

Baseball Over-Under Estimation

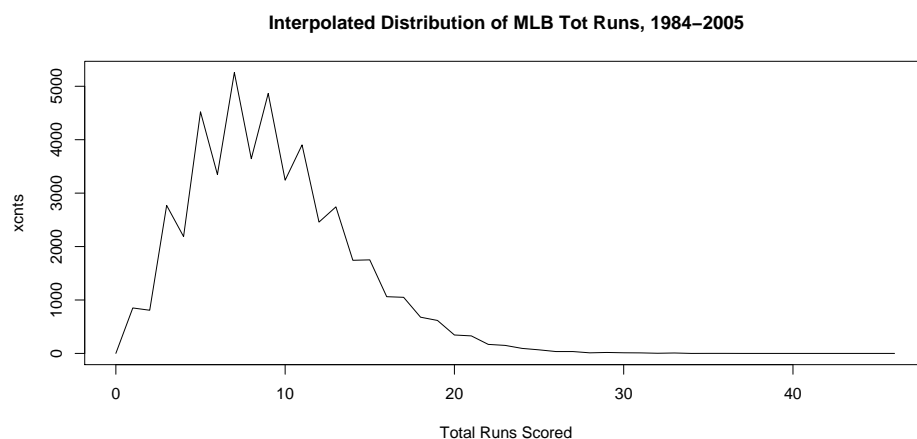
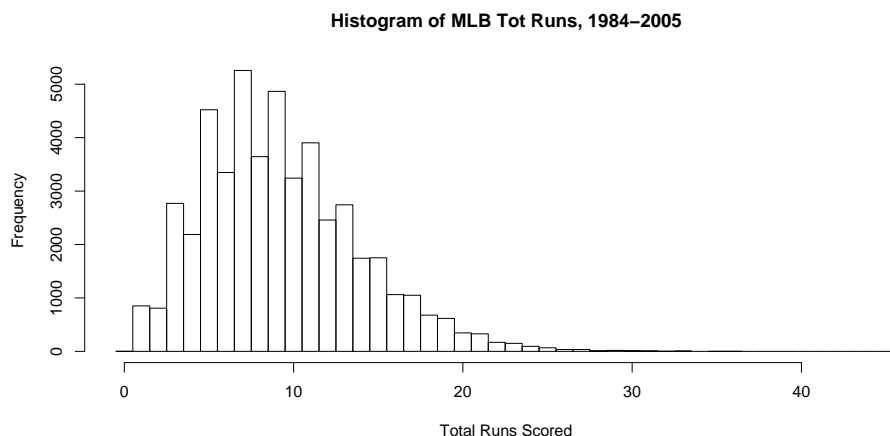
Dave Zes

May 14, 2006

1 Introduction

A baseball (or football, hockey, cricket...) over-under estimate is a “median” value; it is that value which evenly divides the frequency population of total run outcomes which has been augured for some game. That is, the estimate is indifferent to the torque of outliers. From this follows another, perhaps more arcane fact: a frequentist approach to optimizing some system of predicting total runs will rely on the L^1 metric, the absolute difference between the predicted and actual value. By definition, the median value from a list of natural numbers will be some natural number. In our context, this is very dissatisfying, as we are dealing with a coarse frequency, and this integer middle value is not very descriptive of the frequency of total runs near the middle. Put another way, suppose we have for some game a total runs best guess of 9.21, then *we have violated the definition of median value*. What is needed is a modified, Real-valued median value, one optimally descriptive of near-central occupancy. Passing our discrete distribution into a continuous one is hindered slightly because, in baseball, ties are not permitted, so even totals are disfavored. The following method is a facile solution whereby the discrete histogram is interpolated linearly. The resulting middle value is called a “zedian”, and is denoted by ζ .

2 Study



To construct our interpolated distribution, for each x (Tot Runs) we connect the corresponding count with a straight line. The line function connecting (x_0, y_0) to (x_1, y_1) will be given by $a_1x + b_1$. Let us call the Area under our interpolated frequency distribution A_Σ .

Let the particular gap upon which A_Σ 's area will be evenly divided be called I , and defined on the domain by x_1, x_2 , to which y_1, y_2 correspond, and let A_I be its area.

Let A_W be the area west of I , A_E the area east of I .

Let the function over I be called $a_Ix + b_I$. (Do not confuse the I subscript with the number 1 subscripts above).

Let ζ be our “zedian.”

So, we seek that

$$A_W + \int_{x_1}^{\zeta} a_I s + b_I = A_E + \int_{\zeta}^{x_2} a_I s + b_I$$

The integral on the RHS is equal to A_I minus the integral on the LHS, so we have

$$A_W + 2 \int_{x_1}^{\zeta} a_I s + b_I = A_E + A_I$$

$$\begin{aligned} \Leftrightarrow 2 \left(\frac{a_I}{2} \zeta^2 + b_I \zeta - \frac{a_I}{2} x_1^2 - b_I x_1 \right) + A_W - A_E - A_I &= 0 \\ &= a_I \zeta^2 + 2b_I \zeta + (A_W - A_E - A_I - a_I x_1^2 - 2b_I x_1) \end{aligned}$$

So, when we have $A = a_I$, $B = 2b_I$, $C = A_W - A_E - A_I - a_I x_1^2 - 2b_I x_1$, we find our zedian by

$$\zeta = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

So, for example, looking at the whole of MLB baseball games from 1984-2005 inclusively, we have

TotR	Count	a	b	A		
0	0					
1	851	851	0	425.5	425.5	48787
2	808	-43	894	829.5	1255	48361.5
3	2770	1962	-3116	1789	3044	47532
4	2187	-583	4519	2478.5	5522.5	45743
5	4522	2335	-7153	3354.5	8877	43264.5
6	3348	-1174	10392	3935	12812	39910
7	5257	1909	-8106	4302.5	17114.5	35975
8	3644	-1613	16548	4450.5	21565	31672.5
9	4866	1222	-6132	4255	25820	27222
10	3242	-1624	19482	4054	29874	22967
11	3903	661	-3368	3572.5	33446.5	18913
12	2458	-1445	19798	3180.5	36627	15340.5
13	2743	285	-962	2600.5	39227.5	12160
14	1743	-1000	15743	2243	41470.5	9559.5
15	1751	8	1631	1747	43217.5	7316.5
16	1062	-689	12086	1406.5	44624	5569.5
17	1050	-12	1254	1056	45680	4163
18	678	-372	7374	864	46544	3107
19	617	-61	1776	647.5	47191.5	2243
20	345	-272	5785	481	47672.5	1595.5
21	328	-17	685	336.5	48009	1114.5
22	169	-159	3667	248.5	48257.5	778
23	150	-19	587	159.5	48417	529.5
24	94	-56	1438	122	48539	370
25	67	-27	742	80.5	48619.5	248
26	35	-32	867	51	48670.5	167.5
27	35	0	35	35	48705.5	116.5
28	11	-24	683	23	48728.5	81.5
29	18	7	-185	14.5	48743	58.5
30	12	-6	192	15	48758	44
31	10	-2	72	11	48769	29
32	3	-7	227	6.5	48775.5	18
33	8	5	-157	5.5	48781	11.5
34	0	-8	272	4	48785	6
35	1	1	-34	0.5	48785.5	2
36	1	0	1	1	48786.5	1.5
37	0	-1	37	0.5	48787	0.5
38	0	0	0	0	48787	0
39	0	0	0	0	48787	0
40	0	0	0	0	48787	0

where we readily find our values of interest:

$$A_W = 21565$$

$$A_E = 22967$$

$$A_I = 4255$$

$$a_I = 1222$$

$$b_I = -6132$$

$$x_1 = 8$$

And, so, $\zeta = 8.6952$.