Sampling, Probability Models and Statistical Reasoning

Statistical Inference

Mohan Delampady and V R Padmawar

Statistical inference is introduced here as an application of inductive inference. Using examples it is illustrated how random sampling allows data to be modelled with the help of probability models and how these probability models provide the mathematical tools for statistical inference.

Mohan Delampady received his Ph.D. from Purdue University in 1986. After spending five years at the University of British Columbia he joined Indian Statistical Institute in 1991 and has been with its Bangalore centre since then. His research interests include robustness, nonparametric inference and computing in Bayesian statistics.

V R Padmawar is with Indian Statistical Institute, Bangalore. He received his Ph.D. from Indian Statistical Institute in 1984. His research interests lie in the area of sampling theory.

Inductive Inference

Experimentation is a vital ingredient of scientific advancement. We have all conducted experiments in school or college laboratories at one time or another. For example, consider the simple experiment involving a pendulum to determine the gravitational constant. We obtain data on l, the length of the pendulum and t, the time required for some fixed number, say m, of oscillations. We then substitute these values in the following formula to get the value of the gravitational constant g.

$$g = 4\pi^2 \times \frac{l}{(t/m)^2}$$

We repeat the experiment a few times and for each trial compute the value of g. We then compute the average of these values. Incidentally, most instructors would also want us to report the *experimental error*. We compute the standard error of the values of g based on different trials to get an estimate of the *experimental error*. (The meaning of these *error* terms will become clear as you go on!)

Having conducted the experiment we are willing to accept the value of g that we obtained to hold even outside our specific location. In other words, for example, if the experiment were conducted in Hassan rather than in Bangalore we believe that we would still get more or less the same value of g. This seems to be acceptable to us without actually conducting the experiment in Hassan. Scientists often generalise from a particular experiment to a class of 'similar' experiments. This kind of extension from the particular to the general is called *inductive inference*. This is in contrast to *deductive inference*, an example of which is a mathematical proof of a conjecture. Clearly, inductive inference entails a certain degree of uncertainty.

Scientists often generalise from a particular experiment to a class of 'similar' experiments. This kind of extension from the particular to the general is called *inductive inference.*

One need not be disheartened by this element of uncertainty if the degree of uncertainty can be estimated. This can indeed be done provided we follow certain principles. Statistics plays an important role in providing techniques for making an inductive inference as well as for measuring the degree of uncertainty in such an inference. This uncertainty is measured in terms of probability.

Let us now consider an example of inductive inference. A consignment of 10,000 units each having its own identification number comes to the office of a statistical officer in a firm. A unit can be classified as defective or nondefective depending on certain specified standards and criteria. The officer's job is to decide whether to accept or reject the consignment in its totality. As common sense dictates, we would accept the consignment if there weren't 'too many' defective units, or in other words, if the proportion of defective units in the consignment is not very substantial. Let θ denote this unknown proportion of defectives. Our objective now is to determine this θ. One way of achieving this is to test each and every unit by subjecting it to the specified standards and criteria. This would tell us whether a given unit is defective or not. This in turn would give us the total number or equivalently the proportion of defective units in the consignment. Such a procedure is often long drawn and expensive. In many situations such as measuring the breaking strength, we end up destroying the units while testing them. Isn't it futile if at the end of the procedure, we know the exact proportion of defectives θ in the consignment but are left with no usable units?

A thought that comes to mind is whether one can subject only a few units to the specified standards and criteria and based on these few units make a statement about the unknown θ. The answer is that we cannot determine the exact value of θ but can make a probabilistic statement about it provided we select the few units in accordance with certain principles. 'Random sampling' which is discussed later in some examples is one such technique to select a few units from the entire 'population' that satisfies the above requirement. It is shown later that random sampling allows the data obtained to be modelled using probability models. These probability models in turn provide the mathematical tools for statistical inference.

In science as well as in human affairs we often lack resources to study more than a glimpse of the phenomenon that might advance our knowledge!

Moreover, if we adopt the above testing procedure only for a sample from the consignment rather than the whole consignment, we would save on cost, time, as well as effort. Prior to the advent of pressure cookers many of us have seen how our grandmothers would mash a few grains of rice to determine whether the entire rice in the vessel was properly cooked or not. Thanks to their wisdom we were never left starving and we seldom ate undercooked rice. In science as well as in human affairs we lack resources to study more than a glimpse of the phenomenon that might advance our knowledge!

Thus, we select a sample of a few units from the consignment, observe the number of defective units in the sample, and from this knowledge try to predict the unknown proportion of defectives θ in the consignment. We cannot be certain of our answer but we can make a statement about the error in our answer. This gives us an idea about inductive inference.

Even in our problem we could have asked ourselves the following two questions :

1. What is the value of θ?
2. Is $\theta \leq .01$(say)?

The first question leads to the *theory of estimation* whereas the second question leads to the *theory of testing of hypotheses*. In what follows we discuss some simple examples where statistical inference can be done meaningfully.

Probability Models and Statistical Inference

Example 1. Suppose that some of the printed circuit boards manufactured by a company have a certain defect which can only be detected with extensive testing. A "random sample" of n of these boards is chosen for testing. A random sample is one where every unit in the population has the same chance of being included in the sample. Suppose k out of these n turn out to be defective and the rest are fine. What can be said about the proportion θ of defective boards in the entire population of boards that this company has produced, if we can assume that the total number of boards produced is very large?

A random sample is one where every unit in the population has the same chance of being included in the sample.

First, note that θ is also the probability that a randomly chosen board (from the population) is defective. Now, let X denote the number of defective boards in the sample of n boards which were selected. For any integer x between 0 and n, we claim that

$$P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \tag{1}$$

This can be justified as follows. Sampling n units from a population involves n trials of the same experiment, i.e. that of choosing a unit at random. Since the population size is huge, we can assume that the outcomes of different trials are independent. Therefore the probability of obtaining any given sequence of outcomes which consists of x defectives and $n-x$ non-defectives is $\theta^x (1-\theta)^{n-x}$. Since there are $\binom{n}{x}$ such sequences,the probability of the event $\{X = x\}$ must be what we specified above. (Refer to Karandikar, R L 1996, On Randomness and Probability, ***Resonance***, Vol. 1, No. 2, pp. 55-68 for related material.) The probability distribution specified by (1) is called the *binomial distribution*. Using (1) it can be shown easily that the expectation of X is $E(X) = n\theta$ and the its variance is $Var(X) = n\theta(1-\theta)$.

For each θ in the unit interval (0, 1), (1) gives a different binomial probability model for X. What is the model in which observing the event $\{X=k\}$ is most likely? Let us denote by $P(X = x|\theta)$ the probability specified by (1) for the given θ. Let us define the function $L(\theta)$ by

$$L(\theta) = P(X = k/\theta),$$

where k is the observed number of defectives (which is known now). Note that L is a nonnegative function defined on the set of all possible values of θ which is called the parameter space. This function L is called the likelihood function of the unknown quantity θ and clearly this measures how likely is the event $X=k$, if θ is indeed the true value of the proportion of defectives in the entire population.

One appealing method to *estimate* θ is to find the value of θ which maximizes the likelihood function. This is called maximum likelihood estimation and the interpretation of the obtained estimate is that it gives the model, from all the models considered in (1) in which the observed event is most likely.

In the problem above, the maximum likelihood estimate of θ is $\hat{\theta} = k/n$ since this is the value of θ which maximizes $\binom{n}{x}\theta^k(1-\theta)^{n-k}$. Since X denotes the (random) number of defectives in the sample of (fixed) size n, we see that the maximum likelihood estimation method provides X/n as the estimator for θ. Using the *Weak Law of Large Numbers* (see Karandikar, 1996)

Polya's View

"Experience modifies human beliefs. We learn from experience or rather, we ought to learn from experience. To make the best possible use of experience is one of the great human tasks and to work for this task is the proper vocation of scientists. A scientist deserving this name endeavours to extract the most correct belief from a given experience and to gather the most appropriate experience in order to establish the correct belief regarding the correct question."

it can be shown that the maximum likelihood estimator, X/n approaches θ as $n \rightarrow \infty$. It is comforting to note that if a large number of boards chosen at random are tested it is indeed possible by the above method to determine the actual proportion of defectives.

The precision of an estimator can be measured by its *standard deviation* (again refer to Karandikar, 1996 for details) if the estimator's expectation is the quantity it is supposed to estimate. Note that, smaller the standard deviation better the estimate on average. In the present case, $E(X/n) = n\,\theta/n = \theta$ and $Var(X/n) = n\,\theta\,(1-\theta)/n^2 = \theta\,(1-\theta)/n$. Therefore the standard deviation of the estimator X/n is $\sqrt{\theta(1-\theta)/n}$. For the observed data, namely k defectives in the sample of n, θ is being estimated by $\hat{\theta} = k/n$ and hence a measure of precision of the estimate is given by $[\hat{\theta}\,(1-\hat{\theta}\,)/n\}]^{1/2}$.

It is comforting to note that if a large number of printed circuit boards chosen at random are tested it is indeed possible by the maximum likelihood method to determine the actual proportion of defectives.

Example 2. What if θ is very small in the above problem? If n is not very large, most of the samples may show 0 defectives. Then $\hat{\theta}$ is also 0. This is not good enough if we want a good idea about the proportion of defectives in the population. It is clear that the information provided by our experiment is inadequate. How can we modify our experiment to gather more informative data?

One suggestion is to continue (random) sampling until a prefixed number, say r, of defectives is observed. What is the data now? It is simply X = number of sampled boards which obtains r defectives in the sample. It can be seen, arguing as in the above example, that

$$P(X = x) = \binom{x-1}{r-1}(1-\theta)^{x-r}\,\theta^r$$

for any integer $x \geq r$. This probability distribution is known as the *Negative Binomial* distribution. When $r = 1$, it is generally called the *Geometric distribution*. Estimation of θ can be done following the steps of the above example.

Example 3. Suppose we want to estimate the number of fish in a lake or in a particular part of a sea. Let N denote this unknown number of units. Consider the following experiment which is known as the *capture-recapture* method. Catch N_1 fish from this population. Tag all of them and release them. Now fish again. This time suppose that n fish are caught, n_1 of which have tags on them. Clearly $N_1 \le N$ and $n_1 < n$. What is an estimate of N and how good is this estimate? N clearly cannot be less than $N_1 + (n-n_1)$ which is the total number of fish that we have seen in the two stages of fishing. For any value of $N \ge N_1 + (n - n_1)$ what is the probability of observing n_1 tagged fish among n fish caught in the second stage? This is exactly equal to the ratio of *the total number of ways of choosing n units from N units such that n_1 of them came from a fixed set of N_1 to the total number of ways of choosing n units from N units.* Therefore, the probability is $p(N) = \binom{N_1}{n_1}\binom{N-N_1}{n-n_1}/\binom{N}{n}$. The estimation procedure discussed above can be easily implemented here too by first obtaining a likelihood function for the unknown parameter N using the above probability expression. As is discussed in Feller (1993), to find the maximum likelihood estimate of N consider the ratio

$$\frac{p(N)}{p(N-1)} = \frac{(N-N_1)\,(N-n)}{(N-N_1-n+n_1)N}\,.$$

Note that this ratio is greater than or smaller than 1, according as $Nn_1 < N_1n$ or $Nn_1 > N_1n$. Therefore, with increasing N the sequence $p(N)$ first increases and then decreases; it reaches its maximum when N is the integer part of N_1n/n_1, so that the maximum likelihood estimate of N is approximately N_1n/n_1. This is an intuitive estimate since it implies that approximately n/N equals n_1/N_1; it is the same as saying that the observed proportion of tagged fish should be approximately the same as the proportion of sampled fish.

The *capture-recapture* method is an ingenious way to estimate the number of fish in a lake or in a particular part of a sea.

The next example is about elections. Since the elections are just around the corner and various commercial agencies will be making *expert predictions* about which party will win, let us consider the statistical methods they employ!

Example 4. Suppose that we want to know the proportion of eligible voters who support a particular political party. A random sample of size n is selected from this population and suppose k voters support this party. What is a good estimate of the required proportion? How do we obtain a probability model for the experiment just conducted? Let us examine the following simple experiment. Consider a box containing a large number of marbles, a proportion θ of which are red. If n marbles are chosen at random from this box, what is the probability of getting k red marbles? The answer depends on whether the sampling is done with or without replacement. Random sampling with replacement means that the unit sampled is replaced in the box before the next draw is made. In random sampling without replacement the sampled units are not replaced. If the sampling is with replacement the probability distribution of the number (X) of red marbles in the sample is given by the Binomial probability model of (1). If it is without replacement then we get

If the sample size n is very small compared to the population size, the hypergeometric model can be approximated by the binomial model.

$$P(X=x) = \frac{\binom{N\theta}{x}\binom{N(1-\theta)}{n-x}}{\binom{N}{n}}$$

This is the *Hypergeometric probability distribution.* What can be seen is that if the composition of the box doesn't change from draw to draw then the draws are independent and hence we get the Binomial model instead of the Hypergeometric. Note that, if the number of sampled units n is very small compared to the total number N of marbles in the box then the composition of the box doesn't change much from draw to draw even if the sampling is done without replacement. Therefore the Binomial model should be available as the approximation of the Hypergeometric model for large N. This is indeed true as the following result shows.

$$\lim_{N \to \infty} \frac{\binom{N\theta}{x}\binom{N(1-\theta)}{n-x}}{\binom{N}{n}} = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

To prove this, expand all three terms in the left hand side using factorials, factor out the right hand side from there and then note that whatever remains converges to 1 as $N \to \infty$.

Let us return to the question of predicting the outcome of elections. If a random sample of n voters contains k who support a particular political party, then from the exact Hypergeometric model or the approximate Binomial model, one can check that the maximum likelihood estimate of the proportion of voters in the population who support that political party is simply k/n. Of course, the commercial agencies which conduct the surveys may employ more advanced techniques such as dividing the sample size between different cities according to their population and other important factors. When random sampling involves such additional constraints, the probability models for the sampling design need to be modified accordingly.

If a random sample of n voters contains k who support a particular political party, one can show that the maximum likelihood estimate of the proportion of voters in the population who support the political party is simply k/n.

Example 5. Let us see how our school experiment to determine g can be put in the set up of statistical inference. Let y_i be the value of the gravitational constant that we obtain in the ith trial (for $i=1, \dots, n$). Then we can represent or model our data as follows:

$$y_i = g + \varepsilon_i, \quad i = 1, \dots, n,$$

where g is the true (but unknown) value of the gravitational constant and ε_i is the combined error due to various factors which are beyond our control in the ith trial of the experiment. If we assume that there are no systematic errors in the experiment then the expectation of ε is $E(\varepsilon) = 0$. The variance of ε is $Var(\varepsilon) =$

σ^2 which indicates how precise the experiment is. The Gaussian distribution (bell-shaped curve, to be discussed in detail in another article) is normally used to model measurement errors like the ε above. This will imply that our data, $y_1, y_2, \ldots, y_n$ are independent and identically distributed Gaussian (or normal) random variables with expectation g and variance σ^2. It can be shown using the probability density of the Gaussian distribution that the maximum likelihood estimate of g is $\bar{y}$, the average of the observations. This is indeed what we were told to report as the value of g by our instructor, isn't it? To obtain a measure of precision of this estimate, we note that $E(\bar{y}) = g$ and $Var(\bar{y}, y) = \sigma^2/n$. It can be shown again using the probability density of the Gaussian distribution that the maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$. Therefore, an estimate of the experimental error is $\hat{\sigma}/\sqrt{n}$.

An important point to note is that a probability distribution is used in each of the examples to model data. This is the only way to obtain optimal statistical procedures.

In the above discussion we tried to give a flavour of statistical inference using some simple illustrations. The important point to note is that a probability distribution is used in each of the examples to model data. This is the only way to obtain optimal statistical procedures. It can be readily seen that the scope of statistics is much wider than what is discussed above using mostly simple discrete probability models to estimate unknown parameters. One very important topic which is not covered at all is hypothesis testing. Likelihood methods again play a substantial role here as well. The methodology involved here is material for a future article.

Address for correspondence
Mohan Delampady
Indian Statistical Institute,
8th Mile, Mysore Road
Bangalore 560 059, India.

Suggested Reading

G Polya. Induction and Anology in Mathematics. Princeton Univ. Press.1954.

G Polya. Patterns of Plausible Inference. (2nd ed.). Princeton University Press. 1968.

M A Mood, F A Graybill, and D C Boes. Introduction to the Theory of Statistics. McGraw-Hill Kogakusha. (3rd Ed. Intl. Student Edition). 1974.

W Feller. An Introduction to Probability Theory and Its Applications. Wiley- Eastern, New Delhi. Vol.1 (Third Edition).1993.