Baseball VII

Produced by Dr. Mario | UNC STOR 538

Streakiness in Sports

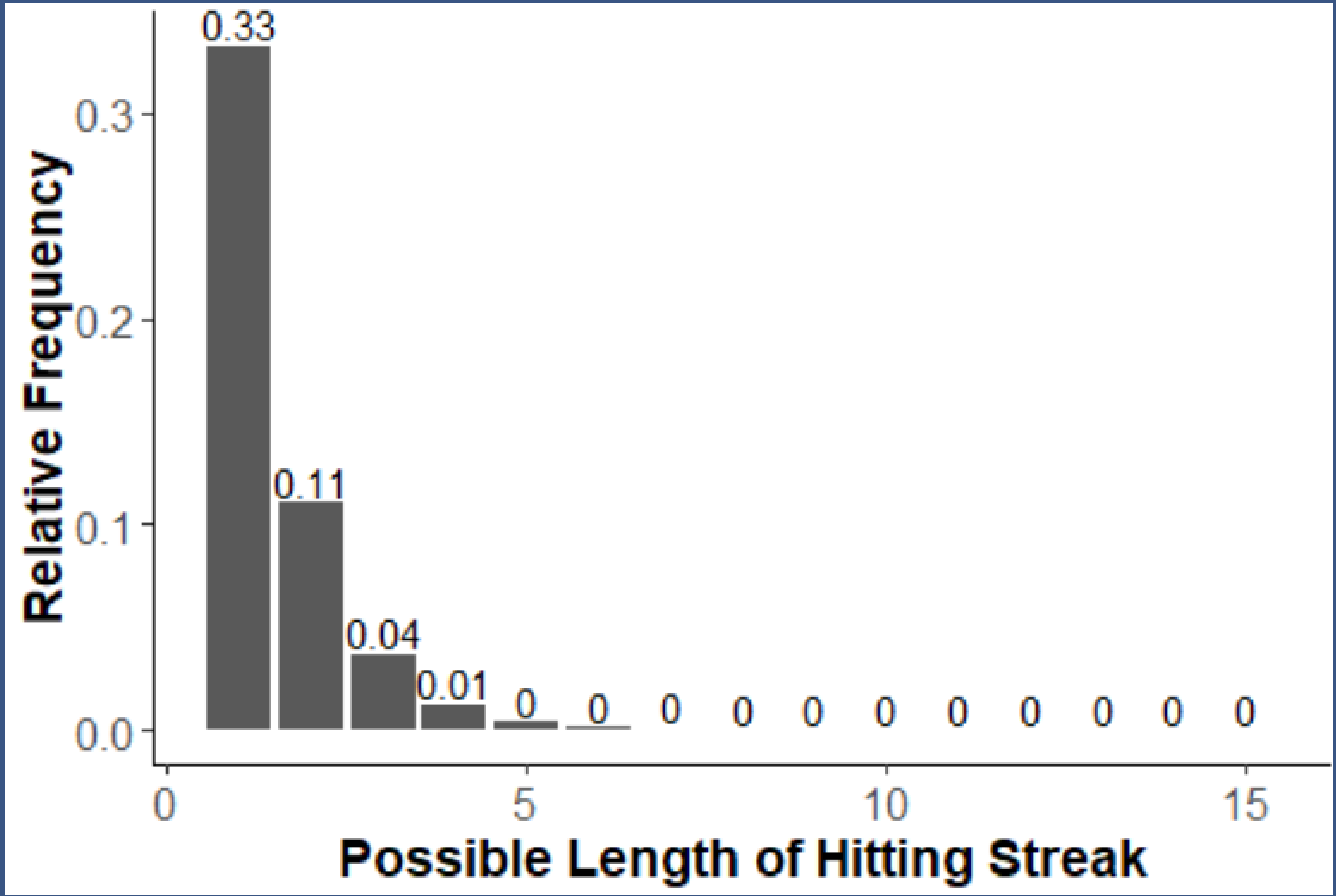
- Question? When Can We Say a Batter is HOT or COLD?
- Hypothetical Batter With Batting Average of 0.333
  - Each Plate Appearance, Batter has a 33.3% Chance of Hitting
  - HOT = Player Has an Unusual # of Consecutive Hits
  - COLD = Player Has an Unusual # of Consecutive Misses
  - Ignore Walks and Hit-by-Pitches
- Simulation
  - 1,000,000 Plate Appearances
  - 33.3% Chance of Hitting & 66.7% Chance of Not Hitting
  - Consider Possible Hitting Streaks and Hitting Slumps of 1 to 15
  - In 1 Million Plate Appearances, What Would be Considered a HOT Hitting Streak and COLD Hitting Slump?

Streakiness in Sports

```r
#Random Simulation of Hitting Streaks of Good Batter
Batting.Average=0.333
hit.sim=sample(x=c(0,1),size=1000000,replace=T,
               prob=c(1-Batting.Average,Batting.Average))
hitting.streak=1:15
streak.count=1:15

for(i in hitting.streak){
  n.streak=0
  count=0
  for(j in 1:(length(hit.sim)-i+1)){
    count=count+1
    if(sum(hit.sim[j:(j+i-1)]==1)==i){
      n.streak=n.streak+1
    }else{
      n.streak=n.streak+0
    }
  }
  hitting.streak[i]=n.streak
  streak.count[i]=count
}
```

Streakiness in Sports

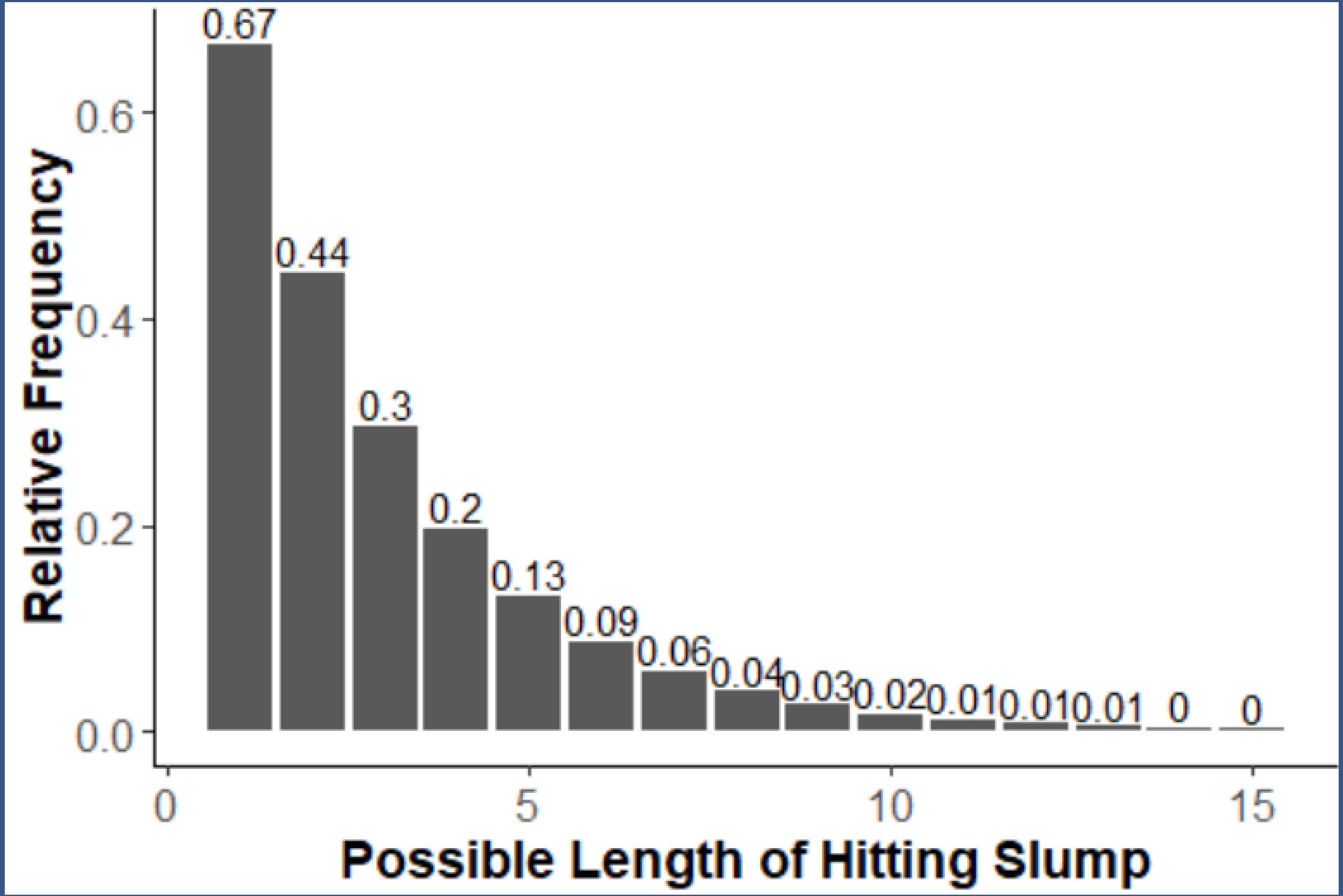


Streakiness in Sports

```r
#Random Simulation of Hitting Slumps of Good Batter
Batting.Average=0.333
hit.sim=sample(x=c(0,1),size=1000000,replace=T,
               prob=c(1-Batting.Average,Batting.Average))
hitting.slump=1:15
slump.count=1:15

for(i in hitting.slump){
  n.slump=0
  count=0
  for(j in 1:(length(hit.sim)-i+1)){
    count=count+1
    if(sum(hit.sim[j:(j+i-1)]==0)==i){
      n.slump=n.slump+1
    }else{
      n.slump=n.slump+0
    }
  }
  hitting.slump[i]=n.slump
  slump.count[i]=count
}
```

Streakiness in Sports



Streakiness in Sports

- R Code for Figures

```
ggplot(data=sim.data)+
  geom_bar(aes(x=length,y=hitting.streak),stat="identity")+
  geom_text(aes(x=length,y=hitting.streak,
                label=round(hitting.streak,2)), vjust=-0.2)+
  xlab("Possible Length of Hitting Streak")+
  ylab("Relative Frequency")+
  theme_classic()+
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14,face="bold"))

ggplot(data=sim.data)+
  geom_bar(aes(x=length,y=hitting.slump),stat="identity")+
  geom_text(aes(x=length,y=hitting.slump,
                label=round(hitting.slump,2)), vjust=-0.2)+
  xlab("Possible Length of Hitting Slump")+
  ylab("Relative Frequency")+
  theme_classic()+
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14,face="bold"))
```

Streakiness in Sports

- Wald Wolfowitz Runs Test (WWRT)
  - Topic Streakiness Pertaining to Wins (W) and Losses (L)
  - Suppose a Teams Record is 5-5 (W-L)
  - Streaky Would Be WWWWWLLLLL (2 Runs)
  - Not Streaky Would Be WLWLWLWLWL (10 Runs)
  - Idea: Fewer Runs = More Streaky
  - Let W=# of Wins, L=# of Losses, and T=W+L
  - According to Wold and Wolfowitz, if X=Number of Runs,

$$E[X] = \mu = \frac{2 \times W \times L}{T} + 1 \qquad VAR[X] = \sigma^2 = \sqrt{\frac{(\mu - 1)(\mu - 2)}{T - 1}}$$

  - For Team with 5-5 Record, $\mu = 6$ and $\sigma = 1.49$

$$Z_1 = \frac{2 - 6}{1.49} = -2.68 \qquad Z_2 = \frac{10 - 6}{1.49} = 2.68$$

Streakiness in Sports

- Hypothesis Test
  - Null: W's and L's are Randomly Distributed
  - Alternative: W's and L's are Streaky
  - Random Variable Z Has Approximate Normal Distribution if Number of Games T is Long Enough
  - If $Z < -2$, We Would Determine That Team is Streaky
  - Suppose in 162 Games, Team is 100-62 with 15 Runs
  - Test Statistic

```
> mu=2*100*62/162+1
> sd=sqrt((mu-1)*(mu-2)/(162-1))
> z=(15-mu)/sd
> print(c(mu,sd,z))
[1]  77.543210   5.992915 -10.436192
```

  - Conclusion: Ultra Streaky Bruh

Evaluating the Greatest Streak

- Joe DiMaggio
  - Played 13 Seasons With the New York Yankees
  - Known for 56 Game Hitting Streak (1941)
  - “Most Enduring Record in Sports” -*New York Times*
- Johnny Vander Meer
  - Known for Time With the Cincinnati Reds
  - No-Hitter Against the Boston Bees (June 11,1938)
  - No-Hitter Against the Brooklyn Dodgers (June 15, 1938)
  - No Other Pitcher Has Matched This
- What is the Most Difficult Achievement?

Evaluating the Greatest Streak

- Modeling Probabilities Using Poisson Distribution
  - Useful for Random Variable $X \in \{0,1,2,3,\ldots\}$
  - Probability Mass Function

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

  - Expected Value

$$E[X] = \lambda$$

  - Usage in R: Super Mario Averages 5 Shrooms Per Day

```
> dpois(7,lambda=5,log=F)
[1] 0.1044449
```

$$E[X] = \lambda = 5$$

$$P(X = 7) = 10.4\%$$

Evaluating the Greatest Streak

- Probability of Independent Events
  - If Events A and B are Independent,

$$P(A \cap B) = P(A) \times P(B)$$

  - Usage in R: Probability Super Mario Fasts for 5 Straight Days
    - Random Variables X_1, X_2, X_3, X_4, X_5
    - Assume They Are Independent and Identically Distributed

$$P(FAST) = P(X_1 = 0) \times P(X_2 = 0) \cdots \times P(X_5 = 0)$$
$$= P(X_1 = 0)^5$$

```
> dpois(0,lambda=5,log=F)^5
[1] 1.388794e-11
```

Evaluating the Greatest Streak

- How Rare Was Joe DiMaggio's Achievement?
  - Assumptions
    - Batters Need At Least 500 At-bats
    - Not Include Hitting Streaks Across Seasons
    - Batters with Over 500 At-bats Averaged 3.5 At-bats Per Game (Equivalent to 3 At-bats for Half Season and 4 At-bats for Remaining
  - Suppose Batter Hits .333 in 1900 (154 Game Season)
  - Probability of Event A3 = Batter Gets a Hit in 3 At-bat Game

$$P(A3) = 1 - (1 - .333)^3 = 70.33\%$$

  - Probability of Event A4 = Batter Gets a Hit in 4 At-bat Game

$$P(A4) = 1 - (1 - .333)^4 = 80.21\%$$

  - Probability of Event A = Hit During 56 Consecutive Games

$$P(A) = P(A3)^{28} \times P(A4)^{28} = 0.000011\%$$

Evaluating the Greatest Streak

- How Rare Was Joe DiMaggio's Achievement?
  - Number of Opportunities to Start Hitting Streak Where Batter is Hitless During the Previous Game

$$154 - 56 = 99\ Opportunities$$

  - Approximate Probability of Event E = Hitless Game

$$P(E) = \frac{(1 - P(A3)) + (1 - P(A4))}{2} = 24.7\%$$

  - Average Number of Opportunities to Start Winning Streak

$$1 + 98 \times 0.247 = 25.21\ Opportunites$$

  - Expected Number of 56 Game Hitting Streaks in a Season

$$25.21 \times P(\mathrm{A}) = 0.0000027$$

Evaluating the Greatest Streak

- How Rare Was Joe DiMaggio's Achievement?
  - Total Number of Batters Between 1900 and 2016 = 8233

```
> library(Lahman)
> Data=Batting %>%
+   filter(yearID>=1900 & yearID<=2016) %>%
+   filter(AB>=500) %>%
+   summarize(n=n())
> Data$n
[1] 8233
```

  - Expected Number of 56 Game Winning Streaks for All Batters

$$\lambda = E[Player_1] + E[Player_2] + \cdots + E[Player_{8233}] \approx .024$$

  - Probability of Event H = At Least 1 Hitting Streak of 56 Games

$$P(H) = 1 - P(\overline{H}) = 1 - \frac{\lambda^0 e^{-\lambda}}{0!} \approx 2.4\%$$

  - Batter With Batting Average of 0.33 Requires 9,926 Seasons to Have a 50% Chance of Getting the 56 Game Winning Streak

Evaluating the Greatest Streak

- How Rare Was Johnny Vander Meer's Achievement?
  - Assumptions
    - All Games Are Started by Pitchers Who Start Exactly 35 Games (Exactly 951 Pitchers Under This Criteria from 1900 to 2016)
    - Assume Probability of No Hitter is 0.062% for All Pitchers for Every Single Game
  - Following Similar Ideas from DiMaggio, the Probability of Event N = At Least 1 Starting Pitcher Would Throw Consecutive No Hitters

$$P(N) = 15.7\%$$

- Both Achievements Are Unlikely But Possible
- Both Achievements Become More Likely As Time Passes

Evaluating the Greatest Streak

- What is the Most Difficult Achievement?

Trick Question…

Lebron James Winning a Championship for Cleveland #216

Final Inspiration

So I’m ugly. So what?

I never saw anyone hit with his face.

-Yogi Berra