Appendix 6J

Comparison of Leak Frequencies in Rural and Urban Segments of the Pipeline

Comparison of the Leak Rates in Urban and Rural Segments of the Pipeline

The purpose of this section is to examine the relationship between the leak rates in the urban and rural areas and to compare the two rates. For the purposes of the analysis here, Harris and Travis counties are treated as urban, and the other counties are treated as rural. This represents only one of the possible ways of comparing urban versus rural conditions. Another approach is recommended at the end of this section.

Table 1 presents the basic data by county on which the analysis was based.

Table 1. Pipeline Miles and Leaks by County with Sums for Urban and Rural Counties

County	Urban or Rural	Pipeline Miles	Leaks in Most Recent 10 Years	Leaks in 29 Years
Harris	Urban	41.10	1	3
Travis	Urban	27.84	1	3
Sum	**Urban**	**68.94**	**2**	**6**
Waller	Rural	13.72	0	1
Austin	Rural	28.80	0	1
Fayette	Rural	27.00	0	0
Bastrop	Rural	33.74	1	7
Hays	Rural	10.09	0	0
Blanco	Rural	26.11	0	1
Gillespie	Rural	23.56	1	1
Mason	Rural	33.25	0	0
Kimble	Rural	34.75	1	3
Menard	Rural	3.74	0	0
Schleicher	Rural	53.53	1	1
Crockett	Rural	25.92	1	1
Reagan	Rural	28.02	0	1
Upton	Rural	33.50	1	2
Crane	Rural	5.00	0	1
Sum	**Rural**	**380.73**	**6**	**20**

The leak rate in leaks per mile per year is computed as follows:

$$LeakRate(leaks/mile/year) = \frac{Leaks}{PipelineLength(miles) * Duration(years)}$$

The leak rates were computed separately for predominantly rural and urban counties. Also, calculations were performed separately for the entire 29-year period and for the most recent ten years. The 29-year period provides a much larger statistical sample size on which to base the analysis. The ten-year period is of interest because it may be more representative of current conditions.

The difficulty, from a statistical point of view, in basing the analysis on the most recent ten-year period pertains to the small number of leaks that occurred during that period. There were only two leaks in the urban counties and six leaks in the rural counties. It is difficult to make precise estimates of leak rates on such small numbers of events. During the 29-year period, six leaks occurred in the urban counties and 20 leaks occurred in the rural counties, which provides a better opportunity to estimate the leak rates.

Table 2 presents the leak rates for the urban and rural counties for the two analysis periods mentioned. For the ten-year period, the leak rates are 0.00290 and 0.00158 leaks/mile/year for the urban and rural areas, respectively. For the 29-year period, the leak rates are 0.00300 and 0.00181 leaks/mile/year for the urban and rural areas, respectively. Thus, the leak rates are higher in the urban areas for both analysis periods.

Table 2. Leak Rates and 95 Percent Confidence Intervals

	29 Years		**Most Recent 10 Years**	
	Urban	**Rural**	**Urban**	**Rural**
Lower Confidence Limit	0.00110	0.00111	0.00035	0.00058
Leak Rate	0.00300	0.00181	0.00290	0.00158
Upper Confidence Limit	0.00653	0.00280	0.01048	0.00343

However, one must take into account random variability when comparing the leak rates estimated from data. Using a methodology described by Hahn and Meeker (1991) a confidence interval can be calculated for an estimated leak rate. The methodology is based on the Poisson distribution. We will first give a general description of a Poisson process, and then we will discuss how this description applies in our context.

If the following assumptions are satisfied, then the number of events that occurs in a time interval with a specified length has a Poisson distribution (Mood, Graybill, and Boes, 1974):

The probability that exactly one event will occur in a short time interval of length h is approximately λh, where λ is the rate of occurrence of the event.

The probability that more than one event will occur in a short time interval of length h is negligible compared to the probability that one event will occur.

The numbers of events that occur in non-overlapping time intervals are independent.

Our situation is more complicated in that we have events that occur at different times and at different locations. The assumptions can be restated as follows to account for the two dimensions, space and time. Our rate λ, as stated earlier, has units of leaks/mile/year. The probability that exactly one leak occurs within a short time interval of length h and within a spatial interval of length d is approximately λhd. The numbers of events that occur in non-overlapping cells in space and time (rather than in time alone) are independent. Thus, the fact that we are concerned with both space and time does not change the basic concepts.

Hahn and Meeker present the following methodology.

The estimate λ of the rate of occurrences (leaks in our case) is as follows:

$$l = \frac{x}{n}$$

The quantity x is the number of leaks. The quantity n in our case is the number of years times the number of pipeline miles. This formulation gives us the leak rate in leaks/mile/year.

The following, then, is the expression for a confidence interval for the leak rate:

$$\left(\frac{G_L}{n}, \frac{G_U}{n} \right)$$

where G_L and G_U are parameters that depend on both the number (x, above) of occurrences of the event and the confidence level. Hahn and Meeker present a table in which G_L and G_U are tabulated as a function of these two parameters.

Table 2 presents the 95% confidence intervals for the four estimates of leak rate. For the ten-year period, the confidence intervals are extremely wide, especially for the urban case, for which there were only two leaks. The confidence interval, 0.00035 to 0.01048 leaks/mile/year, includes the leak rate, 0.00158 leaks/mile/year for the rural case. In fact, the entire confidence interval for the rural case falls within the confidence interval for the urban case. Similarly, the rural confidence interval, 0.00058 to 0.00343 leaks/mile/year, includes the urban estimate, 0.00290 leaks/mile/year. Whether confidence intervals overlap is not a formal criterion for a statistically significant difference. Nevertheless, in this case, it is very apparent that the uncertainties are large compared to the difference between the two leak rates.

For the 29-year period, the leak rates are 0.00181 and 0.00300 leaks/mile/year for the rural and urban cases respectively. Because of the longer period and the larger number of leaks observed, the confidence intervals here are much narrower. The confidence intervals overlap somewhat, which may seem to suggest that the urban and rural leak rates are not significantly different. However, as is discussed above, whether the separate confidence intervals overlap is not a formal criterion for a statistically significant difference.

Hald (1952) gives a rationale for a hypothesis test that can be adapted for this more complex situation, with two continua (pipeline length and years) and with different numbers of mile-year combinations for the urban and rural cases.

The approach can be explained intuitively as follows. We have estimated the leak rate as 0.00199 leaks/mile/year for the pipeline as a whole for the 29-year history. Thus, we want to test whether (1) the difference between the leak rates in the rural and urban areas can reasonably have occurred by chance or (2) the difference is too large to be explained in these terms. In the second case, we would conclude that the difference was statistically significant.

It was expected that, if there were a difference, the leak rate would be higher in the urban areas because of the greater level of activity and greater probability of third-party damage. Thus, we are really interested in testing whether the number of leaks in the urban areas is too large to have occurred by chance if the true leak rate is the same in the urban and rural areas. A test involving data for only leaks caused by third-party damage is not possible because the cause is unknown for about a third of the leaks.

The methodology can be explained in more detail as follows. Under the null hypothesis (to be tested) that the leak rate is constant, the following is the Poisson probability density for the number of leaks in the pipeline as a whole:

$$f_T(x) = \frac{e^{-l_T}\, l_T^x}{x!}$$

where

x = total number of leaks for the pipeline (the actual number is 26),

λ_T = the parameter of the Poisson distribution for the pipeline as a whole, and

$f_T(x)$ = the probability that exactly x leaks occur in the entire 29-year history in the pipeline as a whole.

The parameter λ_T is computed as follows:

λ_T = (leak rate)*(totalpipeline miles)*(years)

= (0.00199 leaks/mile/year)*(449.67 miles)*(29 years)

= 26.

We also need the joint probability density that n_U leaks occur in the urban areas, while n_R leaks occur in the rural areas:

$$f(n_U, n_R) = \frac{e^{-l_U} l_U^{n_U}}{n_U!} \frac{e^{-l_R} l_R^{n_R}}{n_R!}$$

The parameters here are defined similarly.

n_U = total number of leaks in the urban area,

λ_U = the parameter of the Poisson distribution for the urban area, and

n_R = total number of leaks in the rural area,

λ_R = the parameter of the Poisson distribution for the rural area, and

$f(n_U,n_R)$ = the probability that exactly n_U leaks occur in the urban area and n_R leaks occur in the rural area during the entire 29-year history. That is, $f(n_U,n_R)$ is the probability that these two counts of events in the two different areas occur simultaneously, without any restriction on the total number of leaks in the pipeline as a whole.

The parameter λ_U is computed as follows:

λ_U = (leak rate)*(urbanpipeline miles)*(years)

= (0.00199 leaks/mile/year)*(68.94 miles)*(29 years)

= 3.99

The parameter λ_R is computed as follows:

λ_R = (leak rate)*(ruralpipeline miles)*(years)

= (0.00199 leaks/mile/year)*(380.73 miles)*(29 years)

= 22.01

In each instance, λ is the mean of the distribution. For example, λ_T is the mean number of failures along the complete pipeline in a 29-year period. Not surprisingly, this mean value is estimated as the actual observed number of leaks, 26.

The average number of leaks in the urban area is 3.99, which is less than the actual number, 6. The average is based on the null hypothesis, that the true leak rate is the same everywhere in the pipeline, and the difference between the average (3.99) and the actual number (6) can be explained in terms of random variability.

As is indicated above, $f(n_U,n_R)$ is the probability that n_U leaks occur in the urban area and n_R leaks occur in the rural area, without restriction on the total number of leaks. Ultimately, we want to test whether a count as large as 6 could reasonably have occurred in the urban area, given that the total number of leaks was 26. For this purpose, we need the following function, which is called the "conditional probability density function":

$$f(n_U, n_R \mid x) = \frac{f(n_U, n_R)}{f_T(x)}$$

The quantity $f(n_U, n_R \mid x)$ is the probability that n_U leaks occur in the urban area and n_R leaks occur in the rural area, given that the total number of leaks is x. Clearly, the conditional probability has meaning only when $n_U + n_R = x$.

We are interested in knowing whether the number of leaks actually observed in the urban area is too large reasonably to have occurred given the assumption that the true leak rate is the same in the rural and urban areas. We can obtain the probability of occurrence of 6 leaks or more in the urban area by summing the conditional probability density, $f(n_U, n_R \mid x)$, for values of n_U from 6 to 26. If we specify n_U for a given term in the sum, n_R is automatically determined to be 26 - n_U. Then, the probability of 6 or more leaks in the urban area is as follows:

$$\sum_{n_U=6}^{26} f(n_U, 26 - n_U \mid 26)$$

or, equivalently,

$$1 - \sum_{n_U=0}^{5} f(n_U, 26 - n_U \mid 26)$$

The value obtained is 0.20. That is, if the true leak rate were the same in the urban and rural parts of the pipeline and a total of 26 leaks occurred in the 29-year history, then the probability that at least six of the 26 leaks would occur in the urban area would be 0.20, or 20%. Since an event that has a 20% probability is not especially unusual, we conclude that the difference between the observed number of leaks, 6, and the expected number, 3.99, could have occurred by chance. At the 90% confidence level, we do not reject the null hypothesis that the leak rates in the rural and urban counties are the same.

As a check, the sum of the conditional probabilities was computed for values of n_U from 0 to 26. Since n_U must equal one of these values, the sum must be one, and this result was obtained.

It is stated earlier that it is very apparent for the 10-year case that the difference between the urban and rural leak rates is small compared to the large uncertainties. Nevertheless, the methodology described above was applied to the 10-year case, and the difference between the urban and rural leak rates was found to be statistically insignificant.

For completeness, we give the form of the conditional probability density used in the calculations. This function can be reproduced from the equations given above:

$$f\,(n_U\,,n_R\mid x) = c\,l_U^{n_U}\;l_R^{n_R}\left(\frac{x}{n_U}\right)$$

where the quantity c, given below, is constant in the summations described earlier:

$$c = \frac{e^{\,l_T - l_U - l_R}}{l_T^x} = \frac{1}{l_T^x}$$

The following quantity, which appears in the expression above for the conditional probability density, is the binomial coefficient, x over n_U:

$$\left(\frac{x}{n_U}\right) = \frac{x!}{n_U\,!(x - n_U)!}$$

In the analysis above, urban versus rural conditions were compared by treating pipeline segments in Harris and Travis counties as urban and segments in all other counties as rural. It would be of interest to refine this analysis by using the population density in the vicinity of the pipeline as an independent variable. Then it would be possible to assess the relationship between leak occurrence and population density. Notice that it would not be sufficient to have the population density in the vicinity of each leak at the time of the leak. Comparable information would be required for conditions (time-space combinations) with and without leaks, so that the rates of leak occurrences could be assessed as a function of population density. The fact that population densities changed over the 29-year history is a complicating factor that would have to be taken into account.

References

Hahn, Gerald J. and William O. Meeker, *Statistical Intervals*, John Wiley & Sons, New York, 1991.

Hald, A., *Statistical Theory with Engineering Applications*, John Wiley & Sons, Inc., New York, 1952.

Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes, *Introduction to the Theory of Statistics*, Third Edition, McGraw-Hill Book Company, New York, 1974.