The River Problem

Douglas A. Kerr, P.E.

Issue 1
December 19, 2005

ABSTRACT

A recreational problem in statistics describes a situation in which an intelligence operative, wishing to know (on very short notice) the fraction of the inhabitants in a certain city who lived north of a river running through the city, contacted one inhabitant at random and determined that he lived on the north side of the river. From that, it was determined that "the best estimate" of the fraction of the inhabitants living north of the river was 2/3. In this article we discuss the possible meanings of that answer and, choosing one for further study, analytically derive that same value.

THE PROBLEM

Several years ago, Paul Gayet presented to me a recreational problem in statistics. As I remember it, the problem was essentially this:

> A military special operations unit had to perform a clandestine operation in a certain city on short notice. They had essentially no information about the city, other than that a river ran through it roughly west-to-east. It would be very helpful in planning their tactics if they knew what fraction of the inhabitants lived north of the river.
>
> They sent an operative into the city to take a survey of this matter, but travel problems got him to the city with only 10 minutes to make his determination and radio his findings to the unit's headquarters.
>
> The operative chose a name at random from the local telephone book, called that person, and asked if he lived north of the river. The operative reported this to the headquarters. There, they were able to make a best estimate of the fraction of the inhabitants living north of the river.
>
> What is that best estimate?

Paul told me that the answer had been said to be "2/3".

SIGNIFICANCE OF THE ANSWER

I recently decided to address this problem. A major complication was to determine just what was meant by the "best estimate of the fraction of the inhabitants living north of the river."

There are three candidate statistical estimators that come to mind. For convenience, I will refer to the fraction of the inhabitants living north of the river as *p*, since that fraction is of necessity the probability that a randomly chosen inhabitant would live north of the river.

- The *maximum likelihood estimator* of *p*. This is the value of *p* that is most likely to exist for that city, given the results of the one-inhabitant survey. (That value of *p* produces the highest probability that the survey will have the outcome we describe).

  It can be readily shown that in our case, the maximum likelihood value of *p* is 1. My guess is that this interpretation of "best estimate" is not what the puzzler had in mind.

- The *expectation* of *p*. That would be the average of the actual values of *p* we would find over a large number of cities having east-west rivers. Note that to address this problem at all, we must know (or assume) the statistical distribution of *p* over all such cities. As you will see later, I will assume a uniform distribution.

  Having done that, we can determine the expectation of *p* with no reference to the result of the survey: it will be 1/2. My guess is that this interpretation of "best estimate" is not what the puzzler had in mind.

- The *conditional expectation* of *p*. That would be the average of the actual values of *p* we would find if we performed our one-inhabitant survey in a large number of cities and only included in our average the cities for which the survey gave the result we got in the problem: the single inhabitant questioned lived on the north side of the city's river.

  I decided to assume that this was the interpretation of "best estimate" contemplated by the puzzler, and undertook an analytical investigation of the problem on that basis.

ANALYSIS

Our survey is an example of the classical model for the binomial distribution, since each "trial" (the questioning of one inhabitant) can produce only two outcomes, "north" (with probability *p*) and "not-north" (with probability 1-*p*).

Thus, for any given actual value of ***p*** for a city, the probability that a survey of ***n*** inhabitants would produce the outcome "north" for ***t*** of those inhabitants, symbolized as $P(p,n,t)$, is given by:

$$P(p,n,t) = \binom{n}{t} p^t (1-p)^{(n-t))} \qquad (1)$$

The symbol $\binom{n}{t}$ means the number of combinations of n things taken t at a time.

If $n = 1$ and $t = 1$, as in our problem, this becomes:

$$P(p,1,1) = \binom{1}{1} p^1 (1-1)^0 = 1 \cdot p \cdot 0^0 = 1 \cdot p \cdot 1 = p \qquad (2)$$

To address this problem at all, we must know (or assume) the statistical distribution of the probability p over all such cities. We will assume for the moment that, over all cities, the probability distribution of the probability p itself is uniform over the range 0 through 1.

Making the expectation "conditional"

If we limit ourselves to those cities for which the "single-trial" experiments yield the answer "north" ($n = 1$, $t = 1$), then the expectation of p, symbolized $E(p)$, over the universe of such cities is given by:

$$E(p) = \frac{\int_{p=0}^{1} P(p,1,1)\, p\, dp}{\int_{p=0}^{1} P(p,1,1)\, dp} \qquad (3)$$

The numerator adds up the weighted values of p (weighted, that is, by the probability, P, that a certain value of p gets that city into the running at all, by virtue of the single surveyed inhabitant answering "north"). The denominator just adds up the weighting factors, as we need to get the weighted average.

Now, substituting for $P(p,1,1)$ from equation 2, we get:

$$E(p) = \frac{\int_{p=0}^{1} p \cdot p\, dp}{\int_{p=0}^{1} p\, dp} = \frac{\int_{p=0}^{1} p^2\, dp}{\int_{p=0}^{1} p\, dp} \qquad (4)$$

Now, writing the indefinite integrals for the functions being integrated in this equation, we get

$$\int p^2\, dp = \frac{1}{3} p^3 + \mathrm{c} \qquad (5)$$

$$\int p\, dp = \frac{1}{2} p^2 + \mathrm{c} \qquad (6)$$

The symbol c represents the infamous "constant of integration". It will disappear in the next step, but I show it for the sake of rigor, lest I receive a visit from the ghost of my high school calculus teacher, who was very stern about this matter!

The result

Thus, evaluating the indefinite integrals at the limits of p (1 and 0), and subtracting, we get for the expectation of p:

$$E(p) = \frac{\frac{1}{3}\cdot 1^3 - \frac{1}{3}\cdot 0^3}{\frac{1}{2}\cdot 1^2 - \frac{1}{2}\cdot 0^2} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} \qquad (7)$$

This is the result claimed in connection with the problem.

CONFIDENCE

Note that this value. 2/3, is not the value of p for the city of interest. It is only a "best estimate", and in fact the particular "best estimate" we described above: the *expectation* of p, conditioned on the answer "north" being given by the only inhabitant surveyed.

There is associated with such estimation the concept of the "confidence interval". There, we ask, "what is the probability, over the universe of interest, that the actual value of p for the city of interest fell within some stated interval containing our result, $E(p)$?" For example, we might be interested in the question, for the city of interest, "what is the probability that the actual value of p lies in the range of 0.5 through 0.75 (our best estimate being 0.667). Evaluating this is beyond the scope of this article.

OTHER SITUATIONS

Larger sample surveys

Suppose that the operative has time to contact a larger sample of inhabitants, all of whom happened to answer "north" ($n>1$, $t=n$). For this broader case, through the process above it can be shown that:

$$E(p) = \frac{n+1}{n+2} \qquad (8)$$

Other results

For experiments in which only part of the citizens surveyed answered "north" ($t \neq n$), the integrals are more difficult; I leave the derivation for this general case up to the reader.

Non-uniform distribution of *p*

The assumption that the distribution of the probability that an inhabitant lived north of a west-east river, over all cities having such a river, was uniform over the range 0-1 is not actually very reasonable. (It is, however, apparently a requirement if we are to get the "official" answer to the problem.)

Suppose the probability distribution of p over all cities is not uniform (as we had assumed), but rather has a probability density function G(p). That function tells us the probability (over all cities) that the "north-side" probability for that city, p, lies in some particular part of the full possible range of 0 to 1.

Then in the equivalent of equation 3, the weighted probabilities become further weighted by G(p), and the equation becomes:

$$E(p) = \frac{\int_{p=0}^{1} \mathrm{G}(p) P(p,1,1) p \, dp}{\int_{p=0}^{1} \mathrm{G}(p) P(p,1,1) \, dp} \qquad (9)$$

I will not attempt to evaluate that for other distributions we might conjecture that the universe of cities might exhibit.

Note that certain alternative distributions that we might decide to assume are not credible in this situation. For example, the *normal distribution* is not applicable to a situation where the random variable (here, p) is known to fall within a limited range (here, 0 through 1).

But one could, for example, assume a distribution, G(p), based on this model:

- The outline of the city, within which all inhabitants live, is circular.
- The distribution of inhabitants by area is uniform over the city.
- The river follows a straight-line west-east path.
- The probability that the river passes through a certain point on the north-south diameter of the city is uniform over distance along that diameter.

That would be tantamount to the assumption of a certain G(p), whose derivation I leave up to the reader.

#