ISyE 3770 Assignment 4: Introduction to R, Descriptive Statistics, Sampling Distributions

Due date: 11:59 PM, Wednesday, March 13, 2024.

Question 1 (Descriptive Statistics Using R)**.** Students in a statistics course have a mid-term exam. The scores of 15 selected students are: `92, 57, 92, 61, 94, 98, 95, 52, 66, 63, 61, 71, 52, 56, 86`.

(a) Calculate the sample mean, sample variance, median, and sample IQR. (8 points)

(b) Construct a box plot of the data. (8 points)

(c) Construct a stem-and-leaf diagram of the data. (8 points)

(d) Construct a histogram plot of the data. (8 points)

(e) Assume we have a new data point added, which is `0`. Recalculate the sample mean, sample variance, median, and sample IQR. How would you comment on this data point? Which numerical summaries are robust to the outlier and why? (14 points)

(f) Now suppose the average homework scores of those 15 selected students are `96, 65.5, 93, 69.5, 98, 100, 88.5, 57, 60, 67.5, 63, 68, 58.5, 59, 86.5`. Draw a scatter plot to summarize the relation between mid-term and homework scores. Calculate the sample covariance and sample correlation between them. Write an interpretation of the information that you see in the plot and statistics. (14 points)

Question 2 (Sampling Distribution)**.** PVC pipe is manufactured with mean diameter of 1.05 inch and a standard deviation of 0.005 inch. Find the probability that a random sample of $n = 25$

sections of pipe will have a sample mean diameter greater than 1.049 inch and less than 1.051 inch. (10 points)

Question 3 (R practice)**.** Suppose that an experimenter observes a set of variables that are taken to be normally distributed with an unknown mean and variance. Using simulation methods, for given values of the mean and variance, we can simulate the data values that the experimenter might obtain. More interestingly, we can simulate lots of possible samples of which, in reality, the experimenter would observe only one. Performing this simulation allows us to check on sampling distributions of the parameter estimates.

Let us assume that $\mu = 100$ and $\sigma^2 = 9$, which, in fact, the experimenter does not know. In our simulation study, we assume that the experimenter will observe 100 observations, which are normally distributed. To simulate a sample of 100 observations from $N(100, 9)$, which the experimenter might observe, the R command is

```
x = rnorm(100,mean=100,sd=3)
```

The vector x will contain 100 values which are observations from a normal distribution $N(100, 9)$.

1) What is the mean and the variance of this sample? How do the sample mean and sample variance compare to true values of the mean and variance? (10 points)
2) Obtain random samples from the sampling distributions for the sample mean and the sample variance. (5 points)

   Instructions. In order to check the sampling distribution of the sample mean $\hat{\mu}$ and of the sample variance $\hat{\sigma}^2$, we will simulate 100 samples for several times (say 500 times). To simulate 500 times, we run the `rnorm` command within a for loop and create a matrix X with 500 rows and 100 columns, each row corresponding to one sample of 100 observations:

```
n = 100 #number of observations in one sample
S = 500 #number of simulations
X = matrix(0,nrow=S, ncol=n)
for(i in 1:S){
  X[i,] = rnorm(n,mean=100,sd=3)
}
```

To obtain the sample means and sample variances of the 500 samples, we apply the function `apply` as follows:

```
means = apply(X,1,mean)
variances = apply(X,1,var)
```

The vectors `means` and `variances` will contain the 500 sample means and 500 sample variances of the 500 samples.

3) Find the 5-numerical summary for the sample means and sample variances from the 500 samples (using the R functions `summary`). Plot the sample means and sample variances using a histogram. (5 points)

Instructions. The R command for a histogram is `hist`. To divide the figure into two panels each panel with one different plot use the command `par(mfrow=c(2,1))`.

```
par(mfrow=c(2,1))
hist(means)
hist(variances)
```

4) What is the (theoretical) sampling distribution of $\hat{\mu}$ if we know that the 500 samples come from a normal distribution $N(100, 9)$? Does the histogram approximate the sampling distribution for the sample mean? Why? (10 points)