CSE 5520 Fall 2021

Homework 2 (Due midnight September 15, 2021 at HuskyCT) - revised

Basic Statistics Concepts and Visualization

This homework is to help you refresh upon some basic statistics concepts and also to gently introduce basic data visualization fundamentals related to statistical concepts. Please be mindful that many of the topics you will be exposed to from this exercise will be reused in future lecture and homework. You are required to do this exercise in Python.

Part 1: Normal Distribution

Step 1. Download the NBA data (nba.csv) from HuskyCT which is available in the directory, called "Data Files". This file is in csv format.

Step 2. Compute the mean and standard deviation of the heights of NBA players, say, mu and sigma, respectively.

Step 3. Create a histogram summarizing the NBA player's height distribution. Think what would be the desired x-axis's interval for the plot.

Step 4. Show the distribution using boxplot.

Step 5. Show the distribution using violin plot.

Step 6. Convert the histogram you created in Step 3 using z-values for the heights of the players. Think what would be the desired x-axis's interval for the plot.

Step 7. Compute the mean and standard deviation of the heights of NBA players in terms of z-values, say, mu_z and sigma_z, respectively.

Step 8. Show the distribution using boxplot using z-scores.

Step 9. Show the distribution using violin plot using z-scores.

Step 10. Create a normal distribution graph using mu and sigma you created in Step 2 (you are using formula for this) and overlay it on the histogram you created in Step 3. You can do this using matplotlib.pyplots.subplots.

Step 11. Create a normal distribution graph using mu_z and sigma_z you created in Step 7 (you are using formula for this) and overlay it on the histogram you created in Step 6. You can do this using matplotlib.pyplots.subplots.

Step 12. Discuss the difference between the two sets of overlaid graphs you produced from Steps 10 and 11. Are the shapes different? How different? Do they say anything about the difference between histograms and PDFs?

Step 13. Calculate $P(x > 82)$, i.e., what is the probability for a player to have his/her height taller than 82?

Part II: Poisson, and Gamma Distributions

The Connecticut State Museum of Natural History at UCONN holds annual mushroom hunting trips in a mountain in Ashford. It is a half day trip lasting about 4 hours in a weekend in October. It is known that each participant collects about 8 mushrooms in the trip and the event's program fee is $20. You are pondering if you like to participate in this year's event by considering various factors.

Step 1. What is the probability that you collect "at least" 10 mushrooms? Since you are paying $20 fee, you are interested in figuring out if you can recoup your investment by estimating $2 per fresh organic mushroom. As long as the chance is over 60%, you like to register the event. Show your decision with two accompanying probability distribution plots, one PMF and one CDF. Show also the simulated PMF using poisson.rvs. You show three plots in total.

Hint: Study scipy.stats.poisson
Search stackoverflow with "how to plot exponential cdf with specific λ"
https://projector-video-pdf-converter.datacamp.com/14568/chapter3.pdf

Step 2. You are anxious to know how quickly you can collect the first mushroom from the trip. This is because your priority is not recouping your investment for the event but to enjoy the excitement of finding the first fresh wild mushroom and come back to complete the homework after you found one. What is the probability that you can get the first mushroom within one hour of the trip? As long as the chance is over 60%, you like to register the event. Show your decision with an accompanying probability distribution plot.

Step 3. You are exploring a different scenario. Why not recoup some investment and also come back early to complete the homework? What is the probability that you can get the four mushrooms within 2 hours of the trip? As long as the chance is over 60%, you like to register the event. Show your decision with an accompanying probability distribution plot.

Part III: Beta Distribution

Consider back to NBA statistics. Giannis Antetokounmpo having nick name "Greek Freak" was instrumental in helping Milwaukee win 2020-2021 NBA championship. His FG% for the season was 0.569 (Wikipedia).

Step 1. Assume his FG% was from making 1140 goals from 2003 shots (attempts). Create a Beta distribution plot for Giannis Antetokounmpo's FG% for that season.

Step 2. Suppose this Greek Freak did 27 goals from 35 shots during the first game in Fall 2021. Can you estimate his FG% for the upcoming season using beta distribution? You are approximating α and β for this and rely on the python function to create the Beta distribution plot similar to what has been discussed in class, i.e., show both prior and posterior distributions.

You upload your Jupyter notebook in HuskyCT. The file name should be of the following format: HWn_JohnDoe where n is the homework number and John Doe denotes first name followed by last name.

The report can include figures, tables and graphs with type-written in 8 x11 inch paper with 1 inch margin top, bottom and side. There is no page limit in your report.

Instead of having a separate PDF report, you can incorporate your report into Jupyter notebook (.ipynb file) using markdown cells and submit one HW2 report.

HWs and Projects, 5% penalty for one day late submission. No acceptance after 5 days late. Extension is allowed only with the supporting medical record.