By the Numbers

Review

Academic Research: Intangibles and Salaries

Charlie Pavitt

The author describes two academic studies, one trying to determine which players bring the most "intangibles" to the team, and the other testing whether high-, middle-, or low-salary players give the best bang for the buck.

This is one of a series of reviews of sabermetric articles published in academic journals. It is part of a project of mine to collect and catalog sabermetric research, and I would appreciate learning of and receiving copies of any studies of which I am unaware. Please visit the Statistical Baseball Research Bibliography at www.udel.edu/communication/pavitt/biblioexplan.htm. Use it for your research, and let me know what is missing.

Scott Berry, <u>Winning Isn't Everything; It's the Only Thing</u>, Chance, Volume 18 Number 3, Summer 2005

Scott Berry, <u>Budgets and Baseball Concave or Convex: Winning the Salary Game</u>, Chance, Volume 19 Number 1, Winter 2006

I had the opportunity to take my summer tour of libraries that subscribe to some of the more esoteric journals, allowing me to catch up on this past year's offerings in Chance. Happily, Scott Berry is still writing his "A Statistician Reads the Sports Pages" column in a majority of issues, and his research remains as much fun to read as I assume he had in conducting it.

In *Winning Isn't Everything*, Berry takes on the notion that there are some players who are "winners in a way that cannot be captured in statistics" (page 53). Berry doesn't buy it, but analyses it anyway through an iterative process in which players from 1901 through 2004 were credited with the annual winning percentage of every team played for after adjustment for every other player on the team. Unless I missed something, it appears that his method does not control for game appearances. Anyway, the greatest winner of all time was, of all people, Dennis Cook,

whose many teams consistently performed better the seasons he was on them than seasons before and after, to the tune of 4.03 extra wins per year. (I seem to recall reading something analogous on Cook in the past, although I can't remember where; if anyone remembers please let me know.)

Actually, there were some very well known names in the top 50 (Collins, Cochrane, Baker, Frank Robinson, Gehrig, Berra, Mantle, Reese, Doerr) but also some comparable to Cook (Rich Bordi, Rocky Nelson) and some stars that rated very low (Dale Murphy, Chuck Klein). Berry concludes (also page 56) that "despite not believing in the model wholeheartedly, I find the results interesting, and not completely silly, as I originally thought possible." However, the names I mentioned should provide at least part of the reason for these results; most of the players that ranked highly were lucky enough to join a team about when that team registered a significant improvement and leave about when the team collapsed (and vice versa for Murphy and Klein). So, for example, Berra (#17) and Mantle (#50) were joined on the top 50 list by Rizzuto (#19) and Skowron (#25). This was easier to achieve early on, and 39 of the top 50 played all or almost all of their games before 1950. Berry basically realizes this, noting there were "long streaks of domination for teams and less player movement in the first 50 years of the 104 seasons" he studied (page 56).

In *Budgets and Baseball* Berry addressed the issue that the Baseball Prospectus concentrates on: efficient use of available funds in player acquisition and retention. He computed a runs-created type measure normalized to league average for the same number of plate appearances for batters, and analogous earned-runs-saved measure for pitchers normalized for same number of innings pitched, and plotted both against annual salary. Particularly for batters but to some extent for pitchers also, for 1995 through 2004 he found a curvilinear relationship such that high-salaried players were relatively underpaid given their performance when compared to medium-salaried players (a convex production function), implying that teams use their budget more efficiently by acquiring some high-salaried and some low-salaried players rather than by signing all mid-salaried. The reason: middle-range players were not sufficiently more productive than low-range to warrant as much extra money as they were getting, whereas high-range players were as a group markedly better than middle. This, of course, is exactly what the *Baseball Prospectus* folks have been arguing ever since I've been reading their work.

Charlie Pavitt, chazzq@udel.edu ♦

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis - that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any - I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics - confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Ben Baumer	bbaumer@nymets.com	Statistics
Jim Box	jim.box@duke.edu	Statistics
Keith Carlson	kcsqrd@charter.net	General
Dan Evans	devans@seattlemariners.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	Han60Man@aol.com	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Bill Johnson	firebee02@hotmail.com	Statistics
Mark E. Johnson	maejohns@yahoo.com	General
David Kaplan	dkaplan@UDel.Edu	Statistics (regression)
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
Chris Long	clong@padres.com	Statistics
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
John Stryker	john.stryker@gmail.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

The OBP/SLG Ratio: What Does History Say?

Victor Wang

The correct relative contribution of OBP and SLG in "OPS-type" statistics has been the subject of some discussion recently. Here, the author checks historical team run records to see which ratio gives the closest correlation to runs scored.

There has been much debate and research in the past issues of *By the Numbers* about how much more valuable OBP is to SLG. Values ranging from 1.5 to even 3 have been brought up. No one, however, has actually compared the various values of OBP to SLG to the runs scored of a team. To solve this, I took the OPS and runs scored from every team since 1960. I then adjusted the OPS using the different suggested coefficients for OBP. The adjusted OPS I used were OBP weighted by 1.5, OBP weighted by 1.8, OBP weighted by 1.9, OBP weighted by 2.0, and regular OPS. These OBP weights have all been suggested in one place or another.

The results:

OBP Coefficient	Correlation to R
1	0.8386
1.5	0.8394
1.8	0.8408
1.9	0.8407
2	0.8405

We can see that normal OPS has the worst correlation when compared to each adjusted OPS. The correlation keeps improving until the coefficient reaches 1.8, when it starts to decline but still has a higher correlation than with a coefficient of 1. However, the correlations remain very close to each other.

The data shows that the best coefficient to use when weighting OBP is 1.8. This was also confirmed by Tom Tango though I am unaware where his study is located. In fact, *The Hardball Times* currently uses a stat called "GPA," which adjusts OPS using a 1.8 coefficient for OBP and divides by 4 to make the stat on a similar scale to batting average. If anyone is interested in the complete set of data that contains all teams from 1960 and there adjusted OPS with runs scored, please contact me at the e-mail address below.

Victor Wang, atsbuy@yahoo.com ♦

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work are all welcome.

Articles should be submitted in electronic form, either by e-mail or on CD. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (i.e., informal peer review).

If your submission discusses a previous BTN article, the author of that article may be asked to reply briefly in the same issue in which your letter or article appears.

I usually edit for spelling and grammar. If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does.

I will acknowledge all articles upon receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

Send submissions to:
Phil Birnbaum
88 Westpointe Cres., Nepean, ON, Canada, K2G 5Y8
birnbaum@sympatico.ca

Are Career Home Run Trends Changing in the Steroid Era?

Yoshio Muki and John Hanks

Historically, a power hitter's home run rate tended to decline as he got older. These days, it looks like many aging sluggers are increasing their power rate. Is that really happening? If so, is it statistically significant? And are steroids responsible? Here, the authors investigate.

Steroid abuse allegations and vehement denials abound in baseball. Almost everybody who's a fan of the sport has a gut feel that something is not quite right, and anyone who considers himself to be a baseball purist is troubled by the allegations and suspicions that chemistry is skewing the baseball statistics in favor of present day players.

Sabermetrics has yet to develop a strong case for the "steroid effect" (Los Angeles Times, May 28, 2006, Michael A. Hiltzik), even though sabermetricians have been poring over the data for several years.

In this article, we demonstrate that the group of the top 100 all-time home run producers is actually comprised of two distinct populations. We don't claim that steroids are responsible. (In fact, to remove any hint of making an allegation, we'll refer to the pre-steroid era as the "Classical Era", and the current steroid era as "Current Era" or "Present Day Era".) We don't even claim to know all of the existing sabermetric research on the topic. What we do claim is that there is a statistically significant difference in the career patterns of the sluggers in the classical era and those of the current era. In this article we present a different way of distinguishing the two populations of major league sluggers.

The hunch

All investigations begin with a hunch. We didn't look at total home run output or home run to at-bat ratio (HR/AB); instead we looked at the *trends* in the HR/AB ratio over the course of each career. Our hunch was that classical era hitters peak early in their careers and maintain a steady level or demonstrate a steady decline; whereas the current era players demonstrate a steady improvement as they age and peak later in their careers. For our sample, we chose the top 100 MLB home run hitters,

Figure 1 – The 500 Home Run Club				
Player	**Year at 500**	**HR**	**AB**	**HR/AB**
Babe Ruth	1929	714	8399	0.085
Jimmie Fox	1940	534	8134	0.066
Mel Ott	1945	511	9456	0.054
Ted Williams	1960	521	7706	0.067
Willie Mays	1965	660	10881	0.060
Mickey Mantle	1967	536	8102	0.066
Eddie Mathews	1967	512	8537	0.059
Hank Aaron	1968	755	12364	0.061
Ernie Banks	1970	512	9421	0.054
Frank Robinson	1971	586	10006	0.059
Harmon Killebrew	1971	573	8147	0.070
Willie McCovey	1978	521	8197	0.064
Reggie Jackson	1984	563	9864	0.057
Mike Schmidt	1987	548	8352	0.066
Eddie Murray	1996	504	11326	0.045
Mark McGwire	1999	583	6187	0.094
Barry Bonds	2001	734	9507	0.077
Sammy Sosa	2003	588	8408	0.067
Rafael Palmeiro	2003	569	10472	0.054
Ken Griffey Jr.	2004	563	8536	0.068

starting with Hank Aaron (755) to George Brett (317). The sample starts chronologically with Babe Ruth (retired 1935) and includes the 20 active players (as of 2006 season). The difference of this trend between the classical era batters and the current crop of batters will be shown to be both measurable and statistically significant.

Data collection and math model

Data was gathered for each of the 100 batters and tabulated using HR/AB stats for each year in their career. As a sub-sample, the data for the 500 Home Run Club are shown in Figure 1.

For each batter, we chose the best-fit straight line for their career HR/AB trend. A comparison of Frank Robinson to Barry Bonds (Figure 2) clearly shows a difference. The zigs and zags of the data are smoothed out using a curve fit. In each figure, the "best-fit" curve is superimposed over the data. It turns out that Robinson had a slightly downward trend for his career, while Bonds had a strong upward trend.


Figure 2 – HR/AB trend lines for Frank Robinson (left) and Barry Bonds (right)
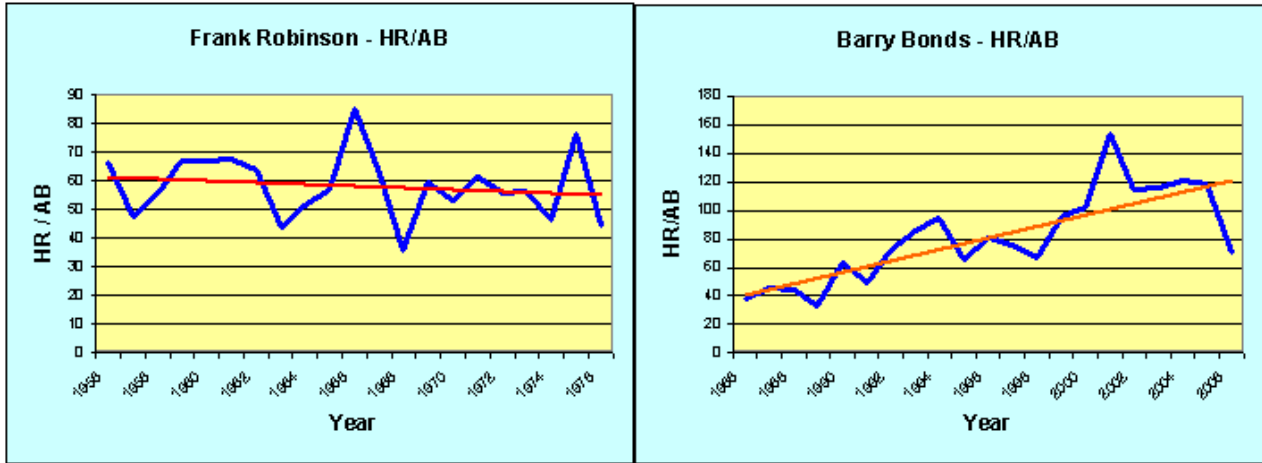
Slopes

The mathematical expression for the curve fit has the form of $y = Mx + b$. Those of you who still remember 8th grade algebra will quickly recognize this as the "slope intercept form" where "M" represents the slope. We'll refer to "M" as the HR/AB trend line slope. The "M" is the crux of our discussion, so it's worthwhile to expand the discussion a little bit. The slope is sometimes referred to as "rise over run", or how much the output (y, or in our case HR/AB) increases with input (x, or in our case, years in the majors). A positive value of M indicates that the player's HR/AB increased as the years in the majors progressed, in other words, as the player got older, his home run efficiency increased. A negative value of M indicates that the players HR/AB decreased with years in the majors, in other words,

Figure 3 – Career slopes for HR/AB, in order of career HR
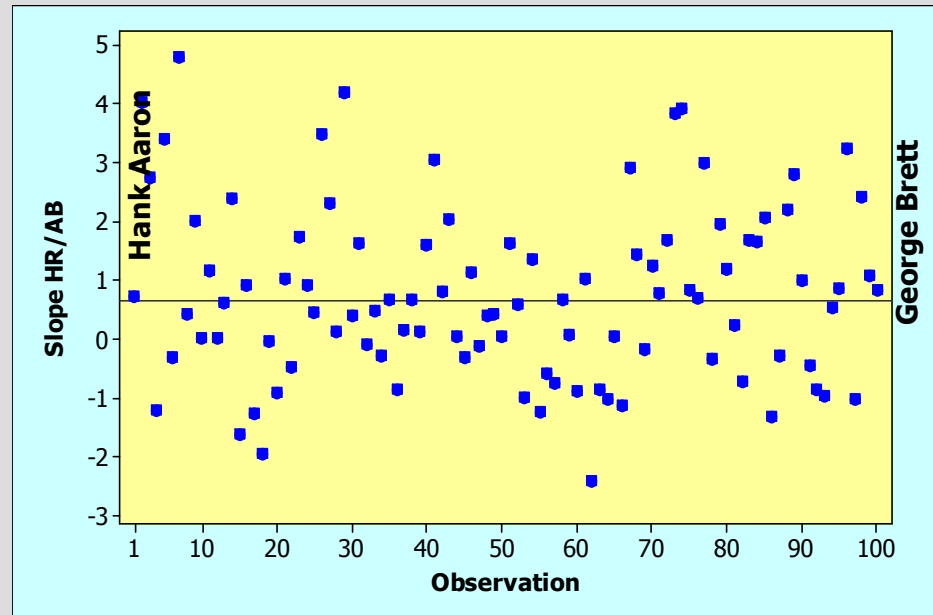
as he got older, it took him more trips to the plate to get a home run.

We developed a list of the "M" or HR/AB trend line slope values for each of the 100 players in this same fashion.

We now present the slopes for each of the hitters in our study. In Figure 3, we present the slope values for all 100 hitters in terms of ranking, starting with Hank Aaron to George Brett. As can be seen, there is no strong recognizable pattern in the series, therefore it can be concluded that the pattern is random. In Figure 4, the same data are presented, but are sorted in order of when the player retired. What is apparent in Figure 4 is the tendency of the current-era players to consistently have positive slopes.

The pattern becomes more distinct if we isolate Figure 4 to only include the 500 HR Club players. This is shown in Figure 5. In that context, the period between Eddie Murray and Mark McGwire indicates the watershed between classical-era and present-day era players.

In Figures 3 and 4, the trends are erratic, and ordering of players that retire in the same year can have an effect on the apparent instability of the curves. To smooth out the curves and to remove the dependency of the retirement of players in any given year, we grouped the players in sets of ten (rational sub-groups). The average values from each group of ten are plotted in Figures 6 and 7 for ranking and retirement years, respectively.

In Figure 6, again, there in no discernable pattern, and in control terms, it can be said that the process is stable. The upper and lower control limits (UCL and LCL) define the 3σ boundary to indicate statistical stability.

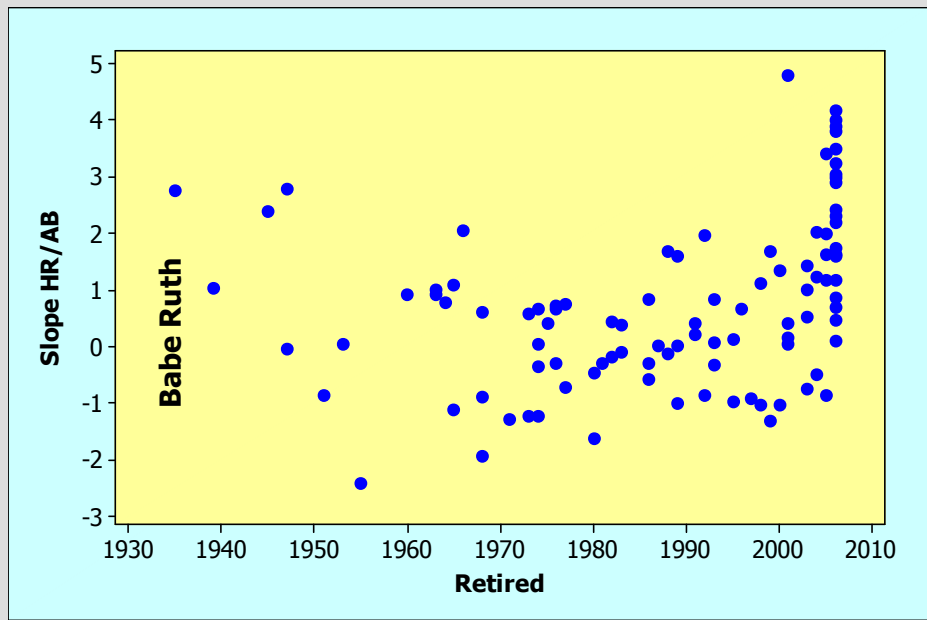Figure 4 – Career Slopes for HR/AB, in order of retirement year

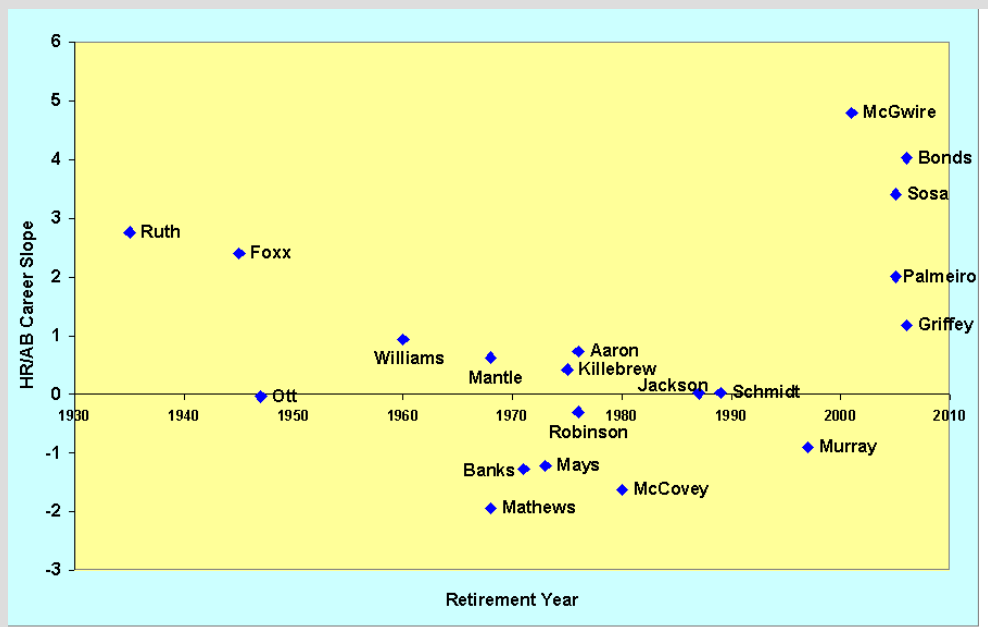Figure 5 – Career Slopes for HR/AB, in order of retirement year, players with 500+ HR

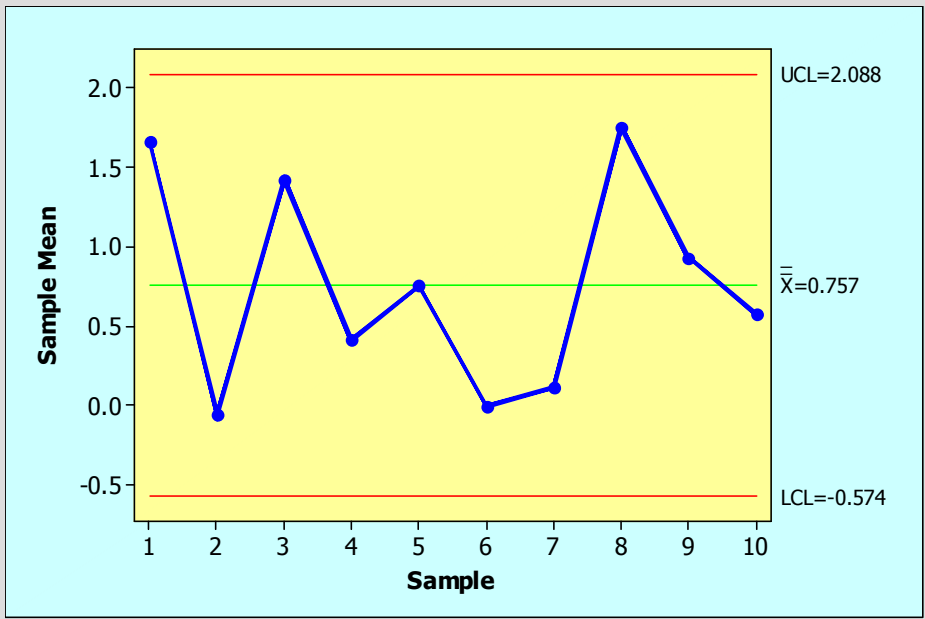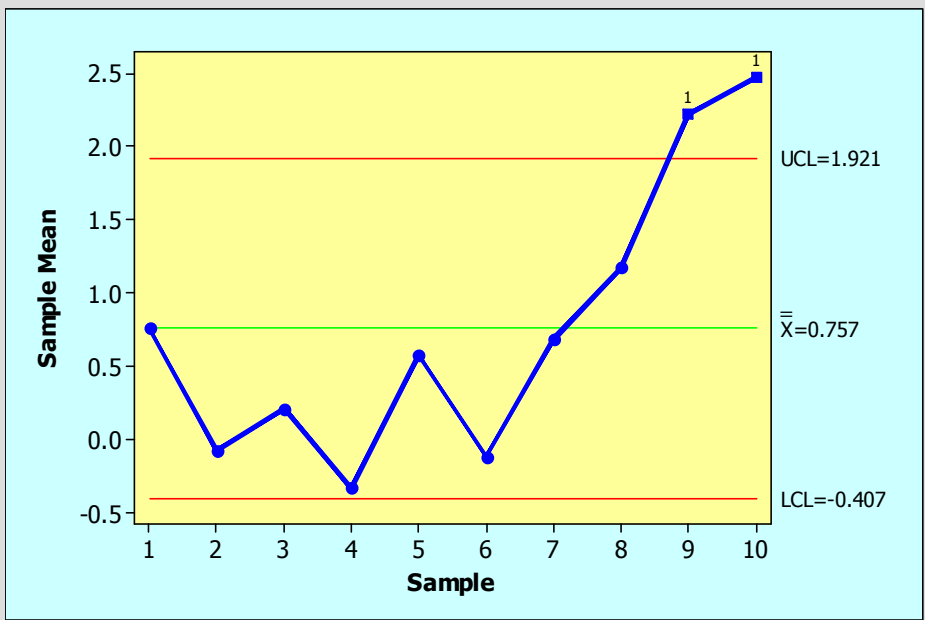Figure 6 – Average HR/AB Slope of subgroups, in order of career HR

UCL=2.088
$\bar{\bar{X}}$=0.757
LCL=-0.574

Figure 7 – Average HR/AB Slope of subgroups, in order of retirement year

UCL=1.921
$\bar{\bar{X}}$=0.757
LCL=-0.407

The same data is presented in Figure 7, but this time in order of retirement year. The noticeable feature of the curve in Figure 7 is the strong upward tendency at the far right (current era). In fact, the tendency is sufficiently strong to violate the 3σ upper control limit which triggers a "Special Cause" anomaly; this occurs twice. Special Cause is a condition that cannot be attributed to a normal random process.

Normality testing

In a natural or random process, a population will show a normal distribution, typically referred to as the "bell curve".

Figure 8 shows the distribution of career HR/AB slopes for all 100 top home run producers, along with the normal curve that's the best fit to the data. Although a best fitting normal curve can always be drawn, there is a test, the "Anderson-Darling Normality Test," that is used to check whether the fit is truly close enough to be able to conclude that the data is actually normal.

Figure 8 – Best-fit normal curve for career slope, top 100 home run hitters
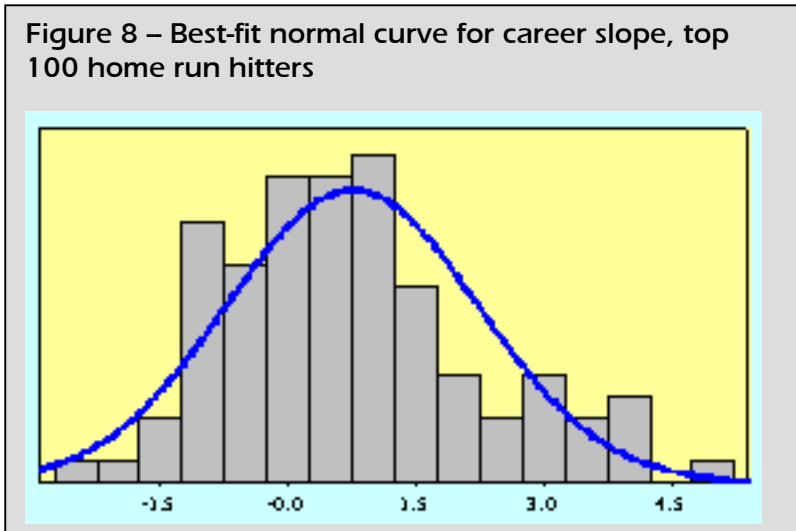
In this case, running this test on the sample gives a p-value of only 0.044. This indicates that this distribution is not a naturally occurring random process. In fact, a close examination of the distribution indicates a bimodal nature; that is, there may be two distinct groups within this sample.

Again, we chose to break up the sample into two groups -- those that retired before 2001, versus those who were still active in 2001 or later. The distributions are shown in Figures 9 and 10.[1]

Running the Anderson-Darling test on the two samples lets us conclude that the classical era players (p=0.688) and the modern day players (p=0.751) show normal distributions when treated as separate samples. This indicates that within each sample, the variation is driven by natural or random causes. The two histograms (distributions) are presented together in Figure 11.

Figure 9 – Best-fit normal curve for career slope, top 100 home run hitters, retired before 2001

---
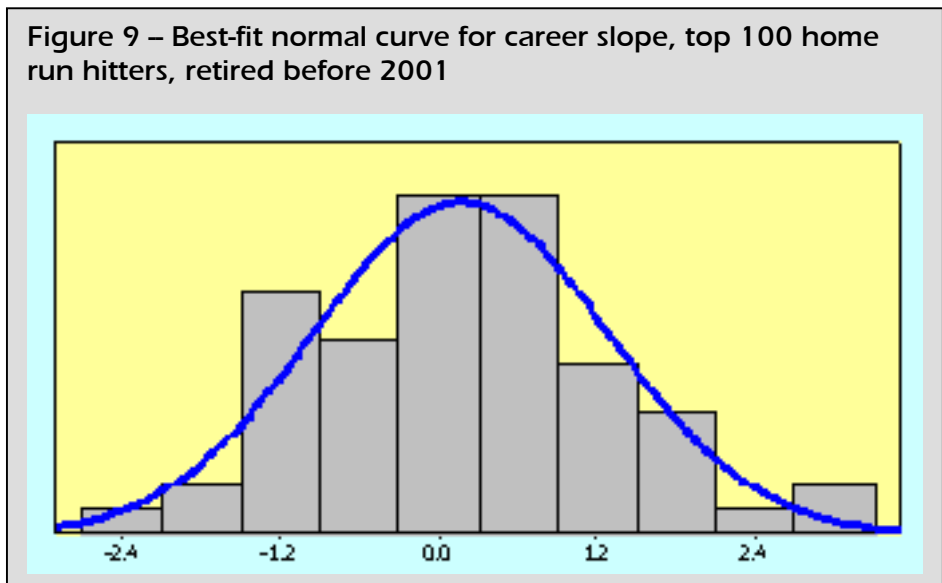
[1] Further statistical properties of the distributions in figures 9 and 10 are available on request from the authors.

Figure 10 – Best-fit normal curve for career slope, top 100 home run hitters, still active in or after 2001
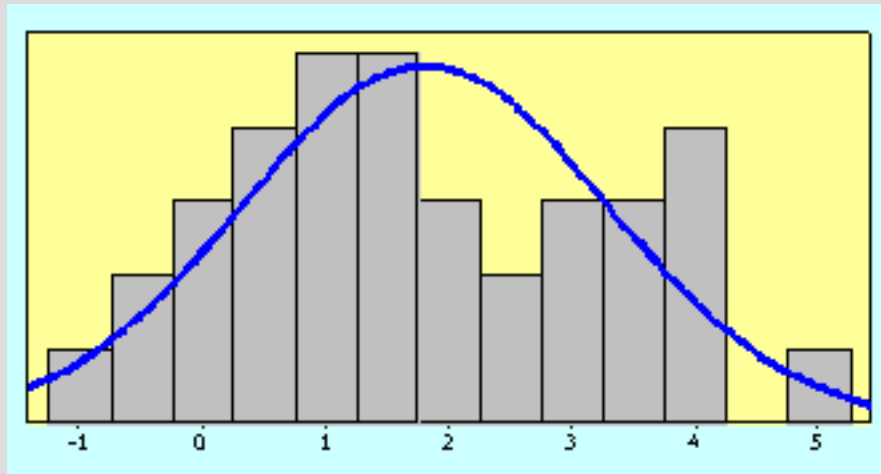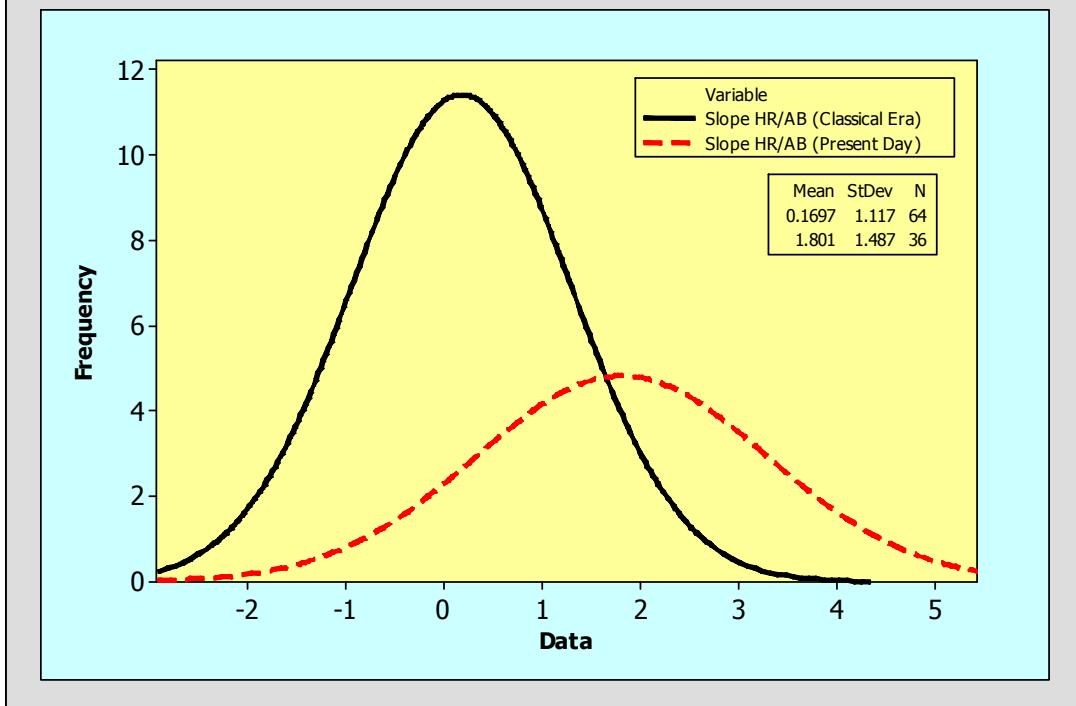

Figure 11 – Smoothed curves for the two subgroups within the Top 100 home run hitters

We've shown that there is a visible difference between the two populations, and that the Anderson-Darling normality test shows both subpopulations are normal. A summary of the two distributions is given below:

Value	Classical Era	Present Day
Samples	64	36
Mean Value	0.170	1.802
Standard Deviation	1.12	1.49
p-value (A-D Normality test)	0.688	0.751

What's the difference?

We started down this path to demonstrate that the HR/AB trends of the two populations were statistically different. The statistical tool that's used to show this is called the Two-Sample-T test. (The Two-Sample-T test is derived from a Student-T test, developed by W. S. Gossett (1876-1937), who was trying to solve quality problems while working at a brewery. See how beer, baseball and statistics all go together?)

The Two-Sample-T test is considered one of the best ways to determine if the averages of two populations are statistically different. Obviously from Figure 11, there is a visible difference between the peaks, but is it statistically significant? The Two-Sample-T test will answer this question.

We submit the two sample populations into the Two-Sample-T test and find the following output:

```
                   N   Mean   StDev   SE Mean
Classical Era     64   0.17   1.12     0.14
Modern Era        36   1.80   1.49     0.25


Estimate for difference between the two means:  -1.63174

99% CI for difference:  (-2.38981, -0.87368)
T-test of difference = 0 (vs not =): p-value = 0.000
```

Through the two-sample t-test, we are trying to determine if two distributions with standard deviations of 1.12 and 1.49 are sufficient to distinguish a difference in means of 1.63. The p-value is an indication of the level of overlap between the two average values. For this evaluation, P is 0.000, indicating there almost no potential for overlap; we can say this with over 99.95% confidence.

With a P value of 0.000 in our test, we reject the null hypothesis that states both samples are statistically the same, and we are left with the alternative hypothesis which says that the two samples are statistically different.[2]

More than just the era effect?

We have demonstrated that an "era effect" exists between classical-era and current-era players. Could there be a further effect?

In Figure 9, the distribution for the classical-era players show a classic bell curve distribution, whereas Figure 10, the distribution for the current-era players, although normal, seems almost bimodal. It may be possible to make the argument that there are actually three distinct populations within the all-time top 100 home run producers:

```
1)  Classical era   (no steroid influence)
2)  Current era     (no steroid influence)
3)  Current era     (with steroid influence)
```

---

[2] A number of active players (as of the close of the 2006 MLB season) are included in this study. There is a possibility that as they approach retirement, their future performance can skew the results. In a side study, the same analysis presented herein was performed excluding all active players, and no significant differences were found.

We have established that there is a statistically significant difference between the first population and the other two. Can we discriminate the population further?

We know that there are eight batters in the current-era sample that have either admitted to, tested positive for, or were implicated in the use of steroids or human growth hormones (HGH).[3]

If we separate those who are not associated with steroid/HGH abuse from those who have been tied to steroid abuse, we get Figures 12 and 13. Both show normality, with p-values 0.609 and 0.633, respectively.

An interesting feature jumps out of these two distributions. That is the hump in the 3.2 bin (ranges from 2.8 to 3.6) in Figure 12, and the gaping hole in the same bin in Figure 13. There are 6 players in the 3.2 bin in Figure 12; there are two players in the bins adjacent to 3.2 in the Figure 13.

We can postulate that some players in the "innocent" category (Figure 12) may have simply not been caught. What happens if we "borrow" some of the elevation in the 3.2 bin in Figure 12 and attribute them to those who have been associated with steroid/HGH abuse? Let's move three from the 3.2 bin in Figure 12 to the 3.2 bin in Figure 13.

In Figures 14 and 15, we show the summaries for the "redistributed" current-era players. (The vertical axes on Figures 14 and 15 are not to the same scale.)

The two adusted samples are closer to normal, with p-values of 0.647 and 0.700, respectively.

A two-sample t-test on these two groups, however, gives a p-value of 0.103, indicating that the two samples, even with the heavy-handed redistribution are not sufficiently different from one another to be statistically significant.

Perhaps steroid and HGH abuse is more wide-spread than suspected.

Perhaps there's some truth to BALCO founder Victor Conti's claim: "It's not cheating if everybody is doing it — and if you've got knowledge of what everyone is doing, and those are the real rules of the game, then you're not cheating." (20/20, ABC News, Friday, December 3, 2004)

Figure 12 – Best-fit normal curve for career slope, current-era hitters in the sample not implicated in steroid/HGH use

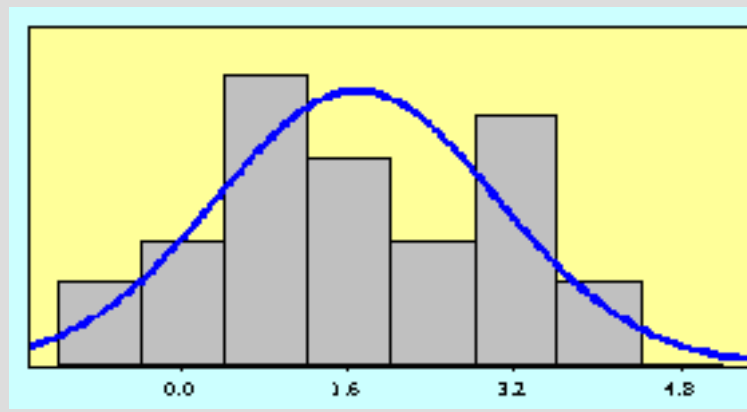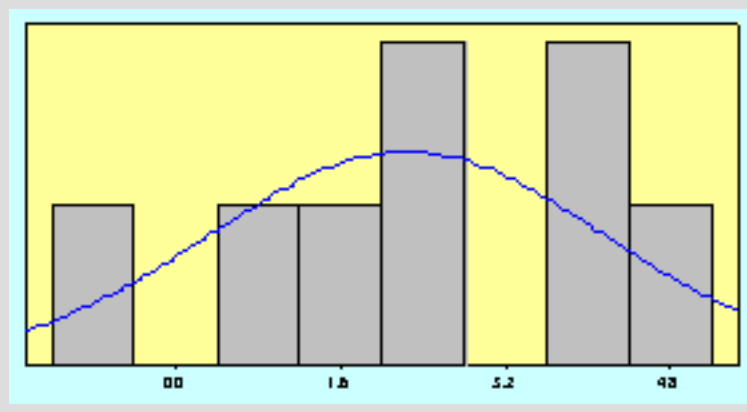Figure 13 – Best-fit normal curve for career slope, current-era hitters in the sample implicated in steroid/HGH use

No finger pointing here!

We've shown through statistical analysis that there exists a measurable and significant difference in the HR/AB ratio trend line slope between the classical-era and present-day era batters of the all-time top 100 home run producers. The difference is clearly measurable; the changeover from the classical era to the present-day era is very well-defined. The classical-era members of the all-time top 100 home run

_____

[3] See http://thesteroidera.blogspot.com/2006/08/list-of-steroid-hgh-users-in-baseball.html .

Figure 14 – Distribution of HR/AB career slopes of current-era players not associated with steroid abuse, with three players removed from the 3.2 bin (shown as the lighter block)
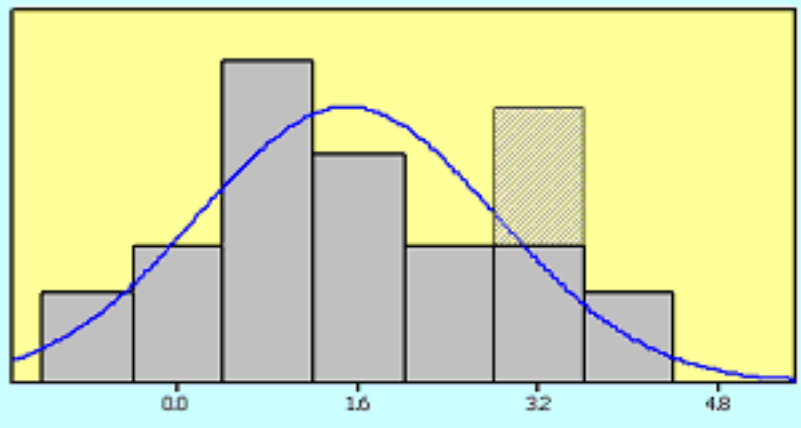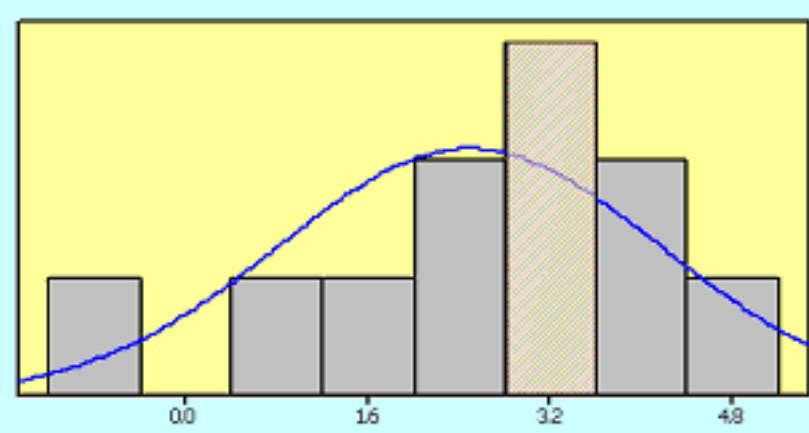

Figure 15 – Distribution of HR/AB career slopes of current-era players associated with steroid abuse, with three players added to the 3.2 bin (shown as the lighter block)

producers show steady or slight improvement or decline in their home run efficiency as their careers progressed, while the present-day era members of the top 100 home run producers show a much greater and measurable improvement in their home run efficiency as their careers progress. Hitting home runs is a very physical activity. The fact that the current sluggers can continue to improve while others in the past has generally maintained steady output or declined certainly raises eyebrows.

We've further shown that there is no statistically significant difference between the current-era players who have and have not been linked to steroid/HGH abuse.

The naysayers to the steroid argument always counter with other possible innocent reasons for the enhanced career trends:

- Diluted pitching
- Home-run friendly ballparks
- Different bats
- Use of improved vitamin supplements and nutrition
- Different workout and preparation routine
- Stronger athletes, better training regimen.

And the short list of not-so-innocent reasons, but not as detrimental as steroid abuse:

- Corked bats
- Juiced baseballs

We are not discounting these influences, but these are all relatively gradual effects. The pitching talent didn't become diluted all of a sudden. League expansion is incremental. Not all ballparks were demolished overnight to be replaced by home-run friendly fields. And bats didn't all go from ash to maple all at once.

Certainly, there is an "era" effect. There are two factors that may have contributed to the sudden change in the HR/AB career trends. The first is the institution of free agency in the mid-1970s; the associated astronomical salaries certainly had a hand in enticing the major league players to maintain or enhance their performance as their careers start to wane.

The other factor is the introduction of performance enhancing drugs and HGH supplements. Free agency may have provided a great motivator to stretch out the productivity of a player in later years, but it seems that steroids and other performance enhancing substances provided the opportunity.

The fact that the historical change (between Murray and McGwire, Figure 5) appears so well-defined suggests something more dramatic and sudden. And once the change occurred, the trend is clearly consistent. The authors can see no overt and sudden changes in pitching, ballparks, training techniques or other influences that can account for such a sudden and dramatic change.

The era effect is easy to prove. When did he retire? That's a matter of public record, but the steroid effect is not as easy to prove. But based on the absence of any difference within the current-era between the players linked to steroid abuse and those not implicated, the Era Effect may very well be the Steroid Effect. We leave that judgment to the reader, and to history.

Yoshio Muki, engineering consultant, and John Hanks, PE, CRE, are engineers with nearly 50 years of combined experience in the engineering field, mostly in aerospace and defense. Mr. Hanks is an avid baseball fan and historian. He is also an active coach for his sons' little league teams. The analysis presented in this article started out as a dinner table discussion about steroid usage in sports and quickly evolved into an exercise to engage his sons in statistics and critical thinking.

Yoshio Muki, YMuki@hrtextron.textron.com
John Hanks, Jhanks@hrtextron.textron.com ⬧