range	98	Range	115
Minimum	15	Minimum	15
Maximum	113	Maximum	130
Sum	2523	Sum	3201

Good

These tables show us some interesting statistics about the data, to begin, the sum of sales for the new website is higher than the old website. This indicates that new website is better than the old one. This data also shows that the new website had the highest single sales in a day. The median is also higher, indicating that there were more sales per day across the entire dataset.

Hypotheses

In order to tell to have a successful analysis of the data we have to specify what indicates that one site is better than the other. To do this, we have to prove something is true. However, there are many values to consider if you are looking for which is a better website.

For example, if one website had a higher median but a low mode, which is the better website. As, there are many variables to consider to work out if a website is good or not, we use a null hypothesis to solve this.

A null hypothesis (H_0) "is a hypothesis which the researcher tries to disprove, reject or nullify." [1] By doing this, we prove the hypothesis or alternate hypothesis (H_1). It is often easier to disprove something to prove something else.

Good

Therefore, the H_1 of this data is that one website is a better website and generates more sales than the other website.

The H_0 of this data is that the old website and the new website had the same sales or New Website Sales = Old Website Sales.

If we can disprove H_0 then H_1 will be correct.

Identifying Suitable Tests

So, from the analysis of the initial data it looks as though the new website has more sales than the old one, based on the sum and the other averages. However, if we want to disprove the H_0 and therefore prove the H_1 we need to work out if the difference between the two sets of data is statistically significant.

Although we can see that there is a difference in the sales, we must prove that the difference is statically significant to wholly disprove the H_0 .

Good

Now that we have identified that we need to check if the difference is statistically significant, we can work out what tests we must conduct in order to prove this.

To determine if the difference is statistically significant, t-tests can be used to find out. "A t-test is used to compare the mean of two given samples" [2] which makes it useful for the data we have.

Testing

As decided in the previous section, t-tests will be used to determine if the difference is statistically significant. To make sure the value is significant however, we need to decide a significance level. This level called Alpha in excel, dictates where the significance becomes non-significant. For example, and Alpha value of 0.05 means a significance level of 5%. This means, that means that our data has a 5% chance of not being significant.

Test 1 (0.10 significance level):

t-Test: Two-Sample Assuming Unequal Variances		
	New Website	Old Website
Mean	80.025	63.075
Variance	1022.999359	988.4301282
Observations	40	40
Hypothesized Mean Difference	0	
df	78	
t Stat	2.390271823	
P(T<=t) one-tail	0.009624153	
t Critical one-tail	1.292499597	
P(T<=t) two-tail	0.019248306	
t Critical two-tail	1.664624645	

Now that we have performed a t-test we can use this data to see if the difference is significant or not. From these values, we can determine if the null hypothesis is disproven.

To disprove the H0 we need the t Stat value to be larger than t Critical two-tail.

In this test, t-stat > t Critical two-tail is true; therefore, we can disprove out H0 with this test.



Test 2 (0.05 significance level):

In this test we use a significance level of 0.05 meaning we are more stringent with our requirements for the statistical difference.

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	Variable 2
Mean	80.025	63.075
Variance	1022.999359	988.4301282
Observations	40	40
Hypothesized Mean Difference	0	
df	78	
t Stat	2.390271823	
P(T<=t) one-tail	0.009624153	
t Critical one-tail	1.664624645	
P(T<=t) two-tail	0.019248306	
t Critical two-tail	1.990847069	

In this test, t Stat > t Critical two-tail = true, therefore the null hypothesis is disproved again.

Test 3 (0.01 significance level):

In this test we use a significance level of 0.01 meaning we are more stringent with our requirements for the statistical difference.

t-Test: Two-Sample Assuming Unequal Variances		
	New Website	Old Website
Mean	80.025	63.075
Variance	1022.999359	988.4301282
Observations	40	40
Hypothesized Mean Difference	0	
df	78	
t Stat	2.390271823	
P(T<=t) one-tail	0.009624153	
t Critical one-tail	2.375110958	
P(T<=t) two-tail	0.019248306	
t Critical two-tail	2.640340015	

In this test, $t \text{ Stat} > t \text{ Critical two-tail}$ = false, therefore the null hypothesis is not disproven. This means that we cannot confirm our hypothesis.

Analysis

Out of the three tests we conducted two said that the null hypothesis was disproven. However, the last one did not. Yet, we can say that a 0.05 or 5% difference is statistically significant.



Considering Outliers

Although, we have disproven the H_0 in the previous section, we are not finished yet as we did not remove the outlier in the data on day 36. To rectify this, we will delete the datapoint on day 35 and check to see if the data still has a statistically significant difference.

This outlying data could have happened for many different reasons. Due to the nature of online sales, it can be influenced by many factors. However, here are a few examples of what could have affected the data.

- National Holiday
- Successful Social Media Post
- Improved ad placement for the day
- A particular social group gaining an interest in the product (E.G. School)
- Error in data handling

Test 1 (0.10 significance level):

t-Test: Two-Sample Assuming Unequal Variances		
	New Website	Old Website
Mean	79.43589744	62.74358974
Variance	1035.673414	1009.932524
Observations	39	39
Hypothesized Mean Difference	0	
df	76	
t Stat	2.304823627	
P(T<=t) one-tail	0.011953967	
t Critical one-tail	1.292790268	
P(T<=t) two-tail	0.023907933	
t Critical two-tail	1.665151353	

With the new values, we can see that the observations show 39 values instead of 40.

In the new data, $t \text{ Stat} > t \text{ Critical two-tail}$ = true therefore, this new test still disproves the null hypothesis and proves the hypothesis.

Test 2 (0.05 significance level):

In this test we use a significance level of 0.05 meaning we are more stringent with our requirements for the statistical difference.

t-Test: Two-Sample Assuming Unequal Variances		
	New Website	Old Website
Mean	79.43589744	62.74358974
Variance	1035.673414	1009.932524
Observations	39	39
Hypothesized Mean Difference	0	
df	76	
t Stat	2.304823627	
P(T<=t) one-tail	0.011953967	
t Critical one-tail	1.665151353	
P(T<=t) two-tail	0.023907933	
t Critical two-tail	1.99167261	

In this new test, $t \text{ Stat} > t \text{ Critical two-tail}$ = true again, therefore the null hypothesis is disproved again with the outlying value removed.

Test 3 (0.01 significance level):

In this new test we use a significance level of 0.01 meaning we are more stringent with our requirements for the statistical difference.

t-Test: Two-Sample Assuming Unequal Variances		
	New Website	Old Website
Mean	79.43589744	62.74358974
Variance	1035.673414	1009.932524
Observations	39	39
Hypothesized Mean Difference	0	
df	76	
t Stat	2.304823627	
P(T<=t) one-tail	0.011953967	
t Critical one-tail	2.376420376	
P(T<=t) two-tail	0.023907933	
t Critical two-tail	2.642078313	

In this new test, $t \text{ Stat} > t \text{ Critical two-tail} = \text{false}$, which is the same as before. Therefore, the null hypothesis is not disproved in this test. This means that we cannot confirm our hypothesis (H1) with this test.

Analysis

With the three new tests without the outlying value, there is a difference in the values. However, the results were the same, with 0.10 and 0.05 disproving the null hypothesis H0 again.

With the final test, the chance for false positives is only 1%. But for the use of out data, a 5% false positive chance is very accurate and enough to disprove the null hypothesis.

Conclusion

In conclusion, it is evident that the new website is a better website, with a difference that is statistically significant to a 5% significance level, which is large enough to give a conclusive answer to the question.

This conclusion has been reached by conducting various t-tests and removing outliers from the dataset. Which in turn disproved the null hypothesis of Old Website sales = New Website sales.

References

[1] M. Shuttleworth, "Null Hypothesis - The Commonly Accepted Hypothesis", Explorable.com, 2020. [Online]. Available: <https://explorable.com/null-hypothesis>. [Accessed: 27- May- 2020]

[2] V. Nigam, "Statistical Tests — When to use Which ?", Medium, 2020. [Online]. Available: <https://towardsdatascience.com/statistical-tests-when-to-use-which-704557554740>. [Accessed: 27- May- 2020]

Data Analysis and Visulaisation - Assingment 2

GRADEMARK REPORT

FINAL GRADE

67/0

GENERAL COMMENTS

Instructor

Student FEEDBACK FORM

This section details the *extent* to which the assessment criteria are demonstrated by you, which in turn determines your mark. The marks available for each category of skill are shown. Lecturers will use the space provided to comment on the achievement of the task(s), including those areas in which you have performed well and areas that would benefit from development/improvement.

Generic Assessment Criteria	Marks available	Marks awarded
1. Engagement with Literature Skills		0
	0	0
2. Knowledge and Understanding Skills		0
	0	0
3. Cognitive and Intellectual Skills		0

Descriptions of the dataset. You included the data descriptive analysis from Excel which is good and drew some initial conclusions, you did not mention that the new website was offering a 15% discount which may have skewed the figures.	10	7
Development of appropriate hypotheses for the analysis. You have included a good statement about the null and alternate hypothesis. This shows that you are able to present, evaluate and interpret qualitative and quantitative data, in order to develop lines of argument and make sound judgements in accordance with basic theories and concepts of statistical analysis.	10	7

Selection (and justification) of appropriate techniques to perform the analysis. You have used appropriate testing methods to perform the analysis. You mention that the tests become – more ‘stringent’, but you do not explain why. You also do not explain why you chose a two tailed test over a one tailed test? A one tailed test would show that the new website generates more sales – a two tailed test however, tests for difference above or below the critical value, more sales or less sales!	15	10
Consideration of outliers and other factors. You have considered the outliers and how they may have occurred. It is unlikely to be a data input error as it occurs in both populations. But we do not know. To improve here you could add visualisations to support you text.	10	6
4. Practical Application Skills		0

Application of selected analysis techniques. You have performed all the selected tests (T-Test two tailed) accurately, and given consideration the different alpha levels.	20	14
Interpretation of the results of the analyses. You have interpreted all your results and taken into account the various alpha levels and then compared the tests with the outlier data removed.	15	12
Presentation of the results in both textual and, where appropriate, diagrammatical form. You have included several visualisations to support your discussion. Although, there is an assumption that the reader understands the graphics, perhaps a little more explanation would help.	10	6
5. Transferable Skills for Life and Professional Practice		0

Presentation of the paper in an appropriate professional format. Correct use of technical English. Suitable labelling and referencing strategies. You have demonstrated that you can communicate the results of your study/work accurately and reliably, and with structured and coherent arguments. To improve on this you should give more consideration to the labelling of visualisations and referencing them – Figure 1 for example	10	5
		67

Assessment Mark (Assessment marks are subject to ratification at the Exam Board. These comments and are to give feedback on module work and are for guidance only until they are >

PAGE 1

PAGE 2

PAGE 3

PAGE 4

QM

Label

Please label your figures and use this label as a reference in your text. e.g. Figure 1



Comment 1

Why would they both have an increase in sales from day one?



Comment 2

This graph excludes the 'blip' figures, you should mention this.

QM

Good

Good

PAGE 5

QM

Good

Good

QM

Good

Good

PAGE 6



Comment 3

The null hypothesis can be rejected

PAGE 7



Comment 4

This could be extended - Why is a 5% difference statistically significant - perhaps a citation here would help.

PAGE 8

PAGE 9

PAGE 10
