Statistics and Confidence

Statistics and probability are complementary to each other. In probability theory, we are given probabilities of events and want to say something about what might actually occur. Conversely, in statistical theory, we are given the events that actually occurred and want to say something about the probability of the events.

This paper is about statistics. We will explore how to derive probabilities from data and develop a measure of how confident we are that the derived probabilities are true.

Let us consider a simple yet remarkably subtle problem. We do *N* flips of a coin and we observe a certain list of outcomes. This specific result out of the many possible is called a *realization*. Given this single realization, what can we say about the probability of landing on each side of the coin and how certain we are about our answer?

First some basics:

1. We label one side of the coin with T (true) and the other with F (false).
2. The number of times the event T is observed will be called  and for F it will be . Since the events are mutually exclusive, .
3. Using standard probability theory:  
     
    

QUESTION: *Given a process that emits 2 mutually exclusive symbols {T,F} with* P(T) = p *and* P(F) = 1-p*, and also given that we observe a sequence of realizations of length N with occurrences of the* T *event, what is the value of* p *that is most likely to generate the observed sequence? We will refer to our estimate of this value as .*

Our approach will be to compute how likely it is for our observed sequence of realizations to occur given different choices of *p* and solve for the maximum.

Since each event in the sequence is independent, the probability of the realization of any particular sequence does not depend on order. This is simply given by:

 (eq. 1)

Note that we explicitly state the conditionals in , but to save typing from here on, some of the conditionals will be implied and we will use the simplified shorthand . The value of *p* most likely to generate our sequence is the value of *p* that maximizes the equation above. To find it, we take the partial derivative with respect to *p*, set it equal to zero and solve for *p*:



First partial:







Second partial:







And finally:

 (eq. 2)

By inspection, it is clear that if  and , then *eq.2* can only equal zero if the last term is zero. Solving, gives us the solution . In the cases where or , *eq. 2* runs into singularities which we could resolve using limit arguments involving the squeeze theorem. Instead, we will use a far simpler argument. When, the sequence *must have*  and our solution holds. Similarly, when , the sequence *must have* and again our solution holds.

*Result 1: Given a process that emits 2 mutually exclusive symbols {T,F} with* P(T) = p *and* P(F) = 1-p*, and also given that we observe a sequence of realizations of length N with occurrences of the* T *event, then the value of p that is most likely to generate that sequence is given by .*

QUESTION: *The value  is called the sample probability. A similar question we can ask is what value of p is most likely to generate a sequence with sample probability?*

This is a subtly different question. We are not asking which value of *p* is most likely to generate a *particular* sequence but *any* sequence with the same. Since there are  equivalent sequences that generate the same, this answer is given by:



The only difference is the multiplied  factor and since it does not depend on *p,* it simply gets propagated to the derivative when we solve for the value of *p* that will maximize this equation. Thus, we can jump straight to the solution:

 (eq. 3)

For the same reasoning as above, we conclude that, and thus.

*Result 2: Given a process that emits 2 mutually exclusive symbols {T,F} with* P(T) = p *and* P(F) = 1-p*, and also given that we observe a sequence of realizations of length N with occurrences of the* T *event, then the value of p that is most likely to generate ANY sequence with events of type T is also given by .*

QUESTION: *A similar but subtly very different question is to NOT ask which value of* p *is most likely to generate the observed sequence, but instead to ask which value of* p *is most likely to have generated the sequence? We will refer to our estimate of this value as .*

We begin by noting that each realization is a single bit of information. With a single bit, one cannot expect to do anything better than to first separate all possible probabilities into 2 *a priori* groups and on acquiring that bit, use it to select among these groups. With 2 bits, we have 4 potential sequences, but ignoring order, we only have 3 unique ones. In fact, when we have *N* bits and ignore order, the differentiator between groups is the number of T events (i.e.), and so the best we can achieve is to separate all possible probabilities into groups and use the observed sequence to select between them. In the discussion below, we will use the value of as a *label* for identifying each group.

We know we need to create  groups in order to estimate the probability, but are these groups contiguous? And how big should each bin be? To answer both of these questions, our approach will be to make choices that minimize the expected error of the estimate we are developing.

=== inline experimentation ===

The expected error is:



Since probs equally distrib:



Substituting:



To minimize subject to constraint , we need to minimize the Lagrangian:



We do this by first calculating all partial derivatives:





And solving for when they all equal zero:



The first equation group can be solved to:



Which means that, even without knowing , we know that all bin sizes must be the same. Plugging into the second equation, we get:



And finally, plugging back into our first result:



=== end ===

If we arrange the bin boundaries such that each bin contains the values of *p* that are more likely to generate a sequence with the label of its group rather than a sequence with the label of some other group, then:



To reduce typing, let us use the following abbreviations:



Then:

 (eq. 4)

The relationship of the first terms of each side is what we’re interested in knowing (because we are trying to determine which probability bin is most likely to have generated the observed sequence) so we need to quantify the second terms so that this equation suits our purpose. This means we need to further characterize the bins. Once we know the bin boundaries, we can compute their sizes. Once we know their sizes, we can compute their probabilities. And once we know their probabilities, we can identify the bin that is most likely to have generated the observed sequence. Again, by the arguments above, identifying the most likely bin is the best any process can do.

First, we must identify the bin boundaries. These boundaries are the values of *p* where the probability of producing a sequence in one group is the same as producing a sequence in the adjacent group. We only care about the adjacent group because (by construction) when *p* crosses these values, the most likely group label to be generated changes to a new label, signifying a bin boundary:











 (eq. 5)

The first (implied) boundary is always at zero since probabilities cannot be negative. By inspection of the equation above, we see that subsequent boundaries are uniformly spaced in  intervals with the last boundary at. See figure below for.

|  |
| --- |
|  |
| Colored curves are probabilities of generating each group label. Dotted lines are bin boundaries and lie at the intersection of the probability curves of adjacent labels. Solid lines represent the group labels. They are placed here at positions  (sample probabilities) as a reference for visualizing the bin boundaries. Colored circles are centers of group bins (more about bin centers below). |

To go further, we need to determine the distribution of the values of *p.* From one point of view, we cannot say we know this. The distribution of values of *p* can in actuality be anything at all. Yet, when referring to our *incremental knowledge of p,* we can intuitively say that when, we have no knowledge whatsoever and thus that we are maximally uncertain of the value of *p.* In the language of Information Theory, this maximally uncertain situation corresponds to the distribution of *p* having maximum *entropy,* which occurs when the distribution of *p* is uniform.

We can view our method of learning about *p* as a process where we start with a maximally uncertain (and hence uniform) distribution for *p* and refine it incrementally with each realization we observe.

=== Took a turn in a different direction here … stuff below needs to be reworked =====

Given that the true probability is uniformly distributed and our result that all bins have the same size, we conclude that probability of being in a bin is the same for all bins. Using this information (and expanding the abbreviations)*,* we have:



Expanding

=================== WE ARE HERE ====================

So they have the same prob and the equation above is resolved and probs being uniform the least error is the center of the bin.

QUESTION: *We have shown that*  *is the most likely value for p to have generated our sequence or an equivalent one with the same, but how useful is this value?*

The motivation for this question is the observation that is simply a ratio that gives no indication of *N.* Yet, it is intuitive, for example, that observing a single T event in a sequence with  events is much less informative that observing 1000 T events in a sequence with  events. In the first case, we would be hard pressed to guess what the next event will be. In the second case, we would be very surprised if we got an *F* event. Yet both sequences generate the same.

TODO: The difference is the density. Each possible labels a set of sequences but also a bin of probabilities. Need to show that is the *CLOSEST* value to *p*. Then maybe do something like this: If this is so, then bins are of known size (maybe equal except at zero and one? Or something else). *P* is uniformly distrib therefore the number of p per bin is proportional to size of bin. Therefore the density is inversely proportional to the size of the bin. It is intuitive that p\_hat approaches p as N increases, therefore eq.2 must be narrowing and the density of p in the bin must be either increasing or decreasing at a slower rate. Figure this mess out mathematically.

SCRAP AREA – RANDOM THOUGHTS OF DUBIOUS VALUE

Curiously, but I don’t know what use it might be, the boundaries coincide with the possible labels of a sequence that is one sample longer. Hmmm … wonder if this means something.

====

Before we have any samples, we have no information about *p.* Intuitively, this translates to stating that all values of *p* are equally likely, or, in other words, that the possible values of *p* are *uniformly distributed.* With this assumption in hand, we can then separate the possible values of *p* into “bins” where each bin contains the values of *p* more likely to generate

QUESTION: *If we consider to be a label for a bin of probabilities, what are the bin boundaries?*

==========

Consider that for a particular *N*, many sequences can generate the same, thuscan also be thought of as a label that identifies a subset of all the possible sequences of length *N.* The value for  will be in the range , thus there are  subsets and associated labels. Usingas a label, we can partition the space of possible sequences of length *N* in two: one subset where  is the label (i.e. the subset of sequences where ), and all the others.

THOUGHT 1

The number of ways in which we can produce a realization that has the same *N* and  as our observed realization can be computed by first noting that this question is the same as asking “how many different ways can I place  events of type T in a sequence that is *N* events long?” Next we note that this can be converted into an N-choose-K problem by using the analogy that we are choosing  “balls” (representing the T positions) from a set of *N* balls representing sequence length. Thus, the number of ways an equivalent realization (i.e. one where only event positions change, not their number) to our actual realization can be achieved is . And further, since , then .

The probability that a particular equivalent realization will occur is given by *eq.1* above. Similarly, the probability that *any one* of the equivalent realizations is the number of equivalent sequences times the probability of a sequence. Plugging in our estimate  we have:

 (*eq.3)*

THOUGHT 2

We talk a lot about binomial processes above (i.e. where probs = {P, 1-P}). Is it true that any process can be expressed in terms of binomial processes? If not, when can’t it? What does *any process* mean anyway?

THOUGHT 3

Let’s say we have two events, *C* and *E,* that we are considering as a potential cause and effect link. *C* occurs before *E* and we are only considering *E* when it occurs a specific time after *C.*

In a single observation, we might observe one of four possibilities: . If we now, observe this many times, let us consider the meaning of certain conditional probabilities when they approach:

|  |  |
| --- | --- |
| Limit | Meaning |
|  | is a *sufficient* *cause* of |
|  | is a *sufficient* cause of |
|  | is a *sufficient* cause of (i.e. inhibits*)* |
|  | is a *sufficient* cause of (i.e. inhibits*)* |
|  | is a *required* cause of |
|  | is a *required* cause of |
|  | is a *required* cause of (i.e. *is required* to inhibit*)* |
|  | is a *required* cause of (i.e. *is required* to inhibit*)* |

Note, however, that  so these limits are mutually exclusive in pairs and thus are a redundant description of a causal link. For example,  implies that .

As a consequence, to completely describe these limiting situations, we only need a set of any four of the independent conditional probabilities. The choice of which 4 is arbitrary with respect to the information being captured but we can use our intuitions about the world and intelligent systems to choose a set that is likely to be more efficient in terms of the implementation of an actual intelligent system.

As mentioned before, the purpose of an intelligent system is to conscientiously intervene in the natural unfolding of world events in order to increase the odds that its goal measure will increase. It is intuitive that environments that benefit from intelligence are those where the correct decisions to intervene are sporadic and/or varied.

Consider an environment where a certain correct intervention is *not* sporadic (i.e. the probability of the intervention being helpful is greater than 50%), such a system would benefit from intervening constantly. Such a system does not need great intelligence; this constant intervention could likely arise through the mechanisms of evolution: by evolving a thicker bone as an intervention against gravity breaking it, for example.

Now consider the converse: the environment does *not* guarantee that a particular intervention is helpful > 50% of the time, instead, only during certain *situations* can we know that an intervention is likely to be helpful. Now we require a form of intelligence; these situations must be identified and we must have a “rule” such that the intervention is only made at those times.

The variety of correct interventions a system can orchestrate is also a factor. A system that can only direct one intervention, for example, is intuitively not as effective and not considered as intelligent as one that can direct a variety of possible interventions depending on the situation.

From these intuitions, we can justify our choice of the set of conditional probabilities an efficient intelligent system should store:

1. Arbitrarily, we will discuss a system with “positive” logic: means an event *has* occurred and  means it has not. This choice does not affect our further conclusions.
2. An intelligent system must focus on events that are correlated to changes in its goal measure. These events are *effects* that the system must attempt to manipulate. If the effect is helpful, the system must seek the causes of  in order to promote it. If it is harmful, the system must seek the causes of  in order to inhibit it.

------- OK up to here ------

1. Since these important events are intuitively sporadic, they are active less often than they are inactive so we will use fewer resources if we only do computations when the event is active. This argument favors keeping track of probabilities involving
2. Given an arbitrary effect, it is intuitive that it is represented as an event out of many the system can detect.