Article From:

The Actuary

March 1986 – Volume No. 20, Issue No. 3

MAJOR LEAGUE BASEBALL — AS A MARKOV CHAIN

By C.L. Trowbridge

The game of baseball has characteristics that suggest its analysis as a Markov chain. Such an analysis may be of interest to statistically minded baseball fans, and may prove to have some practical application.

The team on the offense (i.e., at bat) finds itself in one of 24 "states", depending on how many are out, and the number and position of baserunners. Each half-inning begins with none on and none out (state A) and ends, after three are out, in the 25th or "absorbing" state Z. Between A and Z the half-inning wanders through the alphabet, sometimes forward, sometimes back; sometimes scoring runs, sometimes not.

A code for "State"

Let the letters A thru H represent the 8 no-out states, in the following order: In A the bases are empty; B, C, and D represent *single* baserunners on 1st, 2nd, and 3rd respectively; E, F, and G represent states with *two* baserunners, on 1st and 2nd, 1st and 3rd, and 2nd and 3rd; in H the bases are full.

Let the letters I through P similarly represent the 8 one-out states, and Q through X the 8 states where two are out. As previously noted, Z is the side-out state that ends each half-inning. Y is the only unused letter.

Half-inning Records

A half-inning changes state whenever (a) there is an out, (b) there is a change in the number or position of base runners, or (c) a run is scored. Note that some changes occur as a *result* of a time at bat, while others take place *during* a time at bat. States that occur only briefly during continuous action are disregarded.

A half-inning record is a listing of the letter codes for the successive states, in the order that they appear, supplemented by an indication of the number (if any) of the runs scored in the transition from one state to another. Examples will make this clear:

A I Q Z Three up and three down

A I K KI S Z After an initial out, back to back doubles, then 2 outs

A AI I Q Z Lead off home run, followed by 3 outs

A B F A3 I K S U T2 QI Z A big inning with 6 runs

Note that the letters can repeat, but only if runs score, and if any baserunners are "replaced."

Transition Pairs

The rules of baseball impose some conditions upon the changes from one state to the next.

(a) The number out can never decrease.

(b) Baserunners can stay still, disappear, or go forward; but they can not move back.

(c) In any transition the number of new baserunners cannot exceed 1.

(d) The number out in any transition cannot exceed the initial number of baserunners plus 1.

Because of these limitations, of the 600 potential pairs of successive states some 310 are impossible. Examples of impossible pairs are QA, LM, JP, and BZ.

There are some 90 *extra* transition possibilities. Two or more transitions may be indicated by the same letter pair, the difference lying in the run indicator carried by the second letter:

B C represents a runner's advance from 1st to 2nd with no change in batter but B C1 indicates that the runner has scored as the batter reaches 2nd.

S Z represents a 3rd out leaving a runner on 2nd.

S Z1 indicates that the runner scored prior to the out.

After eliminating the impossible and adding the extras, we find the universe of possible transitions to be about 380.

A Probability Matrix

It is not difficult to visualize a probability matrix, consisting of some 380 entries of the form Pr $s_1 \rightarrow s_2$, where s_1 stands for the initial state, s_2 for its successor. If we can assume that these probabilities are independent of how state s_1 came into being, the Markov model may be valid.

We should be able to get good estimates of the entire matrix via the collection of a large number of half-inning records. (From all the major league games of a single season, a complete set of half-inning records will number approximately 35,000). Count all of the times that state s_1 appears, and compute the proportion of such appearances that are immediately followed by state s_2. This proportion would seem to be a good estimate of Pr $s_1 \rightarrow s_2$.

Since the probabilities so computed are averages, they may well be inappropriate when we have further information. If we know that the batter (or the pitcher) is somehow "unaverage", we must assume that the true probabilities are different.

The Average Expected Value

Let V_s represent the "expected value" or the "run potential" of state s. Turn to the collection of half-inning records and the count of the number of times that each state s appears. Then count the number of runs indicated *after* state s but in the same half-inning. V_s can be estimated as the second count divided by the first. Because these 24 values of V_s too are averages, we will refer to V_s as the average expected value, measured in runs, of state s.

Common sense tells us something about the magnitude of V_A. By its very nature V_A must be close to "average runs per half-inning", which is about 0.45. V_s should decrease with the number of outs ($V_Z=0$); and increase with baserunner advance. The largest of the 24 should be V_H, the smallest V_Q. Actual calculation of the V_ss, which to our knowledge has not been attempted, should remove any speculation as to the relative sizes of the 24 values.

Interrelationships

If it is true that "half-inning records have no memory", and that the Markov model is reasonably accurate, V_s can be expressed in terms of the probability matrix and the average expected values of all successor states. Checks by means of a formula similar to that shown below should prove useful in solidifying the estimates of both the V_s vector and the probability matrix.

$$V_Q = \Pr Q \rightarrow R \cdot V_R + \Pr Q \rightarrow S \cdot V_S + \Pr Q \rightarrow T \cdot V_T + \Pr Q \rightarrow Q1(1+V_Q) + \Pr Q \rightarrow Z \cdot 0.$$

Applications

Interesting as this theory may be, it will have little practical use unless it can be applied to real baseball problems.

As a first application, consider the V_s table and its relationship to baseball strategy. Experienced managers may well have some feel for the chances of a rally after two are out, or what to expect from a good start; but the availability of V_Q in one case, of V_H in the other,

(Continued on page 5)

Major League Baseball

(Continued from page 4)

should be of help to informed judgement. A comparison of V_K with V_B tells something about the wisdom of sacrificing with one on and none out, while the two-on one-out sacrifice can be partially evaluated by comparing V_W and V_N. Such comparisons do not provide definitive answers because they do not take into account the possibilities of either unsuccessful or overly successful sacrifice attempts, but they provide useful information nonetheless.

There is another potential use with greater possibilities. Just as a stock market average permits the comparison of an individual stock against the average, the V_ss make possible the comparison of one offensive player against the average of all, and hence against any other. Moreover, one average can include all of the skills of the baseball offense, baserunning, runs batted in, as well as the more commonly calculated "batting" average.

Let a player's "offensive performance index" be calculated by adding (algebraically) his "values added" and dividing by "times at bat". A value added is V_s for the state after he has batted + any runs indicated by the attachment to $V_s - V_s$ for the state in which he came to bat. Value added will normally be positive if the batter gets a hit, draws a walk, or especially if he drives in runs; but will be negative if he makes an out without advancing a runner, or especially if he hits into a double play.

Baserunning skill too is measured by value added. With two out a runner on 1st successfully steals 2nd. He is credited with $V_S - V_R$, whereas if he fails his value added is $-V_R$. Whenever state changes, but the batter is not involved, the value added is charged or credited to the baserunner, and treated as if it were a part of his earlier time at bat. In relatively infrequent circumstances it may be appropriate to credit part of the value added to the batter, another part to the baserunner. As an example, in state B the batter singles, and the runner on 1st advances to 3rd. If in the opinion of the official scorer the extra base is more the result of the baserunner's speed than the place to which the ball was hit, $V_E - V_B$ might be credited to the batter, $V_F - V_E$ to the baserunner now on third.

Summary

The analysis of the game of baseball, in terms of a Markov chain, shows promise. The Markov theory is based on the premise that the chain has no memory, so that the transitional probabilities starting from state s are independent of what occurred before. Whether this is truly characteristic of the game of baseball may be difficult to determine, but it seems to be a reasonable assumption.

The author knows of no attempts to quantify the transition matrix or the potential run values indicated here by V_s. Such an attempt would seem to be the next logical step. If successful, the spin-offs might be surprising. □

Settlement Annuity

(Continued from page 1)

definiteness of the claim settlement, and that fact that the claim will be lower by the action of interest. The life company finds that substantial premium is produced in big chunks, and the annuity payments have good cash flow characteristics. The Settlement Annuity seems to be one of those synergetic arrangements in which everyone gains.

The life insurance company issuing such annuities must answer two main questions as to pricing. What interest rates and what mortality table are to be used in the rate development, and how are persons with life-impairing injuries to be evaluated? Other questions are expense factors, profit margins, the types of annuities, and any limits that the company may need to impose.

An Actuary or life insurance executive involved with this business, or considering becoming involved, must realize that theoretical assumptions are of no value unless business is produced. To produce business, by far the most important factor is the premium quoted. In this field the lowest premium is vital — perhaps to a greater extent than in most other insurance lines. The premium is being paid by a casualty insurance company, which has more than a "casual" interest in keeping its claim costs low. The casualty company will (and should) demand the lowest possible premium from the life insurance industry.

The agent or broker presenting the rates of a life insurer has not only an ethical responsibility to present the lowest rate (assuming the life companies available are all well rated and with good financial structure), but a very practical reason for doing so. If he doesn't, someone else will! And if someone else does, the agent will not only lose this case, but the casualty company as a customer for future cases. Unless a life company is willing and able to be price competitive, there is no reason for it to be in the field.

This does not mean that a company whose rates are not the lowest will get no business at all; nor that the company with the lowest rates will get all of the business. Service to the agent, flexibility, and cooperation are important, and if rates are reasonably competitive will attract a certain amount of business. Poor service, inflexibility, and lack of cooperation will drive away business despite a low rate. Rate structures are not static, so one life company may be the lowest bidder on one case, another life insurer on the next.

There is one new development. The author is now engaged, full-time, as an expert witness testifying in court trials as to Settlement Annuity costs. This testimony seems to be very successful in reducing huge lump-sum verdicts to more reasonable levels. Judges like it, juries can understand it, and the casualty companies truly appreciate the savings in claim dollars.

The Settlement Annuity field is fast rolling, fun, frightening at times, amazing as to the premium volumes than can be developed — and always exciting. □

WANTED: M.A.A. REPRESENTATIVES

The Subcommittee on Relations with Colleges, Universities, High Schools and Related Matters is seeking additional volunteer FSA's to serve as local representatives for the Mathematical Association of America. The function of these is to promote the math contests sponsored in part by the Society and the MAA. They are called upon from time to time to disseminate MAA information in the local areas. Their purpose is to foster widespread cognizance of MAA activities. Interested persons can contact Bob Musen at CIGNA (203-726-5331). □