Statistical Computing with R – MATH 6382[1,*] Set 6 (Monte Carlo Methods in Statistical Inference)

Tamer Oraby

UTRGV

tamer.oraby@utrgv.edu

[1] Based on textbook.

[*] Last updated November 14, 2016

MC in Estimation

MC in Estimation

- The purpose is to estimate parameters θ by $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, using a random sample $X_1, \ldots, X_n$ from the population or a distribution modeling the population X.
- A sampling distribution of $\hat{\theta}$ could be simulated by repeating the previous step a relatively large number of times.
- A standard error ($se(\hat{\theta})$) of $\hat{\theta}$ is the standard deviation of the repeated sampling in the previous step.

MC in Estimation

Example: Estimate $\theta = \mathbf{E}(|X_1 - X_2|)$ with X_1 and X_2 are i.i.d. $N(0,1)$ r.v.'s. and find the standard error and sampling distribution. (We know that $\theta = \mathbf{E}(|X_1 - X_2|) = \frac{2}{\sqrt{\pi}} = 1.128379$.)

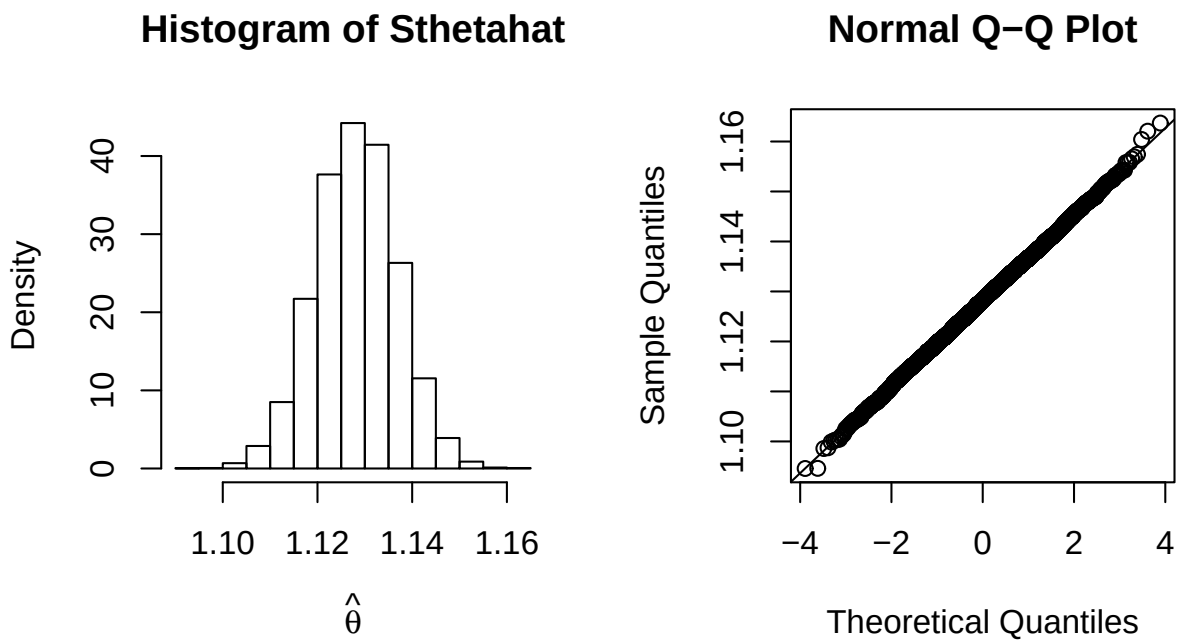
1. Generate a sample of paired (X_1, X_2): $(x_1^{(1)}, x_2^{(1)}), \ldots, (x_1^{(n)}, x_2^{(n)})$
2. Find $\left|x_1^{(1)} - x_2^{(1)}\right|, \ldots, \left|x_1^{(n)} - x_2^{(n)}\right|$
3. Find $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left|x_1^{(i)} - x_2^{(i)}\right|$
4. Repeat steps 1-3 for m number of times and use the outcomes as a representative sample from the sampling distribution of $\hat{\theta}$ to be used to find the standard error and the histogram and empirical cdf of the sampling distribution

MC in Estimation

```
n<-10000;m<-10000
X1<-rnorm(n);X2<-rnorm(n)
thetahat<-mean(abs(X1-X2))
thetahat
[1] 1.137416
sd(abs(X1-X2))/sqrt(n)
[1] 0.008629679
Sthetahat<-replicate(m,{X1<-rnorm(n);X2<-rnorm(n)
mean(abs(X1-X2))})
mean(Sthetahat)
[1] 1.128175
sd(Sthetahat)
[1] 0.008567367
```

MC in Estimation

```
par(mfrow=c(1,2))
hist(Sthetahat,prob=T,xlab=expression(hat(theta)))
qqnorm(Sthetahat);qqline(Sthetahat)
ks.test(Sthetahat,"pnorm",mean(Sthetahat),
sd(Sthetahat))$p
[1] 0.7324885
```



MC Interval Estimation

Thus a $(1-\alpha)100\%$ confidence interval is $\hat{\theta} \pm z_{\alpha/2} se(\hat{\theta})$
So in the last example, 95% C.I. for $\theta = \mathbf{E}(|X_1 - X_2|)$ is

```
thetahat<-mean(Sthetahat)
se<-sd(Sthetahat)
CL<-.95
c(thetahat-qnorm((1+CL)/2)*se,
thetahat+qnorm((1+CL)/2)*se)
[1] 1.111559 1.145149
```

Estimate coverage probability

```
mean((1.111559 < Sthetahat)&(Sthetahat < 1.145149))
[1] 0.9505
```

Empirical 95% C.I. (credible interval)

```
quantile(Sthetahat,c(.025,.975))
    2.5%    97.5%
1.111589 1.145112
```

MC MSE Estimation

To estimate the bias use $bias = \mathbf{E}(\hat{\theta}) - \theta$

To estimate mean squared error $MSE = \mathbf{E}\left[(\hat{\theta} - \theta)^2\right]$ use

$$\hat{MSE} = \frac{1}{m}\sum_{j=1}^{m}\left(\hat{\theta}_j - \theta\right)^2$$

where m is the number of replications of the MC experiments to find $\hat{\theta}_1, \ldots, \hat{\theta}_m$.

In the example,

```
bias<-mean(Sthetahat)-2/sqrt(pi)
bias
[1] 4.280798e-05
MSEhat<-mean((Sthetahat-2/sqrt(pi))^2)
MSEhat
[1] 7.523616e-05
```

MC in Hypothesis Testing

Test of Hypothesis

Testing

$$H_0 : \theta \in \Theta_0 \text{ vs } H_a : \theta \in \Theta_1$$

where the parameters space $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \phi$. The conclusion is to reject H_0 if the

$$p - value = P(\hat{\theta} \in C | H_0 \text{ is true}) < \alpha$$

α is called the level of significance or the probability of committing type I error defined as

$$\alpha = P(\text{ reject } H_0 | H_0 \text{ is true})$$

the probability of committing type II error defined as

$$\beta = P(\text{ accept } H_0 | H_a \text{ is true})$$

and the power of the test is given by

$$power = 1 - \beta = P(\text{ reject } H_0 | H_a \text{ is true})$$

Type I error

Let the power function be defined by

$$\pi(k) = P(\text{ reject } H_0 | \theta = k)$$

for $k \in \Theta$

- then we define

$$\alpha = \sup_{k \in \Theta_0} \pi(k)$$

- also the power of the test when $\theta = k \in \Theta_1$ is $\pi(k)$

Type I error

Algorithm: Pre-assign a level of significant α

1. Generate a random sample of size n from the model of H_0
2. Compute the test statistics TS using the random sample
3. Record $I = 1$ if the test is significant (H_0 is rejected) at the α
4. Repeat the previous steps m times
5. estimate the probability of type I error by $\hat{p} = \frac{1}{m}\sum_{j=1}^{m} I_j$ and its standard error by

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}}$$

Type I error

Recall: Skewness is the degree of asymmetry of a probability distribution measured by Pearson's moment coefficient of skewness

$$\gamma_1 = \mathbf{E}\left(\left(\frac{X - \mu_X}{\sigma_X}\right)^3\right)$$

If $\gamma_1 = 0$ then the density is symmetric, and it is right (positive) skewed if $\gamma_1 > 0$. It is negative (left) skewed if $\gamma_1 < 0$.

Type I error

It is estimated by

$$\hat{\gamma}_{1,n} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{3/2}}$$

Theoretically, if $X \sim N(\mu, \sigma)$

$$\frac{\hat{\gamma}_{1,n}}{\sqrt{6/n}} \to Z \text{ in distribution}$$

as $n \to \infty$ where $Z \sim N(0, 1)$, but the convergence is slow and the test statistic $\frac{\hat{\gamma}_{1,n}}{\sqrt{6/n}}$ might not be close enough to the standard normal distribution for small values of n values.
How does that affect probability of type I error?

Type I error

Example: Test skewness (normality)

$$H_0 : \gamma_1 = 0 \ vs \ H_a : \gamma_1 \neq 0$$

for different values of n and estimate the probability of type I error.

Remember the test statistic $TS = \dfrac{\hat{\gamma}_{1,n}}{\sqrt{6/n}}$ and the test is significant when $|TS| > z_{\alpha/2}$

Type I error

Example: Test skewness (normality)

$$H_0 : \gamma_1 = 0 \ vs \ H_a : \gamma_1 \neq 0$$

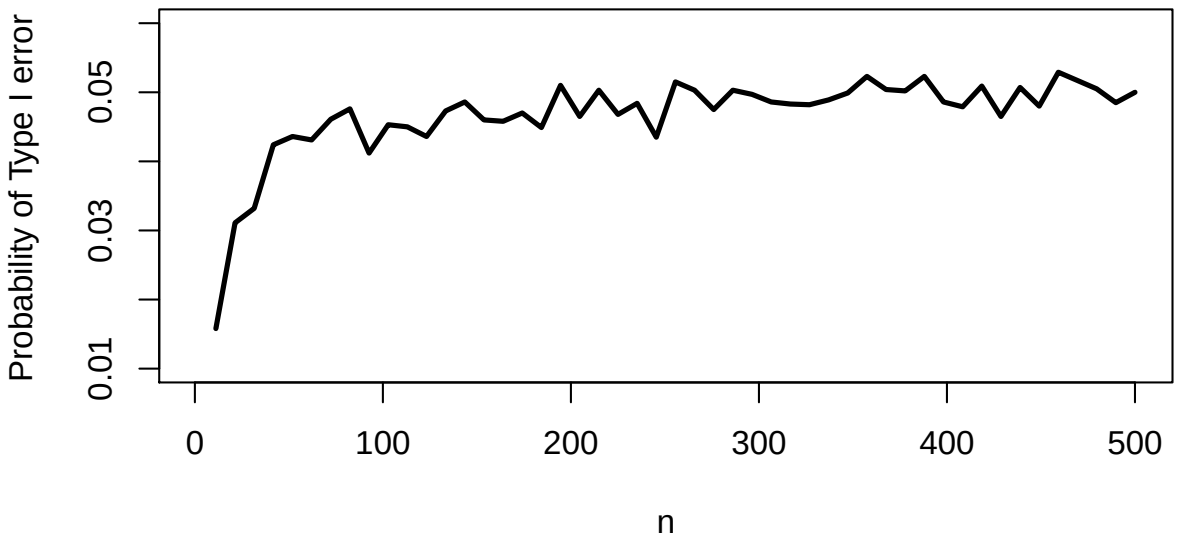
```
alpha<-.05
p<-function(n){
I<-replicate(10000,
{x<-rnorm(n)
xbar<-mean(x)
k3<-mean((x-xbar)^3)
k2<-mean((x-xbar)^2)
gamma1<-k3/(k2)^(3/2)
TS<-gamma1/sqrt(6/n)
(abs(TS)>qnorm(1-alpha/2))})
mean(as.integer(I))}
n<-seq(1,500,length=50)
phat<-sapply(n,p)
```

Type I error

Example: Test skewness (normality)

$$H_0 : \gamma_1 = 0 \ vs \ H_a : \gamma_1 \neq 0$$

```
plot(N,phat,type="l",lwd=2.5,xlab="n",ylab="
Probability of Type I error",ylim=c(.01,.06))
```



But

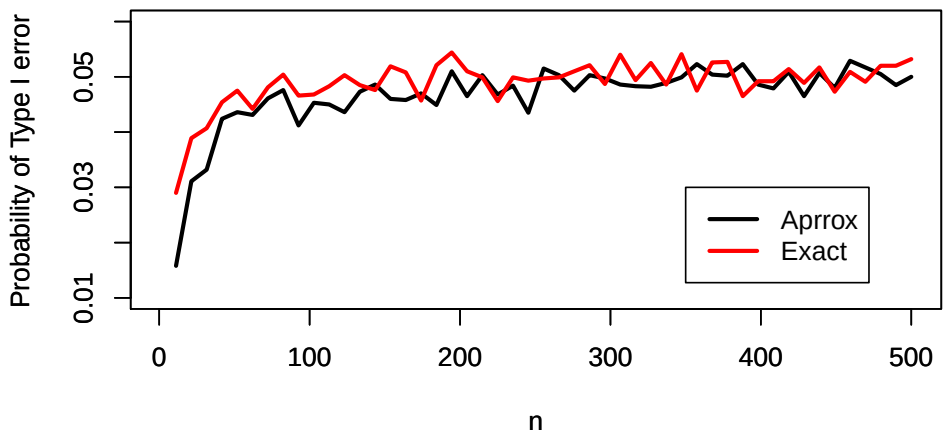
Type I error

The exact form of the variance

$$\mathbf{V}(\hat{\gamma}_{1,n}) = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}$$

Use $TS = \dfrac{\hat{\gamma}_{1,n}}{\sqrt{\mathbf{V}(\hat{\gamma}_{1,n})}}$

```
TS<-gamma1/sqrt((6*n*(n-1))/((n-2)*(n+1)*(n+3)))
legend(350,.03,c("Aprrox","Exact"), lty=c(1,1),
lwd=c(2.5,2.5),col=c("black","red"))
```



Type II error (Power of the test)

- The power function for each $k \in \Theta$ is

$$\pi(k) = P(\text{ reject } H_0 | \theta = k)$$

estimated by $\hat{\pi}(k)$

- Then probability of Type II error is $\beta = 1 - \hat{\pi}(k)$ when $\theta = k \in \Theta_1$

Type II error (Power of the test)

A skewed-normal distribution $N(0,1,\delta)$ is defined by a pdf

$$f(x) = 2\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \cdot \Phi(\delta x)$$

for $-\infty < x < \infty$, where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\,dz$$

Note that $N(0,1,0)$ is the normal distribution $N(0,1)$

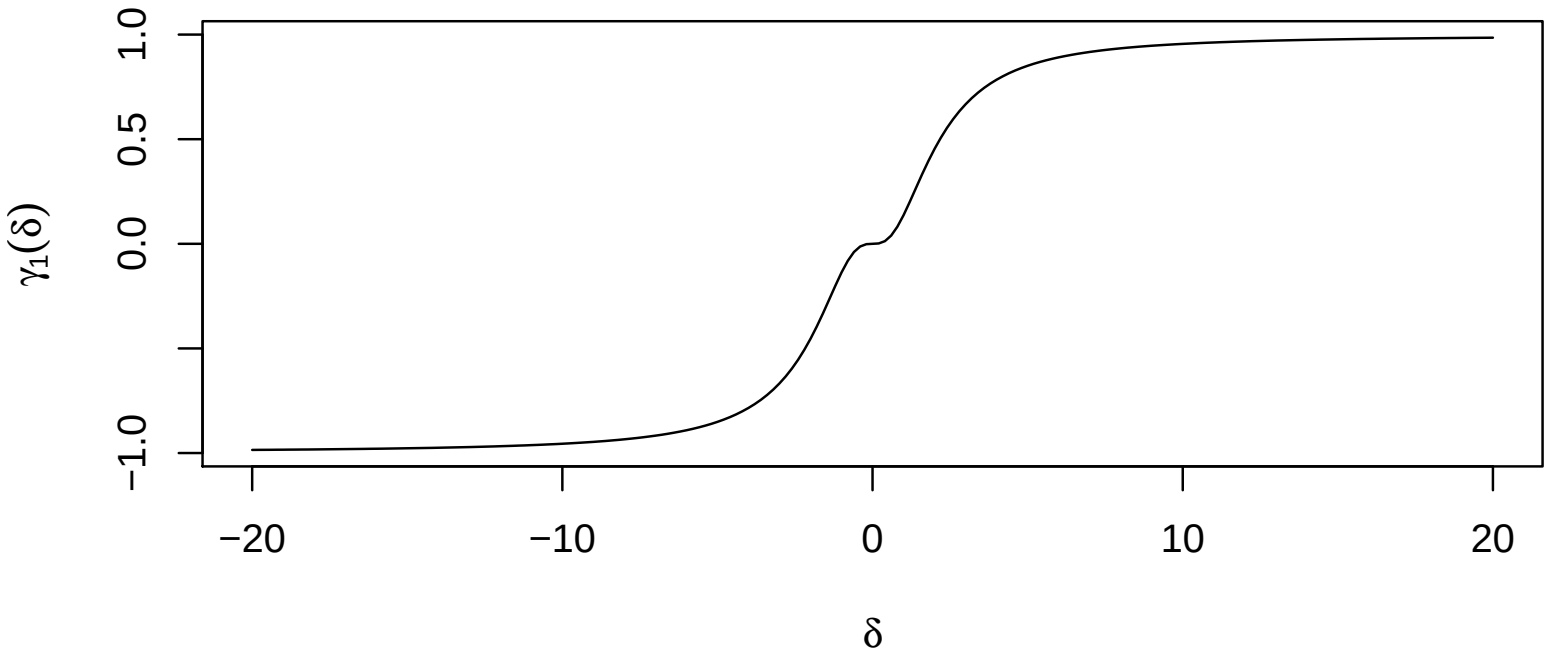
Type II error (Power of the test)

The measure of skewness is given by

$$\gamma_1 = \frac{4-\pi}{2} \frac{(\sqrt{2/\pi})^3}{(1-\frac{2\delta^2}{\pi(1+\delta^2)})^{3/2}} (\frac{\delta}{\sqrt{1+\delta^2}})^3$$

```
gamma1<-function(k)
{((4-pi)/2)*(sqrt(2/pi)*(k/sqrt(1+k^2)))^3
/(1-(2/pi)*((k/sqrt(1+k^2)))^2)^1.5}
k<-100:100
plot(k/5,gamma1(k/5),type='l',xlab=expression(delta),
ylab=expression(gamma[1](delta)))
```

Type II error (Power of the test)



for only $\gamma_1 \in (-1, 1)$
So testing $\gamma_1 = 0$ is equivalent to $\delta = 0$

Type II error (Power of the test)

Example: To test skewness (normality) assuming the alternative is $N(0, 1, \delta)$ with a sequence of nonzero δ real-values, we are testing

$$H_0 : \delta = 0 \text{ vs } H_1 : \delta \neq 0$$

So estimate

$$\pi(k) = P(\text{ reject } H_0 | \delta = k)$$

for $k = -10, -9, \ldots, 9, 10$, use $\alpha = .05$ and $n = 30$.
Remember the test statistic

$$TS = \frac{\hat{\gamma}_{1,n}}{\sqrt{(6 * n * (n-1))/((n-2) * (n+1) * (n+3))}}$$

and the test is significant when $|TS| > z_{\alpha/2}$

Type II error (Power of the test)

Algorithm:

1. For each $\delta = k$ where $k = -10, -9, \ldots, 9, 10$, generate a random sample from $N(0, 1, \delta)$
2. Use that sample to calculate the test statistic TS
3. Record $I = 1$ if the test is significant (H_0 is rejected) when $|TS| > z_{\alpha/2}$
4. Repeat the previous steps m times
5. estimate the power function π at k by $\hat{\pi}(k) = \frac{1}{m}\sum_{j=1}^{m} I_j$ and its standard error by

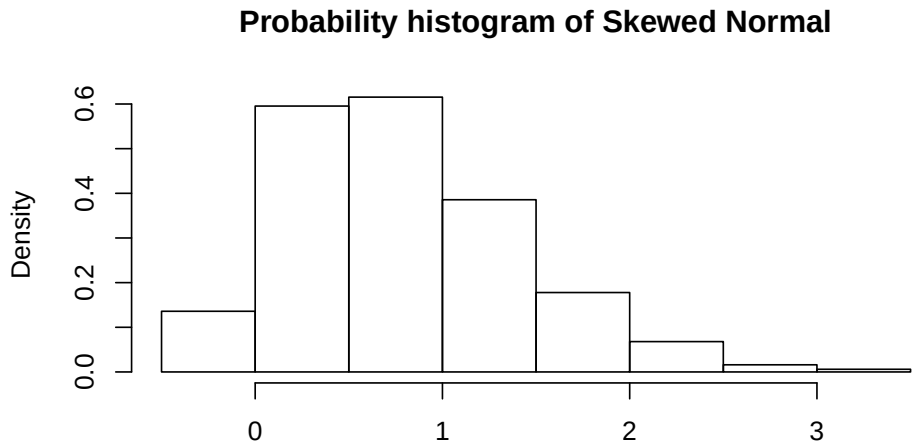
$$se(\hat{\pi}(k)) = \sqrt{\frac{\hat{\pi}(k)(1-\hat{\pi}(k))}{m}}$$

Type II error (Power of the test)

How to generate a random sample from $N(0, 1, \delta)$ (for step 1)?
Use acceptance rejection with acceptance condition $U < \frac{f(x)}{cg(x)}$ and proposal normal pdf $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and $c = 2$ since

$$\frac{f(x)}{g(x)} = 2\Phi(\delta x) \leq 2 \text{ and } \frac{f(x)}{cg(x)} = \Phi(\delta x)$$

```
n<-1000;k<-5;vx<-c()
while(length(vx)<=n){
 x<-rnorm(1)
 u<-runif(1)
 if(u<pnorm(k*x)){
 vx<-c(vx,x) } }
hist(vx,prob=T,main="Probab
histogram of Skewed
Normal")
```



Type II error (Power of the test)

Make a function rskewnorm(n,k)

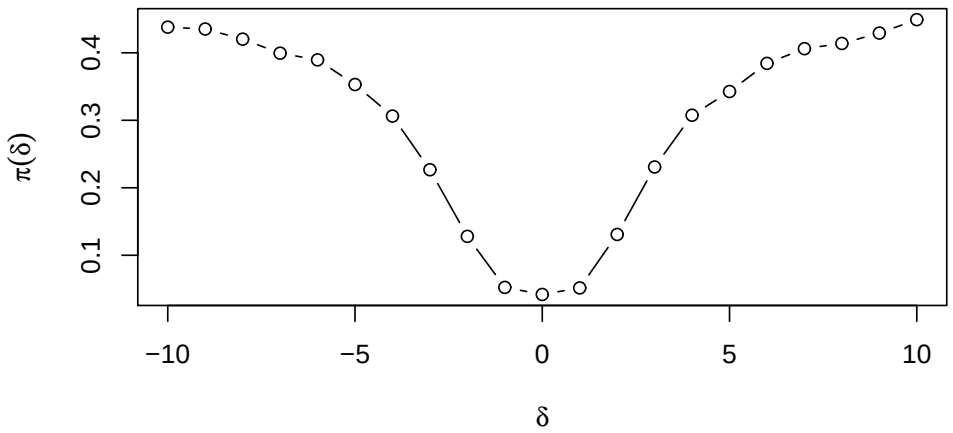
```
rskewnorm<-function (n,k){vx<-c()
while(length(vx)<=n){
 x<-rnorm(1)
 u<-runif(1)
 if(u<pnorm(k*x)){
 vx<-c(vx,x) } }
return(vx) }
```

Type II error (Power of the test)

```
n<-30;m<-10000;alpha<-.05
powfn<-function(n,k){
 I<-replicate(m,{x<-rskewnorm(n,k)
 xbar<-mean(x)
 k3<-mean((x-xbar)^3)
 k2<-mean((x-xbar)^2)
 gamma1<-k3/(k2)^(3/2)
 TS<-gamma1/sqrt((6*n*(n-1))/((n-2)*(n+1)*(n+3)))
 (abs(TS)>qnorm(1-alpha/2))})
 mean(as.integer(I))}
```

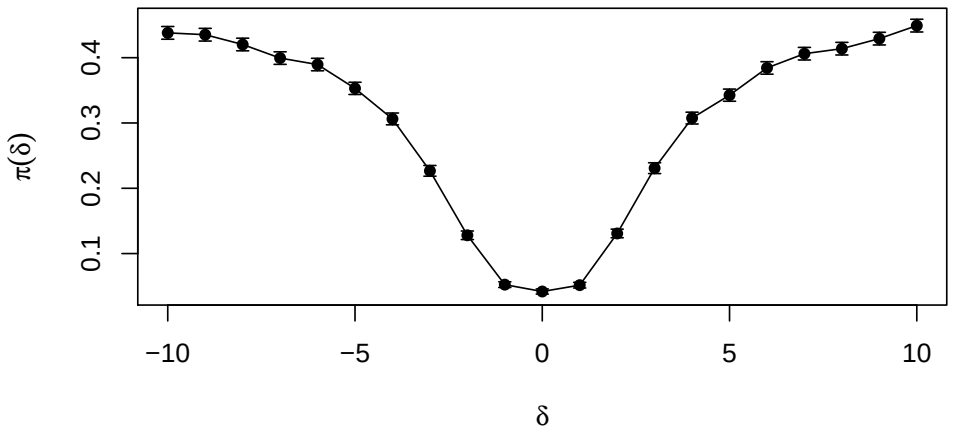
Type II error (Power of the test)

```
k<-seq(-10,10,by=1)
powfnhat<-sapply(k,function(x) powfn(n,x))
plot(k,powfnhat,type="b",xlab=expression(delta)
,ylab=expression(pi(delta)))
```



Type II error (Power of the test)

```
se<-sqrt(powfnhat*(1-powfnhat)/m)
install.packages("Hmisc")
library(Hmisc)
errbar(k,powfnhat,yplus=powfnhat+1.96*se
,yminus=powfnhat-1.96*se,xlab=expression(delta)
,ylab=expression(pi(delta)))
lines(k,powfnhat)
```



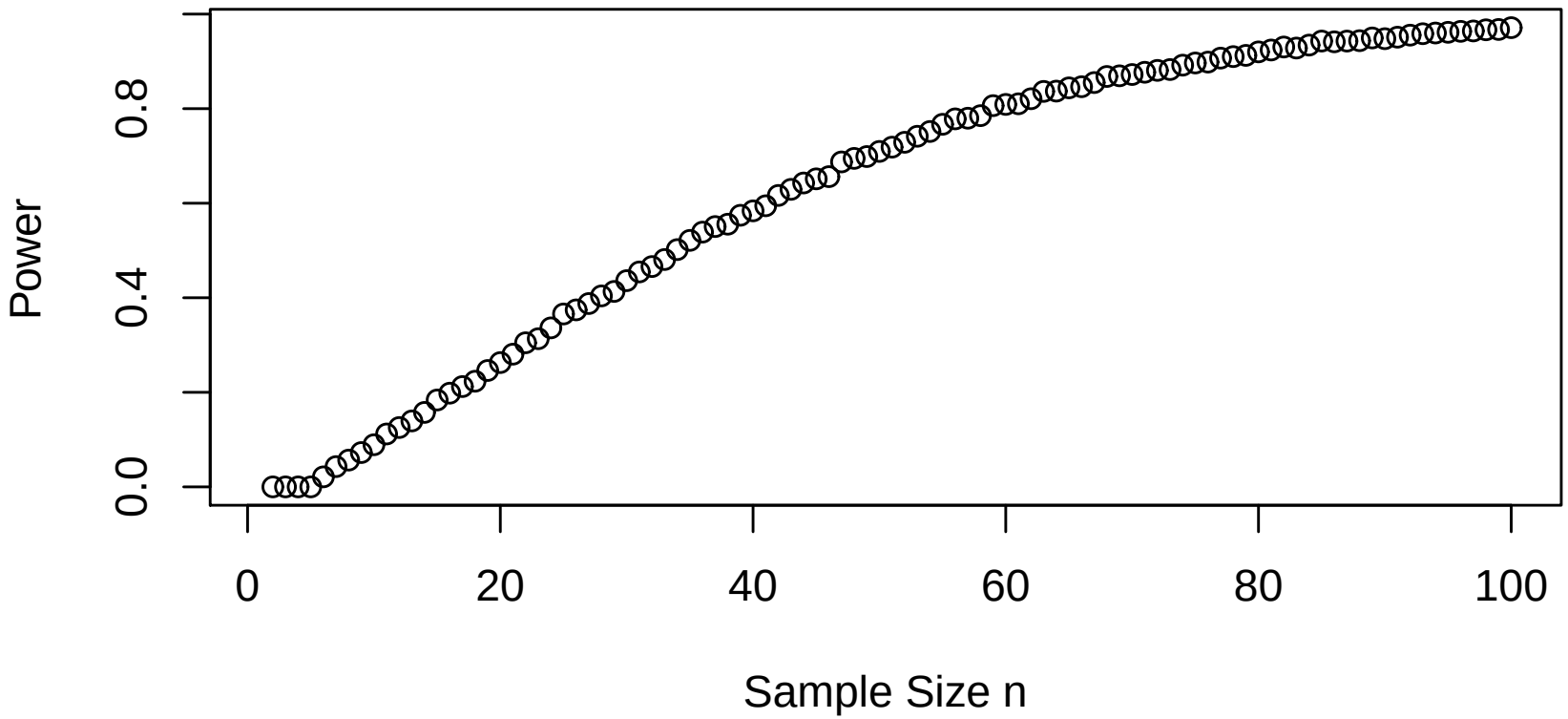
Type II error (Power of the test)

Example: Do power analysis for the sample size *n vs* power of the test at some effect you want to observe, say $e = 10$, at $\alpha = .05$. That means you want to find the appropriate sample size to detect a skewness as large as $\delta = 10$ (assuming skewed-normal alternative). The relation between *n* and power=$\pi(10)$

```
e<-10
n<-seq(1,100,by=1)
powfne<-sapply(n,function(x) powfn(x,e))
plot(n,powfne,type="b",xlab="Sample size n"
,ylab="Power")
```

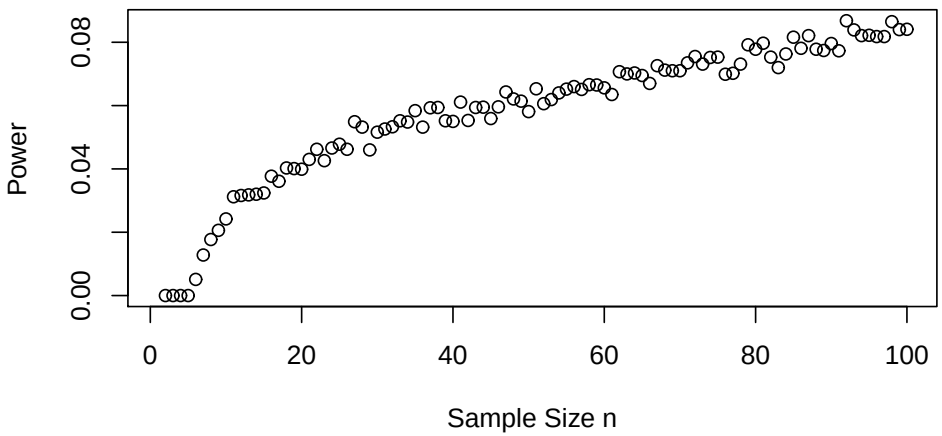
Type II error (Power of the test)

```
n[min(which(powfne>=.8))]
[1] 59
```

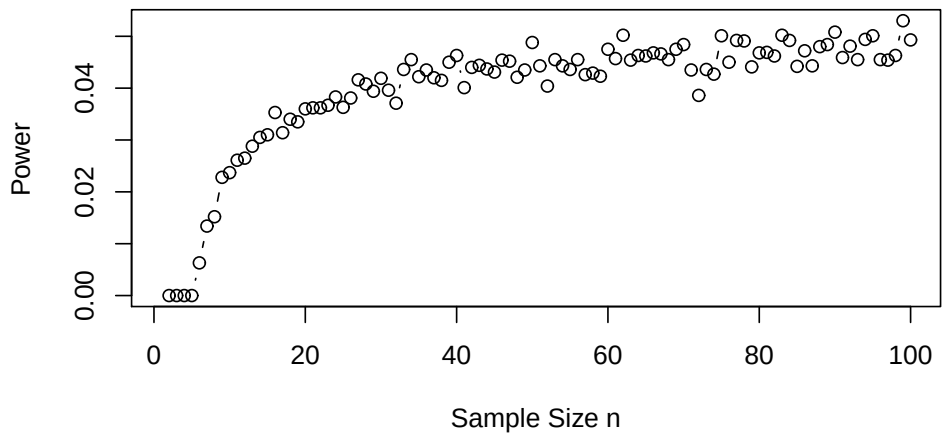


Type II error (Power of the test)

```
e<-1
```




```
e<-.1
```



Maximum likelihood method

Maximum likelihood method

- The likelihood (probability) $L(x|\theta)$ or $L(\theta)$ of observing that simple random sample $x_1, x_2, \ldots, x_n$ of i.i.d. measurements is

$$L(\theta) := L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, \ldots, \theta_k)$$

  by independence

- The maximum likelihood principle (due to Fisher) finds the MLE $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ that maximize the likelihood $L(\theta)$ of observing those observations $x_1, x_2, \ldots, x_n$.

- Invariance property: If $\hat{\theta}$ is an MLE of θ then $g(\hat{\theta})$ is an MLE of $g(\theta)$

Maximum likelihood method

- It is equivalent to maximize $\ell(\theta) := \log(L(\theta)) = \sum_{i=1}^{n} \log(f(x_i|\theta))$ (or minimize $-\ell(\theta)$).
  That is, solve

$$\frac{d\ell(\theta)}{d\theta} = 0$$

  which might be a system of nonlinear equations.

- Fisher score vector (function)

$$s(\theta) = \frac{d\ell(\theta)}{d\theta}$$

- Fisher information matrix

$$\mathcal{I}(\theta) = -\mathbf{E}_\theta(\ell''(\theta))$$

Maximum likelihood method

MLE's are asymptotically normal and so as $n \to \infty$

$$\hat{\theta} \to N(\theta, \mathcal{I}^{-1}(\theta)) \text{ in distribution}$$

and the variance of $\hat{\theta}$ is estimated by $\mathcal{I}^{-1}(\hat{\theta})$.

Maximum likelihood method

Example: Find the MLE of the rate λ in an exponential model given the data $x_1, \ldots, x_n$.
The likelihood function is given by

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-n\lambda \bar{x}}$$

and

$$\ell(\lambda) = n \log(\lambda) - n\lambda \bar{x}$$

and so

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - n\bar{x}$$

and so solving

$$\frac{n}{\lambda} - n\bar{x} = 0$$

gives $\hat{\lambda} = \frac{1}{\bar{x}}$.

Maximum likelihood method

If the sample is given by

$$4.679781, 9.325474, 2.491352, 0.841366, 2.982121$$

(FYI: I used *rexp*(5, .373))
Then $\bar{x} = 4.679781$ and so $\hat{\lambda} = 0.2136852$. (If you increase n you will get better estimate value.)

Maximum likelihood method

That could be automated in one of three ways either

1. Find the root of $\frac{d\ell(\theta)}{d\theta} = 0$. `uniroot` in R.

   OR

2. Optimize (minimize) the function $-\ell(\theta)$. `optimize`or `optim` in R *default* Nelder-Mead method

   OR

3. Use *mle* function in the R library stats4

Maximum likelihood method

Root finding methods (I leave them to Numerical Analysis course)

1. Bisection methods
2. Newton–Raphson method
3. Secant method
4. Aitken acceleration and Steffensen's method

Maximum likelihood method

Optimization methods (I leave them to Numerical Analysis course)

- Local optimization

(a) Golden search method

(b) Steepest descent method

(c) Nelder-Mead simplex (direct search) method

- Global optimization

(a) Genetic algorithms

(b) Simulated annealing

(c) Branch and Bound method

Maximum likelihood method (using root finding)

Example: Find the MLE of the rate $\lambda \in \Theta = \mathbb{R}^+$ in an exponential model given the data $x_1, \ldots, x_n$.

```
sample<-c(4.679781, 9.325474, 2.491352, 0.841366,
2.982121)
dlogLexp<-function(lambda,x){length(x)/lambda-sum(x)}
uniroot(dlogLexp,lower=.0001,upper=10e3,x=sample)
$root
[1] 0.2460903
$f.root
[1] -0.002346432
$iter
[1] 22
$init.it
[1] NA
$estim.prec
[1] 6.103516e-05
```

Maximum likelihood method (using optimization)

Example: Find the MLE of the rate $\lambda \in \Theta = \mathbb{R}^+$ in an exponential model given the data $x_1, \ldots, x_n$.

```
sample<-c(4.679781, 9.325474, 2.491352, 0.841366,
2.982121)
neglogLexp<-function(lambda,x){-1*(length(x)*
log(lambda)-sum(x)*lambda)}
optim(c(.1),neglogLexp,lower=.0001,upper=10e3,
x=sample)
$par
[1] 0.2460632
$value
[1] 12.01086
$counts
function gradient
 12 12
$convergence
```

Maximum likelihood method (using optimization)

Example: Find the MLE of the rate $\lambda \in \Theta = \mathbb{R}^+$ in an exponential model given the data $x_1, \ldots, x_n$.

```
sample<-c(4.679781, 9.325474, 2.491352, 0.841366,
2.982121)
logLexp<-function(lambda,x){length(x)*
log(lambda)-sum(x)*lambda}
optimize(logLexp,lower=.0001,upper=10e3
,x=sample,maximum=T)
$maximum
[1] 0.2460805
$objective
[1] -12.01086
```

Maximum likelihood method (using *mle*)

Example: Find the MLE of the rate $\lambda \in \Theta = \mathbb{R}^+$ in an exponential model given the data $x_1, \ldots, x_n$.

```
library(stats4)
x<-c(4.679781, 9.325474, 2.491352, 0.841366,
2.982121)
neglogLexp<-function(lambda){-1*(length(x)*
log(lambda)-sum(x)*lambda)}
results<-mle(neglogLexp,start=list(lambda=1))
summary(results)
Maximum likelihood estimation
Call:
mle(minuslogl=neglogLexp,start=list(lambda=1))
Coefficients:
 Estimate Std.  Error
lambda 0.2460615 0.1100402
-2 log L: 24.02172
```

Maximum likelihood method (using optimization)

Normality and variance: Fisher information matrix

$$\mathcal{I}(\lambda) = -\mathbf{E}_\lambda(\ell''(\lambda)) = \mathbf{E}_\lambda(\frac{n}{\lambda^2})$$

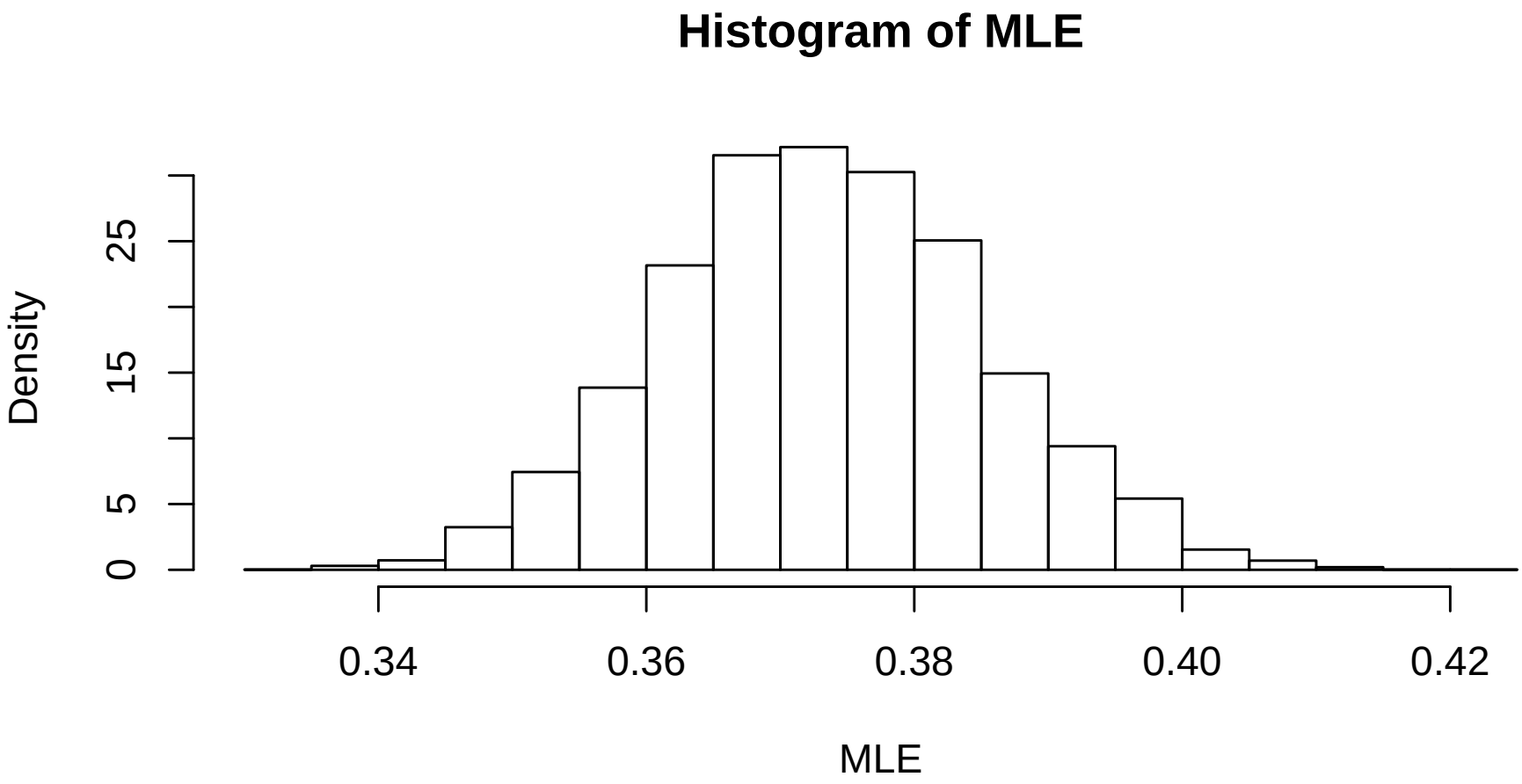
and $\mathbf{V}(\hat{\lambda}) = \mathcal{I}^{-1}(\lambda)$ and

```
MLE<-replicate(10000,{
 sample<-rexp(1000,.373)
 optimize(logLexp,lower=.0001,upper=10e3
,x=sample,maximum=T)$maximum})
mean(MLE)
[1] 0.3733541
var(MLE)
[1] 0.0001409264
(1000*mean(1/MLE^2))^(-1)
[1] 0.0001389718
```

(Note that $.373^2/n = 0.000139129$)

Maximum likelihood method (using optimization)

```
hist(MLE,prob=T)
```



EM algorithm

Expectation – Maximization (EM) algorithm

- EM is used for incomplete data, e.g. missing data, censored data or latent variables.

- If the complete data $X = (O, M)$ where O is the observed data and M is the missing data.

- Note that

$$f(X|\theta) = f(M|\theta, O) \cdot f(O|\theta)$$

That is

$$L(\theta|X) = f(M|\theta, O) \cdot L(\theta|O)$$

- $L(\theta|X)$ is the complete likelihood function and $L(\theta|O)$ is the incomplete likelihood function

Expectation – Maximization (EM) algorithm

and

$$logL(\theta|O) = logL(\theta|X) - log(f(M|\theta, O))$$

so

$$\int logL(\theta|O)f(M|\theta', O)dM = \int logL(\theta|X)f(M|\theta', O)dM$$

$$- \int log(f(M|\theta, O))f(M|\theta', O)dM$$

or

$$logL(\theta|O) = \mathbf{E}_{M|\theta',O}(logL(\theta|X)) - \mathbf{E}_{M|\theta',O}(log(f(M|\theta, O)))$$

Expectation – Maximization (EM) algorithm

EM Algorithm:

1. Start from initial point $\theta^{(0)}$, then for each $k \geq 1$
2. **E** step: Find $Q_k(\theta|\theta^{(k)}) := \mathbf{E}_{M|\theta^{(k)},O}(logL(\theta|X))$
3. **M** step: Find $\theta^{(k+1)} = \text{argmax}_{\theta\in\Theta} Q_k(\theta|\theta^{(k)})$
4. Stop when $\left|\theta^{(k+1)} - \theta^{(k)}\right| / \theta^{(k)} < TOL$

Remark: Convergence is theoretically guaranteed.

Expectation – Maximization (EM) algorithm

Example: Let $x_1, x_2, \ldots, x_n$ be an observed data of completion time at checking out at a grocery store with two cashiers and no waiting lines. They are modeled by a mixture of two exponential distributions with rates λ_1 and λ_2 with probability of selection (mixture weights) p and $1 - p$.
The parameter vector is $\theta = (p, \lambda_1, \lambda_2)$ and $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$.
The incomplete likelihood function

$$L(\theta|x) = \prod_{i=1}^{n} (p \cdot f(x_i|\lambda_1) + (1 - p) \cdot f(x_i|\lambda_2) \cdot)$$

But what we didn't observe is from where each data point is coming from. That corresponds to latent variable $z_1, \ldots, z_n$ for which cashier was selected, encoded as $z_i = 1$ if cashier 1 is selected and $z_i = 0$ if cashier 2 is selected.

Expectation – Maximization (EM) algorithm

By Bayes' theorem

$$p_i := P(Z_i = 1|X = x_i, \theta) = \frac{p \cdot f(x_i|\lambda_1)}{p \cdot f(x_i|\lambda_1) + (1-p) \cdot f(x_i|\lambda_2)}$$

and the complete likelihood function

$$L(\theta|x, z) = \prod_{i=1}^{n} (z_i p \cdot f(x_i|\lambda_1) + (1 - z_i)(1-p) \cdot f(x_i|\lambda_2) \cdot)$$

and

$$\mathbf{E}_{Z|X,\theta}(logL(\theta|x, z)) = \sum_{j=0}^{1} \sum_{i=1}^{n} \log (z_i p \cdot f(x_i|\lambda_1) + (1 - z_i)(1-p)$$

$$\cdot f(x_i|\lambda_2)) \cdot P(Z_i = j|X = x_i, \theta)$$

Expectation – Maximization (EM) algorithm

Thus, E step:

$$Q_k(\theta|\theta^{(k)}) = \mathbf{E}_{Z|X,\theta^{(k)}}(logL(\theta|x,z)) =$$

$$\sum_{i=1}^{n} p_i^{(k)}(\log(p) + \log(f(x_i|\lambda_1))) + (1 - p_i^{(k)})(\log(1-p) + \log(f(x_i|\lambda_2))) =$$

$$\sum_{i=1}^{n} \left(p_i^{(k)} \log(p) + (1 - p_i^{(k)}) \log(1-p) \right) +$$

$$\sum_{i=1}^{n} \left(p_i^{(k)} \log(f(x_i|\lambda_1)) + (1 - p_i^{(k)}) \log(f(x_i|\lambda_2)) \right)$$

where

$$p_i^{(k)} := P(Z_i = 1|X = x_i, \theta^{(k)}) = \frac{p^{(k)} \cdot f(x_i|\lambda_1^{(k)})}{p^{(k)} \cdot f(x_i|\lambda_1^{(k)}) + (1 - p^{(k)}) \cdot f(x_i|\lambda_2^{(k)})}$$

Expectation – Maximization (EM) algorithm

And, M step can be split into
M sub-step 1: Find

$$p^{(k+1)} = \text{argmax}_{p\in(0,1)} \sum_{i=1}^{n} \left(p_i^{(k)} \log(p) + (1 - p_i^{(k)}) \log(1 - p) \right)$$

M sub-step 2: Find

$$\lambda_1^{(k+1)} = \text{argmax}_{\lambda_1\in(0,\infty)} \sum_{i=1}^{n} \left(p_i^{(k)} \log(f(x_i|\lambda_1)) \right)$$

M sub-step 3: Find

$$\lambda_2^{(k+1)} = \text{argmax}_{\lambda_2\in(0,\infty)} \sum_{i=1}^{n} \left((1 - p_i^{(k)}) \log(f(x_i|\lambda_2)) \right)$$

The last two are weighted MLE's.

Expectation – Maximization (EM) algorithm

M sub-step 1: Gives

$$p^{(k+1)} = \frac{\sum_{i=1}^{n} p_i^{(k)}}{n}$$

M sub-step 2: Gives

$$\lambda_1^{(k+1)} = \frac{\sum_{i=1}^{n} p_i^{(k)}}{\sum_{i=1}^{n} p_i^{(k)} x_i}$$

M sub-step 3: Gives

$$\lambda_2^{(k+1)} = \frac{\sum_{i=1}^{n} (1 - p_i^{(k)})}{\sum_{i=1}^{n} (1 - p_i^{(k)}) x_i}$$

where

$$p_i^{(k)} := P(Z_i = 1 | X = x_i, \theta^{(k)}) = \frac{p^{(k)} \cdot f(x_i | \lambda_1^{(k)})}{p^{(k)} \cdot f(x_i | \lambda_1^{(k)}) + (1 - p^{(k)}) \cdot f(x_i | \lambda_2^{(k)})}$$

Expectation – Maximization (EM) algorithm

Example: Use a sample of $n = 1000$ generated from a mixture of $exp(\lambda_1 = .3)$ and $exp(\lambda_2 = .5)$ with probabilities $p = .2$ and $1 - p = .8$, respectively, to estimate p, λ_1 and λ_2.
You consider them as 1000 finishing time of 1000 transactions through two different cashiers that you have collected.
First, generate the 1000 points

```
n<-1000;p<-.2;lambda1<-.3;lambda2<-.5
lambda<-c(lambda1,lambda2)
K<-sample(1:2,n,prob=c(p,1-p),rep=T)
x<-rexp(n,rate=lambda[K])
```

Expectation – Maximization (EM) algorithm

```
TOL<-1e-8;j<-0
pold<-0.9;lambda1old<-0.1;lambda2old<-0.9
pnew<-.1;lambda1new<-1;lambda2new<-.1
vpnew<-pnew*dexp(x,lambda1new)/
(pnew*dexp(x,lambda1new)+(1-pnew)*dexp(x,lambda2new))
while(max(abs(pnew-pold)/pold,
abs(lambda1new-lambda1old)/lambda1old,
abs(lambda2new-lambda2old)/lambda2old)>TOL){
 j<-j+1
 pold<-pnew
 lambda1old<-lambda1new
 lambda2old<-lambda2new
 vpold<-vpnew
 pnew<-mean(vpold)
 lambda1new<-1/weighted.mean(x, vpold)
 lambda2new<-1/weighted.mean(x, 1-vpold)
```

Expectation – Maximization (EM) algorithm

```
 vpnew<-(pnew*dexp(x,lambda1new))/
(pnew*dexp(x,lambda1new)+(1-pnew)*dexp(x,lambda2new))
}
j
[1] 6074
pnew
[1] 0.8089904
lambda1new
[1] 0.512815
lambda2new
[1] 0.2730581
```

Why the switch? Look at the initial values of the parameters.

Practical advice: Use different initial values of parameters λ_1 and λ_2 or p will not get updated ($p^{(k)} = p^{(0)}$ for all k).

Likelihood Ratio Test

Likelihood Ratio (LR) Test

Testing

$$H_0 : \theta \in \Theta_0 \text{ vs } H_a : \theta \in \Theta_1$$

where the parameters space $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \phi$.
Use the test statistic

$$LR(x) := \frac{\sup_{\theta \in \Theta_0} L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)}$$

The conclusion is to reject H_0 if the test statistic

$$LR(x) < c$$

where $0 \leq c \leq 1$ is such that

$$P(LR(X) < c | H_0 \text{ is true}) = \alpha$$

The LR test is the most powerful test according to Neyman–Pearson Lemma.

Likelihood Ratio (LR) Test

But, how can we find that probability

$$P(LR(X) < c | H_0 \text{ is true}) = \alpha$$

if we don't know the probability distribution of $LR(X)$?
The exact probability distribution is hard to find, yet

Theorem (Wilks' Theorem)

The deviance

$$D = -2\log(LR(x_1, \ldots, x_n)) \to \chi^2(df) \text{ in distribution}$$

as $n \to \infty$, given that H_0 is true, where $df = dim(\Theta) - dim(\Theta_0)$.

Wilks' theorem provides approximate test in which H_0 is rejected at level of significance α if $D > \chi^2_\alpha(df)$.

Likelihood Ratio (LR) Test

- First, note that the unrestricted optimization results in $\sup_{\theta \in \Theta} L(\theta|x) = L(\hat{\theta}|x)$ where $\hat{\theta}$ is the MLE of θ. Thus,

$$LR(x) := \frac{\sup_{\theta \in \Theta_0} L(\theta|x)}{L(\hat{\theta}|x)}$$

- If $\Theta_0 = \{\theta_0\}$; that is, we are testing

$$H_0 : \theta = \theta_0 \text{ vs } H_a : \theta \neq \theta_0$$

then

$$LR(x) := \frac{L(\theta_0|x)}{L(\hat{\theta}|x)}$$

Likelihood Ratio (LR) Test

- Testing simple hypothesis

$$H_0 : \theta = \theta_0 \ vs \ H_a : \theta = \theta_1$$

then

$$LR(x) := \frac{L(\theta_0|x)}{\sup_{\theta \in \{\theta_0, \theta_1\}} L(\theta|x)}$$

- or sometimes

$$LR(x) := \frac{L(\theta_0|x)}{L(\theta_1|x)}$$

Likelihood Ratio (LR) Test

- Let Model 1 and Model 2 be nested that is Model 1 is a special case of Model 2
- For example,
  Model 1: $X \sim exp(\lambda)$
  Model 2: $X \sim gamma(r, \lambda)$
  Model 1 is a special case of Model 2 when $r = 1$ and so Model 1 is nested within Model 2.
- To test nested models

$$H_0 : X \sim \text{Model 1 } vs \; H_a : X \sim \text{Model 2}$$

Use the test statistic
$D = -2\log(LR(x_1, \ldots, x_n)) \sim \chi^2(df)$ where
$df = N_{\text{Model 2}} - N_{\text{Model 1}}$ where N_{Model} is the number of parameters of the Model.

Likelihood Ratio (LR) Test

Example: If the sample is given by

$$4.679781, 9.325474, 2.491352, 0.841366, 2.982121$$

(FYI: I used *rexp*$(5, .373)$)
Test

$$H_0 : \lambda = .373 \text{ vs } H_a : \lambda \neq .373$$

Using the test statistic

$$LR(x) := \frac{L(.373|x)}{L(\hat{\lambda}|x)}$$

where $\hat{\lambda} = 0.2136852$ is the MLE, and x is the sample.
But again, what is the distribution of $LR(X)$ when H_0 is true($\lambda = .373$).

Likelihood Ratio (LR) Test

```
sample<-c(4.679781, 9.325474, 2.491352, 0.841366,
2.982121)
logLexp<-function(lambda,x)
{length(x)*log(lambda)-sum(x)*lambda}
lambda0<-.373;lambdahat<-0.2136852;n<-length(sample)
LR<-replicate(1000,{
 x<-rexp(n,lambda0)
 exp(logLexp(lambda0,x))/exp(logLexp(lambdahat,x))})
c<-quantile(LR,.05)
LR0<-exp(logLexp(lambda0,sample))/
exp(logLexp(lambdahat,sample))
if(LR0<c) paste("Reject H0") else paste("Do not
reject H0")
[1] "Do not reject H0"
```

Likelihood Ratio (LR) Test

This time test

$$H_0 : \lambda = .373 \ vs \ H_a : \lambda = 2.383$$

Using the test statistic

$$LR(x) := \frac{L(.373|x)}{\max(L(.373|x), L(2.383|x))}$$

where x is the sample.

Likelihood Ratio (LR) Test

```
sample<-c(4.679781, 9.325474, 2.491352, 0.841366,
2.982121)
logLexp<-function(lambda,x)
{length(x)*log(lambda)-sum(x)*lambda}
lambda0<-.373;lambdahat<-0.2136852;n<-length(sample)
LR<-replicate(1000,{
 x<-rexp(n,lambda0)
 exp(logLexp(lambda0,x))/
max(exp(logLexp(lambda0,x)),exp(logLexp(lambda1,x)))})
c<-quantile(LR,.05)
LR0<-exp(logLexp(lambda0,sample))/
max(exp(logLexp(lambda0,sample)),
exp(logLexp(lambda1,sample)))
if(LR0<c) paste("Reject H0") else paste("Do not
reject H0")
[1] "Do not reject H0"
```

Likelihood Ratio (LR) Test

Example: If the sample is given by

$$4.679781, 9.325474, 2.491352, 0.841366, 2.982121$$

Test

$$H_0 : X \sim exp(\lambda_0 = .373) \text{ vs } H_a : X \sim gamma(r = 4, \lambda_1 = .373)$$

Using the test statistic

$$LR(x) := \frac{L_{exp}(\lambda_0 = .373|x)}{L_{gamma}(r = 4, \lambda_1 = .373|x)}$$

```
sample<-c(4.679781, 9.325474, 2.491352, 0.841366,
2.982121)
```

Likelihood Ratio (LR) Test

```
logLexp<-function(lambda,x)
{length(x)*log(lambda)-sum(x)*lambda}
lambda0<-.373;r<-4;lambda1<-.373;n<-length(sample)
logLgamma<-function(r,lambda,x){
length(x)*(r*log(lambda)-log(factorial(r-1)))
+sum((r-1)*log(x)-x*lambda)}
LR<-replicate(1000,{
 x<-rexp(n,lambda0)
exp(logLexp(lambda0,x))/exp(logLgamma(r,lambda1,x))})
c<-quantile(LR,.05)
LR0<-exp(logLexp(lambda0,sample))/
exp(logLgamma(r,lambda1,sample))
if(LR0<c) paste("Reject H0") else paste("Do not
reject H0")
[1] "Do not reject H0"
```

End of Set 6