A Fresh Shot at Statistics in the Classroom: Three Perspectives using World Cup Soccer Player Data

Brenna Curley

MORAVIAN COLLEGE

Anna Peterson

IOWA STATE UNIVERSITY

USCOTS 2021

The Data: FIFA World Cup & Women's World Cup

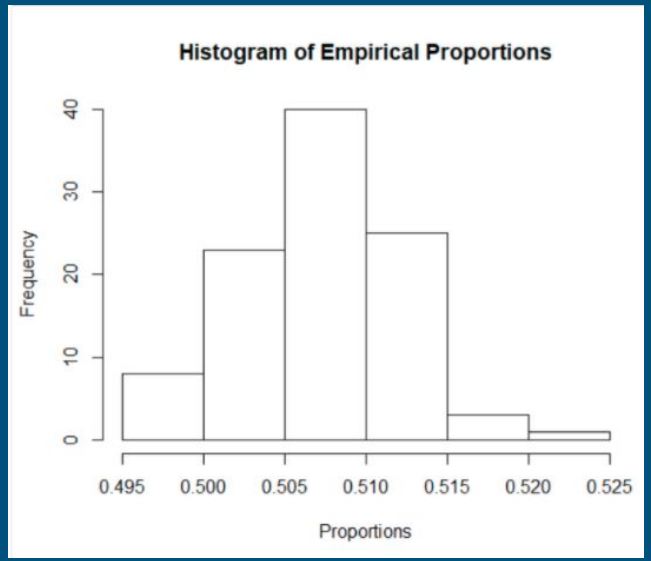
Data are publicly available from Wikipedia:

- 2019 FIFA Women's World Cup squads
- 2018 FIFA World Cup squads

Combined dataset includes n=1288 rows (736 men and 552 women players) and 15 variables

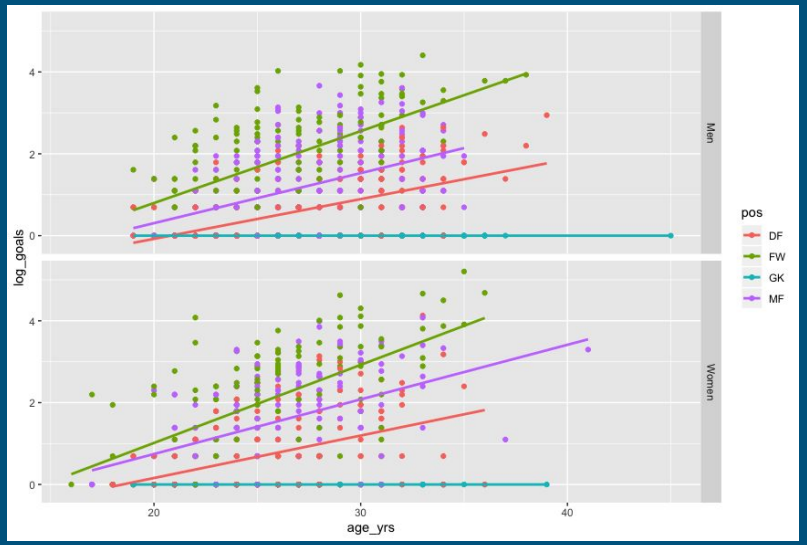
team_id	no	pos	player	age_yrs	caps	goals	club	DOB	Year	Month	Day	country	group	WorldCup
24	10	FW	Carli Lloyd (co-captain)	36	271	107	Sky Blue FC	1982-07-16	1982	7	16	United States	F	Women
23	10	DF	Sunisa Srangthaisong	31	*NA*	*NA*	Bundit Asia	1988-05-06	1988	5	6	Thailand	F	Women
22	10	FW	Sofia Jakobsson	29	101	17	Montpellier	1990-04-23	1990	4	23	Sweden	F	Women
21	10	FW	Yanara Aedo	25	20	9	Valencia	1993-08-05	1993	8	5	Chile	F	Women
24	9	MF	Lindsey Horan	25	66	8	Portland Thorns	1994-05-26	1994	5	26	United States	F	Women
23	9	DF	Warunee Phetwiset	28	*NA*	*NA*	Chonburi Sriprathum	1990-12-13	1990	12	13	Thailand	F	Women
22	9	MF	Kosovare Asllani	29	127	32	Linkoping	1989-07-29	1989	7	29	Sweden	F	Women
21	9	FW	Maria Jose Urrutia	25	7	0	3B da Amazonia	1993-12-17	1993	12	17	Chile	F	Women
24	8	MF	Julie Ertz	27	79	18	Chicago Red Stars	1992-04-06	1992	4	6	United States	F	Women

Three Perspectives



Probability

Multivariable Thinking




```
soccer %>%
  filter(WorldCup == "Women") %>%
  select(Day, Month, team_id) %>%
  group_by(team_id) %>%
  distinct(Month, Day) %>%
  summarise(cnt = n(), cnt23 = (cnt != 23)) %>%
  summarise(num_teams = n(), prop = sum(cnt23)/n())
```

Data Science

USCOTS 2021

A Probability Perspective

Birthday matches in one example?

Birthday matches in a team of 23?

Birthday matches in many teams?

13/24 =0.54

Simulation and Connection to LLN

Country	Group/Student	Circle One	
France		MATCH	ALL DIFFERENT
Nigeria		MATCH	ALL DIFFERENT
Norway		MATCH	ALL DIFFERENT
South Korea		MATCH	ALL DIFFERENT
China PR		MATCH	ALL DIFFERENT
Germany		MATCH	ALL DIFFERENT
South Africa		MATCH	ALL DIFFERENT
Spain		MATCH	ALL DIFFERENT
Australia		MATCH	ALL DIFFERENT
Brazil		MATCH	ALL DIFFERENT
Italy		MATCH	ALL DIFFERENT
Jamaica		MATCH	ALL DIFFERENT
Argentina		MATCH	ALL DIFFERENT
England		MATCH	ALL DIFFERENT
Japan		MATCH	ALL DIFFERENT
Scotland		MATCH	ALL DIFFERENT
Cameroon		MATCH	ALL DIFFERENT
Canada		MATCH	ALL DIFFERENT
Netherlands		MATCH	ALL DIFFERENT
New Zealand		MATCH	ALL DIFFERENT
Chile		MATCH	ALL DIFFERENT
Sweden		MATCH	ALL DIFFERENT
Thailand		MATCH	ALL DIFFERENT
United States		MATCH	ALL DIFFERENT

A Data Science Perspective

The birthday paradox doesn't need to reside solely in a probability course.

Data Wrangling & Summarizing → Conceptual Understanding of Probability → Modeling & Inference

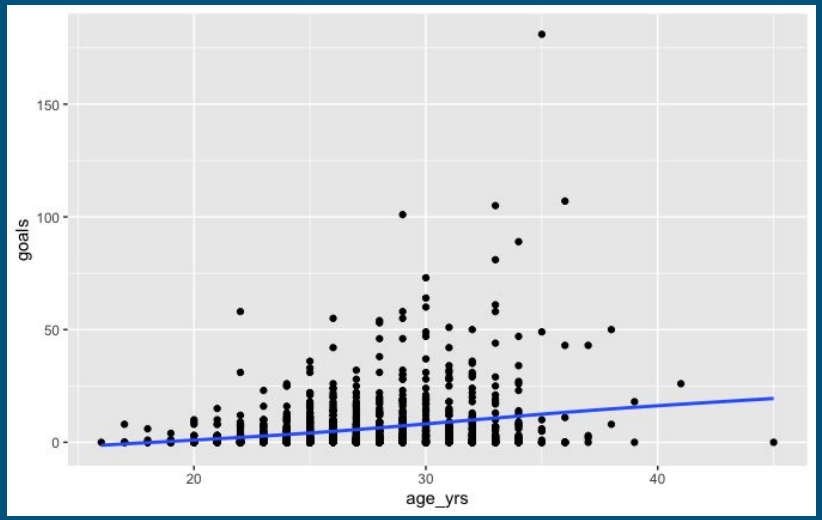
```
soccer %>%
  filter(WorldCup == "Women") %>%
  select(Day, Month, team_id) %>%
  group_by(team_id) %>%
  distinct(Month, Day) %>%
  summarise(cnt = n(), cnt23 = (cnt != 23)) %>%
  summarise(num_teams = n(), prop = sum(cnt23)/n())
```

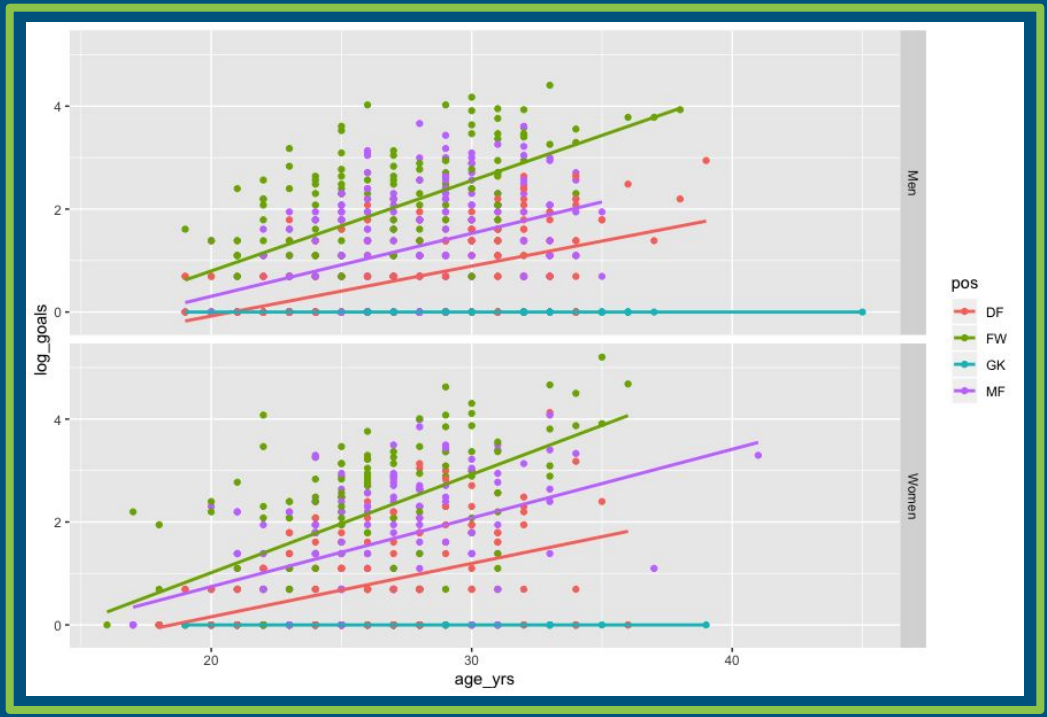
Students will,

- Create a **tidy dataframe** using the *tidyverse* library
- Compute **relative frequencies** for the proportion of teams with less than 23 distinct birthday
- **Compare** their computed proportions to their intuition

A Multivariable Perspective

Example: ***Influence of age on (log) goals scored (by men/women & field position)***





Concluding Remarks

The shared activities align with the revised GAISE college report:

- ➔ Data with a real-world context: **FIFA World Cup soccer players**
- ➔ Problem solving with hands-on experiences:

  e.g., **Empirical & Theoretical Probabilities (Birthday Paradox)**

- ➔ Active learning encouraged as students work in **collaborative groups**
- ➔ Students use the statistical software, **R (base R and *tidyverse*)**

Thank You!

Brenna Curley

Contact:
curleyb@moravian.edu

Anna Peterson

Contact:
ericksad@iastate.edu