STAT 801 Lecture 11

Reading for Today's Lecture:
Goals of Today's Lecture:

- Introduce the likelihood, log-likelihood and score functions.

Likelihood Methods of Inference

Suppose you toss a coin 6 times and get Heads twice. If p is the probability of getting H then the probability of getting 2 heads is

$$15p^2(1-p)^4$$

This probability, thought of as a function of p, is the **likelihood** function for this particular data.

Definition: A **model** is a family $\{P_\theta; \theta \in \Theta\}$ of possible distributions for some random variable X. Typically the model is described by specifying $\{f_\theta(x); \theta \in \Theta\}$ the set of possible densities of X.

Definition: The likelihood function is the function L whose domain is Θ and whose values are given by

$$L(\theta) = f_\theta(X)$$

The key point is to think about how the density depends on θ not about how it depends on X. Notice that X, the observed value of the data, has been plugged into the formula for the density. Notice also that the coin tossing example is like this but with f being the discrete density. We use the likelihood in most of our inference problems:

1. Point estimation: we must compute an estimate $\hat\theta = \hat\theta(X)$ which lies in Θ. The **maximum likelihood estimate (MLE)** of θ is the value $\hat\theta$ which maximizes $L(\theta)$ over $\theta \in \Theta$ if such a $\hat\theta$ exists.

2. Point estimation of a function of θ: we must compute an estimate $\hat\phi = \hat\phi(X)$ of $\phi = g(\theta)$. We use $\hat\phi = g(\hat\theta)$ where $\hat\theta$ is the MLE of θ.

3. Interval (or set) estimation. We must compute a set $C = C(X)$ in Θ which we think will contain θ_0. We will use
$$\{\theta \in \Theta : L(\theta) > c\}$$
for a suitable c.

4. Hypothesis testing: We must decide whether or not $\theta_0 \in \Theta_0$ where $\Theta_0 \subset \Theta$. We base our decision on the likelihood ratio
$$\frac{\sup\{L(\theta); \theta \in \Theta_0\}}{\sup\{L(\theta); \theta \in \Theta \setminus \Theta_0\}}$$

Maximum Likelihood Estimation

To find an MLE we maximize L. This is a typical function maximization problem which we approach by setting the gradient of L equal to 0 and then checking to see that the root is a maximum, not a minimum or saddle point.

We begin by examining some likelihood plots in examples:

Cauchy Data

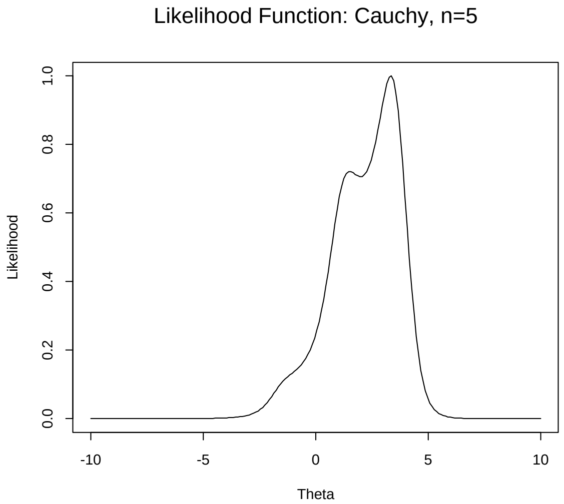
We have a sample $X_1, \ldots, X_n$ from the Cauchy(θ) density

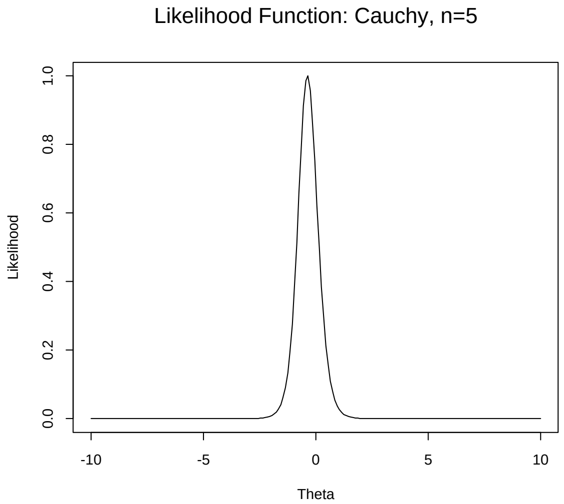
$$f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$$

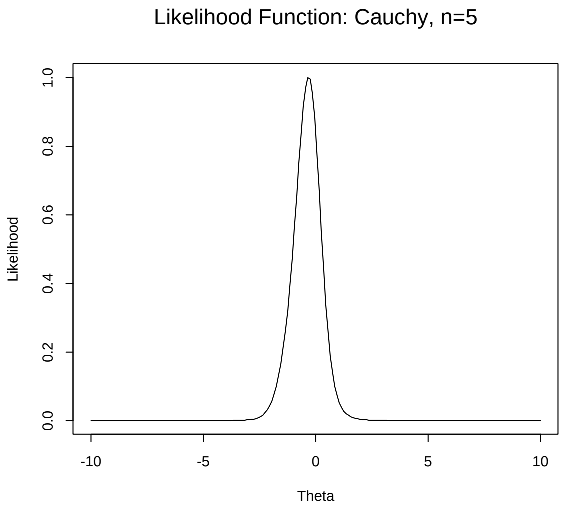
The likelihood function is

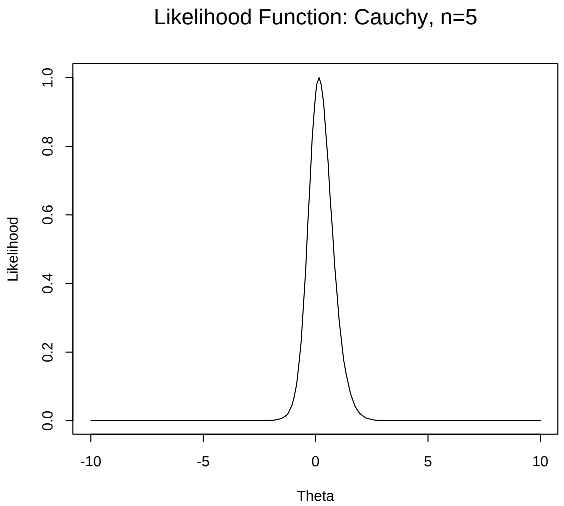
$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\pi(1+(X_i-\theta)^2)}$$

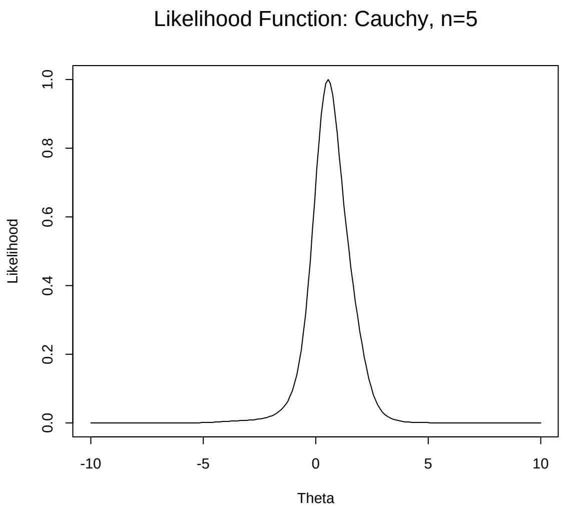
Here are some plots of this function for 6 samples of size 5.

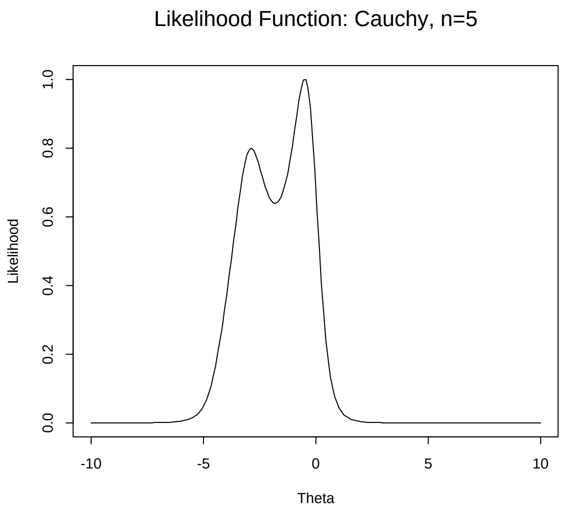




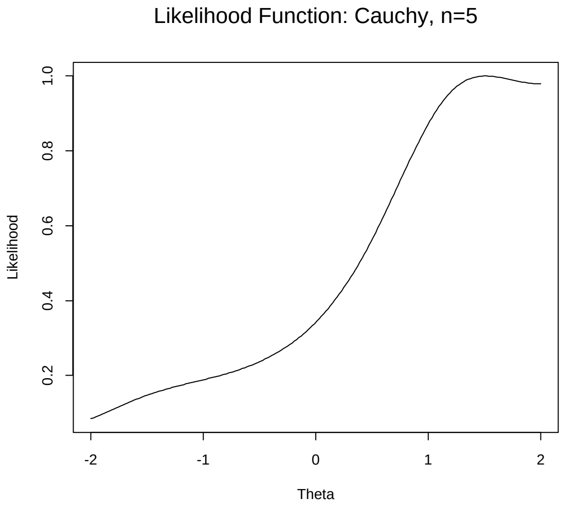


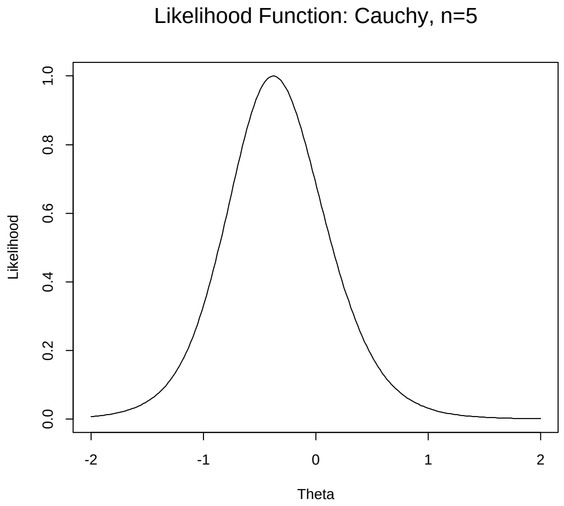


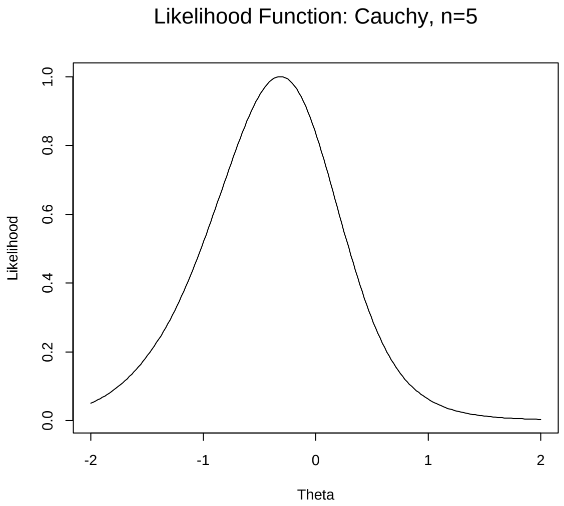


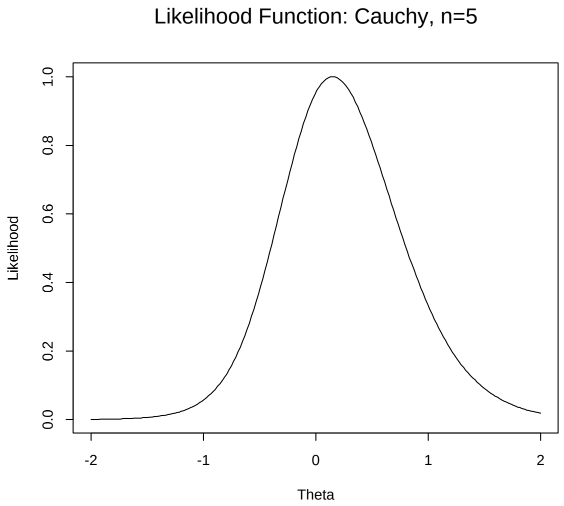


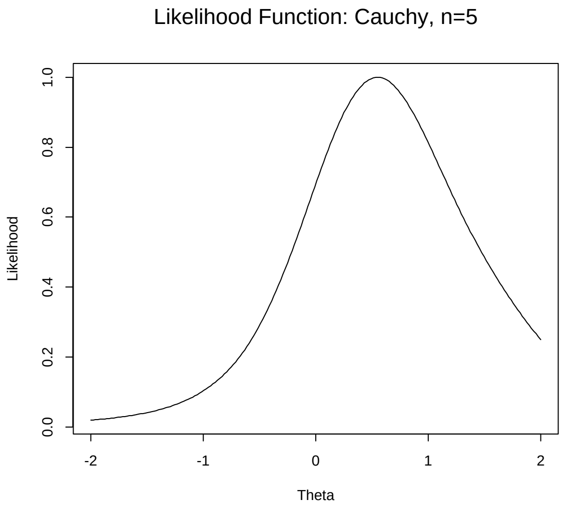
Here are close up views of these plots for θ between -2 and 2.

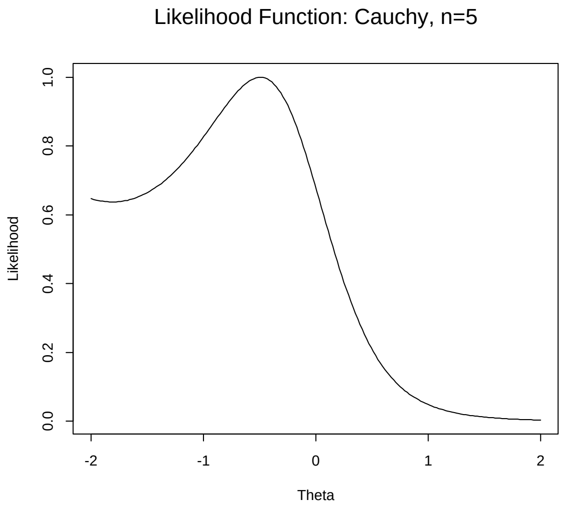




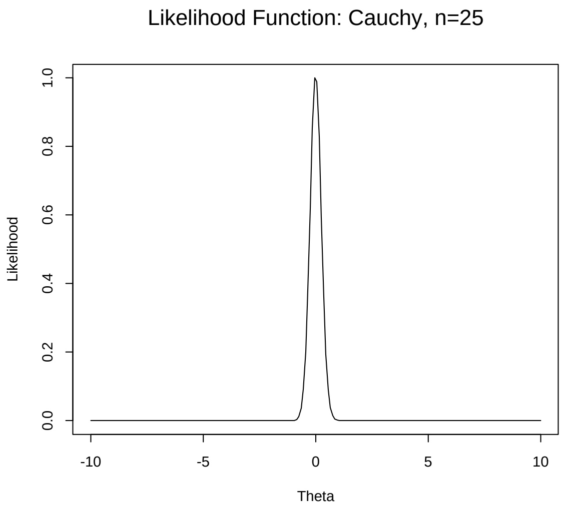


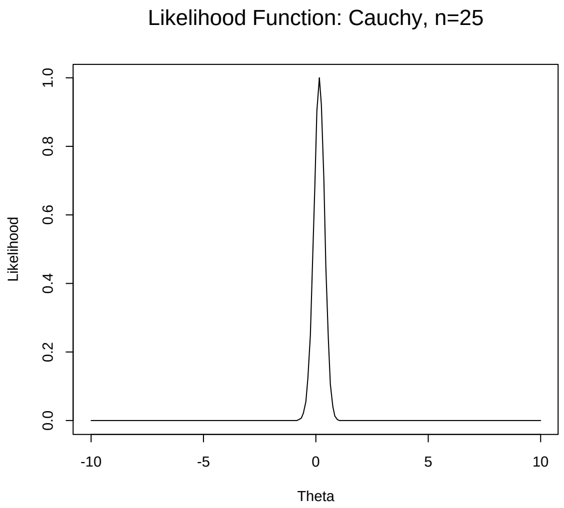


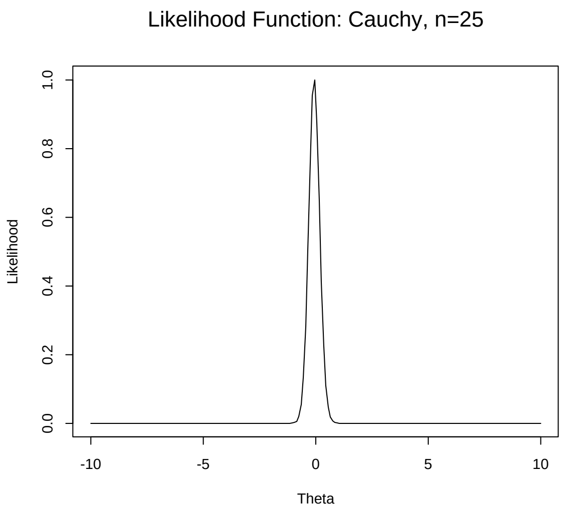


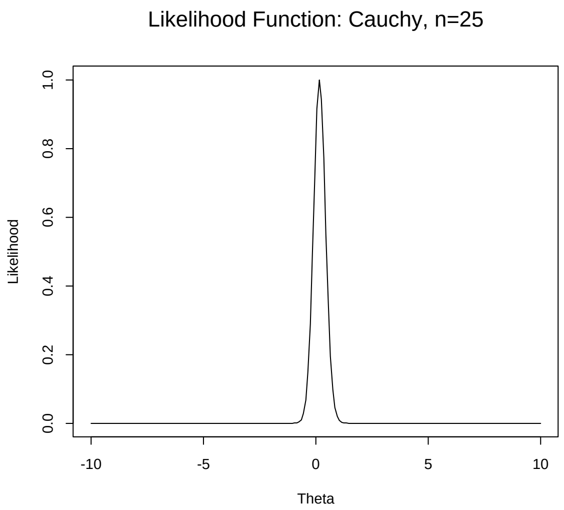


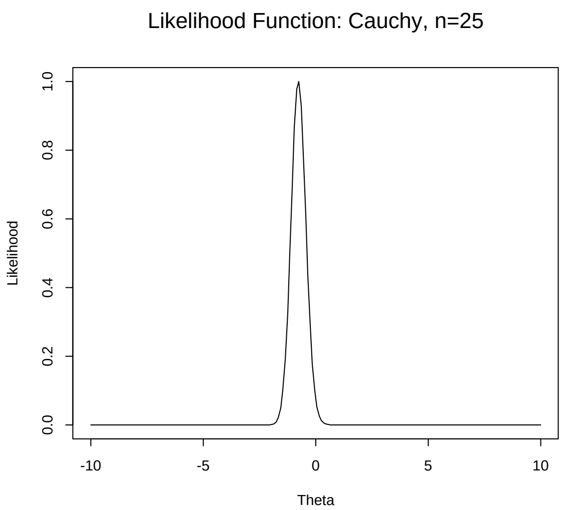
Now for sample size 25.

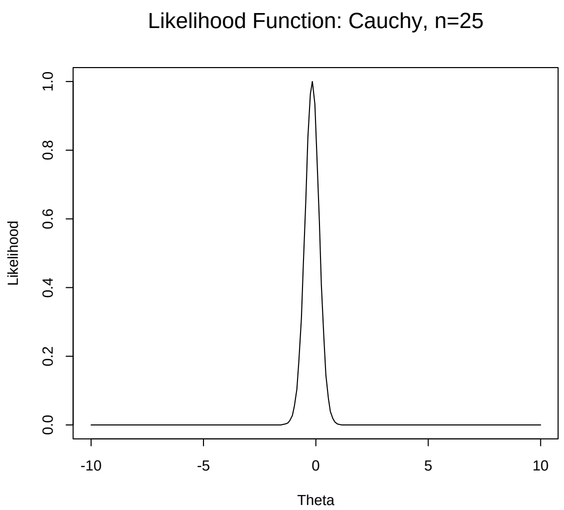




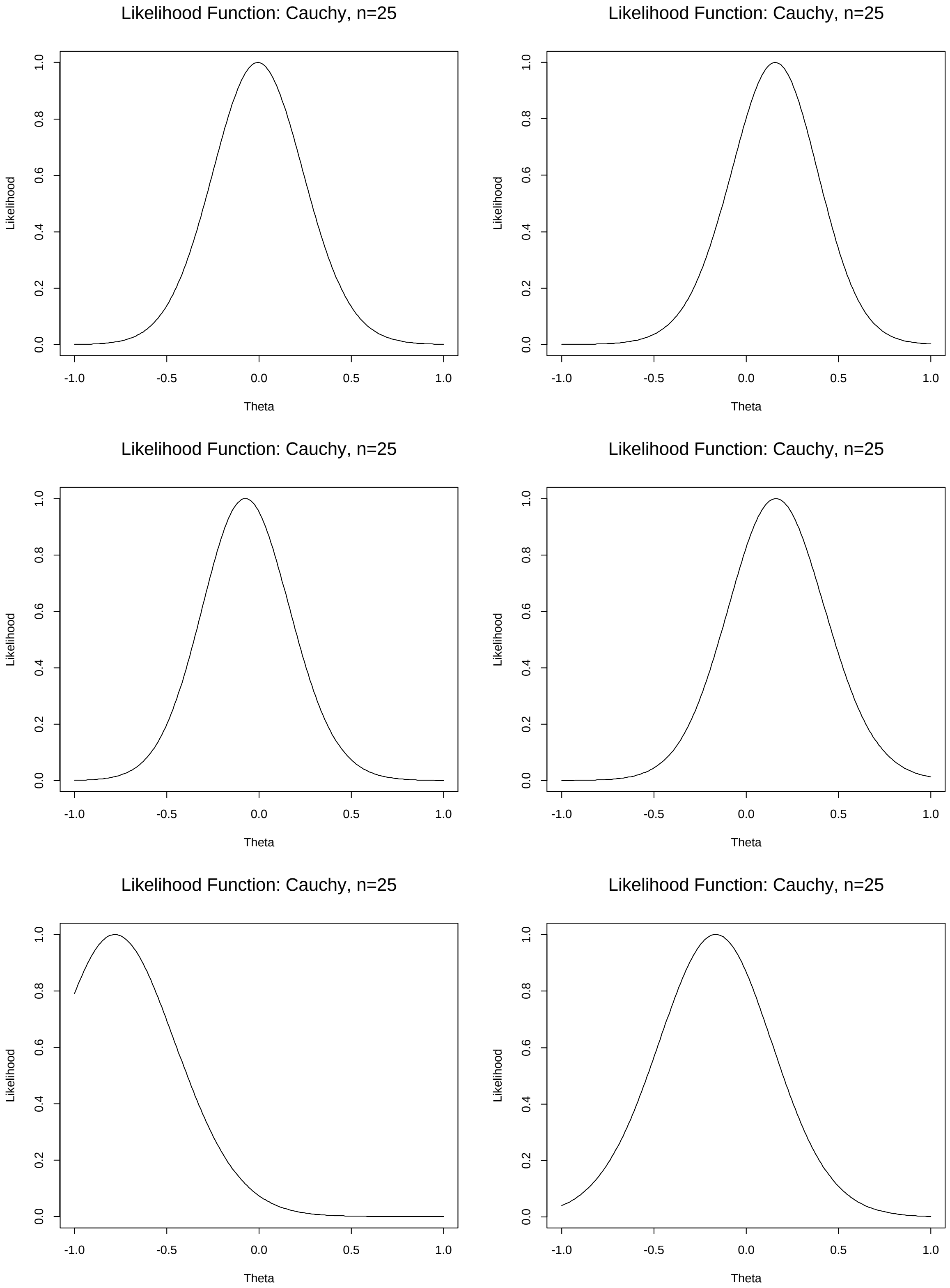








Here are close up views of these plots for θ between -2 and 2.



I want you to notice the following points:

- The likelihood functions have peaks near the true value of θ (which is 0 for the data sets I generated).
- The peaks are narrower for the larger sample size.
- The peaks have a more regular shape for the larger value of n.
- I actually plotted $L(\theta)/L(\hat\theta)$ which has exactly the same shape as L but runs from 0 to 1 on the vertical scale.

To maximize this likelihood we would have to differentiate L and set the result equal to 0. Notice that L is a product of n terms and the derivative will then be

$$\sum_{i=1}^{n} \prod_{j \neq i} \frac{1}{\pi(1+(X_j-\theta)^2)} \frac{2(X_i-\theta)}{\pi(1+(X_i-\theta)^2)^2}$$

which is quite unpleasant. It is much easier to work with the logarithm of L since the log of a product is a sum and the logarithm is monotone increasing.

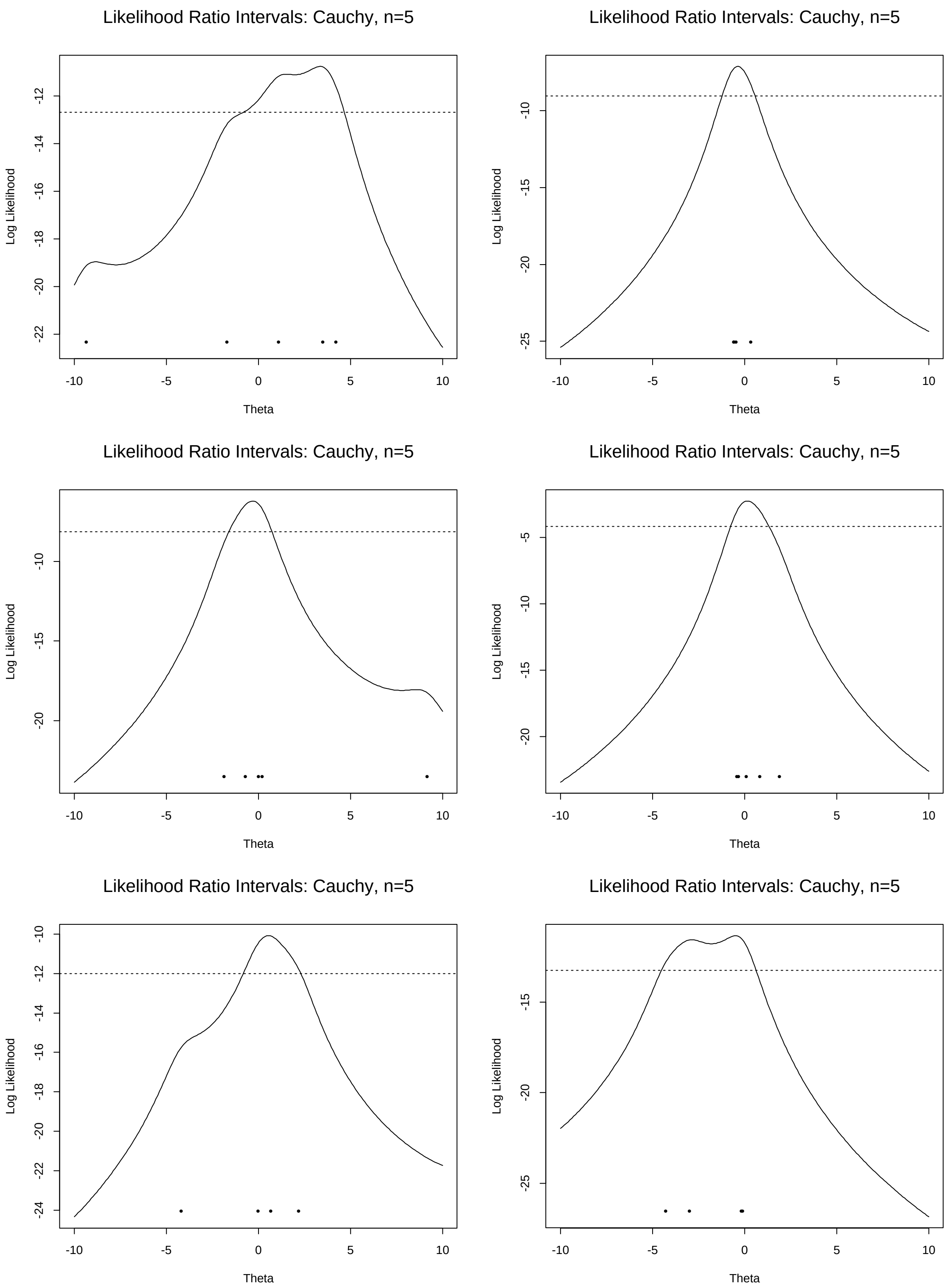
Definition: The **Log Likelihood** function is

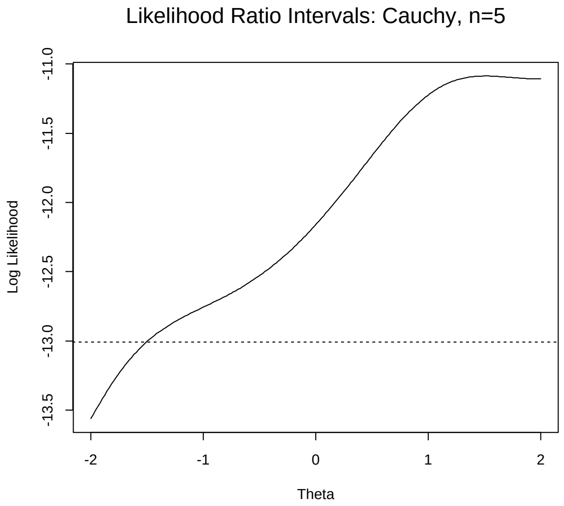
$$\ell(\theta) = \log(L(\theta)) .$$

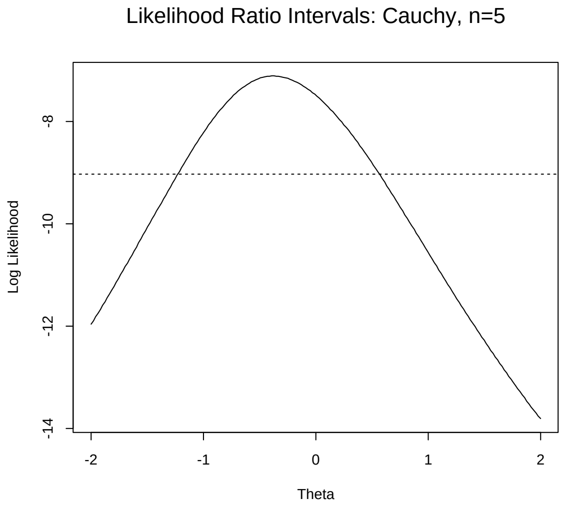
For the Cauchy problem we have

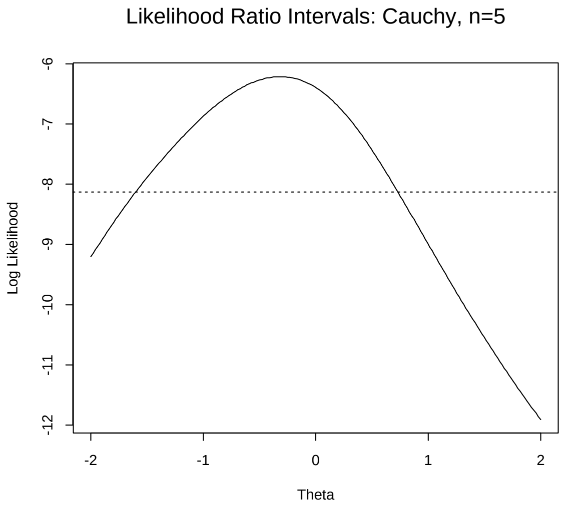
$$\ell(\theta) = -\sum \log(1+(X_i-\theta)^2) - n\log(\pi)$$

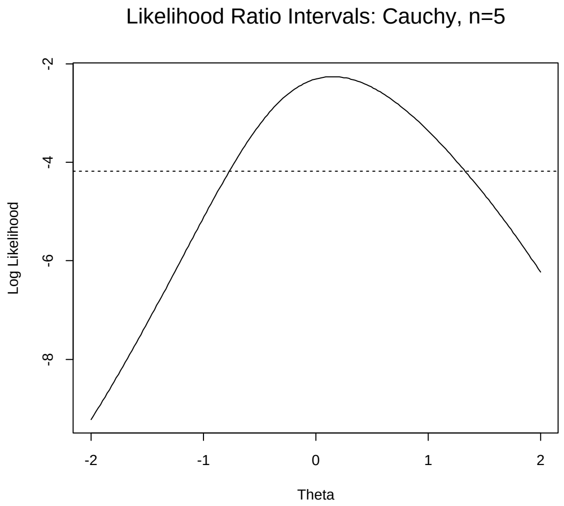
Here are the logarithms of the likelihoods plotted above:

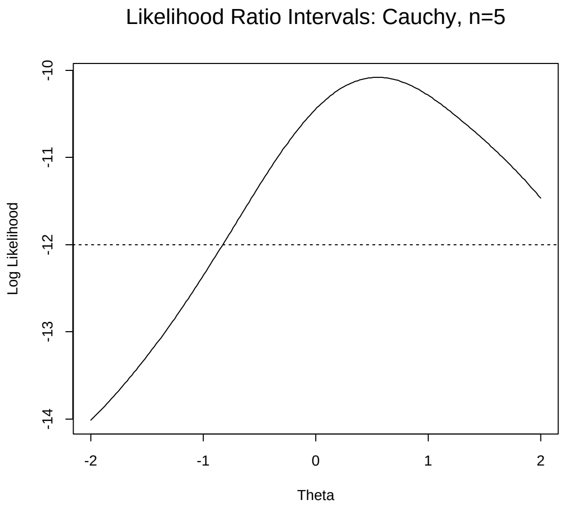

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta

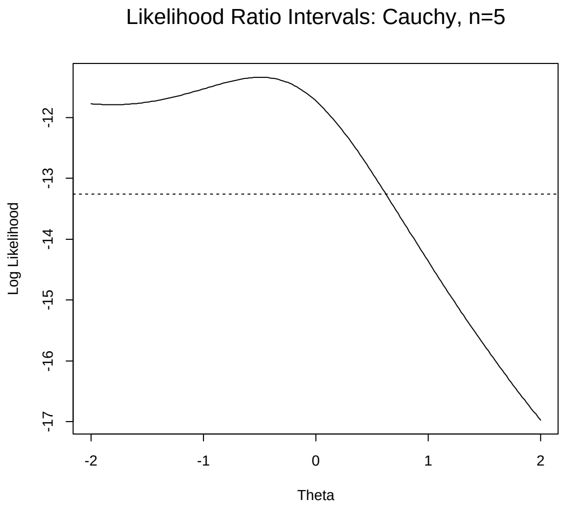

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta

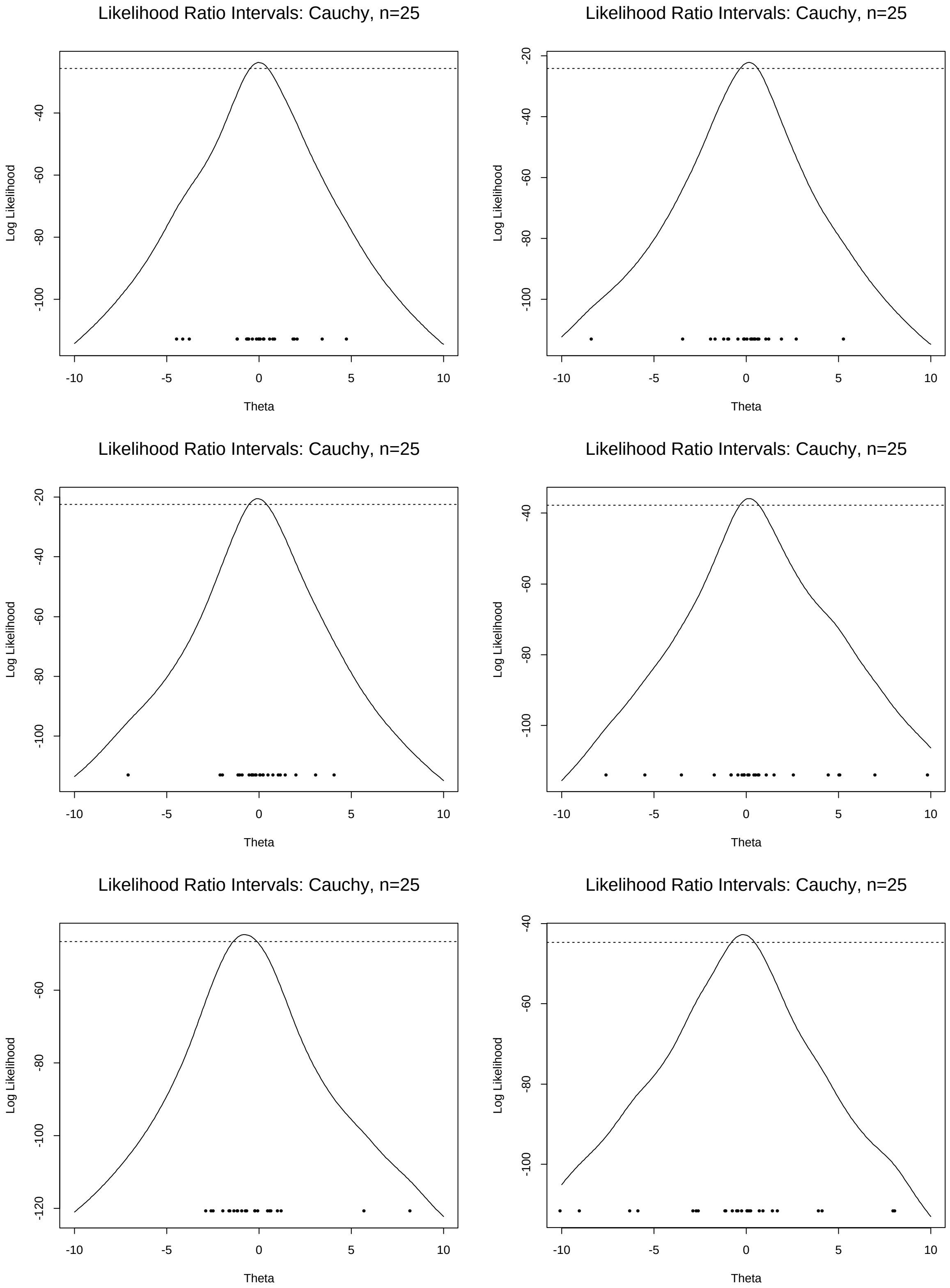

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta

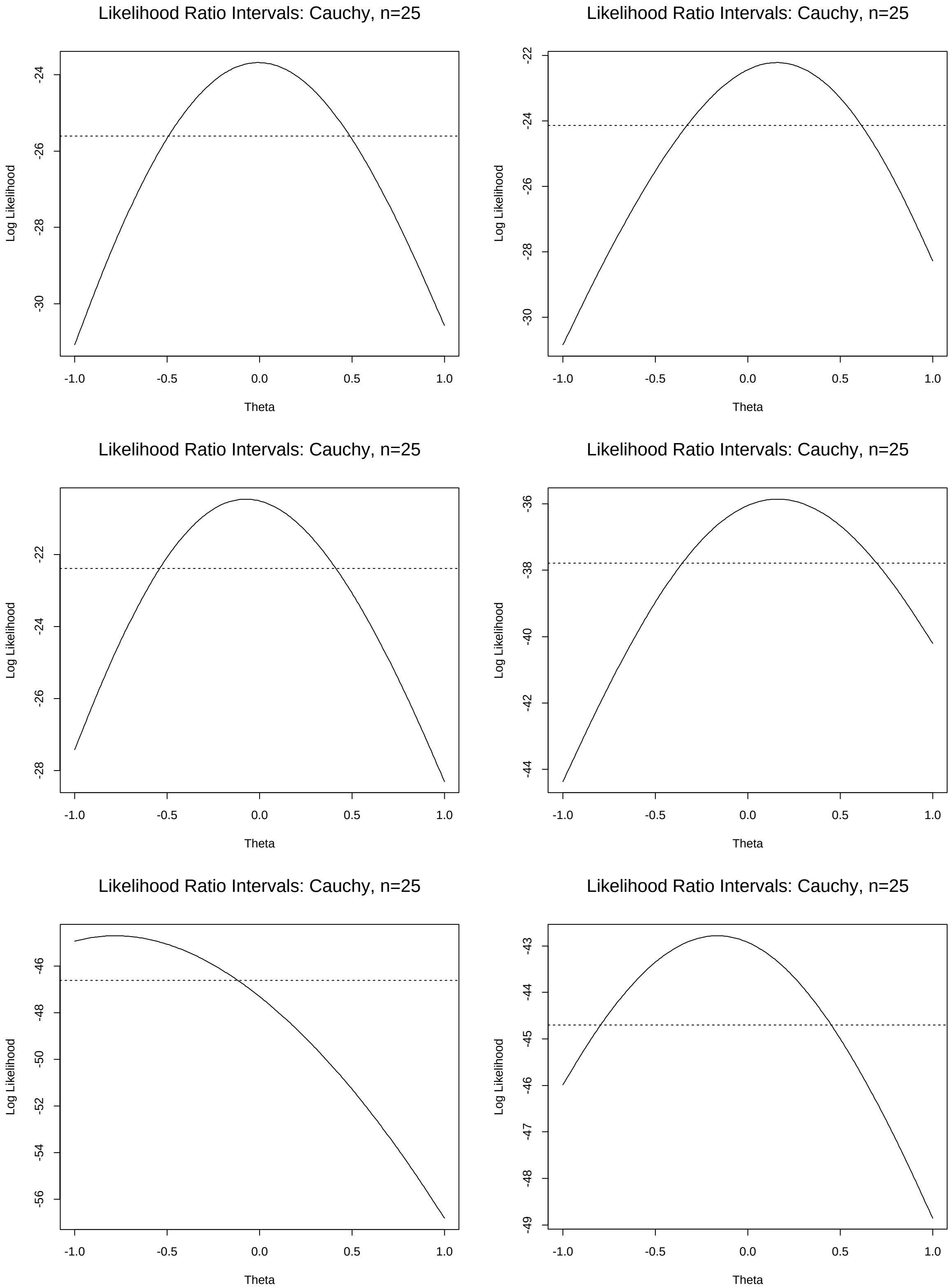

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
Theta


Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
Theta
-10
-5
0
5
10
-40
-60
-80
-100
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
Theta
-20
-40
-60
-80
-100
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
Theta
-20
-40
-60
-80
-100
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
Theta
-40
-60
-80
-100
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
Theta
-60
-80
-100
-120
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
Theta
-40
-60
-80
-100



I want you to notice the following points:

- The log likelihood functions with $n = 25$ have pretty smooth shapes which look rather parabolic.
- For $n = 5$ there are plenty of local maxima and minima of ℓ.

You can see that the likelihood will tend to 0 as $|\theta| \to \infty$ so that the maximum of ℓ will occur at a root of ℓ', the derivative of ℓ with respect to θ.

Definition: The **Score Function** is the gradient of ℓ

$$U(\theta) = \frac{\partial \ell}{\partial \theta}$$

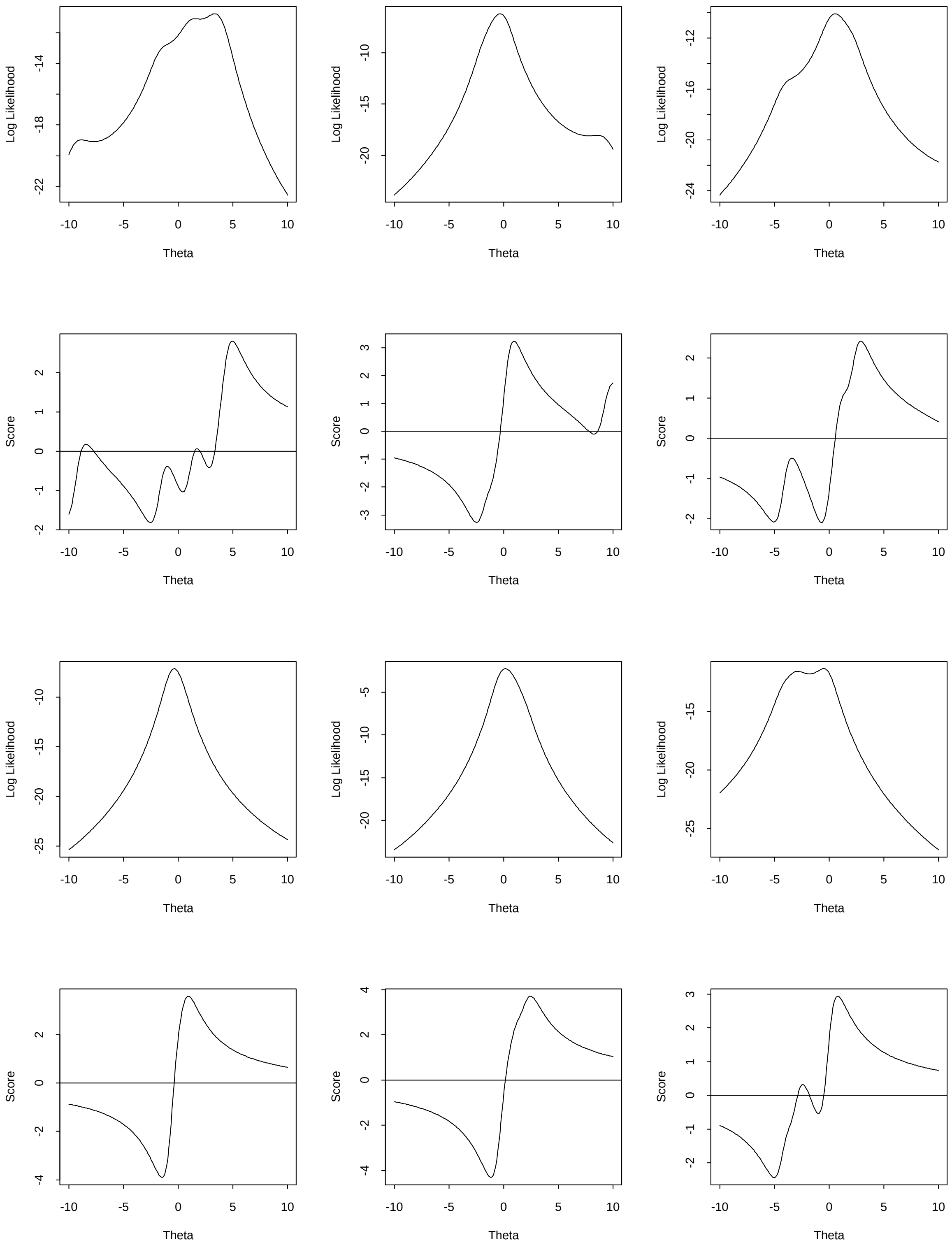
The MLE $\hat{\theta}$ usually solves the **Likelihood Equations**

$$U(\theta) = 0$$

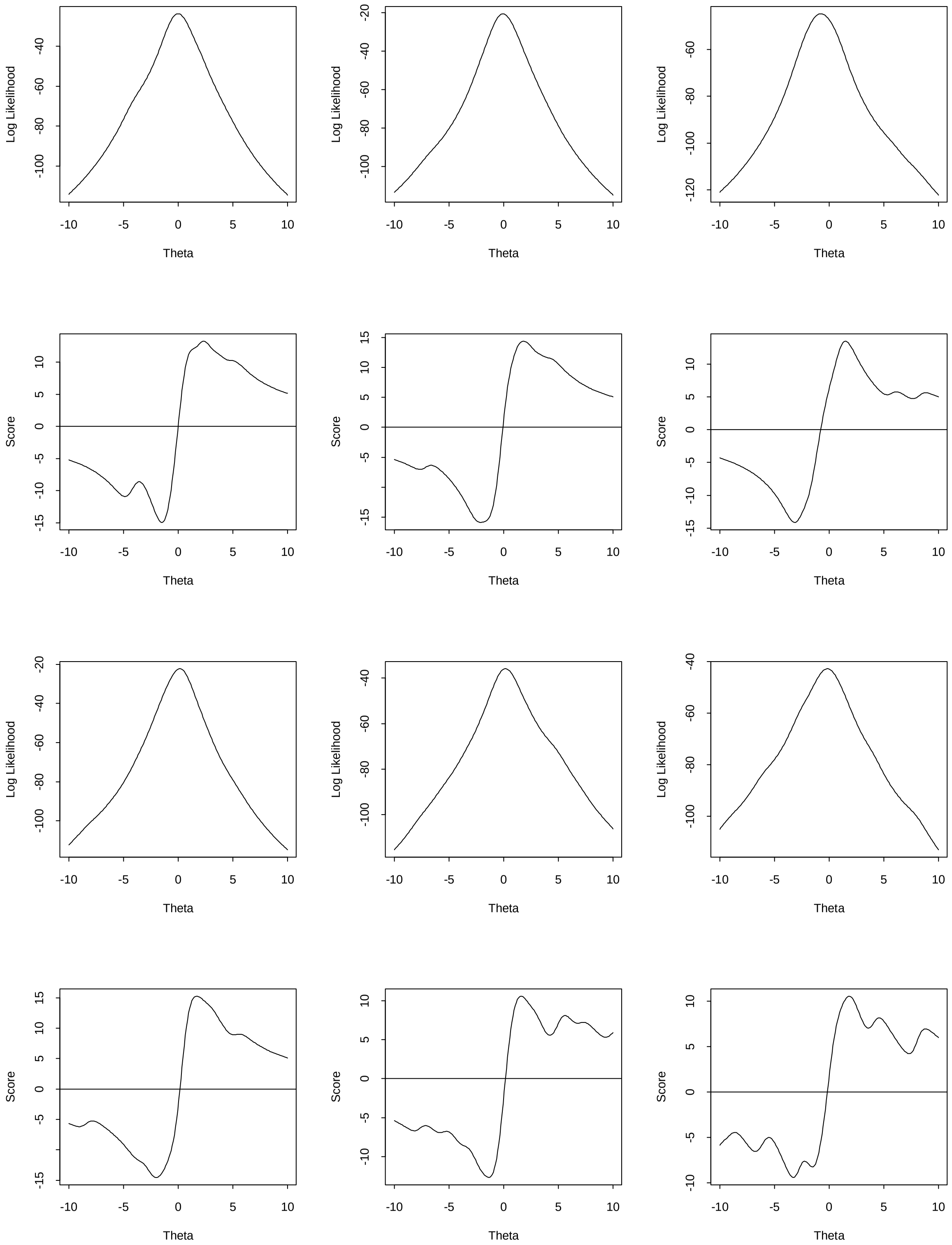
In our Cauchy example we find

$$U(\theta) = \sum \frac{2(X_i - \theta)}{1 + (X_i - \theta)^2}$$

Here are some plots of the score functions for $n = 5$ for our Cauchy data sets. Each score is plotted beneath a plot of the corresponding ℓ.



Notice that there are often multiple roots of the likelihood equations. Here is $n = 25$:



The Binomial Distribution

If X has a Binomial(n,θ) distribution then

$$L(\theta) = \binom{n}{X}\theta^X(1-\theta)^{n-X}$$

$$\ell(\theta) = \log\binom{n}{X} + X\log(\theta) + (n-X)\log(1-\theta)$$

$$U(\theta) = \frac{X}{\theta} - \frac{n-X}{1-\theta}$$

The function L is 0 at $\theta = 0$ and at $\theta = 1$ unless $X = 0$ or $X = n$ so for $1 \le X \le n$ the MLE must be found by setting $U = 0$ and getting

$$\hat\theta = \frac{X}{n}$$

For $X = n$ the log-likelihood has derivative

$$U(\theta) = \frac{n}{\theta} > 0$$

for all θ so that the likelihood is an increasing function of θ which is maximized at $\hat\theta = 1 = X/n$. Similarly when $X = 0$ the maximum is at $\hat\theta = 0 = X/n$.

The Normal Distribution

Now we have $X_1, \ldots, X_n$ iid $N(\mu,\sigma^2)$. There are two parameters $\theta = (\mu,\sigma)$. We find

$$L(\mu,\sigma) = (2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i-\mu)^2/(2\sigma^2)\}$$

$$\ell(\mu,\sigma) = -n\log(2\pi)/2 - \frac{\sum(X_i-mu)^2}{2\sigma^2}$$

$$U(\mu,\sigma) = \left[\begin{array}{c} \frac{\sum(X_i-\mu)}{\sigma^2} \\ \frac{\sum(X_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma} \end{array}\right]$$

Notice that U is a function with two components because θ has two components.

Setting the likelihood equal to 0 and solving gives

$$\hat\mu = \bar X$$

and

$$\hat\sigma = \sqrt{\frac{\sum(X_i-\bar X)^2}{n}}$$

You need to check that this is actually a maximum. To do so you compute one more derivative. The matrix H of second derivatives of ℓ is

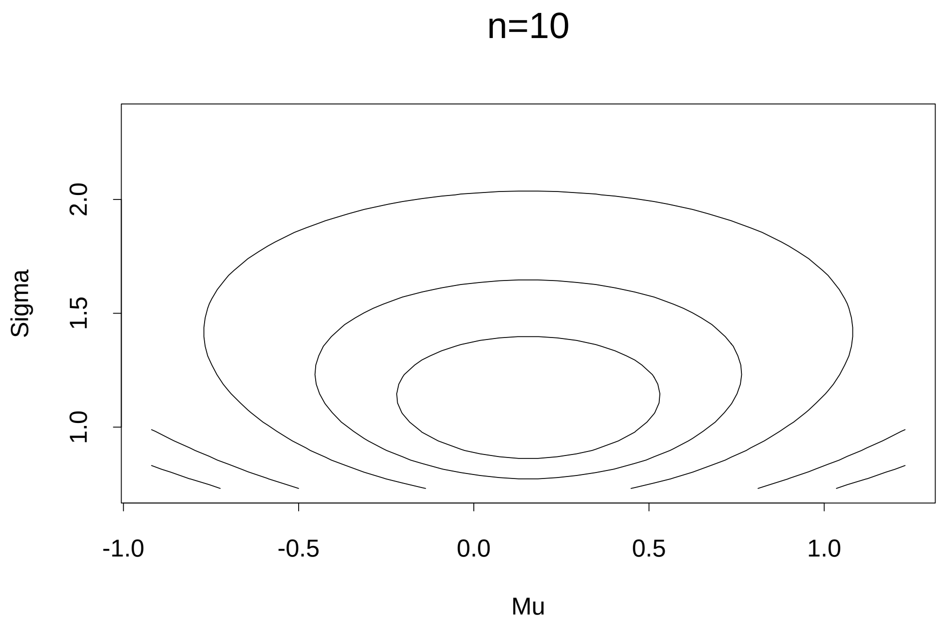
$$\left[\begin{array}{cc} \frac{-n}{\sigma^2} & \frac{-2\sum(X_i-\mu)}{\sigma^3} \\ \frac{-2\sum(X_i-\mu)}{\sigma^3} & \frac{-3\sum(X_i-\mu)^2}{\sigma^4} + \frac{n}{\sigma^2} \end{array}\right]$$

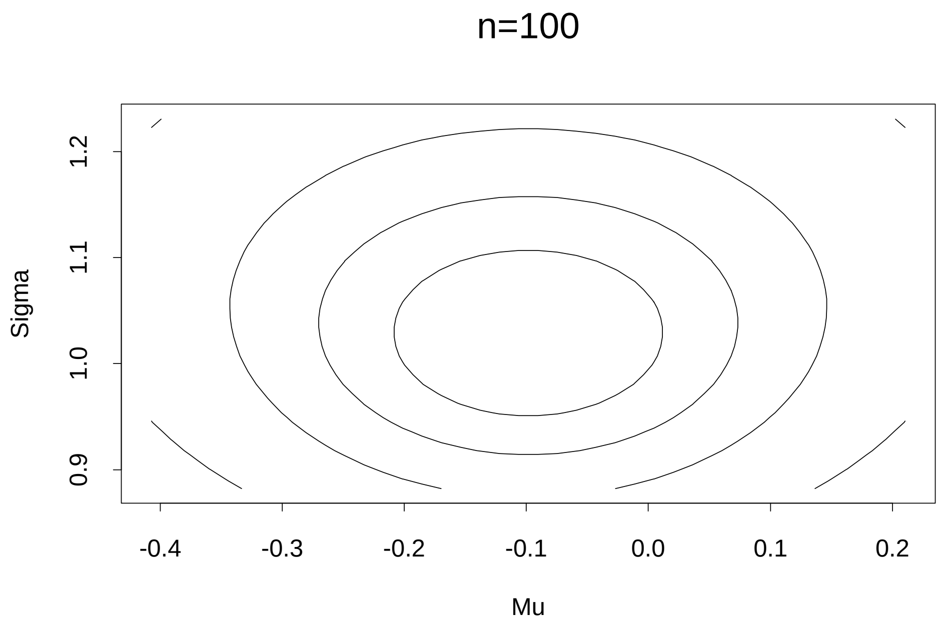
Plugging in the mle gives

$$H(\hat\theta) = \left[\begin{array}{cc} \frac{-n}{\hat\sigma^2} & 0 \\ 0 & \frac{-2n}{\hat\sigma^2} \end{array}\right]$$

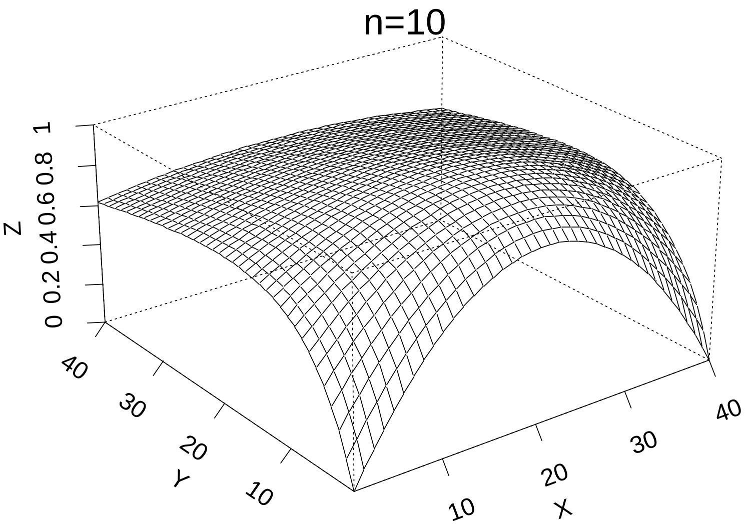
This matrix is negative definite. Both its eigenvalues are negative. So $\hat{\theta}$ must be a local maximum.

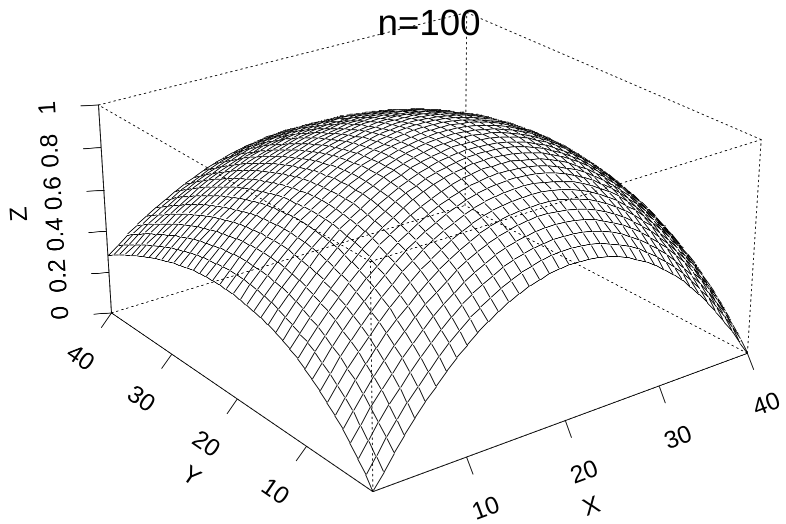
Here is a contour plot of the normal log likelihood for two data sets with $n = 10$ and $n = 100$.





Here are perspective plots of the same.





Notice that the contours are quite ellipsoidal for the larger sample size.

We now turn to theory to explain the features of these plots, at least approximately in large samples.