Homework 1: review and coding questions

your name

Due: 01/26 at 11:59 PM

This assignment is meant to serve multiple objectives:

- You will gain familiarity with R, RStudio, R Markdown, and GitHub
- You will perform at least one iteration of the courses's data science inspired workflow
- You will gain experience with typing mathematics
- You will learn some `dplyr` and `ggplot2` basics

STAT 528 is a collaborative course environment, especially for assignments that involve coding, modeling, and/or data analysis. You are encouraged to ask for help from other students. Coding and data science work flow can be very tedious. Having someone else look over your work or answering a basic question can save you a lot of time. However, direct copying is not accepting. All final work must be your own.

Mathematical review questions

Problem 1: Prove that the Binomial distribution arises as a sum of n iid Bernoulli trials each with success probability p.

Problem 2: Let $l(\theta)$ denote a twice continuously differentiable log likelihood corresponding to an iid sample under density f_θ where n is the sample size. The score function is defined as

$$u(\theta) = \frac{\partial l(\theta)}{\partial \theta},$$

and the Fisher information matrix is defined as

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right),$$

where the expectation is over the assumed distribution for the data when the parameter value is θ. Prove that

$$\mathrm{E}(u(\theta)) = 0 \qquad \text{and} \qquad \mathrm{Var}(u(\theta)) = I(\theta).$$

Coding questions

Problem 3: The data we will use to accomplish this task will come from Lahman's Baseball Database. Thankfully, there is an R package, `Lahman`, that makes importing this data into R very easy. If you have not done so previously, install this package using:

```
install.packages("Lahman")
```

While there many metrics that could be used to determine who is the "best" baseball player, because we are focusing on batters, we will use the on-base plus slugging (OPS) statistic. This statistic measures both a batter's ability to "get on base" and "hit for power."

- YouTube: Moneyball, "He Gets on Base"

Additionally, our definition of "best" will be based on a player's career statistics, but an alternative argument could be made based on single season efforts.

After loading the Lahman package, you will have access to several data frames containing historical baseball data from 1871 - 2022. You will need to interact with the following data frames:

- Schools
- CollegePlaying
- Batting
- People

You should spend some time exploring these datasets and reading the relevant documentation.

You should spend some time exploring these datasets and reading the relevant documentation.

Create a tibble named `illini_mlb_batters` that contains the following elements, in this order:

- playerID
- nameFirst
- nameLast
- birthYear
- G
- AB
- R
- H
- X2B
- X3B
- HR
- RBI
- SB
- CS
- BB
- SO
- IBB
- HBP
- SH
- SF
- GIDP
- PA
- TB
- BA
- OBP
- SLG
- OPS

The rows of the tibble should be sorted from highest OPS to lowest OPS. Each row should represent the career statistics for the player with ID playerID. Only include players that had at least one at-bat and one plate appearance. Except for PA, TB, AVG, OBP, SLG, and OPS, the (sometimes season-level) variables listed can be found in one of the four data frames listed above. The remaining values can be calculated as follows:

- PA = AB + BB + HBP + SH + SF
- TB = H + X2B + 2 * X3B + 3 * HR
- BA = H / AB
- OBP = (H + BB + HBP) / (PA - SH)
- SLG = TB / AB

- OPS = OBP + SLG

Round any rate statistics to three decimals places, as is customary in baseball.

Problem 4: The data we will use to accomplish this task will come from the `Teams` data frame in Lahman's Baseball Database. In this problem we will visualize the Pythagorean Theorem of Baseball. This "Theorem" states that winning percentage is given by the following nonlinear equation:

$$WP = \frac{R^2}{R^2 + RA^2}$$

where

- WP is winning percentage
- R is total runs scored by a baseball team
- RA is total runs allowed by a baseball team

For this problem, plot the estimated number of wins as predicted by the Pythagorean equation and actual wins (denoted W). The estimated number of wins as predicted by the Pythagorean equation

$$162 * \frac{R^2}{R^2 + RA^2}.$$

Provide a line of best fit. Restrict attention to the 1990 season and beyond. Note that there are two shortened seasons that need to be treated separately from the remaining seasons. These seasons are 1994 and 2020. The 1994 season was cut short because of a labor strike. The 2020 season was cut short due to COVID.