1. In this problem we will examine interval estimates for data from a Poisson distribution, where $Y_i \sim Poi(\theta)$ for $i = 1, 2, \dots 80$.

   Copy and paste the following code into R studio, updating the seed to your student ID:

```
#### Problem 1: Poisson interval estimates ####################################
set.seed(XXXXXXXX) #replace XXXXXXXX with your student ID
# generate a random value of theta from a Uniform(5,10) distribution
theta<-round(runif(1,5,10) ,digits=1)
n <-80
y<-rpois(n,theta)   # 80 observations from Poisson(theta) distribution

thetahat<-mean(y)            # maximum likelihood estimate of theta
s<-(thetahat/n)^0.5                # estimate for std error of estimator

#determine an interval of values for plotting relative likelihood function
th<-seq(max(0,thetahat-4*s), thetahat+4*s,0.001)
# You can use th as the x values for your plot once you create the RLF function in R
```

   You will run the code above to generate a value for theta and 80 observations from a Poisson(theta) distribution. Use this output and create additional code based on the R Tutorial for Assignment 3 to answer the questions below. **Include all R code used in this question in your R code file submitted to the LEARN dropbox.**

   For this question we will accept handwritten solutions instead of typed solutions, but you must still include a screenshot of the plots from R.

a) What are the values of θ and $\hat{\theta}$ generated for your ID number?
b) Recall that the estimator $\bar{Y}$ has an approximately gaussian distribution by the CLT. Showing your work/steps, use the CLT to find the approximate probability that the sample mean $\bar{Y}$ will lie within 0.1 units of θ.
c) Showing your work/steps, find the relative likelihood function $R(\theta)$ (in terms of θ and $\hat{\theta}$) for the Poisson(θ) distribution.
d) Create a function for the Poisson relative likelihood function in R and plot the function. Add a line at 12% and insert the plot into your solutions.
e) Use the plot in d) to determine an approximate 12% likelihood interval for θ. Provide the answer in your solutions, i.e use your plot to determine the lower and upper bound.
f) Use the uniroot command in R to determine the solved 12% likelihood interval for θ. Provide the answer in your solutions.
g) A 12% likelihood interval is also an approximate x% confidence interval. Showing your work/steps, find the value of x (the confidence coefficient for the approximate confidence interval corresponding to a 12% likelihood interval).
h) Construct an approximate 99% confidence interval for θ based on the asymptotic Gaussian pivotal quantity. Use the estimate of theta, $\hat{\theta}$, for your standard deviation. Show/explain your steps.

i) For an approximate confidence interval based on the asymptotic Gaussian pivotal quantity in the case of the Poisson(θ) model, determine under what conditions you would obtain an interval consisting of a single point. If the sample size was 5, what is the probability this occurs for your true value of θ?

a) $\hat{\theta} = 7$ and θ is 6.6 according to Fig.1 from R.

n	80
s	0.295803989154981
th	num [1:2367] 5.82 5.82 5.82 5.82 5.82 ...
theta	6.6
thetahat	7
y	int [1:80] 9 8 4 5 9 6 6 3 11 10 ...

Fig.1 Initial variables

b) For any random variables $Y_i \sim$ Poisson(θ, θ), E(Y_i) = θ and Var(Y_i) = θ. Then, for $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ will have $E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}E(\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}n\theta = \theta$

$\mathrm{Var}(\bar{Y}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}n\theta = \frac{\theta}{n}$, so $\sigma_{\bar{Y}} = \sqrt{\frac{\theta}{n}}$

Also, by CLT, $\bar{Y} \sim G(\theta, \sqrt{\frac{\theta}{n}})$ approximately, so we can calculate

$P(-0.1 \le \bar{Y} - \theta \le 0.1) = P(\frac{-0.1}{\sqrt{\frac{\theta}{n}}} \le \frac{\bar{Y}-\theta}{\sqrt{\frac{\theta}{n}}} \le \frac{0.1}{\sqrt{\frac{\theta}{n}}})$ since $\sqrt{\frac{\theta}{n}} > 0$

$= P(\frac{-0.1}{\sqrt{\frac{\theta}{n}}} \le Z \le \frac{0.1}{\sqrt{\frac{\theta}{n}}})$ where Z~G(0, 1)

$= P(\frac{-0.1}{\sqrt{\frac{6.6}{80}}} \le Z \le \frac{0.1}{\sqrt{\frac{6.6}{80}}})$ since n = 80 and θ = 6.6

$\approx P(-0.3481553 \leqslant Z \leqslant 0.3481553)$

$\approx 2P(Z \leqslant 0.3481553) - 1$

$= 2 \times 0.6361382 - 1$

$= 0.2722764$

Therefore, the probability is approximately 0.108 by CLT.

c) $L(\theta) = \prod_{i=1}^{n} \theta^{y_i} e^{-\theta} = \theta^{n\bar{y}} e^{-n\theta}$

$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{n\bar{y}} e^{-n\theta}}{\hat{\theta}^{n\bar{y}} e^{-n\hat{\theta}}} = e^{n(\hat{\theta}-\theta)}(\frac{\theta}{\hat{\theta}})^{n\bar{y}}$

$= e^{n(\hat{\theta}-\theta)}(\frac{\theta}{\hat{\theta}})^{n\hat{\theta}}$ since $\hat{\theta} = \bar{y}$

d) The plot is shown in Fig.2.

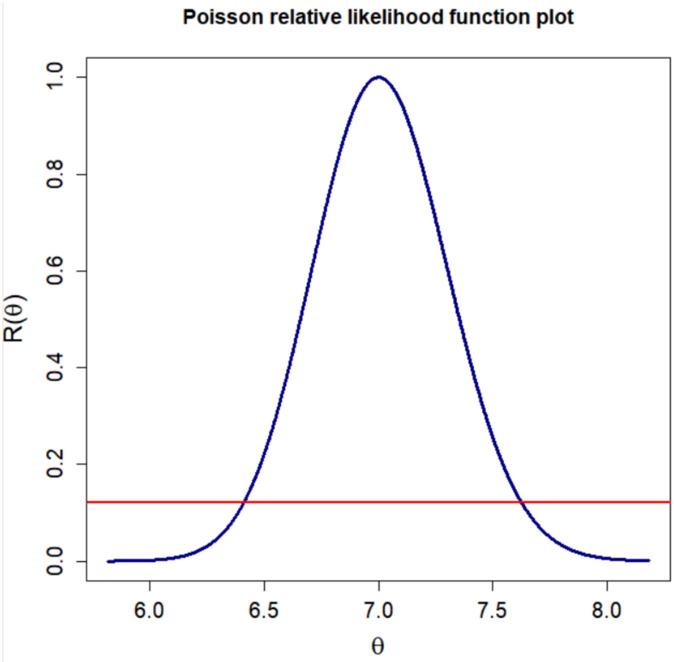

Fig.2 The plot of the Poisson relative likelihood function

e) According to Fig.2, the approximate 12% likelihood interval is about [6.4, 7.65], where 6.4 is the lower bound and 7.68 is the upper bound.

f) According to the R script result shown in Fig.3, the 12% likelihood interval for θ is actually [6.408404, 7.626927].

```
$root
[1] 6.408404

$f.root
[1] 9.436005e-08

$iter
[1] 8

$init.it
[1] NA

$estim.prec
[1] 6.103516e-05
```

```
$root
[1] 7.626927

$f.root
[1] 2.952264e-06

$iter
[1] 8

$init.it
[1] NA

$estim.prec
[1] 6.103516e-05
```

Fig.3 Result of the uniroot function

g) Let V = $2l(\tilde{\theta}) - 2l(\theta)$, then $\{\theta : R(\theta) \geqslant 0.12\} = \{\theta : 2l(\hat{\theta}) - 2l(\theta) \leqslant -2\log 0.12\}$,

where $l(\theta)$ is the log likelihood function of θ.

Since approximately, V~ $\chi^2(1)$ approximately, we have

$$P(-2\log R(\theta) \leqslant -2\log 0.12) \approx P(|Z| \leqslant \sqrt{-2log0.12})$$

$$= 2P(Z \leqslant \sqrt{-2log0.12}) - 1, \text{ where } Z\sim G(0, 1)$$

$$\approx 2\times 0.98 - 1 = 0.96 = x$$

Hence, x should be 0.96.

h) Let $Q_n(Y;\ \theta) = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}}{n}}}$. Note that since for n is large, $Q_n \sim G(0, 1)$, we solve

$0.99 \approx P(-2.575 \leqslant Q_n \leqslant 2.575)$

$= P(-2.575 \leqslant \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}}{n}}} \leqslant 2.575)$

$= P(-2.575\sqrt{\frac{\tilde{\theta}}{n}} - \tilde{\theta} \leqslant -\theta \leqslant 2.575\sqrt{\frac{\tilde{\theta}}{n}} - \tilde{\theta})$

$= P(\tilde{\theta} - 2.575\sqrt{\frac{\tilde{\theta}}{n}} \leqslant \theta \leqslant \tilde{\theta} + 2.575\sqrt{\frac{\tilde{\theta}}{n}})$

Therefore, the 99% confidence interval is approximately $[\hat{\theta} - 2.575\sqrt{\frac{\hat{\theta}}{n}},\ \hat{\theta} + 2.575\sqrt{\frac{\hat{\theta}}{n}}]$

$= [7 - 2.575\sqrt{\frac{7}{80}}, 7 + 2.575\sqrt{\frac{7}{80}}] \approx [6.238305, 7.761695]$

i) To let an approximate confidence interval just contain a single point, we simply let the width be 0.

Width = $2z(\sqrt{\frac{\hat{\theta}}{n}}) = 0$, where z is a non-zero constant corresponds to z values

$\hat{\theta} = 0 = \bar{Y}$ since z, 2, n $\neq 0$

Thus, every data is 0 since there is no negative numbers for Poisson distributions, we have

$P(\bar{Y} = 0) = (\frac{e^{-6.6} * 6.6^0}{0!})^5 = (e^{-6.6})^5 \approx 4.658886\text{e-}15$

Therefore, we need the average of the data is 0 (i.e. every data is 0 since there is no negative numbers for Poisson distributions) to let the interval only contain one point, and the probability is 4.658886e-15 for this to occur with my truth value of θ.

2. Let the random variable Y be the number of packages that reach the customer on time. Assume that Y has a Binomial(n, θ) distribution where n is the 500 observations in your dataset.
   a) What is the maximum likelihood estimate for θ based on your dataset?
   b) Use R to plot the relative likelihood function for θ based on your dataset and insert the plot in your assignment.
   c) Use the function uniroot to obtain a 15% likelihood interval (approximate 95% confidence interval) for θ.
   d) Obtain an approximate 95% confidence interval for θ based on the asymptotic Gaussian pivotal quantity. Show your work/steps in the written solutions.
   e) Provide a written interpretation of what the interval constructed in d) means.

a) Since Y follows binomial distribution, the maximum likelihood estimate is the sample mean of the data. Thus, it is 0.572 according to the result of the R script.

b) By definition, $R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} = \left(\frac{\theta}{\hat{\theta}}\right)^y\left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-y} = \left(\frac{\theta}{0.572}\right)^y\left(\frac{1-\theta}{1-0.572}\right)^{n-y}, 0 \leqslant \theta \leqslant 1$

Then, based on this, we have the plot here in Fig.4.

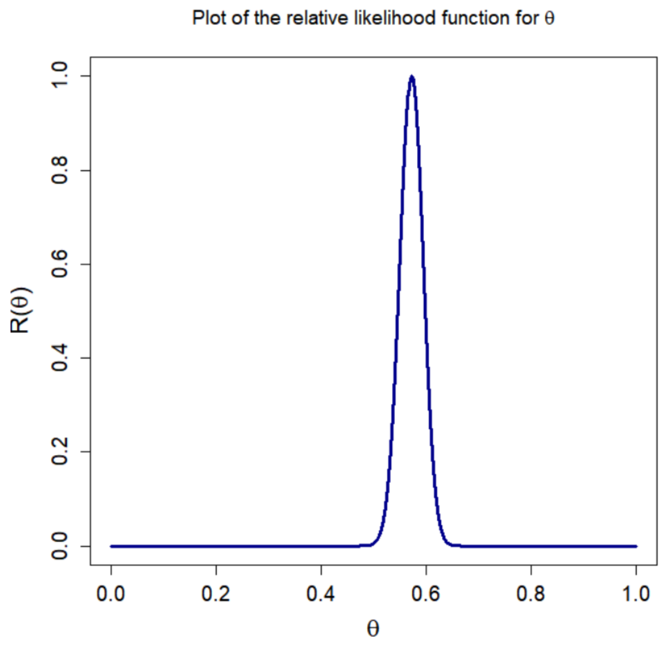

Fig.4 The plot of the relative likelihood function for θ

c) The 15% likelihood interval obtained using uniroot is [0.5285997, 0.6146774] according to Fig.5.

```
$root
[1] 0.5285997
```

```
$root
[1] 0.6146774
```

Fig.5 Result of the uniroot function

d) Let $Q_n(Y; \theta) = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$. Note that since for n is large, $Q_n \sim G(0, 1)$, we solve

$$0.95 \approx P(-1.96 \leqslant Q_n \leqslant 1.96)$$

$$= P(-1.96 \leqslant \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \leqslant 1.96)$$

$$= P(\tilde{\theta} - 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \leqslant \theta \leqslant \tilde{\theta} + 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}})$$

Therefore, the 99% confidence interval is approximately $[\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}] = [0.572 - 1.96\sqrt{\frac{0.572(1-0.572)}{500}}, 0.572 + 1.96\sqrt{\frac{0.572(1-0.572)}{500}}] \approx$ [0.5286298, 0.6153702].

e) If the experiment that gives the data is repeated lots of times, and we construct a 95% confidence interval each time, then about 95% of the constructed intervals would contain the true value of θ. That is to say, we are 95% confident that the true value of θ is within [0.5286298, 0.6153702].

3.

a) For $X \sim \chi^2(10)$ and $Y \sim G(10, \sqrt{20})$, find $P(X \leq 12.5)$ and $P(Y \leq 12.5)$. Are the probabilities close in value? Why or why not?

b) For $X \sim \chi^2(60)$ and $Y \sim G(60, \sqrt{120})$, find $P(X \geq 52.9)$ and $P(Y \geq 52.9)$. Are the probabilities close in value? Why or why not?

c) For $X \sim \chi^2(2)$, find $P(X < 4.3)$ with R commands for **two different probability distributions** (one is chi-squared but what is the other?) that both give the same exact answer.

d) For $X \sim \chi^2(15)$ find a and b such that $P(X \leq a) = P(X > b) = 0.05$.

e) Let $W = \sum_{i=1}^{10} Z_i^2$ where $Z_i \sim G(0,1)$. Find a and b such that $P(W \leq a) = P(W > b) = 0.025$.

a) R code and outputs:

```
> pchisq(12.5, 10)
[1] 0.7470147
> pnorm(12.5, 10, sqrt(20))
[1] 0.7119249
```

According to the output of the R script, P(X $\leqslant$ 12.5) = 0.7470147 and P(Y $\leqslant$ 12.5) = 0.7119249, which are not close. Because when the degrees of freedom k is not much larger than 2, and not close to 30, the probability density function of χ^2(k) will not be close to N(k, 2k) according to the lecture note. Thus, in this example, since 10 is not much larger than 2 and not close to 30 (i.e. 10 much smaller than 30), we say χ^2(10) will not have a similar probability density function as N(10, 20) = Gaussian(10, $\sqrt{20}$). Thus, when we calculate P(X $\leqslant$ 12.5) and P(Y $\leqslant$ 12.5), they will not be close.

b) R code and outputs:

```
> pchisq(52.9, 60, lower.tail = FALSE)
[1] 0.7304257
> pnorm(52.9, 60, sqrt(120), lower.tail = FALSE)
[1] 0.7415523
```

According to the output of the R script, P(X $\geqslant$ 52.9) = 0.7304257 and P(Y $\geqslant$ 52.9) = 0.7415523, which are very close. Because when the degrees of freedom k is large enough (i.e. larger than 30), the probability density function of χ^2(k) will be close to N(k, 2k) according to the lecture note. Thus, in this example, since 60 is much larger than 30, we say χ^2(60) will have a similar probability density function as N(60, 120) = Gaussian(60, $\sqrt{120}$). Thus, when we calculate P(X $\geqslant$ 52.9) and P(Y $\geqslant$ 52.9), they will be very close.

c) R code and outputs:

```
> pchisq(4.3, 2)
[1] 0.8835158
> pexp(4.3, 1 / 2)
[1] 0.8835158
```

According to the output, $P(X < 4.3) = 0.8835158$. And since when $k = 2$, $\chi^2(2)$ will have probability density function of exponential(2), we can also calculate $P(X < 4.3)$ using exponential distribution with parameter 2 (or 0.5 if the parameter is mean), which will give the exactly same answer.

d) R code and outputs:

```
> qchisq(0.05, 15)    # get the value of a
[1] 7.260944
> qchisq(0.05, 15, lower.tail = FALSE)    # get the value of b
[1] 24.99579
```

According to the output, a is 7.260944 and b is 24.99579.

e) Let $Q_i = Z_i^2$ for $i \in \{1, \ldots, 10\}$, we have $Q_i \sim \chi^2(1)$ since $Z_i \sim G(0, 1)$, according to Theorem 30 in the lecture note.

Then, by Theorem 29 in the lecture note,

$W = \sum_{i=1}^{10} {Z_i}^2 = \sum_{i=1}^{10} Q_i \sim \chi^2(10)$

Then, we run the R script to find values of a and b:

```
> qchisq(0.025, 10)    # get the value of a
[1] 3.246973
> qchisq(0.025, 10, lower.tail = FALSE)    # get the value of b
[1] 20.48318
```

Thus, a is 3.246973 and b is 20.48318 according to the script.

4. For parts a) – c) use a 15% likelihood interval:
   a) Change the **sample size** to 20. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value. How many of intervals constructed for the 100 samples contain the true parameter value of 1?
   b) Change the **sample size** to 100. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value. How many of intervals constructed for the 100 samples contain the true parameter value of 1?
   c) Change the **sample size** to 1000. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value. How many of intervals constructed for the 100 samples contain the true parameter value of 1?
   d) Using parts a)-c), explain **how and why** the width of an interval estimate for a given confidence coefficient changes when sample size changes.

a) The screenshot is in Fig.5 below.

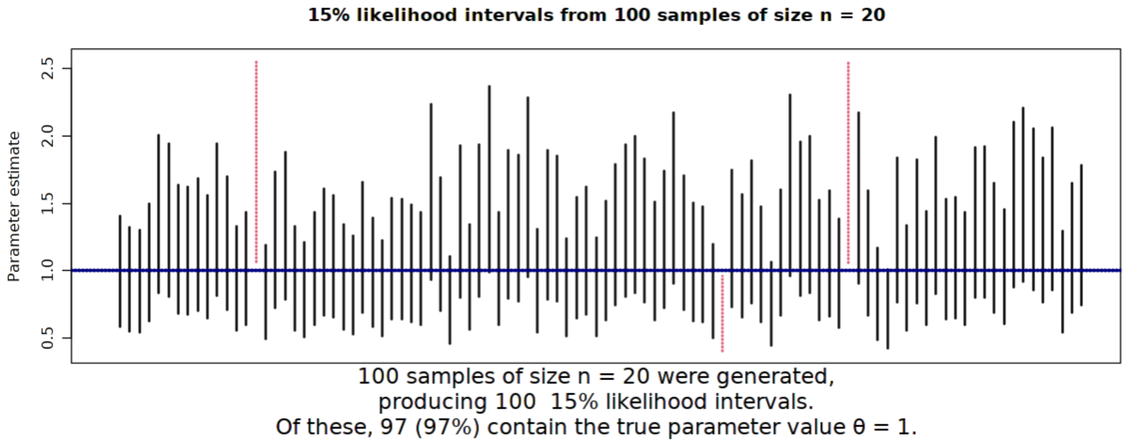

Fig.5 The screenshot for 15% likelihood intervals from 100 samples of size n = 20

Since there are 3 intervals that do not contain the true value, we have 100 – 3 = 97 intervals contain the true value.

b) The screenshot is in Fig.6 below.

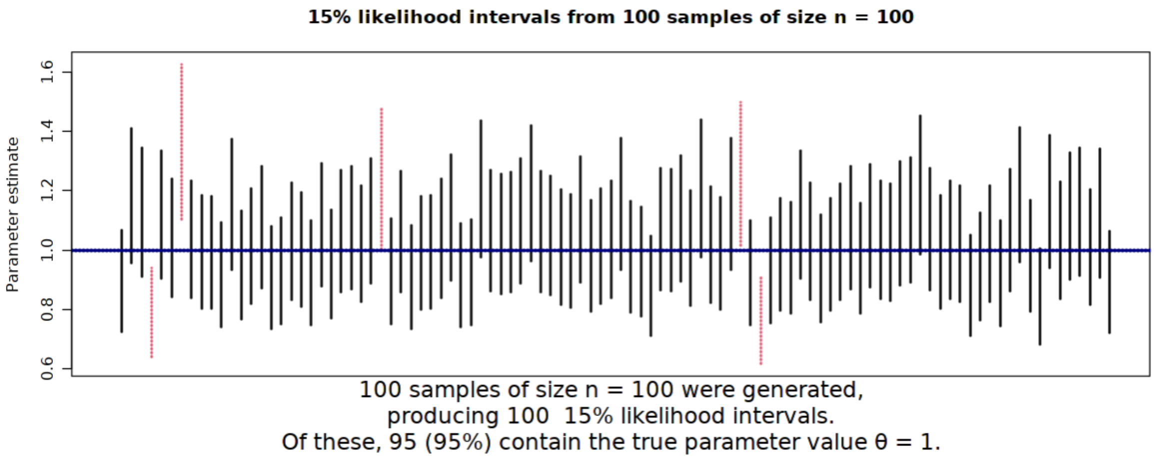

Fig.6 The screenshot for 15% likelihood intervals from 100 samples of size n = 100

Since there are 5 intervals that do not contain the true value, we have 100 – 5 = 95 intervals contain the true value.

c) The screenshot is in Fig.7 below.

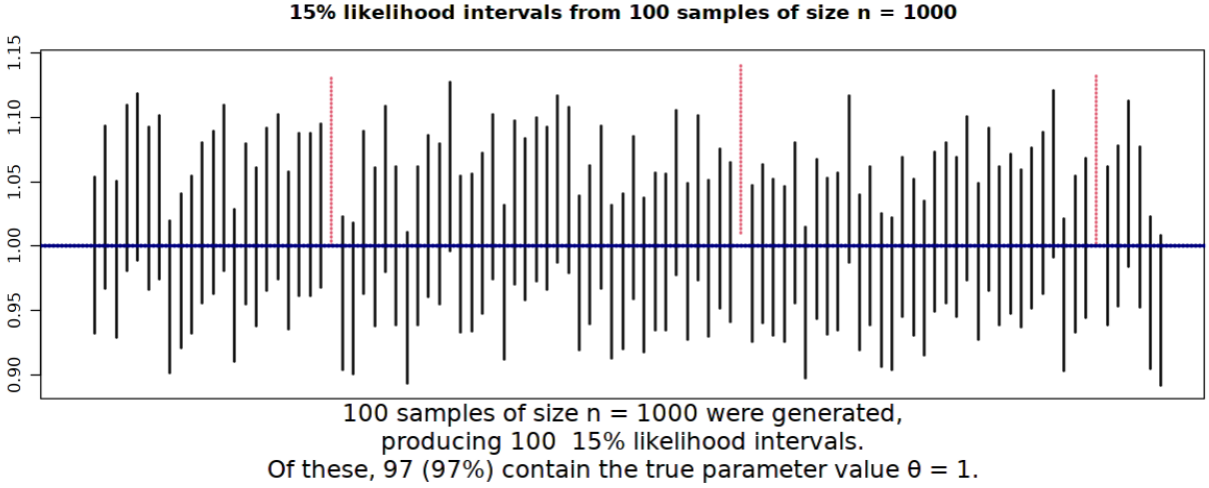

Fig.7 The screenshot for 15% likelihood intervals from 100 samples of size n = 1000

Since there are 3 intervals that do not contain the true value, we have 100 – 3 = 97 intervals contain the true value.

d) According to the graphs from a) to c), the width of interval becomes shorter as the sample size becomes larger.

Let $X_i \sim$ Exponential(1) for any i that all X_i's are independent, then S = $\frac{1}{n}\sum_{i=1}^{n} X_i$ has variance Var(S) = Var($\frac{1}{n}\sum_{i=1}^{n} X_i$) = $\frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$ = $\frac{1}{n^2}$n = $\frac{1}{n}$ since Var(X_i) = 1 for any i.

Hence, sd(S) = $\frac{1}{\sqrt{n}}$.

According to the lecture note, we know the widths of confidence intervals for Exponential distributions are calculated by Width = 2z*sd(S) = 2z$\frac{1}{\sqrt{n}}$, where z is a non-zero constant corresponds to the z-table values, which does not change when the confidence coefficient does not change. Thus, when the sample size n increases and the confidence coefficient does not change, the width becomes shorter.

For parts e) – f), use a sample size of 1000 and a likelihood interval:

e) Change the likelihood level to 5%. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value. How many of intervals constructed for the 100 samples contain the true parameter value of 1?
f) Change the likelihood level to 25%. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value. How many of intervals constructed for the 100 samples contain the true parameter value of 1?
g) Using parts e) and f), explain **how and why** the corresponding approximate confidence coefficient changes when the likelihood level of an interval estimate increases.

e) The screenshot is in Fig.8 below.

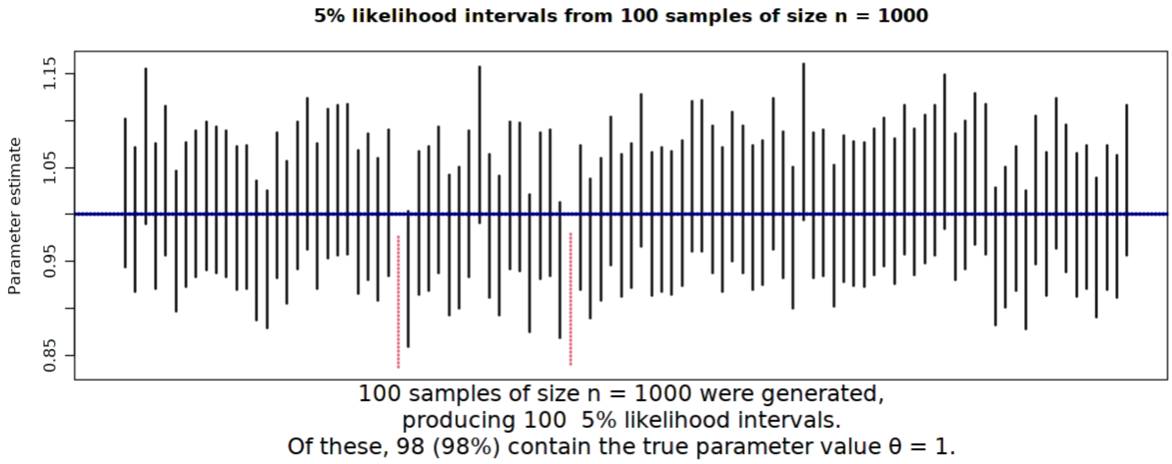

Fig.8 The screenshot for 5% likelihood intervals from 100 samples of size n = 1000

There are 98 intervals contain the true parameter according to Fig.8.

f) The screenshot is in Fig.9 below.

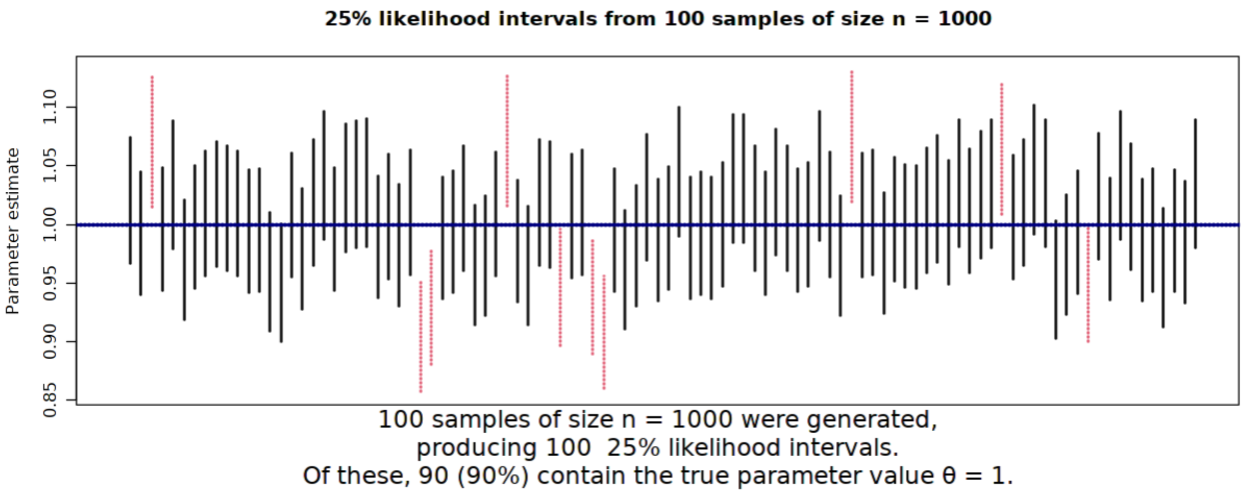

Fig.9 The screenshot for 25% likelihood intervals from 100 samples of size n = 1000

There are 90 intervals contain the true parameter according to Fig.9.

g) The corresponding approximate confidence interval coefficient becomes smaller when the likelihood level of an interval estimate increases.

According to Theorem 34 in the lecture note, A 100p% likelihood interval is an approximate 100q% confidence interval where q = 2P(Z $\leqslant$ $\sqrt{-2logp}$) – 1, where Z~N(0, 1).

Hence, when the likelihood level of an interval estimate increases, p increases, then $\sqrt{-2logp}$ decreases. Then, q = 2P(Z $\leqslant$ $\sqrt{-2logp}$) – 1 = 2pnorm($\sqrt{-2logp}$) – 1, where pnorm is the cumulative density function for N(0, 1). Note that since pnorm is non-decreasing, we know q = 2pnorm($\sqrt{-2logp}$) – 1 must be smaller, which means the corresponding approximate confidence interval coefficient becomes smaller when the likelihood level of an interval estimate increases.

For parts h) – j), use a sample size of 1000 and a confidence interval:

h) Select a 70% confidence interval. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value.
i) Select an 80% confidence interval now. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value.
j) Select a 95% confidence interval now. Insert a screenshot of the interval plot showing the width/location of the intervals and the number containing the true value.
k) Using parts h) -j), explain **how and why** changing the confidence coefficient changes the width of an interval estimate.

h) The screenshot is in Fig.10 below.

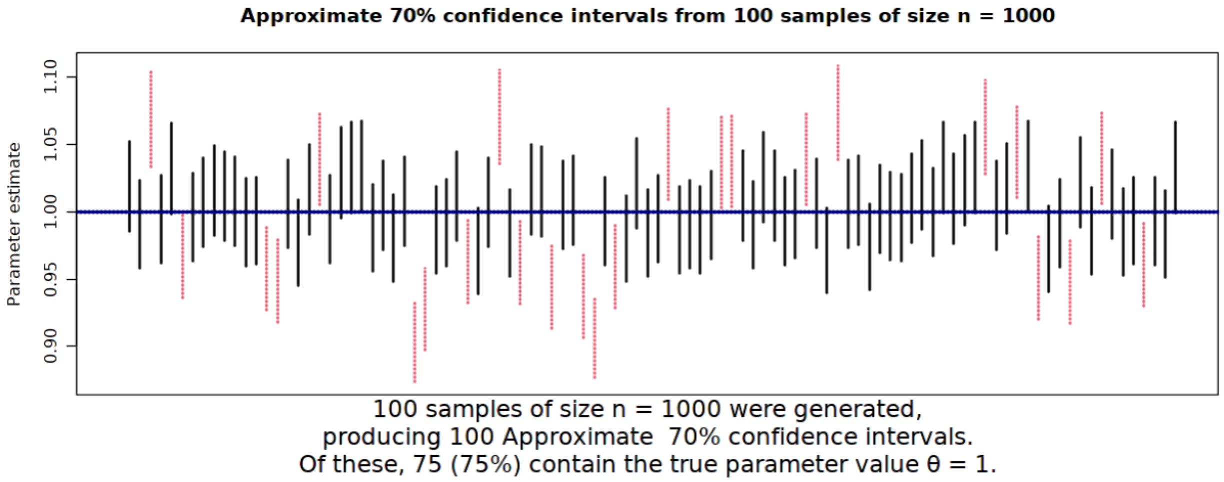

Fig.10 The screenshot for 70% confidence intervals from 100 samples of size n = 1000

i) The screenshot is in Fig.11 below.

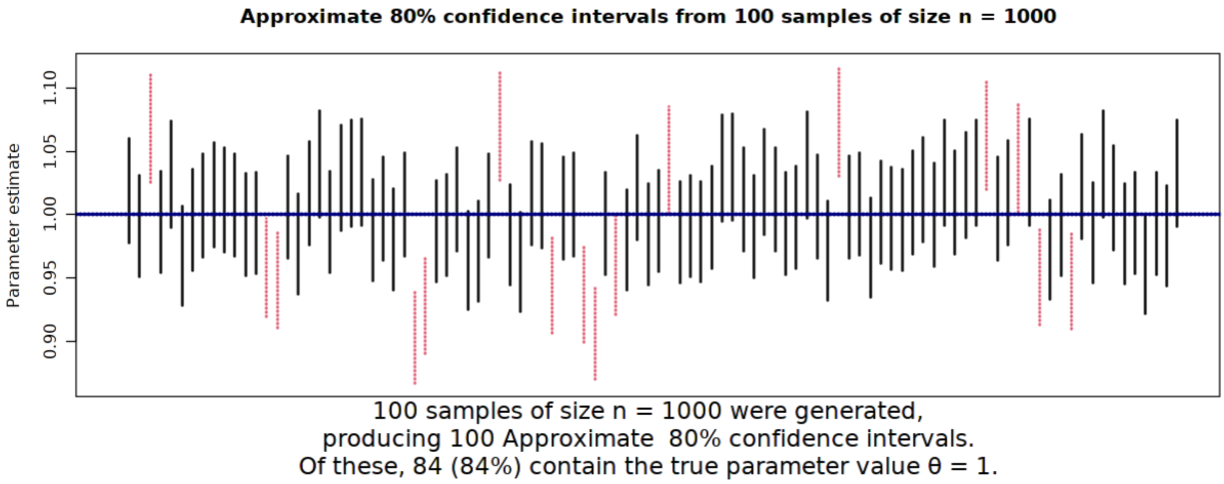

Fig.11 The screenshot for 80% confidence intervals from 100 samples of size n = 1000

j) The screenshot is in Fig.12 below.

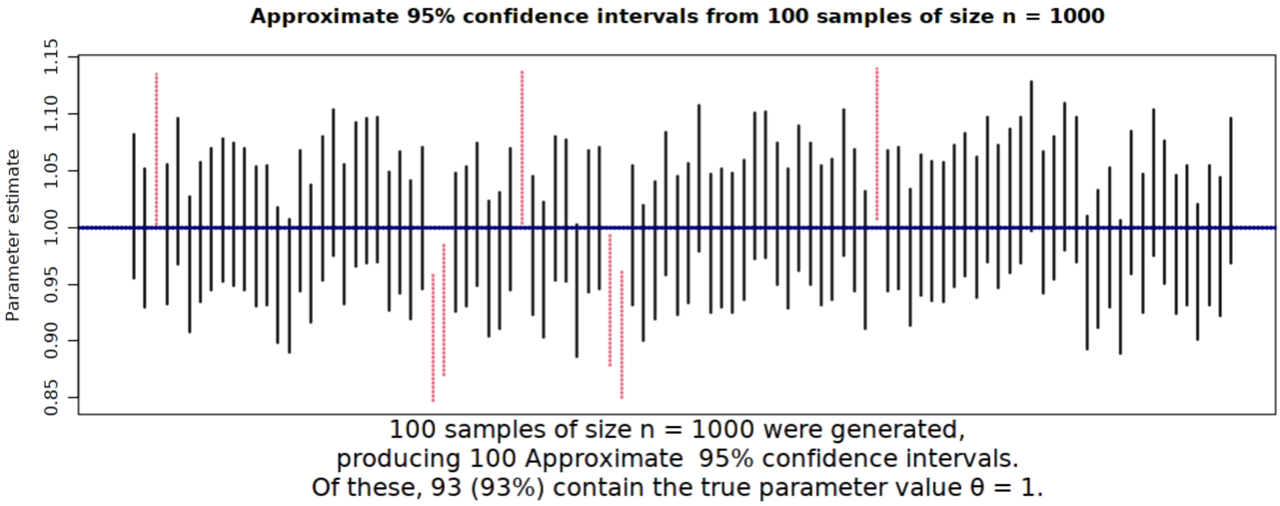

Fig.12 The screenshot for 95% confidence intervals from 100 samples of size n = 1000

k) The width of an interval becomes longer when the confidence coefficient becomes larger. Since we have shown in question 4.d), the widths of confidence intervals for Poisson distributions are calculated by Width = $2a\frac{1}{\sqrt{n}}$. And according to the lecture note, a = qnorm($\frac{1+p}{2}$), where qnorm is the probability density function of N(0, 1) and p is the confidence coefficient. Thus, since the probability density function of N(0, 1) is non-decreasing, a = qnorm($\frac{1+p}{2}$) increases as p increases, which means the width of an interval becomes longer when the confidence coefficient becomes larger.

Question 5 is intended to help you check that you have identified and internalized the important things to learn from this assignment.

5. Write an approximately 1-2 paragraph reflection (in full sentences) that summarizes how you achieved (or not if you're still not confident with them) the intended learning outcomes by completing this assignment. Your response will be graded using the Rubric posted in the Assignments folder on LEARN.

In question 1b, we were given the true value of the parameter before hand, then we calculated the probability for the sample mean, which is also the maximum likelihood estimate, to lie within 0.1 unit from the true population mean value. Through this question, we saw how variable the random variable $\tilde{\theta}$ can be, and the significance for constructing interval estimates in order to quantify the variability. Then, in the rest of questions in question 1, we tried to both obtain likelihood intervals using the graph of the relative likelihood function as well as the numerical method based on R script, and calculate confidence interval based on the asymptotic Gaussian pivotal quantity. Also, we tried to convert specific likelihood interval into the corresponding confidence interval by using the fact that $Z \sim N(0, 1)$ means $Z^2 \sim \chi^2(1)$. Finally, we tried to get the expression of the widths of the confidence intervals for Poisson distributions in term of the standard deviation of Poisson distributions, then we saw the extreme situation when the width of interval is equal to 0 (i.e. there is only one point in the interval). Based on this practice, we tried to check the factors that may affect the widths of the confidence intervals for Exponential(1) in question 4. In question 2, we implemented the interval techniques for quantifying the variability of the maximum likelihood estimator in a real-life example, and tried to interpret the meaning of the confidence interval we got. In question 3, we used R script to see the properties of $\chi^2(k)$ when k is in different ranges. We saw the fact that $Z \sim N(0, 1)$ means $Z^2 \sim \chi^2(1)$, $\chi^2(2) \sim$ Exponential(2), and $\chi^2(k)$ approximately follows $N(k, 2k)$ when k is larger than 30.