# MS4214 Statistical Inference Assignment

**(40% of Module)**

---

## Practical Information

- This assignment must be your own work. Evidence of plagiarism will result in a score of zero, and will be reported to Student Academic Administration.
- Deadline: **Friday, December 11th (Week 11)**.
- Report:
  - The report should be presented in `R` Markdown (install the 'rmarkdown' package in R). The .Rmd file can be knitted in HTML, PDF, or Word. The file must show the code and the output of every chunk (set echo=TRUE).
  - The .Rmd file and its knitted file (HTML, PDF, or Word) to be submitted using the "Assignments" section on Sulis.
  - Some questions need calculations by hand. The calculations must be shown in the report. Use the usual LaTeX language in `R` Markdown (you have to install LaTeX on your computers first, or see https://bookdown.org/yihui/rmarkdown/installation.html).

---

## Assignment

Use the `R` programming language to carry out the tasks below. Produce a document which contains the results, along **with explanations/discussion in your own words.**

1) **Estimation bias and efficiency**

i) Using the function `rpois`, create a function which generates a sample of Poisson data and calculates the mean of this sample (which is an estimator of $\lambda$). Call this function `meanpois`. The inputs of this function should be the value of $\lambda$ and the sample size, $n$.

ii) Generate one sample mean via `meanpois` with $\lambda = 1$ with $n = 10$. Each time you run this, you will get different results. To ensure that I (the examiner) can replicate the exact result which appears in your report, insert `set.seed(IDNUMBER)` in your code just prior to `meanpois` where `IDNUMBER` is your student ID number.

iii) The result of (ii) only applies to *one* sample so does not tell us much about the general properties of the estimator – for this we need replicate samples. Using a `for` loop or `replicate`, repeat part (ii) 1000 times. Now calculate the mean and variance of these 1000 sample means. Also calculate the mean squared error. Note: just prior to your loop, insert `set.seed(IDNUMBER)`.

iv) Repeat step (iii) but for $n = 20$, $n = 50$, $n = 200$, $n = 400$, $n = 1000$ (again insert `set.seed(IDNUMBER)` just prior to each loop). Produce the following graphs (and comment on each):

- bias versus sample size,
- efficiency versus sample size (note: you need to calculate the CRLB by hand first), and
- mean squared error versus sample size.

Question 1 provides the blueprint for carrying out a "simulation study", i.e., repeat an estimation procedure over a number of simulation replicates (typically 1000 replicates as used above), to investigate the properties of this procedure. Note that, for the Poisson example of Question 1, it is also straightforward to calculate bias, efficiency, and mean squared error by hand (i.e., a simulation study in `R` is not really needed) – but, for more complicated situations where nothing can be calculated by hand, a simulation study like in Question 1 is very useful.

Question 1 also highlights the importance of using `set.seed` so that the results you place in your report can be replicated exactly when your code is rerun, e.g., by an examiner, or even by yourself at another time.

The remaining questions are also simulation studies. It is assumed that you now know how to carry out such a study, including the use of `set.seed`, and that you will discuss the findings of each question in your report (i.e., do not simply display results without accompanying discussion).

2) **Central Limit Theorem**

i) Consider data arising from a normal distribution with $\mu = 1$ and $\sigma = 1$. By using histograms and Q-Q plots, show that the sample mean of this data for $n = 5$ is itself normally distributed (as in Question 1, you need to produce 1000 replicate sample means). Show the results for $n = 20$ and $n = 50$ also.

ii) Repeat (i) for the following scenarios:

- Exponentially distributed data with $\lambda = 1$,
- Bernoulli data with $p = 0.5$, and
- Bernoulli data with $p = 0.05$.

In each case again consider sample sizes of $n = 5$, $n = 20$, and $n = 50$. Comment on the distribution of the sample mean in each case (and, in particular, whether or not it looks normally distributed).

3) **Gamma distribution maximum likelihood estimation**

i) Generate *one* sample of size $n = 100$ of Gamma data with $\lambda = 1$ and $\alpha = 3$.

ii) By hand, find the maximum likelihood estimator for the gamma parameter $\lambda$ (assuming $\alpha$ is known).

iii) Now plug $\hat{\lambda}$ into the likelihood function so that it is only a function of $\alpha$. Using `R`, plot the likelihood function for a range of $\alpha$ values to establish the value of $\hat{\alpha}$ to one-decimal place. For this $\hat{\alpha}$ value, what is the associated $\hat{\lambda}$ value?

iv) Now consider the original likelihood which is a function of $\lambda$ and $\alpha$, i.e., the function before you plugged in $\hat{\lambda}$. Find the maximum likelihood estimates $\hat{\lambda}$ and $\hat{\alpha}$ using the numerical optimizer `nlm`. Note: this is done by minimizing *minus* the log-likelihood (which is equivalent to maximizing the log-likelihood).

v) When using `nlm`, set `hessian = TRUE` so that `nlm` stores the so-called "hessian" matrix, i.e., the $2 \times 2$ matrix of second derivatives. Note that since we minimize *minus* the log-likelihood, this is already the observed information matrix. By inverting this matrix in `R`, produce confidence intervals for $\lambda$ and $\alpha$.

Note that Question 3 does not involve replicating results multiple times. It is only required that you discuss the results for one sample, i.e., that which was generated in Q3(i).

4) **Evaluating confidence intervals**

i) Produce a function which generates data from a normal distribution and calculates a 95% confidence interval for $\mu$ via $\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$.

ii) Produce another function which uses $\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$.

iii) Which would you expect to perform better in smaller samples and why?

iv) For normal samples of size $n = 10$ with $\mu = 1$ and $\sigma = 1$, check the proportion of times that the confidence intervals from parts (i) and (ii) include the true $\mu$ value based on 1000 simulation replicates. Comment on the results.

v) Repeat (iv) but with $n = 500$. Comment on the results.

5) **Bootstrapping**

i) Bootstrapping is a useful technique for calculating confidence intervals. It involves resampling with replacement from the original data. Explain briefly in your own words how it works. (This technique is not contained in the lecture notes, so this component is your own independent research.)

ii) Generate a sample of exponential data with $n = 100$ and $\lambda = 1$. Bootstrap this data 1000 times, calculating the the median for the sample in each bootstrapped sample. The 0.025 and 0.975 quantiles of the distribution of bootstrapped medians provides a confidence interval for the median.

iii) In a similar manner to Question 4, repeat part (ii) here 1000 times to evaluate the performance of the confidence interval. Note: here you need to work out the value of the true median first, i.e., the value of $m$ such that $\Pr(X > m) = 0.5$.

iv) By hand, derive an approximate confidence interval (Wald-type) for $\lambda$ and, hence, for the median $m$. Evaluate the performance of this confidence interval.

Note: Typically bootstrapping is used when we do not have explicit formulae for producing a confidence interval. In this question we can derive a formula as in part (iii). However, bootstrapping can be used generally for producing confidence intervals easily for complicated estimators.

---