Exploring Sampling Distributions Using R

A sampling distribution tells us what could possibly happen when we take a sample of size n from a known population and summarize the sample with a statistic (mean, median, maximum, etc.). As we rarely work with known populations in the real world, sampling distributions are rarely directly used without some assumptions. For instance, we assume that a null hypothesis is true when generating a sampling distribution.

1. Read in the tidyverse package and the provided UF2022 dataset.
2. Find the mean number of goals per player during the 2022 season for the Summit team and AlleyCats team individually. (Hint: Use a dplyr function.)
3. Are these values parameters or statistics?
4. In our case, n is the number of players, and this question will explore the sampling distribution of mean goals per player for the Colorado Summit and Indianapolis AlleyCats during the 2022 AUDL season. The code below can be used to generate a sampling distribution of mean goals per player across n players on a certain team. Copy and paste or write the code below into R.

```{r}

n =   
nsim = 5000   
my.means = numeric(nsim)   
  
for (i in 1:nsim)   
{  
 my.sample = UF2022 %>% filter(Team == "team\_name") %>% select(GLS) %>% slice\_sample(n = n, replace = TRUE)

my.means[i] = mean(my.sample$GLS)   
}  
  
hist(my.means, main=paste("Sampling Distribution of the \n Sample Mean when n =",n))

mean(my.means) ##sampling distribution mean

sd(my.means) ##sampling distribution standard deviation

```

1. Modify the code to generate a sampling distribution of mean goals per player across 5, 10, then 25 players on the Colorado Summit. What do you notice about the shape, center (mean) and spread (standard deviation) of the sampling distribution as n increases?
2. Repeat part a) for the Indianapolis AlleyCats.
3. What property/theorem do your findings illustrate?
4. Repeat questions 2-4 for the +/- (`plusminus`) variable instead of the goals (`GLS`) variable. Do you notice any difference in how the sampling distributions act when different variables are used?