2018

Assignment 2

COMPUTATIONAL STATISTICS AND VISUALISATION

NIKHIL JAGATIA - 18055146

MANCHESTER METROPOLITAN UUNIVERSITY

Nikhil Jagatia
18055146

Assignment 2

1

a) A numerical summary of the data by the laboratories:

	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
Mean	4.062	4.00694444	4.00555556	3.92	3.963529	3.960303	4.006486
Median	4.055	4.02	4.01	3.915	3.99	3.98	4.03
Mode	4.04	4.08	4.01	3.88	3.89	4.020000	4
Std Dev	0.03131314	0.08328332	0.02183633	0.03216338	0.0548763	0.06395725	0.07671072
Skewness	0.92325236	-0.7421144	-0.0357983	0.4397193	-0.3806467	-1.1876258	-1.387466
Variance	0.00098051	0.00693611	0.00047683	0.00103448	0.00301141	0.00409053	0.00588453
Q1	4.04	3.96	3.985	3.9	3.89	3.93	4
Q3	4.07	4.08	4.02	3.95	4.01	4	4.05
Inter Q range	0.03	0.12	0.035	0.05	0.12	0.07	0.05
min	4.02	3.85	3.97	3.88	3.89	3.82	3.81
max	4.13	4.11	4.04	3.97	4.02	4.02	4.1

Figure 1: This is a summary of the numerical data from the laboratories

For lab 1 data

```
PROC IMPORT DATAFILE= '/folders/myfolders/sasuser.v94/l1.csv'
DBMS=CSV
OUT=L1;
GETNAMES=YES;
RUN;

PROC PRINT DATA= L1;
RUN;

PROC UNIVARIATE DATA= L1;
FREQ MEAS;
VAR MEAS;
RUN;
```

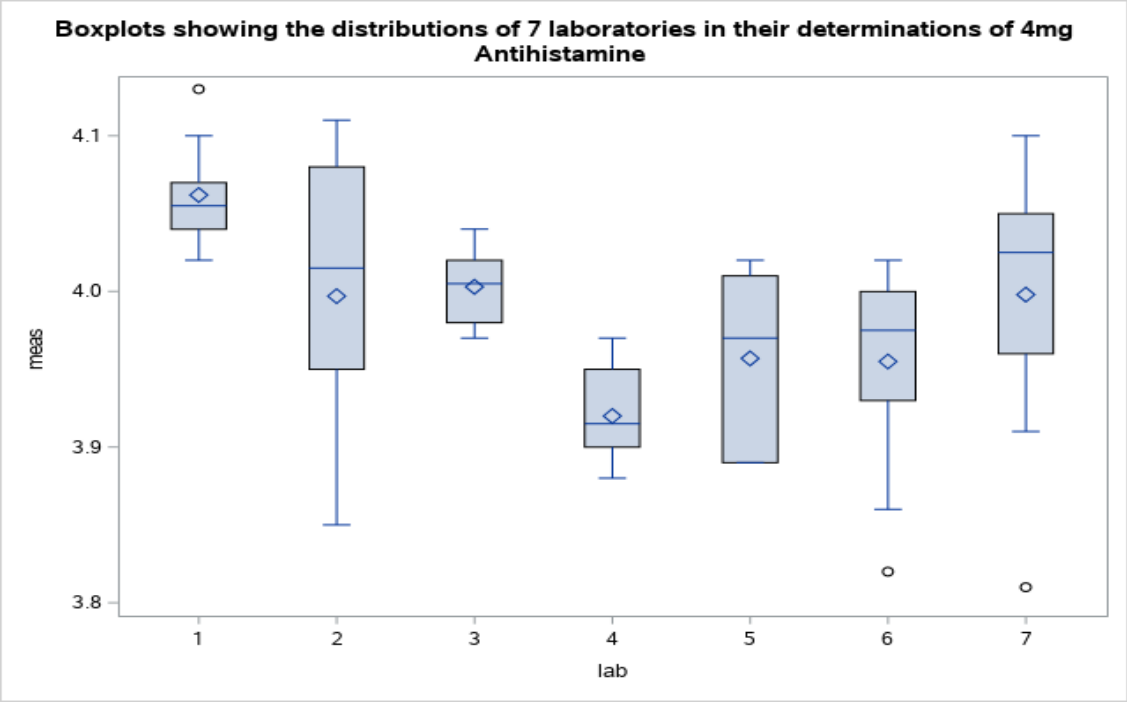
For lab 2 data

```
PROC IMPORT DATAFILE= '/folders/myfolders/sasuser.v94/l2.csv'
DBMS=CSV
OUT=L2;
GETNAMES=YES;
RUN;

PROC PRINT DATA= L2;
RUN;

PROC UNIVARIATE DATA= L2;
FREQ MEAS;
VAR MEAS;
RUN;
```

Similar files were constructed and made and then imported separately like lab 1 and lab 2 data to generate data to be then formed into a separate table such that files were called: l1, l2, l3, l4, l5, l6 and l7 using the same structure code as above but inserting remaining filenames

Nikhil Jagatia
18055146

b) Figure 2: Side by side box-plots of the seven distributions:




```
PROC IMPORT DATAFILE='/folders/myfolders/sasuser.v94/chlor.csv' REPLACE
        DBMS= CSV
        OUT=WORK.IMPORT;
        GETNAMES= YES;
RUN;

PROC SGPLOT DATA= WORK.IMPORT;
        VBOX MEAS / CATEGORY= LAB;
        TITLE 'Boxplots showing the distributions of 7 laboratories in their determinations of
4mg Antihistamine';
RUN;
```

c) Firstly, to comment on these boxplots we can see since we were expecting the antihistamine to be measured to be 4mg the laboratory with the least variation and with the most accurate findings is **laboratory 3**. This is not only where the mean and median are similar, but they both lie around the 4 mark and is only slightly negatively skewed (as shown in the numerical summary). In addition, there were no outliers recorded. Similarly, **laboratory 1** did have one of the least amounts of variation, also showing good validity of their readings however the recordings shown were slightly greater than that of laboratory 3's readings and therefore less accurate to what the true value should be. Yet laboratory 1 has a more symmetrical distribution and lesser interquartile range. Although we can notice there is an outlier amongst the data for laboratory.

Laboratory 2: Here, although the median seems to lie around the 4 value, we can see the greatest range out of all the distributions, identified by the whiskers. This boxplot is negatively skewed with the largest interquartile range. Also this laboratory has the highest value recorded and lowest value recorded of all the laboratories.

Nikhil Jagatia
18055146

Laboratory 4: This boxplot shows symmetry in the distribution of data but it shows the worst mean and median out of all the laboratories which indicates it as being the least accurate.
Laboratory 5: Strikingly this boxplot is negatively skewed with a large interquartile range, with the mean and median still a further distance away from the 4 value which is expected.
Laboratory 6: This has a negative skewness with a smaller interquartile range than laboratory 5 however the averages are similar to laboratory 5's, but still not as close or as accurate as laboratory 3, even with the outlier.
Laboratory 7: This negatively skewed boxplot has a mean which lies very close to 4, but a larger value for the median. In addition there is a very small value which is the smallest of the outliers which is around 3.8, along with the second largest range it is not as accurate as laboratory 3 even with the modal value of 4. Yet it is very negatively skewed.
There are similarities with the interquartile ranges of laboratory 2 and 5, also the maximum values found by laboratories 5 and 6 are identical.

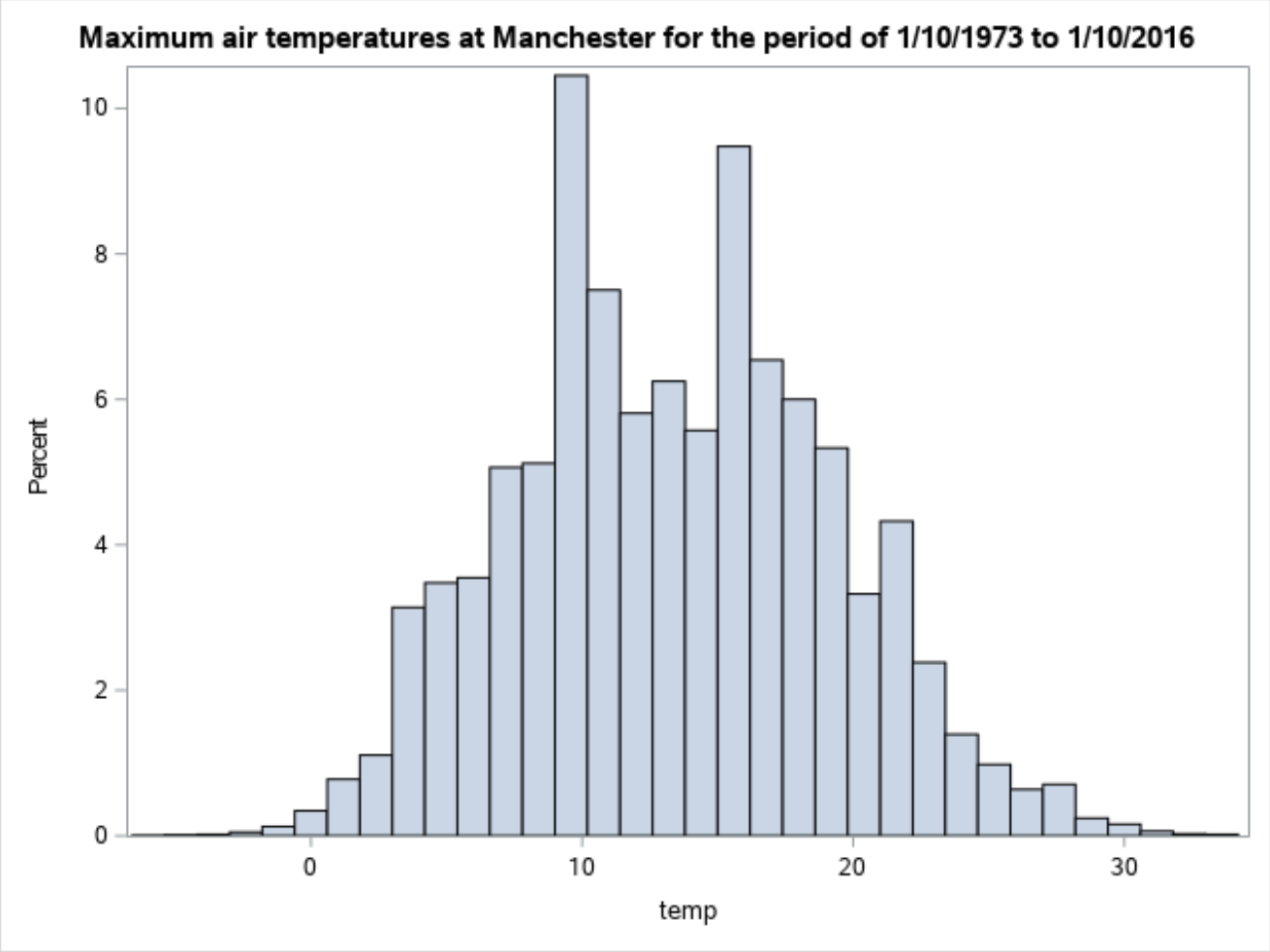
This data representation indicates that laboratory 3 may have better equipment and facilities with less error, range and variation in their measurements compared to the rest of the other laboratories.

Nikhil Jagatia
18055146

2.

a)

A histogram that shows maximum air temperatures at Manchester for the period of 1/10/1973 to 1/10/2016




```
PROC IMPORT DATAFILE= '/folders/myfolders/sasuser.v94/man.csv'
        DBMS=CSV
        OUT= MAN;
        GETNAMES=YES;
RUN;

PROC PRINT DATA=MAN;
RUN;


        PROC SGPLOT DATA= MAN;
        HISTOGRAM TEMP;
        TITLE 'Maximum air temperatures at Manchester for the period
of 1/10/1973 to 1/10/2016';
        RUN;
```

Nikhil Jagatia
18055146

b)

This histogram shows a bimodal distribution with a modal temperature of 10 degrees, with the second most modal of 17 degrees. This moderately symmetrical histogram a 40 degree rang where the maximum temperature recorded at 34 degrees and a minimum temperature recorded at -6 degrees (found using univariate process) and the mean temperature being 13.33 degrees.
The two spikes for the temperatures making is bimodal could represent the seasons where the seasons with the most extreme temperatures winter and summer aren't identified but more of early & late Autumn and early & late Spring. Manchester is not known for its scorching or sub-zero temperatures so this histogram fits into the characteristics of the typical climate of England. With the concept of global warming looming we can suggest that the second most modal value of 17 may overtake the 10 degrees as there is a higher chance of having warmer temperatures recorded.