

Statistical Consulting Center

Department of Statistics

Eberly College of Science

Tel: (814) 863-0281

Fax: (814) 863-7114

The Pennsylvania State University

323 Thomas Building

University Park, Pa 16802

July 28, 2013

XXXXXXX

Department of XXX

The Pennsylvania State University

University Park, PA 16802

Re: SCC # 11-1-042

Dear Coach Popovich,

Thank you for bringing your project to the Statistical Consulting Center. Our recommendations to address the research questions on your project are included in this report. The report is organized in several parts. In the first section, we will summarize our understanding of your study and the statistical questions. Second, we will provide some recommendations to answer your questions. Finally, we will review our findings and provide a list of references.

**1. Project Description**

You came to us to look for evidence basketball players experience a ‘heating up’ effect, whereby they are likely to make many shots in a row or miss many shots in a row, which would suggest that shots are not independent events. You also indicated interest in understanding more about the relationship between shooting percentage (field goal percentage), and several other statistical variables. We discussed previous research on this question, in particular that noted psychologist and mathematician Amos Tversky considered the question from the perspective of the cognitive psychology. He found that people tend to underestimate the ability of strings of random data to contain long “runs” of consecutive successes or failures and consequently overestimate the effect of “heating up”. We are looking to disprove his results using recent data, the entire play by play dataset from the 2011-2012 NBA regular season (excludes playoffs).

Because we are working with is observational data, we will not be able to determine causality from any of our findings. In order to figure out if ‘heating up’ exists, we will consider the number of sets consecutive makes or misses, which we will call a run, and test if the number of runs is different than we would expect to occur if the data were entirely random. We can address this question using the Wald-Wolfowitz runs test. Although we had discussed additional tests in our previous meeting, we believe that the results of this analysis will be satisfactory to answer your research question. In order to answer your question about the relationship between shooting percentage and other variables, we will examine correlations between the variables and build a predictive model to estimate shooting percentage based on the other variables. To establish this model we will use linear regression.

**2. Recommendations**

**2.1 Question 1:** Are shooters likely to make several shots in a row or miss several shots in a row?

**2.1.1 Exploratory Data Analysis (EDA)**

Much of the work for this analysis consisted of cleaning and standardizing the data so that it would be usable. The initial data consisted of approximately 500,000 plays made during the regular season. Of that data, 170,004 of those plays were shot attempts. In order to run the test, we needed the data to be at the level of an individual shooter during a specific game, (we assumed that ‘heating up’ would not carry over from one night to the next) which provided us with 19,300 unique datasets of shooting performance on a given night. We further excluded performances that had less than 5 shot attempts in a night because it is impossible to find a statistically significant result with less than 5 shot attempts, which left us with 14,120 unique shooting performances.

Bearing in mind that we limited the data to only those performances with 5 or more shot attempts, the mean number of shots made, missed, and attempted were as follow:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Made | Missed | Attempted |
| Mean | 4.71 | 6.38 | 11.09 |
| Standard Deviation | 2.81 | 3.29 | 5.14 |

Histograms of the data revealed skewed distributions, histograms are also plotted against a normal distribution for comparison (the histogram of attempts shows no values less than 5):

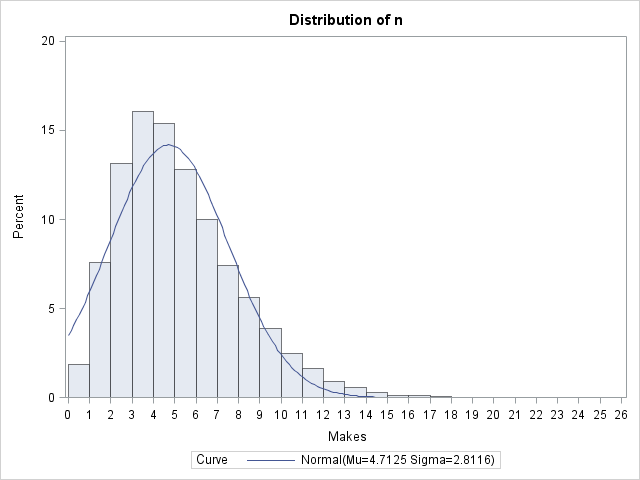


Fig 1. Histogram of successful shots (Makes)

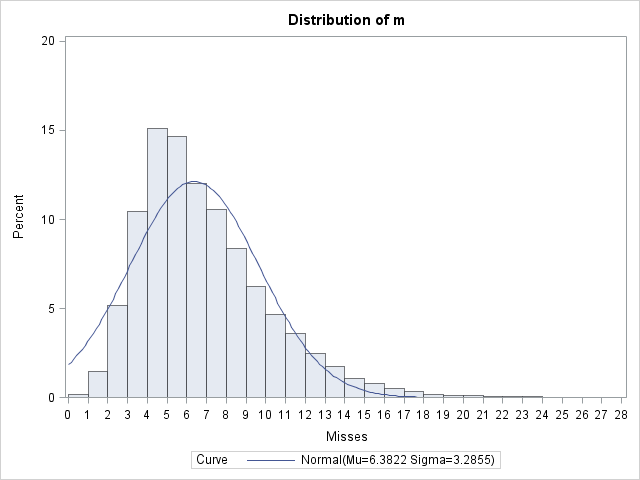


Fig 2. Histogram of unsuccessful shots (Misses)

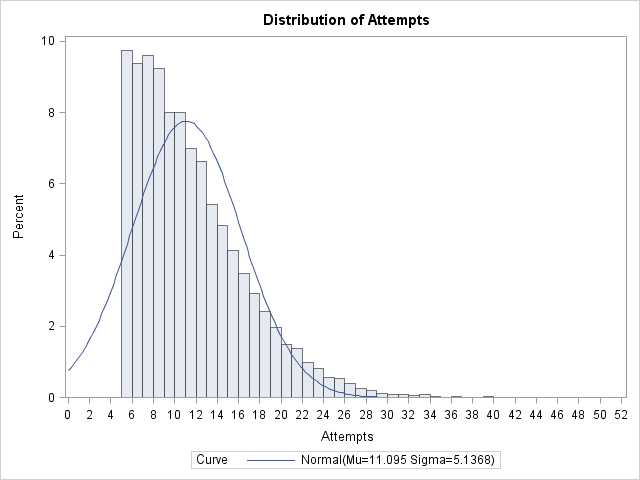


Fig 3. Histogram of shot attempts

**2.1.2 Method**

We chose to analyze the data using the Wald-Wolfowitz runs test. This technique tests whether the number of runs that we observe in a dataset would occur by chance. As an illustration of how this test works consider the following hypothetical examples of a sequence of 20 1’s and 0’s, where we could imagine 1’s representing makes and 0’s representing misses:

Sequence A: 10101010101010101010

Sequence B: 00000000001111111111

Sequence A represents one extreme, wherein the shots alternate every time, which would mean there were 20 runs. Sequence B represents the other extreme, where we have only 2 runs. Our hypothesis is that if a ‘heating up’ effect occurs, we will see a small number of runs, smaller than we would expect to find via random chance. Typically the Walf-Wolfowitz test is a two-sided test. This means that we would get a significant result if we found a significantly large number of runs, or a significantly small number of runs. For our purposes, we are only interested in finding a small number of runs, so we would say we have a significant result if we had a smaller number of runs than expected.

Our data consist of 14,120 sequences, so in order to get an aggregate test statistic, we will examine the data visually and also take the average of our test statistics. Using the standard normal approximation, an individual shooting performance (sequence) would be significant at the .05 level if the z-value associated with that sequence was less than -1.64. Below is the histogram of all of our 14,120 tests:

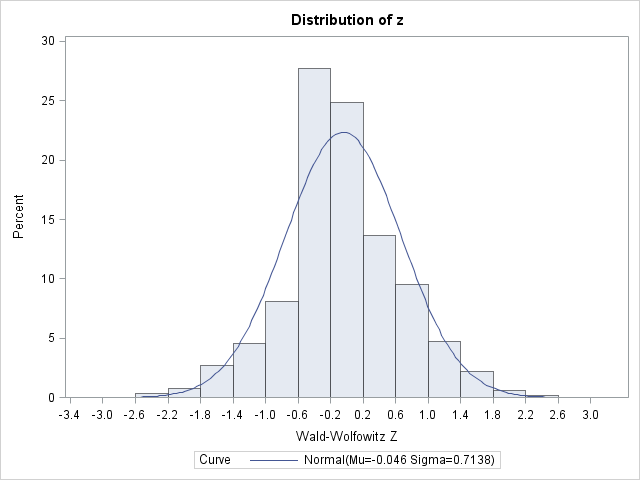


Fig. 4 Histogram of Wald-Wolfowitz Z values

We may be disappointed to find that only a very small portion of our data lies in the region to the left of -1.64. This means that we are not seeing many significant test statistics. The average of the Wald-Wolfowitz test statistics is -.045 with a standard deviation of .714. We can invoke the central limit theorem to find the p-value associated with this statistic which is 0.475. The p-value is the chance of randomly observing a statistic at least as extreme as the one we calculated. This means that the result is not significant at all, it isn’t even close.

**2.1.3 Summary**

Under the assumptions of the model we did not find any evidence of the ‘heating up’ effect, as defined by observing fewer runs than would be expected in random data. Our results are in alignment with those explored in Amos Tversky’s paper ‘The Cold Facts About the “Hot Hand” in Basketball’ which found no evidence of ‘heating up’ and attributed widespread belief in the phenomenon to a cognitive bias. Although we found no evidence for the effect league wide, it is possible that the effect may exist for individual players, although any perceived effect may be the result of cherry-picking particular players and would not be useful in making an inference to the broader population of NBA players.

**2.2 Question 2:** How is shooting percentage related to other player variables?

**2.2.1 Exploratory Data Analysis (EDA)**

In order to answer this question, we combined player level data from all 30 teams in the NBA from 2 sources by merging on player name. In the 2011-2012 season there were 515 players, including duplicates of players where they played for multiple teams during the season (the players had different statistics on each team and were counted as separate observations for the purpose of this analysis. We chose to exclude any players with less than 100 points during the regular season, as these players made less than one shot per game, played very few minutes, and their presence significantly reduced the predictive power of our model. This left us with 370 players. Below are summary statistics for the variables we considered which we believed may be related to shooting percentage (field goal percentage). Note that missing values were generated for Experience and Three Point Percentage:

| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** | **Label** |
| --- | --- | --- | --- | --- | --- | --- |
| Field Goal Percentage | 370 | 0.44455 | 0.05958 | 0.28200 | 0.67900 | Field Goal Percentage |
| Position | 370 | 2.93514 | 1.38951 | 1.00000 | 5.00000 | Position |
| Experience | 326 | 5.87117 | 3.87044 | 1.00000 | 17.00000 | Experience |
| Age | 370 | 26.62162 | 4.17147 | 19.00000 | 39.00000 | Age |
| Minutes Played | 370 | 1234 | 572.86404 | 258.00000 | 2546 | Minutes Played |
| Season Field Goals | 370 | 189.46757 | 123.33848 | 32.00000 | 643.00000 | Season Field Goals |
| Season Field Goals Attempted | 370 | 421.43514 | 261.31639 | 72.00000 | 1336 | Season Field Goals Attempted |
| Three Point Percentage | 331 | 0.28447 | 0.14745 | 0 | 1.00000 | Three Point Percentage |
| Free Throw Percentage | 370 | 0.73809 | 0.11892 | 0 | 1.00000 | Free Throw Percentage |

**2.1.2 Method**

To determine how these variables were related we began by looking at the correlations between the variables of interest. As we can see below, there are a few obvious strong correlations. Age is highly correlated with Experience which could be expected. Minutes Played, Season Field Goals, and Season Field Goals attempted are all correlated because a player who has more minutes has more opportunities to attempt shots, and in general more shot attempts will yield more field goals. A more surprising result is that the correlation of Field Goal Percentage with Free Throw Percentage and Three Point Percentage is actually negative. Additionally, position number (where 1 represents the smallest players, point guards, and 5 represents the tallest players, centers), is positively correlated with Field Goal Percentage, which seems counterintuitive, because smaller guards are generally known to be better shooters than the larger centers. The likely underlying factor explaining these trends is that smaller guards are forced to take jump shots, which are more difficult and generally a less accurate type of shot than the dunks and layups that larger players take. The full correlation matrix is produced below:

| **Pearson Correlation Coefficients  Number of Observations** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **FGpct** | **posno** | **Exp** | **Age** | **MP** | **FG** | **FGA** | **threePpct** | **FTpct** |
| |  | | --- | | **FGpct** | | Field Goal Percentage | | |  | | --- | | 1.00000 | | 370 | | |  | | --- | | 0.52290 | | 370 | | |  | | --- | | -0.14647 | | 326 | | |  | | --- | | -0.10431 | | 370 | | |  | | --- | | 0.23370 | | 370 | | |  | | --- | | 0.28633 | | 370 | | |  | | --- | | 0.13745 | | 370 | | |  | | --- | | -0.31124 | | 331 | | |  | | --- | | -0.24608 | | 370 | |
| |  | | --- | | **posno** | | Position | | |  | | --- | | 0.52290 | | 370 | | |  | | --- | | 1.00000 | | 370 | | |  | | --- | | 0.03636 | | 326 | | |  | | --- | | 0.03970 | | 370 | | |  | | --- | | -0.04952 | | 370 | | |  | | --- | | -0.02160 | | 370 | | |  | | --- | | -0.11501 | | 370 | | |  | | --- | | -0.45847 | | 331 | | |  | | --- | | -0.44930 | | 370 | |
| |  | | --- | | **Exp** | | Experience | | |  | | --- | | -0.14647 | | 326 | | |  | | --- | | 0.03636 | | 326 | | |  | | --- | | 1.00000 | | 326 | | |  | | --- | | 0.90368 | | 326 | | |  | | --- | | -0.03467 | | 326 | | |  | | --- | | -0.04184 | | 326 | | |  | | --- | | -0.02544 | | 326 | | |  | | --- | | 0.10393 | | 293 | | |  | | --- | | 0.02100 | | 326 | |
| |  | | --- | | **Age** | | Age | | |  | | --- | | -0.10431 | | 370 | | |  | | --- | | 0.03970 | | 370 | | |  | | --- | | 0.90368 | | 326 | | |  | | --- | | 1.00000 | | 370 | | |  | | --- | | -0.04363 | | 370 | | |  | | --- | | -0.07315 | | 370 | | |  | | --- | | -0.06490 | | 370 | | |  | | --- | | 0.10171 | | 331 | | |  | | --- | | 0.04355 | | 370 | |
| |  | | --- | | **MP** | | Minutes Played | | |  | | --- | | 0.23370 | | 370 | | |  | | --- | | -0.04952 | | 370 | | |  | | --- | | -0.03467 | | 326 | | |  | | --- | | -0.04363 | | 370 | | |  | | --- | | 1.00000 | | 370 | | |  | | --- | | 0.89658 | | 370 | | |  | | --- | | 0.89577 | | 370 | | |  | | --- | | 0.03636 | | 331 | | |  | | --- | | 0.23904 | | 370 | |
| |  | | --- | | **FG** | | Season Field Goals | | |  | | --- | | 0.28633 | | 370 | | |  | | --- | | -0.02160 | | 370 | | |  | | --- | | -0.04184 | | 326 | | |  | | --- | | -0.07315 | | 370 | | |  | | --- | | 0.89658 | | 370 | | |  | | --- | | 1.00000 | | 370 | | |  | | --- | | 0.98263 | | 370 | | |  | | --- | | 0.00142 | | 331 | | |  | | --- | | 0.27025 | | 370 | |
| |  | | --- | | **FGA** | | Season Field Goals Attempted | | |  | | --- | | 0.13745 | | 370 | | |  | | --- | | -0.11501 | | 370 | | |  | | --- | | -0.02544 | | 326 | | |  | | --- | | -0.06490 | | 370 | | |  | | --- | | 0.89577 | | 370 | | |  | | --- | | 0.98263 | | 370 | | |  | | --- | | 1.00000 | | 370 | | |  | | --- | | 0.06621 | | 331 | | |  | | --- | | 0.32731 | | 370 | |
| |  | | --- | | **threePpct** | | Three Point Percentage | | |  | | --- | | -0.31124 | | 331 | | |  | | --- | | -0.45847 | | 331 | | |  | | --- | | 0.10393 | | 293 | | |  | | --- | | 0.10171 | | 331 | | |  | | --- | | 0.03636 | | 331 | | |  | | --- | | 0.00142 | | 331 | | |  | | --- | | 0.06621 | | 331 | | |  | | --- | | 1.00000 | | 331 | | |  | | --- | | 0.33824 | | 331 | |
| |  | | --- | | **FTpct** | | Free Throw Percentage | | |  | | --- | | -0.24608 | | 370 | | |  | | --- | | -0.44930 | | 370 | | |  | | --- | | 0.02100 | | 326 | | |  | | --- | | 0.04355 | | 370 | | |  | | --- | | 0.23904 | | 370 | | |  | | --- | | 0.27025 | | 370 | | |  | | --- | | 0.32731 | | 370 | | |  | | --- | | 0.33824 | | 331 | | |  | | --- | | 1.00000 | | 370 | |

To gain a deeper understanding of the relationship between Field Goal Percentage and the other variables, we run a linear regression. Linear regression is a technique that allows us to model and predict a dependent variable based on the values of one or more independent variables. We used SAS software to generate the regression model, and we picked the best model based on a combination of predictive power and ease of interpretation. We used 2 model selection algorithms, forward stepwise selection, and maximum R2 to choose the best model. Both methods produced the same model, which has the following equation (where percentage is expressed as a decimal between 0 and 1):

Field Goal % = 0.41679 + 0.00528\*Position + 0.00173\*Season Field Goals - 0.00080254\*Season Field Goals Attempted - .00116\*Experience+0.00002138\*Minutes Played

This model has an R2 of 0.7221. This means that our model explains about 70% of the variation in Field Goal Percentage. Below are diagnostic charts for the model:

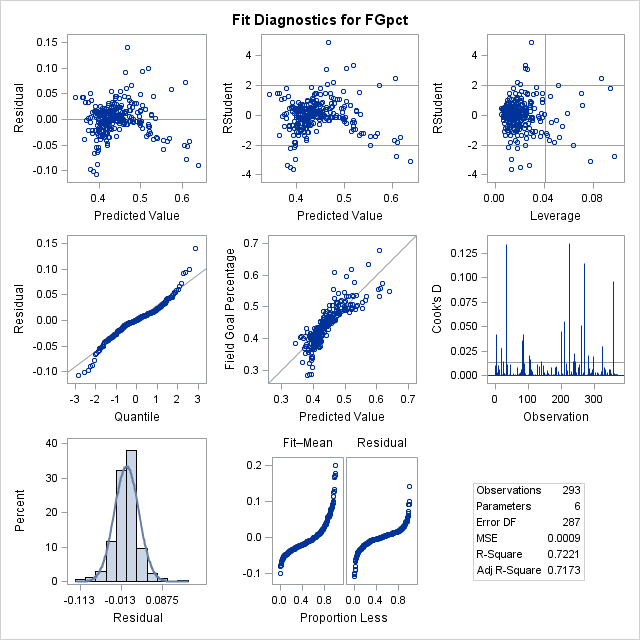


Fig. 5 Fit Diagnostics

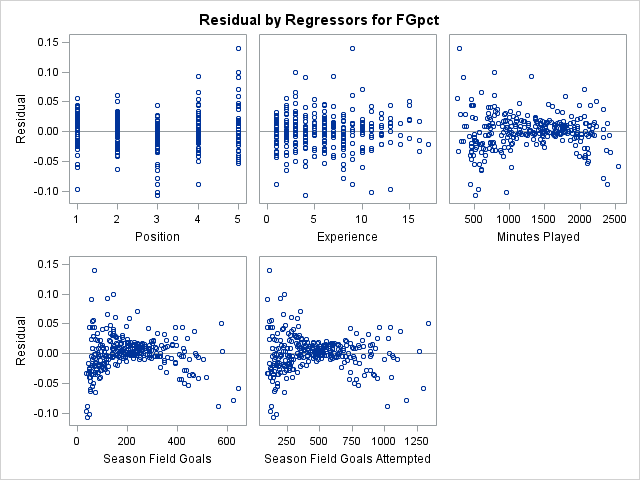


Fig 6. Residuals by regressor

**2.2.3 Summary**

It should not come as a surprise that parameter estimates for Experience and Season Field Goal Attempts are negative. Because the formula for Field Goal Percentage is (Season Field Goals)/(Season Field Goal Attempts) it stands to reason that an increase in attempts will increase the denominator which will decrease Field Goal Percentage. Additionally, we found that Experience had a negative correlation with Field Goal Percentage from our correlation table. Although this model does have a fair amount of predictive power, it is important to realize that there may be underlying factors that explain some of these trends (taller players have access to easier shots).

**3. Conclusions and discussions**

We found no evidence to suggest that a “heating up” effect exists for shooters in basketball. This result is in line with previous research done by Amos Tversky suggesting that the effect is widely perceived to be true even though the data does not support that conclusion. Additionally, we examined the relationship between shooting percentage and several other variables. Although we were able to create a predictive model, we are not able to suggest that differences in our independent variables cause changes in our dependent variable because the study was observational in nature.

Thank you for the opportunity to work on this project with you.

Sincerely,

**Robert Pehlman and the Statistical Consulting Center**

**References**

"NBA & ABA Basketball Statistics & History | Basketball-Reference.com." *Basketball-Reference.com*. N.p., n.d. Web. 28 July 2013. <http://www.basketball-reference.com/>.

"BasketballValue.com Data Files." *BasketballValue.com Data Files*. N.p., n.d. Web. 28 July 2013. <http://basketballvalue.com/downloads.php>.

Tversky, A., and Gilovich, T. (1989). The cold facts about the ‘‘hot hand’’ in basketball. Chance, 2(1),

16–21. Reprinted with permission from CHANCE. Copyright 1989 by the American Statistical Association. All rights reserved.

**Code**

OPTIONS NODATE NONUMBER;

libname library 'C:\Users\Robert\Desktop\Homework\STAT480\Code';

libname out 'C:\Users\Robert\Desktop\Homework\STAT581\data';

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Code for question 1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Reads in data and organizes data for analysis by only retaining shooting attempt plays\*/

**DATA** thesis1 (where=(Made+Missed>**0**) keep=lineno x2 name team gameid time made missed );

length lineno $3. x2 $15.;

infile 'C:\Users\Robert\Desktop\Homework\STAT480\Code\thesis1.csv' firstobs=**2** dlm="," missover;

input gameid $14. lineno time $ x1 $ x2 $ x3 $ x4 $ x5 $ x6 $ x7 $ x8 $ x9 $ x10 $ x11 $ x12 $;

array x(**12**);

team=substr(x1,**2**,**3**);

if count(x2,'.')>**0** then name=compbl(x2||x3);

else name=x2;

Made = **0**;

Missed = **0**;

do i=**1** to dim(x);

Made = Made + count(x(i),'Made','i');

Missed = Missed + count(x(i),'Missed','i');

end;

**run**;

**PROC** **SORT** data=thesis1 out=sortedthesis1;

by gameid name;

**run**;

/\*Creates an indicator variable to separate datasets\*/

**data** countsets;

set sortedthesis1;

by gameid name;

if first.name then count+**1**;

**run**;

**proc** **print** data=countsets (OBS=**80**);

**run**;

\*Determines the number of unique shooting performances;

**data** numsets;

set sortedthesis1;

by gameid name;

if first.name;

**run**;

/\*Macro code to split the data into 19,300 datasets and compute the Wald-Wolfowitz statistic

for each datasets\*/

**%macro** split(ndsn);

%do i = **1** %to &ndsn.;

data out.dsn&i.;

set countsets;

if count=&i;

run;

data runcount;

set out.dsn&i. nobs=nobs end=last;

if made>**0.5** then n+**1**;

if made<**0.5** then m+**1**;

retain runs **0** numpos **0** numneg **0**;

previous=lag(made);

if \_n\_=**1** then do;

runs=**1**;

prevpos=**.**;

currpos=**.**;

prevneg=**.**;

currneg=**.**;

end;

else do;

prevpos=( previous > **0.5** );

currpos=( made > **0.5** );

prevneg=( previous < **0.5** );

currneg=( made < **0.5** );

if \_n\_=**2** and (currpos and prevpos) then numpos+**1**;

else if \_n\_=**2** and (currpos and prevneg) then numneg+**1**;

else if \_n\_=**2** and (currneg and prevpos) then numpos+**1**;

else if \_n\_=**2** and (currneg and prevneg) then numneg+**1**;

if currpos and prevneg then do;

runs+**1**;

numpos+**1**;

end;

if currneg and prevpos then do;

runs+**1**;

numneg+**1**;

end;

end;

if last then output;

run;

data out.waldwolf&i.;

label z='Wald-Wolfowitz Z'

pvalue='Pr > |Z|';

set runcount;

mu = ( (**2**\*n\*m) / (n + m) ) + **1**;

sigmasq = ( (**2**\*n\*m) \* (**2**\*n\*m-(n+m)) ) / ( ((n+m)\*\***2**) \* (n+m-**1**) );

sigma=sqrt(sigmasq);

drop sigmasq;

if N GE **50** then Z = (Runs - mu) / sigma;

else if Runs-mu LT **0** then Z = (Runs-mu+**0.5**)/sigma;

else Z = (Runs-mu-**0.5**)/sigma;

pvalue=**2**\*(**1**-probnorm(abs(Z)));

p2=probnorm(Z);

run;

%end;

**%mend** split;

%***split***(ndsn=**19300**);

\*Dataset containing all of the Wald-Wolfowitz statistics;

**data** out.results;

set out.waldwolf1-out.waldwolf19300;

**run**;

**proc** **print** data=results (firstobs=**19000** obs=**19300**);

var z pvalue name gameid n m;

**run**;

**data** results2;

set results (where=(n+m>**4**));

label n = 'Makes' m='Misses';

Attempts=n+m;

**run**;

\*Produces histograms and summary statistics;

**proc** **univariate** data=results2;

var n m attempts;

histogram / normal endpoints=**0** to **20** by **1**;

**run**;

**proc** **univariate** data=results2;

var z;

histogram / normal endpoints=-**3** to **3** by **.4**;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Code for question 2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** shooting;

length player $**30**;

infile 'C:\Users\Robert\Desktop\Homework\STAT480\Code\shooting.txt' dlm="," dsd missover;

input Rk Player $ Age G GS MP FG FGA FGpct threeP threePA threePpct FT FTA FTpct ORB DRB TRB AST STL BLK TOV PF PTS;

**run**;

**data** players (drop=playerno ht birthdate college);

length player $**30**;

infile 'C:\Users\Robert\Desktop\Homework\STAT480\Code\players.txt' dlm="," dsd missover;

input playerno Player $ Pos $ Ht $ Wt BirthDate $ Exp College $;

height=input(substr(ht,**1**,**1**),**1.**)\***12**+input(substr(ht,**3**,**2**),**2.**);

if pos = 'PG' then posno= **1**;

else if pos = 'SG' then posno= **2**;

else if pos = 'SF' then posno= **3**;

else if pos = 'PF' then posno= **4**;

else if pos = 'C' then posno= **5**;

**run**;

**proc** **sort** data=shooting;

by player;

**run**;

**proc** **sort** data=players;

by player;

**run**;

**data** playershooting (where=(pts ge **100**));

merge players shooting;

by player;

label fgpct = "Field Goal Percentage" posno = "Position" exp = "Experience" age = "Age"

mp = "Minutes Played" fg = "Season Field Goals" fga = "Season Field Goals Attempted"

threeppct = "Three Point Percentage" ftpct = "Free Throw Percentage";

**run**;

\*Creates correlation and regression output;

**proc** **corr** data=playershooting noprob;

var fgpct posno exp age mp fg fga threeppct ftpct;

**run**;

**proc** **reg** data=playershooting;

model fgpct = posno exp age mp fg fga threeppct ftpct / selection = forward;

model fgpct = posno exp age mp fg fga threeppct ftpct / selection = maxr;

**run**;