Class Activity: Sampling Distributions

1. Let us take as the population the baseball players in MLB who, during the 2018 season, had at least 100 at-bats. You can read this data set in with the command:

   ```
   filter(read.csv("https://www.openintro.org/data/csv/mlb_players_18.csv"), AB >= 100)
   ```

   How would you describe the distribution of homeruns? What is the population mean?

2. Now generate an approximate sampling distribution of $\bar{x}$, the sample mean number of homeruns for samples of size $n = 20$. How would you describe this sampling distribution? What are its mean and standard deviation? (There is another name for this standard deviation; do you know it?)

3. If you display your sampling distribution of $\bar{x}$ using `gf_dhistogram()` (usage is like that of `gf_histogram()`), and then follow it with the pipe and extra command as displayed

   ```
   gf_dhistogram( ... ) %>% gf_dist("norm", mean=you supply, sd=you supply)
   ```

   does it appear the sampling distribution is well-approximated by this normal distribution?

4. Play with the "sampling distribution for $\bar{x}$" app at the website `https://shiny.calvin.edu:3838/scofield/cltMeans/`. In particular, select different populations and different sample sizes. Is there a minimal sample size n that always seems to result in a normal-looking sampling distribution?

5. Play with the "sampling distribution for $\widehat{p}$" app at the website `https://shiny.calvin.edu:3838/scofield/cltProportions/`. In particular, select different population parameters p and different sample sizes. Can you find some sort of "rule" for the choice of n and p which seems to ensure a normal-looking sampling distribution?