Special Topics Comp Stat & Pro MAT5999 and Computational Stats & Prob. AIM 5002
Written Assignment 4 (25 points)
2/16/22

Solutions to be submitted on Canvas by the beginning of class on Wednesday, 2/23/22.
This is a long HW. The problems are worth 30 points in total but 25 points will be considered 100%. In other words, you can skip problems worth of 5 points and still get a perfect score.

1. (**7 points** This problem really belongs to week 3 but I did not want to overload HW3 with R problems). Checking normality of data with Q-Q plot

   In many of the above problems, we assume the normality of the data. In practice, the normality of the data is rarely assumed a priori, but instead we are supposed to check it. In this problem, we discuss a somewhat informal graphical method to check normality of a given dataset.

   Assume that Y has standard normal distribution with expectation μ and standard deviation σ. Let $q(p)$ be the p-quantile of Y, that is $\mathbb{P}(Y \leq q) = p$. By standardization we see (check it) that $(q(p) - \mu)/\sigma = z_{1-p}$, where z_{1-p} is the p-quantile of the standard Gaussian distribution. We found a linear relationship between the quantiles of Y and the quantiles of the standard Gaussian distribution.

   Now given a sample, we can test normality by plotting the sample quantiles against the theoretical quantiles of the standard Gaussian distribution (Q-Q plot). If we see a linear graph, we believe that the sample comes from a normal population but if the graph is not linear, we believe that the population is not normal.

   Try the following in R:

   ```
   > data1 <- rnorm(n = 500,mean = 10,sd = 3)
   > qqnorm(data1)
   > data2 <- rpois(n = 500, lambda = 3)
   > qqnorm(data2)
   ```

   You should see two Q-Q plots, the first one is approximately linear, the second one is not.

   Now (a) generate data of size n from exponential distribution with rate $= 1/4$. (b) Compute the average (c) Repeat this process 1000 times to obtain 1000 averages. (d) Create a Q-Q plot to check the validity of the CLT for $n = 1, 5, 30, 100$.

   What do you observe? What do you see? Please turn in your code, Q-Q plots and any conclusion in an R MArkdown file.

2. **(3+3+3 points)** Let $X_1, ..., X_5$ be random integers chosen without replacement from the list $\{1, ..., N\}$. Here N is an unknown parameter that we want to estimate. Let $Y_1, ..., Y_2$ be the order statistics obtained from $X_1, ..., X_5$, that is $Y_1 = \min\{X_1, ..., X_5\}$, Y_2 is the smallest of $\{X_1, ..., X_5\}$ excluding Y_1, etc. That is $Y_1 < Y_2 < ... < Y_5$ is obtained by ordering $X_1, ..., X_5$.

Now assume that

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_2 - Y_1) = \mathbb{E}(Y_3 - Y_2) = \mathbb{E}(Y_4 - Y_3) = \mathbb{E}(Y_5 - Y_4) = N - \mathbb{E}(Y_5) + 1. \quad (1)$$

This is true but a little tedious to prove (we proved the continuous variant of this on class).

(a) Write

$$N = Y_1 + (Y_2 - Y_1) + (Y_3 - Y_2) + (Y_4 - Y_3) + (Y_5 - Y_4) + (N - Y_5 + 1) - 1$$

take expectations and use (1) to derive that $N = 6\mathbb{E}(Y_1) - 1$.

(b) Use (1) to derive that $\mathbb{E}(Y_5) = 5\mathbb{E}(Y_1)$.

(c) Use (a) and (b) to derive that $\hat{N} = \frac{6Y_5}{5} - 1$ is an unbiased estimator of N

Remark: $\hat{N}$ is the MVUE (you do not have to prove this). By using this formula, statisticians reportedly estimated that the Germans produced 246 tanks per month between June 1940 and September 1942. At that time, standard intelligence estimates had believed the number was far, far higher, at around 1,400. After the war, the allies captured German production records, showing that the true number of tanks produced in those three years was 245 per month, almost exactly what the statisticians had calculated, and less than one fifth of what standard intelligence had thought likely.

Emboldened, the allies attacked the western front in 1944 and overcame the Panzers on their way to Berlin. See this link for more details.

3. Let $X_1, ..., X_n$ be a random sample from a normally distributed population with *known* expectation μ and *unknown* variance σ^2.

(a) **(3 points)** Show that $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is an unbiased estimator of σ^2.

(b) **(4 points)** Let $n = 3$. Find the mean squared error of $\hat{\sigma^2}$. *Hint*: You can use the fact that the fourth moment of the standard normal distribution is 3.

4. Let Y_1, Y_2, Y_3 be a random sample from an exponential distribution with parameter $1/\theta$ (that is, with expectation θ).

(a) **(3 points)** Show that $\hat{\theta}_1 = Y_1$, $\hat{\theta}_2 = (Y_1 + 2Y_2)/3$ and $\hat{\theta}_3 = \bar{Y}$ are all unbiased estimators of θ.

(b) **(4 points)** Find the relative efficiencies $eff(\hat{\theta}_1, \hat{\theta}_3)$ and $eff(\hat{\theta}_2, \hat{\theta}_3)$

The following problems form the extra homework. They will not contribute to your final grade and are only included for your entertainment.

5. Prove the following statement: if the MVUE exists, then it is unique. *Hint:* Assume that T_1 and T_2 are two MVUE's. We need to show that $T_1 = T_2$ with probability 1. To this end, prove that

$$V(T_1) \leq V\left(\frac{T_1 + T_2}{2}\right)$$

and use the Cauchy-Schwarz inequality to conclude that $V(T_1) = Cov(T_1, T_2)$. Finally, derive that $V(T_1 - T_2) = 0$.