The Amateur Enthusiast vs. the Bureau of Meteorlogy

A Comparison of Rainfall Data

Sherri McRae: s3117889

Last updated: 22 October, 2017

RPubs link information

- Rpubs link: http://rpubs.com/smcr/321139

Introduction

KM, a resident of Sunbury, Victoria, has been recording the amount of rainfall collected in his backyard rain gauge with enthusiastic precision for more than a decade. This investigation is to determine if there is any difference between KM's records and the data available for the same area from the Bureau of Meteorology.

Problem Statement

Three years of KM's rainfall data, from January 2014 to December 2016, were compared with data for the same region and time period from the website of the Bureau of Meteorology. A paired t-test was used to compare monthly rainfall totals as well as the daily rainfall totals.

Data 1

The original data records from KM were handwritten onto calendars. Four months were excluded from the study due to missing data. These months were February, March, June and July of 2014. The handwritten data was transcribed onto excel spreadsheets. Daily and monthly totals were included.

Open source data was obtained from the Beauru of Meteorology website at: http://www.bom.gov.au/climate/data/index.shtml. The specific weather station chosen for comparison, station number 086282 at Melbourne Airport, was selected because it was the closest to Sunbury with the most complete data set.

The website data is appears in the form of a calendar with columns representing months and individual cells representing days. This data was copied and pasted directly into excel.

Data 2

The Sunbury data was recorded at a location described by the following details: Latitude: 37.57° S Longitude: 144.74° E Elevation: approx 200 m.

The rain gauge has markings up to 33mm but holds much more in an unmarked extension on top. If rainfall exceeded 33mm, the collection was poured into another cylinder in order to take the measurement. The gauge is set up in an open space just beyond the back door of a residential house. Measurements are recorded between 7am and 9am each morning and reflect precipitation occurring in the previous 24 hours.

The Bureau of Meteorology data was obtained from the weather station located at the Melbourne Airport site. The location description is as follows: Latitude: 37.67° S Longitude: 144.83° E Elevation: 113 m. Measurements are recorded at 9am each morning and reflect precipitation occurring in the previous 24 hours.

The two locations are 14.3km apart.

In order to perform a paired t-test two columns were created, one for KM data and one for the Bureau of Meteorology (BOM) data. Rows were equivalent to months on one worksheet and days on another. The difference (d) was calculated by subtracting the BOM observation from the KM observations: $d = KM - BOM$

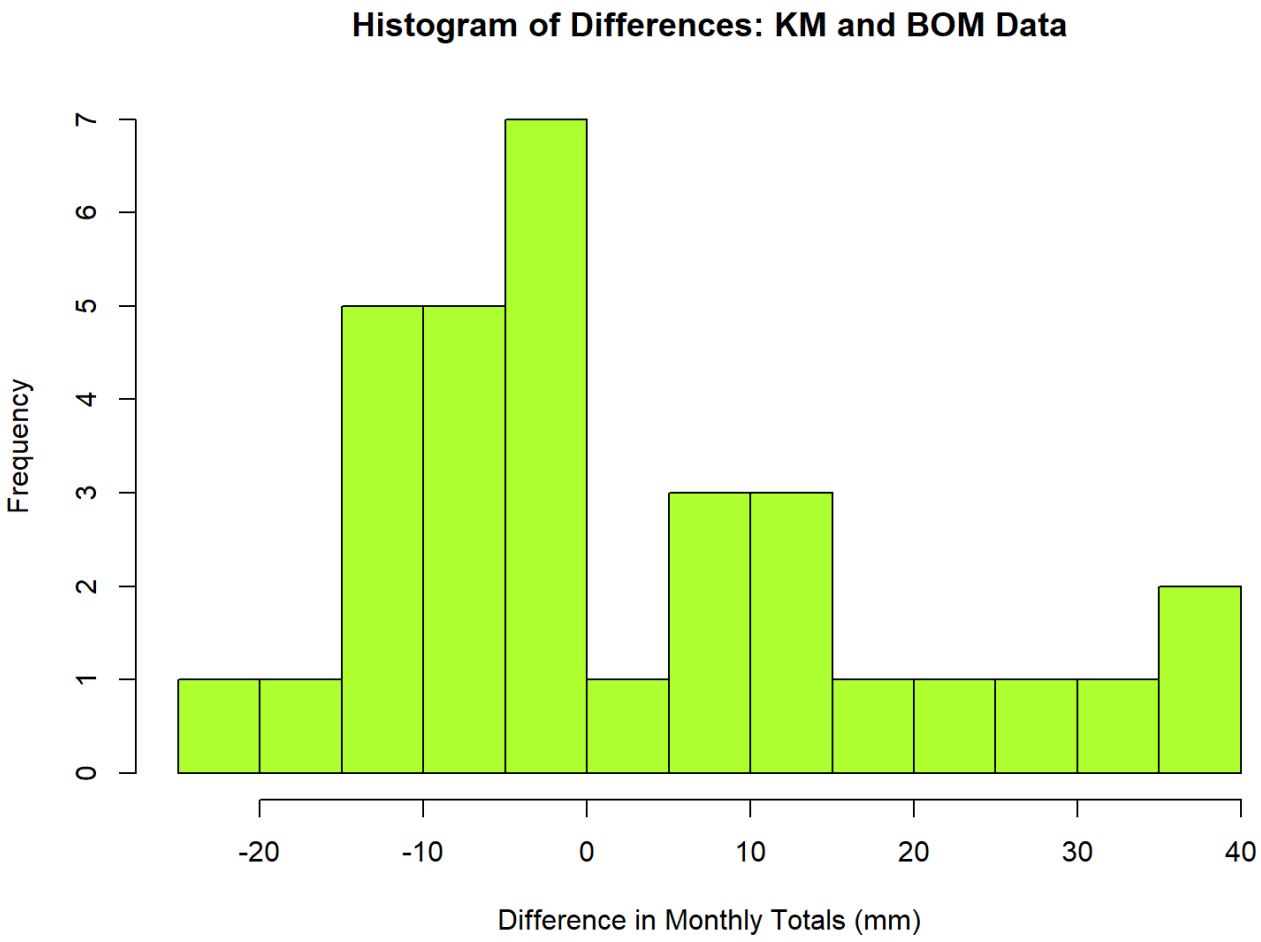
Descriptive Statistics and Visualization (Monthly Data)

A total of 32 months and 976 days were included in the study from January 1, 2014 to December 31, 2016. The years 2014, 2015 and 2016 were selected because KM had these records readily available. Four months were excluded because data was missing. These months were February, March, June and July of 2014.

A summary of the differences between KM and BOM monthly and daily data and histograms appear below:

```
Monthly %>% summarise(Min = min(Monthly$d),
Q1 = quantile(Monthly$d,probs = .25),
Median = median(Monthly$d),
Q3 = quantile(Monthly$d,probs = .75),
Max = max(Monthly$d),
Mean = mean(Monthly$d),
IQR = IQR(Monthly$d),
STD = sd(Monthly$d),
n = n())
```

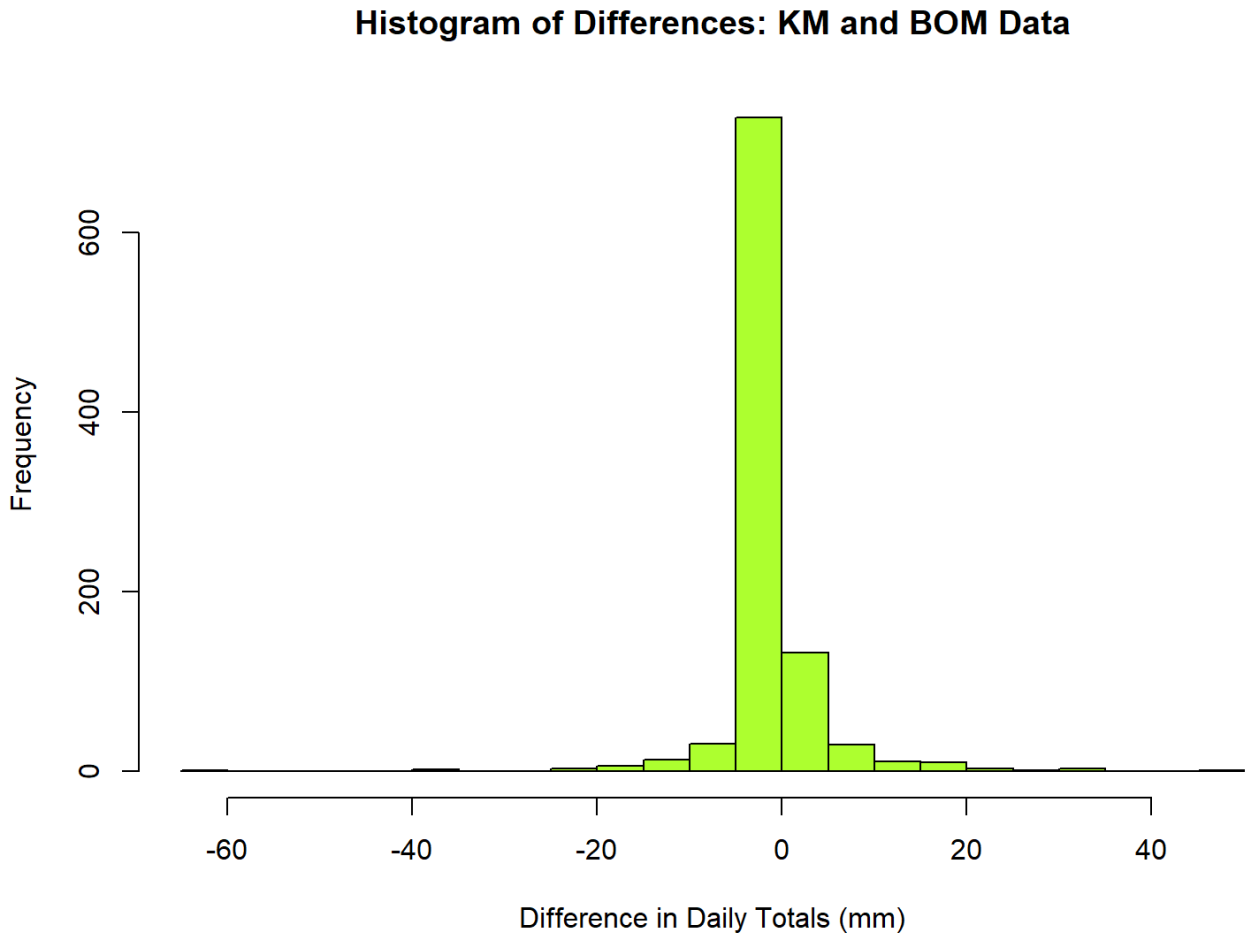
```
Monthly$d %>%  hist(col="greenyellow", breaks=20,
                        xlab="Difference in Monthly Totals (mm)",
                        main="Histogram of Differences: KM and BOM Data")
```


Histogram of Differences: KM and BOM Data
Frequency
0
1
2
3
4
5
6
7
-20
-10
0
10
20
30
40
Difference in Monthly Totals (mm)

Summary Statistics and Visualization (Daily Data)

```
Daily %>% summarise(Min = min(Daily$d),
Q1 = quantile(Daily$d,probs = .25),
Median = median(Daily$d),
Q3 = quantile(Daily$d,probs = .75),
Max = max(Daily$d),
Mean = mean(Daily$d),
IQR = IQR(Daily$d),
STD = sd(Daily$d),
n = n())
```

```
Daily$d %>%  hist(col="greenyellow", breaks=20,
                        xlab="Difference in Daily Totals (mm)",
                        main="Histogram of Differences: KM and BOM Data")
```



Hypothesis Testing I

The null hypothesis is the mean of the differences between the two datasets is equal to zero:

$$H_0 : \mu_d = 0$$

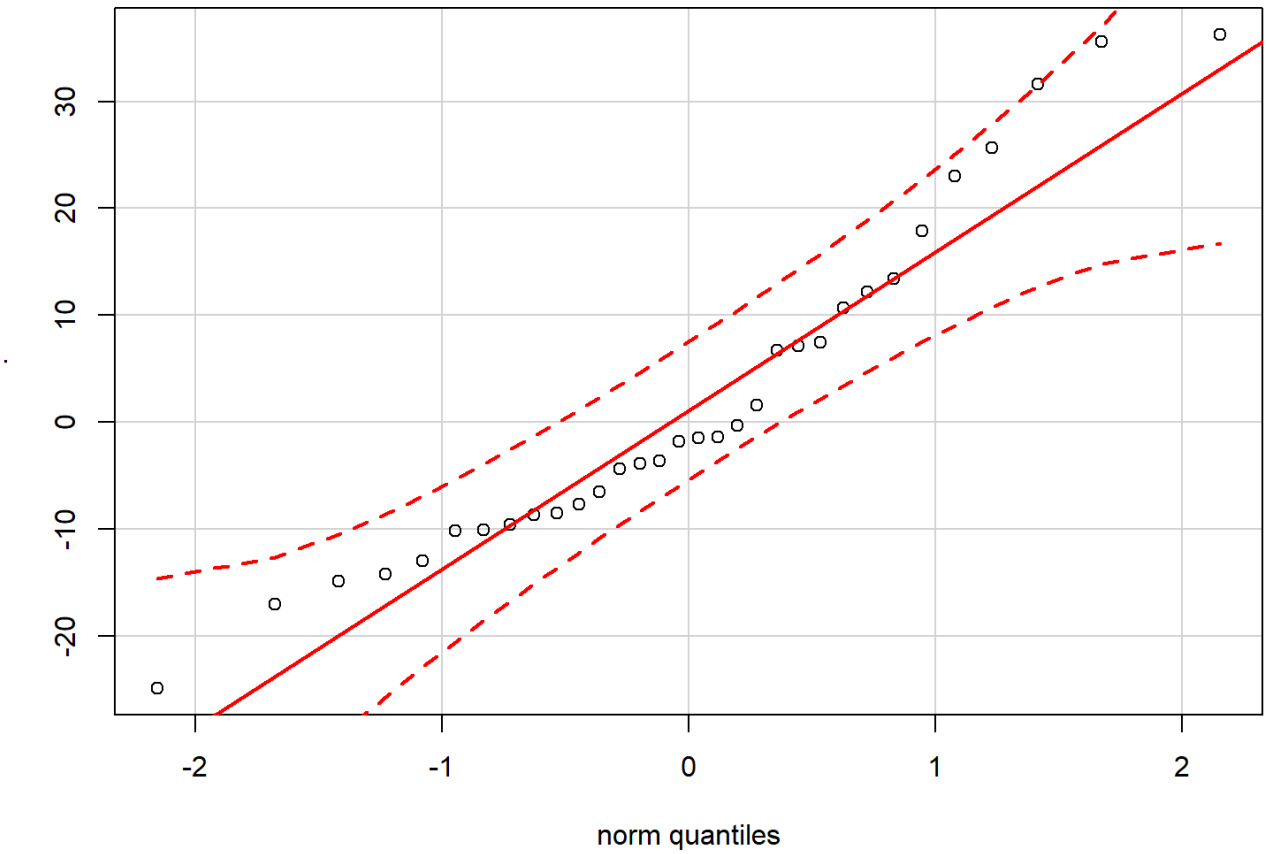
The alternative hypothesis is the mean of the differences is not equal to zero:

$$H_A : \mu_d \neq 0$$

A paired t-test assumes that the data is normally distributed. The daily rainfall data has a sample size of 976. As this is greater than 30, and it can be assumed that the data follows a normal distribution based on the central limit theorem.

The sample size of the monthly data is 32, which is also greater than 30. However, as it is very close to 30 and the histogram does not appear to be normally distributed, it is worth checkint the Q-Q Plot:

```
Monthly$d %>% qqPlot(dist='norm')
```



The Q-Q Plot confirms a normal distribution.

The significance level is 0.05.

Hypthesis Testing 2

```
t.test(Monthly$d,
       mu = 0,
       alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  Monthly$d
## t = 0.75488, df = 31, p-value = 0.456
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -3.568406  7.762156
## sample estimates:
## mean of x
##  2.096875
```

```
t.test(Daily$d,
       mu = 0,
       alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  Daily$d
## t = 0.39373, df = 975, p-value = 0.6939
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2747271  0.4126369
## sample estimates:
##  mean of x
## 0.06895492
```

A paired t-test on the difference between the monthly totals of KM and BOM data had 31 degrees of freedom, a p-value of 0.456 and a 95% CI[-3.57,7.76]. As the p-value was greater than 0.05 and the 95% CI includes zero, the analysis failed to reject the null hypothesis. There is no statistically significant difference between KM and BOM monthly rainfall totals.

A paired t test was used to analyse the difference between the daily rainfall data taken at the weather station at Melbourne Airport and in Sunbury by KM. The p-value=0.6939 which is greater than 0.05. The 95% confidence interval of the mean [-0.27,0.41] includes zero. Therefore, the investigation fails to reject the null hypothesis. There is no statistically significant difference between the two datasets.

Discussion part 1

Inspection of the daily rainfall data from the two datasets revealed possible problems that may affect the statistical analysis.

Firstly, there were a number of occasions where a large amount of rainfall was recorded with a day difference between the Melbourne Airport(BOM) and Sunbury(KM). For example; KM recorded 48 mm on December 29, 2016. The BOM data recorded nothing on the 29th but 62.8 mm on the following day: December 30, 2016. This appears to be the result of the same storm system hitting Sunbury and Tullamarine at different times. As a result, the difference (d) calculated on these days was -62.8 (the minimum in the dataset) for December 30th and +48 (the maximum in the dataset) for December 29th.

As this difference creates similar extremes on either side of zero, it contributes to a mean close to zero. As a result, the time delay of these observations would not have shifted the 95% confidence interval away from zero. The null hypothesis, that the mean difference between the datasets is zero, is therefore supported by this time delay effect - even though it could be argued that the time delay creates a real difference between the datasets and should detract from the null hypothesis, not support it.

Discussion part 2

The second possible problem was that according to the Bureau of Meteorology, there were 669 days out of the 976 included in this study where no rainfall was recorded. This represents 68.5% of the BOM dataset. There were a similar number of days without rainfall in the KM dataset. Zero rainfall is a valid observation, however, if there are many more zeroes recorded than observations greater than zero, the differences between the observations greater than zero may be obscured by the sheer number of observations where both KM=0 and BOM=0. Analyzing the difference between two datasets with so many zeroes could result in a false confirmation of the null hypothesis: that is the mean difference between two datasets is zero. In order to explore this, the zero records from the daily rainfall dataset were removed to create a new dataset called Daily_cleaned2.

Prior to the creation of this new dataset, an excel formula was used to highlight the instances when KM>0, BOM>0 and (KM>0)-(BOM>0)=0. There were 3 such cases. These observations were removed along with the other observations where both KM=0 and BOM=0 as a result of a time saving purge method. This resulted in a loss of approximately 0.6% of rainfall data. It was assumed that this loss would not impact the data analysis.

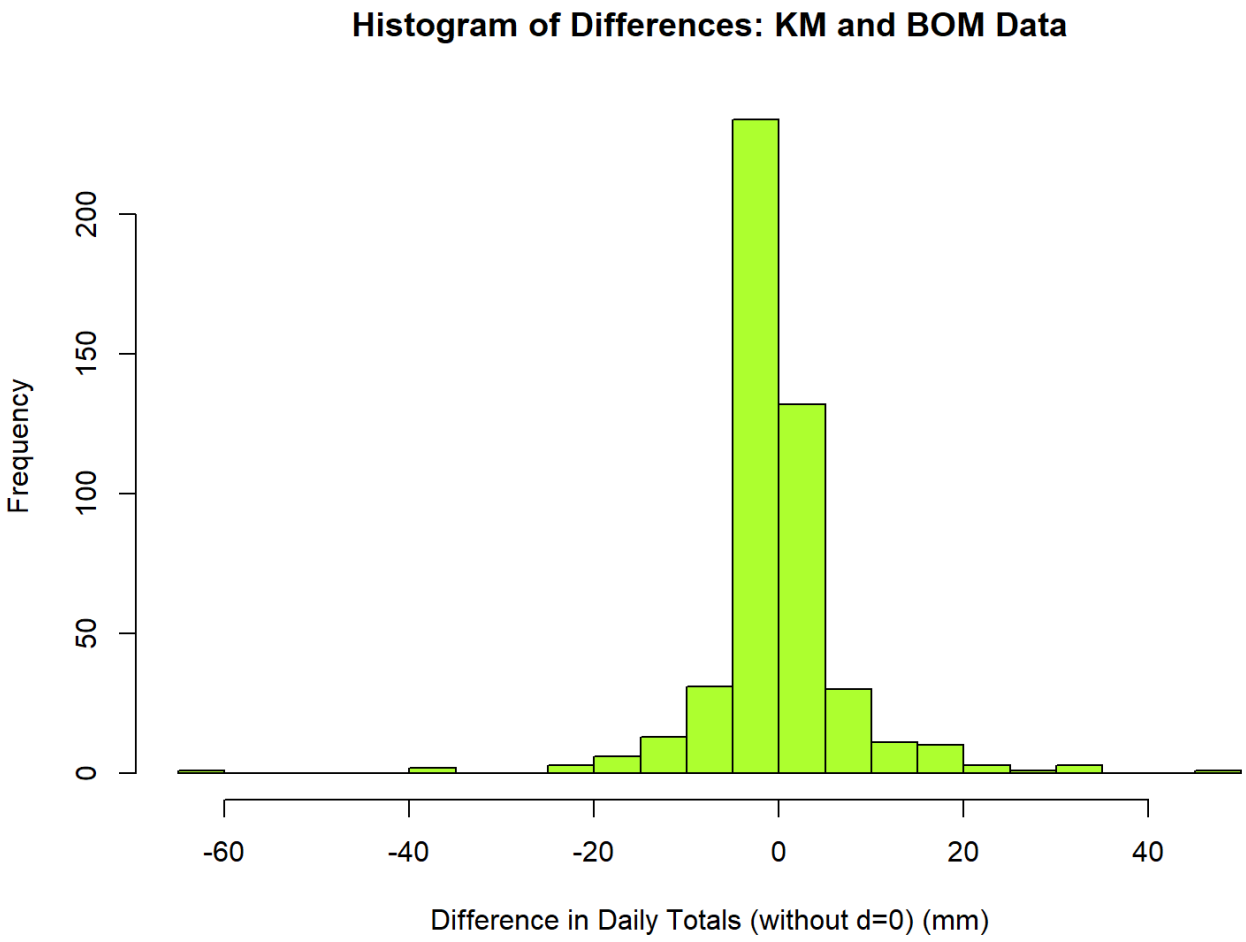
The new zero-free dataset was analyzed in the same way as the previous datasets. As the sample size was 481, (ie, n>30) normal distribution was assumed.

A Final Analysis I

```
Daily_cleaned2 %>% summarise(Min = min(Daily_cleaned2$d),
Q1 = quantile(Daily_cleaned2$d,probs = .25),
Median = median(Daily_cleaned2$d),
Q3 = quantile(Daily_cleaned2$d,probs = .75),
Max = max(Daily_cleaned2$d),
Mean = mean(Daily_cleaned2$d),
IQR = IQR(Daily_cleaned2$d),
STD = sd(Daily_cleaned2$d),
n = n())
```

The median is now -0.4 as opposed to the "Daily" dataset that had a median of 0. The interquartile range has increased from 0.4 to 3.8 and the standard deviation has increased from 5.5 to 7.8. The range remains the same with a minimum of -62.8mm and a maximum of 48mm.

```
Daily_cleaned2$d %>%  hist(col="greenyellow", breaks=20,
                        xlab="Difference in Daily Totals (without d=0) (mm)",
                        main="Histogram of Differences: KM and BOM Data")
```


Histogram of Differences: KM and BOM Data
Frequency
0
50
100
150
200
-60
-40
-20
0
20
40
Difference in Daily Totals (without d=0) (mm)

A Final Analysis 2

```
t.test(Daily_cleaned2$d,
       mu = 0,
       alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  Daily_cleaned2$d
## t = 0.39355, df = 480, p-value = 0.6941
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.5586569  0.8384906
## sample estimates:
## mean of x
## 0.1399168
```

A paired t-test of Daily_cleaned2 shows the degrees of freedom were 480, the p-value was 0.6941 and confidence interval of the mean was [-0.56, 0.84]. The 95% CI is a wider interval because 495 observations at zero were removed thus reducing the certainty that the mean would be close to zero. As zero is still included in this CI and the p-value is greater than 0.05, the null hypothesis is not rejected. Again there is no statistically significant difference between KM and BOM data. The p-value is 0.0002 greater than the p-value of the dataset that contained 495 more zeroes. The zero values in the original dataset did not obscure the results of the analysis.

Conclusion

In conclusion, there is no statistically significant difference between the rainfall data collected by KM and the data collected by the Bureau of Meteorology at a weather station located 14.3km away. This is true for both the monthly and daily totals. However, as there was sometimes a delay in storm systems between the two locations, KM's daily observations would not be able to replace that of the weather station at the Melbourne Airport.

References

- http://www.bom.gov.au/climate/data/index.shtml.