CS57300 Homework	— Homework 2 —	Last update: February 11, 2016
https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/index.html		Due: February 19, 2016

Instructions and Policy: Each student should write up their own solutions independently. You need to indicate the names of the people you discussed a problem with; ideally you should discuss with no more than two other people.

You need to submit your TYPED answer in PDF via Blackboard. LaTeX is typesetting is encouraged but not required. Unless directed you are not required to submit your R and Python codes. Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary.

Q1 (5 pts):
Over the years Democrats (DEM) have argued that **nationally** (disregarding the states) individual donations to their candidates are smaller (on average) than that of Republican (GOP) donations. Your neighbors are in a heated debate about the Democrats' claim. Jimmy supports the Democrats' view while Ronald thinks it is not true. Using data released by the electoral commission for the 2012 election cycle answer the following questions.
Data: GOP 2012 donations `https://goo.gl/dUi3Ef` and DEM 2012 donations `https://goo.gl/uVyI7I` .
The data format: ⟨candidate id⟩, ⟨donation amount⟩, ⟨state of candidate⟩ . **Disregard the candidate ID in questions Q1.i to Q1.iv.**

(i) Jimmy argues the Democrats (DEMs) are right because their (empirical) average donation was $1887 while the GOP (empirical) average donation was $2064. Jimmy obtained these convincing numbers from a news article. Ronald says that proves nothing. Do you agree with Jimmy or Ronald? Justify using statistical arguments (e.g. equations) with an example or counter-example.

(ii) Using the 2012 donation data above, explain in detail how you would verify the Democrats' claims. Give a formal statistical argument and present your conclusions. Also argue whether you can **prove** the DEMs to be right or wrong.
PS: You must remove the negative donations from your data.

(iii) Now consider the DEM claim **per state**. Separate donations by state and show for which states you can support the DEM claims. The data contains 50 U.S. states (note that Wyoming is missing for the Democrats and Vermont is missing for the Republicans). Group the 50 states in ones that you think support the DEM claim and the states that do not (**inside each group list them in alphabetical order**). Give a formal statistical argument for your results and present your conclusions clearly.

(iv) What is the drawback of your solution to Q1.iii? Can you quantify the problem with your approach?

(v) Suppose we reversed the roles and the GOP claimed that GOP donations are on average smaller than DEM donations. Separate donations by state and show for which states you can support this GOP claim.

(vi) Now consider the donations **per candidate**. We are trying to test the claim that DEM candidates in average raise less money than GOP candidates. That is, each data point now is the sum of all donations of a candidate. Redo the analysis in (*i*), (*ii*), and (*iii*) using the total donation to the candidates rather than each individual donation as data points. Give formal statistical arguments (e.g. equations) and present your conclusions.

Q2 (1 pts): Download new datasets at:
`https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/hw/hw2/other_donation_data.zip`
Test each of the following hypotheses separately. Give formal statistical arguments (e.g. equations) and present your conclusions clearly.

(i) Test if "population1**b**.csv" has the same average donation as "population2**b**.csv" assuming donations are independent and identically distributed Bernoulli random variables. What is the average donation in each population?

(ii) Test if "population3**b**.csv" has the same average donation as "population4**b**.csv" assuming donations are independent and identically distributed Bernoulli random variables. What is the average donation in each population?

(iii) Test if "population1**p**.csv" has the same average donation as "population2**p**.csv" assuming donations are independent and identically distributed from an unknown random variable with finite variance. What is the average donation in each population?

(iv) Test if "population3**p**.csv" has the same average donation as "population4**p**.csv" assuming donations are independent and identically distributed from an unknown random variable with finite variance. What is the average donation in each population?

Q3 (4 pts): Multi-armed Bandits (MAB). Using the same dataset used in Q2 answer the following:

(i) Implement UCB1 with four arms. Simulate the reward of the k-th pull of arm $i \in \{1, 2, 3, 4\}$ with the $\langle$donation$\rangle$ of $\langle$user_id$\rangle = k$ of "population$i$**b**.csv". Run UCB1 for 1000 arm pulls.

  (a) Compute the total donation obtained with the MAB (reward).

  (b) In a single plot show in the horizontal axis the time (total number of pulls) and in the vertical axis the cumulative number of times you pull arm $i \in \{1, 2, 3, 4\}$ (each arm should have its own curve).

(ii) Implement UCB1 with four arms. Simulate the reward of the k-th pull of arm $i \in \{1, 2, 3, 4\}$ with the $\langle$donation$\rangle$ of $\langle$user_id$\rangle = k$ of "population$i$**p**.csv". Run UCB1 for 1000 arm pulls.

  (a) Compute the total donation obtained with the MAB (reward).

  (b) In a single plot show in the horizontal axis the time (total number of pulls) and in the vertical axis the cumulative number of times you pull arm $i \in \{1, 2, 3, 4\}$ (each arm should have its own curve).

(iii) Redo Q3.i and Q3.ii using the ϵ-greedy MAB. Is ϵ-greedy better than UCB1 in each of these cases? Why? Give formal statistical arguments.

(iv) Can you explain the difference between Q3.i and Q3.ii using the results in Q2? Give formal statistical arguments (e.g. equations) and present your conclusions clearly.

(v) Redo Q3.i using Thompson Sampling with:

  (a) Beta(1,1) prior per arm.

  (b) Beta(100,100) prior per arm.

  (c) Contrast the results of Q3.v.a against those of Q3.v.b. Give a scenario where the prior in Q3.v.b is a reasonable choice.

  (d) If the prior is Beta(ϵ, ϵ), $\epsilon \approx 0$, what kind of MAB approach does Thompson Sampling degenerate into?