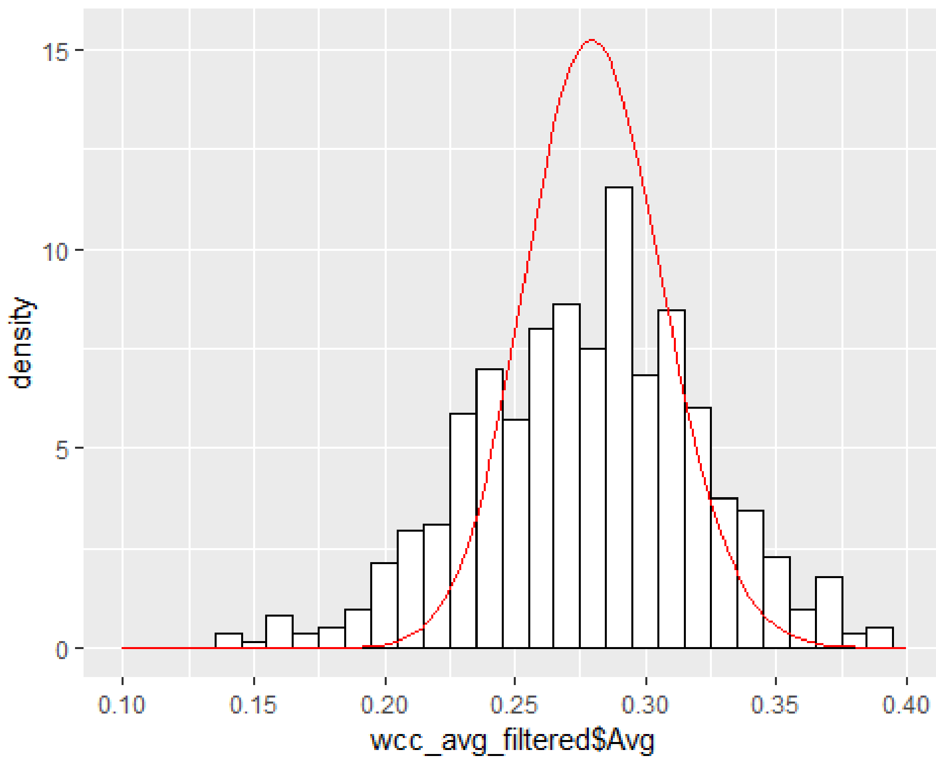
Batting Average and At Bat counts

Zachariah Zanger M.S.

October 2, 2018

```
ggplot(data=wcc_avg_filtered)+geom_histogram(aes(wcc_avg_filtered$Avg, y=..de
nsity..), binwidth = .010, fill="White", colour="black")+
  stat_function(fun=function(x) dbeta(x,alpha_1,beta_1),color="red")+
  scale_x_continuous(limits=c(.1,.4), breaks=c(.1,.15,.20,.25,.30,.35,.40))
```



Introduction

This is an initial attempt to adjust output that can be deceiving due to sample size. The following model needs work. It does some odd things at times, but it will continue to be worked on. The extra reading section explains a few things that explain the premature problems of the model. Batting Average will be used as the featured statistic, although all potential statistics observed are applicable to the model. This analysis is not used to discuss the strengths and weaknesses of the Batting Average statistic, but instead evaluate its distribution of outcomes from 2011 to 2018.

QST is not a decision-making unit in the USF Baseball Program. Although, QST will always produce/suggest research in a respectful fashion for decision makers (Coaching Staff and

Potentially Players). QST sort of throws things up against the wall to see what sticks with decision makers. Decision makers review research and QST keeps working.

The Situation

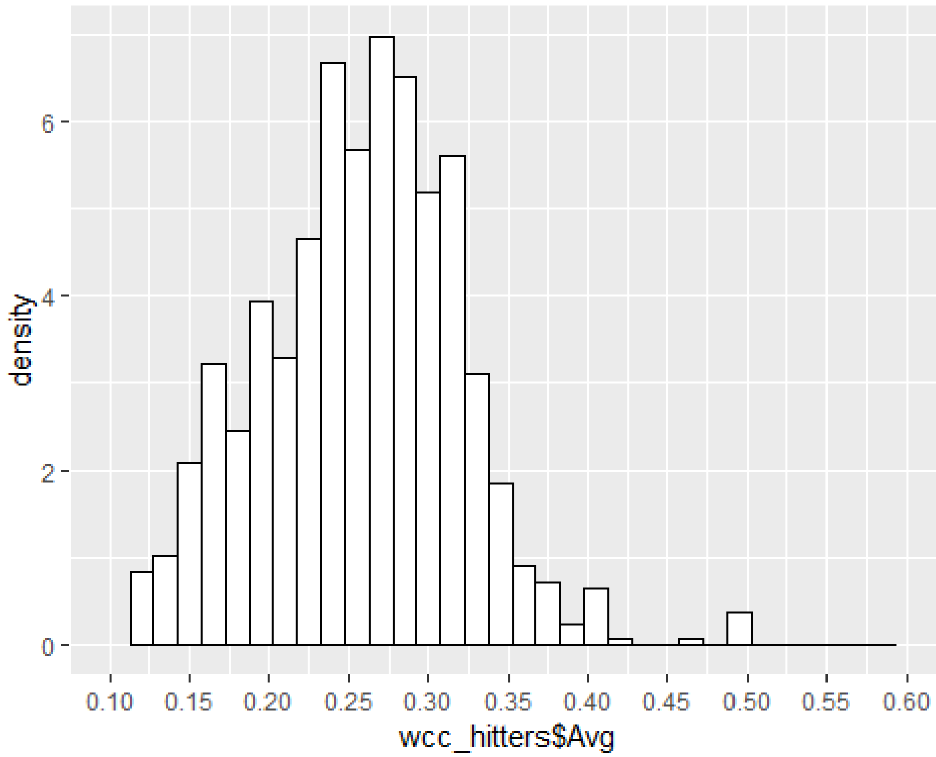
Baseball enhusiasts understand that a .300 batting average in 10 at-bats is not the same as .300 in 100 at-bats. How about a .300 batting average in 300 at-bats versus a .330 average in 250 at-bats? It becomes less clear, as one plays the game of weighing at-bat counts and respective batting averages in their head. Creating a distribution of probabilities assigned to outcomes from an existing dataset, one may be able to adjust observations of at-bats and batting averages. In effect adjusting them, rewarding observations with high at-bat counts, while downgrading observations with low at-bat counts.

If decision makers had a model that could either A) Flag inflated statistics due to sample size, and/or B) objectively compares current players to previous players, decision makers **could have a tool that pairs the subjective (feeling, eye test, look, etc…) with the objective (numbers, data, mathematics, etc…)**.

WCC Background

This is what the distribution of WCC hitters batting average's looks like:

```
ggplot(data=wcc_hitters)+geom_histogram(aes(wcc_hitters$Avg, y=..density..),
binwidth = .015, fill="White", colour="black")+
  scale_x_continuous(limits=c(.10,.60), breaks=c(.10,.15,.20,.25,.30,.35,.40,
.45,.50,.55,.60))
```



The distribution of batting averages in the WCC is telling. This is for a different research piece, but it is easy to see that this distribution is **slightly skewed left.** This means, on the right side of the distribution, the drop off is sharper. Translated, this means it is harder to find a hitter hitting above .260 than below .260. In a different research piece, there could be a measure of how fast the increase/decrease in batting average is for WCC players.

If one were to ask, who the best hitters in the WCC are by batting average, regardless of at-bat count, here they are:

```
head(wcc_hitters %>%
  select(Name, School, Year, AB, Avg) %>%
  arrange(desc(Avg)),20)

##                   Name School    Year AB   Avg
## 1         Blank, Josh    LMU 2014-15  1 1.000
## 2       Haddad, Miles    UOP 2013-14  1 1.000
## 3    Colarossi, Chris    USD 2017-18  3 0.667
## 4       Lengel, James    BYU 2011-12  3 0.667
## 5    Geremia, Anthony    SCU 2014-15  2 0.500
## 6      Erickson, Ryan    LMU 2012-13  2 0.500
## 7      Omahen, Trevor    USD 2017-18  4 0.500
## 8     Dunlap, Spencer    USD 2013-14  4 0.500
## 9       Jauch, Connor    USD 2013-14  8 0.500
## 10  Drongesen, Riley    UOP 2012-13  2 0.500
```

```
## 11 Valentin, Michael    SMC 2014-15 24 0.458
## 12     Meditz, Tyler    SCU 2014-15 26 0.423
## 13    Fornaci, Chris   PEPP 2013-14 42 0.405
## 14       Schafer, Jon    SCU 2015-16  5 0.400
## 15          Diaz, Rob    LMU 2017-18  5 0.400
## 16       Kawano, Ryan    LMU 2016-17  5 0.400
## 17    Kennedy, Kevin    LMU 2013-14  5 0.400
## 18    Kennedy, Kevin    LMU 2012-13  5 0.400
## 19     Haddad, Miles    UOP 2014-15  5 0.400
## 20        Novis, Ryan    SMC 2017-18 35 0.400
```

But where are the Bradley Zimmers, Allen Smoots, Dominic Miroglios of the world?! **The goal of the developed model should be to identify players that have inflated Batting Averages due to low at-bat counts** Those batting averages for those players, and their at-bat counts, need to be discounted quantitatively.

The Model Type

A Beta Distribution is a **Distribution of probabilities for stated outcomes**. Just think that, when using a Beta distribution, you are trying to attach a probability of occurrence to a known outcome among a set of known outcomes. In the case of batting averages, what is the likelihood of observing .500 batting average. There is a very low probability that .500 is observed. Next up is finding a way to look at a player's .500 batting average and either discount the performance (the guy batted twice and got a hit once), or designate the player as an incredible hitter (the guy went 500 for 1000).

Application to Data Set

The parameters of a Beta distribution are Alpha and Beta. Alpha and Beta are what match probabilities to the Distribution of At-Bats and Batting Averages. (Calculation of Alpha and Beta is in the Extra Reading Section)

Now we use Alpha and Beta to Quantitatively Discount or reward combinations of at-bats and batting averages.

Adjusted average = (Hits + Alpha)/(At Bats + Alpha + Beta)

Alpha= 61.29 Beta= 114.64

This is the most important aspect of the exercise. **We take every combination of at-bats and batting average in the WCC from 2011 to 2018, and evaluate them with the equation. This systematically rewards high at-bats with high averages, while discounting low at-bats with low averages.**

Lets evaluate two hypothetical observations: 50 for 100 vs. 75 for 200:

(Hits + Alpha)/(At Bats + Alpha + Beta)

50 for 100 = (50 + 61.24) / (100 + 61.29 + 114.64) = **.403**

125 for 200 = (65 + 61.24) / (200 + 61.29 + 114.64) = **.362**

The model discounted the batter that went 50 for 100 by 97 points! This is likely because it is drawing from the WCC characteristics of at-bats and batting averages. So let's take a look at the WCC. The following table is of the top 40 adjusted batting averages. EB_ESTIMATE is the adjusted average.

```
select(wcc_eb_avg_estimate[order(wcc_eb_avg_estimate$eb_estimate, decreasing=
TRUE),],Name, School, Year,AB,Avg,eb_estimate,)[1:40,]
```

```
##                       Name  School    Year  AB   Avg  eb_estimate
## 487      Bolinger, Royce    GONZ 2011-12 237 0.392    0.3305329
## 510        Lund, Brennon     BYU 2015-16 243 0.387    0.3287066
## 501          Hale, Brock     BYU 2016-17 195 0.395    0.3262144
## 379       Brundage, Beau    PORT 2017-18 209 0.378    0.3211247
## 364      Kalfus, Brenden     SMC 2012-13 194 0.381    0.3207505
## 176       Miller, Austin     LMU 2013-14 206 0.374    0.3190571
## 231       Daniel, Andrew     USD 2013-14 222 0.369    0.3188543
## 40       Zimmer, Bradley     USF 2013-14 220 0.368    0.3181513
## 518     Robinson, Dillon     BYU 2014-15 202 0.371    0.3176018
## 232          Joe, Connor     USD 2013-14 218 0.367    0.3174428
## 552            Sever, Joe   PEPP 2011-12 235 0.362    0.3166952
## 254         Bryant, Kris     USD 2011-12 213 0.366    0.3166308
## 214      Brigman, Bryson     USD 2015-16 191 0.372    0.3165607
## 536            Law, Adam     BYU 2012-13 208 0.365    0.3158027
## 142      Caulfield, Phil     LMU 2016-17 218 0.362    0.3154949
## 124        Harisis, Greg     SCU 2011-12 123 0.398    0.3154280
## 252        LeVier, Corey     USD 2011-12 196 0.367    0.3153746
## 12          Smoot, Allen     USF 2016-17 209 0.364    0.3151765
## 613       Barnett, Aaron    PEPP 2013-14 223 0.359    0.3143808
## 293      Sullivan, Brett     UOP 2013-14 207 0.357    0.3124501
## 470 Cawley Lamb, Payden    GONZ 2013-14 172 0.366    0.3123126
## 503    Anderson, Brennon     BYU 2016-17 260 0.346    0.3114420
## 208        Schuyler, Jay     USD 2016-17 213 0.352    0.3107294
## 533     Robinson, Dillon     BYU 2013-14 149 0.369    0.3104744
## 457   Gunsolus, Mitchell    GONZ 2014-15 207 0.353    0.3104595
## 67           Lavin, Pete     USF 2010-11 236 0.347    0.3104532
## 285      Sullivan, Tyler     UOP 2014-15 211 0.351    0.3099819
## 228        Holder, Kyler     USD 2014-15 224 0.348    0.3099246
## 509     Kringlen, Keaton     BYU 2015-16 141 0.369    0.3092914
## 526       Whitney, Brock     BYU 2013-14 204 0.348    0.3083195
## 249       Daniel, Andrew     USD 2011-12 230 0.343    0.3082884
## 342        Kirtley, Zach     SMC 2014-15 208 0.346    0.3078560
## 540     Hannemann, Jacob     BYU 2012-13 215 0.344    0.3075523
## 524      Nielsen, Hayden     BYU 2014-15 225 0.342    0.3074072
## 202        Schuyler, Jay     USD 2017-18 219 0.342    0.3071048
```

```
## 551         Vincej, Zach    PEPP 2011-12 242 0.339   0.3069868
## 513     Chauncey, Tanner     BYU 2015-16 178 0.348   0.3062413
## 310       Lockwood, Erik     UOP 2011-12 165 0.352   0.3062004
## 344       Villa, Anthony     SMC 2014-15 201 0.343   0.3061536
## 225      Brigman, Bryson     USD 2014-15 218 0.339   0.3057550
```

The model's job is to rank the most impressive batting averages in combination with at-bat counts. **The model will still reward a high batting average, but will adjust due to how many at-bats the player had.** This is the case with Brennon Lund and Brock Hale, ranked 2 and 3 in the table above. Brock Hale has a higher Batting Average than Brennon Lund, but Lund is ranked higher due to a higher at-bat count.

How about strictly USF players:

```
head(wcc_eb_avg_estimate %>%
  filter(School == "USF") %>%
  select(Name, School, Year, AB, Avg, eb_estimate) %>%
  arrange(desc(eb_estimate)),40)

##                     Name  School    Year  AB   Avg  eb_estimate
## 1      Zimmer, Bradley    USF  2013-14 220 0.368    0.3181513
## 2         Smoot, Allen    USF  2016-17 209 0.364    0.3151765
## 3          Lavin, Pete    USF  2010-11 236 0.347    0.3104532
## 4       Perri, Michael    USF  2017-18 232 0.336    0.3052230
## 5    Miroglio, Dominic    USF  2014-15 206 0.340    0.3050949
## 6      Turner, Zachary    USF  2012-13 228 0.333    0.3037343
## 7      Atkinson, Derek    USF  2013-14 215 0.330    0.3016741
## 8       Helland, Riley    USF  2017-18 222 0.329    0.3014581
## 9      Puskarich, Ross    USF  2015-16 154 0.331    0.2981181
## 10     Zimmer, Bradley    USF  2012-13 203 0.320    0.2968986
## 11       Sinatro, Matt    USF  2016-17 226 0.314    0.2953091
## 12      Maffei, Justin    USF  2011-12 200 0.315    0.2946592
## 13        Smoot, Allen    USF  2015-16 156 0.321    0.2945816
## 14         Clear, Adam    USF  2010-11 123 0.325    0.2939152
## 15      Perri, Michael    USF  2016-17 215 0.312    0.2938364
## 16     Allen, Jonathan    USF  2017-18 227 0.308    0.2928294
## 17       Garcia, Aritz    USF  2010-11 107 0.318    0.2906908
## 18     Hofmann, Connor    USF  2014-15 208 0.303    0.2899760
## 19         Clear, Adam    USF  2011-12 198 0.303    0.2897729
## 20       Sinatro, Matt    USF  2015-16 167 0.305    0.2897360
## 21   Hendriks, Brendan    USF  2011-12 119 0.311    0.2895123
## 22          Balog, Nik    USF  2011-12 202 0.302    0.2894530
## 23     Atkinson, Derek    USF  2012-13 135 0.304    0.2880433
## 24      Helland, Riley    USF  2016-17 104 0.308    0.2878664
## 25      Maffei, Justin    USF  2012-13 253 0.292    0.2862396
## 26     Bernatz, Connor    USF  2010-11 212 0.292    0.2857189
## 27   Hendriks, Brendan    USF  2014-15 214 0.290    0.2845970
## 28       Higgs, Travis    USF  2010-11 180 0.289    0.2839160
## 29       Valley, Blake    USF  2015-16 152 0.289    0.2838035
## 30   Hendriks, Brendan    USF  2013-14 209 0.287    0.2834529
```

```
## 31      Cruikshank, Bob    USF 2012-13 210 0.286   0.2828920
## 32       Eaton, Michael    USF 2014-15 168 0.286   0.2826362
## 33    Miroglio, Dominic    USF 2016-17 239 0.285   0.2825105
## 34        Mahood, Jason    USF 2011-12 201 0.284   0.2819774
## 35      Bruce, Harrison    USF 2016-17 145 0.283   0.2815022
## 36          Bate, Brady    USF 2016-17 112 0.277   0.2797582
## 37      Cruikshank, Bob    USF 2013-14 177 0.277   0.2793681
## 38      Atkinson, Derek    USF 2014-15 204 0.275   0.2782808
## 39            Balog, Nik    USF 2010-11 215 0.274   0.2781611
## 40 Ramirez, Jr., Manny    USF 2016-17 101 0.267   0.2774303
```

Extra Reading But not Necessary

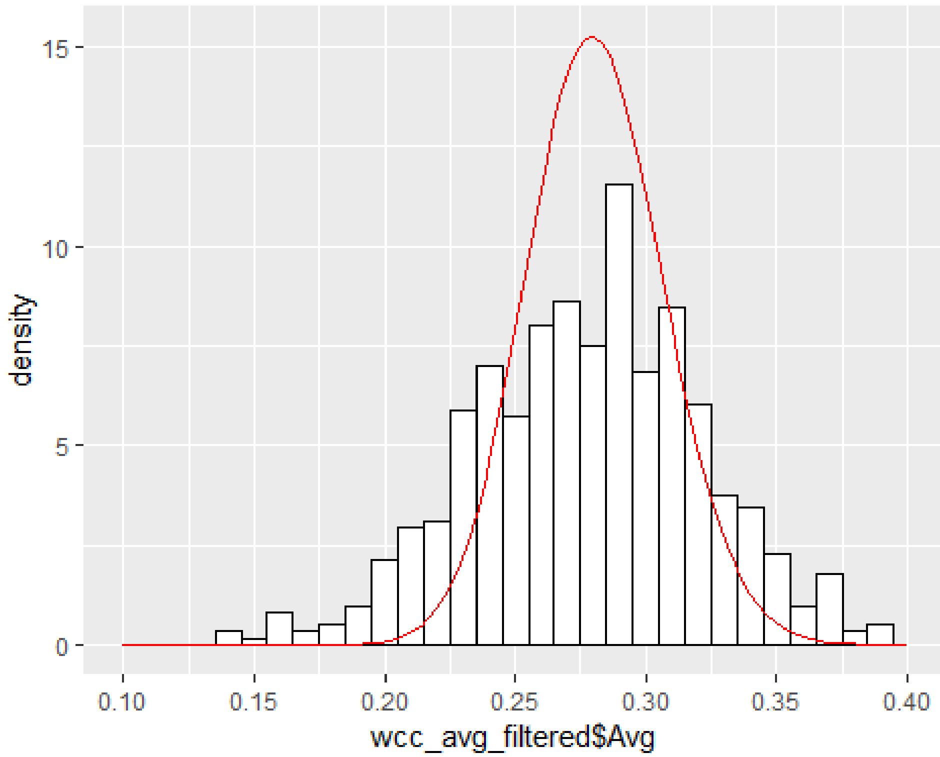
Please find below the calculation of Alpha and Beta:

```
log_likelihood<-function(alpha,beta){
  x<-(wcc_avg_filtered$H)
  total<-(wcc_avg_filtered$AB)
  -sum(VGAM::dbetabinom.ab(x,total,alpha,beta,log=TRUE))
}


m <- mle(log_likelihood, start = list(alpha = 1, beta = 10), method = "L-BFGS
-B",lower = c(0.0001, .1))
pa <- coef(m)
alpha_1= pa[1]
beta_1=pa[2]
```

The model utilizes the existing characteristics of the underlying distribution of Averages. The model calls to the number of hits and the number of at-bats from the distribution. It takes the logarithm of each combination of our beta distribution of averages. Once again, the Beta Distribution is a distribution of probabilities. Setting log=true takes the logarithm of each combination of a t-bats and batting averages.

```
ggplot(data=wcc_avg_filtered)+geom_histogram(aes(wcc_avg_filtered$Avg, y=..de
nsity..), binwidth = .010, fill="White", colour="black")+
  stat_function(fun=function(x) dbeta(x,alpha_1,beta_1),color="red")+
  scale_x_continuous(limits=c(.1,.4), breaks=c(.1,.15,.20,.25,.30,.35,.40))
```



Generally speaking, the model (red line outlining probability on the y-axis) does ok, but the tails need to be fatter to encompass observations outside of it at the extremes. The Log Likelihood function is what I suspect the culprit to be, but need more time to be 100% sure.

The model is not completely powerless. This model captures a large amount of data, and can be applied to the distribution of at-bats and batting averages.