UNIVERSITÉ BISHOP'S UNIVERSITY

Statistical Learning-CS507

Madjid Allili ∼ J125 ∼ ext. 2740 ∼ mallili@ubishops.ca

Assignment 3: Due October 30th, 2017

1. In R, the variable `rivers` gives the lengths (in miles) of 141 "major" rivers in North America as compiled by the US Geological Survey. The following table gives all of the percentiles for the river length variable from 0 to 100 by 5.

Percentile	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Value	135	230	255	276	291	310	330	350	375	392	425

Percentile	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
Value	460	505	545	610	680	735	890	1054	1450	3710

   So, for instance, the 65th percentile is equal to 545 miles. Using just this information answer the following questions:

   (a) how long is the longest river in North America?

   (b) what is the median river length?

   (c) what is the first quartile?

   (d) what is the third quartile?

   (e) approximately how many of these rivers (not the percent but the count) have lengths between the first and third quartile?

   (f) find the two values which encompass the center 60 percent of the river lengths.

   (g) approximately 90 percent of the river lengths are shorter than what length?

   (h) approximately what percent of these rivers are between 505 and 890 miles long?

2. Suppose the random variables X and Y are such that

   - $\mu_X = E(X) = 1$, $\mu_Y = E(Y) = 2$,
   - $\sigma_X = SD(X) = 3$, $\sigma_Y = SD(Y) = 4$,
   - $\rho_{XY} = Corr(X, Y) = 0.5$.

   Calculate

   (a) $E(2X - Y + 5)$

   (b) $SD(2X - Y + 5)$

3. Let X be a random variable such that $\mu_X = 2/3$ and $\sigma_X^2 = 1/18$. Now suppose that $X_1, X_2, \ldots, X_n$ are independent random variables, each having the same distribution as X and let $T = X_1 + X_2 + \ldots + X_{90}$. Find the approximate value of $P(T > 65)$.

4. Let X be the random variable associated to drawing a fair dice with outcomes $1, 2, 3, 4, 5, 6$. Suppose $X_1, \ldots, X_{30}$ be independent and identically distributed (i.i.d) as X.

   (a) Find the distribution of the random variable $\overline{X} = \frac{X_1 + \ldots + X_{30}}{30}$ using CLT.

   (b) Plot this distribution.

   (c) This time do not use the CLT, instead by running a simulation in R in the same way we have done in class, find the mean and variance of $\overline{X}$. You can use the "sample" function in R to roll your dice (i.e., choose an outcome from 1 to 6)!

   (d) Plot the distribution found in the first and second part (histogram) in the same graph.

5. Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. normal random variables with mean μ and variance σ^2. Let $\overline{X}$ be the sample mean and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ be the sample variance. Define the following random variables:

$$V_1 = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}, \quad V_2 = \frac{\overline{X} - \mu}{S/\sqrt{n}}, \quad \text{and} \quad V_3 = \frac{(n-1)S^2}{\sigma^2}$$

   For $n = 15$, use R to calculate the following probabilities:

   (a) $P(V_1 < -1.456)$

   (b) $P(V_2 > 2.155)$

   (c) $P(V_3 > 11.98)$

6. The data set `BodyTemperature.txt` is included in this assignment. In order to load the data into R, you may for instance you could use the command

```
temp<-read.table("BodyTemperature.txt")
```

   if the data file is in your working directory.

   Using the Temperature variable (body temp in Fahrenheit).

   (a) Find the best estimate for the population mean body temp.

   (b) Check that the sample mean for temperature has an approximately normal distribution (use a histogram and any other visual tool to illustrate your response)

   (c) Calculate the the sample mean and sample standard deviation for temperature using R commander.

   (d) Calculate the 95% confidence interval for the population mean temp.

7. The time from first exposure to HIV infection to AIDS diagnosis is called the incubation period. The incubation periods of a random sample of 14 HIV infected individuals is given below (in years):

$$12.0, 10.5, 5.2, 9.5, 6.3, 13.1, 13.5, 12.5, 10.7, 7.2, 14.9, 6.5, 8.1, 7.9$$

   (a) Find the sample mean and the sample standard deviation of this sample.

   (b) Plot a histogram of the data (6 bins, equally spaced from 4 to 16).

   (c) Calculate the 95% confidence interval for the population mean.

   (d) What assumptions are required for the confidence interval to be valid and how do you check each assumption?