Oklahoma Department of Environmental Quality

STATISTICAL DATA FOR HISTORICAL GROUNDWATER QUALITY FROM 1897 TO 1988

FY05/06 106 CARRY OVER, I-006400-05, Project 02
Statistical Evaluation of Public Water Supply Data
Final Data Analysis Report

Prepared by:
Michael S. Houts

December 30, 2004

Executive Summary

The Department of Environmental Quality combined historical data from 1897 to 1988 collected at public water supply wells into a database for analysis. Using a commercial software package called Minitab, basic statistical analysis was performed on a limited set of sixteen parameters.

Analysis of data gathered from historical information that is supported with limited quality assurance information is challenging. None of the data are normally distributed. This is a result of the data being generated as compliance information from public water supply wells and the fact that many results exceed analytical reporting limits. Concentrations of contaminants do exist below the reporting limit, however, they are not quantifiable. If every concentration were quantified (no "censored" data), the resultant distributions would likely be much more normal. Outliers on the high end would still exist, however. Data are skewed toward the lower concentration levels of chemical constituents. One would expect the same for irrigation, domestic, and industrial water wells, as they also are drilled and completed in the major "fresh" water aquifers. There are many wells that have "outliers" or extreme values for chemical constituents; however, the majority of test results fall within normal, or bell-curve, limits for that constituent.

The following determination can be drawn from the project. The median value for aluminum in most of this data set exceeds the secondary water quality standard. Laboratory reporting limits exceed the secondary drinking water standard. If laboratory reporting limits exceed water quality standards, that does not necessarily mean that a sample reported as less than the reporting limit exceeds Water Quality Standards. It simply means that the value is less than the reporting limit. Nitrate values for the median are at the maximum contaminate level (MCL) of ten mg/L in this dataset. Historical data may have not measured this as N but as NO_3. The MCL historically was 45 mg/L for NO_3. Median values for all other constituents were below the MCL or had no MCL. Based on these findings we can conclude that groundwater in major aquifers in Oklahoma is generally of good quality.

Introduction

The Department of Environmental Quality has been collecting data from historical ground water quality monitoring projects for several years. In 2001, we began the process of organizing this historical information into a database using Microsoft Access. This database has been compiled and was used as the data source for the preparation of this report. Using commercially available software packages, basic statistical analyses were performed on certain chemical parameters.

Water Quality Division staff believe that the quality of ground water aquifers across the state would be determined through this effort. The vast majority of the data was collected from water supply wells. Most of these wells produce water that meets the primary and secondary water quality standards for public drinking water supplies.

The statistical analysis was completed using MINITAB statistical software. Mean, median, standard deviation, confidence level and outliers were calculated. Outliers are defined in the boxplot. The data are not normally distributed. The distribution is skewed because the historical data were generated only from drinking water wells and not from the entire population of wells. Non-normal distributions also are due to analytical reporting limits. Concentrations of contaminants do exist below the reporting limit, however, they are not quantifiable. If every concentration were quantified (no "censored" data), the resultant distributions would likely be much more normal. Outliers on the high end would still exist, however.

Disclaimer

Information from this data is to be used with the understanding that these analyses were from different decades and were gathered with varying methods. Statistical bias could present a problem because the wells were not randomly selected as we included all of the available historical data. Comparisons may be less than desired because of the historical differences in data source and test sensitivity. The mean is not the best method for data comparison because of the large number of outliers. The median is a better measure of central tendency. Analysis of data gathered from historical information with minimum quality assurance information is a challenge.

The historical data were derived by water analysis methods that were subjective because the technique of the analyst had significant impact on the results. Analysis methods varied from titration to colorimetric to modern chromatographic methods. The quality assurance methods associated with these historic analyses are not reported in most instances and may not exist. Therefore, care must be taken in using this information and the conclusions drawn.

Project Objectives

The project objectives were:

(a) Utilize the Public Water Supply water quality information from historical data sets.
(b) Based on basic statistical analysis of the historical data, develop some general statements about ground water quality in Oklahoma.
(c) Compile information into a database and a report of the analyses.

The Department of Environmental Quality (DEQ) selected test result data from published reports produced by highly regarded organizations such as the Oklahoma State Department of Health, Oklahoma Geological Survey, Oklahoma Agricultural and Mechanical College, and the Oklahoma Planning and Resources Board. The data used were collected and analyzed with the best available and accepted methods of the time. Samples were from a single well as opposed to "system" samples, which now are collected for drinking water standards compliance. Copies of the reports are available from DEQ.

Methodology

Selected parameters for measurement:

Sixteen chemical and physical parameters that reflect secondary drinking water standards and major ions were selected for statistical analysis.

1. Aluminum
2. Chloride
3. Copper
4. Fluoride
5. Iron
6. Manganese
7. NO2+NO3
8. Sulfate
9. Total Dissolved Solids
10. Zinc
11. Calcium
12. Magnesium
13. Sodium
14. Total Alkalinity
15. Carbonate Alkalinity
16. Potassium

Explanation of the Graphical Summary

Each graphical summary in the following section includes a histogram that summarizes the data, a box plot, a confidence interval graph, and an overall statistical summary for the given parameter, all presented within the same window. The following information is copied from Minitab software help pages and is paraphrased in places.

Histogram of Data with Normal Curve

The histogram of the data is overlaid with a normal curve to assess the normality of the data. A normal distribution is symmetric and bell-shaped.

Boxplot

- Boxplots summarize information about the shape, dispersion, and center of the data. They also can help identify outliers.

- The left edge of the box represents the first quartile (Q1), while the right edge represents the third quartile (Q3). Thus the box portion of the plot represents the interquartile range (IQR), or the middle 50% of the observations.

- The line drawn through the box represents the median of the data.

- The lines extending from the box are called whiskers. Whiskers extend outward to indicate the lowest and highest values in the data set (excluding outliers).

- Extreme values, or outliers, are represented by dots. A value is considered an outlier if it is outside of the box (greater than Q3 or less than Q1) by more than 1.5 times the IQR.

- Use the boxplot to assess the symmetry of the data:

    - If the data are fairly symmetrical, the median line will be roughly in the middle of the IQR box and the whiskers will be similar in length.

    - If the data are skewed, the median may not fall in the middle of the IQR box, and one whisker will likely be noticeably longer than the other.

Confidence Intervals for Mean, Standard Deviation, and Median

A confidence interval is an interval used to estimate a population parameter from sample data. The upper and lower bounds of the confidence intervals for m (mu), s (standard deviation), and the median are displayed in the graphical summary. In addition, the confidence intervals for m and the median are displayed graphically.

Confidence intervals are composed of two basic parts:

- Point estimate - a single value computed from the sample data. This value is considered to be an estimate of the parameter of interest. However, it is unlikely that the point estimate is equal to the parameter. Therefore, to account for the possibility of estimation error, the error margin is included in the confidence interval to provide a range of possible parameter values.

- Error margin - determines the width of the confidence interval through the use of probability. To construct the confidence interval, add and subtract the error margin from the point estimate.

If a 95% confidence interval is selected, the method used to construct the interval has a probability of 0.95 of producing an interval containing the parameter of interest. In other words, you can be 95% confident that the true value of the parameter is within the interval. Thus, if one hundred 95% confidence intervals were constructed, you would expect around 95 of the intervals to contain the parameter.

Table of Statistics

Anderson-Darling Normality Test

The Anderson-Darling normality test can help determine whether the data follow a normal distribution. The "A statistic" that the test provides is used to determine the p-value, which ranges from 0 to 1, and indicates how likely it is that the data follows a normal distribution.

First, it must be determined how low the p-value must be to conclude that the data are not normal. A commonly chosen criterion is 0.1. If the p-value is lower than the criterion, it must be concluded that the data do not follow a normal distribution. Otherwise, there is not enough evidence to conclude that the data do not follow a normal distribution.

Mean and N

Mean

The mean, also called the average, is a measure of where the center of the distribution lies. The mean is determined by calculating the sum of all observations divided by the number of observations. The mean is strongly influenced by extreme values (outliers).

N

N is the number of non-missing values in the data set.

Standard Deviation (StDev) and Variance

The standard deviation and variance measure dispersion, or how far the observations in a sample deviate from the mean. The standard deviation is analogous to an average distance (independent of direction) from the mean. The variance is the standard deviation squared. Like the mean, the standard deviation (as well as the variance) is very sensitive to extreme values.

Skewness and Kurtosis

Skewness

Skewness refers to a lack of symmetry within a data set. A distribution is skewed if one tail, or outer portion of the bell-curve, extends farther than the other. A skewness statistic is provided with the graphical summary:

- A value close to 0 indicates symmetrical data.
- Negative values imply negative/left skew.
- Positive values indicate positive/right skew.

Kurtosis

Kurtosis refers to the peak sharpness of a distribution curve. A kurtosis statistic is provided with the graphical summary:

- Values close to 0 indicate normally peaked data.
- Negative values indicate a distribution peak that is flatter than normal.
- Positive values indicate a distribution with a sharper than normal peak.

Minimum and Maximum

One of the easiest ways to assess dispersion within a data set is to compare the minimum and maximum values. The minimum is the smallest value in a data set, and the maximum is the largest value.

Minimum and maximum values are used to calculate the range, which is a statistic often used to describe dispersion within data sets. The range is the (Maximum)-(Minimum).
The range is very sensitive to extreme values.

First and Third Quartiles (Q1 and Q3)

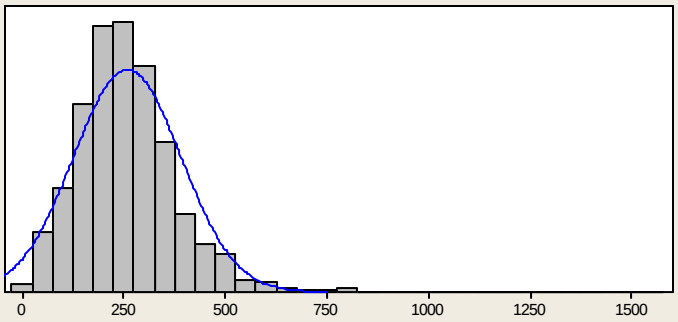
The first quartile (Q1, also called the 25th percentile) is the highest value for the lowest 25% of the observations. The third quartile (Q3, also called the 75th percentile) is the lowest value for the highest 25% of the observations. Q1 and Q3 are often used to calculate the interquartile range (IQR), which also is used to describe dispersion. The IQR is the range of the middle 50% of the values, and is calculated by subtracting Q3 from Q1. The IQR is relatively insensitive to extreme values.

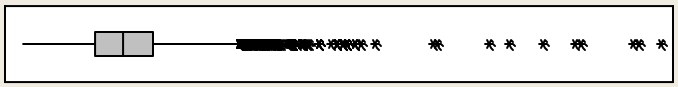
Median

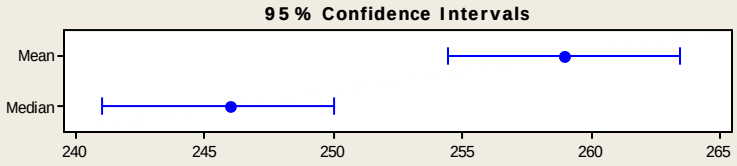
The median, also called the 2nd quartile or 50th percentile, is the middle observation of the data set. The median is determined by ranking the data and locating observation number ([N + 1] / 2). If there are an even number of observations, the median is extrapolated as the value midway between that of observation numbers (N / 2) and ([N / 2] + 1).

The median is less sensitive to extreme values than the mean. Therefore, the median is often used instead of the mean when data contain outliers, or are skewed.

Summary for ALKALINITY







Anderson-Darling Normality Test	
A-Squared	39.44
P-Value <	0.005

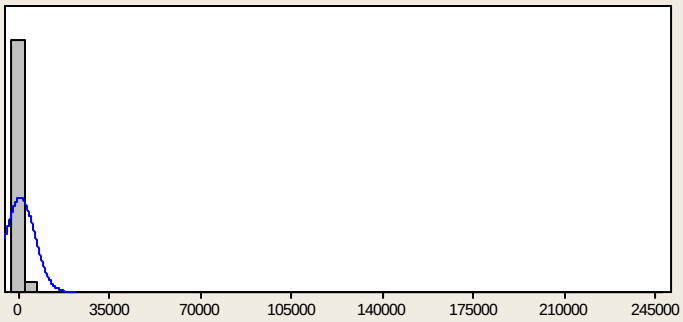
Mean	258.93
StDev	132.30
Variance	17504.37
Skewness	2.1736
Kurtosis	13.8770
N	3318

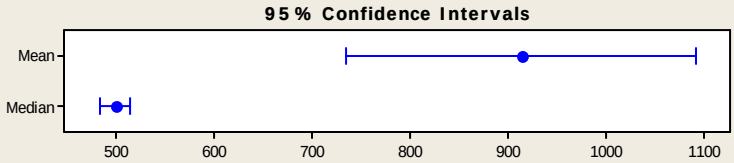
Minimum	2.00
1st Quartile	177.75
Median	246.00
3rd Quartile	320.00
Maximum	1569.00

95% Confidence Interval for Mean	
254.43	263.44
95% Confidence Interval for Median	
241.05	250.00
95% Confidence Interval for StDev	
129.20	135.57

3/1/2004; Year 1 GW Monitoring Project

Summary for TOTAL_SOLI





Anderson-Darling Normality Test	
A-Squared	1140.09
P-Value <	0.005

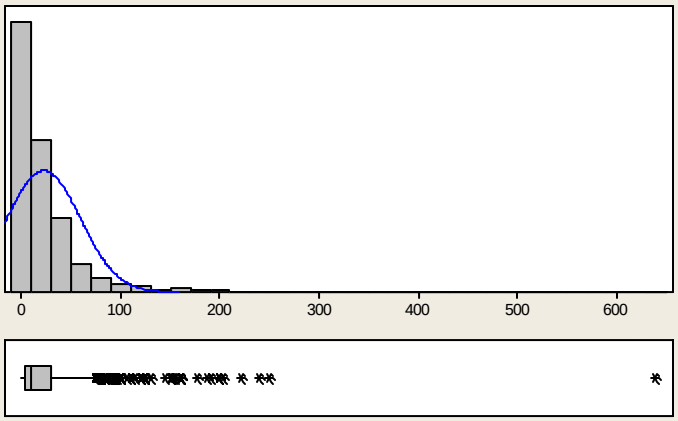
Mean	912
StDev	5566
Variance	30981413
Skewness	38.23
Kurtosis	1570.90
N	3750

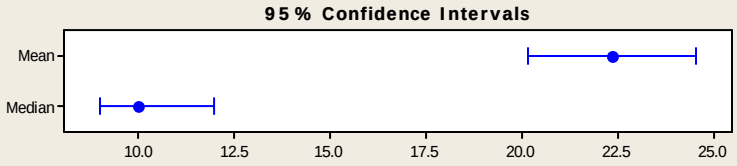
Minimum	0
1st Quartile	330
Median	499
3rd Quartile	789
Maximum	246428

95% Confidence Interval for Mean	
734	1090
95% Confidence Interval for Median	
483	513
95% Confidence Interval for StDev	
5443	5695

3/1/2004; Year 1 GW Monitoring Project

Summary for NO3_PPM

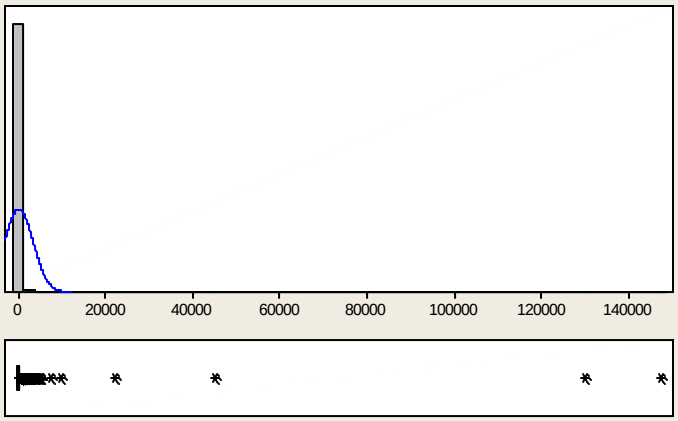


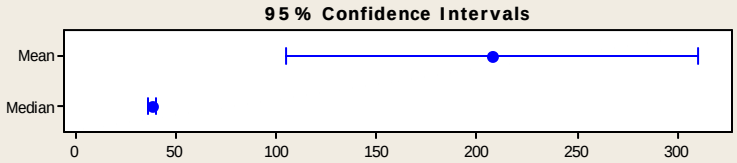


Anderson-Darling Normality Test	
A-Squared	115.62
P-Value <	0.005
Mean	22.339
StDev	36.935
Variance	1364.185
Skewness	6.2352
Kurtosis	77.2005
N	1083
Minimum	0.100
1st Quartile	2.500
Median	10.000
3rd Quartile	30.000
Maximum	638.000
95% Confidence Interval for Mean	
20.137	24.541
95% Confidence Interval for Median	
9.000	12.000
95% Confidence Interval for StDev	
35.442	38.560

3/2/2004; Year 1 GW Monitoring Project

Summary for CL_PPM

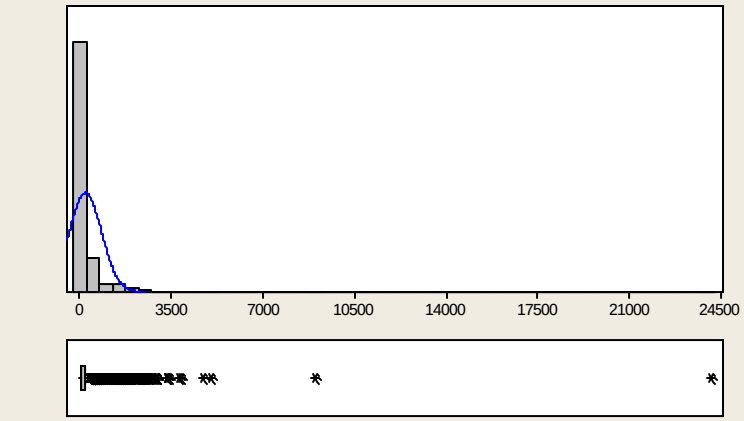


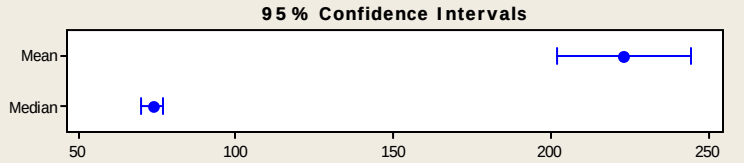


Anderson-Darling Normality Test	
A-Squared	1362.90
P-Value <	0.005
Mean	207
StDev	3270
Variance	10694750
Skewness	40.28
Kurtosis	1699.05
N	3879
Minimum	1
1st Quartile	19
Median	38
3rd Quartile	103
Maximum	147100
95% Confidence Interval for Mean	
105	310
95% Confidence Interval for Median	
36	40
95% Confidence Interval for StDev	
3199	3345

3/2/2004; Year 1 GW Monitoring Project

Summary for SO4_PPM





Anderson-Darling Normality Test	
A-Squared	658.44
P-Value <	0.005

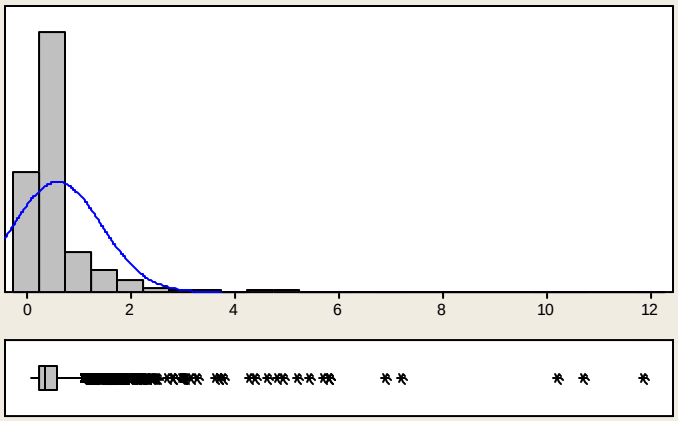
Mean	223.1
StDev	614.7
Variance	377805.6
Skewness	20.252
Kurtosis	714.882
N	3262

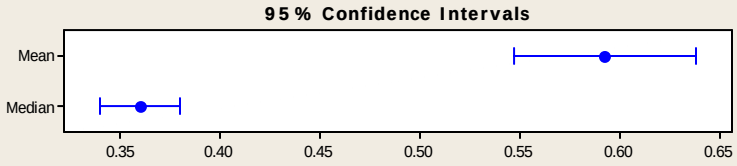
Minimum	0.2
1st Quartile	36.0
Median	73.5
3rd Quartile	168.3
Maximum	24096.0

95% Confidence Interval for Mean	
202.0	244.2
95% Confidence Interval for Median	
70.0	77.0
95% Confidence Interval for StDev	
600.1	629.9

3/2/2004; Year 1 GW Monitoring Project

Summary for FLUORIDE

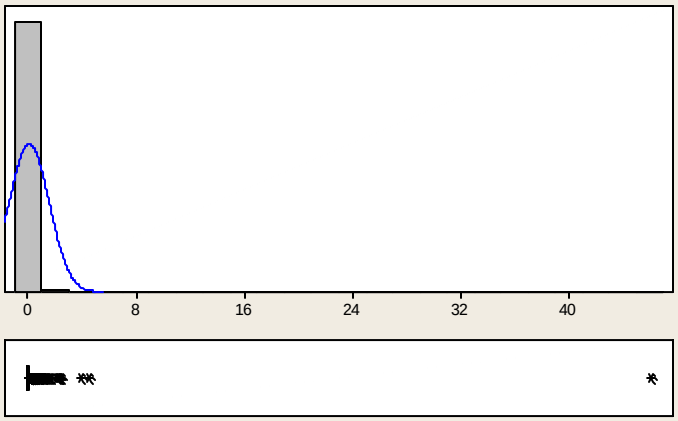


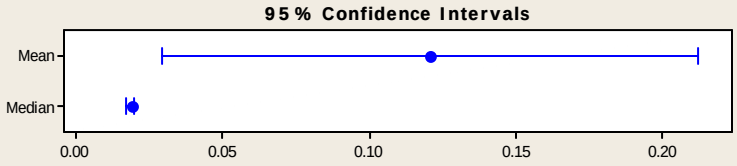


Anderson-Darling Normality Test	
A-Squared	190.99
P-Value <	0.005
Mean	0.5924
StDev	0.8470
Variance	0.7174
Skewness	6.4961
Kurtosis	61.3960
N	1333
Minimum	0.1000
1st Quartile	0.2400
Median	0.3600
3rd Quartile	0.5800
Maximum	11.8500
95% Confidence Interval for Mean	
0.5469	0.6379
95% Confidence Interval for Median	
0.3400	0.3800
95% Confidence Interval for StDev	
0.8160	0.8804

3/1/2004; Year 1 GW Monitoring Project

Summary for COPPER

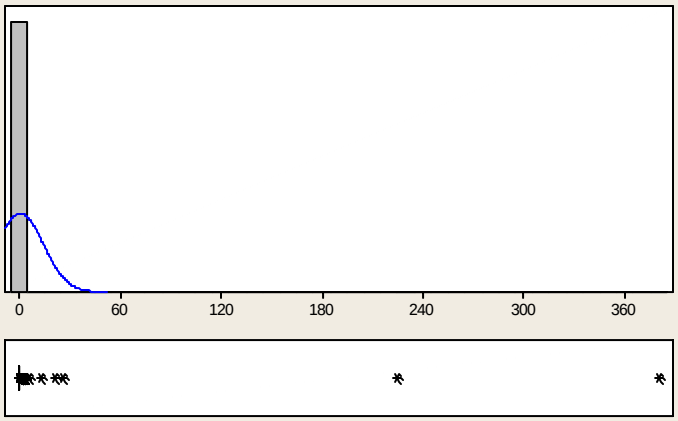


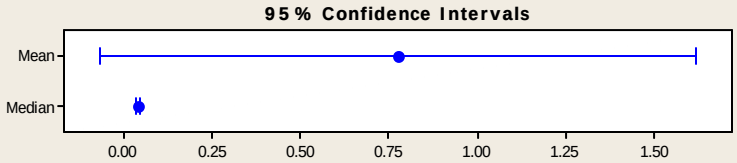


Anderson-Darling Normality Test	
A-Squared	343.35
P-Value <	0.005
Mean	0.1205
StDev	1.4736
Variance	2.1716
Skewness	30.123
Kurtosis	936.716
N	1007
Minimum	0.0040
1st Quartile	0.0100
Median	0.0190
3rd Quartile	0.0470
Maximum	46.0000
95% Confidence Interval for Mean	
0.0293	0.2116
95% Confidence Interval for Median	
0.0170	0.0200
95% Confidence Interval for StDev	
1.4120	1.5410

3/2/2004; Year 1 GW Monitoring Project

Summary for ZINC

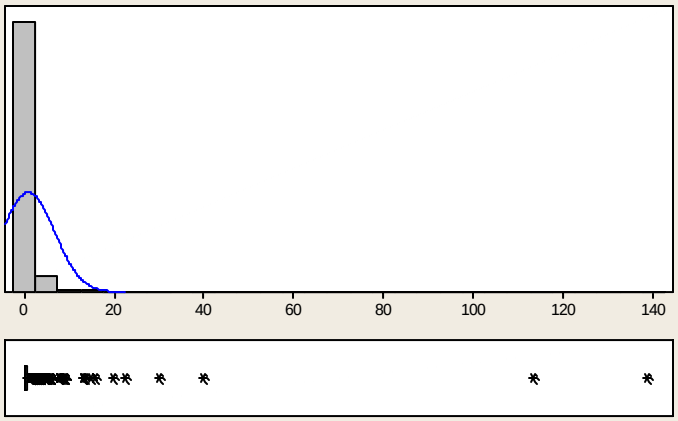


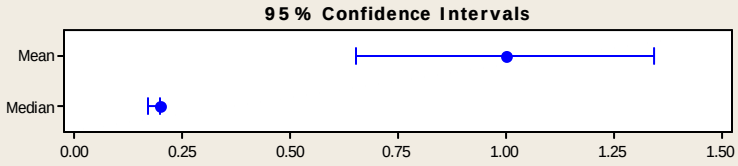


Anderson-Darling Normality Test	
A-Squared	382.29
P-Value <	0.005
Mean	0.775
StDev	13.789
Variance	190.146
Skewness	24.425
Kurtosis	623.889
N	1031
Minimum	0.004
1st Quartile	0.016
Median	0.040
3rd Quartile	0.104
Maximum	380.000
95% Confidence Interval for Mean	
-0.068	1.618
95% Confidence Interval for Median	
0.035	0.045
95% Confidence Interval for StDev	
13.219	14.412

3/2/2004; Year 1 GW Monitoring Project

Summary for FE

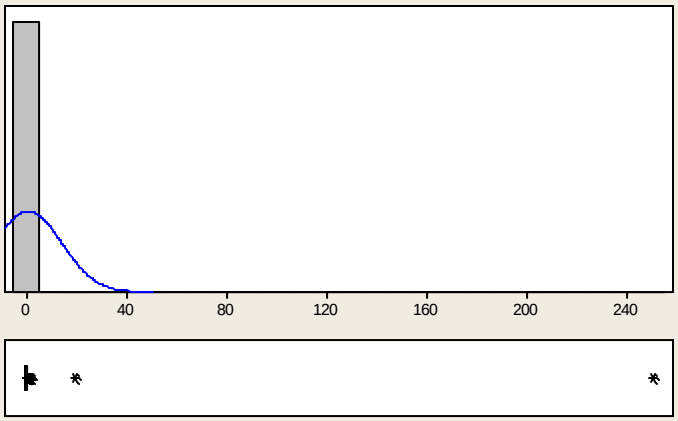


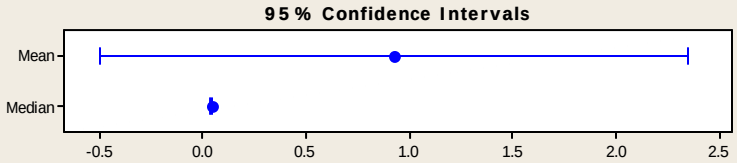


Anderson-Darling Normality Test	
A-Squared	310.30
P-Value <	0.005
Mean	0.998
StDev	5.831
Variance	33.999
Skewness	19.123
Kurtosis	411.052
N	1090
Minimum	0.008
1st Quartile	0.100
Median	0.200
3rd Quartile	0.600
Maximum	138.500
95% Confidence Interval for Mean	
0.652	1.345
95% Confidence Interval for Median	
0.170	0.200
95% Confidence Interval for StDev	
5.596	6.087

3/1/2004; Year 1 GW Monitoring Project

Summary for MN

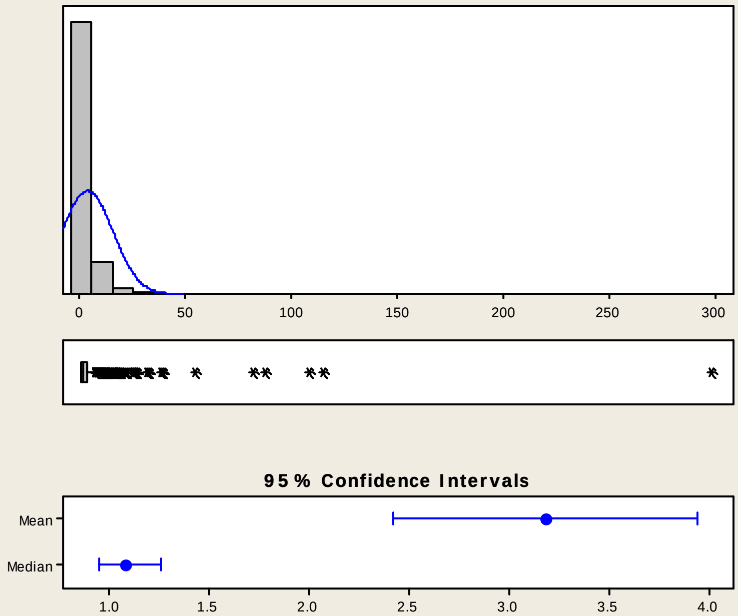




Anderson-Darling Normality Test	
A-Squared	128.17
P-Value <	0.005
Mean	0.923
StDev	13.455
Variance	181.027
Skewness	18.451
Kurtosis	342.371
N	347
Minimum	0.013
1st Quartile	0.020
Median	0.050
3rd Quartile	0.140
Maximum	250.000
95% Confidence Interval for Mean	
-0.498	2.344
95% Confidence Interval for Median	
0.040	0.050
95% Confidence Interval for StDev	
12.523	14.538

3/1/2004; Year 1 GW Monitoring Project

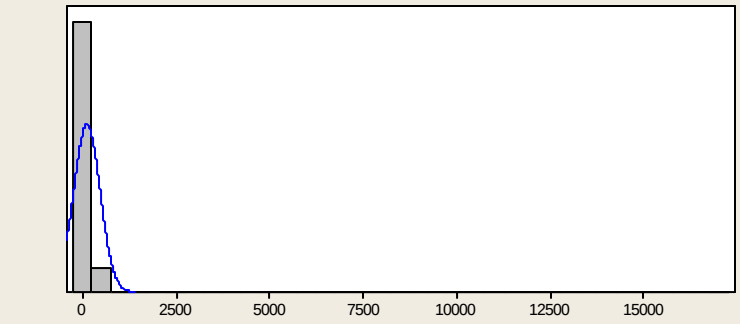
Summary for Al

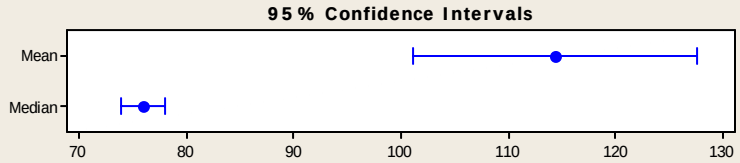


Anderson-Darling Normality Test	
A-Squared	229.38
P-Value <	0.005
Mean	3.183
StDev	12.098
Variance	146.360
Skewness	17.104
Kurtosis	376.587
N	977
Minimum	0.063
1st Quartile	0.378
Median	1.071
3rd Quartile	2.772
Maximum	297.593
95% Confidence Interval for Mean	
2.423	3.942
95% Confidence Interval for Median	
0.945	1.260
95% Confidence Interval for StDev	
11.584	12.660

Worksheet: Worksheet 10; 5/13/2004; Year 1 GW Monitoring Project

Summary for CA_PPM





Anderson-Darling Normality Test	
A-Squared	584.73
P-Value <	0.005

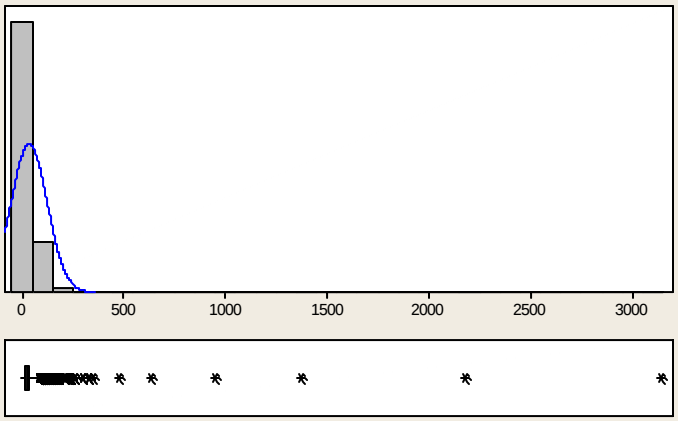
Mean	114.2
StDev	349.9
Variance	122440.0
Skewness	42.06
Kurtosis	2020.26
N	2695

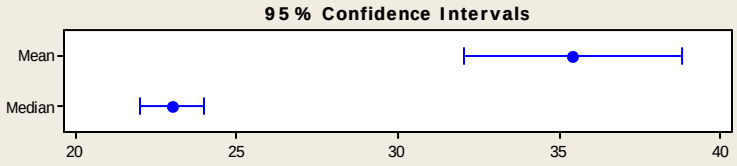
Minimum	0.3
1st Quartile	45.0
Median	76.0
3rd Quartile	115.0
Maximum	17010.0

95% Confidence Interval for Mean	
101.0	127.5
95% Confidence Interval for Median	
74.0	78.0
95% Confidence Interval for StDev	
340.8	359.5

3/2/2004; Year 1 GW Monitoring Project

Summary for MG_PPM





Anderson-Darling Normality Test	
A-Squared	495.57
P-Value <	0.005

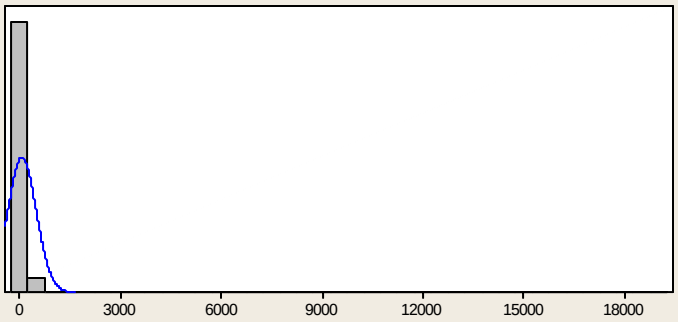
Mean	35.41
StDev	88.07
Variance	7756.20
Skewness	23.910
Kurtosis	732.940
N	2658

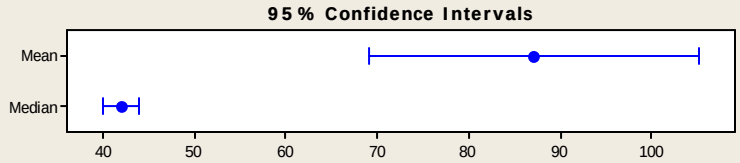
Minimum	0.10
1st Quartile	12.00
Median	23.00
3rd Quartile	39.00
Maximum	3135.00

95% Confidence Interval for Mean	
32.06	38.76
95% Confidence Interval for Median	
22.00	24.00
95% Confidence Interval for StDev	
85.76	90.50

3/2/2004; Year 1 GW Monitoring Project

Summary for NA_PPM





Anderson-Darling Normality Test	
A-Squared	545.43
P-Value <	0.005

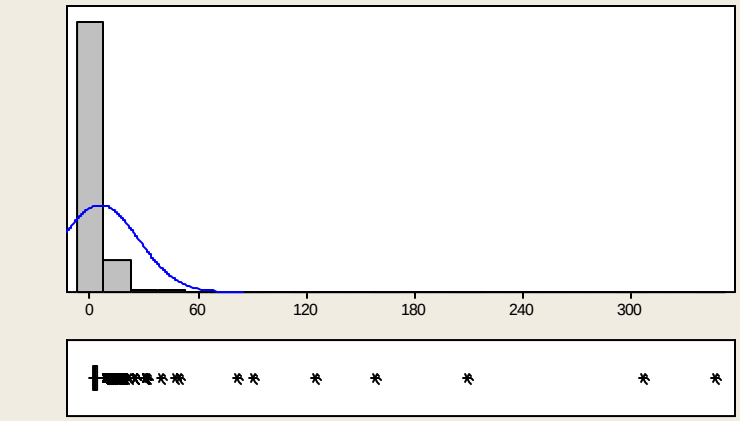
Mean	87.1
StDev	422.0
Variance	178109.3
Skewness	42.21
Kurtosis	1878.82
N	2101

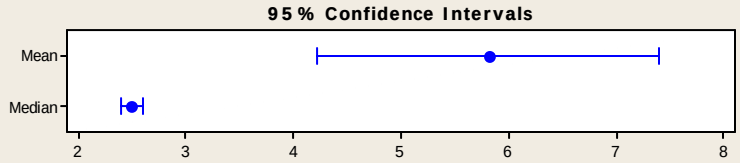
Minimum	2.0
1st Quartile	23.0
Median	42.0
3rd Quartile	96.5
Maximum	18890.0

95% Confidence Interval for Mean	
69.0	105.1
95% Confidence Interval for Median	
40.0	44.0
95% Confidence Interval for StDev	
409.6	435.2

3/2/2004; Year 1 GW Monitoring Project

Summary for K_PPM





Anderson-Darling Normality Test	
A-Squared	186.74
P-Value <	0.005
Mean	5.812
StDev	21.291
Variance	453.304
Skewness	12.142
Kurtosis	166.605
N	697
Minimum	0.200
1st Quartile	1.600
Median	2.500
3rd Quartile	4.500
Maximum	346.000
95% Confidence Interval for Mean	
4.228	7.395
95% Confidence Interval for Median	
2.400	2.600
95% Confidence Interval for StDev	
20.229	22.472

3/2/2004; Year 1 GW Monitoring Project

Conclusions

Although groundwater constituents in Oklahoma are generally lower than the MCL for a specific constituent, any particular well may contain water that exceeds the MCL.

This project was focused on basic statistical analysis of historical public water supply chemical analyses. The purpose of this study was to provide a general look at historical ground water quality data, to provide this information to others with some interpretation that makes it understandable, and to provide some basis to look at past and present conditions in ground water. The available information was accumulated and organized in a database. The information is summarized in this final report and will be made available to the citizens of Oklahoma.

Using commercially available software packages, basic statistical information (mean, median, standard deviation, confidence level) about some chemical analyses were calculated. The median value for aluminum exceeds the secondary water quality standard. This is due to detection limits for the analysis and the calculations needed to convert Al2O3 to Al. Nitrate values for the median are at the maximum contaminate level (MCL) of ten mg/L. Historical data may have not measured this as N but as NO3. The MCL historically was 45 mg/L for NO3. All other median values were below the MCL or had no MCL.

There is a disclaimer in the report explaining the many challenges associated with this type of data accumulation and analysis. Analysis methods varied from titration to colorimetric to modern chromatographic methods. The quality assurance methods associated with these historic analyses are not reported in most instances and may not exist. We assumed that the information is at least accurate enough to be correct 80% of the time. For purposes of this data gathering effort, we assumed the modern data is of equal quality as the historical information. For Water Quality Division staff, the decision is based on the question, “does the water meet primary drinking water standards and secondary standards?” Because the analyses show that the median falls below the maximum contaminant level, we have determined the water is of good quality.