Please provide complete and well-written solutions to the following exercises.

Due October 28, 12PM noon PST, to be uploaded as a single PDF document to blackboard (under the Assignments tab).

Homework 5

Exercise 1. I believe that the number of home runs hit by an MLB baseball player in a single season satisfies a Poisson distribution with some unknown parameter $\lambda > 0$. In this exercise, let's try to find the parameter $\lambda > 0$ that best fits the data, using whatever estimation method you want (e.g. MLE is fine).

Here the data can be found from:

http://www.seanlahman.com/baseball-archive/statistics

I recommend using the 2020 Version, comma delimited version. The data is in a zip file, and home run data can be found in `Core` then `batting.csv` then the column HR.

After fitting the Poisson distribution to the data, compute the total variation distance of the data from the fitted Poisson distribution. If P, Q are two probability laws on e.g. the positive integers, then the total variation distance between P and Q is

$$||P - Q||_{\text{TV}} := \frac{1}{2} \sum_{k=0}^{\infty} |P(k) - Q(k)| .$$

Here P would be the fitted Poisson distribution, and Q would be the probability distribution corresponding to the data. If $||P - Q||_{\text{TV}}$ is close to 0, then the Poisson distribution that you found fits well to the data. If $||P - Q||_{\text{TV}}$ is far from 0 (perhaps close to 1), then the Poisson distribution that you found does not fit the data well.

Try to answer the same question as above for the number of made 3 point shots among 2020 WNBA players. Data can be found here:

(this link).

The data we are particularly interested in is the column `3P`.

Exercise 2. Wikipedia has a list of most played video games with at least 10 million users, sorted by the maximum number of registered players for that game:

(this link)

This is an open-ended question, related to this list.

Plot a histogram of the player counts of this list of games (i.e. the second column of the table). Does this histogram resemble any particular distribution? If so, try to fit that distribution to the data, as in the previous question, and use the total variation distance as a measure of goodness of fit.

Exercise 3. Suppose you flip a coin 1000 times, resulting in 560 heads and 440 tails. Is it reasonable to conclude that the coin is fair (i.e. it has one half probability of heads and one half probability of tails)? Justify your answer.

Exercise 4. Suppose the number of typos in my notes in a given year follows a Poisson distribution. In the last few years, the average number of typos was 15, and this year, I had 10 typos in my notes. Is it reasonable to conclude that the rate of typos has dropped this year? Justify your answer.

Exercise 5. Suppose X is a Gaussian distributed random variable with known variance $\sigma^2 > 0$ but unknown mean. Fix $\mu_0, \mu_1 \in \mathbf{R}$. Assume that $\mu_0 - \mu_1 > 0$. We want to test the hypothesis H_0 that $\mu = \mu_0$ versus the hypothesis H_1 that $\mu = \mu_1$. Fix $\alpha \in (0,1)$. Explicitly describe the UMP test for the class of tests whose significance level is at most α.

Your description of the test should use the function $\Phi(t) := \int_{-\infty}^{t} e^{-x^2/2} dx/\sqrt{2\pi}$, $\Phi\colon \mathbf{R} \to (0,1)$, and/or the function $\Phi^{-1}\colon (0,1) \to \mathbf{R}$. (Recall that $\Phi(\Phi^{-1}(s)) = s$ for all $s \in (0,1)$ and $\Phi^{-1}(\Phi(t)) = t$ for all $t \in \mathbf{R}$.)

Exercise 6. Let $X_1, \ldots, X_n$ be i.i.d. random variables, and denote $X = (X_1, \ldots, X_n)$. Let $\{f_\theta\colon \theta \in \Theta\}$ be a family of multivariable PDFs. Assume that X has distribution f_θ. Suppose Θ consists of two points, i.e. $\Theta = \{\theta_0, \theta_1\}$. Let Z be a sufficient statistic for θ. Consider the likelihood ratio test of the null hypothesis H_0 that $\theta = \theta_0$ versus the alternative H_1 that $\theta = \theta_1$.

Show that the likelihood ratio is a function of Z.

Exercise 7. Suppose X is a binomial distributed random variable with parameters $n = 100$ and $\theta \in [0,1]$ where θ is unknown. Suppose we want to test the hypothesis H_0 that $\theta = 1/2$ versus the hypothesis H_1 that $\theta \neq 1/2$. Consider the hypothesis test that rejects the null hypothesis if and only if $|X - 50| > 10$.

Using e.g. the central limit theorem, do the following:

- Give an approximation to the significance level α of this hypothesis test
- Plot an approximation of the power function $\beta(\theta)$ as a function of θ.
- Estimate p-values for this test when $X = 50$, and also when $X = 70$ or $X = 90$.