STAT 2002 - Probability and Statistics II, Summer 2024
Homework 1 - Descriptive Statistics and Point Estimation
100 points total.

This homework is due **Beijing Time 11:59pm Tuesday, June 18, 2024** on BlackBoard. No late homework accepted.

Please make sure to **SHOW ALL WORK** in order to receive full credit.

1. (20 points.) **Numerical summaries.** The "cold start ignition time" of an automobile engine is being investigated by a gasoline manufacturer. The following times (in seconds) were obtained for a test vehicle: 1.75, 1.84, 2.12, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91, 3.25, 2.83.

   (a) Calculate the sample mean, sample variance, and sample standard deviation.

   (b) Construct a box plot of the data by hand (not using R or other statistical softwares). You need o show the derivation for box plot.

2. (20 points.) **Smiling?** Dale Carnegie stated that smiling helps win friends and influence people. Research on the effects of smiling has backed this up and shown that a smiling person is judged to be more pleasant, attractive, sincere, sociable, and competent than a non-smiling person.

   There is evidence that smiling can attenuate judgments of possible wrongdoing. This phenomenon termed the "smile-leniency effect" was the focus of a study by Marianne LaFrance & Marvin Hecht in 1995. These researchers were interested in two questions: (a) Does smiling really increase leniency? (b) Are different types of smiles differentially effective?

   Data. Subjects in the experiment were asked to assume the role of a student member of a college disciplinary panel and judge a student accused of cheating. Each subject received a file that contained (a) a letter ostensibly from the chairperson of the Committee on Discipline, (b) a summary of the evidence, (c) background information on the suspect including prior academic performance and a color picture portraying one of the four facial expressions, and (d) rating scales to indicate the judgments. Subjects answered five questions about the likelihood of the suspect's guilt and how severe the punishment should be. These questions were combined into one "leniency score." This score is such that the higher is the score, the higher is the lenience.

   Four groups of subjects were tested. Each group saw one of the three types of smiles or a neutral-expression control. Subjects from a sample of 136 college students were randomly assigned to the four conditions with the constraint that there was an equal number of subjects (34) in each group.

   The data are attached together with the homework document. The name of the data file is 'smiles.txt'. Once you have saved the data file in the working directory, read the data in R using the commands

```
smiles=read.table("smiles.txt")
names(smiles)=c("groups","scores")
attach(smiles)
```

The first command reads the data in R, the second assigns names to each column and the third makes the components in the data readable by their specified names.

(**Part 1**) Construct histograms and stem-and-leaf plots for each of the four categories. Comment on the shape of the distribution of the observations in each of the four categories. Interpret and compare.

Instructions on how to use R:

i. To obtain the stem-and-leaf plots for all four groups jointly you can run the following R command:

```
tapply(scores,groups,stem)
```

which applies the function *stem* in R to each of the four groups.

ii. To obtain the histograms for all four groups jointly you can run the following R command:

```
splitgroup=split(scores,groups)
attach(splitgroup)
par(mfrow=c(2,2))
hist(false,main="")
hist(felt,main="")
hist(miserable,main="")
hist(neutral,main="")
```

which first splits the data by group, then divides the figure into four panels and within each panel, plots the histogram of one group. Note that you apply the *hist* function to each of the four groups.

(**Part 2**) Obtain the 5-numerical summaries and the corresponding boxplots for all four categories. Interpret and compare.

Instructions on how to use R:

i. To obtain the 5-numerical summary for each group you will use the R functions *summary*, which summarizes the data without providing the variance, and *var* which will give you the variance. The R code below is only for one group. Repeat for the other three groups.

```
summary(false)
var(false)
sd(false)
```

ii. The R function that you could use to construct a boxplot is *boxplot*. There are two ways to construct the boxplots for all categories jointly for comparison. One way is to use a similar code as you used for histograms.

```
par(mfrow=c(2,2))
boxplot(false)
title("false")
boxplot(felt)
title("felt")
boxplot(miserable)
title("miserable")
boxplot(neutral)
title("neutral")
```

A different approach is to simply use the boxplot function as follows:

```
par(mfrow=c(1,1))
boxplot(scores~groups)
```

Try to understand the input to the boxplot function. One way to read the help menu for a function is to use:

```
par(mfrow=c(1,1))
help(boxplot)
```

(**Part 3**) Based on the descriptive statistics and the different graphical displays, summarize your findings for the data analysis in the context of the problem.

3. (10 points) Of n_1 randomly selected students at CUHKSZ, X_1 owned a Mac, and of n_2 randomly selected students at SUSTech, X_2 owned a Mac. What is the sampling distribution of sample proportion difference $X_1/n_1 - X_2/n_2$? Suppose observed $n_1 = 200$, $X_1 = 150$, $n_2 = 250$, $X_2 = 185$. (Hint: using CLT approximation.)

4. (10 points) Let $\bar{X}_1$ and S_1^2 be the sample mean and variance for a sample of size n_1 from a population with mean μ_1 and variance σ_1^2. Similarly, let $\bar{X}_2$ and S_2^2 be the sample mean and variance for a sample of size n_2 from a population with mean μ_2 and variance σ_2^2.

   (a) Find an unbiased estimator for $\mu_1 - \mu_2$ and find its standard deviation.

   (b) Find the bias of the estimator $\bar{X}_1^2 - \bar{X}_2^2$ for the parameter $\mu_1^2 - \mu_2^2$. What happens to the bias as the sample sizes of n_1 and n_2 increase to ∞?

   (c) Assume that both populations have the same variance; that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Show that
   $$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$
   is an unbiased estimator of σ^2.

5. (20 points.) **Maximum Likelihood Estimation.** Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. exponential random variables with pdf given by
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

We are interested in estimating the parameter λ using a number of methods.

(a) What is the method of moment estimator $\widehat{\lambda}_{\text{mom}}$ of λ?

(b) What is the maximum likelihood estimator $\widehat{\lambda}_{\text{mle}}$ of λ ?

(c) Suppose the following $n = 6$ samples was generated from a exponential distribution with parameter λ as described in this question:

$$3.8 \quad 3.24 \quad 1.4 \quad 1.22 \quad 4.5 \quad 4.6$$

Compute $\widehat{\lambda}_{\text{mom}}$ and $\widehat{\lambda}_{\text{mle}}$.

(d) Suppose that $n > 1$. Is $\widehat{\lambda}_{\text{mle}}$ an unbiased estimator of λ? Why?

6. (10 points) Let $X_1, \ldots, X_n$ independent random variables identically distributed with density function

$$f(x) = \begin{cases} (\theta + 1)x^{\theta} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find a Method of Moments (MOM) Estimator of θ.

(b) Find the Maximum Likelihood Estimator (MLE) of θ.

7. (10 points) Let $X_1, \ldots, X_n$ be i.i.d. random sample from the uniform distribution on the interval (a, b). Find the method of moments estimators of a and b.