# Homework 4: Due on Mar. 3

## Guideline

- Homework should be submitted via Gradescope by Friday midnight (11:59 pm. CDT).
- Homework answers to Simulations and data analysis should be written in R Markdown.
- Please find the following exercise question in [HMC] (Hogg, Mckean, and Craig 2018).
- The first question is worth of 10 points each and the last two is 20 points each.

1. Ex 6.1.8. Remark: Cauchy distribution is also a non-centered $t_1$ distribution. Here $\theta$ is median of the Cauchy distribution, instead of mean (does not exist).

2. Let $X_1, X_2, X_3, \cdots, X_{30}$ be the number of patients that walked into a hospital in the first, second, third,..., 30th day in a month. It is reasonable to assume they are i.i.d. Poisson random variables with parameter $\lambda$, where $\lambda$ describes the intensity rate of the patients flow in a day. The hospital wants to estimate $\lambda$ to decide how many nurses needed in the branch.

   Since the mean and the variance of Poisson random variables are both $\lambda$, one can use either the sample mean or the sample variance to estimate $\lambda$. But which one is better? We use simulations in R to investigate the distribution of the sample variance and compare it with the distribution of the sample mean.

   (a) Generate 1000 random samples of Poisson random numbers with sample size 30 and $\lambda$=20. Record the sample mean and sample variance in each random sample. Draw the histogram of the 1000 sample means, and the histogram of 1000 sample variances. Please compare the two histograms in terms of center and spread.

   (b) For sample mean, we have shown it is an unbiased estimator for $\lambda$, and $Var(\bar{X}) = Var(X_1)/n = \lambda/n$. Therefore, $MSE(\bar{X}) = [E(\bar{X} - \lambda)^2] = \lambda/n$. We check these properties using the simulations from part (a). Calculate the average and the sample variance of the 1000 sample means. What is the simulated MSE of the sample mean? Hint: the simulated (sample version of) MSE of the estimator ($\hat{\lambda}$ ) is: ( sample mean of $\hat{\lambda}^2 - \lambda)^2 +$ sample variance of $\hat{\lambda}$.

   (c) We further investigate the property of the sample variance (as an estimator for $\lambda$). Different from the normal distribution, the distribution of the sample variance of Poisson random variables is not easy to derive. Fortunately, part (a) provides an estimated distribution for the sample variance. What is the simulated MSE of the sample variance? Compare it with the MSE of the sample mean in part (b).

   (d) Based on your result in part (b) and (c), which of the two estimators is a better estimator for $\lambda$? Sample mean or sample variance?

   (e) Please redo a-d for $\lambda$=5,10,30,50,100. Does your conclusion change with different values of $\lambda$?

3. Continue from the previous problem. Besides point estimate, we also want to provide a 95% confidence interval for $\lambda$. In practice, sample size $n = 30$ is often good enough to apply CLT. (The sample mean approximately follow normal distribution.) In this case, $\bar{X} \sim N(\lambda, \lambda/n)$ approximately. (See your histogram in part (a) from the previous problem. )

   (a) Therefore, the hospital may consider a 95% confidence interval for $\lambda$ as $(\bar{x} - 1.96s/\sqrt{30}, \bar{x} + 1.96s/\sqrt{30})$. Please report this 95% confidence interval for $\lambda$ in each of the1000 random samples you simulated in part(a) of the previous problem. (Only print out the first 5 intervals.) How many of them have lower limit smaller than 20 AND upper limit larger than 20? (The number of CIs that include the true $\lambda$.)

   (b) Because sample mean is an estimator for $\lambda$ and the variance of original data is also $\lambda$, one may use sample mean to estimate the variance of data. Hence the hospital can also consider the 95% confidence interval $(\bar{X} - 1.96\sqrt{\bar{X}}/\sqrt{30}, \bar{X} + 1.96\sqrt{\bar{X}}/\sqrt{30})$. Please report this 95% confidence interval for $\lambda$ in each of the1000 random samples simulated in part(a) of the previous problem. (Only print out the first 5 intervals.) How many of the CIs include 20?

   (c) Please plot the CIs in a scatter plot: put the lower and upper limits of the CIs on x-axis and y- axis, respectively. Use black color for the CIs from (a) and blue for (b). Add a vertical line $x = 20$ and a horizontal line $y = 20$. Which part of the plot are for the CIs that contains the true =20 ?

   (d) Based on your findings from (a-c), which type of confidence intervals is better?

   (e) Does your conclusion change with different values of $\lambda$?

   Remark: we have seen that (a) the sample mean is the maximum likelihood estimator for $\lambda$, and (b) the sample variance is a model-free moment estimator for $\lambda$. Both are unbiased estimators. We claimed model-based estimator is better when the assumption is correct. This simulation setup provides an empirical example.