Data Quality Assessment in support of the Risk Assessment for the Southeast Drainage near the Weldon Spring site, Weldon Spring, Missouri

Prepared by

Paul Black
Deborah Carlson

Neptune and Company, Inc.

August 15, 1996

NEPTUNE AND COMPANY, INC.
1505 15th Street, Suite B
Los Alamos, NM 87544
Phone: 505-662-2121 Fax: 505-662-0500

EXECUTIVE SUMMARY

An engineering evaluation/cost analysis (EE/CA) has been prepared to support the proposed removal of contaminated sediment from selected portions of the Southeast Drainage as part of cleanup activities being conducted by the U.S. Department of Energy (DOE) at the Weldon Spring site in St. Charles County, Missouri. The Southeast Drainage (SE Drainage) is a natural channel with intermittent flow that traverses the Weldon Spring Conservation Area from the Weldon Spring Chemical Plant to the Missouri River. The drainage became contaminated as a result of past activities of the U.S. Army and the DOE. The primary contaminants in sediment are radium, thorium, and uranium. The purpose of this document is to evaluate the adequacy of the data collected from the SE Drainage to support human health risk-based decisions.

The risk-based decisions to be made for the SE Drainage depend on the adequacy of the data collected for supporting those decisions. The data consist of activities of radium-226, radium-228, thorium-230, and uranium-238 measured from sediment samples taken from the SE Drainage. Based on the observed data for each radionuclide, the data for radium-226 can be used effectively to drive the decision in each of the four exposure units.

Risk based cleanup criteria for the principal radioactive contaminants have been established at 13 pCi/g for radium-226 and radium-228, 350 pCi/g for thorium-230 and 290 pCi/g for uranium-238, each corresponding to a human health risk of 10^{-5} excess cancers per lifetime for the hypothetical child scenario. Decisions to be made at the SE Drainage site concern comparison for each radionuclide of the average activities with the target risk levels. The comparisons are performed separately for each of the four exposure units, labeled Units A, B, C, and D, that comprise the SE Drainage area. The main purpose of this document is to perform a Data Quality Assessment (DQA) to determine the adequacy of the radionuclide data collected in 1995 for supporting the human health risk-based decisions. The average radium-226 activity exceeds the target risk level of 13 pCi/g in each exposure unit. This is sufficient information to determine that the data are adequate to support the decision that the average activities at this site exceed the target risk levels.

Further data analyses were performed to provide some other insights into the data. Both surface (0-6 inches below ground surface) and subsurface (6-12 inches) data were collected, however, no statistical differences were exhibited between these two sets of data. Also, no statistical differences were indicated between exposure units for any of the radionuclides, however, graphical presentations indicate that activities for some radionuclides in Unit B may be lower than activities in the other Units.

TABLE OF CONTENTS

EXECUTIVE SUMMARY I

TABLE OF CONTENTS II

TABLE OF TABLES III

TABLE OF FIGURES IV

1. INTRODUCTION: DECISIONS BASED ON RISK ASSESSMENT 1

2. STATISTICAL MODEL 3

3. DATA PREPARATION 4

3.1 COMPARING SURFACE AND SUBSURFACE DATA 5
3.2 COLLOCATED SAMPLES 8
3.3 SUMMARY OF DATA PREPARATION 8

4. COMPARISON OF UNITS A, B, C AND D 9

5. DQA FOR THE CHILD RISK SCENARIO 13

5.1 RADIUM-226 13
5.2 RADIUM-228 13
5.3 THORIUM-230 15
5.4 URANIUM-238 15
5.5 SUMMARY 16

6. REFERENCES 17

APPENDIX A. DATA APPENDIX 18

APPENDIX B. STATISTICAL MODEL 20

APPENDIX C. SURFACE VERSUS SUBSURFACE COMPARISON 27

APPENDIX D. SAMPLE LOCATIONS 33

APPENDIX E. COMPARISON OF ANALYTES ACROSS UNITS 34

APPENDIX F. MEAN ESTIMATES BASED ON LOGNORMALITY 38

APPENDIX G. POWER PLOTS AT TARGET RISK LEVELS 39

INTERPRETATION 39
POWER PLOTS FOR A RISK LEVEL OF 13 PCI/G 41
POWER PLOTS FOR A RISK LEVEL OF 350 PCI/G 45
POWER PLOTS FOR A RISK LEVEL OF 290 PCI/G 49

TABLE OF TABLES

TABLE 3-1 COMPARABLE SURFACE AND SUBSURFACE RESULTS....6
TABLE 3-2 SUMMARY STATISTICS AND TEST RESULTS FOR SURFACE AND SUBSURFACE DATA....7
TABLE 3-3 COLLOCATED SAMPLES....8
TABLE 3-4 SUMMARY STATISTICS FOR COLLOCATED SAMPLES....9
TABLE 4-1 SUMMARY OF DATA FOR UNIT A....10
TABLE 4-2 SUMMARY OF DATA FOR UNIT B....11
TABLE 4-3 SUMMARY OF PREPARED DATA FOR UNIT C....11
TABLE 4-4 SUMMARY OF PREPARED DATA FOR UNIT D....11

TABLE B- 1 CONCLUSIONS AND CONSEQUENCES FOR A CLASSICAL TEST OF HYPOTHESES....21

TABLE E- 1 DIFFERENCES BETWEEN UNITS: STATISTICAL TEST RESULTS FOR URANIUM-238....36
TABLE E- 2 DIFFERENCES BETWEEN UNITS: STATISTICAL TEST RESULTS FOR RADIUM-226....36
TABLE E- 3 DIFFERENCES BETWEEN UNITS: STATISTICAL TEST RESULTS FOR RADIUM-228....36
TABLE E- 4 DIFFERENCES BETWEEN UNITS: STATISTICAL TEST RESULTS FOR THORIUM-230....37

TABLE F- 1 MEAN ESTIMATES (MVUE) BASED ON LOGNORMAL ASSUMPTIONS....38

TABLE OF FIGURES

FIGURE 1-1 SAMPLE LOCATIONS FOR THE SE DRAINAGE 2

FIGURE E- 1 BOX PLOTS OF EACH ANALYTE BY UNIT 34

FIGURE G- 1 $\alpha = 0.01$ 41
FIGURE G- 2 $\alpha = 0.05$ 42
FIGURE G- 3 $\alpha = 0.1$ 43
FIGURE G- 4 $\alpha = 0.2$ 44
FIGURE G- 5 $\alpha = 0.01$ 45
FIGURE G- 6 $\alpha = 0.05$ 46
FIGURE G- 7 $\alpha = 0.1$ 47
FIGURE G- 8 $\alpha = 0.2$ 48
FIGURE G- 9 $\alpha = 0.01$ 49
FIGURE G- 10 $\alpha = 0.05$ 50
FIGURE G- 11 $\alpha = 0.1$ 51
FIGURE G- 12 $\alpha = 0.2$ 52

1. INTRODUCTION: DECISIONS BASED ON RISK ASSESSMENT

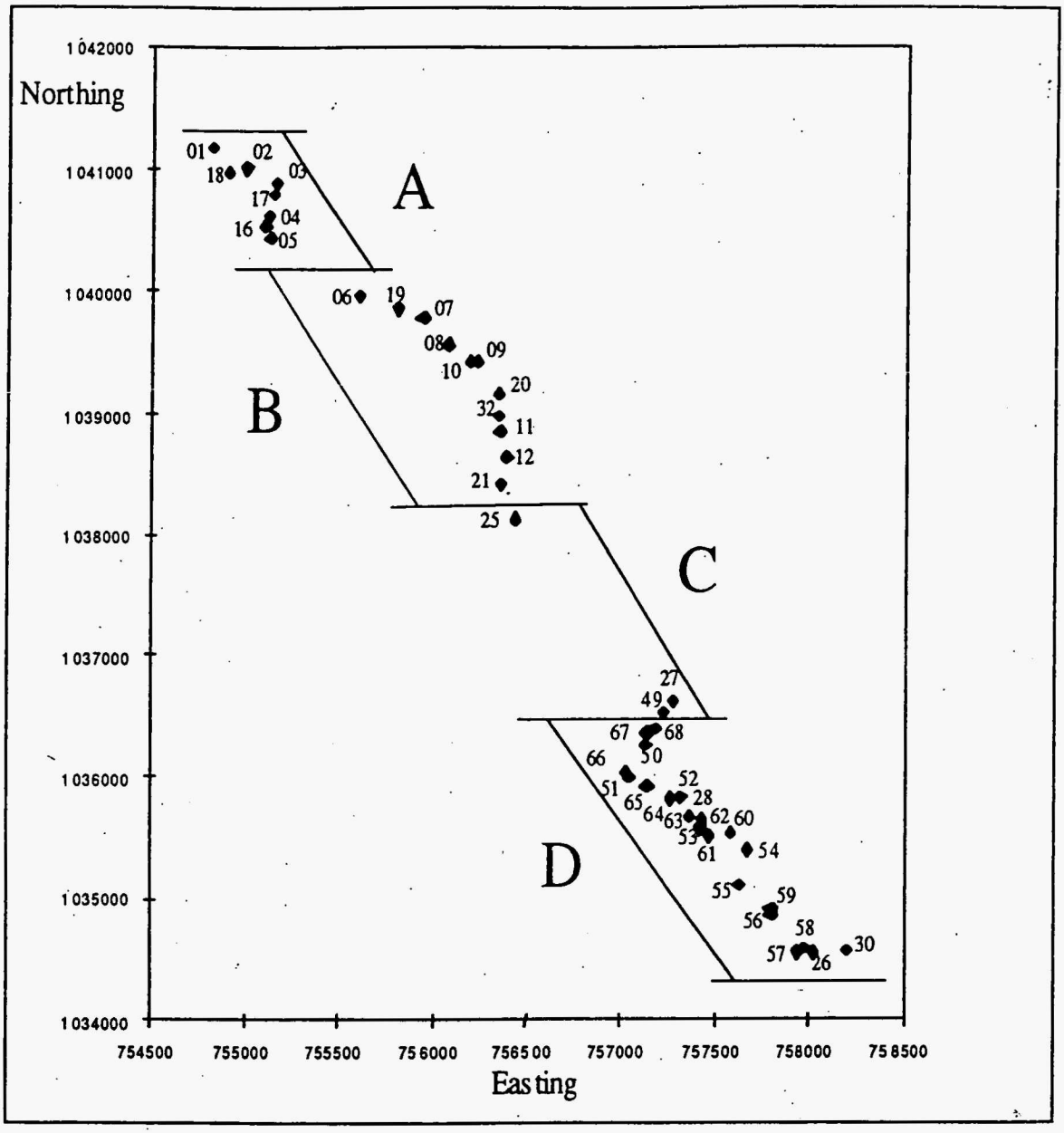
The south-east drainage area (SE Drainage) is a natural drainage with intermittent flow that traverses the Weldon Spring Conservation Area from the Weldon Spring Chemical Plant to the Missouri River. During past operations at the chemical plant, the SE Drainage received discharge from the sanitary and process sewers, and overflow from the raffinate pits. As a result, sediments and soils in the SE Drainage are contaminated with uranium, thorium and radium. Details of the site and the sampling activities can be found in the Weldon Spring Site Remedial Action Plan (1996).

Data has been collected from the SE Drainage area, for the dual purposes of assessing the current human health risk and providing information that may be pertinent to remediation activities. Data is available from 82 samples collected from 44 sampling locations throughout the drainage area (see Appendix A). The data consist of measured activities of uranium-238, thorium-230, radium-226 and radium-228 isotopes for each of the 82 samples. Sample locations are, in general, tens or hundreds of feet apart; however, some of the samples are collocated (i.e., are a few feet apart), and some are subsurface samples (6-12 inches below ground surface) from the same location as a surface sample (0-6 inches) (see Figure 1-1). The purpose of this data quality assessment (DQA) is to evaluate the adequacy of the data collected from the SE Drainage to support human health risk-based decisions for the site.

The risk-based decision model presented in this document uses comparison of upper confidence bounds on mean concentrations derived risk-based cleanup criteria presented in the EE/CA (EE/CA, 1996). These comparisons are performed through the use of one sample *t*-tests, perhaps the most commonly used Classical statistical hypothesis testing mechanism. Risk-based cleanup criteria were derived for a corresponding risk level of 1×10^{-5} for a hypothetical child scenario. The calculated levels are as follows: radium-226, 13 pCi/g; radium-228, 13 pCi/g; thorium-230, 350 pCi/g; and uranium-238, 290 pCi/g.

Data is currently available for the SE Drainage Units A through D (see Figure 1-1 and the Weldon Spring Site Remedial Action Plan, 1995). The principle question to be answered in this DQA is: Given the target risk levels, are the available data adequate for supporting the risk-based decisions of interest? The DQA can also be used to indicate a broader range of (statistical) conditions, or assumptions, under which the data are adequate for supporting risk-based decisions; or, conversely, (statistical) conditions, or assumptions, under which more data would need to be collected.

Figure 1-1 Sample Locations for the SE Drainage



Each of the four radionuclides for which data are available are included separately in this DQA. For each radionuclide, the decision will be made that a unit poses an unacceptable risk if the data for that radionuclide in that unit is unacceptably high (data sufficiently greater than the target risk level). The opposite decision will be made if the data are sufficiently less than the target risk level. However, there is a "gray" region, or a region of indecision, between sufficiently greater and sufficiently less for which the decision may not be clear. The region of indecision is related to the quantity of the

available data, and the magnitude and underlying variability of the data. Increasing the number of data points generally results in a reduction in the size of the effective region of indecision. If the data are far enough removed from the area of indecision then the data adequately support the decision to be made; if, on the other hand, the data fall in the region of indecision then more data may be needed to adequately support the decision to be made. This DQA will indicate if the current data are sufficient for supporting the risk-based decisions of interest, and, if this is not the case, conditions under which the current data would be sufficient for supporting risk-based decisions will be indicated.

2. STATISTICAL MODEL

The decision model used in this DQA relies on Classical statistical hypothesis testing, in particular a one sample *t*-test. In this procedure competing hypotheses are established; the first relates to the possibility that an unacceptable risk is associated with a radionuclide in a Unit; the alternative is that no unacceptable risk exists from that radionuclide in that Unit. In Classical statistical terminology, the former hypothesis is established as the *null* hypothesis. This is the hypothesis that we would like to be able to disprove. The latter hypothesis is established as the *alternative* hypothesis. This is the hypothesis that we would like to establish as "true". Details of the underlying statistical process used for this DQA are presented in Appendix B. A brief description is included in this section.

Formally, it is the null hypothesis that is tested in the Classical testing procedure. If a target risk level for a radionuclide in a given unit is denoted R, and the mean of the concentration distribution for that radionuclide and unit of interest is denoted μ, then the above hypotheses may be translated into the following, more mathematical, statements:

Null Hypothesis: H_o: $\mu > R$

Alternative Hypothesis[1]: H_A: $\mu < R$

Classical statistical testing is structured such that sufficient data must be collected in order to *reject* the null hypothesis (i.e., "prove" the alternative hypothesis). Otherwise the null hypothesis is not rejected. To perform a Classical hypothesis test a *test statistic* is calculated and is compared to a suitable reference probability distribution. In this case, the test statistic is the t statistic, which is compared to the Student t distribution. This comparison indicates the extent to which the data would be considered unusual if the null hypothesis is in fact "true". If the data are deemed unusual in this sense, then the null hypothesis is rejected.

[1] Without affecting the outcome, one of these hypotheses may be established to include equality.

The goal of this DQA exercise is to determine the conditions under which the available data are sufficient for decision making. Adequacy of the data for supporting the risk-based decisions is measured in terms of the *power* of the statistical tests, as described in Appendix B. Presented in the following sections are results of the DQA performed for the SE Drainage area. This DQA uses the available data at face value, assuming, for example, that the necessary QA/QC activities have been performed and the data are ready for their intended use, i.e., risk-based decision making. Separate analyses are performed for each radionuclide for which data are available, and for each Unit (A, B, C, and D).

Before commencing with the DQA, the available data are briefly summarized and exploratory analyses are presented to provide an understanding of the data. In particular, comparisons are made between surface and subsurface data collected; between Units for each analyte; and, some exploratory results are presented for the few collocated samples that were collected. Based on the results of the exploratory analysis, the values used in the ensuing DQA are average values by location of the surface data. Averaging was performed across collocated samples. This represents a conservative approach, resulting in effective sample sizes respectively of 8, 11, 3 and 22 for units A, B, C, and D. More data are available from the multiple observations taken at each location, although the gain in information from the multiple values is difficult to quantify.

A number of assumptions are used as the basis for this DQA. In particular, the data are assumed normally distributed and independent from one another. Given the SE Drainage data, these assumptions can reasonably be questioned. At this time appeals are made to regulatory guidance (e.g., EPA, 1989) and the robustness of *t*-tests. The term robust refers to the capability of a statistical test to withstand substantial deviations from the underlying assumptions. The robustness of the *t* test has been demonstrated repeatedly since its inception in the early 1900s. It may be more appropriate to model the data assuming underlying lognormal distributions (e.g., if the data are skewed to the right) rather than a normal distribution; more complete models may allow for a correlation structure related to the location or comparative proximity of observations; or, more complete models may incorporate aspects of samples taken at different depths. Unless such a need becomes apparent, however, the standard Classical *t*-test is sufficiently robust to provide reasonable results. This is particularly true if the data clearly support the decisions to be made based on this method.

3. DATA PREPARATION

Data are available for radionuclides radium-226, radium-228, thorium-230 and uranium-238 in Units A, B, C, and D. The full set of available data is presented in Appendix A.

A number of issues for the data need to be considered before embarking on the Data Quality Assessment. For example, samples were sometimes collected at the same locations on the surface (0-6 inches) and subsurface (6-12 inches); and, some samples that were collected are collocated in the sense that two or three (surface) samples may have been collected close together, but have been assigned the same location (this includes a few samples that are listed as field duplicates)[2]. The DQA must be performed on a consistent set of data that supports as strongly as possible the underlying statistical model. Decisions need to be made, therefore, about how to handle the surface versus subsurface data and the collocated samples.

3.1 COMPARING SURFACE AND SUBSURFACE DATA

The first step taken during data preparation was to compare the surface and subsurface data. Subsurface data are not available at all locations for which a surface sample was taken, in which case this preparation step considers a subset of the full data set. Table 3-1 provides a list of the data used for this step. Only data directly comparable between surface and subsurface are included. A brief look at the data suggests that there is not much difference between the surface and subsurface results as a whole, although some individual results may be quite different. Figures C-1 through C-4 in Appendix C provide histograms, box plots and simple density estimations that demonstrate the overall similarities. Formal statistical test results for the difference between surface and subsurface data were performed using paired *t*-tests and Mann-Whitney (non-parametric) test procedures. The results are presented in Table 3-2 along with summary statistics for each case. The observed significance levels, or *p*-values, reported for the tests are usually compared to some small probability (typically 0.05) to determine significant effects. Based on the results presented in Table 3-2, there is little evidence of a statistical difference between the surface and subsurface data.

The plots presented in Appendix C indicate that the data are not normally distributed, in which case the nonparametric test results may be preferred. However, the general conclusions are similar regardless of which test results are considered. The summary statistics for uranium-238 indicate that surface concentrations may be marginally greater than subsurface concentrations,

[2] Some sample analyses were repeated resulting in two measurements for the same sample. In general, the duplicate analyses were in close agreement. Data presented include the maximum of two such data points.

Table 3-1 Comparable Surface and Subsurface Results

		U-238		Ra-226		Ra-228		Th-230	
Unit	ID	Surface	Sub-surface	Surface	Sub-surface	Surface	Sub-surface	Surface	Sub-surface
A	001	240.5	166.8	165.4	21.1	15.1	3.4	66.6	50.9
A	002	142.7	93.8	39.7	37.4	8.1	2	23	4.3
A	004	35.5	64.4	8.5	25.8	1.8	3.6	5.2	24.2
A	005	83.1	243.7	11.9	19.3	96.6	288.2	57.5	87.1
A	005	67.6	290.8	23.3	18.8	28.8	326.2	431	158
A	005	235.2	176.7	17.3	6.1	185.3	94.5	208	50.1
A	016	17.6	15.7	5.1	9	1.3	1.62	2.6	27.5
A	017	14.7	14.6	13.8	7.5	1.4	1.3	2.4	0.6
B	007	66.8	31.2	18.9	5.6	6.1	1.8	27.6	3.1
B	019	29	12.7	1.3	35.4	1	1.2	1.7	11.6
B	020	30	2.57	1.4	0.98	1.1	0.76	1.2	0.3
B	021	18.6	9.71	2.9	1.78	1.1	1.33	3.9	1.7
B	032	74.7	39.7	78.5	125.1	1.6	4.5	331.2	1919.6
C	025	741.5	535.6	363.2	199.6	1.8	1.5	455.5	183
C	027	129.2	27.9	11.3	3	35.7	5.6	31	5.1
C	049	25.7	20.5	6.5	3	1.7	1.3	11.6	6.9
D	030	1.6	4.2	1.5	3.3	1.4	1.6	1.5	9.4
D	050	5.8	9.7	7.5	10.6	0.9	1.2	4.4	3.7
D	051	44.2	27.6	10.6	2.8	4.5	1.6	14.3	179.4
D	052	7.5	3.9	2.9	1.5	1.6	1.2	1.7	1.1
D	053	30.5	9.9	9.2	2.1	1.4	0.8	11	1.4
D	054	5.7	0.9	3	1.3	1.3	1	2.6	0.5
D	055	47.9	42.8	17.9	15.9	1.6	2	51	34.7
D	056	17.4	15.1	5.3	2.6	1.3	1.4	10.1	8.3
D	058	123.8	38	78.3	31.1	4.9	2.4	118.8	30.7
D	060	16.6	30.8	6.9	123.8	1.3	1	17.5	51.4
D	065	116.1	277.5	17.5	50.8	16.1	21.6	48.8	235.1
D	066	199.2	213.7	37.9	31.5	4.3	4.7	317.4	2.7

Units are pCi/g.

Table 3-2 Summary Statistics and Test Results for Surface and Subsurface Data

	U-238		Ra-226		Ra-228		Th-230	
Statistic or Test	Surface	Sub-surface	Surface	Sub-surface	Surface	Sub-surface	Surface	Sub-surface
Median	39.9	29.4	11.0	9.8	1.7	1.6	16.0	10.5
Average	91.7	86.4	34.6	28.5	15.3	27.8	80.7	110.4
Std. Dev.	144.6	125.6	73.0	46.3	38.4	81.0	135.4	360.5
Std. Error	27.3	23.7	13.8	8.7	7.3	15.3	25.6	68.1
95% UCL	147.8	135.2	62.9	46.4	30.2	59.3	133.2	250.2
t-test	0.733		0.525		0.349		0.631	
Mann-Whitney	0.094		0.300		0.466		0.406	

Activity Units are pCi/g.
Test results are observed significance levels, or *p*-values.

however, neither statistical test shows a significant difference between the two depths. The box plots and density estimates in Figure C-1 illustrate the similarities between the two sets of data. Correlation plots, found in Appendix C (Figures C-5 and C-6), were also generated to compare the surface and subsurface data. The correlation plots present the data in both the original scale and the logarithmic scale. The logarithmic scale serves to spread out the data so that the regression line is not so sensitive to the large number of low activities reported. These plots again indicate substantial agreement for the surface and subsurface results.

The objectives of the comparison between surface and subsurface data were to determine if there were differences that would need to be accounted for in subsequent analyses, and to determine which data should be used in subsequent analyses. The second objective is of concern due to the different sampling arrangements performed at different locations. For example, some locations have a single sample, others have surface and subsurface samples, and others have more than one surface or subsurface sample (i.e., collocated samples). The issue for the following analyses is one of data comparability. Ideally, data collected under similar circumstances should be used in data analysis. This is required in the context of the underlying statistical model used to support the decision making process. The results of the comparison between surface and subsurface data indicate that the surface data may be used as a surrogate for the subsurface data and surface data combined. This promotes data comparability and allows inclusion of all locations from which samples were collected.

3.2 COLLOCATED SAMPLES

Some collocated data are available at the SE Drainage site. At some locations, the field screening instrument scan of a location indicated that the local area of contamination may be broad enough that two or three samples could be taken in close proximity. Table 3-3 presents data that are collocated. The analytical results from collocated samples are reasonably consistent with those of the initial samples that were taken at these locations; in some instances the first result is larger than the collocated, and in others, vice versa. Table 3-4 presents the summary statistics for the initial samples and the subsequent collocated samples. No meaning need be attached to the order of the samples; the results presented are meant simply to reflect the apparent consistency between results for collocated samples.

Figure D-1 in Appendix D indicates that, for the most part, collocated samples are much closer in proximity than the distinct sample locations. This together with the rationale for collecting the collocated data provides reasonable grounds for averaging across collocated data to provide one activity (concentration) per analyte per location. The limited data analysis of the collocated samples suggests that this approach is reasonable. Averaging is the basic process underlying risk assessment, providing further justification. It is unlikely that averaging the data will result in false negative risk assessment decisions. If anything, risk assessment decisions may be slightly conservative because the number of effective observations included in the risk assessment is less than the number actually collected.

3.3 SUMMARY OF DATA PREPARATION

Based on the findings presented in this section, and on the base rationale for risk assessment, that decisions are made based on average activities, the following data preparation decisions are made to prepare data for the DQA: Subsurface samples are not included in the DQA, introducing a slight conservatism because the effective sample size is reduced; and, averaging is performed across collocated samples, introducing a similar conservatism.

Table 3-3 Collocated Samples

		Initial Results				Collocated Results			
Unit	ID	U-238	Ra-226	Ra-228	Th-230	U-238	Ra-226	Ra-228	Th-230
A	001	240.5	165.4	15.1	66.6	78.7	18.2	3.9	5.4
A	003[a]	327.7	60.5	1.4	37.2	84.5	17.22	1.4	41.7
A	005	83.1	11.9	96.6	57.5	67.6	23.3	28.8	431
A	005					235.2	17.3	185.3	208
A	005[b]	243.7	19.3	288.2	87.1	290.8	18.8	326.2	158
A	005[b]					176.7	6.1	94.5	50.1

D	051	44.2	10.6	4.5	14.3	38	10.7	4.9	43
D	055[b]	42.8	15.9	2	34.7	44	11.5	1.6	48.6
D	060[b]	30.8	123.8	1	51.4	26.9	120.8	1.3	98.8
D	066[b]	79.4	19.7	2.4	344.7	213.7	31.5	4.7	2.7
D	026	10.7	4.5	1.4	24	13.59	5.05	1.88	244
D	026					5.7	1.3	1.09	21.5

Units are pCi/g.
a - Collocated with a subsurface sample.
b - Collocated subsurface samples.
Locations 005 and 026 have multiple collocated samples.

Table 3-4 Summary Statistics for Collocated Samples

	Initial Samples				Subsequent Collocated Samples			
	U-238	Ra-226	Ra-228	Th-230	U-238	Ra-226	Ra-228	Th-230
Median	79.4	19.3	2.4	51.4	73.15	17.26	4.3	49.35
Average	122.54	47.96	45.84	79.72	106.28	23.48	54.63	112.73

Activity Units are pCi/g.

4. COMPARISON OF UNITS A, B, C and D.

Before performing the DQA using the data at this site, it is worth comparing analyte data across the four Units of interest. Because decisions are being made on a Unit by Unit basis, it seems appropriate to consider if there are Unit differences in the data and to further explore the data for a better understanding of the contamination that exists at the site.

Tables 4-1 through 4-4 present the data used in the comparative analyses. Also included in these tables are summary statistics by Unit for each analyte. The data presented and used in the following analyses and DQA consist of the surface results that are averaged across collocated samples as indicated in the previous section.

The data and summary statistics presented in Tables 4-1 through 4-4 indicate initially that there are differences in activities for each analyte across the Units. Box plots, which graphically portray the data distributions, are presented in Figure E-1 in Appendix E. The box plots facilitate comparison between Units for each analyte. For example, the medians across Units for each analyte are relatively consistent, while the variability shows large amounts of fluctuation (medians are shown by white bars across the main

box). Unit C, because of its small sample size (three), exhibits the largest amount of variability for all four analytes, while Unit B exhibits the least. The box plots for Unit C are largely influenced by the one sample (Sample ID 025) that indicates comparatively high activities for uranium-238, radium-226 and thorium-230. Other comparatively high radium-228 and thorium-230 activities are also clearly illustrated in their respective box plots.

Statistical tests were performed to evaluate potential differences between Units. These tests included the *t*-test and the nonparametric Mann-Whitney rank sum, Quantile, and Slippage tests (cf., Gilbert and Simpson, 1992). The data do not, in general, satisfy normality assumptions, in which case the nonparametric tests may be preferred. The *t*-test and the Mann-Whitney test are best suited for measuring complete shifts between distributions, whereas the Quantile and Slippage tests are best suited for assessing partial shifts that may result from a mixture of background and released contaminants. The *t*-test effectively tests the difference between means (assuming normality) of two sets of data; the Mann-Whitney test considers the difference in distributions by ranking the combined data values and comparing the rank sum for each data set; the Quantile test effectively tests for an unusually high proportion of one data set in the upper range of the combined data; and the Slippage test considers the probability of obtaining concentrations from one data set that exceed the maximum concentration from the other. Together these tests provide an indication of the similarity of data sets.

Table 4-1 Summary of Data for Unit A

Sample ID	U-238	Ra-226	Ra-228	Th-230
001	159.6	91.8	9.5	36.0
002	142.7	39.7	8.1	23.0
003	327.7	60.5	1.4	37.2
004	35.5	8.5	1.8	5.2
005	128.6	17.5	103.6	232.2
016	17.6	5.1	1.3	2.6
017	14.7	13.8	1.4	2.4
018	16.2	1.3	0.8	0.2
Average	105	30	16	42
Std. Dev.	109	32	36	78
Std. Error	38	11	13	28
95% UCL	196	56	46	108

Units are pCi/g. Data presented subsequent to data preparation step.

Table 4-2 Summary of Data for Unit B

Sample ID	U-238	Ra-226	Ra-228	Th-230
006	55.6	25.3	2.8	18.2
007	66.8	18.9	6.1	27.6
008	17.0	36	1.5	13.5
009	58.6	111.3	1.7	14.5
010	17.1	20.5	2.2	14.4
011	2.6	0.3	0.7	0.3
012	52.3	42.2	1.6	12.1
019	29.0	1.3	1.0	1.7
020	30.0	1.4	1.1	1.2
021	18.6	2.9	1.1	3.9
032	74.7	78.5	1.6	331.2
Average	38	31	2.0	40
Std. Dev.	24	35	1.5	97
Std. Error	7.2	11	0.5	29
95% UCL	54	55	3.0	105

Units are pCi/g. Data presented subsequent to data preparation step.

Table 4-3 Summary of Prepared Data for Unit C

Sample ID	U-238	Ra-226	Ra-228	Th-230
025	741.5	363.2	1.8	455.5
027	129.2	11.3	35.7	31.0
049	25.7	6.5	1.7	11.6
Average	298.8	127.0	13.1	166.0
Std. Dev.	386.7	204.6	19.6	250.9
Std. Error	223.4	118.1	11.3	144.8
95% UCL	1259.8	635.2	61.8	789.2

Units are pCi/g. Data presented subsequent to data preparation step.

Table 4-4 Summary of Prepared Data for Unit D

Sample ID	U-238	Ra-226	Ra-228	Th-230
026	10.0	3.6	1.5	96.5
028	78.9	21.8	8.8	34.1
030	1.6	1.5	1.4	1.5
050	5.8	7.5	0.9	4.4
051	41.1	10.6	4.7	28.6
052	7.5	2.9	1.6	1.7

Sample ID	U-238	Ra-226	Ra-228	Th-230
053	30.5	9.2	1.4	11.0
054	5.7	3.0	1.3	2.6
055	47.9	17.9	1.6	51.0
056	17.4	5.3	1.3	10.1
057	3.6	2.7	1.3	1.6
058	123.8	78.3	4.9	118.8
059	134.1	54.2	2.5	196.8
060	16.6	6.9	1.3	17.5
061	273.0	76.7	2.5	131.8
062	27.0	14.0	2.3	11.9
063	110.5	48.2	3.3	86.6
064	60.0	20.5	3.1	86.2
065	116.1	17.5	16.1	48.8
066	199.2	37.9	4.3	317.4
067	144.6	30.0	3.5	44.4
068	124.4	23.1	85.8	157.9
Average	71	22	7	66
Std. Dev.	73	23	18	80
Std. Error	16	5	4	17
95% UCL	106	33	15	105

Units are pCi/g. Data presented subsequent to data preparation step.

Tables E-1 through E-4 provide summary statistical test results for determining differences in radionuclide activities between Units. Some observations are also provided in Appendix E on these test results. Considering these observations, it is difficult to support conclusions that there are differences in activities between the Units. The graphical presentations indicate that activities of uranium-238, radium-228 and thorium-230 may be lower in Unit B than in the other Units, however, the statistical test results support this conclusion only marginally. The marginality of the relatively few potentially significant test results could also be a consequence of the relatively small sample sizes, especially from Units A and C.

The test results discussed in this section do not affect the Unit specific DQA that is presented in the next Section. However, particularly in the case of radium-226, for which no between Unit statistical differences were observed, it may be reasonable to consider the Unit data together instead of separately by Unit. This would certainly provide more power because of the increased sample size, but would not allow for Unit specific risk-based decisions.

5. DQA FOR THE CHILD RISK SCENARIO

The first indication of whether the data are adequate for supporting the intended risk-based decisions can be found by comparing upper confidence bounds on the available data to the target risk levels. Tables 4-1 through 4-4 present upper confidence bounds for Units A, B, C and D. If the upper confidence bound is greater than the target risk level, then the null hypothesis will not be rejected at the corresponding significance level. Conversely, if the upper confidence bound is less than the target risk level, then the null hypothesis will be rejected at the corresponding significance level. If the null hypothesis is not rejected, then the data adequately support the risk-based decision to be made (at the given significance level). However, if the null hypothesis is rejected, then the *power of the test* (see Appendix B) must be considered to determine if the number of samples is adequate to support the decision.

5.1 RADIUM-226

In the case of radium-226, the mean concentrations in all Units exceed the 10^{-5} risk level of 13 pCi/g, the lowest mean occurring in Unit D at approximately 22 pCi/g. Consequently, the null hypothesis for radium-226 is not rejected in any Unit. That is, there is sufficient evidence to believe that radium-226 activities are greater than the risk level of interest[3]. Consequently, the decision for radium-226 is clear and the data are sufficient to support the decision. If the target risk level is changed then this conclusion would need to be revisited.

Decisions for the remaining analytes are, consequently, subordinate to the decision for radium-226 in the sense that the overall null hypothesis for this site concerns exceedence of risk levels for any one analyte. That is, because radium-226 activities exceed risk levels, then activities as a whole exceed risk levels. Consequently, the data at the SE Drainage site are adequate to support the decision that site activities exceed human health risk levels for the scenario presented.

The remainder of this DQA focuses on the remaining three analytes to determine if the analyte-specific data are sufficient to support further analyte-specific decisions.

5.2 RADIUM-228

Considering radium-228, the mean activities in Units A and C exceed the target risk level of 13 pCi/g. Consequently, the data are adequate to support the radium-228 specific risk-based decision for these Units.

---

[3] Note that the radium-226 mean activities when estimated based on a lognormal distribution are also greater than the risk level (see Appendix F).

The mean radium-228 activities for units B and D are approximately 2.0 pCi/g and 7.0 pCi/g, respectively, with corresponding 95% upper confidence bounds of 3.0 pCi/g and 15.0 pCi/g. The data for Unit B are therefore adequate to support the decision at a 0.025 significance level, while the adequacy of the Unit D data is not yet determined[4]. The power of the corresponding statistical tests can be considered to determine the range of conditions under which the radium-228 data are adequate to support a decision that the site data are less than the target risk level for these two Units. Figures G-1 through G-4 present power plots corresponding to a risk level of 13 pCi/g. The plots are presented at four different significance levels (0.01, 0.05, 0.1, 0.2) for which the number of samples is varied across the range of the number of samples available in each Unit (n = 3, 8, 11, 22), and across a range of standard deviations that are consistent with the range of standard deviations exhibited for the analyte data by Unit.

For example, Figure G-1 shows the power plots corresponding to a significance level of 0.01. The first plot shows that, with a standard deviation of 3 pCi/g and the estimated mean radium-228 activity in Unit B, three samples are adequate to support a radium-228 specific decision for this Unit. The estimated standard deviation of radium-228 activities in Unit B is approximately 3.3 pCi/g, and the number of available data points is 11, in which case the data adequately support a radium-228 specific risk-based decision for this Unit (i.e., that the radium-228 activities in Unit B are probably below the target risk level).

Given the estimated mean and standard deviation of radium-228 activities in Unit D (7.0 pCi/g and 18.3 pCi/g), Figure G-1 shows that 22 data points are not sufficient to adequately support a radium-228 specific decision at the 0.01 significance level (the second plot indicates that the power at the "true" mean of 7.0 pCi/g, with a "true" standard deviation of 18.3 pCi/g is approximately 0.3, corresponding to a 70% false positive rate). This finding is corroborated by considering the upper confidence bound of 15.3 pCi/g radium-228 activity presented in Table 4-4. That is, the data are not sufficient to support the decision at the 0.025 significance level. If the significance level is increased to 0.2, then the corresponding power is approximately 0.9; conditions that may be considered adequate to support a radium-228 specific decision. Overall, it appears that insufficient data are available to support such a decision. Given the estimated mean radium-228 activity in this Unit, either more data are needed, a lower standard deviation is needed, or greater tolerance for decision errors are required.

[4] The significance level is half of one minus the confidence level because the confidence bound is based on a two sided analysis.

5.3 THORIUM-230

The target risk level for thorium-230 at the SE Drainage area is 350 pCi/g. The mean thorium-230 activities respectively in Units A, B, and D are approximately 42 pCi/g, 40 pCi/g, and 68 pCi/g, and the 95% upper confidence bounds in these Units respectively are 108, pCi/g, 105 pCi/g, and 105 pCi/g. Consequently, the decision in each of these Units appears to be adequately supported by the data at a 0.025 significance level.

Given the sample sizes in these three Units, the power plots presented in Figure G-5 confirm the findings based on the upper confidence bounds presented. The available data are adequate to support thorium-230 specific risk-based decisions at the 0.01 significance level. Figure G-6 through G-8 provide further power plots at alternate significance levels.

Given the estimated mean and standard deviation of thorium-230 activities in Unit C (170 pCi/g and 250 pCi/g), Figure G-8 clearly shows that 3 data points are not sufficient to adequately support a thorium-230 specific decision at the 0.2 significance level. This finding is corroborated by considering the upper confidence bound of 790 pCi/g thorium-230 activity presented in Table 4-3. Overall, insufficient data are available to support a thorium-230 specific decision. Given the estimated mean and standard deviation of thorium-230 activity, more than 20 samples would be needed to adequately support a thorium-230 specific risk-based decision in Unit C. The conclusions for Unit C are, however, highly affected by the one comparatively high observed activity of 460 pCi/g (Sample ID 025). The estimated mean and standard deviation for this Unit are greater than for the other Units because of this value. In light of the site specific conclusions for radium-226, the thorium-230 activities in the other three Units (which are not statistically different than Unit C), and the occurrence of the statistical outlier, there is no apparent need to collect more thorium-230 data for this Unit.

5.4 URANIUM-238

The target risk level for U-238 at the SE Drainage area is 290 pCi/g. The mean U-238 activities respectively in Units A, B, and D are approximately 105 pCi/g, 38 pCi/g, and 71 pCi/g, and the 95% upper confidence bounds in these Units respectively are 196, pCi/g, 54 pCi/g, and 106 pCi/g. Consequently, the decision in each of these Units appears to be adequately supported by the data at a 0.025 significance level.

Given the sample sizes in these three Units, the power plots presented in Figure G-9 confirm the findings based on the upper confidence bounds presented. The available data are adequate to support thorium-230 specific risk-based decisions at the 0.01 significance level. Figure G-10 through G-12 provide further power plots at alternate significance levels.

Considering uranium-238, the mean activities in Unit C exceeds the target risk level of 290 pCi/g. Consequently, the data are adequate to support the uranium-238 specific risk-based decision for these Units. The conclusion for Unit C is, however, highly affected by the one comparatively high observed activity of 742 pCi/g (Sample ID 025). The estimated mean and standard deviation for this Unit are greater than for the other Units because of this value. In light of the site specific conclusions for radium-226, the conclusion reached based on the data for uranium-238 activities in Unit C are not in conflict with those for radium-226, in which case the presence of the statistical outlier does not influence the overall conclusions for this site.

5.5 SUMMARY

The decisions at this site are driven by mean radium-226 activities, which consistently exceed the target risk level of 13 pCi/g. In which case, the overall decision for each Unit, that the target risk levels are exceeded and further action needs to be considered, are supported adequately by the available data.

Analyte specific conclusions are also, in general, supported by the available data. With the exception of radium-226 these conclusions usually indicate that the radioactivities at the site are not of unacceptable human health risk concern. The main exception occurs for Unit C, for which only three data points are available, one of which might be considered a statistical outlier.

The analyte specific conclusions do not affect the overall conclusions that are driven by radium-226 results.

It should always be recognized that there are a number of assumptions underlying the analysis presented that may be violated to some degree (especially the normality assumption and independence assumptions), and that the data have been prepared in a conservative way (because the effective sample size was substantially reduced) to produce these results. Mean concentrations based on lognormal distributional assumptions tend to be reasonably in line with the simple averages that are presented, for which normal assumptions are in effect (see Appendix F). Consequently, there is good reason to believe that the results presented, at least qualitatively, provide reasonable conclusions for this site.

6. REFERENCES

ANL (Argonne National Laboratory), August, 1996, "Engineering Evaluation/Cost Analysis for the Proposed Removal Action at the Southeast Drainage near the Weldon Spring Site, Weldon Spring, Missouri," prepared for the U.S. Department of Energy, DOE/OR/21548-584, Environmental Assessment Division, Argonne, Illinois.

EPA (US Environmental Protection Agency), December 1989. "Risk Assessment Guidance for Superfund, Volume I, Human Health Evaluation Manual (Part A)," Interim Final, EPA 540/1-89/002, Office of Emergency and Remedial Response, Washington, DC.

Gilbert, R.O., 1987. "Statistical Methods for Environmental Pollution Monitoring," Van Nostrand Reinhold Publishing Company, 115 Fifth Avenue, New York, NY 10003.

Gilbert, R.O., and Simpson, J.C., December 1992. "Statistical Methods for Evaluating the Attainment of Cleanup Standards - Volume 3: Reference-Based Standards for Soils and Solid Media," prepared for the U.S. Environmental Protection Agency under a Related Services Agreement with the U.S. Department of Energy, Contract DE-AC06-76RLO 1830, Pacific Northwest Laboratory, Richland, Washington 99352.

Weldon Spring Site Remedial Action Project, April 1995. "Southeast Drainage Soils Review Sampling Report: Revision A (DRAFT)," prepared by MK-Ferguson Company and Jacobs Engineering Group for the U.S. Department of Energy, Oak Ridge Operations office, Under Contract DE-AC05-86OR21548.

APPENDIX A. DATA APPENDIX

Table A- 1 SE Drainage Data

Exposure Unit	Location ID	Depth/ Collocation ID	U-238	Ra-226	Ra-228	Th-230	cpm
A	001	1A	240.5	165.4	15.1	66.6	131000
A	001	1B	78.7	18.2	3.9	5.4	131000
A	001	2A	166.8	21.1	3.4	50.9	131000
A	002	1A	142.7	39.7	8.1	23	50000
A	002	2A	93.8	37.4	2	4.3	50000
A	003	1A	327.7	60.5	1.4	37.2	75000
A	003	2B	84.5	17.22	1.4	41.7	75000
A	004	1A	35.5	8.5	1.8	5.2	31000
A	004	2A	64.4	25.8	3.6	24.2	31000
A	005	1A	83.1	11.9	96.6	57.5	60000
A	005	1B	67.6	23.3	28.8	431	60000
A	005	1C	235.2	17.3	185.3	208	60000
A	005	2A	243.7	19.3	288.2	87.1	60000
A	005	2B	290.8	18.8	326.2	158	60000
A	005	2C	176.7	6.1	94.5	50.1	60000
A	016	1A	17.6	5.1	1.3	2.6	NA
A	016	2A	15.7	9	1.62	27.5	NA
A	017	1A	14.7	13.8	1.4	2.4	15000
A	017	2A	14.6	7.5	1.3	0.6	15000
A	018	1A	16.2	1.3	0.8	0.2	9800
B	006	1A	55.6	25.3	2.8	18.2	20000
B	007	1A	66.8	18.9	6.1	27.6	30000
B	007	2A	31.2	5.6	1.8	3.1	30000
B	008	1A	17	36	1.5	13.5	27500
B	009	1A	58.6	111.3	1.7	14.5	0
B	010	1A	17.4	20.5	2.2	14.4	0
B	011	1A	2.6	0.3	0.7	0.3	0
B	012	1A	52.3	42.2	1.6	12.1	20000
B	019	1A	29	1.3	1	1.7	NA
B	019	2A	12.7	35.4	1.2	11.6	NA
B	020	1A	30	1.4	1.1	1.2	0
B	020	2A	2.57	0.98	0.76	0.3	0
B	021	1A	18.6	2.9	1.1	3.9	0
B	021	2A	9.71	1.78	1.33	1.7	0
B	032	1A	74.7	78.5	1.6	331.2	NA
B	032	2A	39.7	125.1	4.5	1919.6	NA
C	025	1A	741.5	363.2	1.8	455.5	210000
C	025	2A	535.6	199.6	1.5	183	210000
C	027	1A	129.2	11.3	35.7	31	61000

Exposure Unit	Location ID	Depth/ Collocation ID	U-238	Ra-226	Ra-228	Th-230	cpm
C	027	2A	27.9	3	5.6	5.1	61000
C	049	1A	25.7	6.5	1.7	11.6	62000
C	049	2A	20.5	3	1.3	6.9	62000
D	026	1A	10.7	4.5	1.4	24	12000
D	026	1B	13.59	5.05	1.88	244	12000
D	026	1C	5.7	1.3	1.09	21.5	12000
D	028	1A	78.9	21.8	8.8	34.1	30000
D	030	1A	1.6	1.5	1.4	1.5	NA
D	030	2A	4.2	3.3	1.6	9.4	NA
D	050	1A	5.8	7.5	0.9	4.4	13000
D	050	2A	9.7	10.6	1.2	3.7	13000
D	051	1A	44.2	10.6	4.5	14.3	26000
D	051	2A	27.6	2.8	1.6	179.4	26000
D	051 DU	1B	38	10.7	4.9	43	NA
D	052	1A	7.5	2.9	1.6	1.7	13000
D	052	2A	3.9	1.5	1.2	1.1	13000
D	053	1A	30.5	9.2	1.4	11	14000
D	053	2A	9.9	2.1	0.8	1.4	14000
D	054	1A	5.7	3	1.3	2.6	12000
D	054	2A	0.9	1.3	1	0.5	12000
D	055	1A	47.9	17.9	1.6	51	28000
D	055	2A	42.8	15.9	2	34.7	28000
D	055 DU	2B	44	11.5	1.6	48.6	NA
D	056	1A	17.4	5.3	1.3	10.1	16000
D	056	2A	15.1	2.6	1.4	8.3	16000
D	057	1A	3.6	2.7	1.3	1.6	17000
D	058	1A	123.8	78.3	4.9	118.8	40000
D	058	2A	38	31.1	2.4	30.7	40000
D	059	1A	134.1	54.2	2.5	196.8	38000
D	060	1A	16.6	6.9	1.3	17.5	10000
D	060	2A	30.8	123.8	1	51.4	10000
D	060 DU	2B	26.9	120.8	1.3	98.8	NA
D	061	1A	273	76.7	2.5	131.8	49000
D	062	1A	27	14	2.3	11.9	NA
D	063	1A	110.5	48.2	3.3	86.6	35000
D	064	1A	60	20.5	3.1	86.2	40000
D	065	1A	116.1	17.5	16.1	48.8	90000
D	065	2A	277.5	50.8	21.6	235.1	90000
D	066	1A	199.2	37.9	4.3	317.4	55000
D	066	2B	79.4	19.7	2.4	344.7	55000
D	066 DU	2A	213.7	31.5	4.7	2.7	NA
D	067	1A	144.6	30	3.5	44.4	62000
D	068	1A	124.4	23.1	85.8	157.9	114000

DU indicates field duplicate.

APPENDIX B. STATISTICAL MODEL

As discussed in Section 2, the decision model used in this document relies on Classical statistical hypothesis testing, in particular, a one sample *t*-test. If a target risk level for a radionuclide in a given Unit is denoted *R*, and the mean of the concentration (activity) distribution for that radionuclide and Unit is denoted μ, then the null and alternative hypotheses may be written as follows:

Null Hypothesis: H_o: $\mu > R$

Alternative Hypothesis[5]: H_A: $\mu < R$

Classical statistical testing is structured such that sufficient data must be collected in order to *reject* the null hypothesis (i.e., "prove" the alternative hypothesis). Otherwise the null hypothesis is not rejected. To perform a Classical hypothesis test a *test statistic* is calculated and is compared to a suitable reference probability distribution. This comparison indicates the extent to which the data would be considered unusual if the null hypothesis is in fact "true". If the data are deemed unusual in this sense, then the null hypothesis is rejected. The reference distribution is selected based on the underlying (assumed) statistical process. In this DQA, each observation (radionuclide activity) within a Unit is treated as an independent realization of the same (but unknown) normal distribution. For the purposes of performing a human health risk assessment, decisions are often made based on the mean concentration (activity), $\bar{x}$, of a contaminant. Under these assumptions the appropriate reference distribution is the Student *t* distribution[6]. The test statistic, *t*, is calculated as follows (where *s* is an estimate of the standard deviation of the distribution of activities for a radionuclide in a Unit, and *n* is the number of independent data points):

$$t = \frac{\bar{x} - R}{s/\sqrt{n}}$$

This test statistic is compared to the *t* distribution with *n*-1 degrees of freedom. The comparison is performed at a specified *significance level*, α, that represents the probability of making a *Type I Error*. For the hypotheses specified above a Type I Error corresponds to a false negative decision error, i.e., the probability of concluding that the null hypothesis should be rejected when in fact it should not be rejected.

---

[5] Without affecting the outcome, one of these hypotheses may be established to include equality.

[6] Assumes unknown variance

A *Type II Error*, i.e., the probability of concluding that the null hypothesis should not be rejected when in fact it should be rejected is also considered. For the hypotheses under consideration, a Type II Error corresponds to a false positive decision error. This probability may be specified distinctly for each possible value of μ.

Table B-1 presents the possibilities in terms of making a correct decision or making an incorrect decision.

Table B- 1 Conclusions and Consequences for a Classical Test of Hypotheses

CONCLUSION	"TRUE" STATE OF NATURE	
	H_o "true"	H_A "true"
H_o "true"	Correct decision	Type II Error[b] (probability b)
H_A "true"	Type I Error[a] (probability a)	Correct decision

a - false negative error rate for the hypotheses given
b - false positive error rate for the hypotheses given

The *power function* is related directly to Type II Error rates. The power of the hypothesis test at a given value of μ is simply the probability of concluding that the null hypothesis should be rejected when in fact it should be rejected (i.e., a correct decision), and this is 1-Type II Error. A typical power function for a one-sided *t*-test of the type used for this investigation is depicted in Figure B-1. Figure B-2 provides a representation that more clearly translates to desired performance characteristics, or Data Quality Objectives (DQOs) that may be specified during the planning process.

Desired performance characteristics of a data collection activity are measured through inputs that reflect "allowable power functions". These characteristics include specifications of Type I and Type II error rates, i.e., probabilities that represent the decision makers tolerance for making an incorrect decision. For example, in Figure B-1, α is specified as 0.05, which corresponds to a 5% chance of making a false negative (in this case) decision or of rejecting the null hypothesis when in fact it should not be rejected. Also specified in Figure B-1 is a value of 0.9 that corresponds to acceptable power given a specified "true" mean that falls well to the left of the alternative hypothesis space. Equivalently, this value corresponds to a 10% probability of making a false positive decision error (i.e., specifying that the null hypothesis should not be rejected when in fact it should be rejected) at a given hypothesized value of the "true" mean. Through this mechanism of specifying acceptable limits on decision errors given values of the "true" mean the allowable class

of power curves can be derived. An optimal sample size can then be calculated by determining the single power curve that most closely satisfies the constraints specified by the desired performance characteristics.

The process of determining optimal sample size, or equivalently of determining whether available data are sufficient for decision making purposes is better explained by considering Figure B-2, as well as Figure B-1. Although the hypotheses are specified at a level R, the decision point for a Classical hypothesis test is termed the *Critical Value* of the test, denoted C. In the case of the one sided hypothesis presented for the risk-based decision here, the critical value is less than the level specified in the null hypothesis, i.e., $C < R$. This is because the onus of this one sided testing strategy is on proving that the null hypothesis is false, and some data less than the hypothesized value R is considered not sufficient to overturn the belief in the null hypothesis. Once the mean of the observed data falls below the critical value then the null hypothesis can be rejected. The critical value corresponds to a power of 0.5, or equivalently, to a 50% probability of making a Type II Error. This is why it is the effective decision point. The desired performance specifications required to compute an optimal sample size and the critical value must include at least the following elements:

1.	A (risk-based) threshold value	R
2.	A desired detectable difference	δ, or $(R - x)$
3.	Type I Error Rate	α
4.	Type II Error Rate that corresponds to the detectable difference at x	$\beta(x)$
5.	Estimated Standard Deviation	s

This minimal list allows for specification of two points (α and $\beta(x)$) on the desired performance graph. More points can be specified, each of which adds another constraint that affects the optimal sample size calculations. Note that the desired detectable difference is related to x in this presentation. In effect the desired detectable difference is defined, for the hypotheses specified, in terms of the largest value of the "true" mean at which a specification of power (or probability of false positive decision) is made.

Figure B- 1 A Typical Power Curve for a One Sided *t*-test

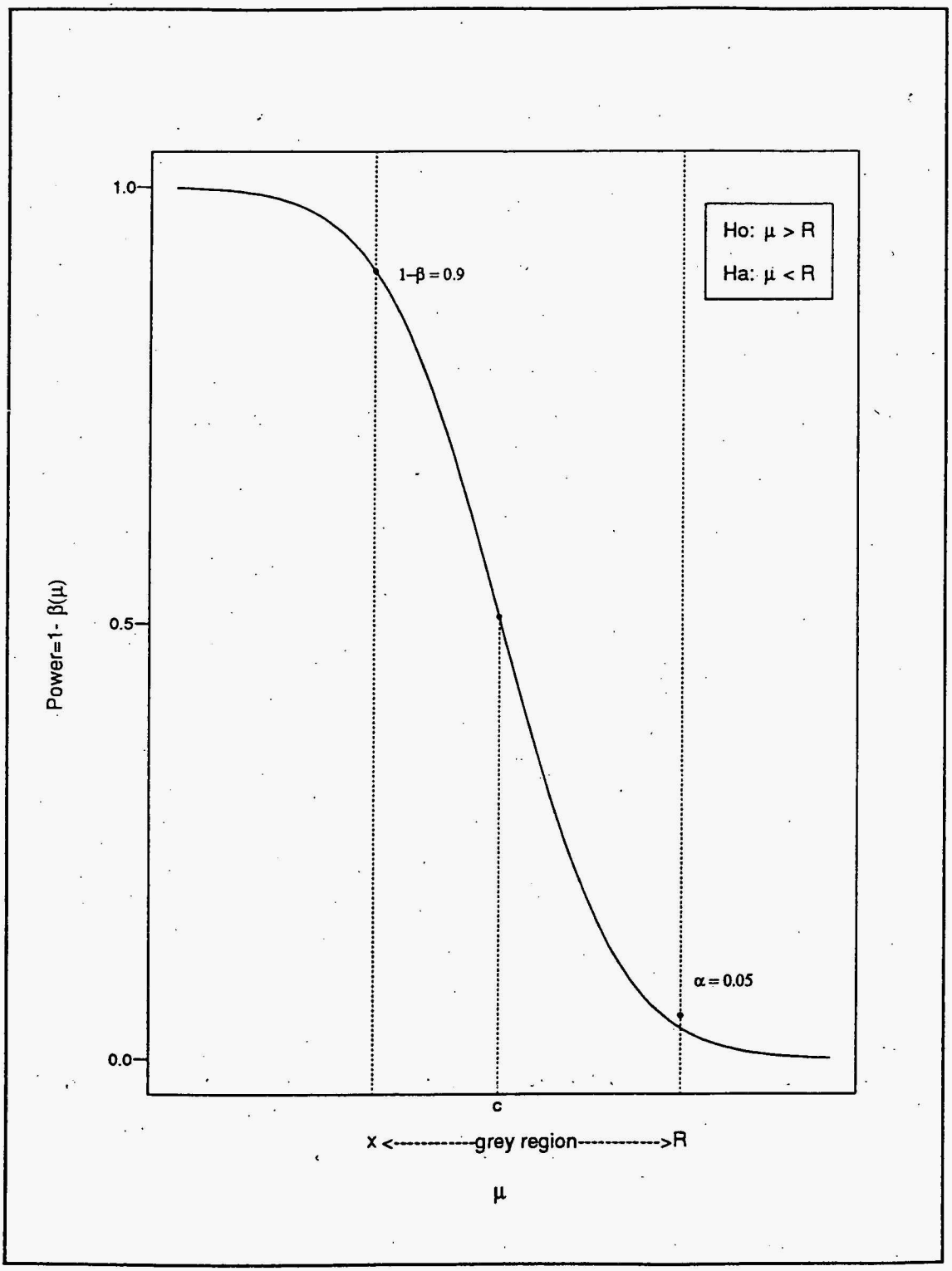
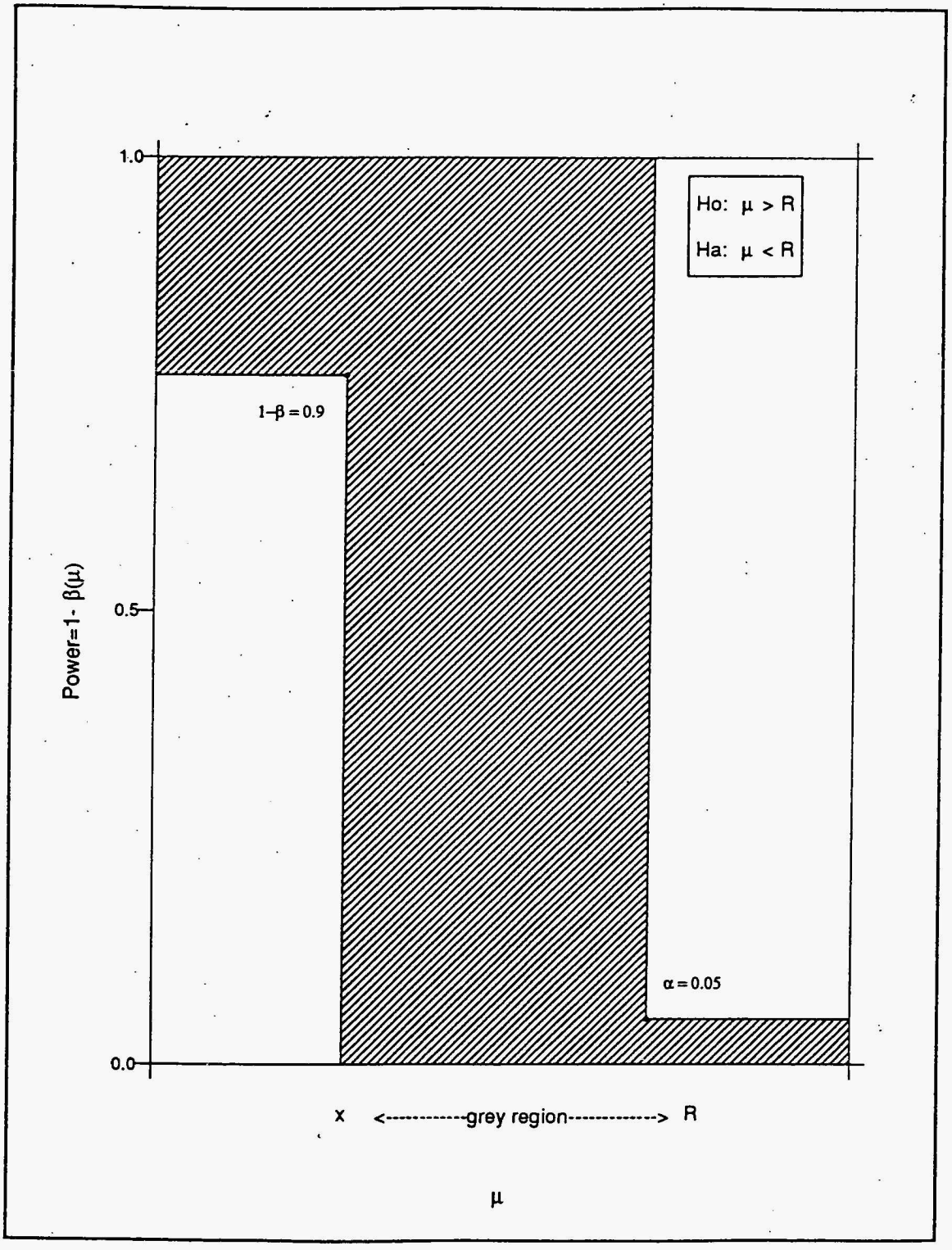


Figure B- 2 A Typical Specification of Desired Performance Characteristics



Under the statistical assumptions of independent observations from the same normal distribution, the following statements lead to a formula for calculating the sample size that will satisfy, in expectation, the specified desired performance constraints. First, the power function at a given "true" value of the mean x is given by the following probabilistic relationship:

$$Power(x) = \Pr(\overline{X} < C \mid \mu = x)$$

Using this foundation and the statistical assumptions indicated above, the specified Type I and Type II Error rates can be translated into the following statistical equations:

$$\alpha = \Pr(\overline{X} < C \mid \mu = R)$$

$$= t\left(\frac{C-R}{s/\sqrt{n}}\right)$$

i.e.,

$$t_{\alpha,n-1} = \frac{C-R}{s/\sqrt{n}} \tag{B.1}$$

and:

$$1-\beta(x) = \Pr(\overline{X} < C \mid \mu = x)$$

$$= t\left(\frac{C-x}{s/\sqrt{n}}\right)$$

i.e.,

$$t_{\beta(x),n-1} = \frac{C-x}{s/\sqrt{n}} \tag{B.2}$$

Given specifications of the desired performance characteristics, this pair of equations can be solved for n, and C. After some manipulation, the following result provides the mechanism by which n is calculated:

$$n = \frac{s^2}{\delta^2}(t_{\alpha,n-1} + t_{\beta(x),n-1})^2.$$

Because n appears on both sides of this equation, the solution is obtained by iteration. Once the sample size n is obtained, the critical value can be determined by substitution into either power equation (Equations B.1 or B.2).

Rather than calculating optimal samples size based on this approach, the goal of this DQA exercise is to determine the conditions under which the available

data are sufficient for decision making. The basic formulation is the same. However, the objective is to vary each of the input parameters to determine effect on sample size determination, and hence to determine if the data collected are adequate for supporting the risk-based decisions of interest.

Underlying the statistical process described above are assumptions of normality and independence. For some data sets the validity of these assumptions can be questioned. At this time appeals are made to the robustness of *t*-tests, the capability of which to withstand substantial deviations from the underlying assumptions. The robustness of the *t*-test has been demonstrated repeatedly since its inception in the early 1900s. More complete models may allow for a correlation structure that takes into account the specific characteristics of the Weldon Spring data. These site specific characteristics include the comparative proximity of observations, assumptions of a positively skewed distributions (as opposed to the normal distribution assumption of symmetry), and samples taken at different depths. Unless such a need becomes apparent, however, the standard Classical *t*-test is sufficiently robust to provide reasonable results. This is particularly true if the data clearly support the decisions to be made based on this method.

Of more importance may be the effect of performing one-sided Classical hypothesis testing in this framework. For the hypotheses specified above, if the mean concentration estimated from the data is greater than the hypothesis threshold (actually greater than the critical value is all that is required), then the data are considered sufficient to support the decision. Even if, for example, the data consist of three observations (e.g., Unit C)! If the estimated mean is less than the critical value then appeals are made to the power of the test to determine sufficiency. This procedure is overly protective of the null hypothesis for pure decision making purposes. One should realize that in this Classical framework the null hypothesis may be proved to be false (i.e., by collecting sufficient data that the mean concentration is far enough below the hypothesis value being tested), but it can never be proved to be true. Just because the alternative hypothesis cannot be proved does not mean that the null hypothesis is proven! It simply means that insufficient information has been collected to prove the alternative hypothesis. This is a continuing source of dilemma for Classical procedures that is exacerbated by using one-sided testing procedures. The testing procedures do not adequately translate to decision rules. It is somewhat preferable to perform two-sided tests for this reason (at least then power is nearly always considered), although further departures from Classical methods may be more preferable (e.g., Bayesian decision based methods).

APPENDIX C. SURFACE VERSUS SUBSURFACE COMPARISON

Figure C- 1 Comparable Surface and Subsurface Data Plots for Radium-226

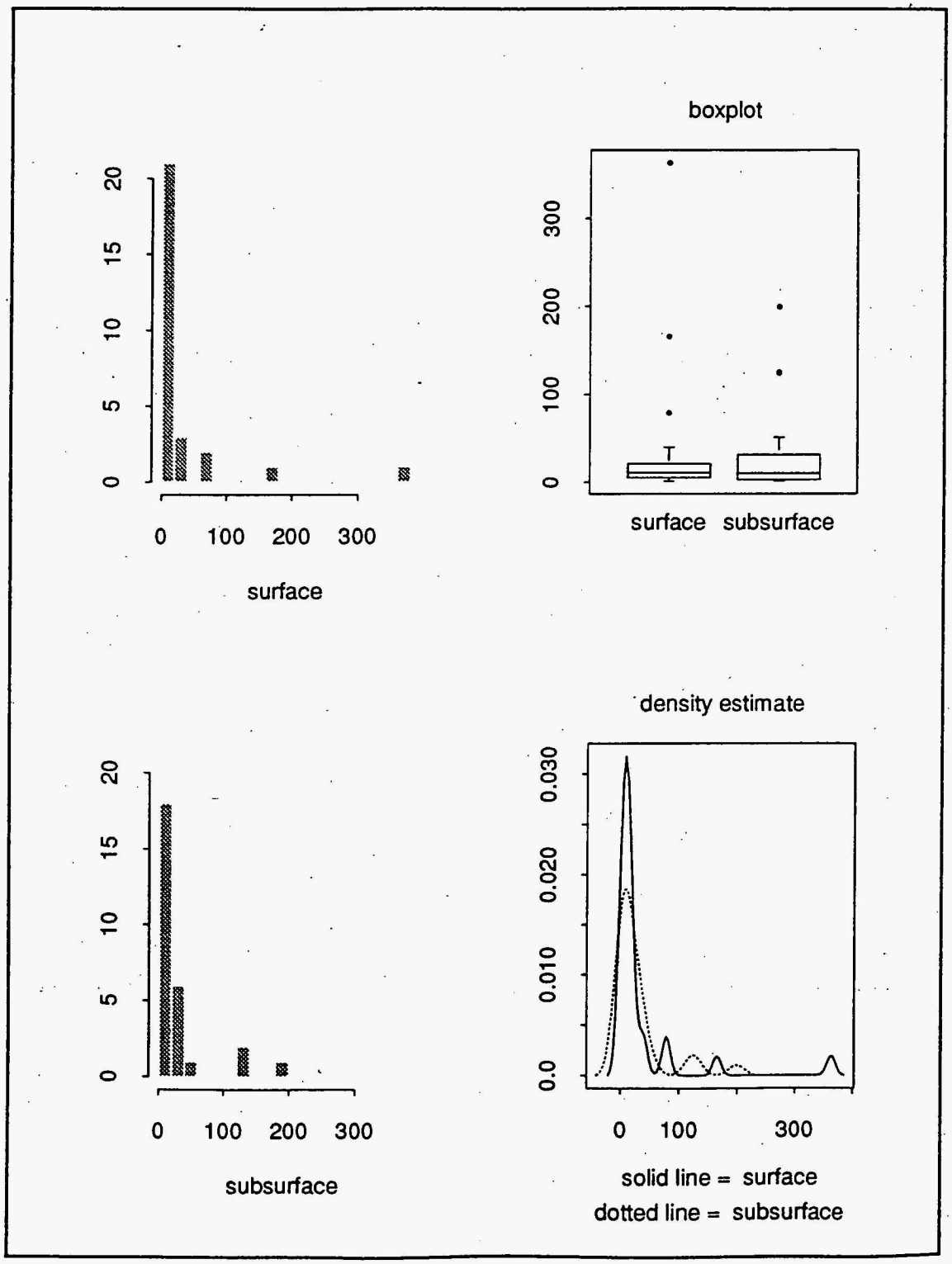


Figure C- 2 Comparable Surface and Subsurface Data Plots for Radium-228

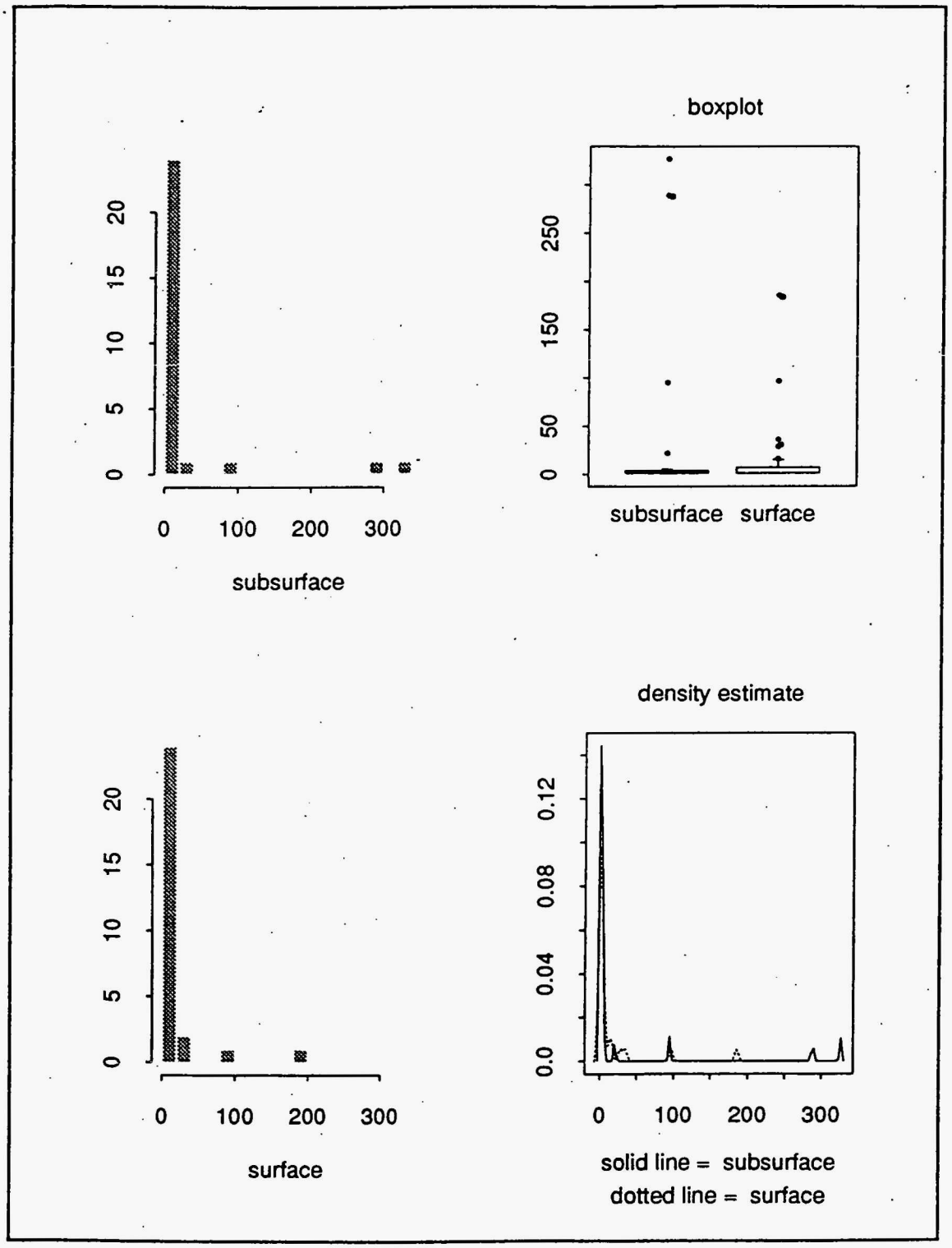


Figure C- 3 Comparable Surface and Subsurface Data Plots for Thorium-230

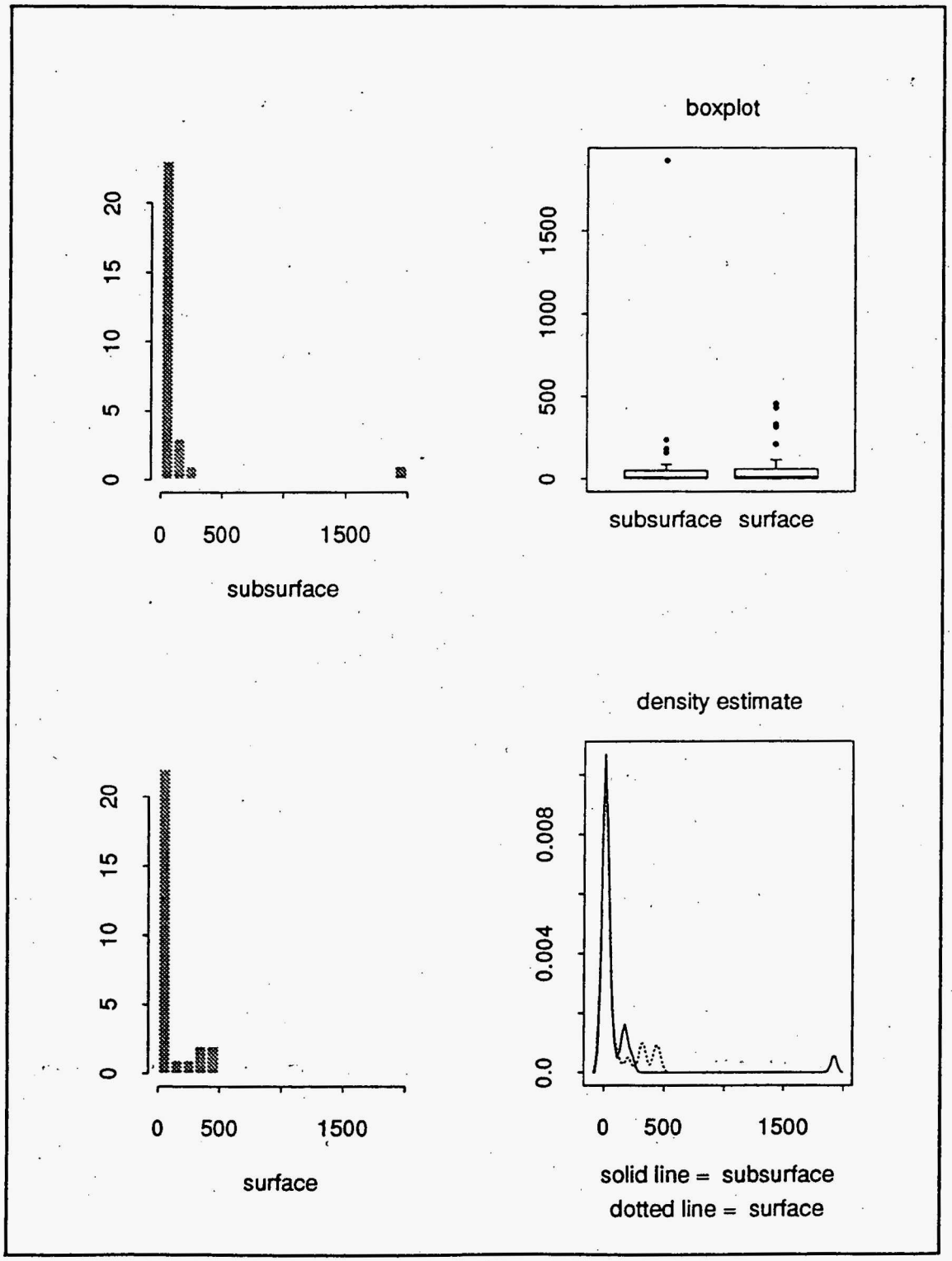


Figure C- 4 Comparable Surface and Subsurface Data Plots for Uranium-238

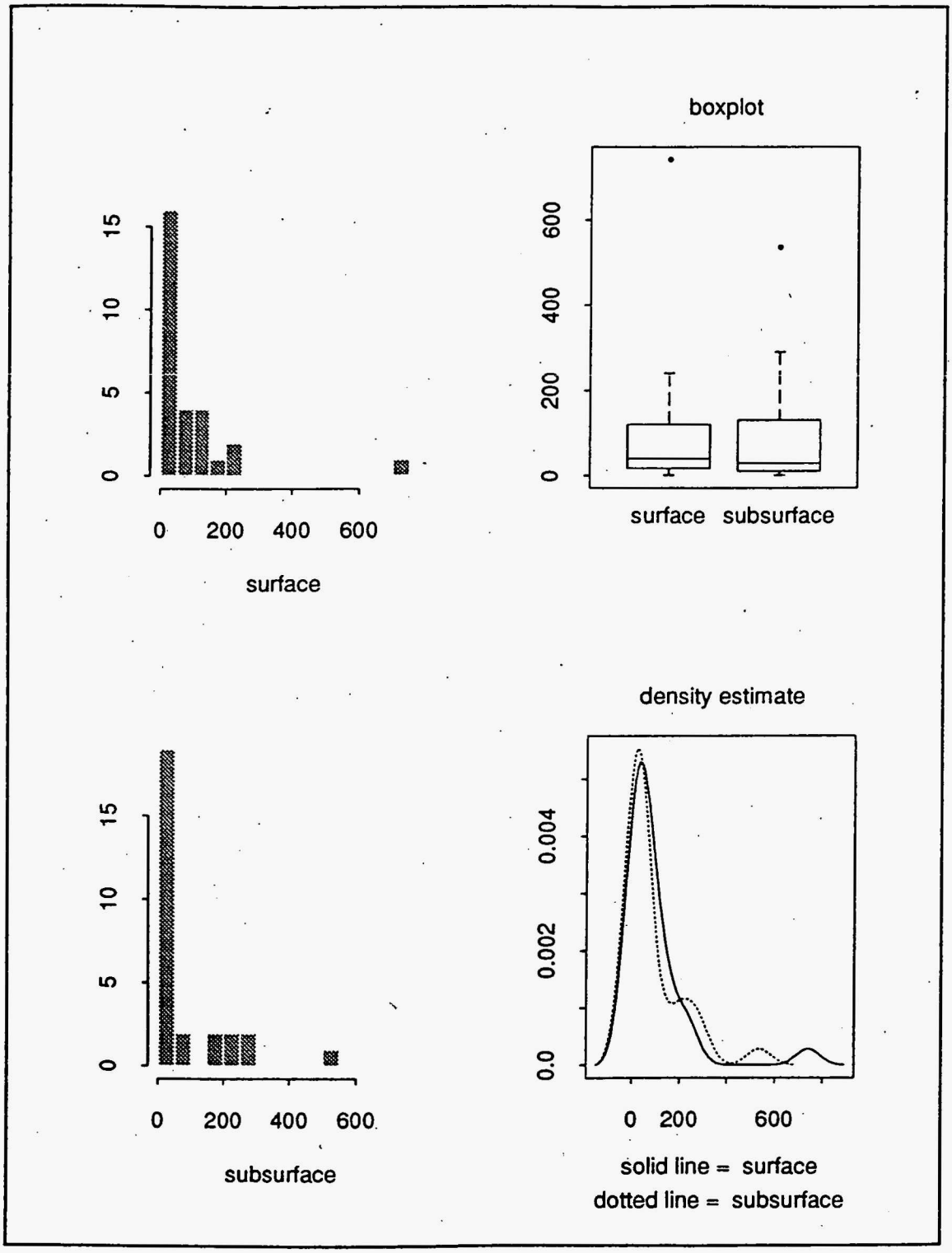


Figure C- 5 Correlation Plots for Surface and Subsurface Data

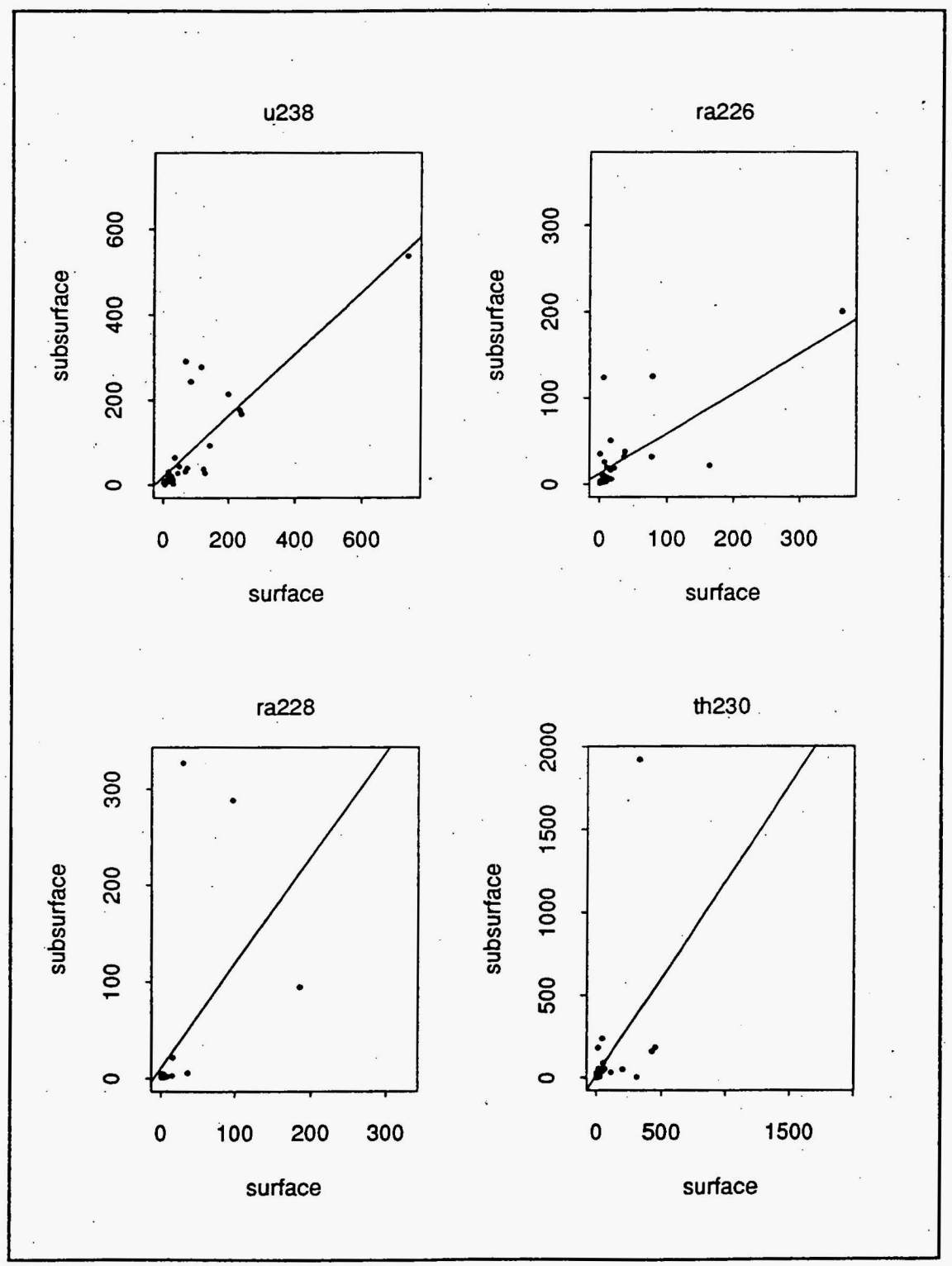
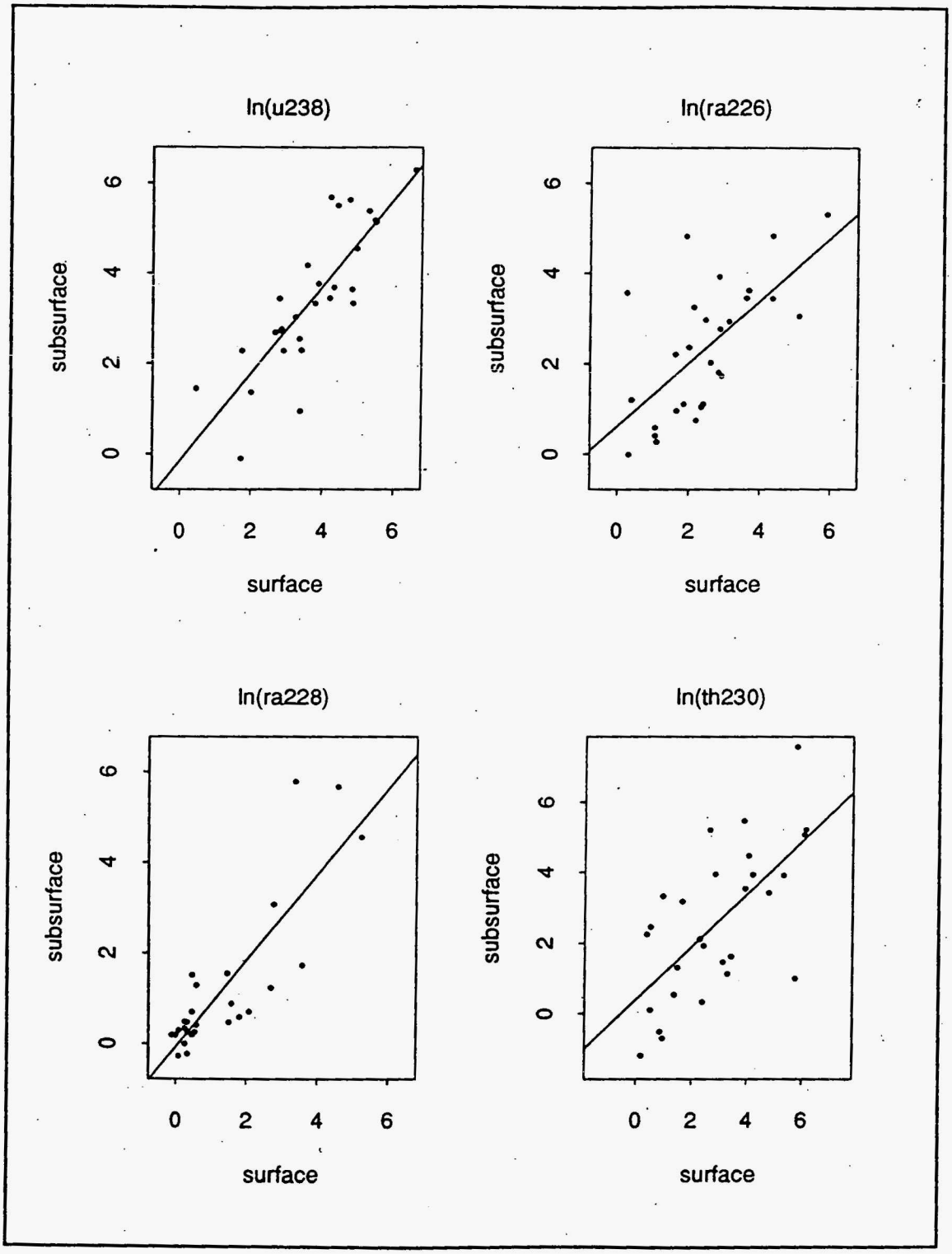
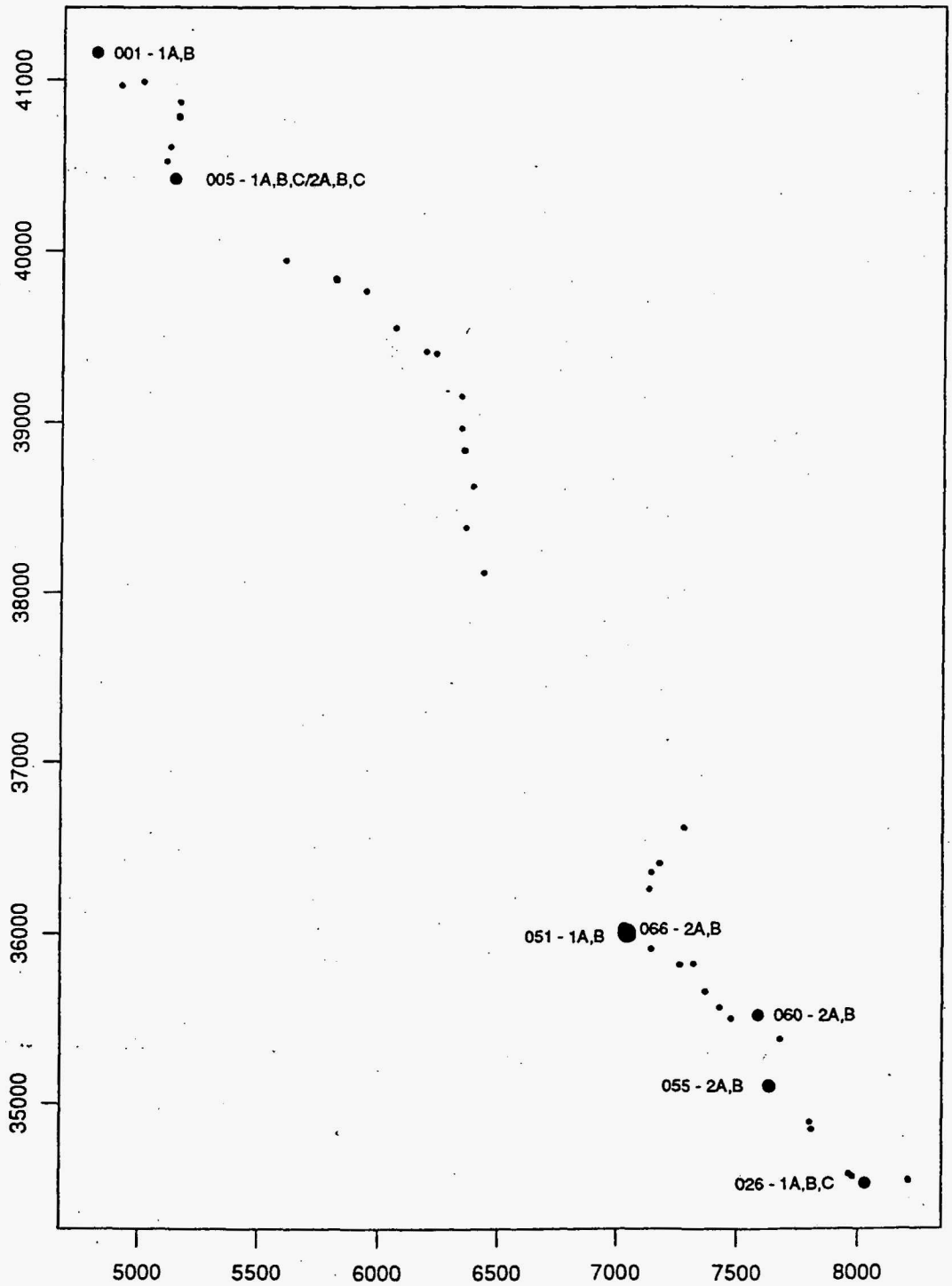


Figure C- 6 Correlation Plots for Surface and Subsurface Data - Log. Scale



APPENDIX D. SAMPLE LOCATIONS

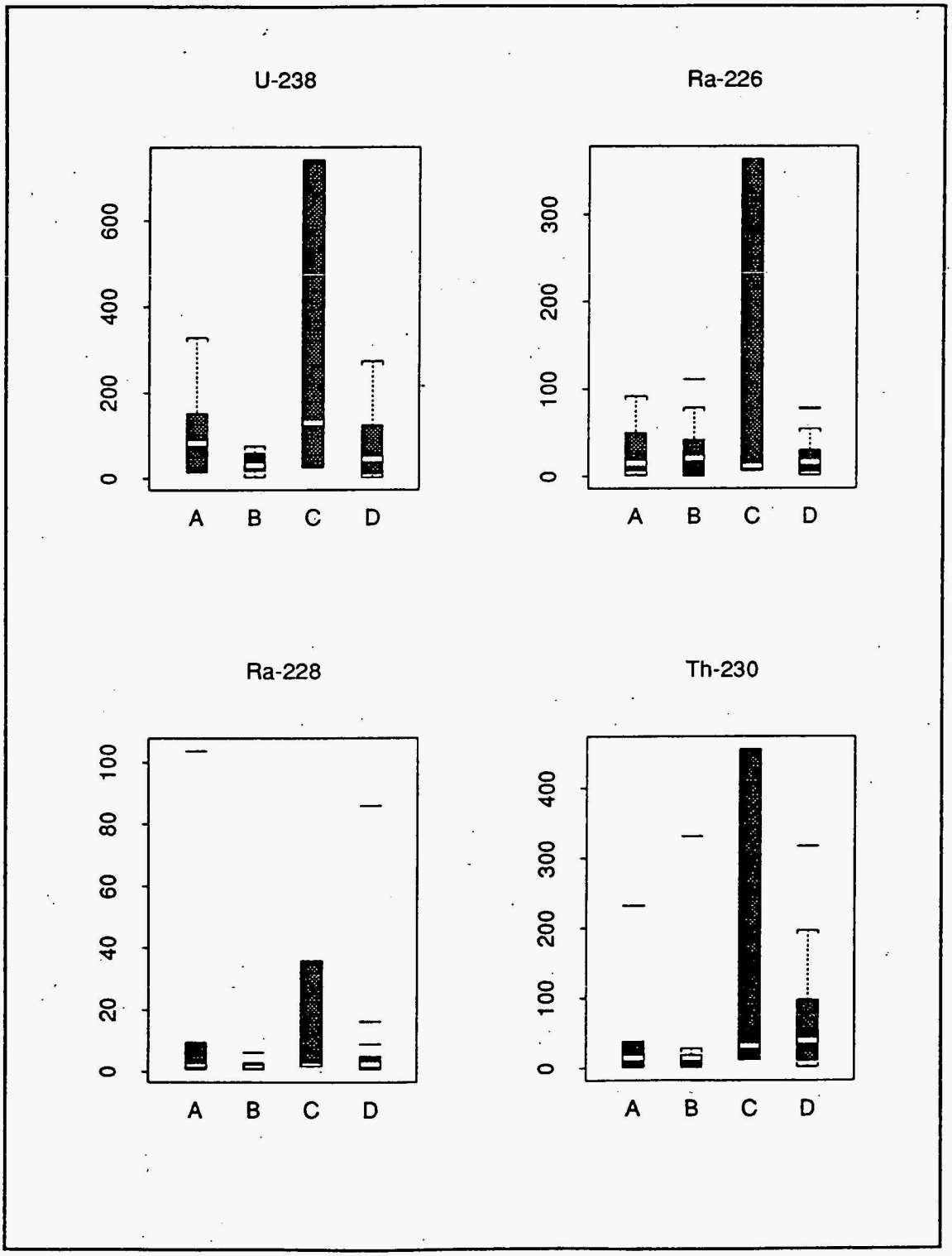
Figure D- 1 Data Locations at the SE Drainage Area



Large dots indicate locations with collocated samples.

APPENDIX E. COMPARISON OF ANALYTES ACROSS UNITS

Figure E- 1 Box Plots of each Analyte by Unit



The following Tables (Tables E-1 through E-4) provide summary statistical test results for determining differences in radionuclide activities between Units. Some observations should be made prior to interpreting these results. The first is that the non-parametric tests that were performed are not symmetric in their test output. In particular, these tests are not two-sided, and their results reflect, in part, the order of the inputs to the tests. Each of these tests is marked with an asterisk for identification. The results of these tests presented in the tables are those *p*-values that most closely indicate a difference between Units. The best interpretation can be given by considering these test results in conjunction with the box plots presented in Figure E-1.

For example, the Quantile test for uranium-238 indicates that the *p*-value for the difference between Units A and B is approximately 0.02. Notice that the box plot indicates that the uranium-238 activities may be greater in Unit A than Unit B. The value of 0.02 represents the *p*-value for the hypothesis that part of the distribution of uranium-238 activities in Unit A is greater than the distribution in Unit B. The *p*-value for the reverse null hypothesis is essentially 1, but the result 0.02 is the only one of these two results presented in Table E-1. In summary, the box plots indicate the likely direction of any potential differences between Units.

The second observation that should be made is that this form of statistical testing (i.e., performing many tests on the same set of data) may result in identification of significant results at a fixed α level (say 0.05 or 5%) due primarily to performing so many tests. It may be more appropriate to make corrections to the reported *p*-values, or to compare the *p*-values to a more stringent significance level (say, 0.01 or smaller) to adjust for the number of tests that are performed on the same data.

The final observation that can be made is that statistical tests are, perhaps, most appropriately used to verify observations made graphically about the data. Figure E-1, for example, indicates that uranium-238 and radium-228 activities in Unit B may be less than activities of these radionuclides in the other Units. If the test results support any differences at all they are these differences (there are more marginal indications that similar differences exist for thorium-230 activities in Units C and D compared to Unit B).

It may be difficult to support the graphical conclusions based on the statistical tests for the following reasons: the large number of tests performed; the tests were performed after seeing the data; the two variable non-parametric test results presented are those for the potentially favorable direction of differences; and the overall Kruskal-Wallis tests indicate no differences between Units. At best, activities in Unit B may be somewhat lower than activities in the remaining Units, with the exception that there appears to be no statistical difference in activities between Units for the radium-226 isotope.

Table E- 1 Differences Between Units: Statistical Test Results for Uranium-238

	Difference between Pairs of Units					
Test	A and B	A and C	A and D	B and C	B and D	C and D
t	0.13	0.48	0.44	0.36	**0.08**	0.42
Mann-Whitney*	0.20	0.25	0.16	**0.08**	0.32	0.69
Quantile*	**0.02**	0.49	0.19	**0.09**	**0.06**	0.52
Slippage*	**0.02**	0.27	0.28	**0.03**	**0.02**	0.13
Kruskal-Wallis* overall test for differences between Units						0.38

Numbers presented are observed significance levels or *p*-values.

Table E- 2 Differences Between Units: Statistical Test Results for Radium-226

	Difference between Pairs of Units					
Test	A and B	A and C	A and D	B and C	B and D	C and D
t	0.95	0.50	0.57	0.50	0.49	0.47
Mann-Whitney*	0.50	0.46	0.37	0.38	0.45	0.34
Quantile*	0.57	0.49	0.54	0.55	0.67	0.52
Slippage*	0.58	0.27	0.28	0.21	0.11	0.13
Kruskal-Wallis* overall test for differences between Units						0.96

Numbers presented are observed significance levels or *p*-values.

Table E- 3 Differences Between Units: Statistical Test Results for Radium-228

	Difference between Pairs of Units					
Test	A and B	A and C	A and D	B and C	B and D	C and D
t	0.30	0.87	0.51	0.43	0.23	0.65
Mann-Whitney*	0.22	0.27	0.51	**0.07**	**0.09**	0.78
Quantile*	0.18	0.49	0.19	0.55	0.31	0.52
Slippage*	**0.06**	0.73	0.28	0.21	0.42	0.88
Kruskal-Wallis* overall test for differences between Units						0.39

Numbers presented are observed significance levels or *p*-values.

Table E- 4 Differences Between Units: Statistical Test Results for Thorium-230

	Difference between Pairs of Units					
Test	A and B	A and C	A and D	B and C	B and D	C and D
t	0.95	0.48	0.45	0.48	0.42	0.57
Mann-Whitney*	0.36	0.19	0.13	0.11	**0.07**	0.31
Quantile*	0.18	0.49	0.46	**0.09**	0.31	0.52
Slippage*	0.58	0.27	0.72	0.21	0.34	0.13
Kruskal-Wallis* overall test for differences between Units						0.29

Numbers presented are observed significance levels or *p*-values.

Notes on the testing procedures:

The *t*-test tests for a difference between mean concentrations for two data sets.

The Mann-Whitney test involves ranking the combined data and determining if the sum of ranks for one data set is significantly different than the sum of ranks for the other data set.

The Quantile test also involves ranking the combined data set, but then considers if there are a disproportionate number of observations from the separate data sets in the top 20% of the combined data (any quantile can be used; 20% was used for the tests presented above).

The Slippage test determines if the number of observations in one data set that exceed the maximum observation on the other data set is disproportionate.

The Kruskal-Wallis test is a generalization of the Mann-Whitney test that ranks the combined data from all four (in this case) data sets, and then determines if the sum of ranks for the individual data sets are significantly different.

APPENDIX F. MEAN ESTIMATES BASED ON LOGNORMALITY

The radiological data indicate a positive skew (skew to the right) across radionuclides and Units. Consequently, it may be more appropriate to consider the data using underlying lognormal assumptions as opposed to underlying normal assumptions. Table F-1 presents summary statistics for the radionuclides, by Unit, that were generated using lognormal distribution theory. The summary statistics were generated according to a procedure described in Gilbert (1987), in which, for example, the mean estimate presented is a *minimum variance unbiased estimate* (MVUE), generated according to the iterative formulas offered in Gilbert (1987, Ch. 13).

Note that the estimated means for radium-226 are, again, all greater than the target risk level for this radionuclide of 13 pCi/g, in which case the earlier decisions made based on normal assumptions are corroborated when lognormal assumptions are used instead.

Table F- 1 Mean Estimates (MVUE) Based on Lognormal Assumptions

Unit	Uranium-238	Radium-226	Radium-228	Thorium-230
A	110	34	10	53
B	43	49	1.9	34
C	300	100	11	140
D	91	24	4.7	92
ALL	94	37	5.0	89

Units - pCi/g.

APPENDIX G. POWER PLOTS AT TARGET RISK LEVELS

Interpretation

Table 1-1 provided target risk levels for the child scenario for the radionuclides for which data are available. The parameters that are available for this DQA are described in Appendix B - i.e., α, β, R, δ, C, s, and n. The plots included in this Appendix provide, for each radionuclide, power curves for given target risk levels, R, and Type I Error rates, α.

For example, the power curves in Figures G-1 through G-4 portray the effect on acceptable Type II errors of changing the sample size and standard deviation for a fixed target risk level of 13 pCi/g (for radium-226 or radium-228) and fixed Type I Error rate of 1%, 5%, 10%, and 20%. Four sample sizes are depicted, 3, 8, 11, and 22 (varied within plot), corresponding to the actual sample sizes for Units A, B, C and D, and several possible values of standard deviations are included (varied from plot to plot).

Notice that as the sample size increases (for example, in Figure G-1a), the effective region of indecision shrinks to reflect that there is more information available from the increased sample size. There is less uncertainty as the sample size increases. Also notice the effect of increasing the standard deviation is to increase the size of the effective region of indecision; there is greater uncertainty as the standard deviation increases. The series of power curves depicted in Figures G-1 through G-4 also demonstrate that as the significance level, α, increases the effective region of indecision shrinks. In other words, as the tolerance for making Type I Errors increases in probabilistic terms (allowing more decision errors to be made), then for other parameters fixed, the probability of making a Type II Error decreases. This effect is produced by the trade off between allowable decision errors. If all other parameters are fixed, then for a given sample size, as Type I Error is allowed to increase, Type II Error will decrease, and vice versa.

The first point to be recognized is that if the estimated mean of the data is greater than the risk threshold of interest, or more appropriately the critical value, C, then the decision that the null hypothesis cannot be rejected is supported by the data. That is, the site exhibits comparatively high radioisotope activities. The problem is more complex if the estimated mean concentration is lower than the critical value. Under such circumstances the power (or the probability of making a false positive decision error) must be considered. The following series of related examples may help interpretation of the power plots. Appendix B provides some further discussion along these lines.

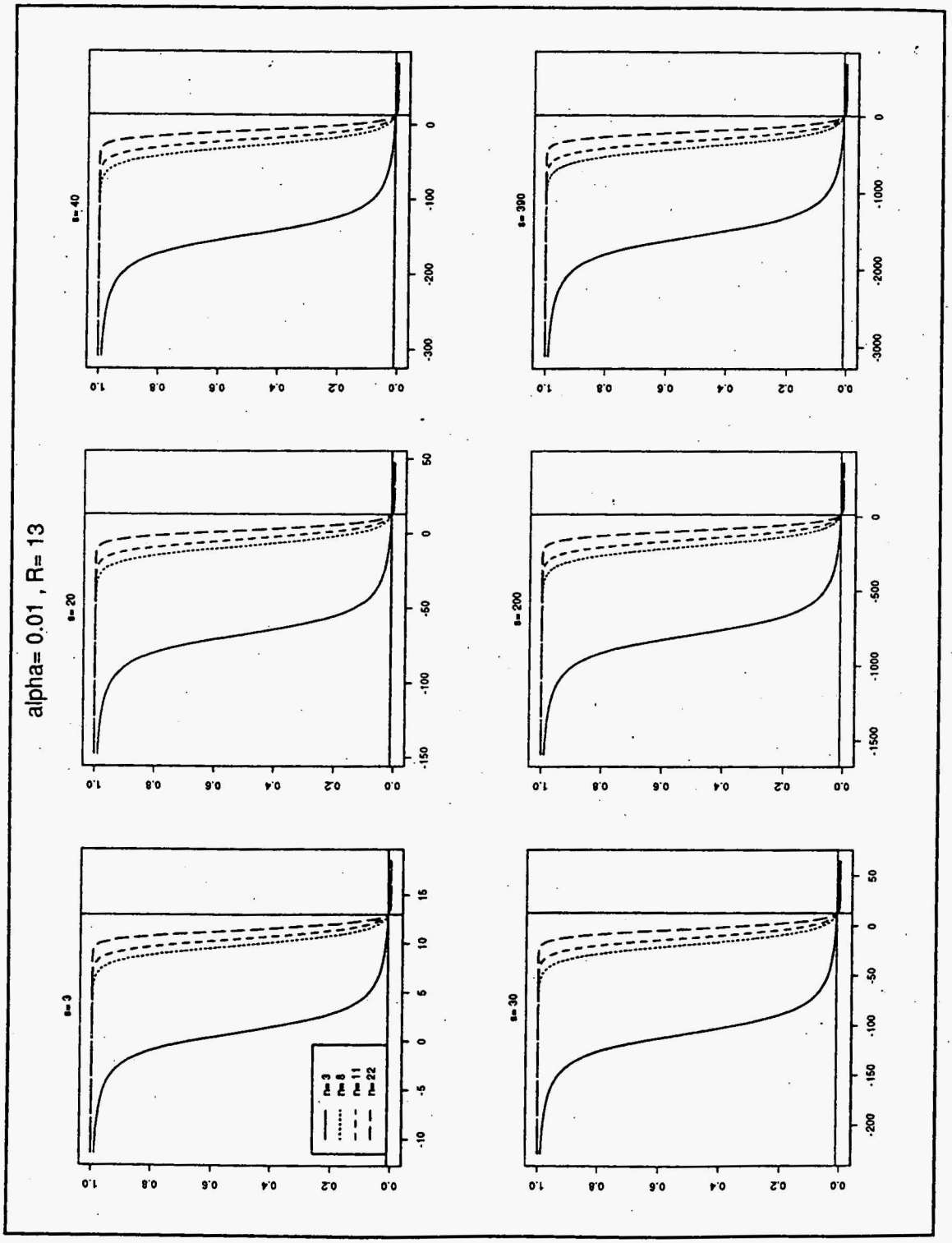
If a mean concentration of 3 pCi/g and a standard deviation of 3 pCi/g were to be considered for supporting a risk-based decision based on a target risk level of 13 pCi/g, then the first plot in Figure G-1 indicates that 8 samples are more than adequate to support such a decision at the 0.01 significance level, but 3 samples are not. The power for 8 samples is very close to 1, whereas for 3 samples the power is approximately 0.2 (corresponding to a false positive decision error rate of 80%!). Some number of samples between 3 and 8 is optimal depending on the tolerance for making false positive decision errors. Notice that as the standard deviation (or variability) increases, more and more samples are required to adequately support the decision. At a standard deviation of 20 pCi/g, even 22 samples are not sufficient to support a decision at the 0.01 significance level.

If the tolerance for making false negative decision errors (i.e., the significance level) can be relaxed, then 22 samples, for example, may be adequate. Figure G-2 depicts power curves under the same conditions except that the significance level is relaxed to 0.05 (corresponding to a 5% false negative decision error rate). The power, under the conditions given (same mean and standard deviation with 22 samples) has increased to approximately 0.7. If the significance level is relaxed again to 0.1 (Figure G-3) or even 0.2 (Figure G-4) then the power (based on 22 samples) increases to approximately 0.9 and 0.95. In general there is a trade off between probabilities of false negative and false positive decision error rates that can be tolerated. Figure G-5 through G-8 and G-9 through G-12 have similar interpretations, but the target risk levels are different (to reflect target risk levels for thorium-230 and uranium-238).

The Figures provide presentations of power curves covering a wide range of conditions. These power curves indicate conditions under which data collected may, or may not be, adequate for supporting decisions based on mean radioisotope activities.

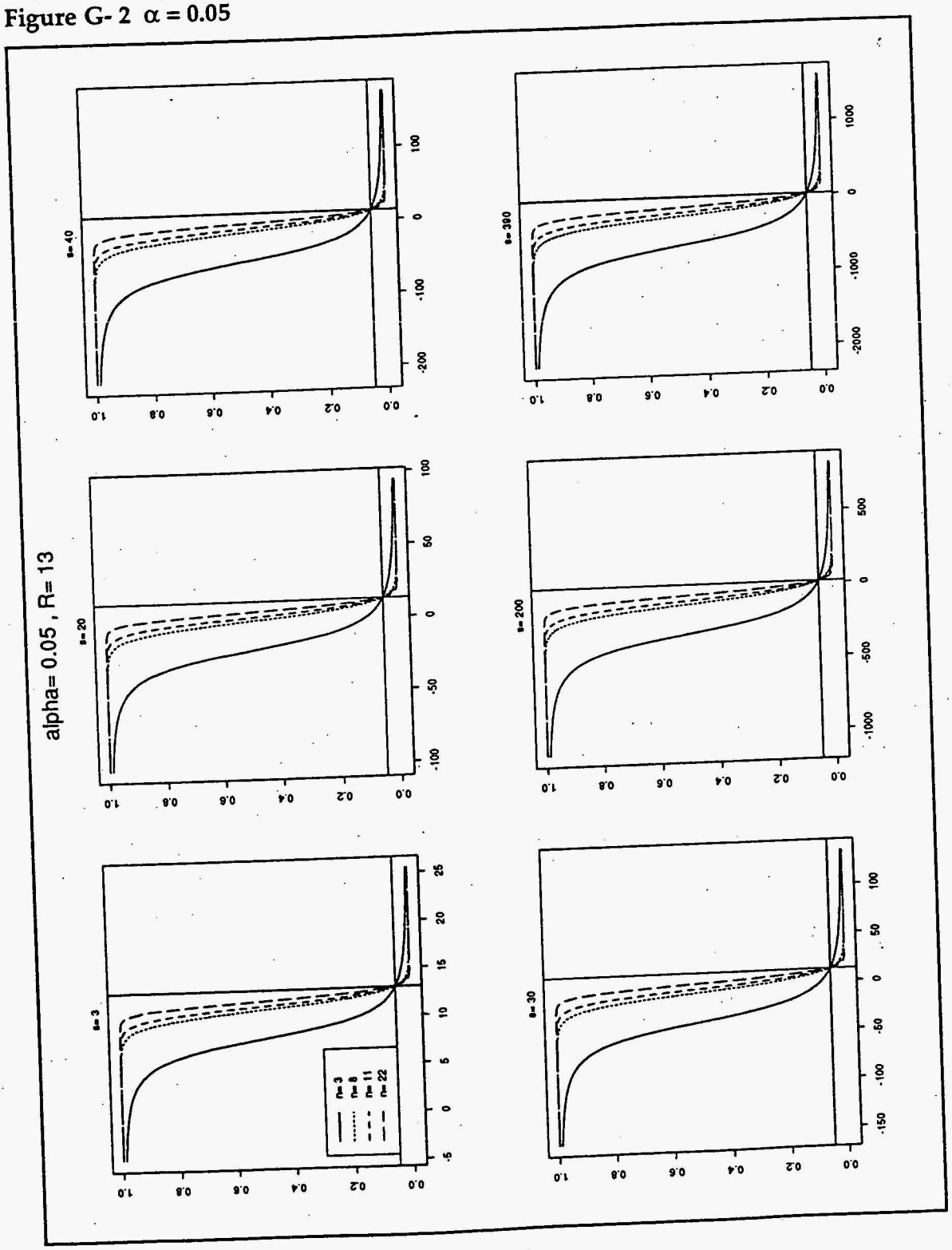
POWER PLOTS FOR A RISK LEVEL OF 13 pCi/g

Figure G- 1 $\alpha = 0.01$



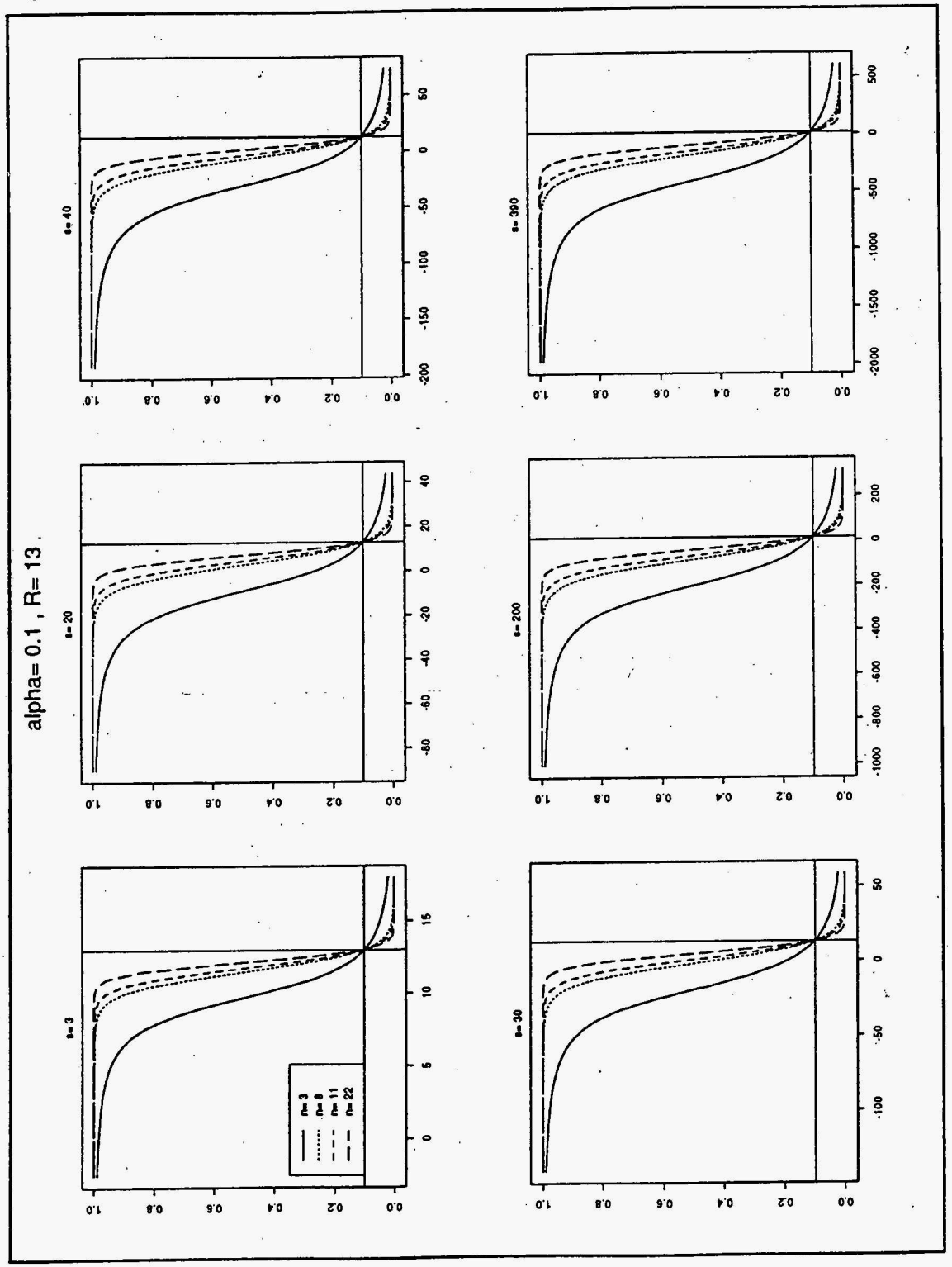
POWER PLOTS FOR A RISK LEVEL OF 13 pCi/g

Figure G- 2 $\alpha = 0.05$



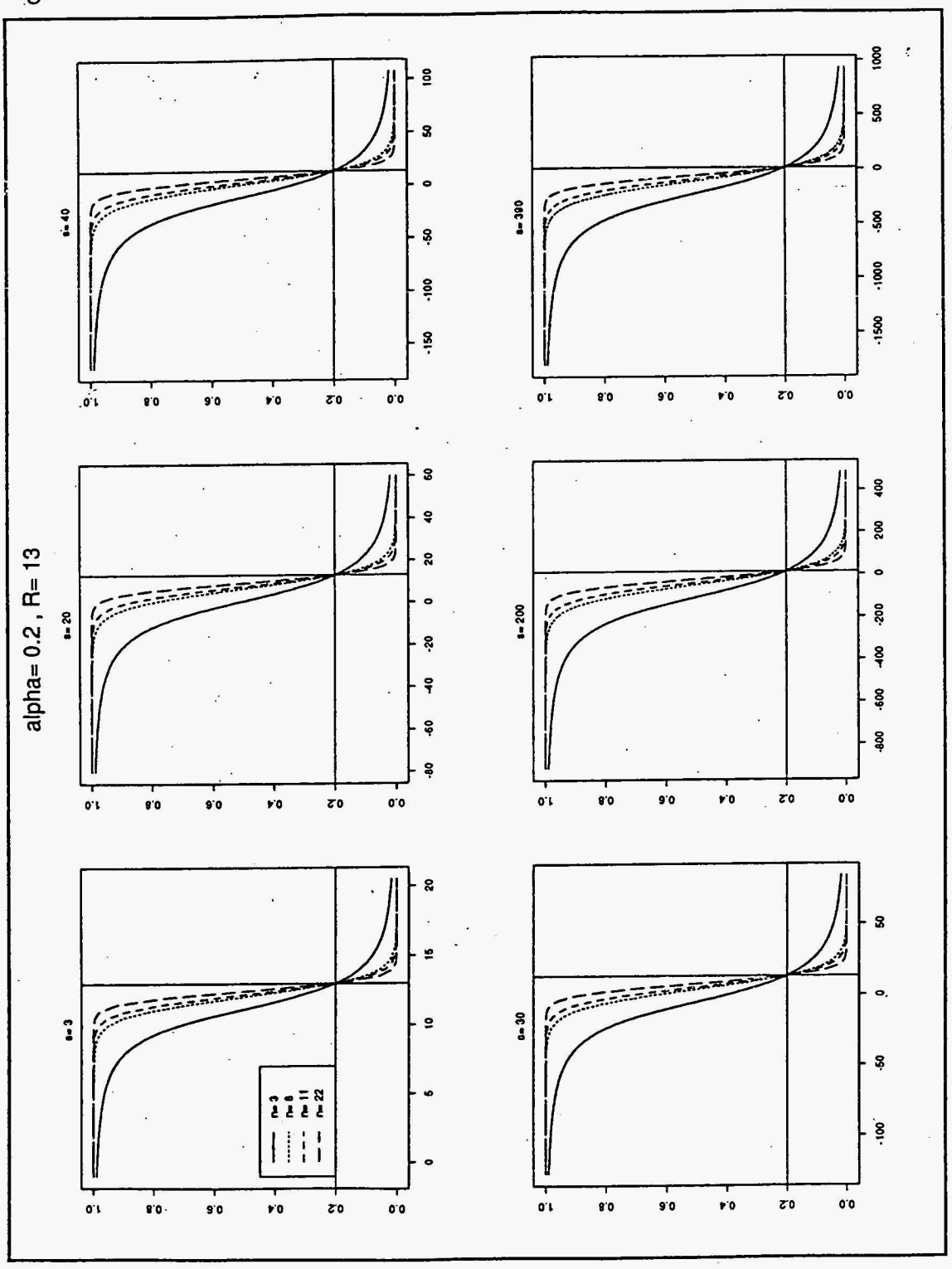
POWER PLOTS FOR A RISK LEVEL OF 13 pCi/g

Figure G- 3 $\alpha = 0.1$



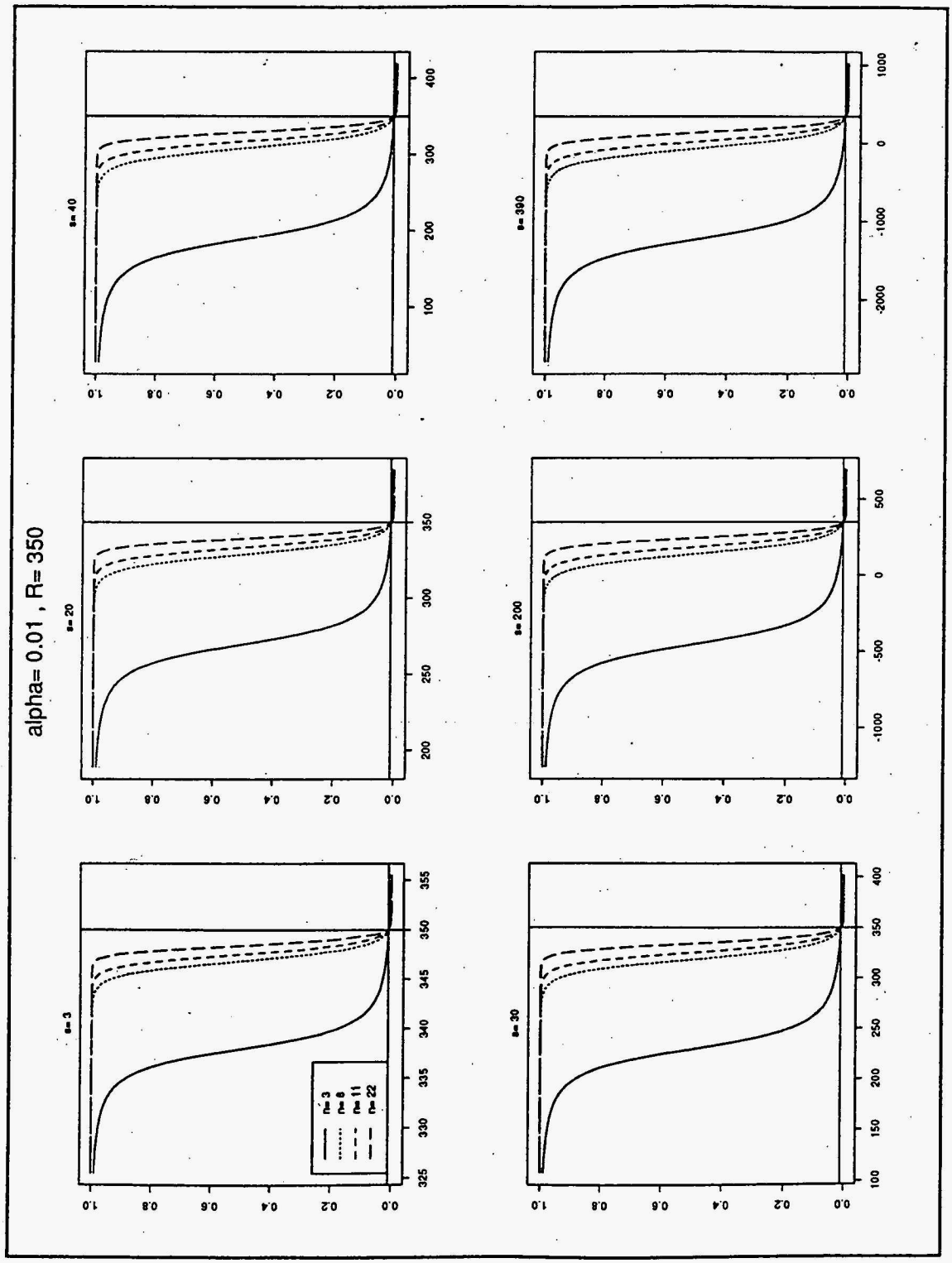
POWER PLOTS FOR A RISK LEVEL OF 13 pCi/g

Figure G- 4 $\alpha = 0.2$



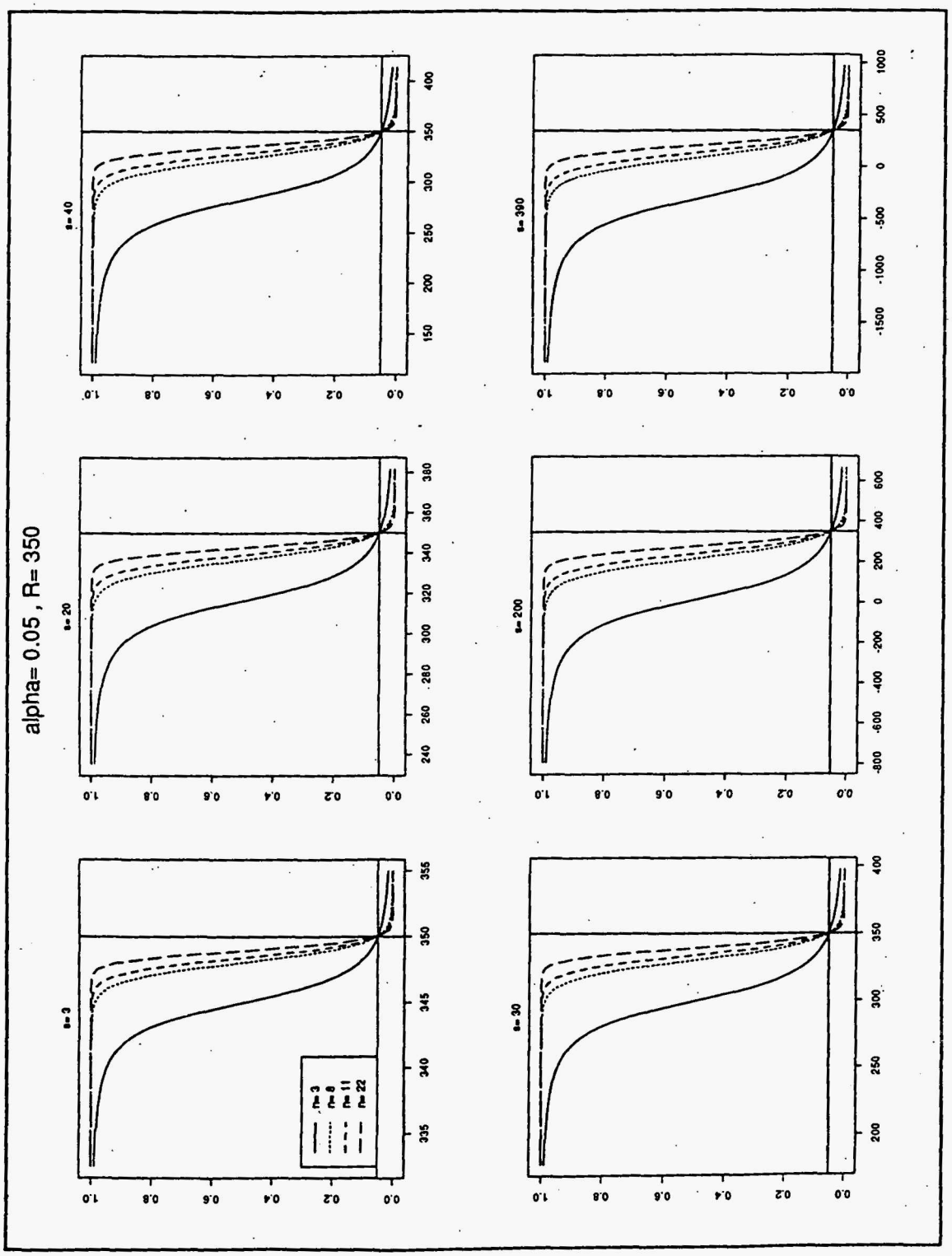
POWER PLOTS FOR A RISK LEVEL OF 350 pCi/g

Figure G- 5 $\alpha = 0.01$



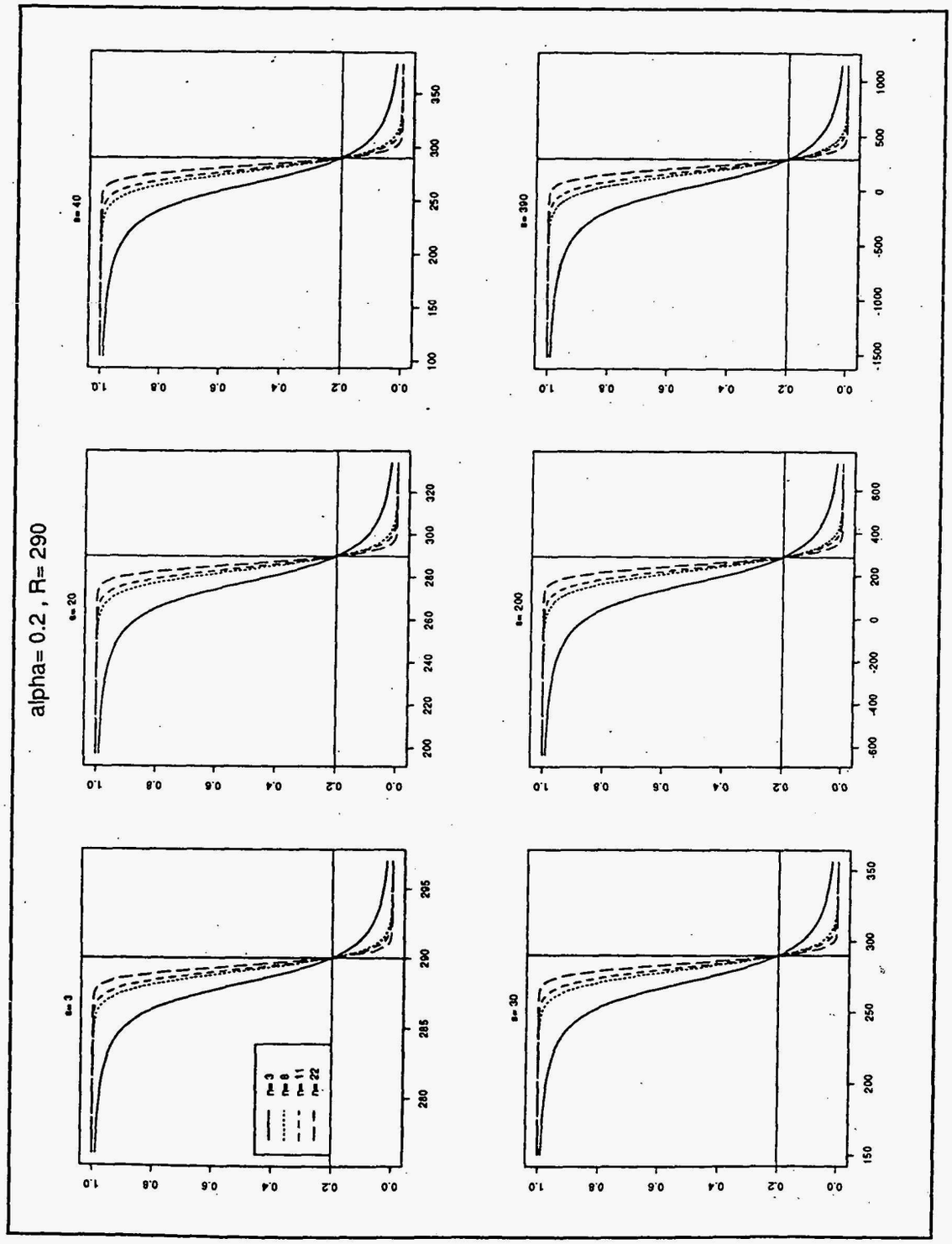
POWER PLOTS FOR A RISK LEVEL OF 350 pCi/g

Figure G- 6 $\alpha = 0.05$



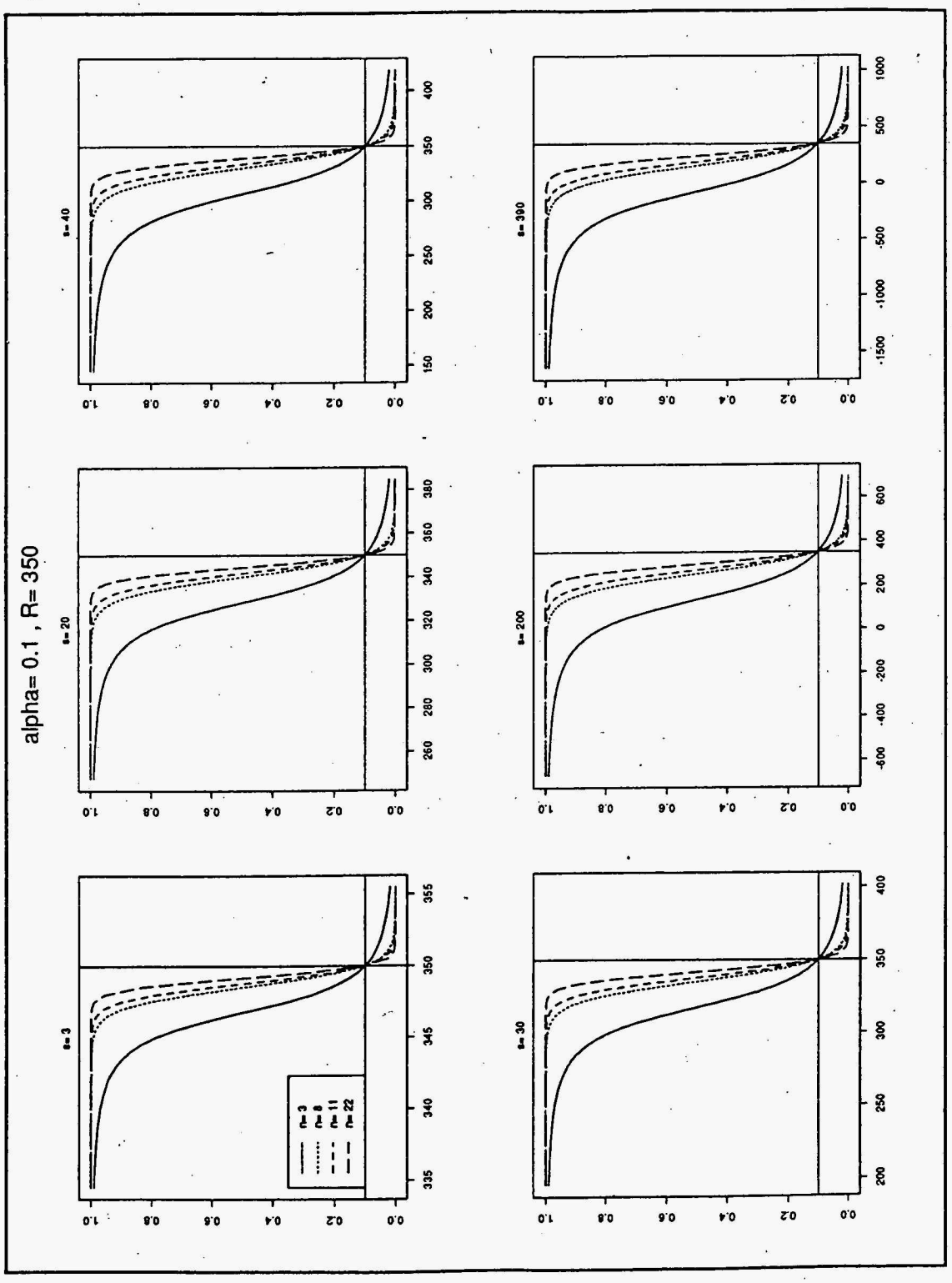
POWER PLOTS FOR A RISK LEVEL OF 290 pCi/g

Figure G- 12 $\alpha = 0.2$



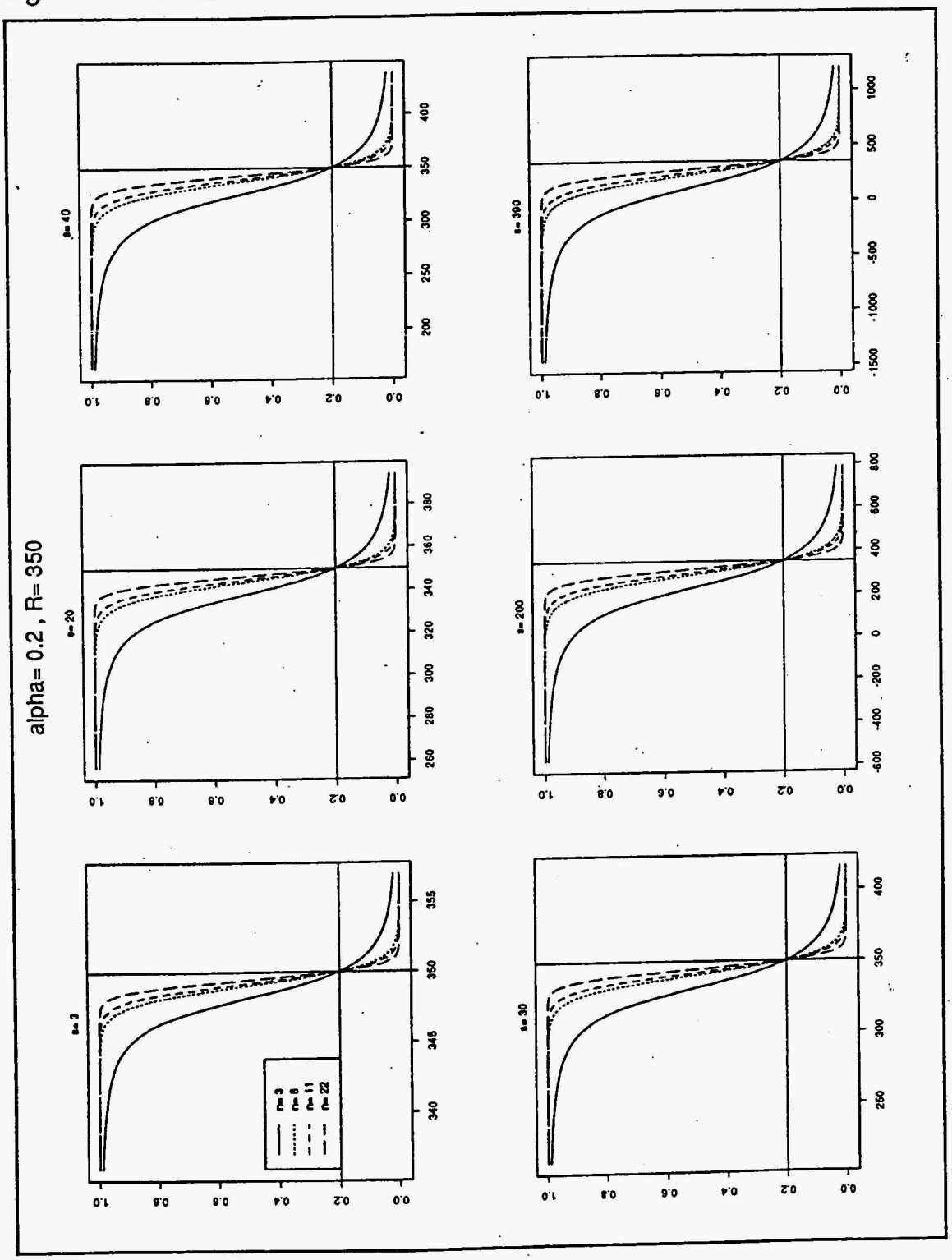
POWER PLOTS FOR A RISK LEVEL OF 350 pCi/g

Figure G- 7 $\alpha = 0.1$



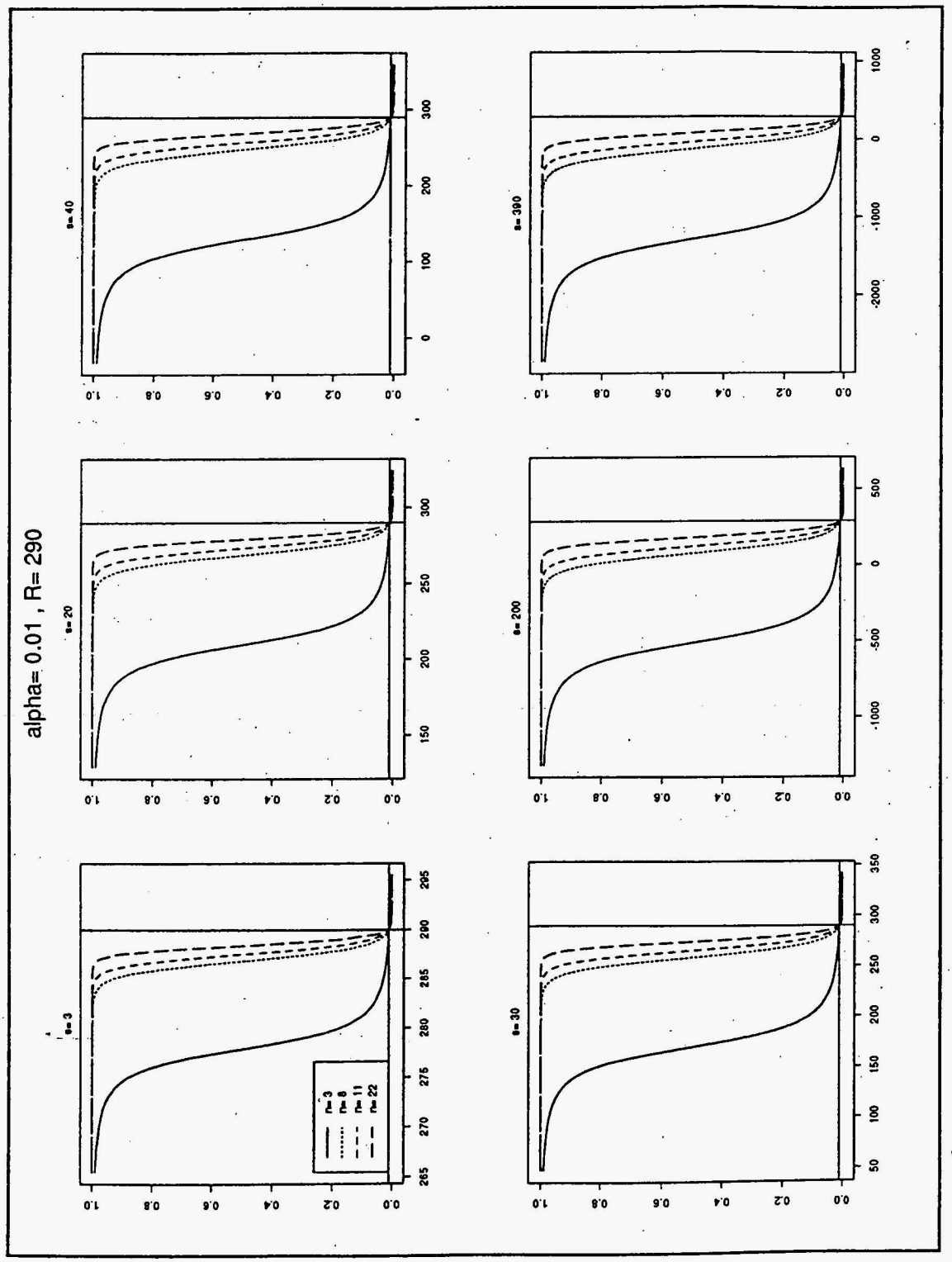
POWER PLOTS FOR A RISK LEVEL OF 350 pCi/g

Figure G-8 $\alpha = 0.2$



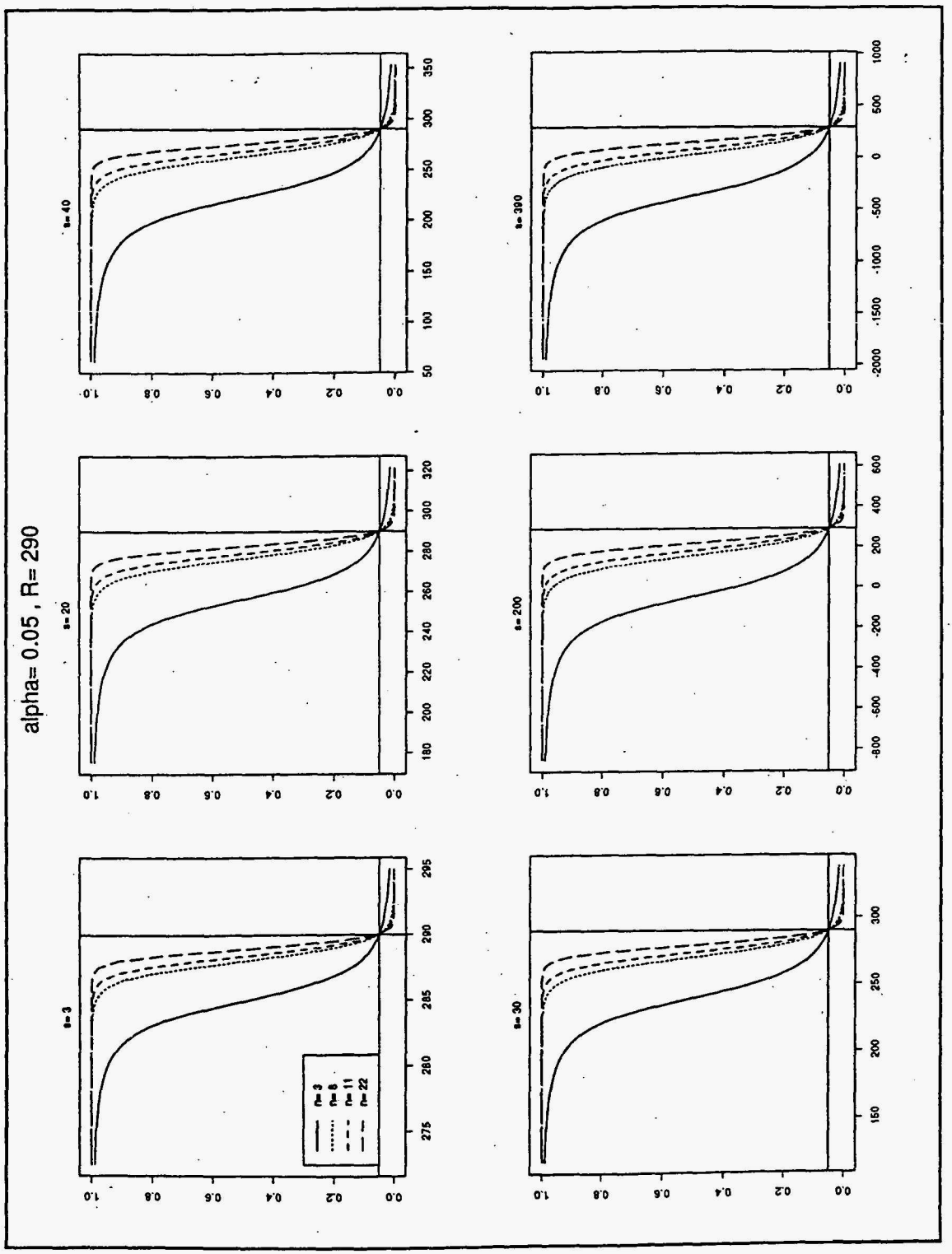
POWER PLOTS FOR A RISK LEVEL OF 290 pCi/g

Figure G- 9 $\alpha = 0.01$



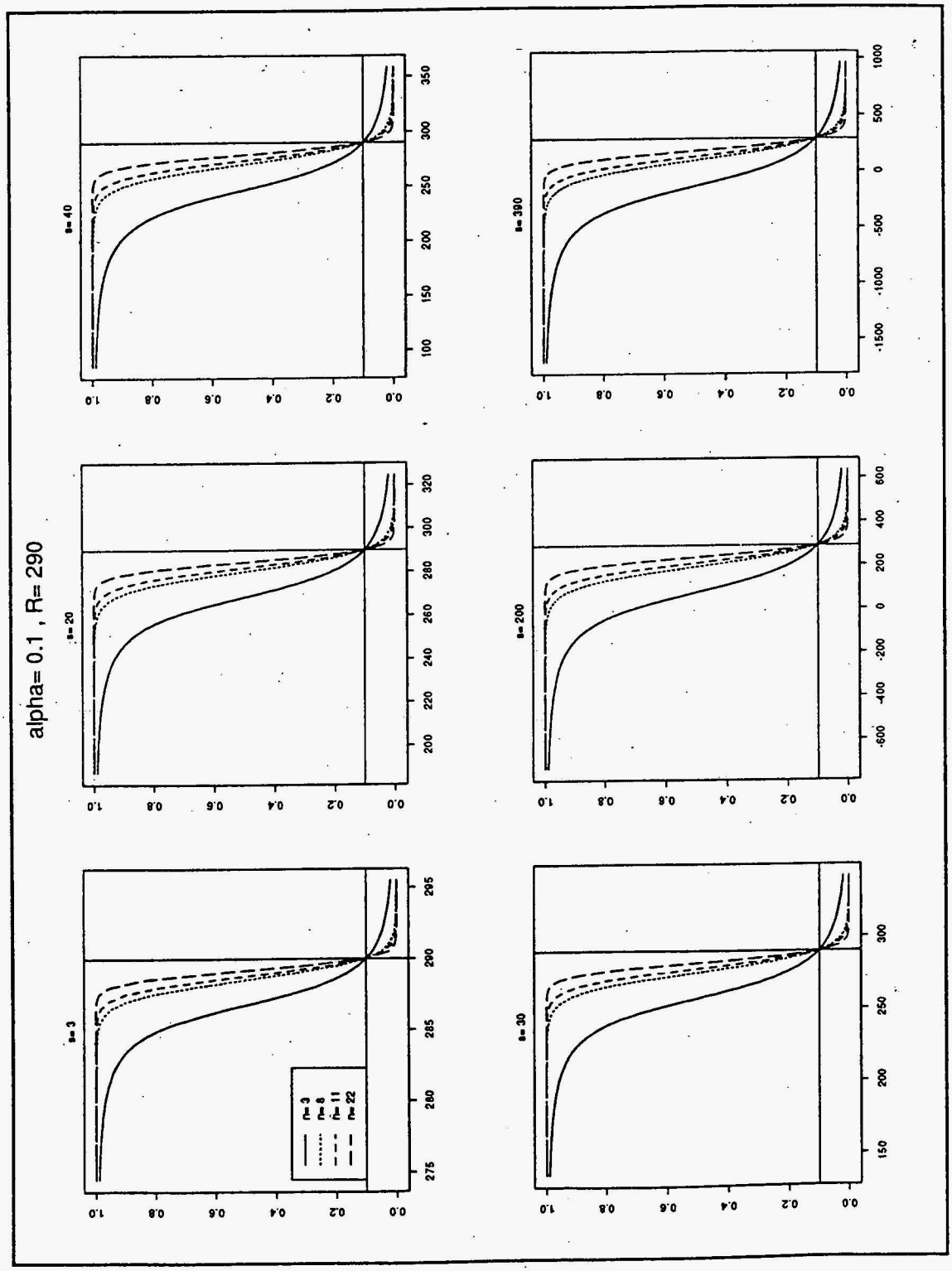
POWER PLOTS FOR A RISK LEVEL OF 290 pCi/g

Figure G- 10 $\alpha = 0.05$



POWER PLOTS FOR A RISK LEVEL OF 290 pCi/g

Figure G- 11 $\alpha = 0.1$



POWER PLOTS FOR A RISK LEVEL OF 290 pCi/g

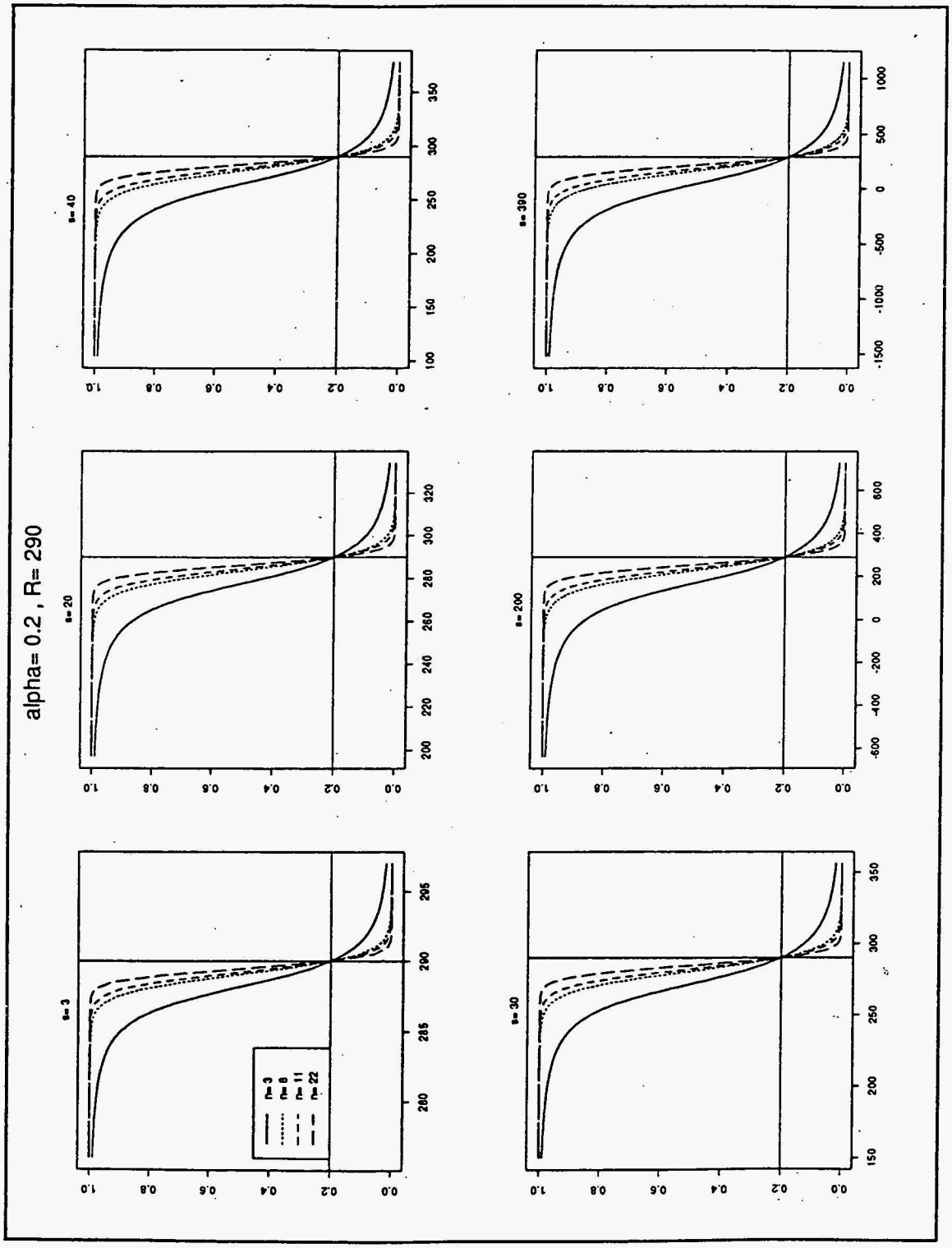



Figure G- 12 $\alpha = 0.2$