HW2 Stat-comp (due Wed, Oct 19th in D2L)

1) Maximum likelihood estimation and inference with the exponential distribution

The density function of an exponential random variable is

$f(x_i|\lambda) = \lambda e^{-\lambda x_i}$

where $x_i \geq 0$ is the random variable, and $\lambda > 0$ is a rate parameter.

The expected value and variance of the random variables are $E[X] = \frac{1}{\lambda}$ and $Var[X] = \frac{1}{\lambda^2}$.

The following code simulates 50 IID draws from an exponential distribution

```
set.seed(195021)
x=rexp(n=50,rate=2)
```

The maximum likelihood estimate of λ has a closed form, indeed

$L(\lambda|x) = \lambda^n e^{-\lambda n\bar{x}}$

Thus, $l(\lambda|x) = nlog(\lambda) - \lambda n\bar{x}$, therefore

$\frac{dl}{d\lambda} = \frac{n}{\lambda} - n\bar{x}$. Setting this derivative equal to zero, and solving for $\hat{\lambda}$ gives $\hat{\lambda} = \frac{1}{\bar{x}}$

1.1) Use `optimize()` to estimate λ compare your estimate with $\frac{1}{\bar{x}}$.

```
negLogLik=function(y,lambda){
    n=length(x)
    xBar=mean(x)
    logLik=n*log(lambda)-lambda*xBar*n
    return(-logLik)
}
fm=optimize(f=negLogLik,y=x,interval=c(0,100))

fm$minimum
```

```
## [1] 3.247199
```

```
1/mean(x)
```

```
## [1] 3.2472
```

```
fm$objective
```

```
## [1] -8.889658
```

```
negLogLik(y=x,lambda=fm$minimum)
```

```
## [1] -8.889658
```

1.2) Use numerical methods to provide an approximate 95% CI for your estimate.

Hint: `optimize()` does not provide a Hessian. However, you can use the `hessian()` function of the `numDeriv` R-package to obtain a numerical approximation to the second order derivative of the logLikelihood at the ML estiamte. To install this package you can use

```
#install.packages(pkg='numDeriv',repos='https://cran.r-project.org/')
library(numDeriv)
H=hessian(func=negLogLik,y=x,x=fm$minimum)
H
```

```
##          [,1]
## [1,] 4.741898
```

```
VAR=1/H # since H is scalar, we just use 1/H
SE=sqrt(VAR)

SE
```

```
##           [,1]
## [1,] 0.4592233
```

```
CI=fm$minimum+c(-1,1)*as.vector(1.96*SE)
round(CI,3)
```

```
## [1] 2.347 4.147
```

2) CIs for Predictions from Logistic Regression

Recall that in a logistic regresion model, the log-odds are parameterized as

$$log[\frac{\theta_i}{(1-\theta_i)}] = \mathbf{x}_i'\beta = \eta_i \tag{1}$$

The sampling variance of $\mathbf{x}_i'\beta = \eta_i$ is $Var(\eta_i) = \mathbf{x}_i'\mathbf{V}\mathbf{x}_i$, where $\mathbf{V}$ is the (co)variance matrix of the estimated effects; therefore, a SE and an approximate 95%CI for η_i can be obtained using

$SE(\eta_i) = \sqrt{\mathbf{x}_i'\mathbf{V}\mathbf{x}_i}$ and $CI : \mathbf{x}_i'\hat{\beta} + / - 1.96 \times SE(\eta_i)$.

Because the inverse-logit is a monotonic map, we can then obtain a 95% CI for the predicted probabilities by applying the inverse logit,$\theta_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, to the bounds of the CI for the linear predictor.

- Using the gout data set, fit a logistic regression for gout using sex, age, and race as predictors (for this you can use `glm()`, don't forget the link!).
- From the fitted model, and using the formulas presented above, compute the predicted probability of gout for each of the following cases, and the corresponding 95% CI for the predicted risk.

Race	Sex	Age	Predicted Risk	95%CI
White	Male	55		
White	Female	55		
Black	Male	55		
Black	Female	55		

```
DATA=read.table('https://raw.githubusercontent.com/gdlc/STAT_COMP/master/DATA/goutData.txt',header=TRUE)

DATA$y=ifelse(is.na(DATA$gout),NA,ifelse(DATA$gout=='Y',1,0))

table(DATA$y,DATA$gout)
```

```
##
##       N   Y
##   0 370   0
```

```
##   1   0  30
```

```
fm=glm(y~race+sex+age,data=DATA,family=binomial)
```

Once we fitted the model, I:

- create the incidence matrix (X) for the cases we want to predict,
- evaluate the linear predictor (eta=Xb), and its SE,
- use the above to get a CI for eta
- map it, using the inverse-logit, into a CI for the predicted probability

```
## Incidence matrix
X=cbind('int'=1,'raceW'=c(1,1,0,0),'sexM'=c(1,0,1,0),'age'=55)
X
```

```
##      int raceW sexM age
## [1,]   1     1    1  55
## [2,]   1     1    0  55
## [3,]   1     0    1  55
## [4,]   1     0    0  55
```

```
eta=X%*%coef(fm)
VCOV.ETA=X%*%vcov(fm)%*%t(X)
SE.ETA=sqrt(diag(VCOV.ETA))

invLogit=function(eta){
    exp(eta)/(1+exp(eta))
}

LOW=invLogit(eta-1.96*SE.ETA)
UP=invLogit(eta+1.96*SE.ETA)

ANS=data.frame('race'=c('W','W','B','B'),'sex'=c('M','F','M','F'),'age'=55,'prob'=invLogit(eta),'lower-

ANS
```

```
##   race sex age       prob lower.bound upper.bound
## 1    W   M  55 0.03568382 0.014337159  0.08603923
## 2    W   F  55 0.02307028 0.008963444  0.05807762
## 3    B   M  55 0.07219612 0.027323395  0.17732751
## 4    B   F  55 0.04730937 0.018421435  0.11613854
```

3) Bootstrap

Use 1,000 bootstrap samples to estimate the SE and 95% CIs for the probabilities reported in Question 2. Compare your bootstrap results with those reported in Question 2.

Here I used three approaches to compute CI with Bootstrap, for grading any of those will be considered correct.

```
nSamples=10000
ESTS=matrix(nrow=nSamples,ncol=nrow(X),NA)

for(i in 1:nSamples){
   tmp=sample(1:nrow(DATA),size=nrow(DATA),replace=TRUE)
   bootstrap_data=DATA[tmp,]
   fm=glm(y~race+sex+age,data=bootstrap_data,family=binomial)
   eta=X%*%coef(fm)
```

```
  prob=invLogit(eta)
  ESTS[i,]=prob
}


# Percentile method  (this method may be slightly bias, for bias correction see note below)
ANS2=ANS
ANS2[,'lower.bound']=apply(FUN=quantile,prob=0.025,MARGIN=2,X=ESTS)
ANS2[,'upper.bound']=apply(FUN=quantile,prob=0.975,MARGIN=2,X=ESTS)


## Now assuming normality and using Bootstrap to compute SE
## This one works well if the sampling distribution is approximately normal
SE=apply(FUN=sd,X=ESTS,MARGIN=2)
ANS3=ANS
ANS3[,'lower.bound']=ANS[,'prob']-1.96*SE
ANS3[,'upper.bound']=ANS[,'prob']+1.96*SE


ANS
```

```
##   race sex age       prob lower.bound upper.bound
## 1    W   M  55 0.03568382 0.014337159  0.08603923
## 2    W   F  55 0.02307028 0.008963444  0.05807762
## 3    B   M  55 0.07219612 0.027323395  0.17732751
## 4    B   F  55 0.04730937 0.018421435  0.11613854
```

```
ANS2
```

```
##   race sex age       prob lower.bound upper.bound
## 1    W   M  55 0.03568382 0.011970009  0.07587995
## 2    W   F  55 0.02307028 0.006998701  0.05048751
## 3    B   M  55 0.07219612 0.023251023  0.14609911
## 4    B   F  55 0.04730937 0.014276462  0.09718134
```

```
ANS3
```

```
##   race sex age       prob  lower.bound upper.bound
## 1    W   M  55 0.03568382 0.0034920382  0.06787560
## 2    W   F  55 0.02307028 0.0007848451  0.04535571
## 3    B   M  55 0.07219612 0.0092792600  0.13511298
## 4    B   F  55 0.04730937 0.0048350577  0.08978368
```

Let's look at CIs on top of the Bootstrap distribution

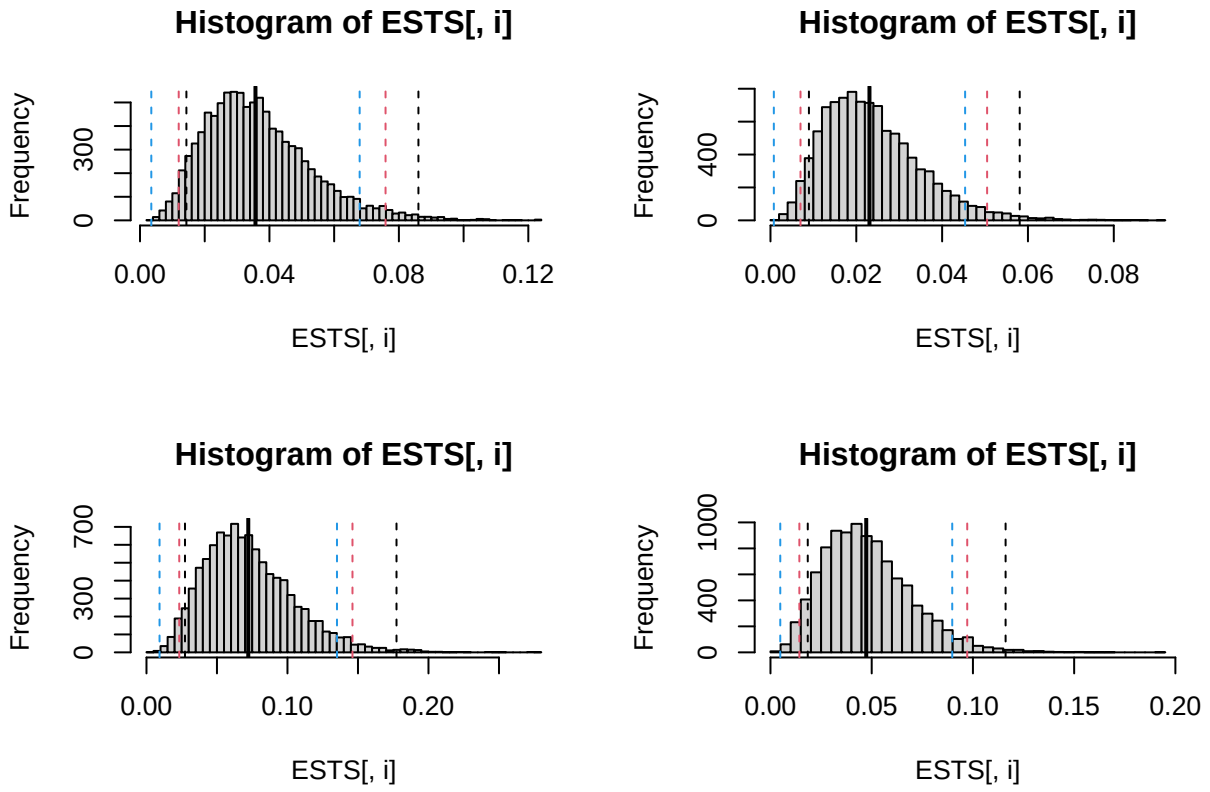
```
par(mfrow=c(2,2))
 for(i in 1:4){
     hist(ESTS[,i],50)
     abline(v=ANS[i,'prob'],lwd=2)

     abline(v=ANS[i,'lower.bound'],col=1,lty=2)
     abline(v=ANS[i,'upper.bound'],col=1,lty=2)

     abline(v=ANS2[i,'lower.bound'],col=2,lty=2)
     abline(v=ANS2[i,'upper.bound'],col=2,lty=2)

      abline(v=ANS3[i,'lower.bound'],col=4,lty=2)
```

```
    abline(v=ANS3[i,'upper.bound'],col=4,lty=2)
}
```



Note: You may also want to compare with `predict.glm()` this function can give you predictions and SEs for the predictions, see INCLASS-assingment.

Remarks

- The sampling distribution appears not to be normal,
- There are some differences in CIs (the first one appears shifted towards the right, the percentile method seems to capture a bit better the skewed nature of the distribution)
- For more details about different ways of estimating CIs using Bootstrap I refer you to Chapter 11 of the CASI, a book that you can download from here.