Lab 2: Probability

Hot Hands

Basketball players who make several baskets in succession are described as having a "hot hand." Fans and players have long believed in the hot hand phenomenon, which refutes the assumption that each shot is independent of the next. However, a 1985 paper by Gilovich, Vallone, and Tversky collected evidence that contradicted this belief and showed that successive shots are independent events[1]. This paper started a great controversy that continues to this day, as you can see by Googling "hot hand basketball."

We do not expect to resolve this controversy today. However, in this lab we'll apply one approach to answering questions like this. The goals for this lab are to (1) think about the effects of independent and dependent events, (2) learn how to simulate shooting streaks in SAS, and (3) compare a simulation to actual data in order to determine whether the hot hand phenomenon appears to be real.[2]

Saving Your Code

As you go along, copy and paste your code into SAS and save it. This is a good way to keep track of your code and be able to reuse it later. To run your code from this document, you can simply highlight the code and click the **Submit** button on the toolbar, or highlight the code, right-click, and select **Submit**.

You'll also want to save this code in a SAS syntax (.sas) file. To do so, click the disk icon. The first time that you select **Save**, a dialog box appears that asks for a filename and location. You can name the file anything that you want.

Getting Started

Our investigation will focus on the performance of one player: Kobe Bryant of the Los Angeles Lakers. His performance against the Orlando Magic in the 2009 NBA finals earned him the Most Valuable Player award, and many spectators commented on how he appeared to show a hot hand. Let's load some data from those games and look at the first several rows.

If you are using SAS University Edition, you need to ensure that interactive mode is turned off. To do this, click the button to the right of **Sign Out** in the upper right corner of the window and then click **Preferences**. In the Preferences window, on the General tab, the bottom check box (located next to the text **Start new programs in interactive mode**) should ***not*** be selected. If the box is selected, you need to clear it and save your change.

.

[1] "The Hot Hand in Basketball: On the Misperception of Random Sequences." Gilovich, T., Vallone, R., Tversky, A., 1985. Cognitive Psychology, 17, pp. 295-314

```
filename kobe url 'http://www.openintro.org/stat/data/kobe.csv';

proc import
   datafile=kobe
   out=kobe(drop=time)
   dbms=csv
   replace;
   getnames=yes;
   guessingrows=max;
run;

proc print data=kobe (obs=10);
run;
```

In this data set, every row records a shot taken by Kobe Bryant. If he hit the shot (made a basket), a hit, H**,** is recorded in the column named **basket**. Otherwise, a miss, M, is recorded.

Just looking at the string of hits and misses, it can be difficult to gauge whether it seems like Bryant was shooting with a hot hand. One way we can approach this is by considering the belief that hot hand shooters tend to go on shooting streaks. For this lab, we define the length of a shooting streak to be the ***number of consecutive baskets made until a miss occurs***.

For example, in Game 1, Bryant had the following sequence of hits and misses from his nine shot attempts in the first quarter:

H M | M | H H M | M | M | M

To verify this, use the following PROC PRINT syntax:

```
proc print data=kobe;
   where game=1 and quarter='1';
   var basket;
run;
```

Within the nine shot attempts, there are six streaks, which are separated by a | above. Their lengths are 1, 0, 2, 0, 0, and 0 (in order of occurrence).

Exercise 1: What does a streak length of 1 mean—that is, how many hits and misses are in a streak of 1? What about a streak length of 0?

The SAS macro **%calc_streak** can be used to calculate the lengths of all shooting streaks and then look at the distribution. This macro can be loaded using the FILENAME and %INCLUDE statements shown below. The syntax for the macro is shown at the end of this lab. The macro creates a new data set named **kobe_streak** that contains one variable, **streak**.

```
filename cstrk url
"http://www.openintro.org/stat/data/calc_streak.sas";
%include cstrk;

%calc_streak(dset=kobe, streakvar=basket, outset=kobe_streak);

proc sgplot data=kobe_streak;
   title "Basket Streaks by Kobe Bryant in 2009 (133 attempts)";
   vbar streak;
run;
```

Note that instead of making a histogram, we chose to make a bar plot from a table of the streak data. A bar plot is preferable here because our variable is discrete (counts) instead of continuous.

Exercise 2: Describe the distribution of Bryant's streak lengths from the 2009 NBA finals. What was his typical streak length? How long was his longest streak of baskets?

Compared to What?

We've shown that Bryant had some long shooting streaks, but are they long enough to support the belief that he had hot hands? What can we compare them to?

To answer these questions, let's return to the idea of *independence*. Two processes are independent if the outcome of one process doesn't affect the outcome of the second. If each shot that a player takes is an independent process, having made or missed your first shot will not affect the probability that you will make or miss your second shot.

A shooter with a hot hand will have shots that are ***not*** independent of one another. Specifically, if the shooter makes his first shot, the hot hand model says he will have a ***higher*** probability of making his second shot.

Let's suppose for a moment that the hot hand model is valid for Bryant. During his career, the percentage of time Bryant makes a basket (i.e. his shooting percentage) is about 45%, or in probability notation,

$$P(\text{shot 1 = H}) = 0.45$$

If he makes the first shot and has a hot hand (***not*** independent shots), then the probability that he makes his second shot would go up to, let's say, 60%,

$$P(\text{shot 2 = H} \mid \text{shot 1 = H}) = 0.60$$

As a result of these increased probabilities, you'd expect Bryant to have longer streaks. Compare this to the skeptical perspective where Bryant does ***not*** have a hot hand, where each shot is independent of the next. If he hit his first shot, the probability that he makes the second is still 0.45.

$$P(\text{shot 2 = H} \mid \text{shot 1 = H}) = 0.45$$

In other words, making the first shot did nothing to effect the probability that he'd make his second shot. If Bryant's shots are independent, then he'd have the same probability of hitting every shot regardless of his past shots: 45%.

Now that we've phrased the situation in terms of independent shots, let's return to the question: how do we tell whether Bryant's shooting streaks are long enough to indicate that he has hot hands? We can compare his streak lengths to someone without hot hands: an independent shooter.

Simulations in SAS

We don't have any data from a shooter we know to have independent shots, but that sort of data is very easy to simulate in SAS. In a simulation, you set the ground rules of a random process, and then the computer uses random numbers to generate an outcome that adheres to those rules. As a simple example, you can simulate flipping a fair coin with the following DATA step by using the RAND function in SAS:

```
data coin (keep = outcome);
   flip = rand("Bernoulli",0.5);
   if flip = 1 then outcome="heads"; else outcome="tails";
run;

proc print data=coin;
run;
```

The variable **flip** contains the result of a single Bernoulli trial, and it takes on a value of 1 with probability 0.5. We then convert the 0/1 values to tails and heads, respectively.

Run the code shown above several times. Just like when flipping a coin, sometimes you'll get a heads, sometimes you'll get a tails. But in the long run, you'd expect to get roughly equal numbers of each.

If you wanted to simulate flipping a fair coin 100 times, you could either run the program 100 times or, more simply, modify the DATA step to produce a data set containing the outcomes from 100 flips.

```
data sim_fair_coin(keep=outcome);
   do i = 1 to 100;
      flip = rand("Bernoulli",0.5);
       if flip = 1 then outcome="heads"; else outcome="tails";
       output;
   end;
run;
```

We can then create a frequency table to see the number of heads and tails in this simulation.

```
proc freq data=sim_fair_coin;
   tables outcome;
run;
```

We have specified that the probability that we flip a coin and it lands heads is 0.5. Now let's say that we're trying to simulate an unfair coin that we know lands on heads only 20% of the time. We can adjust for this by simply changing the Bernoulli probability in the RAND function.

```
data sim_unfair_coin(keep=outcome);
   do i = 1 to 100;
      flip = rand("Bernoulli",0.2);
      if flip = 1 then outcome="heads"; else outcome="tails";
      output;
   end;
run;
```

Exercise 3: In your simulation of flipping the unfair coin 100 times, how many flips came up heads?

Simulating the Independent Shooter

Simulating a basketball player who has independent shots uses the same mechanism that we use to simulate a coin flip. To simulate 100 shots from an independent shooter with a shooting percentage of 50%, we use the following DATA step:

```
data sim_basket(keep=basket);
   do i = 1 to 100;
      flip = rand("Bernoulli",0.5);
      if flip = 1 then basket="H"; else basket="M";
      output;
   end;
run;
```

To make a valid comparison between Bryant and our simulated independent shooter, we need to align both their shooting percentage and the number of attempted shots.

Exercise 4: What change needs to be made to the DATA step so that it reflects a shooting percentage of 45%? Make this adjustment, and then run a simulation to sample 133 shots.

Note that we've named the new data set **sim_basket**, the same name that we gave to the previous data set reflecting a shooting percentage of 50%. In this situation, SAS overwrites the old data set with the new one, so always make sure that you don't need the information in an old data set before running a new DATA step with the same name.

With the results of the simulation saved in the data set **sim_basket**, we have the date necessary to compare Bryant to our independent shooter. We can look at Bryant's data alongside our simulated data by merging the two data sets. Because the outcome variable is named **basket** in each of the data sets, we rename those variables as part of the merge.

```
data kobe_sim (keep=kobe sim);
   merge kobe(rename=(basket=kobe)) sim_basket(rename=(basket=sim));
run;

proc print data=kobe_sim;
run;
```

Both data sets represent the results of 133 shot attempts, each with the same shooting percentage of 45%. We know that our simulated data is from a shooter that has independent shots. That is, we know that the simulated shooter does not have a hot hand.

On Your Own

Comparing Kobe Bryant to the Independent Shooter

Using **%calc_streak**, compute the streak lengths of **sim_basket**.

1. Describe the distribution of streak lengths. What is the typical streak length for this simulated independent shooter with a 45% shooting percentage? How long is the player's longest streak of baskets in 133 shots?

2. If you were to run the simulation of the independent shooter a second time, how would you expect its streak distribution to compare to the distribution from the question above? Exactly the same? Somewhat similar? Totally different? Explain your reasoning.

3. How does Kobe Bryant's distribution of streak lengths from page 2 compare to the distribution of streak lengths for the simulated shooter? Using this comparison, do you have evidence that the hot hand model fits Kobe's shooting patterns? Explain.

4. What concepts from the textbook are covered in this lab? What concepts, if any, are not covered in the textbook? Have you seen these concepts elsewhere (for example, lecture, discussion section, previous labs, or homework problems)? Be specific in your answer.

The syntax for the **%calc_streak** macro is

```
%macro calc_streak(dset=, streakvar=, outset=);

data s1(keep=x);
   set &dset;
   if &streakvar="H" then x=1; else x=0;
run;

data xadd;
   x=0;
run;

data s2(keep=n);
   set xadd s1 xadd;
   n=_N_;
   if x=0;
run;

data &outset;
   set s2;
   retain lastn;
   if _N_=1 then lastn=n;
   else do;
      streak=n-lastn-1;
      lastn=n;
   end;
   if _N_=1 then delete;
run;

%mend;
```