Homework I: Preliminaries on Probability Concepts

Solutions to this assignment are to be submitted in myCourses via Dropbox. The submission deadline is **Saturday September 21, 2019 at 11:59pm**. You should submit a zipped file containing a pdf with written answers and a Jupyter notebook with any code you write. Use comments to explain your code. All code and plots should be in the notebook while descriptions should be in the PDF.

1. (25 points) **Statistics with discrete distributions [PDF]** Consider the drug Zyban which is meant to create the urge to quit smoking. In clinical trials 35% of a study participants experienced insomnia when taking 300mg of Zyban per day [*Source*: GlaxoSmithKline]

   A random sample of 25 Zyban users is obtained and the number who experienced insomnia is recorded. Using the table provided, provide detailed answers to questions (a)-(e):

x	P(X=x)
0	2.10297E-05
1	0.0003
2	0.0018
3	0.0076
4	0.02224
5	0.0506
6	0.0908
7	0.1327
8	0.1607

   (a) Find the probability that exactly 8 users experienced insomnia as a side-effect

   (b) Find the probability that less than 4 users experienced insomnia as a side-effect

   (c) Find the probability that 5 or more users experienced insomnia as a side-effect

   (d) Find the probability that 20 users do not experience insomnia as a side-effect

   (e) Lastly, is the probability of observing 2 or fewer users having insomnia an unusual event? Explain why or why not using probabilities.
   NOTE: In this context, we say that an event with a probability less than 2% is unusual

2. (25 points) **Simulating a long-run relative frequency [Notebook+PDF]**
   Goal: Toss a coin N times and compute the running proportion of heads.
   Suppose we want to know the long-run relative frequency of getting heads from a fair coin. All we know is that there's some underlying process that generates an 'H' or a 'T' when we sample from it. The process has a parameter called θ, whose value is $\theta = 0.5$. If that's all we know, then we can approximate the long-run probability of getting an 'H' by simply repeatedly sampling from the process. We sample from the process N times, tally the number of times an 'H' appeared, and estimate the probability of H by the relative frequency: $\hat{\theta} = \frac{\#H}{N}$. Show your proportion of heads versus flip number (flip number is on a log scale to show details of the first few flips)

   - Set the"seed" for the Numpy random number generator.
   - Sample the input vector [0,1] and uniform distribution sampling probability (with replacement) using the numpy command:
     `flipsequence = np.random.choice(a=[0,1] , p=[.5,.5] , size=N , replace=True)`

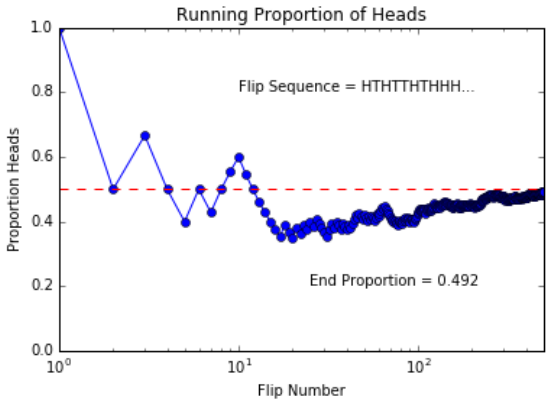

Figure 1: Running proportion of heads when flipping a coin. The x-axis is plotted on a logarithmic scale to show the details of the first few flips and also the long-run trend after many flips.

- Compute the running number of heads in your `flipsequence`
  NOTE: Numpy has the in-built function `cumsum` that can be used to compute this, but it is not required you use it
- Compute the running proportion by doing {`cumsum`}/{`N`}, component-by-component.
- Graph the running proportion against the flip number to see how the probabilities. For better visualization, set the x-axis to logarithmic.

  See example in Figure 1.

(a) Write your code in the notebook, showing your graph and the steps to get there. Display the first 10 sequences and show the value of your last proportion.

(b) Discuss your observations from the graph

(c) Increase N and explain how this changes your previous final results

(d) Change the probability of H from 0.5 to other values and observe how this changes the plot. Discuss this change.

3. (15 points) **Graphing the Gaussian probability density [Notebook+PDF]**
   The exact mathematical formula for the normal probability density is:

$$p(x) = \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right) \tag{1}$$

(a) Write your code in the notebook, showing your graph of the normal probability density function, with meanμ at 0.0 and standard deviation, σ at 0.2. Record the maximum value of the graph and explain why this is greater than 1, given that this is probability density function.

(b) Reduce your standard deviation and discuss the changes to the graph including the maximum value.

(c) Change the value of the mean from zero, leaving everything else the same. Discuss your observations of the graph.

4. (35 points) **Getting familiar with common distributions from the exponential family [PDF]**
   Write out the mathematical formula for each of the distributions below, stating which variable(s) represents the parameter(s) of the distribution. Indicate whether it models discrete or continuous variables. Also, in one or two sentences, describe when this distribution would come in handy during a modeling task. They are:

   (a) Gaussian; (b) Bernoulli; (c) Binomial; (d) Multinomial; (e) Exponential; (f) Poisson; (g) Dirichlet.