CSC591: Foundations of Data Science
HW2: Probability distributions, Expectation, Maximum Likelihood Estimation, Sampling Distribution, Central Limit Theorem, Confidence Intervals, Hypothesis Testing.

Released: 9/20/18
Due: 10/1/18 (23:55 PM) (**absolutely no late h/w are allowed**, because solution will be released on 10/2; and midterm will be on 10/3)

Student Name:
Student ID:

Notes

- Filename: Lastname_StudentID.pdf (only pdf).
- You can also submit scanned hand written solution (should be legible, TA's interpretation is final).
- This h/w is worth **4**% of total grade.
- **Separate R mini project will be released after 1st midterm (to facilitate more time for exam preparation)**.
- A subset of the questions will be graded by the TAs (these questions will be decided by the instructor).
- All submission must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
- No makeups or bonus; for regarding policies, refer to syllabus and 1st day lecture slides.
- You are encouraged to do research, study online materials; discuss with fellow students; BUT ANSWERS SHOULD BE YOUR OWN. Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.

Q#	Max Points	Your Score
1	10	
2	10	
3	10	
4	10	
5	10	
6	10	

7	10	

Q1. Generic (**10 points**)

(a) List formulae for sample mean, mode, variance, standard deviation.Is the sample variance an unbiased estimator of population variance? Why or why not?**(5 points)**
(b) Define Central Limit Theorem and state assumptions. **(5 points)**

Q2. (Expected Values) **(10 points)**

Remember the following (i) and (ii) as you may need them for answering some of the questions:

(i). If X and Y are two random variables with finite expected values,
then $E(X+Y) = E(X) + E(Y)$.

(ii) If X and Y are independent, the $E(XY) = E(X)E(Y)$.

Answer (a) – (h).

(a) Define Expected Value of discrete (numerical) random variable. **(1 point)**
(b) Suppose in an experiment a fair coin is tossed 4 times. Let X denotes the number of tails that appeared in the experiment. Then what is E(X). **(2 points)**
(c) Recall the discussion on Bernoulli distribution. Let S_n be the number of success in n Bernoulli trials with probability p for success on each trail. Then what is $E(S_n)$. **(3 points)**
(d) A coin is tossed twice. Let $X_i = 1$ if the i^{th} toss is heads and 0 otherwise. Then what is $E(X_1X_2)$? **(2 points)**
(e) Let X be a random variable with expected value $\mu = E(X)$, then show that the Variance, $V(X) = E(X^2) - \mu^2$. **(2 points)**

Q3. (Expected Values) (**10 points**)
Using the principles from Q2 (if needed), please answer the following questions

(a) Let X be an exponentially distributed r.v. with parameter λ. Then the density function of X is given by: $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Compute E(X) and V(X), where V stands for variance. **(3 + 3 = 6 points)**
(b) Let us say an insurance company pays $500 for lost luggage or a cancelled flight. Historical data shows that the company ends up paying 1 out of 100 policies it sells. What premium should the company charge in order to make profit? (**4 points**)

(Q4) Continuous Distributions (**10 points**)

(a) Let us assume that the life of pen drives before failure is normally distributed with mean = 10 years and a standard deviation of 2 years. Find the probability that the pen drive fails between 9 years and 11 years. (**4 points**)

(b). (**6 points**) Let us assume that CSC-591 FDS class final numerical grades (maximum 100) are values of a continuous r.v. X that follows a normal distribution with mean 75 and s.d. 15. Students are assigned letter grades as following: A ($X \geq 90$); $B(80 \leq X < 90); C(70 \leq X < 80); D(60 \leq X < 70),$ *and* $F(X < 60)$. Answer following:

(i) If a student is chosen at random then compute the probability that the student earns a given letter grade

(ii) Compute the expected proportion of students in each letter grade

(Q5) Maximum Likelihood Estimation (MLE) (**10 points**)

(a) Concisely describe MLE procedure for single parameter **(2 points)**

(b) The Pareto distribution is sometimes used to model heavy tailed distributions. Consider a Pareto distribution with density function given by:
$f(x;\theta) = (\theta - 1)\,x^{-\theta}$ *if* $\theta > 2$ *and* $1 \leq x < \infty$

If $X_1, X_2, X_3, \ldots, X_n$ are i.i.d with density function given by $f(x;\theta)$, calculate MLE for θ.
(8 points)

(Q6) CI (**10 points**)

(a) Define Confidence Interval for population mean. **(2.5 points)**

(b) Outline the procedure for finding C.I. **(2.5 points)**

(c) The following data for a sample of 40 users from a social media site shows number of friends for each user. Compute the 97% CIfor the point estimate of mean, and margin of error. **(5 points)**

28 32 45 28 65 45 29 31 23 34
35 31 23 54 34 25 23 15 65 38
64 65 46 56 36 45 67 65 54 66
45 56 57 45 38 48 25 26 34 36

(Q7) Hypothesis testing **(2x5 = 10 points)**

(a) A student researcher claims that the average cost of an engineering book is less than $80. He selects a random sample of 36 books from University engineering book stores, where cost of each book in $s is listed below:

50 95 120 85 45 90 70 60 70 50 40 80 70 90 75 60 90 90 75 85 80 60 110 65 80 85 85 45 60 95 110 70 75 55 80 55.

Assume $\sigma = 19.2$. Is there enough evidence to support the student researchers claim at $\propto = 0.10$?

(**b**) The mean age of graduate students at a University is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.