Exam 2 review

Name: _____

November, 2018

Topics for in class exam

- Applications of risk aversion/minimax/prospect theory
- How to calculate expected points/expected points added
- When do field goals go in?
- Hockey analytics terms: Corsi, CorsiRel, CF60, PDO
- James-Stein estimator vs. MLE
- Mean absolute error
- James-Stein Paradox
- Density curves and interpretation

Topics for take-home exam

- Data manipulation skills using `group_by`, `summarise`, `top_n`, `slice`, `filter`
- NFL data (see labs and HWs)
- NHL data (see labs and HWs)

Sample exam questions

We use data similar to what was used in HW 6. In this case, we look at all rushing plays during the 2017 season. **Note: please put the URL on one line when reading into RStudio**

```
library(RCurl)
library(tidyverse)
url.17 <- getURL("https://raw.githubusercontent.com/ryurko/nflscrapR-data/master/
                 play_by_play_data/regular_season/reg_pbp_2017.csv")
pbp <- read.csv(text = url.17)
head(pbp)
rushes <- pbp %>% filter(play_type == "run")
rushes <- rushes %>%
  select(game_id, home_team, away_team, posteam, yardline_100, game_seconds_remaining,
         qtr, down, ydstogo, yards_gained, shotgun, no_huddle, run_gap,
         run_location, epa, wpa, rusher_player_name) %>% na.omit()
```

Compare only James White (`rusher_player_name == "J.White"`) and Kareem Hunt (`K.Hunt`) on their rushes.

1. What is the average epa and median epa on rushes for White and Hunt? What do you notice?
2. Make density curves of EPA on rushes by Hunt and White. Describe differences in the center, shape, and spread of each curve, and be sure to tie your answer into your findings in Question 1.
3. Is there a difference in each players' number of carries? Why would that matter?
4. Using all players (not just White and Hunt), make density curves comparing the `ydstogo` based on `factor(down)` and identify one interesting aspect of the chart.

5. Among players with at least 100 carries, identify the three with the highest epa and identify the three with the lowest epa. What is roughly the difference in the change in expected points per player between the top three and bottom three players?

6. A football coach notes that comparing running backs based on expected points added is particularly unfair. Why might be be correct? Consider expected points added in the framework of how we judge metrics (recall: there are three properties)

7. The following code creates a new varible called `success.rush`, which is a 1 if a run has a positive EPA and a 0 otherwise. It also creates a new data set, `rushes.1`, which uses players with at least 10 carries, and a summary data set, `player.sum`, that has success rates among players who rushed at least 10 times.

```
rushes <- rushes %>% mutate(success.rush = ifelse(epa > 0, 1, 0))
rushes1 <- rushes %>% group_by(rusher_player_name) %>% summarise(n.carries = n()) %>%
  filter(n.carries >= 10) %>% inner_join(rushes)

player.sum <- rushes1 %>%
  group_by(rusher_player_name) %>%
  summarise(ave.success.rate = mean(success.rush), n.carries = n())
```

7a. Which player has the highest success rate?

7b. Calculate the maximum liklihood estimator of average success rate among all players with at least 10 carries. You do not need to show this output, but you should be identify how this is done.

7c. Calculate the James-Stein estimator of average success rate among all players with at least 10 carries. To do this, calculate each step of the process as defined in Lab 7: `p.bar`, `p.hat`, `N`, `sigma.sq`, `c`. Find the shrinkage constant `c` for the first player in `player.sum` (A.Abdullah) and interpret what it means. For most players, is success rate more linked to player skill or random chance?

Solutions

In words.

1. Hunt has higher EPA (0.06 versus -0.03) but lower median EPA (-0.22 versus -0.01) when compared to White.

2. White's curve is slightly shifted to the right when compared to Hunt. This matches his higher median EPA. However, Hunt has a few outliers to the right. These carries likely shift his overall average up so that it is is higher than White's. Both distributions are skewed right. The peak for Hunt is a bit higher around 0 than it is for White.

3. Yes: White has 24 carries, Hunt with 157. This means we are more certain of Hunt's overall performance than we are of White's.

4. Answers will vary. Interesting that the third down curve has a longer curve to the right when compared to other downs.

5. Highest three are Kamara, Lewis, Elliot: lowest three are Abdullah, Powell, Murray. The difference is about 0.25 EPA per carry.

6. Changes in a plays expected points are attributed to all players on offense, not just running backs. If certain running backs have bigger holes to run through, or play worse defenses, that is not neccessarily something attributable to running backs. Other factors are likely having an influence on EPA.

7a. Matt Ryan

7b. The MLE is each players' average success rate.

7c. See code. Typical c values vary between 0.5 and 0.9. This implies success rate is more skill than luck – each players J-S estimate is between 50 percent of their past MLE and 90 percent of their past MLE, with the rest shrunken towards the league average.

Code

```
hunt.kamara <- rushes %>% filter(rusher_player_name == "K.Hunt"|rusher_player_name == "J.White")
hunt.kamara %>% group_by(rusher_player_name) %>%
  summarise(ave.epa = mean(epa), median.epa = median(epa), n.carries = n())

ggplot(data = hunt.kamara, aes(x = epa, colour = rusher_player_name)) +
  geom_density()

ggplot(data = rushes, aes(x = yards_gained, colour = factor(down))) +
  geom_density()

rushes %>%
  group_by(rusher_player_name) %>%
  summarise(n.carries = n(), med.epa = median(epa)) %>%
  filter(n.carries >= 100) %>%
  top_n(3, med.epa)

rushes %>%
  group_by(rusher_player_name) %>%
  summarise(n.carries = n(), med.epa = median(epa)) %>%
  filter(n.carries >= 100) %>%
  top_n(3, -med.epa)


player.sum %>% arrange(-ave.success.rate)

p.bar <- mean(player.sum$ave.success.rate)
p.bar
p.hat <- player.sum$ave.success.rate
p.hat

N <- player.sum$n.carries
N
sigma.sq <- sd(p.hat)^2
sigma.sq

c <- (N/0.25)/(N/0.25 + 1/sigma.sq)
c

player.sum$c <- c
player.sum$sr.MLE <- player.sum$ave.success.rate
player.sum$sr.JS <- p.bar + c*(p.hat - p.bar)
```