MANCHESTER METROPOLITAN UNIVERSITY

MSC DATA SCIENCE

6G7Z1020

COMPUTATIONAL STATISTICS AND VISUALISATIONS

ASSIGNMENT 2: -
ANTIHISTAMINE & MANCHESTER TEMPERATURES

DANIEL RAY
18053479

04/11/18

INTRODUCTION

Continuing from the previous assignment, we are tasked with using the powerful statistical analysis program SAS and its proprietary language, SAS Language, to employ several descriptive statistical techniques to obtain a deeper understanding of some dataset.

In Assignment Two, we have been given two datasets; chlor.csv – which contains data of 10 determinations of an antihistamine from seven different labs. The second dataset, man.csv, containing observations of the daily minimum temperature in Manchester from 1st October 1973 to 1st October 2016.

ANTIHISTAMINE

In the first question, we have been given the task of describing a dataset which contains 10 determinations of samples containing 4mg of an antihistamine from seven different labs. In the dataset, we have two variables, meas – which is the determinations of the 10 samples along with the variable lab which contains the id of the lab, ranging from 1 to 7.

The first part of the question asks us to produce a numerical summary of the data by each laboratory. Within the summary, to gain an insight to the data of each labs testing, we would want to have measures of central tendency:

Analysis Variable: meas							
lab	N Obs	Mean	Median	Minimum	Maximum	Range	Std Dev
1	10	4.0620000	4.0550000	4.0200000	4.1300000	0.1100000	0.0325918
2	10	3.9970000	4.0150000	3.8500000	4.1100000	0.2600000	0.0896970
3	10	4.0030000	4.0050000	3.9700000	4.0400000	0.0700000	0.0231181
4	10	3.9200000	3.9150000	3.8800000	3.9700000	0.0900000	0.0333333
5	10	3.9570000	3.9700000	3.8900000	4.0200000	0.1300000	0.0571645
6	10	3.9550000	3.9750000	3.8200000	4.0200000	0.2000000	0.0670406
7	10	3.9980000	4.0250000	3.8100000	4.1000000	0.2900000	0.0848266

Table 1 – Statistical Summary of the 10 Samples for Each Lab

Looking at this table, we can see a statistical breakdown of each of the labs 10 samples. As expected, each of the seven labs have 10 observations. We can see that the mean ranges from 3.92 (Lab 4) to 4.06 (Lab 1). The median values for all seven labs fall within the range of 3.92 (Lab 4) and 4.05 (Lab 1). The min and max values for the seven labs again fall within the range of 3.81 (Lab 7), the smallest value and 4.13 (Lab 1) which is the largest value. The smallest max value is 3.97 (Lab 4) and the largest min value is 4.02 (Lab 1).

One of the most important statistical measurement for this specific dataset is Range. Range allows us to see the disparity between the smallest value and largest value for each lab, the smallest range is from Lab 4 with only a variation of 0.07 from the smallest value to largest. The lab with the largest range is Lab 7 with 0.29.

Finally, the last measurement of the summary table is standard deviation. Standard Deviation, much like the range, allows us to see the spread of the data. However, the standard deviation is the measurement to tell how spread out the values are from the mean, a low standard deviation means the range of number is very close to the average. Looking at the summary table, we can see that Lab 3 has the smallest standard deviation with 0.23 and Lab 2 with 0.9.

To gain a deeper understanding of the data, especially the distribution of values per lab, we can employ the descriptive statistical technique called a box plot. A box plot is a visual representation of a dataset containing several important measurements; min value, lower quartile, median, mean, upper quartile, max value and outliers. Quartiles are a type of quantile; the lower quartile represents the middle value from the min value and the median. The upper quartile being the middle value between the median and the highest value of the dataset.

Using the SAS procedure SGPLOT, we can produce several different types of box plots. Using the key words, vbox or hbox, allows us to create either a vertical or horizontal box plot. We can then use the category or group key word to group the data into several different box plots on the same graph. This is ideal for the dataset we are examining as we want to group the data by labs. Below are both a vbox and hbox of the antihistamine dataset:

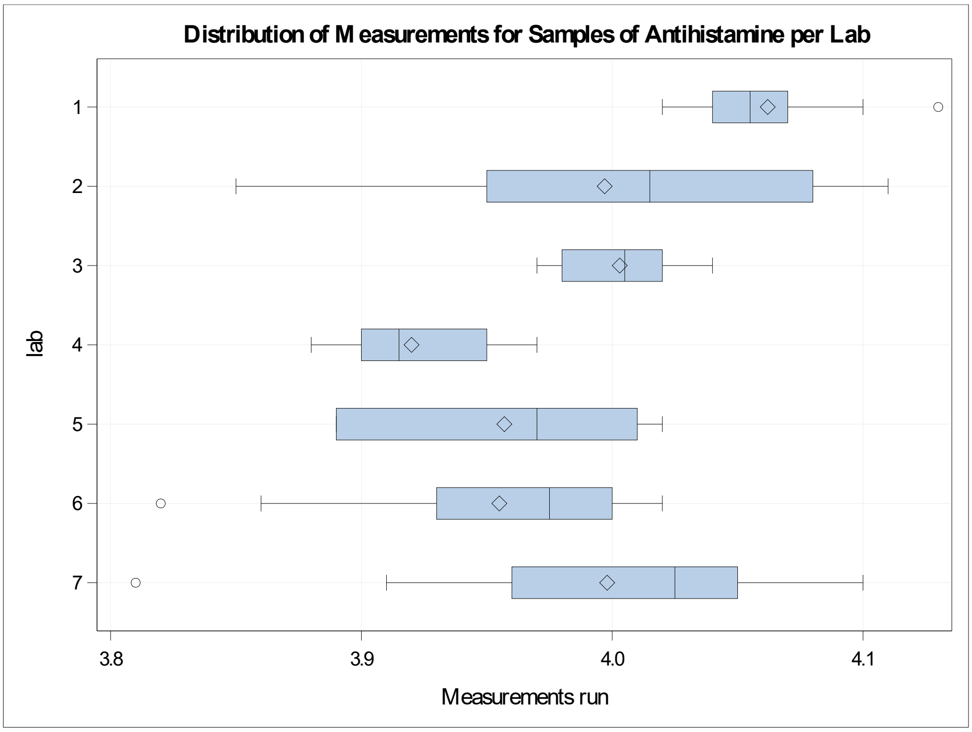

Figure 1 - Vertical Box Plot of The Distribution of Measurements for Samples of Antihistamine per Lab

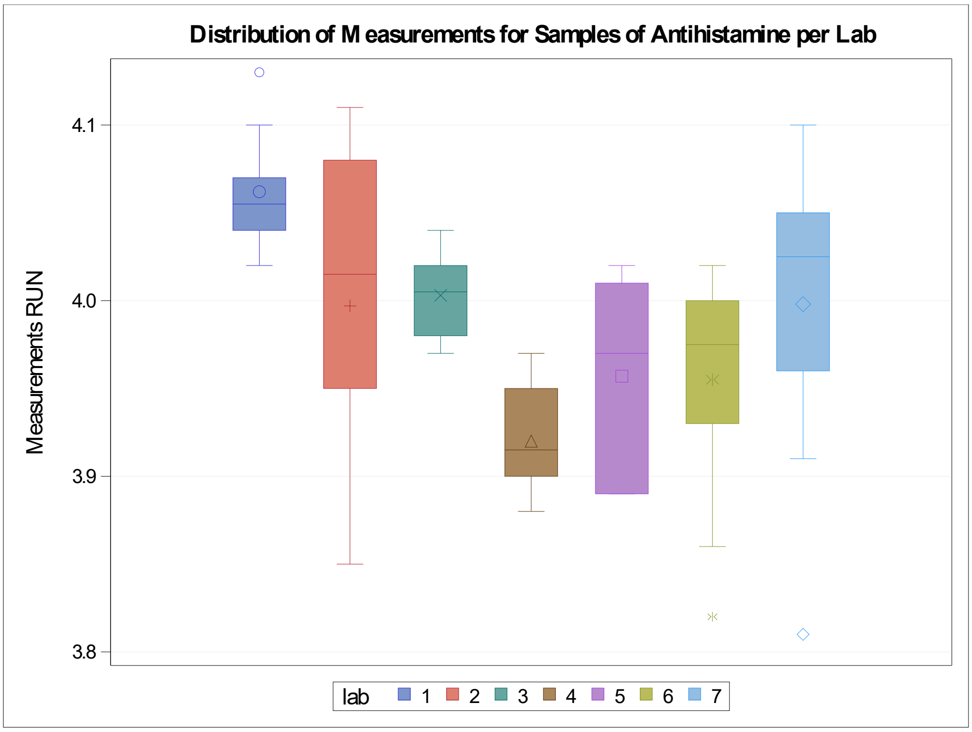

Figure 2 - Horizontal Box Plot of The Distribution of Measurements for Samples of Antihistamine per Lab

Looking at both graphs, we can see there is a lot of information to digest. To explain with precision, we must define what each aspect of the box plot is:

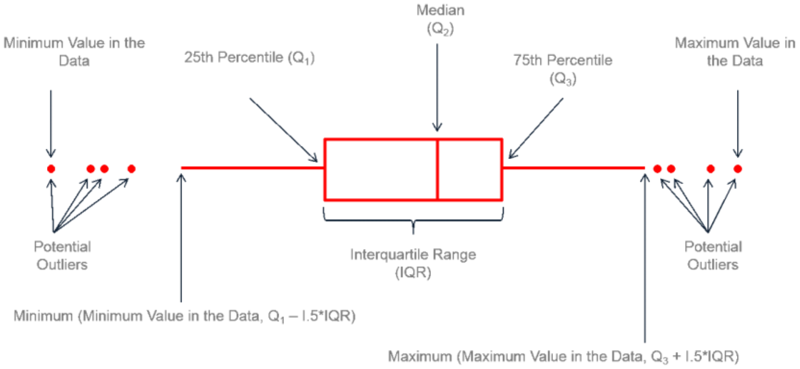
Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

Figure 3 - Explanation of the Aspects of a Box Plot
https://www.leansigmacorporation.com/box-plot-with-minitab/

As we can see, the dots outside of the box plot represent the potential outliers of the data, the lines coming from the box, known as whiskers represent the min and max values. The lines of the box represent the Q1 and Q3 quartiles of the data set. The line in the middle of the box represents the median. In the SAS produced graphs, the symbol within the box represents the mean value of the data.

Now that we can digest a box plot correctly, we can analyse the Antihistamine dataset. Looking at the graphs, we can say that Lab 2 has the largest distribution of measurements of the samples. We can also say that Lab 3 had the smallest distribution of measurements, thus making it the most accurate lab.

Delving deeper, we can say that Lab 1 had the smallest Inter Quartile Range whilst Lab 2 had the largest. Lab 3's Mean and Median are the closest together4, strengthening our argument for Lab 3 being the most accurate lab. Interestingly, Lab 5's minimum value is also its Q1 value. We could also argue that Lab 4, although having a small distribution, is the furthest away from 4, which is measurement of each sample. Now comparing both graphs, I have to say that, I find the vertical box plot easier to digest as the comparisons are done on the horizontal axis rather than the vertical.

CODE

```
/* Import DATA*/
PROC IMPORT DATAFILE="/folders/myfolders/2/data/chlor.csv"
      DBMS=CSV
      OUT=anti;
      GETNAMES=YES;
RUN;
/* Stat Summary*/
PROC MEANS DATA=anti MEAN MEDIAN MIN MAX Range STD;
    CLASS lab;
    VAR meas;
RUN;
/* Horizontal Box Plot*/
proc sgplot data=anti;
      hbox meas / category=lab;
      XAXIS GRID;
      YAXIS GRID;
      Title 'Distribution of Measurements for Samples of
Antihistamine per Lab ';
      Label meas = 'Measurements'
run;
/*Vertical Box Plot*/
PROC SGPLOT DATA=anti;
    VBOX meas / GROUP=lab;
      YAXIS GRID;
```

MANCHESTER TEMPERATURES

In the second question of the assignment we are given a dataset containing the maximum air temperatures in Manchester from 1/10/1973 to 1/10/2016 taken from the National Climatic Data Centre. Using SAS, we can produce a statistical summary of the dataset:

Analysis Variable : temp		
Mean	Median	Std Dev
13.3256683	13.0000000	5.8597092

Table 2 – Statistical Summary of the Temperature in Manchester

We can see that the mean value is 13.33 °C, whilst the median is 13 °C. The standard deviation of the dataset is 5.86 °C. Using this data, we can say that the most frequent temperature in Manchester is roughly 13 °C. To gain a deeper understanding of the distribution of frequencies of temperature in Manchester, we can employ the descriptive statistical method called a Histogram, which visualises the shape and spread of continuous sample data.

Using SAS, we can create a Histogram using two procedures. Like the box plots, we can call the SGPLOT function to create a histogram of the temperature in Manchester. We can also call the UNIVARIATE function to create the histogram. I will first use the UNIVARIATE procedure to produce the histogram:

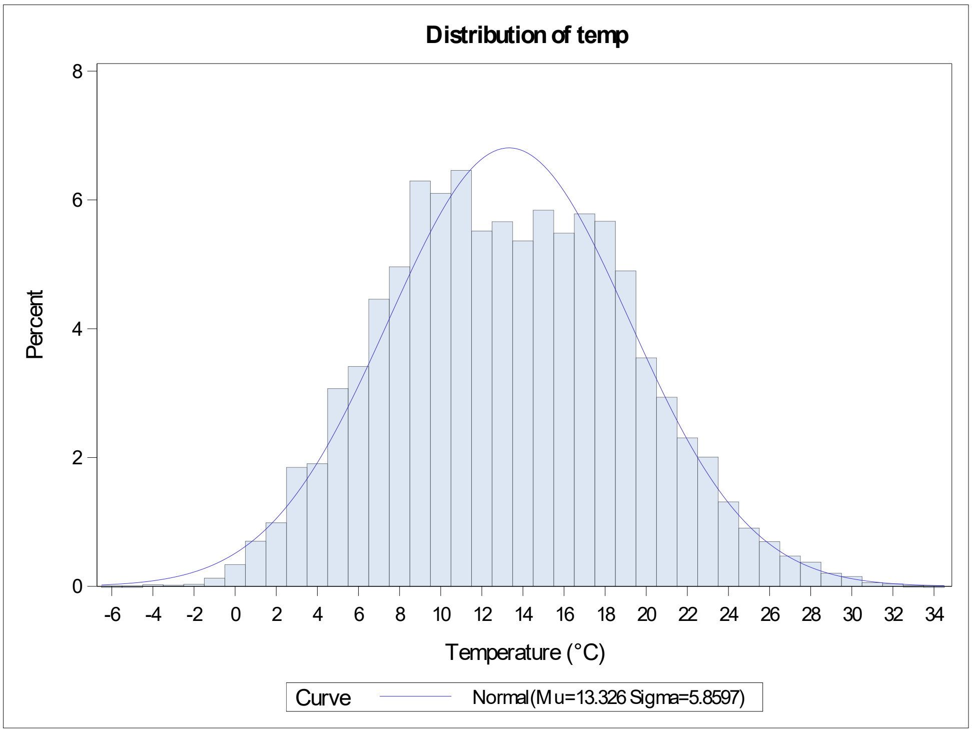

Figure 4 - Histogram for Temperatures in Manchester using UNIVARIATE Procedure

As we can see from the graphical representation of the dataset, the range of temperatures in Manchester are from -6°C to 34°C, with 13°C being the middle value – which we discovered in the statistical summary above. We can also see from this graph that 11°C is the most frequent temperature in Manchester. In the range from 6°C to 20°C, we can say these are the most frequent temperatures. Using SAS, we can also overlay the normal distribution of the dataset, which shows that most of the values are clustered in the middle and taper off the further away from the centre.

To delve deeper into the dataset, and extract more information, we can use SAS to change the midpoints of the histogram, or more specifically how the midpoints are incremented. The graph above has midpoints that are incremented by 1°C. Below I will produce a histogram with midpoints that increment by .5°C to see if we can gain any more information:

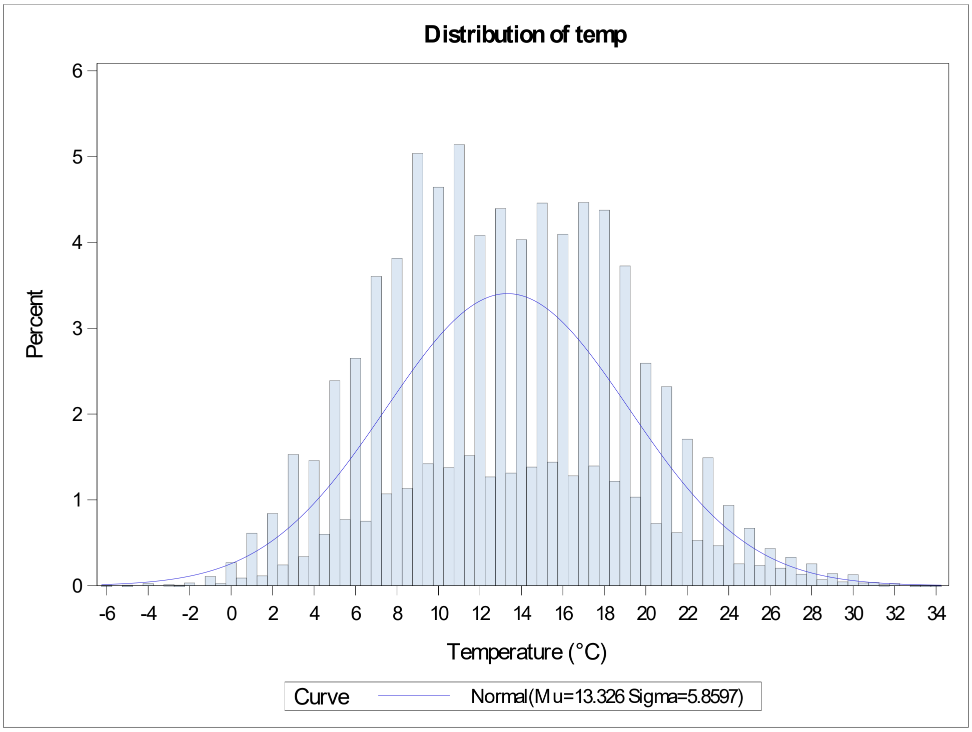

Figure 5 - Histogram for Temperatures in Manchester using UNIVARIATE Procedure

As we can see, the histogram with smaller midpoints follows the same shape as the previous graph, having the most frequent temperatures clustered in the middle and as the temperature moves away from the middle the less frequent the temperatures get. However, the most interesting discovery from the graph is the disparity between temperatures close together. For example, if we look at 10°C, we can see the frequency is roughly 4.5% and 10.5°C is roughly 1.5%. This pattern is common through the distribution. Knowing this knowledge, we could argue that the measurement device that collected the data may not have the accuracy to measure float values and instead measures integer values or that the data has been rounded to the nearest integer.

Both graphs above were created using the UNIVARIATE procedure in SAS, however, this is not the only method of creating a histogram using SAS. We can call the SGPLOT procedure:

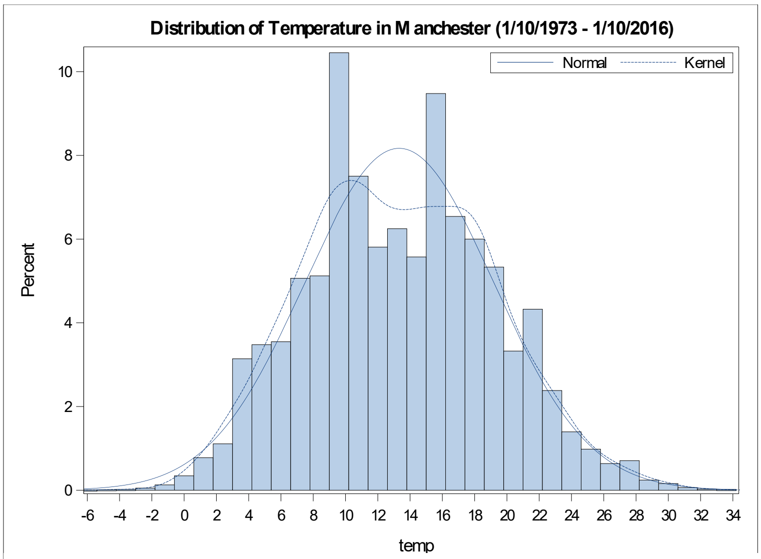

Figure 6 - Histogram for Temperatures in Manchester using SGPLOT Procedure

As shown in the previous two graphs, the temperatures follow somewhat of a normal distribution. Using the SGPLOT procedure in SAS, we can also overlay the Kernel Distribution of the dataset, which is a nonparametric representation of the probability density function of a random variable.

CODE

```
/* Import DATA*/
PROC IMPORT DATAFILE="/folders/myfolders/2/data/man.csv"
     DBMS=CSV
     OUT=man;
     GETNAMES=YES;
RUN;
/* Stat Summary*/
PROC MEANS DATA=man MEAN MEDIAN STD;
    VAR temp;
RUN;
/*Univ Histogram inc by 1 */
PROC univariate DATA=man;
   HISTOGRAM temp / normal
                    ctext = blue
                    midpoints=-6 to 34 by 1
                    ;
Label temp = 'Temperature (°C)';
run;
/*Univ Histogram inc by .5 */
PROC univariate DATA=man;
   HISTOGRAM temp / normal
                    ctext = blue
                    midpoints=-6 to 34 by .5
                    ;
Label temp = 'Temperature (°C)';
run;
/*sgp Histogram inc by 0.2 + Kernal */
proc sgplot data=man;
     histogram temp;
     xaxis values=(-6 to 34 by 0.2);
     density temp;
     density temp / type=kernel;
     keylegend / location=inside position=topright;
     Title 'Distribution of Temperature in Manchester (1/10/1973 -
1/10/2016)';
run;
```

CONCLUSION

In this assignment, we were tasked at representing two dataset using various statistical methods to gain a deeper understanding of the data, essentially transforming data into information by giving it contextual meaning. The themes of this week's assignment were distribution, trying to see if the data was compactly or sparsely distributed. Not only could we represent the data to see the size of the distribution, but we were also able to see the shape.

In Question 1, we were able to compare the distribution of measurements for each lab which allowed us to conclude which lab was the most accurate as well as the lab that was the least accurate by comparing the box plots of each lab's dataset, seeing the size of the distribution.

In Question 2, we were able to compare the distribution of temperatures in Manchester from 1/10/1973 to 1/10/2016 using Histograms to see not only the size of the distribution but also the shape. We were successful in seeing where the majority of the temperatures were clustered and were able to suggest that it followed a normal distribution – having the majority of frequencies of temperatures clustered in the middle.

SOURCES

Giagos, V. and Shea, B. (2018). *Descriptive Statistics*. Manchester: Manchester Metropolitan University.

Lean Sigma Corporation. (2018). *Box Plot with Minitab - Lean Sigma Corporation*. [online] Available at: https://www.leansigmacorporation.com/box-plot-with-minitab/ [Accessed 31 Oct. 2018].

Support.sas.com. (2018). *Base SAS(R) 9.2 Procedures Guide*. [online] Available at: http://support.sas.com/documentation/cdl/en/proc/61895/HTML/default/viewer.htm#a002473539.htm [Accessed 31 Oct. 2018].

Support.sas.com. (2018). *Base SAS(R) 9.4 Procedures Guide: Statistical Procedures, Second Edition*. [online] Available at: http://support.sas.com/documentation/cdl/en/procstat/66703/HTML/default/viewer.htm#procstat_univariate_overview.htm [Accessed 31 Oct. 2018].

Support.sas.com. (2018). *SAS/GRAPH(R) 9.2: Statistical Graphics Procedures Guide, Second Edition*. [online] Available at: http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/viewer.htm#sgplot-ov.htm [Accessed 31 Oct. 2018].

Support.sas.com. (2018). *SAS/STAT(R) 9.3 User's Guide*. [online] Available at: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_boxplot_sect014.htm [Accessed 31 Oct. 2018].

www.tutorialspoint.com. (2018). *SAS Box Plots*. [online] Available at: https://www.tutorialspoint.com/sas/sas_boxplots.htm [Accessed 31 Oct. 2018].

www.tutorialspoint.com. (2018). *SAS Histograms*. [online] Available at: https://www.tutorialspoint.com/sas/sas_histograms.htm [Accessed 31 Oct. 2018].