Applying quality control techniques to baseball statistics

Matthew Gardner

Traditional baseball projections generally take in to account years of data, whereas it is difficult to make inferences using data about players during the season for a variety of reasons. Monitoring processes of quality control are a convenient way to compare small portions of the season in an attempt to detect changes in the expected performance of individual players— especially in finding small shifts that are difficult to find otherwise. This paper focuses on exploring detecting those smaller shifts using the cumulative sum (CUSUM) control chart and comparing their usefulness to that of using Shewhart control charts. In addition, we discuss the choice of what a significant shift is, and what kind of expectation should be used in this framework.

Introduction

Russell Carleton's work on the stabilization of baseball statistics (Carleton, 2012a) demonstrates the unreliability of small samplings of a baseball player's performance, even as far as indicating no clear stabilization in some skills at all. Best used in retrospect, "they are not nearly as powerful in predicting future performance," Carleton (Carleton, 2012b) later said in regard to his research. Stemming from this idea, reliability of the data is the key issue that teams face when determining if a subset of performance is a significant change from what the expectation was.

Control charts are a methodology used in statistical process control to monitor a process, generally with some parameters by which it is required to be limited by. In a production line, this may be a technical limitation of a produced part, but in this application, we may think of it as detecting if a particular player is playing significantly different from a previous sample. Potentially violated in this framework is the assumption of independent samples. In the practical sense, there are a variety of reasons that this may be the case including: injuries, quality of pitching/hitting, weather effects, etc. M. H. Lee discusses applying the CUSUM control chart to cases where there are correlated samples (Lee, 2010).

In particular, their discussion on the cost minimization of a genetic algorithm is an enlightening take on optimal selection of the parameters associated with the CUSUM control chart. Additionally, they cite Neuhardt (Neuhardt, 1987), saying:

> ...(they) considered the effects of correlated data within a sample that have been defined for the purposes of statistical process control. Such correlation may arise if the grouping is accomplished because of simplicity in data collection, such as multiple but similar measurements on a single product or multiple station machines. The effect of correlated measurements within a sample is shown to increase the Type I error rate for $\bar{X}$ chart.

This is a particularly important detail, as the selection of parameters for the CUSUM control chart will then be more optimistic.

Chen and Hsieh define a multivariate control chart extension for Hotelling's T^2 with the addition of varying sample sizes (Chen and Hsieh, 2006). While this will not be discussed further in this paper, it is a natural extension of the univariate analysis, and would be better suited for CUSUM control charts that keep track of statistics that contain differently distributed components, such as WAR and various defensive metrics.

For dramatically large breakouts akin to the end of the Dead-ball era, there are many existing, less sophisticated tools to signal a large jump forward or backward in a player's statistics— it's just not that difficult to recognize that Babe Ruth hitting 29 home runs in 1919 is "out-of-control." Therefore, the goal is to develop a process to detect relatively small shifts, such that we could attempt to determine if Ichiro Suzuki's, a current baseball player who's career started in 2001), rising strikeout rate (K%) is an out-of-the-ordinary change, or just due random variation, and to determine what that change would be. This will be further explored in the **Implementation** section further down.

Materials and Methods

Baseball is a sport made up of individual *plate appearances* of a batter against an opposing pitcher. Each of these appearances results in a variety of outcomes, but all of which may be described as either a success (a hit, walk, etc.) or a failure (strikeout, groundout, etc.), with some additional measures of value as needed. Throughout an entire baseball season, a starting offensive player may be expected to reach 600 plate appearances, depending on health and the proportion of 162 games played.

Retrosheet is an online database of baseball statistics dating back to 1871. Specifically, it supplies play-by-play accounts of Major League Baseball games. The methodology and R code for pulling specific data and cleaning it can be attributed to Max Marchi and Jim Albert (2014) from their book cited in the references. Essentially, the "Chadwick expanded event descriptor" is used to read-in and paste together data from every plate appearance in every game in a given year. Available variables include everything from pitch sequences and batted ball types to runner information and weather.

To go further, we provide some details on the control charts that will be used. Similar to a confidence interval, the $\bar{X}$ -chart is centered around the expectation based on some prior data, with the variance also estimated from that data. In a baseball context, this could be based on a couple of things:

- Results from the previous year(s)
    - This may include either single year averages, weighted averages, or some other type of variation
- Projections for the current year
    - Possible rational for the performance of player being significantly different from another year
        - injury, positional adjustment, regression via age, etc
    - May or may not allow for an estimate of variance to be taken from the projection, so would substitute with an assumption of equal variance

These assumptions will also be used in the framework of the CUSUM. Then, the 3-sigma limits of the $\bar{X}$ -chart are obtained by,

$$\textit{Center line}\,(CL)=\bar{\mathrm{X}}$$

$$\text{Upper control limit}\,(UCL)=\mathrm{CL}+3\cdot\sqrt{{}^{1(1-p)}\!\big/_{\text{sample group size}}}$$

$$\text{Lower control limit}\,(UCL)=\max\left(0,\mathrm{CL}+3\cdot\sqrt{{}^{1(1-p)}\!\big/_{\text{sample group size}}}\right)$$

where the 3-sigma limits are expected to contain approximately 99% of the variation and *p* is the probability of a success of a given outcome.

$\bar{X}$ -charts, and the closely related R and S -charts, are more useful for detecting large shifts from the expectations, which could meaningfully be described in baseball as a “breakout.” Graphically, any point plotted outside of the defined control limits must either be explained by an anomaly, or assumed to be as a result of an out-of-control process. This would be good for easily identifying breakouts from a large number of players, though, the problem is, breakouts are not that hard to identify. As was identified in the abstract, we instead target finding smaller shifts in the CUSUM model.

Using the framework described in Montgomery (2013), we can attempt to fit small collections of data into a cumulative sum control chart.

The way the CUSUM works is, each sample is compared to an expected value, in the same way that the $\bar{X}$ -chart was. In addition to that, a reference value, K, is determined by how much a shift in the mean you want to track. These comparisons happen in sequence, so for a shift to be detected, a series of samples have to show some amount of difference that indicates that the sample population does not follow the expected population. In equation form, you keep track of a sum going toward the upper and lower limits:

$$C_i^+ = max\left(0, x_i - (\mu_0 + K) + C_{i-1}^+\right)$$

$$C_i^- = max\left(0, (\mu_0 + K) - x_i + C_{i-1}^-\right)$$

where either C_i^- or $C_i^+ > H$ indicates that the process has gone out of control. The upper and lower limits, H, are selected and dependent on the amount of power you want the determination to have, or the desired average run length (ARL). The ARL is just a measure of how powerful the detection process is, where an ARL of 1000 is associated with there being one wrong classification in 1000 samples. This is a particularly involved calculation process, that Montgomery (Montgomery 2013, pg 423) provides via the recommendation of Woodall and Adams (1993). The approximation described by Siegmund (1985) will be used below.

Implementation

Ichiro Suzuki is a Japanese baseball player that has been active in the major leagues from 2001-present. For the purposes of implementation the methodologies described above, we will examine his strikeout rate (K%) in the 2011 season. Referenced figures are available in the section **Figures** after **References.**

Figure 1 and *Figure 2* are Shewhart $\bar{X}$ -charts that use the same 2011 strikeout rate data. The difference is that *Figure 1* uses the prior year as its estimation of strikeout rate and variance, whereas *Figure 2* uses 2001 (Ichiro's career-best year for this metric). Another observation is that the rational sample sizes are 30 (which will be contrasted in less detail in *Figure 6)*. You'll notice that neither of these charts detect a large shift, though *Figure 2's* data has a noticeably large amount of samples above the expectation. This is the case when you expect you may see something on a CUSUM control chart.

Choosing a *k* value of $1/2$ (looking for a shift of 1 standard deviation) and a *h* of 5, we calculate the CUSUM limits for both groups in *Figure 4* and *Figure 5*. Using Siegmund's approximation, we get an in-control ARL of 461.1, which is fairly reliable, with the note that it may be less so given correlation between rational groups. We also see the major difference between the basis of 2010 and 2001: *Figure 5* goes out-of-control. Stemming from the observation made that there was a large amount of points plotted above the expectation, we see that they accumulate to the point of going out-of-control.

In fact, if we looked at year-to-year strikeout rate of Ichiro, we could see that his strikeout rate had leveled out between 9 and 11 through the 2011 season, so it would not be appropriate to to use 2001 data. Regardless, this demonstrates a control technique that could have been used to monitor Ichiro in 2014, when his K% was a career-high 17.7.

Additional observations come from *Figure 3*, where we see that Ichiro's K% is roughly normally-distributed, confirmed by the Shapiro-Wilk test. This validates the assumption of roughly equally distributed data, despite interdependence.

Figure 6 demonstrates the consequences of changing the rational group size. This goes back to the concept introduced in the introduction by Carleton, where, the smaller the rational sample size, the less reliable the measurement will be. Most importantly, we see that this smaller drawing (of roughly 5 games, rather than 7.5) results in a large shift detected in what would be *Figure 2*.

Discussion and Summary

To begin with, there isn't anything even close to this that has been applied in the sabermetric community. That's not to say that it hasn't been applied in baseball, but if it has been, I haven't found it. To that point, I do not believe that this technique is particularly useful in its present state. It takes a lot of computational power to apply the framework defined in this paper to split, calculate, and continuously monitor groups, especially when you don't have a good way to select optimal rational groups.

That being said, it does have potential as a scouting tool. One of the biggest concerns mentioned in this paper was the interdependence of rational groups. To work off of that concern, there is also a worry that, while there could be a shift in the mean detected relatively early on, there will always be an opposition that is equally trying to make adjustments to your adjustments, potentially throwing out any conclusions you make.

This would be something more apparent the more you work with more data, which is something I intend to do, though goes beyond the scope of this exploration. To potentially counter this, the multivariate framework that I brought up from Chen and Hsieh could account for changes in approach with equal results, but there are also cases of one-skill players that succeed for awhile, then are figured out and never return to baseball.

Beyond that, implementing exponentially weighted moving average (EWMA) charts could be a benefit, though in practice does not make that much of a difference. Nonparametric control charts could be an alternate approach as well, but that would be a separate framework from this.

References

Carleton, Russell A. (2012a). *It's a Small Sample Size After All*. Retrieved from Baseball Prospectus website: http://www.baseballprospectus.com/article.php?articleid=17659

Carleton, Russell A. (2012b). *It Happens Every May*. Retrieved from Baseball Prospectus website: http://www.baseballprospectus.com/article.php?articleid=17742

Chen, Yan-Kwang, Hsieh, Kun-Lin (2007). *Hotelling's T 2 charts with variable sample size and control limit*. European Journal of Operational Research, 182(3) (pp. 1251-1262). doi: 10.1016/j.ejor.2006.09.046

Lee, M. H. (2010). *Economic Design of Cumulative Sum Control Charts for Monitoring a Process with Correlated Samples.* Communications in Statistics – Simulation and Computation, 39 (pp. 1909-1922). doi: 10.1080/03610918.2010.524333

Marchi, M., Albert, J. (2014). *Analyzing Baseball Data with R*. Boca Raton, FL: CRC Press.

Montgomery, Douglas C. (2013). *Statistical Quality Control, 7th Edition* (pp. 414-429) Hoboken, NJ: John Wiley & Sons, Inc.

Retrosheet: http://www.retrosheet.org

Neuhardt, J. B. (1987). *Effects of correlated sub-samples in statistical process control.* IIE Transactions 19(2):208–214.

Siegmund, D. (1985). Sequential Analysis: Tests and Confidence Intervals, Springer-Verlag, New York.

Woodall, W. H., and B. M. Adams (1993). *The Statistical Design of CUSUM Charts.* Quality Engineering, Vol. 5(4), pp. 559–570.

R code beyond what is outlined in the *Appendix* is available at: http://github.com/m-b-gardner/sabr

Figures

Figure 1: Ichiro - K% (2011) - Based on 2010

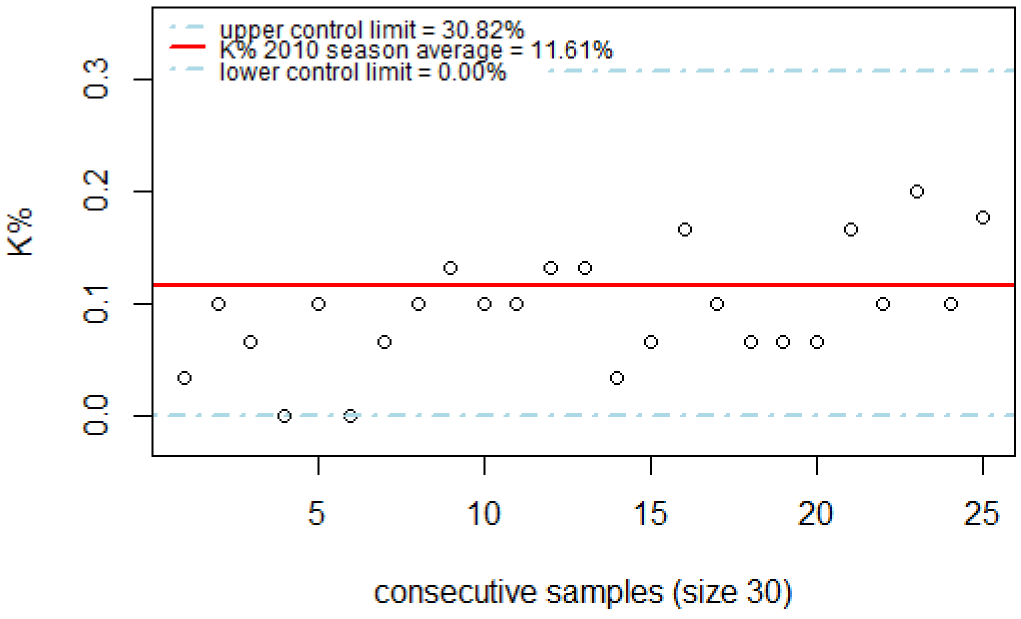


Figure 2: Ichiro - K% (2011) - Based on 2001

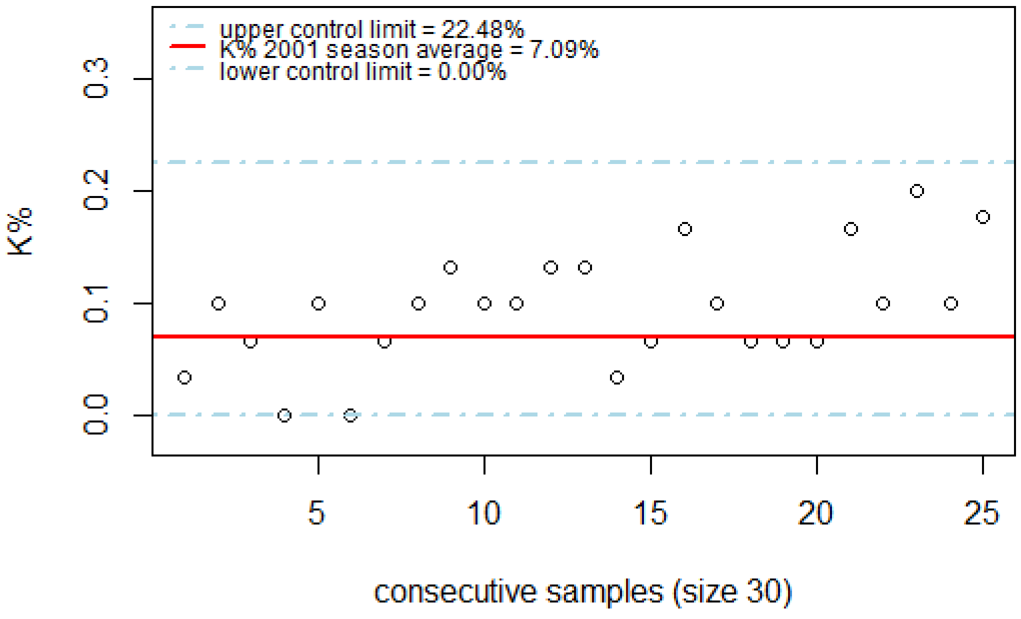


Figure 4: CUSUM - Ichiro - K% (2011) - Based on 2010

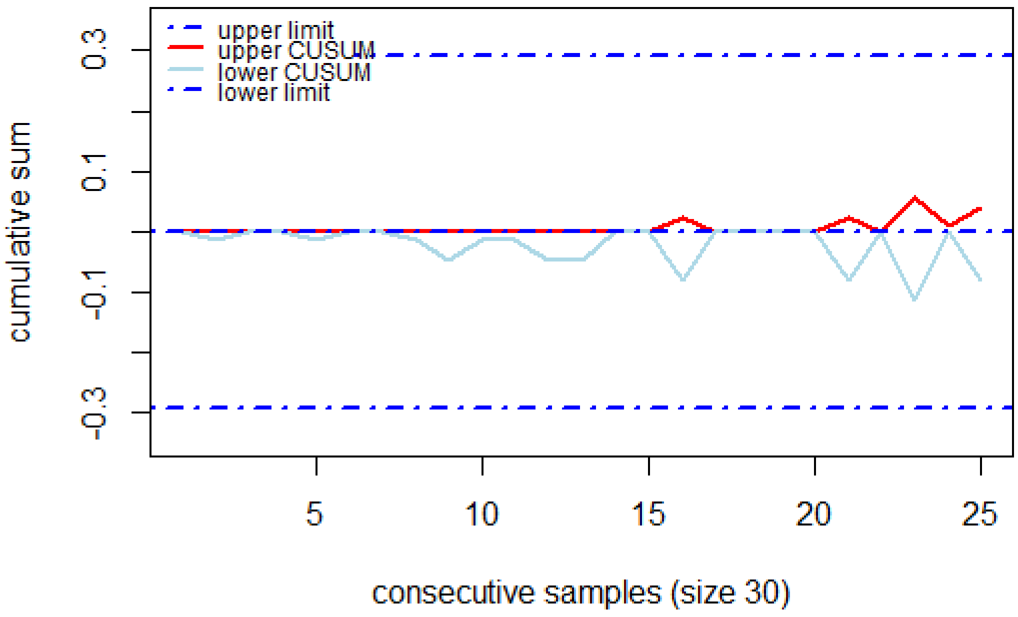


Figure 5: CUSUM - Ichiro - K% (2011) - Based on 2001

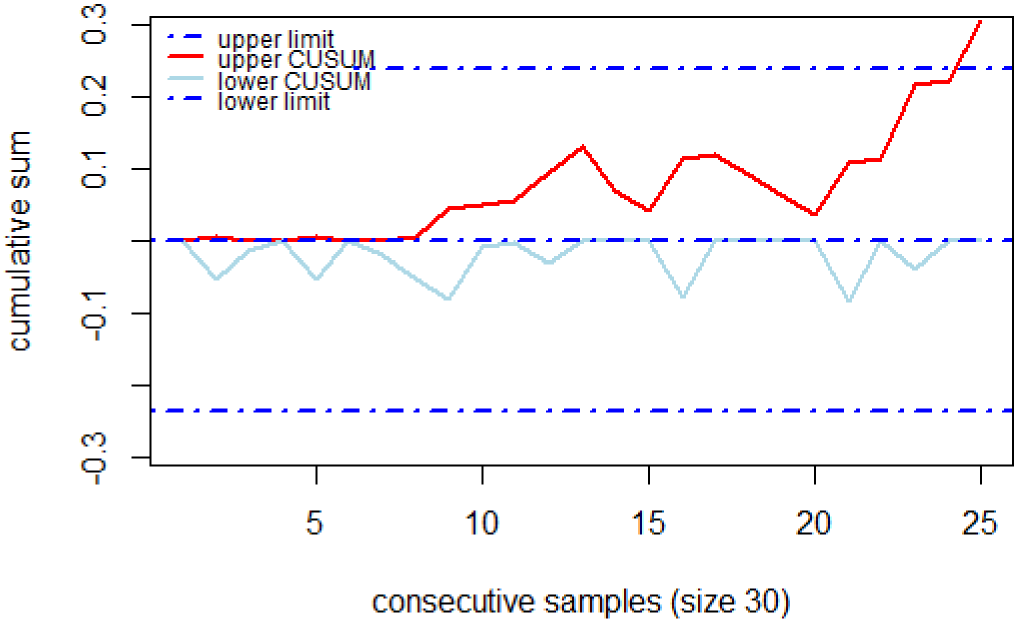


Figure 3: Normal Q-Q Plot: Ichiro - K% (2011)

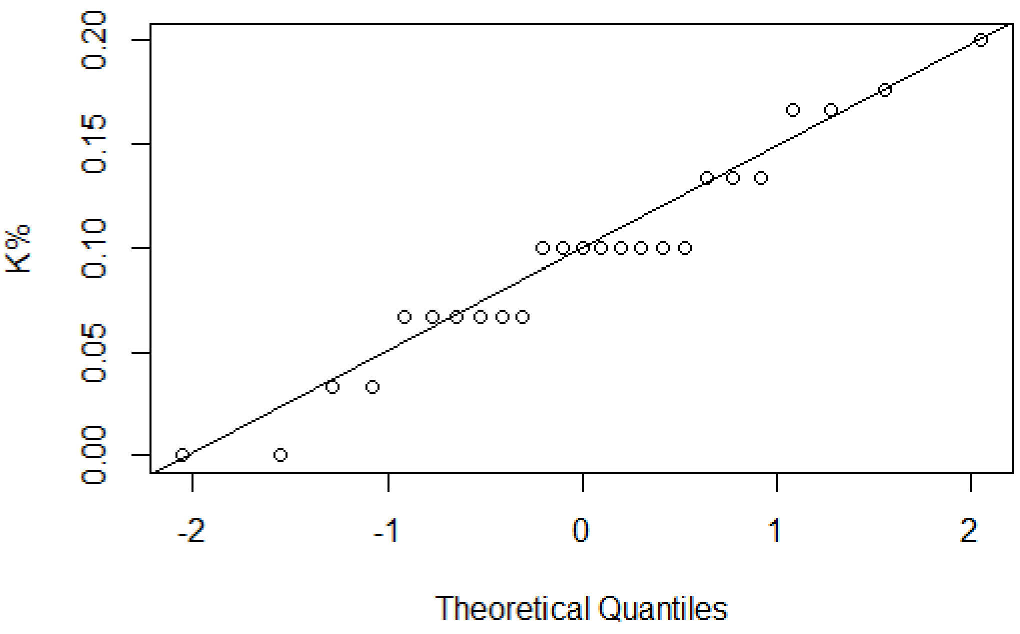
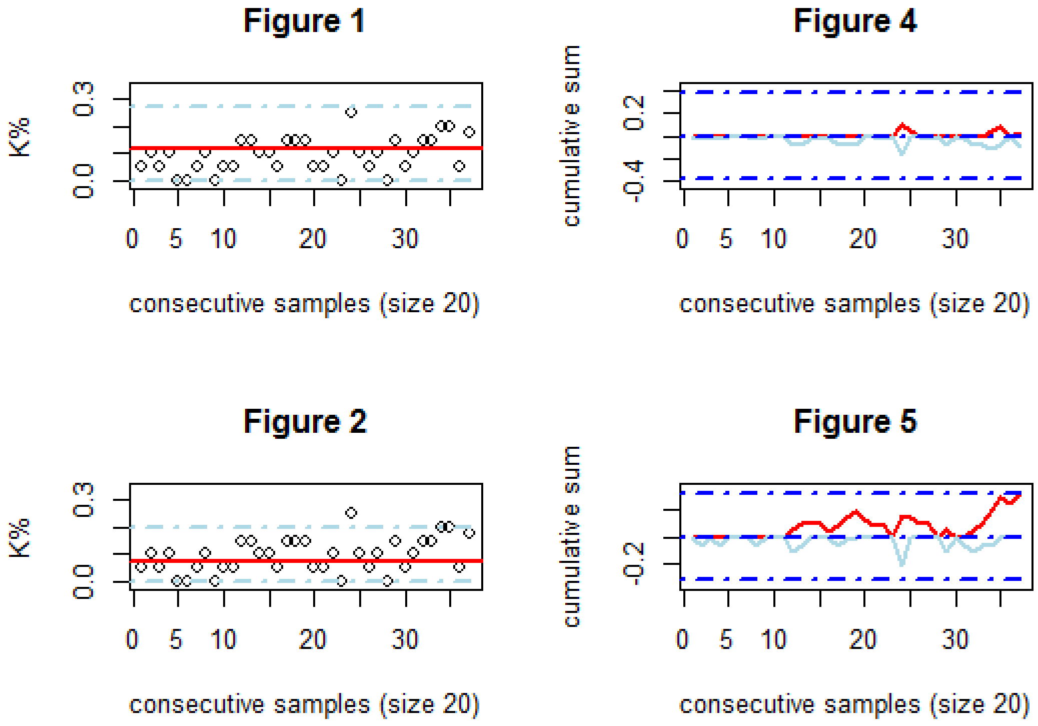


Figure 6: Figures With Different Rational Groups



Appendix

- selections of *R* code, more available through link in References

```
library(devtools)
source_gist(8892981)  # reads in parse.retrosheet2.pbp
parse.retrosheet2.pbp("1999") #also read in for 2000-2014
# manually move .csv files from folder to working directory

# self-imposed limit on years
years <- c(2000:2015)
for (i in years){
   assign(paste0('dat', i),read.csv(paste0(paste0("all",i),".csv"), header=F))
}

# Takes data and divides it into n-sized rational groups of var
chart_div <- function(data, n, var){
   samp = c(0)
   check = floor(nrow(data)/n)
   remainder = nrow(data)-check*n
   col_num = which( colnames(data)==var)
   for (i in 1:(check+1)){
      if(i == (check + 1)){
         samp[i]= sum(((data[,col_num])[(1+n*(i-1)):(n*i-n+remainder)]))/remainder
      }
      else{
         samp[i]=sum(((data[,col_num])[(1+n*(i-1)):(n*i)]))/n
      }
   }
   # return the divisions
   samp
}

# Quicker subset (option to limit to at-bats)
sub_set <- function(data, player_id, ab = FALSE){
   if (ab == TRUE){
      (subset(data, BAT_ID == player_id & AB_FL == TRUE))
   }
   else{
      (subset(data, BAT_ID==player_id))
   }
}
```

```
# DATE creation and sort
dates <- function(data){
   data$DATE <- substr(data$GAME_ID, 4, 12)
   data[order(data$DATE),]
}

# Binary K
alt_K <- function(data){
   ifelse(data$EVENT_CD == 3, 1,0)
}

# CUSUM looking for a shift of 1% K% (base 2010)
k = 1/2
h = 5
K = k*SD
H = h*SD
ARL1 = (exp(-2*(0-k)*(h+1.166))+2*(0-k)*(h+1.166)-1)/(2*(0-k)^2)
(ARL = (1/ARL1+1/ARL1)^(-1))
Cp = c(0)
Cm = c(0)
for(i in 1:(length(samp2011))){
   Cp[i] = max(0,samp2011[i]-(Dbar+K)+Cp[i-1])
   Cm[i] = min(0,(Dbar-K)-samp2011[i]+Cp[i-1])
}

plot(1:(length(samp2011)), Cp, typ="l", col="red", ylim=c(-H-.05,H+.05),
        ylab="cumulative sum",lwd=2,
    xlab = "consecutive samples (size 30)", main="Figure 4: CUSUM - Ichiro - K%
        (2011) - Based on 2010")
par(new=T)
plot(1:(length(samp2011)), Cm, typ="l", col="lightblue", axes=F, ylim=c(-H-.05,H+.05),
        ylab="", xlab="",lwd=2)
abline(h=0, lty=4, col= "blue", lwd=2)
abline(h=H, lty=4, col="blue",lwd=2)
abline(h=-H, lty=4, col="blue",lwd=2)
legend("topleft", c("upper limit","upper CUSUM","lower CUSUM","lower limit"),
     lwd=c(2,2,2,2), cex=0.75
     ,col=c("blue","red","lightblue","blue"), lty=c(4,1,1,4),
     box.lwd = 0,box.col = "white",bg = "white")
box(which = "plot", lty = "solid")
par(new=F)
```