Week 4. Maximum likelihood

Fisher information

Read Section 6.2 "Cramér-Rao lower bound" in Hardle & Simar.

Let $Y_1, Y_2, ..., Y_n$ be iid sample from a general population Y distributed with pdf $f(y;\theta)$.

Definition 1 ***Fisher information****. Fisher information in a single observation is defined as*

$$I_1(\theta) = E\left(\frac{d\ln f(Y;\theta)}{d\theta}\right)^2.$$

Theorem 2 *The following holds:*

$$E\left(\frac{d\ln f(Y;\theta)}{d\theta}\right) = 0,$$

and therefore

$$I_1(\theta) = var\left(\frac{d\ln f(Y;\theta)}{d\theta}\right).$$

Theorem 3 *Fisher information can be derived from second derivative,*

$$I_1(\theta) = -E\left(\frac{d^2\ln f(Y;\theta)}{d\theta^2}\right).$$

Definition 4 *Fisher information in the entire sample is*

$$I(\theta) = nI_1(\theta).$$

Remark 5 *We use notation I_1 for the Fisher information from one observation and I from the entire sample (n observations).*

Theorem 6 ***Cramér-Rao lower bound.*** *Let $Y_1, Y_2, ..., Y_n$ be iid (random sample) from a general population with pdf $f(y;\theta)$ and $\widehat{\theta} = \widehat{\theta}(Y_1, Y_2, ..., Y_n)$ be an* ***unbiased*** *estimator. Then*

$$var(\widehat{\theta}) \geq \frac{1}{I(\theta)} = \frac{1}{nI_1(\theta)}.$$

Definition 7 *We say that unbiased estimator $\widehat{\theta}$ is efficient for θ if its variance reaches the Cramér-Rao lower bound.*

Example 8 *Prove that (a) the average $\widehat{\mu} = \overline{Y}$ of n iid $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known is efficient for μ using three definitions of information.*

Proof. The pdf of the general population is

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}.$$

Therefore for $Y \sim \mathcal{N}(\mu, \sigma^2)$ we have

$$\ln f(Y;\mu) = -\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(Y-\mu)^2$$

and

$$\frac{d\ln f(Y;\mu)}{d\mu} = \frac{1}{\sigma^2}(Y-\mu).$$

The information matrix is

$$I_1(\mu) = E\left(\frac{d\ln f(Y;\mu)}{d\mu}\right)^2 = E\left(\frac{1}{\sigma^2}(Y-\mu)\right)^2 = \frac{1}{\sigma^4}E(Y-\mu)^2 = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2}$$

and

$$I = \frac{n}{\sigma^2}$$

and therefore

$$var \geq \frac{1}{I} = \frac{\sigma^2}{n}.$$

The Cramér-Rao lower bound for any unbiased estimator is

$$var(\widehat{\mu}) \geq \frac{\sigma^2}{n}.$$

But

$$var(\overline{Y}) = \frac{\sigma^2}{n}.$$

That is, the variance of the mean equal to the Cramér-Rao lower bound and therefore $\overline{x}$ is efficient in the family of unbiased estimators.

Using formula

$$I_1(\theta) = var\left(\frac{d\ln f(Y;\theta)}{d\theta}\right),$$

we have

$$I_1(\mu) = var\left(\frac{1}{\sigma^2}(Y-\mu)\right) = \frac{1}{\sigma^4}var(Y) = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2}.$$

Using second derivative

$$I_1(\theta) = -E\left(\frac{d^2\ln f(Y;\theta)}{d\theta^2}\right),$$

we have

$$\frac{d^2\ln f(Y;\mu)}{d\mu^2} = \frac{d}{d\mu}\left(\frac{1}{\sigma^2}(Y-\mu)\right) = -\frac{1}{\sigma^2}$$

and

$$I_1(\mu) = -\left(-\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}.$$

Theorem 9 *The binomial proportion $\widehat{p} = m/n$ is an unbiased efficient estimator of probability p in n Bernoulli experiments.*

Proof. The pdf/pmf is written as

$$f(Y;p) = p^Y(1-p)^{1-Y},$$

where $Y = 0$ or $Y = 1$. We have

$$\frac{d\ln f(Y;p)}{dp} = \frac{Y}{p} - \frac{1-Y}{1-p}.$$

We have

$$\begin{aligned} I_1(p) &= var\left(\frac{Y}{p}-\frac{1-Y}{1-p}\right)=var\left(\frac{Y}{p}+\frac{Y}{1-p}\right)=var\left(Y\left(\frac{1}{p}+\frac{1}{1-p}\right)\right)=\frac{p(1-p)}{p^2(1-p)^2} \\ &= \frac{1}{p(1-p)} \end{aligned}$$

and

$$var(\widehat{p}) \geq \frac{p(1-p)}{n}.$$

But

$$var(\frac{m}{n}) = var(\overline{Y}) = \frac{p(1-p)}{n}$$

Multiple parameters

The true unknown parameter vector, $\boldsymbol{\theta}$ is the $m \times 1$ vector, $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)'$ and $\widehat{\boldsymbol{\theta}}= (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_m)'$ is the $m \times 1$ estimator vector.

Definition 10 ***The multidimensional Mean Square Error*** *is the $m \times m$ expected matrix,*

$$MSE(\boldsymbol{\theta}) = E[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'].$$

In particular, the sum of diagonal elements, $tr(MSE(\boldsymbol{\theta}))$ is called the total MSE.

The (i, j)th element of this matrix is $E[(\widehat{\theta}_i-\theta_i)(\widehat{\theta}_j-\theta_j)]$ and the (i, i)th diagonal element is the standard MSE of the ith component, $MSE_i = E[(\widehat{\theta}_i - \theta_i)^2]$. For an unbiased estimator, $E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ and therefore

$$MSE=cov(\widehat{\boldsymbol{\theta}}),$$

and the total MSE turns into the sum of variances, or the total variance.

Definition 11 *We say that an estimator $\widehat{\boldsymbol{\theta}}_1$ is no less efficient than an estimator $\widehat{\boldsymbol{\theta}}_2$ if for all true values of $\boldsymbol{\theta}$ we have*

$$MSE_1(\boldsymbol{\theta}) \leq MSE_2(\boldsymbol{\theta}),$$

i.e. the difference between the right- and the left-hand side is a nonnegative definite matrix (the eigenvalues of the difference are nonnegative), or in other words the MSE of any linear combination of $\widehat{\boldsymbol{\theta}}_1$ is smaller than the MSE of the linear combination of $\widehat{\boldsymbol{\theta}}_2$. In particular, an estimator $\widehat{\boldsymbol{\theta}}$ is an efficient estimator of $\boldsymbol{\theta}$ if the difference between its MSE and the MSE of another estimator is a nonnegative definite matrix.

Definition 12 ***Fisher information.*** *Let Y_i have common pdf $f(y; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the unknown m-dimensional parameter vector. The $m \times m$ Fisher information matrix in a single observation is defined as*

$$\mathbf{I}_1(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \ln f(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right].$$

Theorem 13 *The following holds:*

$$E\left(\frac{\partial \ln f(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = \mathbf{0},$$

and therefore

$$\mathbf{I}_1(\boldsymbol{\theta}) = cov\left(\frac{\partial \ln f(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right).$$

Theorem 14 *Fisher information can be derived from the second derivative*

$$\mathbf{I}_1(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln f(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right),$$

called the expected Hessian.

Definition 15 *Fisher information in a sample of size n is defined as*

$$\mathbf{I}(\boldsymbol{\theta}) = n\mathbf{I}_1(\boldsymbol{\theta}).$$

Theorem 16 ***Cramér-Rao lower bound for the covariance matrix.*** *Let $Y_1, Y_2, ..., Y_n$ be iid (random sample) from a general population with pdf $f(y;\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(Y_1, Y_2, ..., Y_n)$ be an unbiased estimator. Then*

$$cov(\widehat{\boldsymbol{\theta}}) \geq \frac{1}{n}\mathbf{I}_1^{-1}(\boldsymbol{\theta}).$$

Definition 17 *We say that unbiased $\widehat{\boldsymbol{\theta}}$ is efficient for $\boldsymbol{\theta}$ if its covariance matrix reaches the Cramér-Rao lower bound. We say that a component of $\widehat{\boldsymbol{\theta}}$ is efficient if its variance is equal the respective diagonal element of the Cramér-Rao lower bound.*

Theorem 18 *The average $\overline{Y}$ of n iid $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2* ***unknown*** *remains efficient for μ.*

Proof. We have

$$\boldsymbol{\theta}^{2\times 1} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}.$$

and

$$\frac{\partial \ln f(Y;\mu,\sigma^2)}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \ln f}{\partial \mu} \\ \frac{\partial \ln f}{\partial \sigma^2} \end{bmatrix}$$

where

$$\ln f(Y;\mu,\sigma^2) = -\ln(\sqrt{2\pi}) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(Y-\mu)^2.$$

We have

$$\begin{bmatrix} \frac{\partial \ln f}{\partial \mu} \\ \frac{\partial \ln f}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}(Y-\mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2 \end{bmatrix}$$

The 2×2 Fisher information matrix is

$$\begin{aligned}
\mathbf{I}_1(\mu,\sigma^2) &= E\left(\begin{bmatrix} \frac{1}{\sigma^2}(Y-\mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2 \end{bmatrix}\begin{bmatrix} \frac{1}{\sigma^2}(Y-\mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2 \end{bmatrix}'\right) \\
&= E\left(\begin{bmatrix} \frac{1}{\sigma^4}(Y-\mu)^2 & \frac{1}{\sigma^2}(Y-\mu)\left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2\right] \\ \frac{1}{\sigma^2}(Y-\mu)\left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2\right] & \left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2\right]^2 \end{bmatrix}\right) \\
&= \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & E\left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2\right]^2 \end{bmatrix}
\end{aligned}$$

Therefore,

$$cov \geq \frac{1}{n}\mathbf{I}_1^{-1}(\mu,\sigma^2) = \frac{1}{n}\begin{bmatrix} \sigma^2 & 0 \\ 0 & 1/E\left[-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-\mu)^2\right]^2 \end{bmatrix}$$

and again

$$var(\widehat{\mu}) \geq \frac{\sigma^2}{n}.$$

Maximum likelihood estimation

Reading: Section 6.1 of Hardle and Simar.

Let $f(\mathbf{y};\boldsymbol{\theta})$ be the joint density of random vector of observations $\mathbf{Y}^{n\times 1}$ with unknown parameter vector $\boldsymbol{\theta}^{m\times 1}$. The likelihood is defined as

$$L(\boldsymbol{\theta}) = f(\mathbf{Y};\boldsymbol{\theta}).$$

Note that now we switch our attention from distribution of $\mathbf{Y}$ to function of $\boldsymbol{\theta}$ where $\mathbf{Y}$ (data) is held fixed/known. In the case of **iid** we have

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta})$$

where f is the common density of individual observation y_i.

The **log-likelihood function** is defined as

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln f(\mathbf{Y};\boldsymbol{\theta})$$

and in the iid case

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(Y_i;\boldsymbol{\theta}).$$

Definition 19 *The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}_{ML} = \widehat{\boldsymbol{\theta}}_{ML}(\mathbf{Y})$, as a function of data, that maximizes the likelihood function, or equivalently, the log-likelihood, i.e.*

$$\widehat{\boldsymbol{\theta}}_{ML} = \arg\min l(\boldsymbol{\theta}).$$

In most cases this is equivalent to saying that $\widehat{\boldsymbol{\theta}}_{ML}$ is the solution of the **score equation** (the first-order condition for maximization)

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Requirements:

- Parameters are identifiable: $f(\mathbf{y};\boldsymbol{\theta}_1) = f(\mathbf{y};\boldsymbol{\theta}_2)\ \forall \mathbf{y}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.
- The support of f is independent of $\boldsymbol{\theta}$. For example, uniform distribution with unknown upper limit, $\mathcal{R}(0,\theta)$ does not comply.

Example 20 *The proportion of successes to the number of trials in n Bernoulli experiments is the MLE of the probability, p.*

Solution. Let $Y_1, Y_2, ..., Y_n$ be the Bernoulli trial outcomes, i.e. $\Pr(Y_i = 1) = p$ and $\Pr(Y_i = 0) = 1 - p$. The probability for individual trial i can be written as

$$f(Y_i;p) = p^{Y_i}(1-p)^{1-Y_i}.$$

Hence the log likelihood function is

$$l(p) = \sum_{i=1}^{n} (Y_i \ln p + (1 - Y_i)\ln(1-p)) = m \ln p + (n-m)\ln(1-p),$$

where $m = \sum_{i=1}^{n} Y_i$ is the number of successes. To find maximum differentiate l and set it zero,

$$\frac{dl}{dp} = \frac{m}{p} - \frac{n-m}{1-p} = 0$$

Solving for p yields

$$\widehat{p}_{ML} = \frac{m}{n}.$$

This is the maximum point because function

$$l(p) = m \ln p + (n-m) \ln(1-p)$$

is a concave function,

$$\frac{d^2 l}{dp^2} = \frac{d}{dp}\left(\frac{m}{p} - \frac{n-m}{1-p}\right) = -\left(\frac{m}{p^2} + \frac{n-m}{(1-p)^2}\right) < 0.$$

Properties of MLE and hypothesis testing

MLE has optimal asymptotic properties.

Theorem 21 *Asymptotic properties of the MLE with iid observations:*

1. **Consistency**:

   $$\widehat{\boldsymbol{\theta}}_{ML} \to \boldsymbol{\theta}, \quad n \to \infty$$

   with probability 1. This implies weak consistency: $p\lim \widehat{\boldsymbol{\theta}}_{ML} = \boldsymbol{\theta}$.

2. **Asymptotic normality**:

   $$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) \simeq \mathcal{N}(\mathbf{0}, \mathbf{I}_1^{-1}(\boldsymbol{\theta})), \quad n \to \infty$$

   where I is the Fisher information matrix

   $$\mathbf{I}_1 = cov\left(\frac{\partial \ln f(y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right).$$

3. **Asymptotic efficiency**: if $\widetilde{\boldsymbol{\theta}}$ is any other asymptotically normal estimator such that

   $$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \simeq \mathcal{N}(\mathbf{0}, \mathbf{C}) \quad n \to \infty$$

   then

   $$\mathbf{C} \geq \mathbf{I}_1^{-1}$$

   or in other words,

   $$cov(\widetilde{\boldsymbol{\theta}}) \geq cov(\widehat{\boldsymbol{\theta}}_{ML})$$

   for large n. In words, ML estimator reaches the Cramér-Rao lower bound in large sample (asymptotically).

Example 22 *Establish asymptotic statistical properties of the binomial proportion using the above theorem.*

Solution.

1. Since $Y_1, Y_2, ..., Y_n$ are Bernoulli iid trial outcomes

$$\widehat{p}_{ML} = \frac{m}{n} = \widehat{Y}$$

and therefore $\widehat{p}_{ML}$ is **consistent,**

$$p \lim_{n\to\infty} \frac{m}{n} = p.$$

2. The binomial proportion is **asymptotically normally distributed**

$$\sqrt{n}\left(\frac{m}{n} - p\right) \simeq \mathcal{N}(0, p(1-p))$$

because

$$I_1 = \frac{1}{p(1-p)}.$$

3. $\widehat{p}_{ML} = \frac{m}{n}$ is **asymptotically efficient**, i.e. for every other $\widetilde{p}_n$ such that

$$\sqrt{n}\,(\widetilde{p}_n - p) \simeq \mathcal{N}(0, V(p))$$

we have

$$V(p) \geq p(1-p).$$

Likelihood-based Wald confidence intervals

If $\widehat{\theta}_{ML}$ is the MLE then the asymptotic double-sided Wald CI for a **single** parameter θ is

$$\widehat{\theta}_{ML} \pm Z_{1-\alpha/2} \frac{1}{\sqrt{nI_1(\theta)}}.$$

We infer that in large sample, $n \to \infty$,

$$\Pr\left(\widehat{\theta}_{ML} - Z_{1-\alpha/2} \frac{1}{\sqrt{nI_1(\widehat{\theta}_{ML})}} < \theta < \widehat{\theta}_{ML} + Z_{1-\alpha/2} \frac{1}{\sqrt{nI_1(\widehat{\theta}_{ML})}}\right) \simeq 1 - \alpha$$

When $n \to \infty$, the CI shrinks to θ.

For **multiple** parameters, the CI for the ith component of $\boldsymbol{\theta}$ is

$$\widehat{\theta}_i \pm Z_{1-\alpha/2} \frac{1}{\sqrt{n}} \sqrt{I_1^{ii}(\widehat{\theta}_{ML})}$$

where

$$I_1^{ii} = (\mathbf{I}_1^{-1})_{ii}.$$

Example 23 *10 families report the number of children: 2,0,3,1,3,2,4,1,3,2. (a) Assuming that the number of children in the family follows a Poisson distribution with parameter λ, find the MLE (b) Find the 95% Wald CI for the average number of children in the family. (c) Use simulations to approximate the true coverage probability using specific values for λ and n. (d) Use simulations to demonstrate that the coverage probability improves with n. (e) You are visiting a new family and you want to present a postcard to each child. What is the number of postcards you want to bring so that each child will get a postcard with confidence probability* 0.75.

Solution. (a) Find the MLE of λ in the Poisson distribution:

$$\Pr(Y=k;\lambda)=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,...$$

Therefore if Y_i are the iid counts (observations) the likelihood is the product

$$L(\lambda)=\prod_{i=1}^{n}\Pr(Y_i;\lambda)$$

and the log-likelihood is

$$l(\lambda)=\sum_{i=1}^{n}\ln\Pr(Y_i;\lambda)=\sum_{i=1}^{n}(Y_i\ln\lambda-\ln Y_i!-\lambda)=const+\ln\lambda\times\sum_{i=1}^{n}Y_i-n\lambda.$$

Find the MLE by differentiation and equation to zero,

$$\frac{\sum_{i=1}^{n}Y_i}{\lambda}-n=0$$

and finally

$$\widehat{\lambda}_{ML}=\overline{Y}.$$

> x=c(2,0,3,1,3,2,4,1,3,2)
> mean(x)
[1] 2.1
Thus for our example

$$\widehat{\lambda}_{ML}=2.1$$

(b) Find Fisher information using the fact that $var(Y_i)=\lambda$,

$$I_1=var\left(\frac{d\ln\Pr(Y;\lambda)}{d\lambda}\right)=var\left(\frac{1}{\lambda}Y-1\right)=\frac{1}{\lambda^2}var\left(Y\right)=\frac{\lambda}{\lambda^2}=\frac{1}{\lambda}.$$

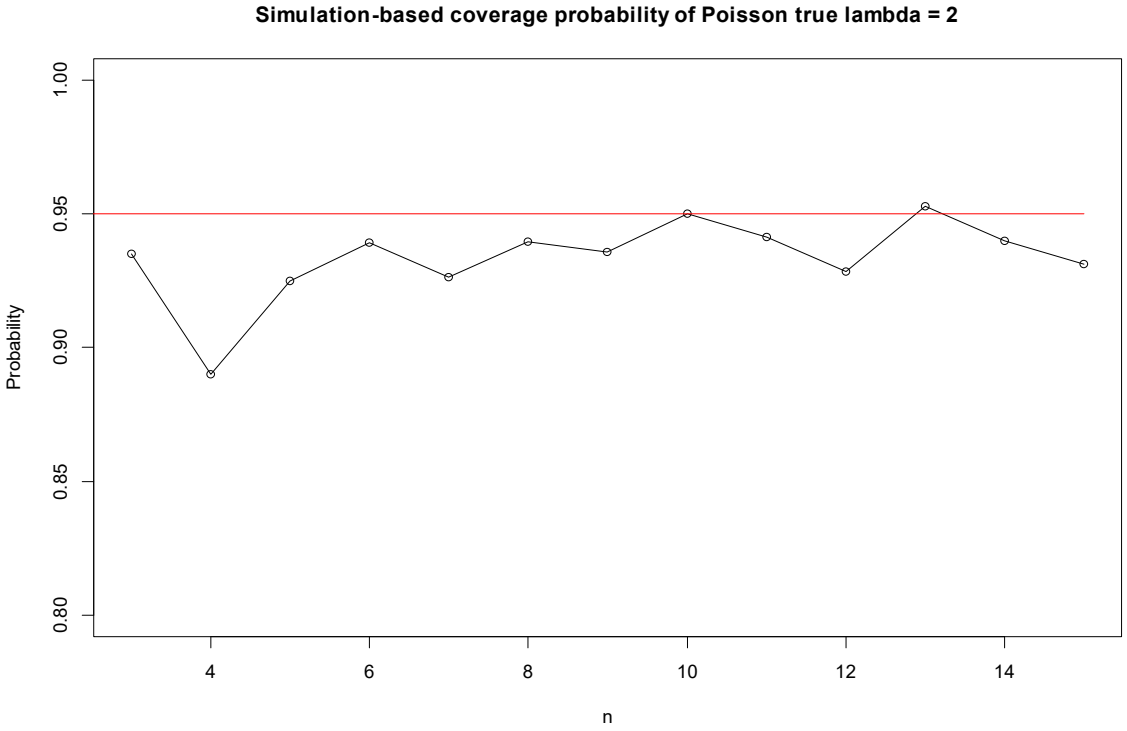
Find 95% Wald CI

$$\begin{aligned}\widehat{\lambda}_{ML}\pm Z_{1-\alpha/2}\frac{1}{\sqrt{nI_1(\widehat{\lambda}_{ML})}} &= 2.1\pm1.96\times\frac{1}{\sqrt{10/\widehat{\lambda}_{ML}}}=2.1\pm1.96\times\sqrt{\frac{\widehat{\lambda}_{ML}}{10}}=2.1\pm1.96\times\sqrt{\frac{2.1}{10}}\\ &= 2.1\pm0.9\end{aligned}$$

Answer: the 95% CI of the **average** number of children in the family is from 1.2 to 3.
(c)

```
ciPOIS=function(job=1,lambda.true=2,n=10,alpha=.05,nSim=10000)
{
dump("ciPOIS","c:\\M7019\\ciPOIS.r")
Z1a=qnorm(1-alpha/2)
if(job==1)
{
cover=0
for(isim in 1:nSim)
```

```
{
Y=rpois(n,lambda=lambda.true)
l.hat=mean(Y)
lowCI=l.hat-Z1a*sqrt(l.hat/n)
upCI=l.hat+Z1a*sqrt(l.hat/n)
if(lambda.true>lowCI & lambda.true<upCI) cover=cover+1
}
return(paste("Cover. prob =",cover/nSim))
}
if(job==2)
{
nseq=3:15
LN=length(nseq)
covn=rep(NA,LN)
for(i in 1:LN)
{
cover=0
for(isim in 1:nSim)
{
Y=rpois(nseq[i],lambda=lambda.true)
l.hat=mean(Y)
lowCI=l.hat-Z1a*sqrt(l.hat/nseq[i])
upCI=l.hat+Z1a*sqrt(l.hat/nseq[i])
if(lambda.true>lowCI & lambda.true<upCI) cover=cover+1
}
covn[i]=cover/nSim
}
plot(nseq,covn,ylim=c(.8,1),type="o",xlab="n",ylab="Probability",
main=paste("Simulation-based coverage probability of Poisson true lambda =",lambda.true))
segments(0,1-alpha,max(nseq),1-alpha,col=2)
}
}
> ciPOIS()
[1] "Cover. prob = 0.9501"
```
(d)



(e) Given $\lambda = \widehat{\lambda}_{ML} = 2.1$ find $k < K$ as the quantile `qpois(p=0.75,lambda=2.1)=3.`

Maximum likelihood-based hypothesis testing

Reading: Sections 6.3-6.5 of Hogg et al. and Section 7.1 of Hardle and Simar.

Let $\{Y_i, i = 1, 2, ..., n\}$ be an iid sample from a general population with the density $f(y; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the $m \times 1$ vector of parameters to be estimated using the sample. Suppose that the null hypothesis is formulated as

$$H_0 : \theta_1 = \theta_{10} \tag{1}$$

with the alternative $H_A : \theta_1 \neq \theta_{10}$, where θ_1 is the first component of vector $\boldsymbol{\theta}$ and θ_{10} is a specified number. The remaining $m - 1$ components $\boldsymbol{\theta}_2 = (\theta_2, \theta_3, ..., \theta_m)$ may be any; sometimes they are referred to as 'nuisance' parameters, so that $\boldsymbol{\theta} = (\theta_{10}, \boldsymbol{\theta}_2)$.

0.0.1 Wald test

From the maximum likelihood theory we know that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) \simeq \mathcal{N}\left(\mathbf{0}, \mathbf{I}_1^{-1}(\boldsymbol{\theta})\right)$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the $m \times m$ Fisher information matrix from the individual observation data, namely,

$$\begin{aligned}
\mathbf{I}_1(\boldsymbol{\theta}) &= E\left[\left(\frac{\partial \ln f(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right] && (2)\\
&= cov\left(\frac{\partial \ln f(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) && (3)\\
&= -E\left(\frac{\partial^2 \ln f(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right). && (4)
\end{aligned}$$

The approximate variance of $\widehat{\theta}_1$ is the $(1,1)$th element of matrix $(n\widehat{\mathbf{I}}_1)^{-1}$ where $\widehat{\mathbf{I}}_1 = \widehat{\mathbf{I}}_1(\widehat{\boldsymbol{\theta}}_{ML})$,

$$var(\widehat{\theta}_1) \simeq \left[(n\widehat{\mathbf{I}}_1)^{-1}\right]_{11}.$$

The key observation is that under the null hypothesis

$$Z = \frac{\widehat{\theta}_1 - \theta_{10}}{\sqrt{var(\widehat{\theta}_1)}} \simeq \mathcal{N}(0, 1), \quad n \to \infty.$$

Therefore, if $|Z| > \Phi^{-1}(1 - \alpha/2)$, where α is the significance level, then we reject the null hypothesis (α is the error of rejecting H_0 when it is actually true). The p-value $= 2(1 - \Phi(|Z|))$. Usually we reject H_0 if the p-value < 0.05.

Example 24 ***Boston shooting***. *35 days of shooting was reported in Boston last year. The national rate is 30 shooting per year in a city like Boston. Test the hypothesis that the difference is due to natural variation.*

Solution. The probability that a shooting occurs on a particular day in a US city is $p_0 = 30/365 = 0.0822$. Let p be the probability of shooting in Boston. The null hypothesis is $H_0 : p = p_0$. First, we derive the log-likelihood and the MLE for p (note that the notation p is used instead of the above θ).The number of shooting is the sum of iid Bernoulli experiments, $Y_i \in \{0, 1\}$ for days $i = 1, 2, ..., n = 365$ with $\Pr(y_i = 1) = p$ (sometimes p is referred to as binomial probability). The likelihood for individual i can be

written as $p^{Y_i}(1-p)^{1-Y_i}$ and the for all i is written as $L(p) = \prod_{i=1}^{n} p^{Y_1}(1-p)^{1-Y_i}$. The log-likelihood takes the form

$$l(p) = \sum_{i=1}^{n} [Y_i \ln p + (1 - Y_i) \ln(1-p)].$$

The MLE for p turns the score equation to zero, $dl/dp = 0$. Solving

$$\frac{dl}{dp} = \sum_{i=1}^{n} [Y_i p^{-1} - (1 - Y_i)(1-p)^{-1}] = 0$$

for p yields

$$\widehat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{m}{n},$$

where m is the number of shooting days over the year. This means that the estimate for Boston is $\widehat{p}_{ML} = 35/365 = 0.0959$. Question: is the difference between 0.0959 and 0.0822 statistically significant? In order to apply the Wald test we need to compute the Fisher information matrix (in our case a scalar). We use three formulas and show that they all give the same result.

1. Expected squared. Since in the previous notation $f(y;p) = p^y(1-p)^{1-y}$ we have $\ln f(y;p) = y \ln p + (1-y)\ln(1-p)$ and

$$\frac{d \ln f}{dp} = \frac{y}{p} - \frac{1-y}{1-p}$$

we have

$$E\left(\frac{d \ln f}{dp}\right)^2 = E\frac{Y^2}{p^2} - 2E\frac{Y(1-y)}{p(1-p)} + E\frac{(1-y)^2}{(1-p)^2}.$$

But

$$\begin{aligned} Ey^2 &= p, \\ E[Y(1-Y)] &= E(Y) - E(Y^2) = 0 \\ E(1-Y)^2 &= 1-p \end{aligned}$$

that gives

$$\begin{aligned} E\left(\frac{d \ln f}{dp}\right)^2 &= \frac{p}{p^2} + \frac{(1-p)}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} \\ &= \frac{1}{p(1-p)} \end{aligned}$$

2. Variance. We have

$$\begin{aligned} var\left(\frac{d \ln f}{dp}\right) &= var\left(\frac{Y}{p(1-p)} - \frac{1}{1-p}\right) = var\left(\frac{Y}{p(1-p)}\right) \\ &= \frac{var(Y)}{p^2(1-p)^2} = \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)}. \end{aligned}$$

3. Expected second derivative. We have

$$\frac{d^2 \ln f}{dp^2} = -\frac{Y}{p^2} - \frac{1-Y}{(1-p)^2}$$

so that

$$E\left(\frac{d^2 \ln f}{dp^2}\right) = -\frac{p}{p^2} - \frac{1-p}{(1-p)^2} = -\frac{1}{p(1-p)}$$

All three equations give the same result, the Fisher information matrix of the binomial probability is

$$I_1(p) = \frac{1}{p(1-p)}.$$

The Z-score test is

$$Z = \sqrt{n}\frac{\widehat{p}_{ML} - p_0}{\sqrt{p_0(1-p_0)}}$$

For our data

$$Z = \sqrt{365}\frac{0.0959 - 0.0822}{\sqrt{0.0822 \times (1-0.0822)}} = 0.953$$

with the p-value `2*(1-pnorm(0.953))=0.34.` Thus we cannot reject the hypothesis that the Boston rate of shooting is at the national level. The difference is due to natural variation.

Sometimes the information matrix in the Z-test is evaluated at the ML estimate. Then

$$Z_m = \sqrt{n}\frac{\widehat{p}_{ML} - p_0}{\sqrt{\widehat{p}_{ML0}(1-\widehat{p}_{ML0})}} = \sqrt{365}\frac{0.0959 - 0.0822}{\sqrt{0.0959 \times (1-0.0959)}} = 0.889.$$

with the p-value = 0.37 and leads to the same conclusion.

Homework 4

1. Observations $\{(X_{1i}, X_{2i}), i = 1, ..., n\}$ are independently drawn from a population with the bivariate density is $f(x_1, x_2; \lambda_1, \lambda_2) = \lambda_1\lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2}$. (a) Derive the Fisher information matrix in the entire sample. (b) Find the MLE for λ_1 and λ_2. (c) Find the asymptotic standard error for $\widehat{\lambda}_{1ML}$ and $\widehat{\lambda}_{2ML}$. (d) Test the hypothesis that $\lambda_1 = \lambda_2$. (e) Four pairs of similar houses in area A and B have been sold after being on the real estate market for (108,87), (23,35), (210,120), (14,23) days. Assuming that the time follows the above distribution, test the hypothesis that the sale rates in the two areas are the same (compute the p-value).

2. Is the average efficient for λ in the Poisson distribution? Show the math..

3. The annual number of sunny days in Hanover is 198. There were 12 sunny days in March. Test the null hypothesis that March is a typical month. Show the math.