Teacher's Guide

Using a Statistical Procedure To Search A Database For Out of the Ordinary Values (Outliers)

Lesson Description

This lesson will allow your students to perform a yearly peer review of a viable Internet database. The data in the Project Watershed database www.projectwatershed.org, like most Internet material, is not subject to rigid testing to be sure it is absolutely 100% correct. While the chemical testing has a rigorous quality assurance/quality control maintained and performed by adult supervisors, as with any protocol, there may be some problems in the process from the collection to the reporting of data.

In this procedure, we will investigate the values given in the database and using a statistical application we will identify outliers. Outliers by definition appear to be outside of the range of the data. Outliers may be the result of possible input errors or for some good scientific reason they may require further investigation. The outliers may, however, be the product of some interesting event or phenomenon that is worth investigating in another study. The ozone hole at the South Pole was so far outside of the expected values that it was ignored for several years because it was thought to be an outlier. For these reasons, all reported outliers should be investigated.

By applying the statistical formula that locates values outside of the probable range of the collected data, we can determine if the extremes of the ranges are unrealistic

Science Concepts Introduced

We will be performing a peer review of a database collected by our fellow teachers and students. Our task will be to identify values (outliers) for further investigation. This lesson will be posted on the website and be made available to math classes for practice using box plots and 5# summaries to locate outliers. Their peer review will perform quality assurance for the site that is the largest volunteer stream study publicly accessible on the Internet. Stress to students that the values change from year to year and that they perform a service of peer review as they monitor the data in this website.

Process Skills Emphasized

- Internet retrieval of a specific database
- Interpreting data
- Applying statistics
- Quality Assurance
- Computer analysis or other technology
- Cooperative learning
- Oral class reports
- Peer review
- Composing letters to public officials

Technology Used

- Internet
- MS Excel

MS Word

Mini Tab Calculator

TI-80 series calculator

MST Standards

Standard 6: A and B

Standard 7: A and B

Learning Outcomes

Students will be able to:

- Access the www.projectwatershed.org site on the Internet.
- Open a spreadsheet, copy and paste (download) the values they investigate.
- Save the original database and make a backup copy to discover the outliers. That way a good computer protocol is performed. Backup data frequently.
- Sort all data
- Use MS Excel to find the minimum and maximum, median, 1st and 3rd quartiles.
- Create a box plot and find outliers.
- Identify and report the suspected outliers to the Project Watershed website.

Time Requirement

1-2 Periods or a full period in a block-scheduling format.

Instructional Strategies:

Cooperative learning groups
Individual learning
Direct instruction
Class presentations

Background

- This exercise involves the use of five-number Box plots to determine if there are numbers that are far outside of the probable range of accuracy.
- The columns are highlighted and the function tabs can be used to select Median, Max, Min and Q1 and Q3. These numbers will allow students to determine the Outliers.
- Find the Inter-Quartile Range or IQR
- The IQR = Q3-Q1.
- When you find the IQR, multiply your value by 1.5.
- After performing this step, take 1.5 *IQR and add it to Q3 and then subtract it from Q1.
- These values are then the upper and lower limits for outliers for this set of data.
- Any data point that is higher than Q3+ 1.5 *IQR or lower than Q1- 1.5 *IQR will now be considered an outlier.
- Teach with the three samples from Butternut Creek as examples.
- Once students have mastered the process, move to the peer review exercise to study the entire database for outliers.

Assessment

- Continual observation by the teacher provides immediate assessment.
- Cooperative groups will be responsible for presenting their findings to the class.
- Peer review of their work will be performed as a check on their evaluation at the time of their oral presentation.
- Students will also be required to hand in a hard copy of their oral report for grading. (A rubric will be provided to them)
- If outliers are discovered and the material is checked and confirmed by the other students (their peers), a letter will prepared to inform the host of the website.
- Finally, the usual assessment on this knowledge will show up on tests and quizzes.

Rubric for Box Plots and Outliers:

100-point lab: analyze these sets of data (pH, temp, DO, BOD, phosphates, nitrates, chlorides, turbidity, fecal coliform and total dissolved solids); make each analysis worth 10 points for each correct box plot and outlier. Within each analysis expect the following:

A 3.5 in floppy disk will be submitted to hand in:

- a folder on the disk called Project Watershed in which all student documents will be stored.
- a data analysis on each of 10 different variables of the given stream to determine the outlier limits and identify the previously existing outliers.

Students will need to:

1.) Develop a spreadsheet of each variable containing FORMULAS for finding the 5 number summary, the IQR, the 1.5*IQR and finally the outlier limits (upper and lower).

a) 5 # Summary 2pts
b) IQR & 1.5*IQR 2pts
c) Outlier limits 3pts

2.) Create a box-plot for each set of data on a piece of graph paper and denote all existing outliers and outlier limits. Correct explanation of outlier limits is needed to receive full credit for that part of the assignment.

d) Box-plot and outliers 3pts

3) 10 points a piece and 10 variables to study allow credit to be given to each part and add up to a total of 100 points.

Extensions/Options

In this investigation, we assumed that all outliers were the result of data collection and entry errors. Is this a good or bad assumption? Justify your answer.

Key Terms

Outliers
Box Plots
Quality Assurance (QA)
Quality Control (QC)
Database
Parameters

Prerequisite Knowledge

- Firm understanding in the measures of Central Tendency (Mean, Median, Mode)
- Students must have an understanding of how to use any of the acceptable statistical tools, for example calculators or computer programs
- Basic knowledge of water quality parameters

Equipment Needed

A computer with Microsoft Excel
Mini Tab
TI-80 series calculator
Microsoft Excel
Mini Tab
TI-80 series calculator
Handouts

References For Teachers

Moore, David The Basic Practice of Statistics 2nd Edition 2000 W.H Freeman and Company

Websites:

www.projectwatershed.org

References For Students

McClave/Sincich Statistics 9th Edition 2003 Prentice Hall Inc

Handouts

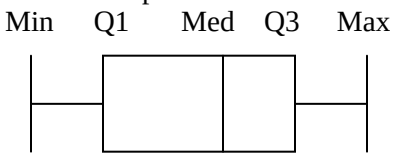
Student Name _____ **Date** _____
Statistics **Chapter 2**

Box-plots / Outlier Project

Directions Go to this site and copy all information stated below into an excel spreadsheet. Today we will be studying the pH, dissolved oxygen and water temperature of Butternut Creek. Copy them from http://watershed.syr.edu/chemdisplay.php .

Step 1: Use the formula options in excel to find the five number summary of each set of data. Remember that the 5 number summaries consist of {Minimum, Quartile 1, Median, Quartile 3, and Maximum}.

Step 2: Using the numbers from your 5 number summary of each variable, make a box-plot on a separate piece of graph paper and completely label it and the axis. Excel cannot make box plots, but software like mini-tab or a graphing calculator can. Use whatever technology you have to complete this task.



Step 3: Using your 5 number summaries, find the Inter-Quartile-Range. Recall that the I.Q.R. is = (Q3 – Q1).

Step 4: Multiply the IQR by 1.5 and write down that number. You now need to add this number to your Q3 and subtract this number from your Q1. This will give you the outlier limits.

Q3 + 1.5*IQR = your upper outlier limit:
Q1 - 1.5*IQR = your lower outlier limit:

This means that given your set of data, any data point above the upper limit or any data point below the lower limit is considered too far away from the median to be considered a good observation. Usually any observation that falls outside these two ranges would be disregarded when you study data.

Step 5: Find the upper and lower outlier limits and find any outliers that may exist in your given data set.

Step 6: Give some reasons why you think we could get outliers in this study. When you look at the sets of data, are there any numbers that don't make sense at all? What should we do with those?

Step 7: Class project and Peer review of the entire database. Break up into groups of three to five students in each group. Select one or more of the nine physical or chemical tests included in the Project Watershed database. Determine the box plots and find any possible outliers in your portion of the database.

Step 8: At the same website each group will select the chemical or physical test they have chosen to evaluate. Once the page has opened select every stream in the Oneida-Seneca-Oswego rivers watershed and hit the submit button. Highlight all of the data by using the Ctrl+A keys, then copy and paste it to your spreadsheet. Follow the same directions to find the outliers.

Step 9: Once outliers are found, make a report to the class.

Step 10: Mail a report to the Host of the website.

Butternut Creek	pH	Dissolved O2 (mg/L)	Water Temp (C)
	4.67	8.2	23.1
	5.89	7.9	13.9
	5.98	12.5	3
	6.02	8.5	11.5
	6.03	11	6.8
	6.15	8.4	20
	6.25	8	2.3
	6.25	9.7	14.2
	6.35	9.7	24
	6.39	5.7	15.7
	6.44	8.1	7.2
	6.5	14.7	16.5
	6.54	11.9	11.2
	6.59	9.2	9.2
	6.7	6.7	0
	6.77	8.9	19.7
	6.79	7.9	17.2
	6.84	7.1	4.4
	7	13	19.8
	7.02	5.8	23.7
	7.24	11	8
	7.31	9.3	6
	7.36	14.1	21
	7.38	6.2	10.9
	7.56	7.8	19
	7.58	7.9	0
	7.59	9.2	5.1
	7.61	10	4.3
	7.65	9	17.5
	7.74	9.8	15.7
	7.77	8.7	17.6
	7.77	10.3	9.1
	7.89	10.9	20.3
	8.09	9.9	8.2
	8.1	9.3	17.6
	8.11	9.5	23.6
	8.12	10.3	4.5
	8.18	8	17.2

	8.19	12.6	14.4
	8.24	8.2	9
	8.26	10.3	9.7
	8.3	10	14
	8.3	9.8	22.5
	8.35	10	4.3
	8.37	10.9	16
	8.55	11.3	
	8.67	11.8	
	8.67	11	
	8.69	13.4	
	8.81		
	8.95		
	9.01		
Min	4.67	5.7	0
Quartile1	6.58	8.20	7.20
Median	7.58	9.7	14
Quartile3	8.20	10.90	17.60
Max	9.01	14.7	24
IQR	**Inter-Quartile Range**		
(Q3 - Q1)	1.63	2.70	10.40
1.5*IQR	2.43	4.05	15.6
Lower Outlier Limit	4.14	4.15	-8.40
Upper Outlier Limit	10.64	14.95	33.20
Denoted Outliers	**None**	**None**	**None**

Student's Guide

Using a Statistical Procedure To Search A Database For Out of the Ordinary Values (Outliers)

Introduction

This lesson will show you how to detect possible outliers out of ordinary values from a given set of data. The method that we will use will introduce and utilize our knowledge of box-plots and 5 number-summaries. The reason for doing this type of investigation is that outliers severely affect the values for the average and standard deviations of sets of data.
This website is a summary of information accumulated by students, their teachers and volunteers from our CNY region who have worked to produce a database that represents a summary of water quality in streams. This database is the largest volunteer stream study available in New York State that is publicly accessible through the Internet.

This is a dynamic database that has yearly input with monitoring by students in the area. Your investigation is part of an ongoing process of quality assurance (QA) for this website. We need to consistently identify outliers in order to get as close as possible to the real story the data are supposed to tell us. By applying the statistical formula that locates values outside of the probable range of the collective data, we can determine if the extremes of the ranges are unrealistic.

The outliers may, however, be the product of some interesting event or phenomenon that is worth investigating in another study. The ozone hole at the South Pole was so far outside of the expected values that it was ignored for several years because it was thought to be an outlier. For these reasons, all reported outliers will be investigated.

Learning Outcomes

Students will be able to:

- Access the www.projectwatershed.org site on the Internet.
- Open a spreadsheet, download the values
- Save the original database and make a backup copy. Backup data frequently.
- Sort all data
- Use MS Excel to find the minimum and maximum, median, 1st and 3rd quartiles.
- Create a box plot and find outliers.
- Identify and report the suspected outliers to the Project Watershed website.

Skills Required

- Use computer technology of choice to analyze a data set.
- Work cooperatively in groups.
- Make a short class presentation
- Write and send an informative letter

New Terms

- Quartile
- Parameter
- Quality Assurance (QA)
- Mean

- Median
- Mode
- Box-plot
- Outlier
- Upper Limit
- Lower Limit
- 5# Summary

Quest

Actually you are a cyberspace bounty hunter (a predator) in the habitat of the Project Watershed website. Your prey are the outliers that are in the data on this website. Your hunting techniques will be to use computer technology to stalk the prey. I would suggest a manufacturers suggested retail price (MSRP) of $1000.00 Dead or Alive for a bounty on each outlier you find. Of course, he MSRP will carry little weight with your teacher. You can negotiate the final payment for each outlier.

Materials

- A computer with Microsoft Excel
- Mini Tab Calculator
- TI-80 series calculator
- Microsoft Excel
- Handouts

Procedure

1. Divide into groups of three to five students per group. The exercise will use Butternut Creek: dissolved oxygen, temperature and pH.
2. Use some form of mathematical technology (ex. MS Excel) to find the 5-number summaries. That summary consists of: minimum value, Q1, median, Q3, maximum value.
3. Find what is called the Inter-Quartile Range or IQR. The IQR = Q3-Q1.
4. When you find the IQR, multiply your value by 1.5.
5. After you get this value, take 1.5 X IQR and add it to Q3, and then subtract it from Q1.
6. These values are your upper and lower limits for outliers for this set of data.
7. Any data point that is higher than Q3 + 1.5 IQR or lower than Q1- 1.5 X IQR will now be considered an outlier.
8. Access the www.projectwatershed.org site on the Internet.
9. Select a set of data from the website database.
10. Download the values into an open spreadsheet.
11. Save the original database and make a backup copy to discover the outliers. Follow good computer protocol. Backup your data frequently.
12. Follow the directions using MS Excel (or other technology of choice) to find the minimum and maximum, median, 1st and 3rd quartiles (the five numbers for the box plot.)
13. Create a box plot and identify outliers.
14. Report the outliers to the host website so they can investigate.
15. Once you have set up the box plot for a set of data, it will be relatively easy to divide the class into smaller groups and each group will be responsible for finding outliers in one or more of the physical and chemical data sets.
    - Water temperature
    - Dissolved oxygen
    - Turbidity
    - Nitrates
    - Phosphates

- Biochemical Oxygen Demand,
- pH
- Chlorides
- Total dissolved solids
- Fecal Coliform

16. Report your information to the class for a class peer review of your 5 number summaries. If you have found outliers and the class agrees with your results, submit that information to the website host.

Extensions/Options

- In this investigation we assumed that all outliers were the result of data collection and entry errors. Is this a good or bad assumption? Justify your answer.

Assessment

- Continual observation with Q and A by the teacher will provide initial assessment.
- Cooperative groups will be responsible for presenting their findings to the class.
- Students will be required to hand in a hard copy of their report for grading. (See rubric below.)
- A separate letter should be prepared to notify Project Watershed that you have discovered a possible outlier, and they can review the data to verify the results and investigate the problem to see if it is valid.
- Finally, the usual assessment on this knowledge will show up on tests and quizzes.

Rubric

100 Point lab exercise will be broken down this way: Students will be analyzing these sets of data (pH, temp, DO, BOD, phosphates, nitrates, chlorides, turbidity, fecal Coliform and total dissolved solids). Each analysis is worth up to 10 points. Within each piece the following will be expected:

1. A 3.5 in floppy disk to hand in.
2. A folder on the disk called Project Watershed and in this folder is where all documents will be stored.
3. 10 different data analysis experiments on each variable of your given stream to determine the outlier limits and remove any already existing outliers.

You will also need to:

4. Create a spreadsheet of each variable containing FORMULAS for finding the 5 number summary, the IQR, the 1.5*IQR and finally the outlier limits (upper and lower).
    a) 5 # Summary **2pts**
    b) IQR & 1.5*IQR **2pts**
    c) Outlier limits **3pts**
5. Create a box-plot for each set of data on a piece of graph paper and there denote all the existing outliers and outlier limits. Correct explanation of outlier limits is needed to receive full credit for that part of the assignment.
    d) Box-plot and outliers **3pts**

A total of 100 points will be divided into 10 points for each of the 10 variables. All groups will be responsible for performing box plots and discovering outliers if they can be found. Reports will be given and compared to other groups work.

Handouts

Box plots/Outlier project