Clutch and Choke Hitters in Major League Baseball: Romantic Myth or Empirical Fact.

By: Elan Fuld

1st Released draft May 19th 2005

ABSTRACT

For a long time now, sabermetricians (baseball statisticians) believed that, at the Major League level, hitters who are truly clutch or choke performers[1] do not really exist, even while the many managers, players, and announcers insist that they do exist. We performed statistical analyses to determine if clutch and/or choke hitters exist in the major leagues using all regular season play during the 1974-1992 seasons. We use a Gompertz regression[2], a regression similar to logistic regression, but without symmetry around the point of inflection. Our independent variable is a probabilistically constructed measure of the *ex ante* importance of each plate appearance, and our dependent variable is based upon the value that OPS (on-base percentage plus slugging percentage) would assign to the outcome of the plate appearance.[3] We find that clutch/choke effects do exist at the major league level. We also find that a relatively small proportion of major leaguers exhibit such effects. According to our findings, only about 2% of major leaguers are in fact either clutch or choke hitters.[4] Furthermore, we find that clutch hitters are not always the players commonly thought to be clutch hitters. In some cases, they are the last people thought to be. For example, one of only two men to show up as clutch at the .001 significance level, regardless of how we treat reaching on an error in the regression, is a man best known for his choke performance defensively: Bill Buckner. The other one, Eddie Murray, comes as less of a surprise. The choke hitters tend to have short careers in which they don't see much playing time, a tendency that agrees with our intuition, since their choke performance diminishes their value to the team. We will follow our introduction with an in-depth discussion of our methodology. We will then discuss our results and conclusions, along with the potential implications for future sabermetric research.

[1] A note to those unfamiliar with the terms "clutch" and "choke." A clutch hitter is one who performs better when the situation is more critical and a choke hitter is one who under-performs when the situation is more critical.
[2] Also known as the complementary log-log regression.
[3] There are some slight modifications made to the way that OPS would be calculated. These will be explained and justified later. For formulas for these statistics see Appendix A.
[4] This only includes major leaguers who get at least 100 PA's in at least two different seasons, as will be explained later. The true number of clutch and choke hitters may be more or less than 2%. However, as will be shown later, finding 2% to be clutch or choke is sufficient to say with a high level of confidence that at least some clutch or choke hitters really do exist.

1. Introduction

Listening to a baseball game on most TV or radio stations, one might get the impression that the existence of clutch hitters is not only an unquestioned fact, but that any half-decent team has at least two or three such hitters in their lineup. Reading the sabermetric[5] literature, however, one gets a very different picture. Celebrated sabermetrician, Bill James, has long held that clutch hitters do not exist in Major League Baseball (abbr. MLB), attributing the widespread belief of their existence to self-aggrandizing players and hero worshipping fans, and calling clutch hitting a "bullshit dump" used by people to justify what they want to believe.[6] Recently, he has softened his stance. In an article, appearing in *The Baseball Research Journal*, he called the existence of clutch hitting "an open question," that cannot be shown to exist with the methods used in the past, but which may still be lurking in the shadows.[7] Even mainstream sports journalism, which has traditionally sided with the broadcasters, is beginning to regard clutch hitting as an open question. In their 2004 baseball preview issue, Sports Illustrated wrote an article that frames this controversy between the traditionalists (who say it exists) and the statisticians (who either say it doesn't exist or say it is unknown) as an open question, closing the article with quotes from two statistically-oriented General Managers (Theo Epstein of the Boston Red Sox and Paul DePodesta of the Los Angeles Dodgers) who are on the fence about whether clutch hitting exists. These GM's seem torn between the numbers and the intuition that the human element somehow factors in when the heat is on.[8] That is the tone that Sports Illustrated chooses to leave us with: Clutch hitting is an unresolved controversy, one of the great mysteries of the baseball universe. This study uses new techniques to investigate whether clutch hitting, and its opposite choke hitting, indeed exist. The study finds evidence for the existence of clutch hitters in MLB, but does not find evidence for the existence of choke hitters in MLB.

Strictly speaking, we are looking for clutch and choke hitters not clutch or choke hitting, but we will sometimes use clutch/choke hitting to mean clutch/choke hitters. The difference is important. In a sense, no one questions the existence of clutch/choke hitting: Carlton Fisk's

[5] Sabermetrics, is a term for the discipline of statistical analysis of baseball.
[6] Bill James, The New Bill James Historical Baseball Abstract The Free Press. New York. 2001. pp. 348-349

[7] Bill James, "Underestimating the Fog." *The Baseball Research Journal.* No. 33. The Society for American Baseball Research. Cleveland, OH. 2005. pg. 31-32

[8] Tom Verducci, "Does Clutch Hitting Truly Exist?" *Sports Illustrated.* Vol. 100 No. 14, April 5, 2004. pp 60-62.

homerun in the 12th inning of game 6 of the 1975 World Series was a clutch hit. It does not get any more important than that, and you can not do much better than a homerun; certainly, a homerun bests anybody's average performance in less important (or all) situations. Even though Fisk's homerun is indisputably clutch, it does not necessarily make Fisk a clutch hitter. Even without clutch hitters, baseball would still have its share of clutch hits. Was Fisk's performance in key situations somehow different than the rest of the season? Fisk did hit his fair share of homeruns during the regular season. It is fairly reasonable to think he was hitting the same way he had all year and that this homerun just happened to come at an opportune moment. After all, clutch hitters or no clutch hitters, in a tie game, someone has to hit the game winner. Games do not end in ties. So while we all know of clutch hits, this does not mean that there are necessarily clutch hitters, batters who are more likely to hit better in those key spots. To determine if there are such hitters requires more investigation.

The first step in investigating whether clutch and choke hitting exists is defining it. In a vague sense, a clutch hitter is a hitter who tends to perform better when things are more important, whereas a choke hitter would tend to perform worse when things are more important. Of course, we need to better define what we mean by "when things are more important." We could define a clutch/choke hitter as: One who performs better/worse in bigger games, one who performs better/worse in at more important points of the game (i.e. plate appearances that come at points that are more pivotal to determining who wins the game) or as one who performs better/worse at more important points of bigger games. Between these three methods there is only one that is a real option for the purposes of this study, and that is the second definition. Any definition that includes performing better in "bigger" games as part or all of its definition would create a serious flaw with any study that attempted to use it: You can't numerically measure how much more important one game is over another. Is a playoff game 3 times as important as a regular season game? Is a World Series game 1.4 times as important as an ALDS game? Is game 6 of the World Series 1.1 times as important as game 2? Is game 2 of the World Series 1.07 times as important as game 7 of the NLCS, or is game 7 of the NLCS, in fact more important? When the Red Sox play the Yankees is that 1.213 times as important as the Sox other regular season games? What about Cubs-Cardinals? What about September in a close pennant-race versus April? I could make up a bunch of numbers for all these and more, so that I could say how much more important I thought one game is over another, but I'd just be making up a

bunch of numbers. Making up arbitrary numbers on a whim is no way to conduct a study, it's a way to run a political campaign, and since I don't plan on running for public office any time soon, I think I'd better stick to the facts.

At this point, the reader may be wondering how it is any less arbitrary to try and define "more important points in the game." This can actually be approximated fairly well if we base our calculations on one simple axiom that I think everyone (except maybe new-agey parents of little leaguers) can agree with: The object of a baseball game is to win. If we freeze the game right before a particular plate appearance, we can see how important the plate appearance is to winning the game. This requires further elaboration. Using simulation techniques and league data, we can get a good approximation of the probability of either team winning the game at any given point in a given game. Going back to our frozen game we can find the importance of the plate-appearance through the use of a series of "what-if" probabilities. What is the probability of winning if the batter strikes out? What if he walks, or singles,[9] or doubles, etc. Now using league data we know how often players walk or strikeout, or double, etc. So we can now look at our frozen game and say: If this game were between two average teams and the man at the plate were an average hitter (for the league that year) on average, how much will he improve the chances of his teams winning the game over an automatic out.[10] This method of determining the importance of a situation to winning the game will be further developed and justified later in this paper. This is just a taste of things to come.

As mentioned above, my project is an investigation into whether "clutch" or "choke" hitters can be statistically shown to exist, and if so, how "clutch" or "choke" hitters can be identified. Much research has been on this topic, but no one has yet produced conclusive evidence of the existence of such a phenomenon. I intend to take a different tact than most who have attempted to answer this question. I will take a probabilistic approach, rather than arbitrarily defining the term "clutch." The two most commonly used arbitrary dividing line definitions are a situation where there is a runner on 2nd and/or 3rd base (the Runners In Scoring Position definition) or the game is in the seventh inning or later, and the batting team is either

[9] Here we must make runner advancement assumptions which will be discussed later.
[10] Although, there is no player in the majors who is truly an automatic out, the author of this paper (facing Major League pitching) would be, at the very least, an excellent approximation of this.

leading by a run, tied, or has the potential tying run on base, at bat, or on deck.[11] However, some recent work done by Jahn Hakes and Raymond Sauer does attempt to measure clutch hitting in a probabilistic fashion.[12] Sauer and Hakes use several methods; I will discuss the one that is most similar to the methodology I intend use to best illustrate how my methodology departs from what has been used in the past. They define a "clutch" situation as one where the impact of a players performance (if he hit a double then the impact of a double is used, if he made an out then the impact of an out is used etc.) on the probability of victory is more than twice the impact of the same performance in an average situation; this ends up including 10.9% of all plate appearances in their sample. They similarly define a low impact, or "relatively meaningless," situation as one in which the impact of the players performance was less than ¼ of what it would be in an "average" situation; this category encompassed 16% of the plate appearances in their sample. They then looked at whether players LWPGP (the average probability impact associated with their actual performance) was higher in "clutch" situations versus other situations and whether their LWPGP was higher in "meaningless" situations versus other situations.

I believe that there are two main problems with this metric that my clutch hitting metric does not have. The first problem stems from the fact that there are situations where the impact of a homerun, triple or double might be more than twice that of an average situation, but the impact of a single walk or strikeout are not. In these situations, Sauer and Hakes will count this Plate Appearance as clutch if the player happens to hit a double, but not if the player "chokes" and strikes out. Thus, whether a situation is clutch or not will often be defined by the performance of the player. However, since we are trying to measure the performance of the player in the clutch, as opposed to other situations, having a definition of clutch that is performance dependent is highly problematic.[13] The second problem with this metric is one that pervades nearly every attempt at measuring clutch performance that I have seen to date. Although it uses probabilistic measures to define clutch, it still arbitrarily draws a line in the sand and defines everything to one side as clutch and everything to the other side as not clutch. Where this line is drawn can potentially have a sizeable effect on the outcome. The arbitrary divider problem is one that

[11] The Close & Late definition. Also known as the Late Inning Pressure Situation definition. These two nomenclatures are interchangeable.

[12] Jahn H. Hakes and Raymond D. Sauer, "Are Players Paid for 'clutch' performance?" John E. Walker Dept. of Economics, Clemson University. Preliminary Draft. June 30, 2003.

[13] This same problem exists with the other probabilistic methods Sauer and Hakes use.

looms much larger over many of the popular methods of looking at clutch hitting. This involves defining some set of circumstances under which something is clutch, such as runners-in-scoring-position or late inning pressure situation[14] This type of method then looks at a player and compares his batting average, On-base percentage, or some other statistic in these "clutch" situations to his value of that statistic in other situations. In this case, not only are the circumstances an arbitrary divider, but they do not even divide strictly along lines of importance; some non-LIPS situations are more important than some LIPS situations!

Some variation of one or both of these problems is extant in every study on clutch hitting to date that I have seen. To clarify my intent in the preceding paragraphs was not to rail against Sauer and Hakes, but to use their work to illustrate the difficulties with measuring clutch and choke hitting. As one of the better studies I have seen on this topic, I thought it seemed a good one to use to illustrate the problems that pervaded such studies in general.

The problem with defining clutch or choke through a statistic that mixes performance (outcome) and the importance of the pre-existing situation merits further discussion, because addressing this problem by using separate measures for performance and importance is a critical distinguishing characteristic of my study. If a player hits a double, such an index only measures how critical it was for him to hit a double, not how critical it would have been had he done something else. The flaw with such a system can best be demonstrated through an example. The situation is tie game, bottom of the 9th bases empty two outs. If the batter hits a homerun he has improved his team's chances of winning greatly. Before, his team had a slightly better than 50-50 chance.[15] Now his team just won (with probability 1) so he made nearly a 50% difference in probability of winning. It would seem he came through big-time in the clutch. Let's say instead he got out. Before his team had a slightly better than 50-50 chance of winning; now they have about a 50-50 chance of winning. His out had a moderate to small impact. If his hitting a homerun counts as coming through "big-time" in the clutch, we'd think a strikeout should be choking "big-time" in the clutch. But instead 2 outs, bottom of the 9th, bases empty, tie game, appears as being of crucial importance when he comes through in the "clutch" with a Homerun, and as being of moderately small importance when he makes an out. So with this type of

[14] Late inning pressure situation is a term that is defined by a clear-cut set of circumstances: In or after the 7th inning where the game is tied, a one-run game, or the tying run is on-base, at the plate, or in the on-deck circle. (This definition was obtained from www.enlexica.com)

[15] Assuming teams are evenly matched and home field provides little to no advantage so that extra innings is practically a coin-toss

method, our definition of what is a "clutch" situation is largely dependent on whether the batter comes through in the "clutch" situation. Therefore, in many cases, coming through (or choking) in the "clutch" situation is what makes it a "clutch" situation. But shouldn't the situation (in our example: tie game, bottom of the 9th, 2 outs, bases empty) have some specific level of importance attached to it? The underlying problem that causes this paradox is that such an index fails to measure the inherent importance of a situation and the outcome a player produces in that situation separately. Hence, importance (of situation) and performance get muddled together[16] into one metric making it impossible to truly see if as importance increases, a player's performance rises.[17] Indeed, this doesn't even give us separate numbers for importance and performance to relate to each other, just one number that is dependent upon some combination of the two.

My new method for measuring clutch hitting effects solves these problems handily, and provides a more effective probabilistic method for measuring clutch performance. I devised a performance independent "importance index" to measure the inherent importance of a situation, not taking into account the particular outcome that follows (See Appendix B for the details of the calculation of this index). This index uses the impact on the probability of winning which hypothetical performances (e.g. HR, BB, 1B, 2B, 3B) in the current plate appearance would have over that of a hypothetical strikeout (the operational definition of strikeout being an out with no runner advancement.) It then weights the impact of each hypothetical performance according to its league-wide relative frequency in the present season. Thus, whether the player ends up striking out or hitting one out of the park, the situation he batted in is assigned the same value on the importance index because the importance index measures only the importance of the situation, and doesn't take into account the performance of the player in that situation. After all, this index is measuring the only importance of the opportunity, not how well it was capitalized on.

Rather than arbitrarily choosing a value of the importance index and making that value the boundary between a clutch and a non-clutch situation, I will create a scatter plot of every plate appearance a batter has had. The X coordinate will be the value of the importance index for that Plate appearance and the Y-value will be the OPS entry for that plate appearance (OPS is

[16] the statistical term for this is confounding.
[17] Or in the case of a choke hitter falls.

on-base percentage plus slugging percentage. Thus, the OPS entry can take on 6 values : out=0, walk/Hit by pitch=1,single=2, double=3, triple=4, Homerun=5.) OPS is a simple measure of productivity that is very closely linked to run production;[18] For these reasons, I use it as the measure of performance used in my analysis.[19] I will then run a regression on the scatter-plot. Since the X-axis is a measure of importance of situation only, and the Y-axis is a measure of the performance of the batter only[20] only, if the slope coefficient is positive and significantly different from 0 (using significance tests for the regression coefficients using a particular α)[21], then this analysis would indicate that this player is "clutch;" he tends to do better in more important plate appearances than in less important ones. Similarly, if the slope is significantly negative then this analysis would indicate that this player is a "choke" player; he tends to do worse in more crucial plate appearances than in less critical ones. However, in a significance test using α, $100\alpha\%$ of all players will appear to be either "clutch" or "choke," even if in reality none of them are "clutch" or "choke." So the real test of whether such effects exist will be to see if the same players who are found to be "clutch" in one year tend to be "clutch" in subsequent years as well, and if the same players who are found to be "choke" in one year tend to be "choke" in subsequent years. If enough players are found to be "clutch" in a majority of their seasons (or "choke" in a majority of their seasons) then we will be able to reject the null hypothesis that they are all really pressure neutral hitters who are showing up "clutch" or "choke" by mere chance variation. Thus, this method will serve as both the tool for detecting the existence of "clutch" or "choke" effects in MLB, and as the tool for determining which players exhibit such effects.

2. The Data

My data, from Retrosheet.org, consists of every regular season game played in both the National and American Leagues for all seasons 1974-1992, including the strike-shortened 1981

---

[18] See Jim Albert and Jay Bennett, Curveball: Baseball, Statistics, and the Role of Chance in the Game Copernicus Books. New York. 2001 pg. 230. To get a fuller development of this topic, read the entire chapter (chapter 8 beginning pg. 207)

[19] I actually use a modified version of OPS, which will be discussed later in the paper.

[20] Performance may not be precise enough. It is really a (rough) measure of how helpful what the player just did is in general.

[21] The determination of an appropriate value of α is trickier than it seems and is discussed fully later in the paper.

season. The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at 20 Sunset Rd., Newark, DE 19711.[22] All players in Retrosheet's database are identified using the Retrosheet player ID consisting of the first 4 letters of the player's last name, the first letter of the first name, and a 3 digit number. When my results were obtained and I wanted to link these ID codes back to the players' full names I used the database found at astrosdaily.net.[23] This database does link the Retrosheet ID's to the players' full names and so I used the Astrosdaily.net data to get the player names in my results tables. All of the actual analysis, however, was conducted using the Retrosheet data. I went up until 1992 because that is the farthest that Retrosheet made free play-by-play data available online. I only went back to1974 because that is the earliest year that Retrosheet's play-by-play data is not missing games in either league. I would like to thank Retrosheet and Astrosdaily.net for making this data available free of charge.

3. The clutch/choke regression model

In this section, we describe the details of and reasoning behind the construction of the regression model we designed to detect patterns of clutch or choke hitting. Sub-section 3.1 provides an overview of the model and the sub-sequent sub-sections discuss the details of its construction.

3.1 An Overview of the Model

Most examinations of clutch hitting in the past have focused on defining some set of situations that were deemed especially important and seeing if players performed significantly differently in those situations than in all other situations, which were all lumped together. The most well known of these is the Close & Late definition, also known as LIPS (Late Inning Pressure Situation.) Defined as any situation in or after the 7th inning where the game is tied, a one-run game, or the tying run is on-base, at the plate, or in the on-deck circle.[24] There are two problems with this. Firstly, the dividing line is inherently arbitrary, so there is a lot noise when you use such a comparison, which could mask the effects being looked for. Secondly, these situations aren't even always more important than other ones, thus the line is also inconsistent.

[22] Inserting this statement verbatim, starting with "the information used…" Is the one thing that Retrosheet requires of people who use their data in the making of a publication.
[23] For the 19 seasons I used, astrosdaily.net derives its database from Retrosheet.
[24] This definition was obtained from www.enlexica.com

For example, coming to the plate in the 6th inning in a tie game with the bases-loaded and two outs, would seem to be a more critical situation than coming up in the 7th with the tying run on-deck and no outs, yet LIPS defines the 2nd situation as critical and the 1st as not.

Our regression model solves this problem. We use for our independent variable a measure of the importance of the situation in which the plate-appearance occurs that is performance independent, meaning that the importance of the situation would be rated the same whether the result was a strikeout, a homerun, or anything in between. For our dependent variable, we use the entry into OPS of the plate appearance, which is a simple and straightforward measure of the performance at the plate, independent of the situation. [25] OPS entry assigns a 0 for an out, 1 for a walk, 2 for a single, 3 for a double, 4 for a triple, and 5 for a homerun.

Now using these two variables we can create a scatter-plot for each player's plate appearances in each season, and run a regression on it. If, in a given player-season, the regression indicates an upward trend, this indicates a pattern of clutch hitting by that player in that season. Similarly, a downward trend indicates a pattern of choke hitting. This interpretation of our regression model follows naturally from its definition. The X-variable (i.e independent variable) measures importance of situation and the Y-variable (i.e. dependent variable) measures the performance. Therefore, an upward trend means that as the importance of the situation increases, the level of the performance tends to increase. This is exactly what clutch hitting is, a tendency to do increasingly well at the plate as the situation gets increasingly important. By the same token, a downward trend means that as the importance of the situation increases, the level of the performance tends to decrease, thus showing a pattern choke hitting.

The sign of the slope,[26] while indicating a pattern, does not necessarily indicate a significant one. Even if a hitter truly performs at the same level regardless of the situation, the slope of his regression is almost never going to be perfectly flat because of chance variation. Our methodology for using our regression results to determine which players, if any, have statistically significant clutch or choke tendencies, and whether the number of significant results observed in the sample is the result of the presence of true clutch and/or choke hitter(s), or if it could feasibly be the result of chance is the topic of Section 4 of our paper.

3.2 The Construction of the Importance Index

[25] We actually make some modifications to OPS-entry that will be discussed in full detail later in this paper.
[26] Since we use non-linear regression, the regression curve does not, strictly speaking, have a constant slope, but it is either monotonic increasing, monotonically decreasing, or flat (in practice it is rarely exactly flat)

The sole object of a baseball game is to win. Any index that measures the importance of a game situation should reflect this self-evident fact. Any plate appearance, where the probability of winning if the hitter struck out would be roughly the same as if he hit a homerun is clearly not a very important one. There are, of course, things besides hitting a homerun that would increase the probability of a hitter's team winning some examples are: walk, get a single, a double, or a triple. Sauer and Hakes, as well as others who have tried probabilistic methods, did not use separate measures for importance of situation and outcome, using a single metric instead. For each plate appearance, they saw how much the actual outcome changed the probability of victory over the average amount that the same outcome changes the probability of victory. [27] Thus, coming up with one metric that measures some combination of situational importance and outcome. However, when trying to detect clutch and/or choke hitting, intertwining importance of situation and value of outcome is a major fallacy. For the reason that the question being asked is no longer ' Is this guy a clutch or choke hitter?' but a somewhat subtly different question: 'Does this guy wind up improving his team's chances of winning games more or less than we would expect someone who got the same number of Plate appearances, outs, walks, singles, doubles, etc. as he did.' This question is different because the situation's importance is partly being defined by the player's performance, which is part of the outcome, not the situation. Furthermore, the player's performance is also partly defined by the situation. An example should serve to clarify this somewhat esoteric, but vitally important, point. If a hitter steps up to the plate, tie game bases empty and two outs in the bottom of the 9th, that situation should be assigned a certain level of importance, regardless of the performance. However, using Sauer and Hakes methodology, the situation would be treated as very important if the player were to hit a homerun, and not very important if he were to strikeout. The problem here is that the deed that the batter does is being confused with the context in which he does it. By having one number that attempts to measure both the deed and the context, it ends up measuring neither. If I were the batter in this situation, I would strike out every time.[28] But that doesn't really make the situation less important; it just means that I'm not good enough to face major

[27] Jahn H. Hakes and Raymond D. Sauer, "Are Players Paid for 'clutch' performance?" John E. Walker Dept. of Economics, Clemson University. Preliminary Draft. June 30, 2003.

[28] The true probability of me not striking out against a major league pitcher is not, of course, exactly 0. But I would estimate that 0 is accurate to at least 8 or 9 decimal places.

league pitching, regardless of how important the situation is. To really answer the clutch/choke hitting question: "Does this guy tend to perform better/worse as the situation gets more important?" we must have separate measures of performance and importance, and not one measure for both. That would be like a thermometer that responds to changes in temperature and humidity; you'd observe the level, but you wouldn't know how hot it was, because you wouldn't know what temperature-humidity combination had caused that level. Analogously, our importance index cannot use the probability impact of a player's actual performance to measure importance. We need a measure of importance that is performance independent.

To construct the importance index, we use a weighted combination of possible outcomes of the plate appearance. However, since a plate appearance always results in a change in the game state[29] we look at the value of possible outcomes over an out without base-runner movement, rather than over the prior game state, since a return to a prior game state is not a possibility. The formula of the importance index is:

Let $f(EventX_{ij}) = \overline{EventX}[\Pr(\text{batting team wins} \mid EventX_{ij}) - \Pr(\text{batting team wins} \mid SO_{ij})]$

Where $\overline{\text{EventX}}$ = the league average frequency of EventX per plate appearance.

and the subscripts ij denoting the player i 's jth plate appearance.

$$\text{Clutch index}_{ij} = \frac{f(HR_{ij}) + f(3B_{ij}) + f(2B_{ij}) + f(1B_{ij}) + f(BB_{ij}) + f(HBP_{ij})}{\overline{HR} + \overline{3B} + \overline{2B} + \overline{1B} + \overline{BB} + \overline{HBP}}$$

Note :the clutch index yields values on the open interval (0,1). However, since I truncated the distributionof runs at 12, the index, in practice, is on the semi - open interval [0,1).

Note : See appendix for list of abbreviations

We call this importance index "Potential Impact Over Generic Out" (Hereby referred to as PIOGO). This formula requires some explanation. Let us start with our what is meant by the probability of victory given a certain event. Those of you familiar with baseball may be wondering: How do we know where any men already on base will wind up in the event of a single, double, or out?[30] We used a set of base-running assumptions devised by D'esopo and Lefkowitz that has become the standard one that is used in Markov-models of baseball such as

[29] Number of outs, half inning, which bases are occupied, the score, or some combination of these.
[30] Base-runner advancement is completely predictable on Homeruns, walks and triples.

those used to find mathematically optimal lineups, among other things. The assumptions are as follows:

The assumptions are as follows:

1) All outs are treated as strikeouts meaning that one out is recorded and a base-runners do not advance.

2) On a single the batter goes to first, a runner on first advances to second, and runners on 2nd and 3rd advance to home.
3) On a double the batter goes to 2nd, a runner on first advances to 3rd and runners on 2nd and 3rd score.
4) On a triple the batter goes to 3rd and all other base runners score
5) On a homerun the batter and all base runners score[31]

Thus, we can now calculate, for example $\Pr(\text{batting team wins} \mid 1B_{ij})$ by (possibly) adjusting the bases occupied and score[32] using the D'esopo-Lefkowitz assumptions and then get the probability of winning from that hypothetical state. This brings us to the matter of how we went about calculating the probability of a team winning from a given game state. In order, to accomplish this, we made a number of simplifying assumptions. First, we assumed that all teams were league average teams, meaning that the distribution of runs any would score in the remainder of any half-inning, in a given base-out situation[33] was the distribution of runs scored from that base-out situation in the that league-year.[34] Second, we assumed that scoring in all half-innings, are independent identically distributed, having the same distribution as all other half-innings. Third, we truncated the distribution of runs in the remainder of a half-inning at 12, meaning we assumed that the number of runs scored in the remainder of a half-inning would be no more than 12, any part of the run distribution that was to the right of 12 was moved to 12. This move changed very little since innings of 13 or more runs were rare in the sample, most league-years did not even contain one such inning. Finally, the possible change in a team's

[31] D.A D'Esopo and B. Lefkowitz, "The Distribution of Runs in the game of baseball." In Optimal Strategies in Sports, 1977. Amsterdam; New York: North Holland Pub. Co., pp 55-62.

[32] For other events, the number of outs and possibly the current half-inning might need adjustment as well.
[33] A base out situation is a situation which describes the number of outs in the half-inning and which bases, if any, are occupied.
[34] A league-year specifies a year of play in one of the two Major Leagues (American or National)

lead,[35] for the remainder of the game, cannot exceed 12. This last assumption means that 0 is used as an approximation for the probability of victory for a team that trails by more than 12 runs, the other team winning with probability 1 since ties are not allowed. This does not mean that the importance index is necessarily 0. If the hitter could potentially make it a 12-run game with a walk, single, double, triple, or homerun. Then the probability of winning, given those events, might be non-zero, thus the index would have a non-zero (albeit a near-zero) index value.

The last two assumptions are mainly to simplify calculations and have negligible impact on the accuracy of the true probability, especially since 13-run innings are rare and 13-run comebacks are unheard of.

Even though the assumptions that the two teams are league average and that the half-innings are independent may cause considerable deviations from the true probabilities, we believe they are justified given the context in which they are used. Clearly, if we were really interested in a given team's true chances of winning this would not be a reasonable assumption, but what we are truly interested in is the overall importance of a situation; probability of winning given potential outcomes is only the way we measure it. Therefore, we believe that using aggregate league data for this purpose does not detract significantly from this goal. A team that scores a lot and has a better chance than most of the league to comeback in the situation that would follow a walk or hit, will also have a better chance than most of the league to comeback in the situation that would follow an out, unless it was the last out. So the impact of this assumption would be more minor than it first appears, since a positive event (for a good team) would add less value, and a negative event would take away more value. When we subtract the probability of winning given a bad event (an out) from the probability of winning given a good event (walk or hit) both numbers are off in the same direction. Thus, their difference actually is more accurate than either of the two numbers being differenced. Furthermore, even though these assumptions yield probabilities that are inaccurate in a given game, since our purpose is to find the general importance of a situation, (in the league that year) assuming the two teams conform to the league average data is an acceptable simplifying assumption. After all, all this assumption really does is make this a measure of the importance of a given situation in a given league for a given season. Finally, we will address the assumption of independent identical distribution. This means that the distribution of runs scoring in a half-inning do not depend on what happened

[35] If the team is trailing or tied this simply means the lead is negative or 0.

in other half-innings or what inning it is. Lindsey shows that the distribution of runs is greater in some innings than others[36] even before the advent of the modern closer. But by trying to establish this difference we would likely do more harm than good. This is because instead of using the data from a league-year to generate 24 distributions, one for each base-out situation, we would need 216 distributions, one for each base-out-inning; we might even need 432 distributions if we distinguished between half-innings. For many of these distributions, the sample size would be too small, and the distributions arrived at might end up much further from the true distribution than if we didn't distinguish between innings. Therefore, we do not distinguish between different innings. Modeling dependence of scoring between innings in the same game is even trickier. Therefore, this independent identical distribution assumption was adopted as the best feasible methodology.

The calculation of the probability of victory from a given game state was done very straightforwardly. For each league-year, we recorded for every plate appearance the base-out situation and number of runs scored from that point through the end of the half-inning, and tabulated them to create distributions for each base-out situation in each league-year. Any time 13 or more runs score by the end of the half-inning, we count that entry as if 12 runs scored. Since every half-inning begins with bases-empty no outs, we can now string together the distribution of the present half inning with all future ones. Any runs scored in the tops of innings are counted as negative runs because they subtract from the home teams lead, and any in the bottom are counted as positive runs. We then jointly distribute the present base-out situation with ones where there is nobody out and nobody on for each remaining half-inning. Unlike a typical discrete joint distribution, the number of distribution points will not grow exponentially because any combinations of scoring in various half-innings summing to the same number are lumped together. We do this joint distribution process recursively until the end of the 9th is reached. Any parts of the distribution less than –12 are moved to –12 and any parts greater than 12 are moved to 12. We now have a distribution for the change in the home team's lead by the end of the game. The probability that the home team wins is the probability that the sum of the

---

[36] George R. Lindsey, "The progress of the score during a baseball game." *Journal of the American Statistical Association,* Vol. 56, Issue 295, Sep. 1961. pp. 703-728

home team's current lead and the change in their lead for the remainder of this game is greater than 0, plus half the probability that this sum is 0. The away team's probability of victory is simply the complement of the home team's. The reason for the second term (half the probability that the aforementioned sum equals 0) is that, under our assumption that teams in the league have identical run distributions, the chance of winning a game that is tied at the end of any inning, including the 9th and subsequent innings, is exactly .5; it's a coin toss. Therefore, when we are tied at the end of 9 innings, we count this as half a win, since each team has precisely a 50-50 chance of winning from this point, under our model. Similarly, each extra inning situation can be treated as a 9th inning situation, since, just like the 9th inning, at the end of the current inning either the game is over, or it is a tie-game coin toss. Thus for any combination of league-year, base-out, half-inning, and home team's lead, we have a probability of home team victory, whose complement is the away team's probability of victory.

To conclude the discussion of the importance index, I would like to discuss what it doesn't do. Many people think of a clutch hitter as a big-game hitter. This is not what my importance index measures. In fact, my dataset doesn't even include playoffs. All my importance index measures is the importance of a situation to winning that particular game. No games are given extra weight over other games. This is because any weighting scheme I could come up with would be completely subjective and arbitrary, as discussed in the introduction. Therefore, I am not looking for the guy who hits better in the big games, but rather for the guy who hits better at big points in the game.

3.3 The treatment of outcomes in the regression model.

OPS is a measure of a batter's overall productivity at the plate that became much more widely known after the publication of Moneyball, a book about the general managerial exploits of Oakland A's General Manager Billy Beane. It is obtained by adding the player's On-base percentage (OBP) and slugging percentage. (SLG) [37] The entry that a plate-appearance makes into OPS is 0 for an out, 1 for a walk or other non-hit that counts as an on-base event 2 for a single, 3 for a double, 4 for a triple, 5 for a homerun. There were some types of plate appearance outcomes that, for reasons that will be discussed in this section, I did not simply make the OPS entry the value of the dependent variable and throw the point into the regression. I will go

[37] For a description of what on-base percentage and slugging percentage are see the appendix on common baseball terms and abbreviations

through each of outcome types that I treated differently than I would have had I used an unmodified OPS entry, discussing what the outcome type is, how I treated this outcome type, and the justification for this treatment.

There are two types of outcomes whose result is functionally equivalent to a standard walk but I treated differently in the regression: Intentional base-on balls (IBB), and hit-by pitch (HBP). The intentional walk is a play where the opposing team decides that rather than face a hitter they will walk him, and the catcher stands up and sets the target high and 2-3 ft off the plate, just to make sure it's not remotely hittable. After 4 balls it's a walk just like any other walk. I exclude any plate appearances that result in an intentional walk because this tells us nothing about a player's clutch/choke ability because he walked solely due to the other team's decision. Although, some players (e.g. Barry Bonds) may get a lot of IBB's in key situations, unless you believe in Jedi mind tricks, that doesn't mean they are clutch hitters, just that other teams are afraid of them. Whenever a batter is hit by a pitch (assuming he is inside the batters box) he gets a free-pass to first, and any runners that this forces forward advance as well, just like a walk. Although there are some factors that may contribute to some players being hit with greater frequency than others: unique batting stances, unusually large body mass, underdeveloped self-preservation reflexes etc. None of these are things that a batter changes in the clutch. Certainly, a hitter will not alter the batting stance that works for him in a key situation, and when the game is on the line no one gains 50 pounds. Perhaps one could argue that when the game is on the line a player somehow can suppress his self-preservation reflexes in a key situation and "take one for the team." But reflexes are precisely what are needed when hitting, so I find this argument rather unconvincing. In addition, the relatively low frequency of HBP's means that even if I am incorrect in my rejection of this argument, it is unlikely that the regression results would be altered in any meaningful way if I included HBP's at a value of 1, the same as a walk.[38]

A sacrifice bunt is a play where a player bunts in order to advance a base-runner at the cost of being thrown out at first himself. Since the decision whether or not to attempt a sacrifice bunt is generally made by the manager, I decided to exclude them from the regression. Although, some players might be better at successfully executing a bunt in a key situation, the dataset does

[38] Since HBP's are functionally equivalent to walks, if they were to be included, this would be the only sensible value to use.

not indicate when a player bunted a ball foul and then (usually by managerial instruction) stopped attempting to bunt. Thus, the guys who are bad bunters may not show up as such, it all depends so much on the manager that I decided including it would not significantly help uncover any true clutch or choke hitters, and would likely mask them by adding largely irrelevant data points to our regression scatter-plot. Even if sacrifice bunts were included, there would be issues with what number to assign them in the modified OPS entry. In some cases, managers order a sacrifice bunt in a scenario where a sacrifice bunt is harmful to a team's chances of winning.[39] Occasionally, Retrosheet will have down a flag indicating sacrifice hit that is neither a sacrifice bunt nor a sacrifice fly, I have looked through most of the instances where this occurs in the data and have not seen a single instance where this scores a run; this effectively is something that accomplishes the same thing as a sacrifice bunt. But since the hitter wasn't bunting this was likely not his intention. In addition, this category seems particularly vulnerable to home-town official scorer nepotism; especially since it all depends on the scorer's ability to read a player's mind and determine if he intended to ground out to advance the runner. Thus, I ignore these non-bunt sacrifice hit designations and treat the outcomes as regular outs.

A hitter can reach on an interference call if the catcher reaches out over the plate to catch the ball, not giving him a chance to hit it, or if a fielder blocks the baseline, not giving the runner a chance to beat out any potential throw. What would have happened had there been no interference varies quite a bit. Due to this, and the fact that interference calls occur too infrequently to have any discernable impact on the regression results, I exclude the plate appearances with interference from the regression.

Any event which does not result in a new batter is excluded from the regression, simply because it is problematic to have two different outcomes for the same plate appearance. If a base runner advances or gets out in the middle of a plate appearance the game-situation that we use to determine the importance value of a plate-appearance is the game-situation immediately before the pitcher throws the final pitch of the plate-appearance. If a base-runner makes the third out in the middle of a plate appearance, then the situation that we count when the batter returns to the plate to leadoff the next inning, if the team does bat again then the plate-appearance is ignored. In the case of a strike-em-out-throw-em-out-double-play, since the caught stealing occurred after

[39] Discussions of this can be found in Curveball ,Lindsey's "an investigation of strategies in Baseball" in some of the works of Bill James, and in the best-selling MoneyBall complete bibliographic info on these works is in the works cited section.

the final pitch of the plate-appearance was thrown, the batter is considered to have struckout in the pre-steal game-situation.

There were two event types that I thought could be reasonably treated in more than one way in the regression: errors and Sacrifice flies. Therefore I ran the regression for each of the combinations of treatments of these two event types that I considered reasonable. An error is when a fielder makes a mistake, as determined by the official scorer, that either prevents his team from making an out on the play, or allows base-runners to advance further than they would have without the mistake, again as determined by the official scorer. When the batter gets, for example, a double and reaches third on an error, I simply count it in the regression as a double.[40] Also, if the catcher drops strike 3 and the hitter reaches, I simply count that as a strikeout in the regression. The reason for this is the hitter didn't even make contact, so saying that him standing at first is the result of some sort of clutch-like ability to create errors seems patently flawed. However, when the hitter puts the ball in play and reaches safely on an error, then I am unsure what to do. Since an error is a highly subjective official scorer decision, I am in a quandary. Errors are most frequently called in cases where a fielder either touches the ball but can't hold on to it or throws significantly off target to attempt a play. So it is feasible that by running harder in a key situation, a batter could induce a rushed throw, thus inducing more errors in key situations. Similarly, a if there are batters who are true a clutch/choke hitters, they might very well hit the ball harder/softer and maybe with more/less topspin in key situations, resulting in more hits, but also making the balls hit within the fielders' range harder to field cleanly, thus resulting in a higher proportion of errors. That said, errors do seem to be mostly attributable to fielders. Thus I came up with three possible ways to treat reaching on errors on balls put in play. In the regression: Exclude those data points, count the OPS entry as a 0, (as an out; this is how it is treated in batting average) and count the OPS entry as a 1 (as a walk, or put another way, this is like giving him credit for half a single.) I did not consider counting either errors or sacrifice flies in OPS entry as something other than the integers that OPS entry already uses.[41] This is because too much tinkering with assumptions by trying numbers like .4, 1.23, or .724 could amount to rigging the results. So I restrict myself to the integers. I could not logically count errors as a 2 or higher; even if a hitter has some ability to affect errors, the hits are still a better indicator of

[40] The same getting an extra unearned base off a single, or triple.
[41] 0,1,2,3,4,5

his performance independent of the fielders. Therefore, these three choices were my only options.

Sacrifice flies are when a hitter hits a fly out to the outfield, that scores a runner from 3rd.[42] There are many situations where this play has value, sometimes it can win the game, and it is much more likely to be intentional, or semi-intentional[43] than a non-bunt sacrifice hit. Logically, I could not count it in OPS entry as a 2 or higher, since it is unequivocally worse than a single, which would score the runner without costing an out. It is also unequivocally better than an out that doesn't advance runners, since it scores a run, so counting it as a 0 would also be highly questionable. But since it might done unintentionally, especially in games that aren't very close, I might also want to exclude it from the regression. Given that, I am constraining myself to using integers in my modified OPS entry, I am left with 2 possible ways to treat sacrifice flies in my regression: exclude those data points, and count them in OPS entry as a 1. This leaves me with 6 possible sets of treatments to run a regression with, one for each possible combination of one of my 3 sets of error assumptions with one of my 2 sets of sacrifice fly assumptions. So I run my regression model and hypothesis test on each of the 6 possible ways separately and present results for all 6, along with analysis and interpretation.

3.4 Restrictions on player-years for inclusion in the regression

My data set, as mentioned above included all seasons in both Major leagues from 1974 through 1992. I ran a regression on each player-year. If a player switched leagues mid-year, he still had one regression for that year. However, his importance index values were calculated according to AL data of that year while in the AL, and NL data of that year while in the NL. Regressions were only done on player-years that met two criteria. First, that the player had at least 100 PA's during that year. Second, that the player had at least 2 seasons with 100 or more PA's in our sample. The 100 PA condition is to ensure that we do not run regressions on samples that are too small to be meaningful. The second condition has to do with the methodology of our hypothesis test and will be explained more fully in section 4. Once these restrictions are applied we end up running regressions on 6784 player-years, belonging to 1075 distinct players.

---

[42] By definition, this requires there to be fewer than two outs in the inning as well. It is possible that a runner scores from second without an error, but such occurrences are very rare.

[43] E.g. the batter tries to "get under" the ball a little more so that even if it is an out, it has a good chance of scoring the guy on 3rd

3.5 Choosing the form of the Regression Model

Now that I have data to run regressions on, I must choose what form the regression curve should take. The standard linear model is out of the question. This is because there are upper and lower bounds on the data. If we were to use a linear model it might say that, in a very important situation, a strongly-choke hitter would be expected to somehow do worse than get out every single time.[44] Also, a linear model would say that whenever importance increases by a fixed amount, expected OPS-entry increases by a fixed amount. Upon closer examination this seems highly illogical. It implies that as importance increases a clutch hitter's performance keeps going up at the same rate, regardless of how high it has already climbed. A model with upper and lower bounds would solve both these issues by placing limits on how much better/worse a clutch/choke hitter could get in a super important situation. Another problem with linear regression is that our dependent variable is discrete, only taking on 6 distinct values, so the assumption that the residuals are normally distributed with a constant variance is necessarily violated, since the normal distribution is continuous. Ordinal logistic regression solves these problems and works with a discrete dependent variable with more than 2 possible values. It also uses the fact that our dependent variable has a natural order to fit a better model. [45] In a logistic regression model, the rate of change first gets bigger and bigger and then gets smaller and smaller. However, logistic regression has the restriction of being symmetric around the point of inflexion, so the (absolute) rate of change is required to get bigger and bigger on one end, just as quickly as it gets smaller and smaller on the other end. I ran a logit model, but the results did not provide significant evidence for the existence of clutch and/or choke hitters. However, I thought that a model that allowed for asymmetry around the point of inflexion would be a better fit. In the Gompertz, the underlying distribution is asymmetric. The magnitude of change initially gets bigger more slowly than in logistic, then increases more rapidly, and then gets smaller faster. [46] This fits the intuition that the bulk of the clutch/choke effect of a clutch/choke hitter wouldn't begin to kick-in in any meaningful way until the level of importance got pretty high, then the effect becomes a noticeable difference fairly quickly, but eventually flattens out when it reaches

[44] Remember I don't count Double Plays differently from outs in my model, so this is literally impossible.
[45] Homerun is better than triple, is better than double, is better than single, etc. This differs from some discrete variables which represent things like different flavors of ice cream, ethnicities, or regions; things that don't really have an inherent ordering.
[46] SAS 8.2 Online Documentation. The probit procedure. Overview.

a limit. This "intuition" is more intuitive than it sounds. A baseball player doesn't calculate probabilities, but he probably (from his experience playing baseball) has a reasonably good idea of the importance of the situation. If clutch or choke hitting exists, it would have to be some sort of mentally triggered phenomenon, since realization of importance is mental. Therefore, it is very reasonable that any clutch/choke abilities a hitter might have would kick-in once the perceived level of importance crosses some psychological threshold. This threshold wouldn't be an exact level of importance because players don't calculate probabilities, and because a player might have slightly different thresholds on different days, depending on his mood. However, any clutch or choke abilities that a player might have should be kicking-in in his threshold neighborhood. In other words, there is an approximate level of importance at which the player decides (maybe subconsciously) "this is really important" at which point any clutch/choke abilities he might possess to kick-in. The Gompertz better models this by changing more rapidly in a smaller region, and being flatter outside that region, and not requiring symmetry around that region. Additionally, a hitter who chokes under pressure would perform worse in a very important situation than in a very unimportant one, but the difference in his performance between an unimportant situation and a modestly important one would likely be somewhere between tiny and non-existent. If what we mean by a clutch hitter is someone who comes through in the very important situations, rather than someone who merely slacks off in unimportant ones, than this kind of asymmetry is also needed to describe the effect. It should be noted that the Gompertz regression is equivalent to using the complementary log-log transformation. However, the procedure I used in SAS refers to it as Gompertz rather than the complementary log-log, so I will refer to it as Gompertz.[47]

4. The Test for the Existence of Clutch/Choke Effects

In this section, we detail our procedures for testing for the existence of true clutch and/or hitters. Our procedures test for both whether clutch and/or choke hitters exist, and for who they are.

4.1 The hypothesis test and test statistic for each player-year regression

[47] There is another Procedure in SAS that refers to it as the complementary log-log, but I found out about it only after coding it this way, and since they both work equally well, I stuck with this one.

Once we have a scatter-plot for each player-year,[48] we must decide how to test each player year. Since a positive coefficient on the importance index indicates that as importance goes up, performance tends to go up, a positive coefficient should indicate clutch ability. Similarly, since a negative coefficient on the importance index indicates that as importance goes up, performance tends to go down, a negative coefficient should indicate choke ability. But even if no such thing as clutch or choke hitting existed, the coefficients would almost never equal exactly zero. So we do a 2-tailed hypothesis test for the coefficient with H_0: $\beta=0$ H_A: $\beta 0$. When we reject H_0 we also want to know whether the coefficient was significantly positive or significantly negative to determine if this was a choke season or a clutch season. Now for a test statistic there are three choices: the Wald test, the Score test, and the Likelihood test. The Wald test exhibits erratic behavior when samples are not sufficiently large, failing to reject H_0 when it should.[49] The likelihood test works better over a wider variety of sample sizes, although for large samples all three tests become asymptotically equivalent.[50] Besides Hosmer-Lemeshow, the other literature I have read recommends the likelihood test for small-samples, and it doesn't seem to have any real disadvantages when sample size is not small.[51] The likelihood test involves taking twice the difference of between log-likelihoods of my model with the included importance index and the with the importance index excluded.[52] Since my only independent variable is the importance index, the model without it is an "empty" model containing only 5 intercepts, since there are 6 possible outcomes. This empty model is basically like the random spinner model described by Albert and Bennett,[53] in this empty model the situation is completely ignored and each player has

[48] Actually, 6 sets of scatterplots for the 6 sets of error and sacfly assumptions as described above. It is also only for players with 2 or more years of 100 or more PA's and only for those years in which they had 100 or more PA's

[49] David W. Hosmer and Stanley Lemeshow, Applied Logistic Regression. New York. Wiley, 2000. pp. 16

[50] ibid pp.16

[51] German Rodriguez. Lecture notes for WWS509 Generalized linear models: A.2 Tests of hypotheses. Princeton University. Other literature on the topic was examined but I stopped printing them out when I realized that they were all saying the same thing about the use of these 3 tests.

[52] It can also be –2 times the difference depending on which model you subtract from the other. The correct choice always yields non-negative numbers, since it has a chi-squared distribution, which is non-negative.

[53] Jim Albert and Jay Bennett, Curveball: Baseball, Statistics, and the Role of Chance in the Game Copernicus Books. New York. 2001 pp. 60

a circular spinner divided into 6 sections: out, walk, single, double, triple, homerun.[54] Each player, depending on his general hitting abilities has a fixed chance of doing each of these things every time he gets to the plate. The model with the importance index is one where the lines between the spinner's sections move depending on the value of the importance index. The test statistic attempts to answer the question: Is the moving spinner really a better description of what's going on here, or is the regular spinner just as good?

The test statistic has a chi-squared distribution[55] with, in our model, one degree of freedom. Since the chi-squared distribution is built into SAS, getting the p-values is relatively easy. We run our model twice, once with the importance index, once empty. We take twice the difference and find the test statistic. Then we find the amount of probability mass to the right of our test statistic on the chi-squared distribution with one degree of freedom. That is the p-value of our test.

In order to determine whether the p-value we get ought to be considered significant we must choose a player-year significance level α. Choosing α is not as straightforward for purposes of our test as it is for most hypothesis tests. Our methodology for choosing α and justification for it, are the topic of the following section.

4.2 Finding single-season significance levels.

Although we run a regression on each player-year, our real interest is not in finding clutch/choke single-season performances, but in finding players who are themselves clutch/choke. Even a hitter who has one clutch/choke season in a 10 year career probably has no real clutch/choke ability, just a (un)lucky year. As a player shows the same effect in more and more seasons, luck looks less and less like a plausible explanation. Consequently, the significance test we really want to run is a career significance test that looks at each of a player's single season regression results and tests the overall picture they paint, rather than looking at them in isolation. Therefore, when we select the fixed significance level to test each hitter for clutch or choke ability, the relevant significance level to test all players at is not the single-season significance level α but some career significance level denoted by $\aleph$. (pronounced:

[54] If you recall from the discussion above, the categories walk and out. They actually include what we are counting as OPS entry of 1 and 0 respectively, and are not necessarily walks or outs, but usually are.
[55] Hosmer Lemeshow pp.14

Aleph) We want to use this same career significance level ℵ for all players in the sample whether they have 2 seasons in the sample (our minimum requirement) or all 19.

As we stated earlier the more seasons a hitter shows the same effect, meaning either clutch or choke, the more convincing the evidence would seem. Now we must ask how many seasons must a hitter show the same effect to be deemed to truly possess that effect. Logically, this minimum can be no less than a simple majority of the seasons of his in-sample career. The reason is best illustrated by example. If we were to be just slightly more lenient and say that a hitter show up clutch in half or more (instead of more than half) of the years of his in-sample career to be deemed clutch, and show up choke in half or more of those same years to be deemed choke, then we set ourselves up for a logical contradiction. That is a player who has played 2 years in the sample could show up clutch in 1 year and choke in the other, yielding the patently absurd result that this man is both a clutch and a choke hitter! By requiring a majority of seasons pointing in the same direction to draw a conclusion, we avoid this logical contradiction. For that reason, we establish this majority of seasons rule: In order for a player to be deemed clutch, he must show up as clutch in at least a simple majority of the seasons he plays in our sample; in order for a player to be deemed choke, he must show up as choke in at least a simple majority of the seasons he plays in our sample.

Since we are really examining whether a player is clutch or choke over a whole career, one may ask why we do separate regressions for each year when we could just do one regression that uses all qualifying data points in a player's career. One reason why we treat seasons separately concerns the manner in which the importance index is constructed. To get its probabilities of victory and its frequencies of various events, the importance index uses data from that league in that season. So the same situation in different league-years can, and typically does, have different values of the importance index.[56] Thus, lumping different years of data into a single regression takes data points with importance index values calculated using different empirical data, and treats them as equivalent, when they were really calculated on different scales. Essentially, this amounts to comparing apples and oranges on the same scatter-plot.

---

[56] Between adjacent seasons these differences tend to be small. However, over the course of 10 or more years the shift in the probabilities and frequencies can become noticeable.

We now seek to apply our majority of seasons rule in a way that yields the same career significance level $\aleph$ for all players, regardless of how many seasons they appear in our sample for. If each season of a given player is tested at α, that means that if there the hitter were actually neither clutch nor choke the probability of him showing up as either clutch or choke is α. Additionally, the chance of him showing up clutch would be $\frac{\alpha}{2}$, and the chance of him showing up choke would also be $\frac{\alpha}{2}$. The career significance level $\aleph$ has a similar interpretation, the probability of a player who is neither clutch nor choke being mistakenly deemed to be clutch or choke is $\aleph$. If we were to use the same α for a player with 3 seasons of data as for a player with 19 seasons then the player with 3 seasons would have a higher value of $\aleph$ than the player with 19.[57] Pressure-neutral player who played 3 seasons and need only show a non-existent effect 2 out of 3 seasons, Therefore, if we are to make it so such a pressure neutral player would have the same value of $\aleph$, regardless of whether he played 3 seasons or 19, then each player must have an α that depends not only on the value of $\aleph$ chosen for all players, but also the number of seasons played by that particular player. Determining the one-year significance level α that yields career significance level $\aleph$ is simply an extension of the basic formula for the distribution of a binomial random variable. As stated above, the chance that he ends up in the significant part of the clutch tail is $\frac{\alpha}{2}$, and of course the probability of winding up in the significant part of the choke tail is also $\frac{\alpha}{2}$. Since a player must end up in the significant part of the same tail for at least a simple majority of his n seasons to be deemed to truly be clutch/choke, we denote the smallest integer that is a simple majority of n with the letter k.[58] So the probability of a player ending up significantly clutch is:

$$\sum_{i=k}^{n} \left(\frac{\alpha}{2}\right)^{i} \left(1-\frac{\alpha}{2}\right)^{n-i} \binom{n}{i}$$

[57] Technically this only holds true if $\alpha < .5$, but this condition is trivial in practice. The standard significance level in most applications of statistics is $\alpha = .05$ (although really the significance level we fix here is $\aleph$.) and all of the solutions for α that I use are below .5. In addition, since I require that the result be significant in the same tail for a majority of seasons, the real relevant figure is $\frac{\alpha}{2}$, and that figure never even comes close to .5.

[58] For the mathematically finicky a formula for k is k(n)=INT((n/2) +1)

The equation for the probability of a true importance-indifferent[59] player appearing in the significant part of the choke tail for at least a majority of his seasons is, of course, identical. Hence, the career-significance-level $\aleph$, representing the probability that a player who in reality satisfies the null hypothesis (i.e. is neither clutch nor choke) will be determined to be either clutch or choke by our test is given by:

$$\aleph = 2\sum_{i=k}^{n}\left(\frac{\alpha}{2}\right)^{i}\left(1-\frac{\alpha}{2}\right)^{n-i}\binom{n}{i}$$

Given a player's n and a desired value of $\aleph$, we can now find the appropriate single-season significance level α to use for any given player, such that all players will have the same career significance level $\aleph$. All that remains to be done is to choose the desired career significance level $\aleph$.[60]

4.3 Testing for the Existence of Real Clutch and/or Choke Hitters in the Population

There were 1075 players that met the criteria to be included in our regression.[61] Therefore, at any value of $\aleph$ that was even remotely reasonable, there was a very high probability that some players would appear as clutch and/or choke even if there were no true clutch or choke hitters in the population. This means that even once we have tested the individual players for clutch and choke ability, we must test the population to see if such results are unlikely to be occur in a population composed entirely of hitters who are neither clutch nor choke. The way we do this is by getting a p-value for the existence of at least some clutch and/or choke hitters in the population. Our null hypothesis is that no one in the population is clutch or choke, and the alternative hypothesis is that one or more hitters in the population are truly clutch and/or choke. This p-value represents the probability that, if none of the 1075 players are truly clutch or choke (i.e. if the null hypothesis were correct), that we would get at least as many players appearing clutch and/or choke as we actually did. If none of the players were indeed clutch or choke, then the probability of any one of them showing up as significantly clutch or choke is

[59] This term is used interchangeably with the term pressure neutral.

[60] See table 16 for the solutions of α for each value of $\aleph$ and n.

[61] That is they had at least 2 seasons of at least 100 PA's in our sample.

given by our chosen value of א. Thus, applying the binomial distribution again, the p-value for the existence of any clutch and/or choke hitters is given by the equation:

$$\sum_{j=N}^{1075} א^{j}\left(1-א\right)^{n-j}\binom{1075}{j}$$

4.4 Choosing the Career Significance Level

There is a distinct possibility that if there are really clutch and/or choke hitters that they are few in number. The fact that many other sabermetricians have, after careful examination, concluded that no such effects existed in any meaningful way would seem to make it less likely that they are too numerous, or the effects too pronounced. Otherwise, their presence would have already been detected in a statistically significant way. Therefore, our choice of א should be relatively small to allow us a good chance of getting significant results for the existence of clutch/choke hitters should only a handful of such players actually exist. Therefore, for my initial test, I chose $א = .001$ this is the level that I used for the non-Gompertz logistic regression and also for my first pass through Gompertz. Getting results whose significance varied greatly across the 6 sets of assumptions regarding sacrifice flies and errors. I then tested again using $א = .01$.[62] The reasons for doing this were firstly, that my $א = .001$ test had very little power, and probably failed to pick up many true clutch and/or choke hitters, and secondly that perhaps clutch/choke effects existed among more than just a few players but were too small in magnitude to be accurately picked up at the very stringent $א = .001$ level, and perhaps this was the reason why they had not been previously detected. The results at the $א = .01$ level not only yielded more significant results, but also painted a clearer picture about the sensitivity of my results to my assumptions on sacrifice flies and errors. My results will be discussed in full in section 5. The necessity of a small value of א was my reason for excluding players with one year of data. For such a player the solution for α is $\alpha = א$ when $א = .001$ this makes finding a true clutch or choke hitter difficult even if someone were one, because it would require the effect to be much

[62] Both of these are more stringent than the standard .05 significance level.

larger than it would likely be even if it were present.[63] With two years, both must be in the same tail so already it's much better. For example, for $\aleph = .001$ a two year career already requires $\alpha \approx .044$ instead of the excessively stringent .001 for a one year significance level. When $\aleph = .01$, it is more likely that a true clutch or choke hitter who played one year would show an effect that is that extreme. However, it is also more likely that a pressure neutral player would. Especially, if that player had just a little over 100 PA's, in which case it is likely that he had just a handful of high-importance plate appearances, and that our whole determination about if he was clutch/choke would be unduly influenced by a only a few data points. Therefore, we require at least 2 years of data for a player to be included.

5. Discussion and analysis of my results.

This section discusses my findings and is meant to be read while referring to the appropriate tables at the end of the paper. Please note that much of the discussion in 5.2 is speculative.

5.1 The Main results: Does Clutch Hitting exist? Does Choke hitting?

In Table 1, which contains, my main results for the existence of clutch hitting, there are 12 different sets of results. The first 6 give the results for $\aleph = .001$, for each of the 6 possible combinations of how I treated sacrifice flies and errors, the last 6 do the same thing for $\aleph = .01$. For purposes of the table "count as out" means use an OPS entry of 0, "count as walk" means use an OPS entry of 1, and "exclude" means leave these points out of the regression altogether. The generally accepted standard for considering results significant is a p-value<.05, and that is what I will mean when I say significant. Often times when a test is done multiple ways people lower the required p-value because there are many chances to get the required p-value making it more likely that this will occur by chance. The most well known, and most conservative, of these is the Bonifarri multiple comparison method, which says that if a significance level of .05 is desired, then our requirement for any one of the tests to be declared significant is

[63] In order for us to detect a true clutch hitter's clutch effect as significant in one season he would need to hit like a journeyman most of the time, but when the game is on the line he hits like Barry Bonds. Such a large effect would strain credulity, and hence I won't even bother looking for it.

p-value$< \frac{.05}{\text{number of tests}}$. I think in this case such a multiple comparison method is not only unwarranted, but actually less accurate. This is because this method is meant for comparing analyses on data sets that, even if not independent, are very different. Here the overlap is enormous, sacrifice flies and errors combined constitute less than 2% of all plate appearances qualifying to be in the regression.[64] So even when we compare two treatment combinations that differ in their treatment of both sacrifice flies and errors, more than 98% of the data points are still the same. Therefore, a multiple comparison method will vastly overstate how much we should raise our standards of evidence. In reality, we probably ought to require something slightly stricter than p-value<.05, but rather than create a formalized process for this I will offer a caveat to treat any borderline significant results with a slightly higher degree of suspicion, since the data sets are at least 98% the same. As far as running the model for $\aleph = .001$ and $\aleph = .01$, this might require some additionally tightening of standards, but here the data is identical and only one of the parameters of our testing procedures changes, so here too a multiple comparison method seems unnecessary. Perhaps a little more suspicion of borderline results is in order though.

In the portion of the table for $\aleph = .001$, each time we exclude sacrifice flies our results are insignificant, in fact they are not even close. Similarly, when $\aleph = .01$, each time we exclude sacrifice flies we get results that are not nearly significant. However, when we include sacrifice flies as walks, the numbers tell a very different story. In the $\aleph = .001$ portion of the table, when we include sacrifice flies as walks, 2 out of our 3 results are significant (p-values: .00492, .0239) and the 3rd is somewhat close to being significant (p-value: .0945.) In the $\aleph = .01$ part of the table, when we include sacrifice flies as walks, all three of our results are highly significant (p-values: .000724, .0000197, .00709.)[65] While the evidence provided by the p-values for existence of clutch and/or choke hitters is convincing when sacrifice flies are included, especially when $\aleph = .01$, our results are clearly sensitive to our inclusion of sacrifice flies in the regression. However, if one believes that sacrifice flies are a valid component of offensive production, then

[64] This is a percentage of PA's qualifying when errors and sacrifice flies are included. The actual figure is 1.95%, with sacrifice flies constituting 0.79% of PA's and errors constituting 1.16% (figures rounded to nearest 0.01%).
[65] The p-values cited here are rounded to the nearest value of the 3rd non-zero digit. Also, it should be noted for a few elite players whose OPS is actually above 1.000, that a sacrifice fly or a walk in a key situation would actually contribute to pulling their regression in the direction of choke. This is not a weakness for

this sensitivity is not necessarily a weakness in the argument for clutch hitting. If a hitter has ability to execute a sacrifice fly more frequently in key situations than in relatively unimportant ones, it is perfectly sensible to consider that a valid component of clutch hitting, because it is one part of an overall trend of increased offensive contribution at times of increased importance.[66] Therefore, the sensitivity to this assumption merely means that sacrifice flies are a more important part of the clutch hitting story than we might have initially thought. Another interesting result, is that while the inclusion of sacrifice flies as walks does not increase the number of choke hitters, it greatly increases the number of clutch hitters. In the case where $\aleph = .01$, the inclusion of sacrifice flies causes clutch hitters to become the dominant effect, and the number of choke hitters are so few that their appearance could easily be due to chance variation, and the number of clutch hitters alone to become pretty convincing evidence. Another, more minor, sensitivity to assumptions seems probable. At both $\aleph = .001$ and $\aleph = .01$, when we include sacrifice flies as walks the highest (least convincing) of the 3 p-values amongst the 3 error assumptions is counting errors as walks. For the other 2 error assumptions a different one has a higher p-value for each of the 2 values of $\aleph$. This pattern might indicate a slight sensitivity to giving some positive value to reaching on an error, but it might just be chance variation. In either case, this sensitivity, if it isn't just chance variation, is relatively minor, especially in comparison to the large sensitivity to our treatment of sacrifice flies.

When we use $\aleph = .01$ and include sacrifice flies the evidence that clutch hitting exists is tremendous. But how prevalent is this effect? If clutch/choke effects didn't exist then we could expect that, on average, 10.75 people in our sample would show up as clutch or choke. Since, in reality, fractional people can't do much of anything, we round this figure to 11. If we include the choke hitters we have about 2% of the population (over the 3 sets of assumptions and parameters being discussed; depending on which case you pick it is between 1.86-2.51%.) Excluding the

[66] It should be noted that for a few elite players whose OPS is actually above 1.000, that a sacrifice fly or a walk in a key situation would actually contribute to pulling their regression in the direction of choke. This is because, for these exceptional hitters, this is actually lower than their average production, so they are actually performing worse, not better in a key situation. Even so, the overwhelming majority of hitters have OPS values below 1, and a sacrifice fly or walk in a key situation pulls their regression in the clutch direction.

chokers, it is still close to the 2% mark, just tending more towards the low side. If everyone in our sample were pressure neutral 1% hitters would show up clutch or choke on average. Also note that it is only the pressure neutral hitters that have a 1% chance of showing up clutch or choke, and that they have an equal chance of showing up either. Because the sample is so large, the percentage of true pressure neutral hitters that show up clutch or choke should be very close to 1% by the normal approximation to the binomial. We can therefore estimate what percentage of our results were false positives; it is a little less than 1%. So therefore a lower-bound estimate of the number of true clutch hitters[67] is about 1% of hitters in the majors. Alternately, the percentage of pressure neutral hitters that we expect show up clutch only is .5%. If we subtract that from the percentage of hitters who showed up clutch only, we get a lower bound closer to 1.5%. However, just as our test can mistakenly pick up clutch hitters, it can also mistakenly miss them. I do not know of a way to estimate what percent of players our test labeled pressure neutral when they were really clutch or choke, but it seems incredibly unlikely that our test didn't miss a single clutch hitter. Therefore, the 1% figure is a low-end estimate, and we can assert fairly confidently that the proportion of true clutch hitters in baseball is at least 1%.

5.2 Clutch and Choke hitters of Note: The Real McCoy's

As our p-values show, when sacrifice flies our included in the regression as walks, we can safely reject the null hypothesis that there are no clutch or choke hitters in our sample. It is extremely unlikely that so many pressure neutral hitters would come up as either clutch or choke by chance. Nevertheless, it is also very unlikely that all of those observed as clutch or choke are true clutch or true choke hitters. This is because, as we mentioned above, we

[67]Since the number of chokers is small and likely due to chance, we will mostly disregard them here.

would expect about 11 hitters to show up clutch or choke even if no one was. Even if we picked up every single true clutch or choke hitter, the chances are somewhere in the neighborhood of 11 pressure neutral players would also show up as clutch or choke. So while we are confident clutch hitters exist, we can not be so confident about which individuals are actually clutch or choke. We are confident that at least some of the players we found to be clutch hitters are indeed true clutch hitters, and possibly that some of those we found to be choke hitters are indeed true choke hitters, but we can not be so confident about which of the players we reported to be clutch or choke are one of the ones who actually are clutch or choke.

Despite a certain degree of uncertainty as to which of the hitters we found to be clutch or choke really are, there are some players whose results provided more powerful evidence than others. In light of this, we will now discuss some of the players reported to be clutch or choke who seem more likely to, actually be so. There was no one in our sample who came out clutch using one treatment of errors and sacrifice flies, and choke under another. Table 2 lists all players who come up clutch or choke in at least under at least one of the 12 sets of assumptions and significance levels listed in Table 1. As well as whether they were clutch or choke, and in how many of the 12 were they reported as such. Thus, table 2 gives us an idea of how robust a player's result of clutch/choke was to changes in significance levels and to changes in treatments of sacrifice flies and errors.[68] When we use the stricter .001 career significance level, clutch hitters are only found when sacrifice flies are included, only 2 men

[68] To see which ones of the 12 sets a player showed up clutch or choke in, see tables 3-14 for a lists of players who showed up clutch and choke for each set.

appear as clutch for all 3 possible treatments of errors: Eddie Murray and Bill Buckner.[69] Ironically, Buckner is best known for his fielding error in game 6 of the 1986 World Series that cost the Red Sox the world championship.[70] This, however, was not an instance of choking at the plate; our study does not encompass fielding. In addition, a single blunder in a key situation is anecdotal, not statistical evidence. Eddie Murray is by far the most robustly clutch hitter in our sample; he is reported as clutch in 9 of the 12 sets of results. Not only is he reported as clutch in the 3 instances mentioned above, but he also is reported as clutch at the .01 significance level for all 6 sets of assumptions. He is the only player to show up clutch more than 6 times. Therefore, we are fairly confident that Eddie Murray is a true clutch hitter.

Buckner shows up as clutch whenever we include sacrifice flies, at both the .001 and .01 levels, for a total of 6 times. When we don't include sacrifice flies though, his results become inconclusive. He is the only other player Besides Eddie Murray to show up clutch at the .001 level 3 times, so he is one of the players that we are more sure is a true clutch hitter.

There are 3 other players who show up clutch 6 times: Darryl Hamilton, Frank Duffy, and Luis Gomez. All three of these players are reported as clutch hitters at the .01 significance level, but not at the .001 significance level. Yet, at the .01 significance level they are clutch regardless of our treatment of errors and sacrifice flies. Darryl Hamilton and Luis Gomez each play 4 qualifying seasons. They are both significantly clutch in 3 of them and not significantly choke in any of them. Frank Duffy plays 5 qualifying seasons, is significantly clutch in 3 of them and is not significantly choke in any of them. These 3

[69] A 3rd hitter, Craig Reynolds, appears as clutch at the .001 for only one of the 6 combinations of sacrifice fly and error treatments.
[70] However, it was not the sole cause of that collapse, just the one best suited for a highlight reel.

players are also, among those who we are more confident are true clutch hitters. The robustness of their clutch designation to changes in error and sacrifice fly assumptions is encouraging. That they don't show up as clutch at the .001 significance level is not as discouraging as it seems. They all had 4 or 5 year careers, and when $\aleph = .001$, the α for careers of 4 and 5 years are .075 and .101 respectively.[71] It is likely that these players possess true clutch effects, but that these effects are of too small a magnitude to be picked up at those significance levels.

Besides Buckner and Murray, the only player to appear clutch at the .001, significance level, is Craig Reynolds. Reynolds appears clutch once at the .001 level and 4 times overall. At the .001 level he appears clutch only when sacrifice flies count as walks and errors are excluded. At the .01 level he appears clutch whenever sacrifice flies count as walks, regardless of how we treat errors. Reynolds might be appearing clutch by chance alone or he might be a true clutch hitter, it's tough to say.

Only 2 players appear as clutch 5 times: Lee May and Jose Canseco. Both appear as clutch at the .01 significance level in for every set of assumptions except for exclude sacrifice flies and count errors as walks. As I mentioned earlier, out of the 3 possible error assumptions counting them as walks is the only one that seems to mask clutch and choke effects. So when sacrifice flies are excluded at the .01 level these two players are borderline clutch. It's possible these guys appear clutch by chance, but I would still put them among those that we are more sure are true clutch hitters.

[71] From Table 16

Out of the 30 players that appear clutch at least once, 15 appear 2 times or fewer. These 15 players are those that seem more likely to be true pressure-neutral hitters that appear as clutch or choke by chance, as we would expect some to do, although which ones in fact are clutch is tough to say.

There was one hitter who showed up as choke all 12 times: Joe Strain. He was probably a choke hitter. But, since he only had 2 qualifying seasons in the sample, meaning that he appeared choke in both his seasons all 12 times, there is some chance that he was a pressure neutral hitter who showed up choke by coincidence. Jeff Stone showed up choke 10 times, the second most. The only two sets of assumptions he was not reported as choke were the two sets at the .001 significance level that counted errors as walks. All 10 times that he was reported as choke he was significantly choke 3 out of his 4 seasons, and significantly clutch in none of them. Since we got so few choke hitters it is tough to say if these guys were actually choke, they could be the ones we'd expect to see by chance. However, the fact that they are less sensitive to our assumptions than any of our clutch hitters, would tend to make me think it likely that they are true choke hitters.

One other choke hitter, Rich Schu, is also worth mentioning, he comes up significantly choke at the .01 significance, under all 6 sets of assumptions, and not at all at the .001 level. Schu is significantly choke in 4 of his 6 seasons all 6 times he comes up choke. Most of the time one of his seasons comes up significantly clutch as well. Due to the low number of choke hitters, and the fact that we'd expect 1 or 2 chokers to show up by chance, I am not so confident that Rich Schu is a true choke hitter.

Although the data we based this sub-section's discussion on were obtained through the detailed statistical analysis described in sections 3 and 4 of the paper, the discussion itself is largely speculative. The charts provided in the back should allow the reader to draw his/her own conclusions. I did not include the season-by-season break down of parameter estimates and p-values of all players who came up significant, because those tables would have been so long that they would have been unreadable to any but the most obsessed reader.

6. Concluding Remarks and Possibilities for Further Research.

Although the number of choke hitters found in the Major Leagues are not very different from what one would expect to occur by chance, the number of clutch hitters are. When we consider sacrifice flies as one piece of the overall picture of batting performance, we find the evidence in favor of the existence of true clutch hitters in the Major Leagues to be overwhelming. Although we use regression models to test for significant clutch or choke effects, we did not use them to model the magnitude of the effects, even though this has potential applications. This is because the noisy nature of the data would make the magnitudes unreliable indicators of the true magnitude of the effect. As difficult as it would be to develop more refined methods to determine with a higher degree of confidence if a player is a clutch hitter, it would be much harder to say "how clutch" a player is. (i.e. how much better does he get in key situations.) That said, I doubt the magnitude of these effects are too large, because if they were, the existence of clutch hitting would have become statistical consensus long before I wrote this paper.

At the $\aleph = .01$ level, we get 20-something clutch hitters when we include sacrifice flies.[72] On average, a combined total of about 11 clutch hitters and choke hitters would be observed by chance, when no such effects existed in the population.[73] Hence, some of the players we observed as clutch are probably pressure-neutral hitters who happened to do better in key situations. By the same token, there are probably true clutch hitters that our test mistakenly labels as pressure neutral. Determining which ones these are is an inexact science, and was the subject of some discussion in the previous section of this paper.

These results do not deny that people who crumble under pressure exist, but our findings suggest that these people tend not to become big league hitters; they probably go into lines of work that are more forgiving of such weaknesses. On the other hand, there do seem to exist Major League hitters who perform better in high pressure/importance situations. However, it seems likely that the overwhelming majority of major league hitters are pressure neutral.[74] This makes sense, because if a player is giving their best effort every time, then it would seem weird if they could do better at especially critical times. It seems likely that someone with the competitive drive to be a MLB player would be giving his best effort all the time.

But then what is the phenomenon that occurs in clutch hitters? Our study cannot directly say anything about the cause(s) of this phenomena, be it psychological, physiological, sociological or otherwise; we can only speculate as to what the underlying cause of this is. One possible explanation for this is that there are some hitters who have a level of effort or

[72] Depending on the error assumptions we get either 16,20, or 24 clutch hitters (See Table 1.) However, this averages out to 20 anyhow and they are all pretty close.
[73] Actually, the expected number of clutch and choke hitters that would be found by chance if no effect existed would be 10.75. Since we can only detect effects in whole numbers of players, I am using 11 since it is the closest whole number.
[74] Since we don't how many true clutch hitters our test failed to detect, we can't assert this with certainty. However, if most (or even a lot) major league hitters were clutch and the remainder were not choke, one would imagine that someone would have found clutch hitting in the League aggregate data by now.

concentration that is their best maintainable level of effort and another that is their best temporarily achievable level of effort. Put differently, they can give more effort than they sometimes do but they cannot keep that level up, whether for mental reasons (like level of concentration) or physical ones. Therefore, they play at their best maintainable level of effort most of the time and at their best temporarily achievable level of effort sometimes. Having to choose (probably subconsciously) times to put in the higher temporary level of effort, they would naturally tend to choose the times where this additional effort could greatest affect their team's chances of winning, which is essentially what my importance index calculates. Other explanations are also available. Perhaps some players play slightly lazily most of the time, but when the situation is important enough they decide (maybe sub-consciously) to put in more effort. There are many possible speculative explanations for why and how clutch hitters exist. This study is unable to lend support to one of them over another.

While this study examines the existence of clutch and choke hitting, it still leaves many unanswered questions for future research. One such topic is the development of methods of identifying clutch or choke hitters that are more accurate (or at least have more quantifiable accuracy) on an individual level.

Another possible topic is a follow-up to the lack of evidence of choke hitters' existence in MLB. We all know people in our everyday lives who under-perform when under pressure, yet we can not find any evidence that any of these people make it into an MLB batter's box. The question is: Where are these people getting weeded out? I would imagine that a choke hitter who is otherwise talented enough to reach the Majors, would still be a starter on his high school team. What about a division 1 college though? If we find that choke hitters are present at these lower levels of baseball, from which players are often drafted, then developing methods to identify

them could help MLB teams avoid wasting draft picks on these choke hitters, who are not only less likely to make the Majors, but are also less valuable in the unlikely event that they do.

There is no reason why future research using an importance index to investigate clutch or choke abilities has to be limited to hitting: We can use similar methods for pitching and fielding. Perhaps an appropriately modified importance index might be used, but the idea would be the same. Instead of the dependent variable being OPS, it could be something else. For fielding, the most simplistic thing to use as a dependent variable is error or no error, but there are many problems with this. For one, errors are sparse, making clutch or choke effects tough to pick up. Another problem is that errors are an imprecise way of measuring fielding performance. Other, more complicated, methods have been developed by others, and I will not go into them now. A dependent variable based such methods might be an appropriate way to measure clutch/choke fielding. Similarly, using a dependent variable appropriate to measure pitching performance (on a batter-by-batter basis) we could attempt to measure clutch pitching.[75] Concluding that clutch hitters exist in MLB hardly ends the debate on clutch hitting. It actually raises more questions than it answers. Where are all the choke hitters hiding? Are they flying under the statistical radar in the Majors? Are they languishing in the minors? Did they peter out in college? High school? Pony league? Who are the clutch hitters? How can we better identify them? And before we get too excited about all this, does it even matter that there are clutch hitters? Maybe they improve by so little that it shouldn't affect personnel or managerial decisions at all. If clutch hitting should affect strategy or personnel decisions, how should it do so? For now, I leave all these and more as open questions.

---

[75] If the dependent variables in either case were continuous (rather than discrete as in OPS) then a different regression technique would need to be used.

Appendix A: Formulas for some basic baseball statistics

$$\text{On - base Percentage} = \frac{\text{hits} + \text{walks} + \text{hit by pitch}}{\text{At Bats} + \text{walks} + \text{hit by pitch} + \text{Sacrifice flies}}$$

$$\text{Slugging Percentage} = \frac{\text{singles} + 2 * \text{doubles} + 3 * \text{triples} + 4 * \text{Homeruns}}{\text{At Bats}}$$

Table 1: Main Results.

This table contains the results from my Gompertz regression. The 1st column shows the career significance level used to test each player in the sample of 1075 players. The 2nd and 3rd columns show our treatment of sacrifice flies and errors respectively. The 4th 5th and 6th columns show how many players were reported as significant for either clutch or choke ability, and how many of these were clutch and how many were choke. The final column shows the p-value of existence: The probability that, if there were no true clutch or choke hitters in our sample that we would have at least as many players be reported as significantly clutch or choke as we actually did, at the given career significance level. A very low p-value of existence makes it very likely that there are at least some true clutch and/or choke hitters in our sample, for if there weren't such an extreme result would be highly unlikely.

career significance level	sacrifice fly assumptions	error assumptions	significant players	clutch	choke	p-value existence
0.001	exclude	exclude	2	0	2	0.29182090
0.001	exclude	count as out	2	0	2	0.29182090
0.001	exclude	count as walk	1	0	1	0.65888577
0.001	count as walk	exclude	5	3	2	0.00491641
0.001	count as walk	count as out	4	2	2	0.02386749
0.001	count as walk	count as walk	3	2	1	0.09450975
0.01	exclude	exclude	11	7	4	0.51041064
0.01	exclude	count as out	13	8	5	0.28361016
0.01	exclude	count as walk	11	4	7	0.51041064
0.01	count as walk	exclude	23	20	3	0.00072368
0.01	count as walk	count as out	27	24	3	0.00001972
0.01	count as walk	count as walk	20	16	4	0.00708777

In this chart, a yellow p-value indicates a result that is significant at the .05 significance level. An orange p-value indicates a result that is significant at the .01 significance level. Finally, a red p-value indicates a result that is significant at the .001 significance level. This may remind some of you of the Department of Homeland Security terror alert level system. However, unlike that system, these categories are well-defined, not vague. In addition, I sincerely hope that you will find this color-coding system somewhat less ominous.

Table 2: Overall list of Players with Significant Results

This table lists all players who showed up significantly clutch, and all players who showed up significantly choke under at least one set of sacrifice fly assumptions, error assumptions, and career significance levels. There were 2 alternate significance levels, 2 alternate sacrifice fly assumptions and 3 alternate error assumptions, making for 12 different sets. Each line of the table lists the player's name, whether the player was significantly choke or significantly clutch (no one came up significantly clutch in one of the 12 sets and significantly choke in another so no player's name appears twice on this list.) and finally in how many of the 12 sets the player in question showed up as significant. Tables 3-14 give separate lists of significant players for each of the 12 sets, to show which of the sets these players were significant in. However, this table only says under how many sets each player was significant.

name	result	number of times appear significant (out of 12)
Eddie Murray	clutch	9
Bill Buckner	clutch	6
Darryl Hamilton	clutch	6
Frank Duffy	clutch	6
Luis Gomez	clutch	6
Jose Canseco	clutch	5
Lee May	clutch	5
Craig Reynolds	clutch	4
Garth Iorg	clutch	4
Alan Trammell	clutch	3
Gary Ward	clutch	3
Kent Hrbek	clutch	3
Phil Garner	clutch	3
Scott Fletcher	clutch	3
Bernard Gilkey	clutch	2
Dane Iorg	clutch	2
Hubie Brooks	clutch	2
Mike Easler	clutch	2
Scott Bradley	clutch	2
Denny Doyle	clutch	1
Jay Johnstone	clutch	1
Jim Dwyer	clutch	1
Oscar Gamble	clutch	1
Pete Rose	clutch	1
Ray Knight	clutch	1
Rickey Henderson	clutch	1
Rico Petrocelli	clutch	1
Rod Scott	clutch	1
Tim Raines	clutch	1
Joe Strain	choke	12
Jeff Stone	choke	10
Rich Schu	choke	6
Sam Horn	choke	4
Joey Cora	choke	1
Lance Parrish	choke	1
Leon Lacy	choke	1
Rex Hudler	choke	1

Tables 3-14: Significant players under each set of assumptions

Each of tables 3-14 lists players who were significantly clutch or choke when using a given set of career significance level, error assumptions, and sacrifice fly assumptions. Each line of these tables first lists the name of the player. The next field gives the result, that is, whether the player was found to be clutch or choke. The next three fields list, the number of qualifying seasons that a player has in our sample, the number of seasons that the player was found to be significantly clutch and the number of seasons that the player was found to be significantly choke.

Table 3

This table uses a career significance level of .001, and it excludes both errors and sacrifice flies from the regression.

name	result	years career	years clutch	years choke
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2

Table 4

This table uses a career significance level of .001. It excludes sacrifice flies from the regression and counts errors as an OPS entry of 0, like an out.

name	result	years career	years clutch	years choke
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2

Table 5

This table uses a career significance level of .001. It excludes sacrifice flies from the regression and counts errors as an OPS entry of 1, like a walk.

name	result	years career	years clutch	years choke
Joe Strain	choke	2	0	2

Table 6

This table uses a career significance level of .001. It excludes errors from the regression and counts sacrifice flies as an OPS entry of 1, like a walk.

name	result	years career	years clutch	years choke
Bill Buckner	clutch	16	9	2
Eddie Murray	clutch	16	9	2
Craig Reynolds	clutch	13	7	4
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2

Table 7

This table uses a career significance level of .001. It counts errors as an OPS entry of 0, like an out, and counts sacrifice flies as an OPS entry of 1, like a walk.

name	result	years career	years clutch	years choke
Bill Buckner	clutch	16	9	2
Eddie Murray	clutch	16	9	2
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2

Table 8

This table uses a career significance level of .001. It counts errors as an OPS entry of 1, like a walk, and counts sacrifice flies as an OPS entry of 1, like a walk.

name	result	years career	years clutch	years choke
Bill Buckner	clutch	16	9	2
Eddie Murray	clutch	16	9	2
Joe Strain	choke	2	0	2

Table 9

This table uses a career significance level of .01. It excludes both errors and sacrifice flies from the regression.

name	result	years career	years clutch	years choke
Jose Canseco	clutch	8	5	1
Frank Duffy	clutch	5	3	0
Luis Gomez	clutch	4	3	0
Darryl Hamilton	clutch	4	3	0
Sam Horn	choke	4	0	3
Garth Iorg	clutch	8	5	0
Lee May	clutch	8	5	1
Eddie Murray	clutch	16	9	3
Rich Schu	choke	6	1	4
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2

Table 10

This table uses a career significance level of .01. It counts errors as an OPS entry of 0, like an out, and excludes sacrifice flies from the regression.

name	result	years career	years clutch	years choke
Jose Canseco	clutch	8	5	0
Joey Cora	choke	4	0	3
Frank Duffy	clutch	5	3	0
Scott Fletcher	clutch	10	6	1
Luis Gomez	clutch	4	3	0
Darryl Hamilton	clutch	4	3	0
Sam Horn	choke	4	0	3
Garth Iorg	clutch	8	5	0
Lee May	clutch	8	5	1
Eddie Murray	clutch	16	9	3
Rich Schu	choke	6	0	4
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2

Table 11

This table uses a career significance level of .01. It counts errors as an OPS entry of 1, like a walk, and excludes sacrifice flies from the regression.

name	result	years career	years clutch	years choke
Frank Duffy	clutch	5	3	0
Luis Gomez	clutch	4	3	0
Darryl Hamilton	clutch	4	3	0
Sam Horn	choke	4	0	3
Rex Hudler	choke	5	0	3
Leon Lacy	choke	13	2	7
Eddie Murray	clutch	16	9	3
Lance Parrish	choke	15	0	8
Rich Schu	choke	6	1	4
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2

Table 12

This table uses a career significance level of .01. It excludes errors from the regression and counts sacrifice flies as an OPS entry of 1, like a walk.

name	result	years career	years clutch	years choke
Scott Bradley	clutch	6	4	0
Hubie Brooks	clutch	12	7	1
Bill Buckner	clutch	16	9	3
Jose Canseco	clutch	8	5	1
Frank Duffy	clutch	5	3	0
Mike Easler	clutch	8	5	0
Scott Fletcher	clutch	10	6	1
Oscar Gamble	clutch	12	7	0
Phil Garner	clutch	13	7	1
Bernard Gilkey	clutch	2	2	0
Luis Gomez	clutch	4	3	0
Darryl Hamilton	clutch	4	3	0
Kent Hrbek	clutch	11	6	0
Dane Iorg	clutch	8	5	2
Garth Iorg	clutch	8	5	0
Lee May	clutch	8	5	0
Eddie Murray	clutch	16	11	3
Craig Reynolds	clutch	13	7	4
Rich Schu	choke	6	1	4
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2
Alan Trammell	clutch	15	8	2
Gary Ward	clutch	10	6	1

Table 13

This table uses a career significance level of .01. It counts errors as an OPS entry of 0, like an out, and counts sacrifice flies as an OPS entry of 1, like a walk.

name	result	years career	years clutch	years choke
Scott Bradley	clutch	6	4	0
Bill Buckner	clutch	16	9	2
Jose Canseco	clutch	8	5	0
Denny Doyle	clutch	4	3	0
Frank Duffy	clutch	5	3	0
Jim Dwyer	clutch	15	8	1
Scott Fletcher	clutch	10	7	1
Phil Garner	clutch	13	7	1
Bernard Gilkey	clutch	2	2	0
Luis Gomez	clutch	4	3	0
Darryl Hamilton	clutch	4	3	0
Kent Hrbek	clutch	11	6	0
Dane Iorg	clutch	8	5	2
Garth Iorg	clutch	8	5	0
Jay Johnstone	clutch	9	5	0
Ray Knight	clutch	11	6	1
Lee May	clutch	8	5	0
Eddie Murray	clutch	16	10	3
Tim Raines	clutch	12	7	1
Craig Reynolds	clutch	13	7	4
Pete Rose	clutch	13	7	0
Rich Schu	choke	6	1	4
Rod Scott	clutch	5	3	0
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2
Alan Trammell	clutch	15	8	2
Gary Ward	clutch	10	6	1

Table 14

This table uses a career significance level of .01. It counts errors as an OPS entry of 1, like a walk, and counts sacrifice flies as an OPS entry of 1, like a walk.

name	result	years career	years clutch	years choke
Hubie Brooks	clutch	12	7	1
Bill Buckner	clutch	16	9	4
Jose Canseco	clutch	8	5	1
Frank Duffy	clutch	5	3	0
Mike Easler	clutch	8	5	0
Phil Garner	clutch	13	7	0
Luis Gomez	clutch	4	3	0
Darryl Hamilton	clutch	4	3	0
Rickey Henderson	clutch	14	8	0
Sam Horn	choke	4	0	3
Kent Hrbek	clutch	11	6	0
Lee May	clutch	8	5	0
Eddie Murray	clutch	16	10	3
Rico Petrocelli	clutch	3	2	0
Craig Reynolds	clutch	13	7	4
Rich Schu	choke	6	1	4
Jeff Stone	choke	4	0	3
Joe Strain	choke	2	0	2
Alan Trammell	clutch	15	8	2
Gary Ward	clutch	10	6	1

Table 15: Results of standard logistic model.

Before achieving success with the asymmetric Gompertz (A.K.A complementary log-log) regression model, a standard logistic regression, with its restrictive symmetry, was used but did not yield significant results. This table shows the results from that logistic model. With the logistic model, only a .001 career significance level was used, and it was tried with all 6 combinations of error and sacrifice fly assumptions. The table lists the names of significant players for each of the 6. Since this table is really 6 distinct lists, a double line is used to separate players significant under one set of assumptions from players significant under another. Since this table is structurally similar to a composite of tables 3-8, some players' names appear multiple times. This table is set up the same way as tables 3-14 except for the addition of the columns 'sacfly assumptions' and 'error assumptions' which denote the treatment of errors and sacrifice flies used here. This information was not included in tables 3-14 because it was already given in the captions preceding each table.

NAME	sacfly assumptions	error assumptions	result	years career	years clutch	years choke
Jeff Stone	exclude	exclude	choke	4	0	3
Joe Strain	exclude	exclude	choke	2	0	2
Jeff Stone	exclude	count as out	choke	4	0	3
Joe Strain	exclude	count as out	choke	2	0	2
Rich Schu	exclude	count as walk	choke	6	0	4
Joe Strain	exclude	count as walk	choke	2	0	2
Eddie Murray	count as walk	exclude	clutch	16	9	2
Jeff Stone	count as walk	exclude	choke	4	0	3
Joe Strain	count as walk	exclude	choke	2	0	2
Jeff Stone	count as walk	count as out	choke	4	0	3
Joe Strain	count as walk	count as out	choke	2	0	2
Rich Schu	count as walk	count as walk	choke	6	0	4
Joe Strain	count as walk	count as walk	choke	2	0	2

Table 16: Chart of single-season significance levels

This chart lists the single season significance level that is used as a function of the career significance level being used and the number of qualifying seasons that the given player has. The career significance level (A.K.A value of Aleph) is given by the column, and the number of seasons in the sample is given by the row.

Value of Aleph		0.001	0.01
Number of seasons	2	0.04472	0.14142
	3	0.02593	0.08280
	4	0.10130	0.22177
	5	0.07510	0.16566
	6	0.15697	0.28719
	7	0.12788	0.23541
	8	0.20632	0.33940
	9	0.17670	0.29211
	10	0.24910	0.38183
	11	0.22000	0.33862
	12	0.28618	0.41706
	13	0.25807	0.37741
	14	0.31852	0.44685
	15	0.29157	0.41029
	16	0.34697	0.47247
	17	0.32122	0.43856
	18	0.37220	0.49478
	19	0.34762	0.46320

Works Cited

German Rodriguez. Lecture notes for WWS509 Generalized linear models: A.2 Tests of hypotheses. Princeton University. URL: http://data.princeton.edu/wws509/notes/als2.html

Newsom

SAS 8.2 online Documentation. Probit Procedure. Overview. URL: http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap54/sect1.htm

David W. Hosmer and Stanley Lemeshow, Applied Logistic Regression. New York. Wiley, 2000.

Jahn H. Hakes and Raymond D. Sauer, "Are Players Paid for 'clutch'performance?" John E. Walker Dept. of Economics, Clemson University. Preliminary Draft. June 30, 2003.

D.A D'Esopo and B. Lefkowitz, "The Distribution of Runs in the game of baseball." In Optimal Strategies in Sports, 1977. Amsterdam; New York: North Holland Pub. Co., pp 55-62.

Mark D. Pankin, "Subtle Aspects of the Game" presented at the SABR XXIII conference, San Diego, CA, June 25, 1993.

Mark D. Pankin, "Do Base Stealers help the next batters?" Presented at the SABR XXXII conference, Boston, MA, June 27, 2002.

Jay Bennett, "Did Shoeless Joe Jackson Throw the 1919 World Series?" *The American Statistician*, Vol. 478 No. 4, Nov., 1993. pp. 241-250

George R. Lindsey, "An Investigation of Strategies in Baseball." *Operations Research,* Vol. 11, Issue 4, Jul.-Aug., 1963. pp. 477-501

Jim Albert and Jay Bennett, Curveball: Baseball, Statistics, and the Role of Chance in the Game Copernicus Books. New York. 2001

Bill James, The New Bill James Historical Baseball Abstract The Free Press. New York. 2001. pp. 348-350

Tom Verducci, "Does Clutch Hitting Truly Exist?" *Sports Illustrated.* Vol. 100 No. 14, April 5, 2004. pp 60-62.

George R. Lindsey, "The progress of the score during a baseball game." *Journal of the American Statistical Association,* Vol. 56, Issue 295, Sep. 1961. pp. 703-728

ENLEXICA-Baseball Dictionary. http://www.enlexica.com

Bill James, This time let's not eat the bones. Villard Books. New York. 1989 pg. 250

Michael Lewis Moneyball: the art of winning an unfair game. W.W. Norton and Company. New York. 2003

All data from: www.retrosheet.org and www.astrosdaily.net

Bill James, "Underestimating the Fog." *The Baseball Research Journal.* No. 33. The Society for American Baseball Research. Cleveland, OH. 2005. pg. 29-33