U.S. DEPARTMENT OF THE INTERIOR
U.S. GEOLOGICAL SURVEY

Probabilistic and Statistical Relationships Between Number of Vehicles and Number of Visitors at a Geologic Site in a National Park

by

Robert A. Crovelli[1]

Open-File Report 94-560

This report is preliminary and has not been reviewed for conformity with U.S. Geological Survey editorial standards and stratigraphic nomenclature.

[1]U.S. Geological Survey, Box 25046, MS 971
Denver Federal Center, Denver, Colorado 80225 U.S.A.

INTRODUCTION

We are interested in probabilistic and statistical relationships between the number of vehicles and the number of visitors at a site in a national park. For example, let us consider the Arches National Park within which is the developed area of the Wolfe Ranch parking lot approximately 1.5 miles from the geologic site of Delicate Arch (see Figure 1). In this study the developed area will refer to the immediate area encompassing the Wolfe Ranch parking lot and the Delicate Arch. A visitor of the developed area arrives in a vehicle parked in the Wolfe Ranch parking lot, walks around in the developed area, and may or may not be at the Delicate Arch site at any one point in time.

Let N : Number of vehicles at one time in the parking lot.

N is a discrete random variable that can take on the values $n = 0, 1, 2, \ldots$

Let X : Number of visitors at one time at the site.

X is a discrete random variable that can take on the values $x = 0, 1, 2, \ldots$

Consider the conditional probability

$$P(X \geq x \mid N = n) = \alpha$$

Two problems of interest are

1. Given n and x, find α.

   Find the probability α that there are at least x visitors at a time at the site when n vehicles are in the parking lot.

2. Given x and α, find n.

   Find the number of vehicles n in the parking lot such that the probability of at least x visitors at one time at the site is equal to α.

   Application of this methodology would be the following situation:

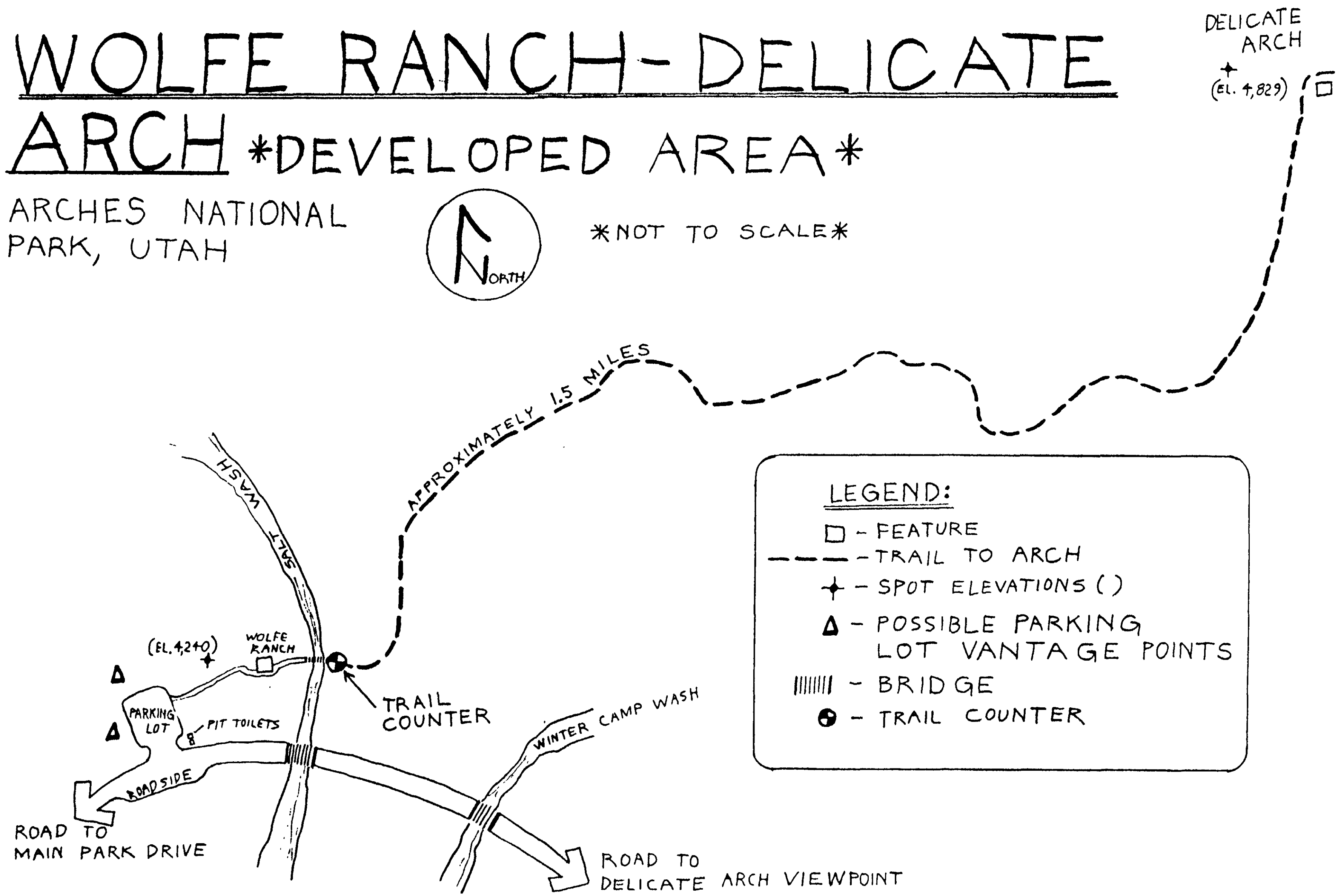

Figure 1. Developed area of Wolfe Ranch-Delicate Arch.

Based on survey data, the National Park Service has found that many visitors feel unacceptably crowded if there are 30 or more people at one time at the arch. Limit the number of vehicles *n* allowed in the parking lot such that 30 or more people are at the arch only 10% of the time.

MODELS AND METHODS

STATISTICAL APPROACHES

Statistical relationships between *N* and *X*.

Empirical Method

Define the visitor-vehicle ratio $R = \frac{X}{N}$.

The empirical complementary cumulative (at least) distribution

$$P(R \geq r)$$

is given in Table 1 for 84 observations (data from National Park Service).

Example: $P(R \geq 0.84) = 0.10$

Define the conditional visitor-vehicle ratio $R_n = \frac{X}{n}$.

Note that $P(X \geq x | N = n) = P\left(R_n \geq \frac{x}{n}\right)$.

Assume that *R* and *N* are independent, i.e., $P(R_n \geq r) = P(R \geq r)$.

Example: Given $n = 49$ and $x = 15$, then

$$P(X \geq 15 | N = 49) = P\left(R_n \geq \frac{15}{49}\right) = P(R \geq 0.31) = 0.73$$

Thus, 73% of the time there are at least 15 visitors at Delicate Arch when 49 vehicles are in the parking lot.

<u>Now consider the problem of given *x* and α, find *n*.</u>

Let r_α be the value of *R* such that $P(R \geq r_\alpha) = \alpha$.

Equating $r_\alpha = \frac{x}{n}$ and solving for *n*, we get $n = \frac{x}{r_\alpha}$.

Table 1. The cumulative (at least) distribution of the visitor-vehicle ratio at Delicate Arch from the National Park Service

Ratio at Delicate Arch				
Visitor: Vehicle Ratio	Cumulative Percent		Visitor: Vehicle Ratio	Cumulative Percent
1.28	1%		0.47	51%
1.00	2%		0.45	52%
0.96	4%		0.45	54%
0.88	5%		0.44	55%
0.88	6%		0.43	56%
0.86	7%		0.42	57%
0.85	8%		0.42	58%
0.84	10%		0.39	60%
0.83	11%		0.39	61%
0.83	12%		0.39	62%
0.83	13%		0.38	63%
0.81	14%		0.36	64%
0.80	15%		0.36	65%
0.80	17%		0.36	67%
0.79	18%		0.34	68%
0.75	19%		0.33	69%
0.71	20%		0.33	70%
0.68	21%		0.32	71%
0.67	23%		0.31	73%
0.67	24%		0.30	74%
0.65	25%		0.29	75%
0.65	26%		0.29	76%
0.64	27%		0.28	77%
0.63	29%		0.27	79%
0.61	30%		0.22	80%
0.61	31%		0.21	81%
0.61	32%		0.20	82%
0.59	33%		0.20	83%
0.58	35%		0.18	85%
0.58	36%		0.17	86%
0.53	37%		0.16	87%
0.52	38%		0.14	88%
0.52	39%		0.13	89%
0.52	40%		0.12	90%
0.51	42%		0.11	92%
0.50	43%		0.11	93%
0.50	44%		0.10	94%
0.50	45%		0.10	95%
0.49	46%		0.08	96%
0.49	48%		0.08	98%
0.48	49%		0.03	99%
0.47	50%		0.03	100%

Example: Given $x = 30$ and $\alpha = 0.10$, $r_\alpha = 0.84$, then

$$n = \frac{30}{0.84} = 36.$$

Therefore, when 36 vehicles are in the parking lot, then 10% of the time there are at least 30 visitors at Delicate Arch.

Theoretical Method

The distribution of R is approximately a normal probability distribution in the range of interest as can be observed in the plot of the Delicate Arch data on normal probability paper in Figure 2. A linear pattern of plotted points from the empirical cumulative distribution suggests a normal distribution (Devore, 1991). The pattern of points is quite linear in the range from 2% to 95%. The National Park Service is interested in percentages well within this range. For comparative purposes, the data were also plotted on lognormal probability paper in Figure 3. The nonlinear pattern of points in this case suggests that the distribution is not lognormal.

Thus $R = \frac{X}{N}$ is approximately distributed as normal (μ_R, σ_R).

Assume $R_n = \frac{X}{n}$ is approximately distributed as normal (μ_R, σ_R).

Then $X = n R_n$ is approximately distributed as normal $(n\mu_R, n\sigma_R)$.

Note: A normal random variable multiplied by a constant is also a normal random variable.

Let $Z = \frac{X - n\mu_R}{n\sigma_R}$.

Hence Z is approximately distributed as normal (0, 1).

Therefore $P(X \geq x \mid N = n) \doteq P\left(Z \geq \frac{x - 0.5 - n\mu_R}{n\sigma_R} \right)$

where the integer x is decreased by 0.5 because the normal distribution is a continuous distribution that is being used as an approximate model for a discrete random variable X.

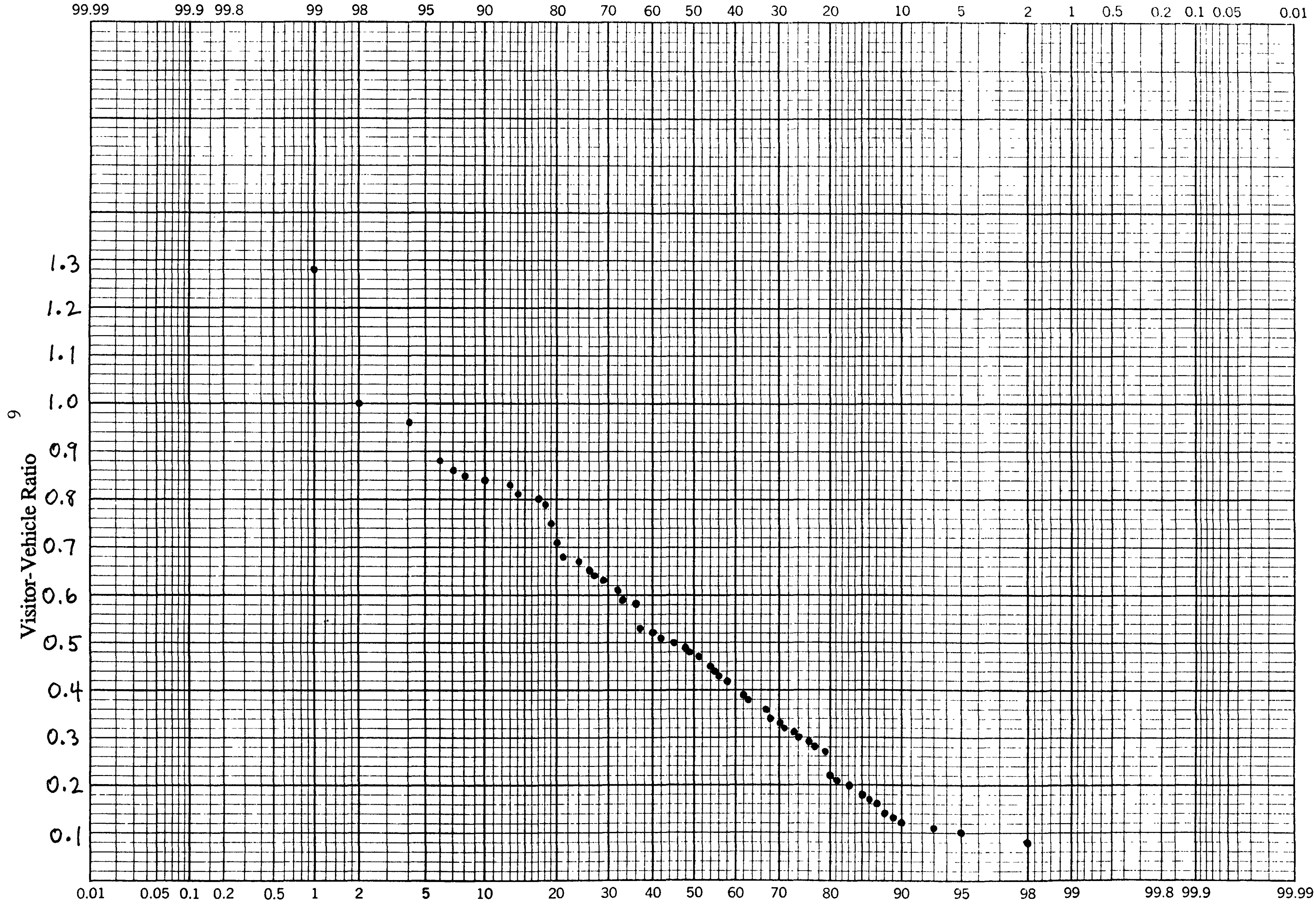

Figure 2. Visitor-vehicle ratio data for Delicate Arch plotted on normal probability paper.

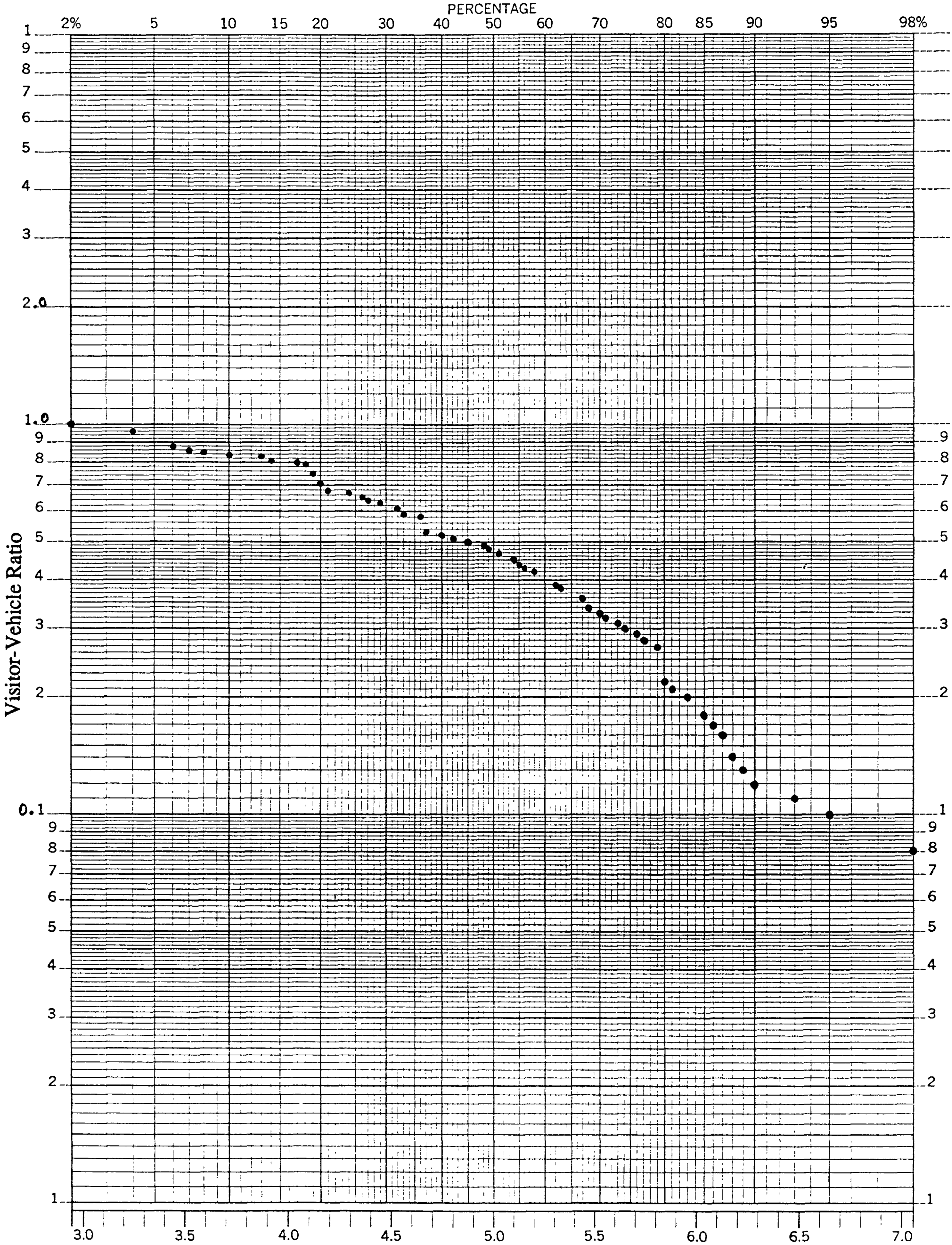

Figure 3. Visitor-vehicle ratio data for Delicate Arch plotted on lognormal probability paper.

The reason is the same as in the case of the normal approximation to the binomial distribution (Devore, 1991).

Example: Using the Delicate Arch data, we have

Sample mean is $\hat{\mu}_R = 0.4767$.

Sample standard deviation is $\hat{\sigma}_R = 0.2626$.

For $n = 49$, we get

Estimated mean of X is $\hat{\mu}_X = n\hat{\mu}_R = 49(0.4767) = 23.36$.

Estimated standard deviation of X is $\hat{\sigma}_X = n\hat{\sigma}_R = 49(0.2626) = 12.87$.

$$P(X \geq 15 | N = 49) \doteq P\left(Z \geq \frac{15 - 0.5 - 23.36}{12.87}\right) = P(Z \geq -0.69) = 1 - 0.2451 = 0.75,$$

compared to the empirical value of 0.73.

Note: The probability of 0.2451 is obtained from a standard normal table.

<u>Now consider the problem of given x and α, find n.</u>

Let z_α be the value of Z such that $P(Z \geq z_\alpha) = \alpha$.

Equating $z_\alpha = \dfrac{x - 0.5 - n\mu_R}{n\sigma_R}$

Solving for n we get $n = \dfrac{x - 0.5}{\mu_R + z_\alpha \sigma_R}$.

Example: Using the Delicate Arch data, we have for given $x = 30$ and $\alpha = 0.10$:

From a standard normal table $P(Z \geq 1.28) = 0.10$.

Hence $n = \dfrac{30 - 0.5}{0.4767 + 1.28(0.2626)} = 36$.

This value for n of 36 vehicles is the same as the empirical value, which results in excellent agreement.

PROBABILISTIC APPROACHES

Probabilistic relationships between N and X.

Analytic Method

Let Y: Number of visitors per vehicle.

Y is a discrete random variable that can take on the values $y = 1, 2, \ldots$

Mean of Y is denoted by μ_Y.

Standard deviation of Y is denoted by σ_Y.

Example: For the parking lot of the Delicate Arch, we have from the National Park Service:

Estimated mean of Y is $\hat{\mu}_Y = 2.7$

Estimated standard deviation of Y is $\hat{\sigma}_Y = 0.2$

Let M: Number of visitors at one time in the developed area.

M is a discrete random variable that can take on the values $m = 0, 1, 2, \ldots$

Mean of M is denoted by μ_M

Standard deviation of M is denoted by σ_M

For a given number of vehicles in the parking lot, $N = n$, the number of visitors in the developed area is the sum of the visitors per vehicle, i.e.,

$$M = \sum_{i=1}^{n} Y_i$$

We have the following parametric relationships for given n:

Mean of M is $\mu_M = n\mu_Y$

Variance of M is $\sigma_M^2 = n\sigma_Y^2$

Example: For $n = 49$ vehicles in the parking lot of the Delicate Arch, we get

Estimated mean of M is $\hat{\mu}_M = n\hat{\mu}_Y = 49(2.7) = 132.3$

Estimated standard deviation of M is $\hat{\sigma}_M = \sqrt{n}\hat{\sigma}_Y = 7(0.2) = 1.4$

From the Central Limit Theorem, we know that for large n,

M is approximately distributed as normal (μ_M, σ_M).

Thus $P(M \geq m | N = n) \doteq P\left(Z \geq \frac{m - 0.5 - n\mu_Y}{\sqrt{n}\sigma_Y}\right)$.

Example: For $n = 49$ vehicles in the parking lot of the Delicate Arch, we have

$$P(M \geq 135 | N = 49) \doteq P\left(Z \geq \frac{135 - 0.5 - 132.3}{1.4}\right) = P(Z \geq 1.57) = 1 - 0.9418 = 0.0582$$

Let P: Proportion of visitors in the developed area who are at the site.

P is a continuous random variable that can take on the values $0 \leq p \leq 1$.

Mean of P is denoted by μ_P

Standard deviation of P is denoted by σ_P

Define $P = \frac{X}{\mu_Y N} = \frac{1}{\mu_Y} R$

Mean of P is $\mu_P = \frac{\mu_R}{\mu_Y}$

Standard deviation of P is $\sigma_P = \frac{\sigma_R}{\mu_Y}$

Example: Using the Delicate Arch data, we have

Estimated mean of P is $\hat{\mu}_P = 0.4767/2.7 = 0.1766$

Estimated standard deviation of P is $\hat{\sigma}_P = 0.2626/2.7 = 0.09726$

Consider the probabilistic relationship for given n

$$X = P \cdot M$$

If we take the case of P and M are independent, then it can be shown that

Mean of X is $\mu_X = \mu_P \cdot \mu_M$

Variance of X is $\sigma_X^2 = \left(\mu_P^2 + \sigma_P^2\right)\left(\mu_M^2 + \sigma_M^2\right) - \left(\mu_P \mu_M\right)^2$

Example: For $n = 49$ vehicles in the parking lot of the Delicate Arch, we have

Estimated mean of X is $\hat{\mu}_X = (0.1766)(132.3) = 23.36$

Estimated variance of X is

$$\hat{\sigma}_X^2 = \left[(0.1766)^2 + (0.09726)^2\right]\left[(132.3)^2 + (1.4)^2\right] - [(0.1766)(132.3)]^2 = 165.878$$

Estimated standard deviation of X is $\hat{\sigma}_X = 12.88$

These estimates of the mean and standard deviation of X are the same values as in the case of the theoretical method of the statistical approach.

Alternative Analytic Methods

Consider the probabilistic relationship

$$X = \sum_{i=1}^{M} I_i$$

where I_i is the indicator random variable defined by

$$I_i = \begin{cases} 1 & \text{if the } i\text{th visitor at one time in the developed area is at the site} \\ 0 & \text{otherwise} \end{cases}$$

Note that I_i equals 1 or 0 depending on whether or not the ith visitor is at the site.

The random variable I_i has a Bernoulli distribution with:

Mean of I_i is $\mu_i = p_i$

Variance of I_i is $\sigma_i^2 = p_i(1 - p_i)$

The parameter p_i, $0 \le p_i \le 1$, is the probability that the ith visitor is at the site, or, equivalently, the proportion of the time that the ith visitor is at the site.

Case 1: $M = m$

For given $M = m$ visitors at one time in the developed area, we get

$$X = \sum_{i=1}^{m} I_i$$

Assuming independence among the visitors, then we have from the Central Limit Theorem for independent random variables

X is approximately distributed as normal (μ_X, σ_X)

Loosely put, the theorem states that the sum of a large number of independent random variables has a distribution that is approximately normal.

If in addition $p_i = p$,

X is distributed as binomial (m, p)

which is well known to be approximately normal for large m.

Case 2: $p_i = p$

Suppose all of the visitors in the developed area are at the site the same proportion of the time, $p_i = p$. The probabilistic relationship

$$X = \sum_{i=1}^{M} I_i$$

is the sum of a random number M of identical random variables I_i.

Recall that:

Mean of M is $\mu_M = n\mu_Y$

Variance of M is $\sigma_M^2 = n\sigma_Y^2$

Mean of I is $\mu_I = p$

Variance of I is $\sigma_I^2 = p(1-p)$

From probability theory (Ross, 1993) we have

$$\begin{aligned}\mu_X &= \mu_M \mu_I = np\mu_Y \\ \sigma_X^2 &= \mu_M \sigma_I^2 + \mu_I^2 \sigma_M^2 = np(1-p)\mu_Y + np^2\sigma_Y^2 \\ &= n\left[p(1-p)\mu_Y + p^2\sigma_Y^2\right]\end{aligned}$$

Let us assume that

X is approximately distributed as normal (μ_X, σ_X)

Then

$$P(X \geq x \mid N = n) \doteq P\left(Z \geq \frac{x - 0.5 - \mu_X}{\sigma_X}\right)$$

Problems:

1. Given n and x, find α.

$$P(X \geq x \mid N = n) \doteq P\left(Z \geq \frac{x - 0.5 - \mu_X}{\sigma_X}\right) = P(Z \geq z_\alpha) = \alpha$$

2. Given x and α, find n.

Equate
$$z_\alpha = \frac{x - 0.5 - \mu_X}{\sigma_X}$$
$$= \frac{x - 0.5 - np\mu_Y}{\sqrt{n}\left[p(1-p)\mu_Y + p^2\sigma_Y^2\right]^{1/2}}$$

Solve for n in the equation
$$p\mu_Y n + z_\alpha\left[p(1-p)\mu_Y + p^2\sigma_Y^2\right]^{1/2}\sqrt{n} + 0.5 - x = 0$$

which has the form $an + b\sqrt{n} + c = 0$

where $a = p\mu_Y$, $b = z_\alpha\left[p(1-p)\mu_Y + p^2\sigma_Y^2\right]^{1/2}$, and $c = 0.5 - x$.

Therefore $n = \left[\dfrac{-b + \sqrt{b^2 - 4ac}}{2a}\right]^2$

Monte Carlo Simulation Method

The Monte Carlo simulation method is another probabilistic approach to studying the relationship between N and X. This method is based upon the idea of repeatedly sampling from the random variables on which the random variable X is functionally dependent in order to calculate many possible observations of X from which an approximate distribution of X can be made. Often the number of samplings is in the order of several thousands.

REFERENCES

Devore, J.L., 1991, Probability and statistics for engineering and the sciences: Monterey, California, Brooks/Cole Publishing Company, 3rd ed., 704 p.

Ross, S.M., 1993, Introduction to probability models: San Diego, California, Academic Press, Inc., 5th ed., 556 p.