Determining the Proper Statistic to Identify Contaminated Groundwater

Michael Fagan

1 Introduction

Residents living in a region close to a power plant are concerned about the arsenic level of their groundwater. Arsenic levels become unsafe when they reach 50 micrograms per liter ($\mu g/l$). Groundwater is considered at a standard level (i.e., clean) when it is normally distributed with a mean of 25 $\mu g/l$ and a standard deviation of 5 $\mu g/l$. Groundwater is considered at a contaminated level when it is normally distributed with a mean of 35 $\mu g/l$ and a standard deviation of 5 $\mu g/l$.

To ensure the safety of the water, the residents will regularly test the arsenic levels of the groundwater. Two plans have been proposed, both of which collect five individuals vials of groundwater. The first plan will compute the mean of the five vials, and the groundwater will be considered contaminated if the mean exceeds a predetermined amount. The second plan will record the largest (max) value of the sample, and the groundwater will be considered contaminated if this value exceeds a predetermined amount. We will henceforth refer to this predetermined amount as the critical value.

In this report, we will investigate the efficacy of these two plans.

2 The Problem

Our first task was to determine what an optimal solution should look like. We do not have knowledge of the cost of treating contaminated water or the residents' risk tolerance. Lacking that information, we determined that the optimal combination of method and critical value is that which would maximize the difference between correctly identifying contaminated groundwater and misidentifying clean groundwater as contaminated.

We then simulated two populations of groundwater, one clean and one contaminated, using the respective parameter values given in the introduction. From each population, we took two sets of 100,000 samples, with each sample consisting of five "vials." One set averaged the value of the five "vials" in each sample; the other set found the largest value of each sample. Distributions of these samples can be seen in Figures 1 and 2.

These sets of sample statistics were then compared to a set of critical values, and the percentage of the sample statistics greater than each critical value were recorded. This data can be seen in Table 1. Plots were created comparing these percentages by statistical method. These plots can be seen in Figures 3 and 4.

Finally, difference of the percentages by critical value were plotted and can be seen in Figures 5 and 6. We then identified the critical value at which the maximum difference occurred.

3 Recommendation

We recommend the residents use the mean method and a critical value of 30. This method and critical value produces the largest difference between correctly identifying contaminated groundwater and misidentifying clean groundwater as contaminated. Residents will identify contaminted groundwater 98.73% of the time while incorrectly identifying clean groundwater as contaminated only 1.48% of the time, for a difference of 97.25%. By contrast, the best performing critical value for the max method (36) correctly identifies contaminated groundwater 93.23% of the time while misidentifying clean groundwater as contaminated 7.78% of the time, for a difference of 85.45%.

4 Conclusion

In this report, we've found the optimal method and critical method to maximize the difference between correctly identifying contaminated groundwater and misidentifying clean groundwater as contaminated. We might come to a different recommendation with additional knowledge. For instance, if residents wanted to guarantee they never failed to identify contaminated groundwater, we would recommend the max method with a critical value of 29, though we would note that they would also misidentify clean groundwater as contaminted 70.60% of the time. And if the residents wanted to guarantee they never misidentified clean groundwater as contaminted (perhaps because treating contaminated groundwater is expensive), we would recommend the mean method with a critical value of 35, again noting that they would only identify contaminted groundwater 48.55% of the time. That said, given reasonable costs and risk tolerance, our recommendation should remain close to optimal.

Table 1 - Percentage of Sample Statistics Rejected							
	Mean Method				Max Method		
CV	Contam.	Clean	Diff.	CV	Contam.	Clean	Diff.
23.5	100.0	75.83	24.17	28.5	100.0	74.99	25.01
24.0	99.99	68.33	31.66	29.0	100.0	70.60	29.40
24.5	99.99	60.15	39.85	29.5	99.99	64.84	35.15
25.0	99.99	51.43	48.56	30.0	99.99	56.64	40.36
25.5	99.99	42.56	57.43	30.5	99.97	53.26	46.71
26.0	99.99	34.12	65.87	31.0	99.95	47.20	52.75
26.5	99.99	26.36	73.63	31.5	99.91	42.09	57.82
27.0	99.98	19.63	80.35	32.0	99.81	36.28	63.53
27.5	99.96	14.14	83.82	32.5	99.69	30.97	68.72
28.0	99.91	9.87	90.03	33.0	99.44	26.51	72.93
28.5	99.82	6.62	93.10	33.5	99.12	22.62	76.50
29.0	99.66	4.19	95.47	34.0	98.58	18.95	79.63
29.5	99.32	2.51	96.81	34.5	97.86	15.23	82.63
30.0	**98.73**	**1.48**	**97.25**	35.0	96.75	12.12	84.63
30.5	97.76	0.84	96.92	35.5	95.22	9.97	85.25
31.0	96.28	0.48	95.80	**36.0**	**93.23**	**7.78**	**85.45**
31.5	93.95	0.28	93.67	36.5	90.44	6.11	84.33
32.0	90.77	0.12	90.65	37.0	87.36	5.05	82.31
32.5	86.51	0.06	86.45	37.5	83.41	4.16	72.25
33.0	80.90	0.03	80.87	38.0	79.12	3.12	76.00
33.5	74.01	0.02	73.99	38.5	73.73	2.46	71.27
34.0	66.21	0.01	66.20	39.0	68.31	1.82	66.49
34.5	57.61	0.004	57.606	39.5	62.12	1.41	60.71
35.0	48.55	0.00	48.55	40.0	56.22	1.04	55.18