THE DISTRIBUTION OF INDOOR RADON CONCENTRATIONS IN A LARGE POPULATION OF HOMES

Willard E. Hobbs
Radon Reduction & Research

Abstract-- One of the least understood qualities of radon and its associated risk is the relationship between its concentration at the source and the statistical distribution of concentrations found in indoor environments. In this presentation I will clarify several of the important factors and provide a simple statistical model for discussing and communicating distribution qualities of our primary source of ionizing radiation. A few of these are as follows: (1) Radium, like most other minerals in the earth's crust, has an overall log-normal distribution in concentration; (2) The factors which affect the concentration of indoor radon do so in a multiplicative fashion; (3) The measured data sets of indoor radon concentrations follow a log-normal distribution which is characterized by a skewed distribution with a long tail to hazardous levels; (4) A fundamental comparison property for such a distribution is the geometric ratio of levels and not the arithmetic difference; and (5) The response of the soil gases to the building stock and ventilation parameters thereof provide an additional factor of approximately 2. These facts provide critical information for developing health strategies for states, communities and individual homeowners.

Introduction

In this paper, information on the distribution of radon concentration in California homes will be discussed. Section II summarizes qualitative information on the lognormal distributions. (Note: An appendix at the end of the paper provides some technical details on this distribution.) Section III discusses the requirement for a scientific sample in studying radon distributions is discussed. Section IV explains how various factors result in the wide spread of observed radon concentrations and their relative magnitudes. In Section V data from the California counties and from Summerland, California, provide understanding of the variance in the observed radon levels. Finally, in Section VI the practical implications of these results and how they can aide in the reduction of radon exposure to the general population.

The Lognormal Distribution

It is well-known that the range of radon concentrations found in US homes is very broad. The result is that the specific concentration in a given home, chosen randomly, cannot be accurately predicted. The best way to determine the radon concentration in a home is to make a measurement. When analyzing data sets, the distribution of radon concentrations is not a common distribution, but a right-skewed one called a lognormal distribution. A lognormal distribution is simply one where the logarithms of the data values have a normal (Gaussian) distribution. (The distribution name is therefore descriptive.) Since the distribution is strongly skewed it has a long tail extending to high magnitudes. Thus, there is usually a reasonable probability that any house could contain a radon concentration that is unacceptable.

The normal distribution is characterized by two parameters: the mean (location of the center) and the standard deviation (the relative degree of the spread). For a lognormal distribution, these are called the geometric mean and the geometric standard deviation. The geometric mean for a sample of n data points is defined as the n^{th} root of the product of values. This is equal to the exponential of the average of the logs of all the data points (any units and any base can be used). The geometric standard deviation (GSD) does not have a simple arithmetic algorithm like the geometric mean. It is defined as the exponential of standard deviation of all the log-data. This is a dimensionless factor which has the same value no matter what units or exponent/log base is used. Conventional units (pCi/liters) and natural logarithm base = 2.718281828 are used here.

The role of the distribution parameters are reflected in the shape of the lognormal distribution. Let (μ_g,σ_g) be the mean and standard deviation of the log (data values) respectively; the geometric mean and geometric standard deviation are M=exp(μ_g) and GSD=exp(σ_g). For a normal distribution about 68% of the distribution is found within one standard deviation of the mean. Thus, if a house is chosen at random, there is about probability P = 68% that the log of its radon concentration (henceforth called logradon) will be in the interval

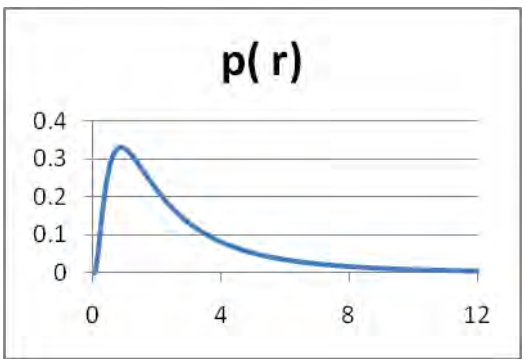
$$\mu_g - \sigma_g < \ln r < \mu_g + \sigma_g \qquad (1)$$

or equivalently

$$M / GSD < r < M \times GSD \qquad (2)$$

There is about 95% probability that the log-radon value will be in the interval $\mu_g \pm 2\sigma_g$, or equivalently the radon concentration itself would be within a factor of GSD^2 of M.

The geometric mean is equal to the median or 50^{th} percentile, the population center of the distribution. To illustrate the lognormal distribution using a specific distribution, suppose the median is 2 pCi/liter and the geometric standard deviation is 2.5. The radon concentration in a randomly chosen house would have 68% probability of being in the interval 0.8 to 5.0 pCi/liter and 95% probability of being the interval 0.32 to 12.5 pCi/liter. This distribution has rather high radon levels; the probability of a house having radon greater than 4 pCi/liter (z-value = [ln(4)-ln(2)]/ln(2.5) = 0.756) is 22.5%. It is convenient to think of the log of the radon level as the basic data unit. The distribution of the logs will be near-normal and the various normal statistical applications are applicable. For example, it is appropriate to do such things as estimate the population parameters from sample data or compare two populations using sample data. In Fig. 1, graphs of the probability distributions for $P(\mu_g=\ln(2),\sigma_g=\ln(2.5))$ are shown:



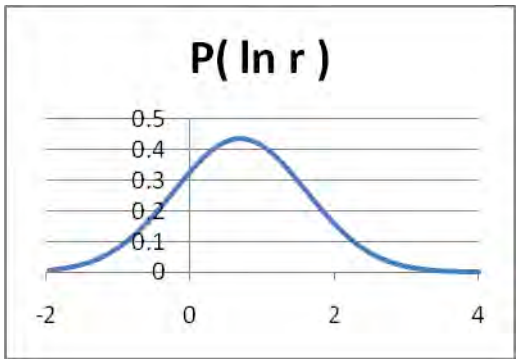


Figure1. Lognormal distribution of radon with median M=2 pCi/l and GSD=exp(σ_g)=2.5. Increasing the geometric standard deviation flattens the lognormal curve just like increasing the standard deviation expands, and flattens, the more familiar bell-shaped normal curve. It greatly increases the probability that a portion of the distribution will be found at high magnitudes. The GSD is therefore a critical parameter for the lognormal curve.

As an example, we know about the standard normal curve with μ=0 and σ=1. If these data were log values, then we would have e^{μ}=1 and e^{σ}=e=2.718. The curves for this appear very similar to those shown in Fig. 1. The bell-shaped curve is shifted over about one unit to the left. For the skewed lognormal curve, values for the horizontal axis are multiplied by 0.5 and the vertical axis is multiplied by 2.0 (the area under the curve remains 1=100%).

To emphasize the skewness of the lognormal distribution here are some other parameters of the distribution shown in Fig. 1. The most likely value (the mode) is equal to 0.86, the median (given) is 2.0, and the arithmetic mean is 3.0, all in pCi/liter. The first quartile (25%-tile) is 1.1 pCi/l and the third quartile (75%-tile) is 3.7 pCi/l (note: a factor of 1.85 on either side of 2.0). The standard relation for the normal curve is ln $r = \mu_g + z\ \sigma_g$; this allows the z-value to be related to the probability value for a range of radon levels. For radon between ln(r_1) and ln(r_2),

$$P(\ r_1 < r < r_2\ ) = normalcdf\ (\ln r_1,\ \ln r_2,\ \mu_g,\ \sigma_g\ ) = normalcdf\ (\ z_1,\ z_2,\ 0,\ 1\ ) \qquad (3)$$

where *normalcdf* is the cumulative distribution function for a normal curve. z-values are used with the standard normal curve and the probabilities are easily evaluated using a table, a calculator, or the internet.

Scientific sampling

The discussion in Section II, while certainly valid, assumes knowledge of the population geometric mean and geometric standard deviation which are fundamentally unknowable quantities. This technically involves information about the whole population, and, if that were known, there would be little reason to calculate the summary values. So, it seems to have limited value. The remedy is to use scientific sampling to estimate the required parameters from a manageable number of data.

The gold-standard for a scientific sample is called a “simple random sample.” This is a sample created where all samples of a given size n have an equal chance of being chosen. It is analogous to putting the names of all members of the population into a hat, shaking them up, and drawing

the sample. This is impractical for large populations. It requires that the whole population be delineated. A multistage sampling procedure is usually taken where demographics are considered and an honest attempt to achieve a representative sample is made. This is called an unbiased or scientific sample.

Various collections of radon data have been made, but most of them involve what is called voluntary or self-selected data. Voluntary samples are notoriously bad. One of the most dramatic cases was illustrated by Ann Landers in her newspaper column of Nov 3, 1975. She had surveyed her readers to determine the fraction that regretted having their children. An overwhelming number (~70%) of those who replied said that they did regret having their children. Of course, most of those who replied were unhappy so naturally they were negative. And clearly, this is not representative of the whole population.

A self-selection process is at work in radon measurements also. I use as an example in my own zip code 93105. As we have determined (Hobbs, 1996), by counting houses from aerial photographs, there are 10,167 houses in this zip code with 275 (2.7%) on the high-uranium Rincon formation. (This is a little higher ratio than that the total city of Santa Barbara: 27,225 homes with 325 on the Rincon.) Based on this, we calculated that as many as 5% of the homes in 93105 would be above 4 pCi/liter. Yet the California DHS data base (Blood, 2002) shows that 54% of the homes tested in 93105 are above 4. To their credit, the web site does caution readers that these results were voluntarily submitted and they do not constitute a scientific survey. They urge everyone to do their own test.

The values for 93105 are biased by the over-representation of the Rincon values. I know this because I contributed many of the values in this data base for 93105. There is a large Rincon downwash area in the eastern part of the zip code that has most of the high radon homes. Most of the homes I test are for real estate transactions. The real-estate agents and geologists know that this is an area of high radon so they emphasize testing to their clients. Whenever I am asked about a home in this area, I reply, "This is a high potential area, you definitely should test for radon."

In proving foresight may be vain:
The best laid schemes of mice and men
Go often askew,
And leave us nothing but grief and pain,
For promised joy!

In the early 90's, as part of the Radon Abatement Act, the California Department of Health Services did undertake to survey all the counties in California to assay the average radon concentration in homes. They have a special office in Sacramento which has statistics professionals and we were told the survey would provide accurate results. This was released after the discovery of the high-uranium Rincon Formation in Santa Barbara County. We were expecting big numbers. The Santa Barbara results were average: geometric mean 0.63 pCi/l, arithmetic mean 1.25 pCi/l, and GSD = 3.35, using n=120 measurements.

They did publish all of their results and identified each of the measurements by zip code (Liu, 1990). We have a nearby zip code that is exclusively built on Rincon formation soil, the tourist

town of Summerland, zip 93067. It was where Carlisle (Carlisle, 1993) had found the first high levels of radon in California and the levels there are consistently high. So, I studied the output and realized, "There are no data from Summerland." I called Steve Hayward of the State Health Service and he investigated the situation. He said that the California data base for homes is from the property tax rolls (Hayward, 1995). They had reduced the base by only considering homes that are occupied by their owners by checking to see if the tax address agreed with the property address. There are none in Summerland because they do not have home delivery of US mail; they all have free boxes at the post office. It turns out that this is not the end of the story. Some years later we found another community built exclusively on exposed Rincon, Los Alamos 93440. This place also has consistently high indoor radon concentrations. You guessed it: they also do not have home delivery so none of their homes were considered in the State survey. So the State had innocently systematically excluded two high-radon communities in Santa Barbara County from their analysis, a definite cause of bias.

Defining your survey population is important in making a proper survey. I believe it is important to have geologists identify different areas of potential concern before making a survey. This was partially done in the State survey by having more measurements done in counties where high levels were anticipated. For example, Ventura County (once part of Santa Barbara County) was sampled n=159 times, the largest in the state. It had been a region of intense uranium prospecting in the 1950's and many mining claims were filed. This State survey (Liu, 1990) was used in making the EPA radon map which is ubiquitous.

A relatively simple proposal for a random sampling would be to chose geographic coordinates randomly and find the nearest houses. We mentioned drawing names from a hat earlier; this process would be analogous to throwing darts at a map on the wall (blind-folded, of course). As shown in the Appendix, a scientific sample of 40 to 50 homes would give a good estimate of the median radon level.

Variation in the Distribution

As Nero, *et al.*, (Nero, 1986) emphasize in their seminal paper on radon distributions, radon data in various categories are very accurately represented by lognormal distributions. They perform various statistical tests to verify this. They also provide discussion as to why this is the case.

A lognormal distribution results when the magnitude of the data results from a product of factors. If there are enough, it does not matter what the shape of the distribution of the factors themselves is. For lognormal factors, there doesn't have to be that many factors. The process is similar to the generation of the more familiar Gaussian distribution by additive factors. Nero, *et al.*, (Nero, 1986) note that the analysis of indoor radon shows that it results from numerous factors, many are multiplicative but some additive. Because of the apparent closeness of the data to a lognormal curve, the multiplicative factors dominate the process.

Suppose we have the magnitude of the indoor radon in some randomly chosen house is r. Conceptually, we can collect the factors for this level into three groups and combine them

$$r = f_s f_w f_b \qquad (4)$$

where f_s is the combination of the factors related to the soil (radium concentration, permeability, *etc.*), f_w is the combination of the factors related to the weather (wind speed, rain amount, *etc.*), and f_b is the combination of the factors related to the building (leakage area, footprint area, *etc.*) We know that radium concentration in soil samples is itself distributed according to a lognormal distribution (Wedpohl, 1969). It is convenient to think of the magnitude of the radon as coming from the soil and the weather and building factors spreading the indoor concentrations (like filters). In fact, it is reasonable to assume that each of these factors itself is a lognormal distribution over the homes in an arbitrary area (containing at least a thousand homes). The product radon level r will most likely have a lognormal distribution.

Remember the log of the radon concentration has a normal distribution, and mathematically

$$\ln r = \ln f_s + \ln f_w + \ln f_b \qquad (5)$$

All these logarithms have simple Gaussian normal distributions and we know how to work with these. In particular, the variance of the logradon values can be written as a quadrature sum

$$\sigma_g^2 = \sigma_s^2 + \sigma_w^2 + \sigma_b^2 \qquad (6)$$

The variation in the product is the sum of the variation in the logarithms of the various factors. The geometric standard deviation of the radon concentrations is written

$$GSD = \exp\sqrt{\sigma_s^2 + \sigma_w^2 + \sigma_b^2} \qquad (7)$$

Recall from Eq. (2) that this is the factor of the distribution median (×÷) that will make an interval containing the 68% of the distribution. A rough estimate of the magnitude of the GSD may be obtained by averaging the ratio of a few pairs of randomly chosen data points. This is because the GSD is related to the average ratio of the data points, similar to the normal standard deviation being related to the arithmetic difference of data points. This is quick, but very rough, method to estimate the GSD. Analytic procedures for estimating distribution parameters are described in the Appendix. Next, I show that values for GSD are often in the range of 2 to 3.

Observations of Radon Variation in California

It is provocative to ponder the life of a radon-222 atom. As it progresses from birth in the soil as radium daughter to its final death in decay as it transmutes into polonium-218, there are numerous physical factors which influence its path. While it may be possible to estimate the branching ratio for the many possibilities, I chose instead to simply look at the final indoor concentration as measured by a radon detection device. In the fall of 1990, The California

Department of Health Services performed numerous (>2000) short-term measurements of radon concentrations in homes in California (Liu, 1990).

County	number	Median	GSD		County	number	Median	GSD
Orange	41	0.8	1.52		Santa Clara	84	0.8	2.87
Tuolumne	30	1.1	1.86		San Joaquin	30	1.1	2.94
Tulare	73	1.5	1.89		Placer	102	0.5	3.04
Kern	122	1.1	1.92		Marin	68	0.5	3.13
San Diego	73	0.6	2.07		Sonoma	101	0.3	3.15
Contra Costa	73	0.7	2.07		Santa Barbara	120	0.6	3.35
Los Angeles	89	0.7	2.11		Napa	33	0.6	3.49
Fresno	123	0.9	2.28		Solano	59	0.5	3.50
Alameda	79	0.7	2.29		San Mateo	50	0.4	3.57
Siskiyou	37	0.7	2.32		Shasta	96	0.4	3.73
Butte	51	0.5	2.39		Nevada	32	0.6	4.07
Ventura	159	0.7	2.48		Sacramento	68	0.3	4.49
El Dorado	45	0.8	2.83		Humboldt	50	0.1	5.42

TABLE I. Summary values for radon concentrations in selected California counties

These data have been published and the central portions of the distributions appear to be well approximated by lognormal distributions. The logarithms of all the data points have been used to calculate their mean and standard deviations in each of the counties. The exponentials of these values have been evaluated and these are listed as the medians (geometric means) and GSD's (geometric standard deviations) in Table I. These are statistics and I treat them as unbiased estimates best estimates of the corresponding population parameters. A more complete discussion would involve finding a confidence intervals. The counties have been ordered in ascending GSD.

An attractive feature of these data was that they were all taken at approximately the same time of year (the fall). According to Nero (Nero, 1986) an important source of radon variation is the outdoor temperature. Thus, the radon measurements in the winter are about a factor of 2 larger than the measurements in the summer. The ability to average over variations due to weather (temperature, wind and rain) is part of the rationale in making a long-term (1-year) measurement. On the scale of a county, taking all the measurements at the same time would also reduce the variation due to weather; all the homes would have approximately the same ambient conditions. So the GSD's in Table I are primarily due to variations in the buildings and the soils. Most of the GSD factors have values between 2.0 and 3.5.

These statistics are calculated by simple direct manipulation of the county data. In a few cases it is clear that there is more than one lognormal population in a county. If the data are ordered and plotted on lognormal paper, most of the points will be in a straight line until the last few points.

The points diverge in a systematic fashion. This indicates the occurrence of a high-radon subpopulation, a geologic radon hot spot. This was discussed in some detail in a previous paper by Meada and myself (Hobbs, 1996). The logradon data clearly follows a dominant Gaussian distribution until well out in the tail, then the last few data transition to another Gaussian: a bump on the tail. The GSD can always be calculated, but in a multimodal case, its magnitude is abnormally increased by the bump. Several California counties show this evidence of a subpopulation hot spot.

This is the case in Santa Barbara County. Here we have 95% of the homes on nominal California radium soil (about 1-2 ppm eU) and the remaining 5% on Rincon soil (about 25 ppm eU) (Rosen, 2002). The official California survey completely omitted Summerland which is built exclusively on this soil. I have been taking measurements in Summerland for the last 20 years. Here are my 36 measurements in Summerland.

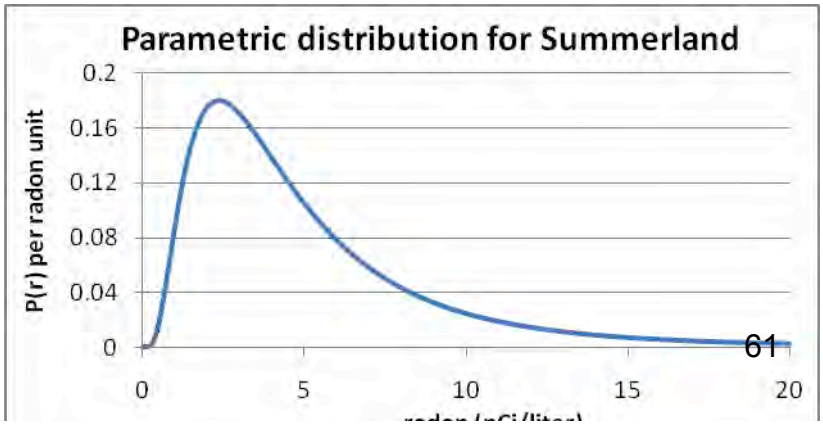
9.5	1.9	3.8	4.7	2.7	3.5
2.0	6.1	1.1	5.9	12.3	2.0
30.1	11.1	1.9	6.0	5.5	4.7
21.1	3.5	5.1	6.8	5.4	3.4
4.3	3.5	2.4	1.0	1.3	10.6
2.9	2.1	3.6	4.5	1.4	1.8

TABLE II. Hobbs radon measurements (pCi/l) in Summerland, CA 93067; 1990-2010.

These measurements were often made pursuant to real-estate transactions and were short-term (2-4 days). Sometimes the houses were empty, but sometimes there were occupants. They were made throughout the years, but I have more business in the springtime. Summerland is right on the coast and doesn't have much weather (marine temperate); the temperature is almost always 45-85ºF. What <u>all</u> the houses have in common is the black Rincon soil.

The logradon data is reasonably represented by a Gaussian distribution or equivalently the radon data is reasonably represented by a lognormal distribution. The data and curve are compared.

	data	curve
geometric mean	4.0	4.0
geometric SD	2.18	2.18
median	3.7	4.0
arithmetic mean	5.2	5.5



min	1.0	0
first quartile (Q_1)	2.05	2.3
median (Q_2)	3.7	4.0
third quartile (Q_3)	5.95	6.5
max	30.1	no bound
(3 data w/ 3.5 pCi/l)		
mode	3.5	2.4

Figure 2. Summerland radon distribution

The geometric mean and standard deviation are calculated from the radon measurements. These statistics become the parameters to generate the continuous probability distribution. In the table the characteristics of the parametric curve are compared with the basic statistics. There are small discrepancies between the data and the theoretical distribution which are expected.

The soil of Summerland is quite homogeneous. In addition to its high uranium content, another important property is that it is very expansive when wet. It is usually dry and forms a spider-web of cracks. This is actually its most identifying characteristic for me. The numerous cracks would lead to enhanced gas-emanation. This, in turn, enhances the probability that homes on the Rincon will contain unacceptable levels of radon. Even so, all the homes in Summerland have essentially identical soils. I estimate that the soils lead to a small variation in the indoor radon levels; less than 10%, so $\sigma_s^2 < [\ln(1.1)]^2 = 0.009$.

The temperatures of Summerland are very mild. Because of the fog and chill at night and mornings, it is common for heating year round in the homes. (We have just completed a record summer in Southern California. There were numerous records set for low temperatures. The high temperature in July was a cool 64^{o}F.) There are occasional breezes, inversions, and showers but I estimate that the weather may contribute less than a 30% variation in radon level; this would mean $\sigma_w^2 < [\ln(1.3)]^2 = 0.069$.

The remaining source of uncertainty in the indoor radon comes from the characteristics of the homes and their operation. This is related to the magnitude of soil gas infiltration rate compared to the total ventilation rate of the house. If this were expressed as a percentage, we would expect a distribution similar to the Summerland radon distribution shown in Fig. 2 above. At this point we can use the data and our assumptions to make an estimate of the building-factor uncertainty, σ_b^2. Rearranging Eq. 6,

$$\sigma_b^2 = \sigma_g^2 - \sigma_s^2 - \sigma_w^2 \qquad (8)$$

From the measurements $\sigma_g = \ln(\text{GSD}) = \ln(2.18) = 0.779$, substituting

$$\sigma_b = \sqrt{0.607 - 0.069 - 0.009} = 0.727 \qquad (9)$$

The exponential of this is a geometric factor representing the building effects of radon, a geometric standard deviation for the building

$$\text{GSD}_b = \exp(0.727) = 2.07 \qquad (10)$$

Building design features and flaws along with the variation in the operation of the building contribute, on the order of, a factor of 2.0 to the indoor radon level. In addition, this value is consistent with various data from the State of California (Table I).

Conclusion and Summary

The distribution of indoor radon is important to understanding its nature and to developing a strategy to protect citizens from radon exposure.

A scientific survey is critical in generating radon data. Many data sets are self-selected (voluntary) and are essentially useless in making inferences. Self-selected data can be used to indicate that radon measurements have been made, but little else.

The specific radon concentration in a home is the result of **numerous factors**. As such the central part of the distribution is well approximated by a lognormal distribution. The logarithms of the data form a Gaussian "bell-shaped" profile. The fundamental parameters of the Gaussian curve are the mean (log **geometric mean**) and standard deviation (log **geometric standard deviation**). When plotted against radon (no log), the distribution is strongly skewed to the right, containing points with relatively high magnitude, several times the mean.

The broad radon distribution emphasizes some platitudes about testing. Since the radon concentration results from many factors, **the best way to determine a radon level is to make a measurement**. Also, **radon levels may vary from house to house in seemingly unpredictable ways**. And most important, **almost any house could possibly have a radon level greater than the EPA action level of 4.0 pCi/liter.**

Much of the tedious work with radon data can be obviated by simply taking the **logarithm of the radon** level first and using the logradon values as the basic data. The logradon values have a simple normal distribution. If you have a random sample you can find confidence intervals for the parameters, compare samples from different populations, etc. A reasonable value for the logradon standard deviation is simply 1.0 (or a geometric standard deviation of about 2.7). These are reasonable values for back-of-the-envelope calculations. It is easy to make a histogram of the logradon values and check if it appears consistent with a normal distribution.

The many factors which contribution to the specific radon concentration in a home can be collected into three groups: **the soil factors, the weather factors, and the building factors.** Each of these is itself a complicated collection of factors and parameters. Thus, it is reasonable to deduce that each of these factors is well represented by a lognormal distribution. This leads to the quadrature formula (Eq. 6) for the log of the geometric standard deviation

$$\sigma_g^2 = \sigma_s^2 + \sigma_w^2 + \sigma_b^2$$

where the terms on the right are the squares of the logs of the geometric standard deviations (GSD's) of the factor groups.

By looking at sets of radon data collected under different conditions, it appears that each of the geometric standard deviations for the groups can be as large as 2 or even larger. This means that each group of phenomena could cause **variation in radon concentration on the order of a factor of 2**. For a lognormal distribution a factor one GSD contains 68% of the data. A simple example is the comparison of the 1-year radon measurements compared with the 4-day measurements. Long-term measurements average over the ensuing radon variation due to meteorological phenomena and consequently have a smaller GSD. Nero, *et al.*, (Nero, 1986) discuss this further.

Both the mean and the standard deviation contribute the indoor radon level. Carlisle & Azzouz. (Carlisle, 1993) note that as the housing stock ages, there is more soil-gas infiltration resulting in a rise of the mean value. The houses do not all age the same way and the distribution will be broadened. I have personal experience with this phenomenon. About 15 years ago, I was asked to make measurements of a cross section of homes in the Winchester Canyon area of Goleta, California. This is near an area of Rincon soil. All of the measurements had magnitudes less than 1.1 pCi/l. Recently, I have had occasion to test some of these homes for real-estate transactions. The recent measurements have increased and fall in the range 1.2-4.5 pCi/l. Both of these sets had a few data points (about 5), but they seem to validate the earlier results (Carlisle, 1993).

The building set of radon factors is the one over which we have most control. We severely limit the amount soil gas infiltrating into a building by changing the building characteristic with a mitigation system. An understanding of the distribution of indoor radon concentrations, as outlined in this paper certainly validates the requirement of building with radon-resistant construction techniques in high potential regions. These are regions where the other sets of factors give rise to a high probability that the home would have a high average radon

concentration, particularly areas of low ambient temperatures (weather) and high soil radium content (soil).

These techniques include such things as a soil-gas ventilation system built into the foundation. There is a tremendous need to develop requirements for energy efficient homes. This should be done taking into consideration the total indoor environment including all the various indoor air pollutants including radon. With this in mind, it is hoped that the various radon-resistant features of home construction will become common place in the future. It is also important that they be built as robust as possible to continue preventing radon entry for the life of the house.

References

1. W. E. Hobbs and L. Y. Maeda, "Identification and Assessment of a Small, Geologically Localized Radon Hot Spot," Environment International Vol. 22, pp. S809-S817, Elsevier Science Ltd, 1996.

2. R. Blood, Radon Database for California, October 2002. http://www.consrv.ca.gov/cgs/minerals/hazardous_minerals/radon/DHS_Radon_Database.

3. D. Carlisle and H. Azzouz, "Discovery of radon potential in the Rincon Shale, California – A case history of deliberate exploration," Indoor Air 3, pp 131-142. Munksguard, 1993.

4. K. S. Liu, S. B. Hayward, J.R. Girman, B. A. Moed, and F. Y. Huang, "Survey of residential indoor radon concentrations in California," Final Report CA/DOH/AIHL/SP-53, Berkeley, CA, 1990.

5. S. B. Hayward, private communication.

6. A. V. Nero, M. B. Schwehr, W. W. Nazaroff and K. L. Revzan, "Distribution of Airborne Radon-222 Concentrations in U.S. Homes," Science, Vol 234, pp. 992-997, 21 Nov 1986.

7. K. H. Wedepohl (ed.), *Handbook of Geochemistry,* Berlin, Springer-Verlag, 1969.

8. Art Rosen, Physics Professor, Cal Poly San Luis Obispo. Several canisters of Rincon Shale soil were assayed using the gamma spectrometer at the radiological lab of Cal Poly SLO in 2004. There was also a canister from San Luis Obispo County which showed the same as Santa Barbara County; ppm eU mean parts per million equivalent uranium (by mass).

Problem: What is the mean of the population of homes with a nominal lognormal radon distribution of GSD = 2.7 (σ_g=1) that has 20% of the homes with radon levels greater than 4.0 pCi/l. The *z*-value for this is *inverseNorm*(1-0.2) =0.84. Using Eq. A5, from $\ln r = \mu_g + \sigma_g z$ we have $\ln 4 = \mu_g + (1)(0.84)$, so $\mu_g = 0.545$; the median is $e^{0.545} = 1.7$ pCi/l and the arithmetic mean is calculated

$$\overline{r} = \exp\left(\mu_g + \sigma_g^2 / 2\right) = 2.8\ pCi / l$$

Appendix

Evaluating parameters for a radon lognormal distribution

The notation and methods of working with the lognormal distribution are reviewed. The lognormal distribution is a probability distribution $P(r)$ of the form

$$P(r)dr = \frac{\exp\left[\dfrac{-(\ln r - \mu_g)^2}{2\sigma_g^2}\right]}{r\sqrt{2\pi\sigma_g^2}}dr \qquad \text{or} \qquad P(\ln r)d(\ln r) = \frac{\exp\left[\dfrac{-(\ln r - \mu_g)^2}{2\sigma_g^2}\right]}{\sqrt{2\pi\sigma_g^2}}d(\ln r) \quad \text{(A1)}$$

The expression on the left is in normal units (e.g. pCi/l) and has a distribution skewed to the right. The expression on the right is in logradon units and is perfectly symmetrical. They are equivalent. In these μ_g is the natural log of the geometric mean and σ_g is the natural log of the geometric standard deviation; both are in logradon units. The geometric mean is also the median (also called the 2nd quartile or 50th percentile). The arithmetic mean is

$$\ln \overline{r} = \int_0^\infty r\, P(r)\, dr = \mu_g + \sigma_g^2 / 2 \qquad \text{(A2)}$$

In the linear theory of cancer, the risk is proportional to $\overline{r} = \exp\left(\mu_g + \sigma_g^2 / 2\right)$. There values in the tail are accentuated by the skewness through the σ_g term. The mode r_m is found at $dP/dr = 0$

$$r_m = \exp\left(\mu_g - \sigma_g^2\right) \qquad \text{(A3)}$$

This means the densest part of the distribution will be less than the median. It will be common to make a measurement that results in a value at this low level concentration.

The specific probability of a given interval of radon concentrations for population (μ_g, σ_g) can be found using standard normal curve values, the normal cumulative distribution function. In particular, it is interesting to estimate the proportion of houses in a population that would have indoor radon levels greater than 4.0 pCi/liter.

$$P(r > 4) = normalcdf(\ln 4, \infty, \mu_g, \sigma_g) = 1 - normalcdf(-\infty, \ln 4, \mu_g, \sigma_g) \qquad \text{(A4)}$$

Again, we point out that the magnitude of the term σ_g is important in this evaluation.

Evaluation of the geometric parameters—is straightforward working from the premise that the data is from a lognormal population. The first step is to take the logarithms of all the radon measurements. As discussed in the body of the paper, the factors contributing to the magnitude of the indoor radon level generally combine through multiplication and this results in a lognormal distribution. We assume that the radon data is from a lognormal population.

If all the data are known (like in Summerland), then the mean and standard deviation of the radon logarithms are estimates of the parameters (μ_g, σ_g). Usually, however, the low-magnitude measurements are only roughly known. This is sometimes indicated by being below a limit. For

example, it is common to see a radon measurement simply stated as less than 0.5 pCi/l. This makes the direct calculation impractical. Other methods for estimating the mean and standard deviation of the distribution make use of the standard normal curve. The dimensionless *z*-value for a radon datum r is defined so that

$$\ln r = \mu_g + \sigma_g z \tag{A5}$$

The *z*-value itself is associated with a probability or position in the distribution. Suppose, for example, there are 55% of the data with values less than r_1 = 1.0 pCi/l and 20% of the data between 1.0 and r_2 = 2.0. The *z*-values are inputs to the cumulative standard normal distribution, so they can be found with the inverse-cdf. The values are z_1 = inverse-normal 55% = 0.1257 and z_2 = inverse-normal 75% = 0.6745. Combining with the logs: ln 1 = 0 and ln 2 = 0.6931, the equations are

$$\begin{aligned} 0 &= \mu_g + (0.1257)\sigma_g \\ 0.6931 &= \mu_g + (0.6745)\sigma_g \end{aligned} \tag{A6}$$

The solution is $\mu_g = -0.1587$ and $\sigma_g = 1.2630$. These are logarithms and their exponentials are the geometric mean M = 0.9 pCi/l and GSD = 3.54.

For the California county data, there were scientific samples of reasonable size taken for most of the counties. The magnitudes of these data less than one were simply given as "< 1.0". These were counted and the *z*-value associated with their fraction were the first point as above. Then the remaining data were ordered and the *z*-values associated with their cumulative fraction calculated. When the log of the data is plotted as a function of the *z*-value, invariably the main body of the distribution would be characterized by a surprisingly straight line. The log of the geometric mean μ_g is its *y*-intercept and the log of the geometric standard deviation σ_g is its slope. These can be evaluated by the least-squares method. This is a good approach since it gives equal weight to each data point and the errors in the position of individual points would be averaged out when many are used (an application of the law of large numbers) .

The population of interest is assumed to have a lognormal distribution. This allows confidence intervals to be calculated for the various parameters. Consider Ventura County in California as an example. There were 159 randomly chosen homes with radon data; 90 had radon measurements less than 1.0 pCi/l. When the other 69 data points are plotted on lognormal probability graph paper they clearly form a straight line. The *y*-intercept and slope are found to be $\mu_g = -0.21$ and $\sigma_g = 0.91$ which are the unbiased estimates. The 95% confidence intervals are found to be

$$\begin{aligned} -0.4325 < \mu_g < -0.1475 \\ 0.819 < \sigma_g < 1.023 \end{aligned} \tag{A7}$$

The confidence intervals go inversely with $\sqrt{n}$, and 40-50 data points results in a reasonable bound. The best estimates for the arithmetic mean of Ventura County homes is $\overline{r}$ =1.1 pCi/and the percentage of homes above 4 pCi/l is p_4 = 3.2%. The 95% confidence intervals for these statistics are

$$\begin{aligned} 0.9 < \overline{r} < 1.46 \ pCi/liter \\ 2.2\% < p_4 < 4.7\% \end{aligned} \tag{A8}$$