# CS 418: Introduction to Data Science
# Homework Assignment 01
# *Fall 2019*

## Instructions

**This assignment is due Wednesday, September 11, at 11:59PM (Central Time).**

**Answers for this assignment must be entered on *Blackboard*. You have 1 attempt to submit the assignment. This attempt is not timed. The assignment will close automatically at the due date and no late submissions will be accepted.**

**This assignment is individual. Offering or receiving any kind of unauthorized or unacknowledged assistance is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.**

## Part I: Descriptive Statistics (30 pt.)

### Problem 1 (20 pt.)

**Given the following table with the length of 84 computer science conferences:**

| Length (in days) of conference | Number of conferences |
|---|---|
| 2 | 4 |
| 3 | 36 |
| 4 | 18 |
| 5 | 19 |
| 6 | 4 |
| 7 | 1 |
| 8 | 1 |
| 9 | 1 |

**Use the information in the table to answer the following questions:**

a. **(4 pts.) Calculate the sample mean and standard deviation.**

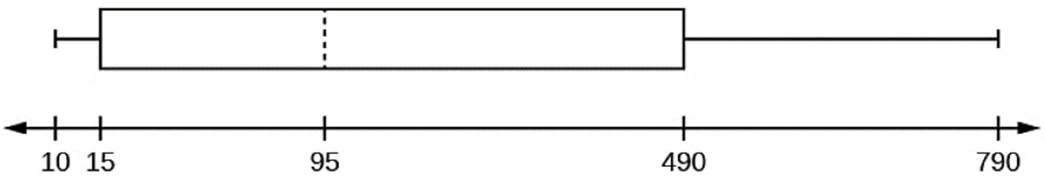
b. **(12 pts.) Find the median, mode, minimum, maximum, range, first quartile, third quartile, and 10th percentile. Try doing this without using any statistical software!**

c. **(2 pts.) Suppose that the organizers of the 8-day conference decide to make the conference last 9 days instead. Which of the previous measures would be affected by this change?**

d. **(2 pts.) Briefly justify your previous answer.**

### Problem 2 (10 pt.)

**Given the following box plot:**



**Use the information in the box plot to answer the following question:**

a. **(6 pts.) Find the median, minimum, maximum, first quartile, third quartile, and interquartile range. Note that the interquartile range is a measure of dispersion given by the difference between the third and the first quartile.**

b. **(2 pts.) Are there more observations in the interval [10, 95] or in the interval [95, 490]?**

c. **(2 pts.) Briefly justify your previous answer.**

## Part II: Probability (30 pt.)

### Problem 3 (15 pt.)

**A recent survey of licensed U.S. drivers found the following information:**

**"The percent of licensed U.S. drivers that are female is 48.5%. Of the females, 5.3% are age 19 or under, 81.2% are age 20-64, and 13.5% are age 65 or over. Of the licensed U.S. male drivers, 6.5% are age 19 or under, 82.3% are age 20-64, and 11.2% are age 65 or over."**

**Use the information in the survey to answer the following questions:**

a. **(4 pts.) What is the probability that a driver is female? What is the probability that a driver is age 65 or over given that the driver is female?**

b. **(3 pts.) What is the probability that a driver is age 65 or over and female?**

c. **(3 pts.) What is the probability that a driver is age 65 or over?**

d. **(2 pts.) Are being age 65 or over and being female independent events?**

e. **(3 pts.) Briefly justify your previous answer.**

### Problem 4 (15 pt.)

**Suppose that 62% of emails classified as spam contain the word "sale" and 83% of emails classified as not spam do not contain the word "sale." If 30% of emails are spam, answer the following questions:**

a. **(3 pts.) What is the probability that a spam email does not contain the word "sale"? What is the probability that a non-spam email contains the word "sale"?**

b. **(6 pts.) What is the probability that an email with the word "sale" is spam?**

c. **(6 pts.) What is the probability that an email without the word “sale” is spam?**

## Part III: Inferential Statistics (40 pt.)

### Problem 5 (20 pt.)

**A tire company claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 7,500. A survey of owners of this tire was conducted. From the 30 tires surveyed, the mean lifespan was 48,500 with a standard deviation of 9,250. Perform a hypothesis test at the $\alpha = 0.05$ significance level to determine whether the claim of the tire company is supported by the evidence and answer the following questions:**

a. **(4 pts.) What are the null and alternative hypotheses of the test?**

b. **(2 pts.) Which distribution do you use to perform the test (normal or Student’s $t$)?**

c. **(2 pts.) Is the test right-, left-, or two-tailed?**

d. **(8 pts.) Calculate the test statistic and the $p$-value.**

e. **(4 pts.) What conclusion do you make from the test?**

### Problem 6 (20 pt.)

**A recruiting office believes that the mean entry-level salary for data scientists is higher than the mean entry-level salary for software engineers. The recruiting office randomly surveys 30 entry level data scientists and 40 entry level software engineers. Their mean salaries were $82,500 and $79,500, respectively. Their standard deviations were $4,750 and $3,250, respectively. Perform a hypothesis test at the $\alpha = 0.05$ significance level to determine whether the belief of the recruiting office is supported by the evidence and answer the following questions:**

a. **(4 pts.) Suppose that $\mu_1$ is the mean entry-level salary for data scientists and $\mu_2$ is the mean entry-level salary for software engineers. What are the null and alternative hypotheses of the test?**

b. **(2 pts.) Which distribution do you use to perform the test (normal or Student’s $t$)?**

c. **(2 pts.) Is the test right-, left-, or two-tailed?**

d. **(8 pts.) Calculate the test statistic and the $p$-value.**

e. **(4 pts.) What conclusion do you make from the test?**