Yesterday we…

- Discussed the theoretical guarantees of MLE
  - Why do we prefer MLE to other estimators?

- Fitted a logistic regression via MLE

- Introduced Likelihood-Ratio test
  - hypothesis testing for nested models

Today

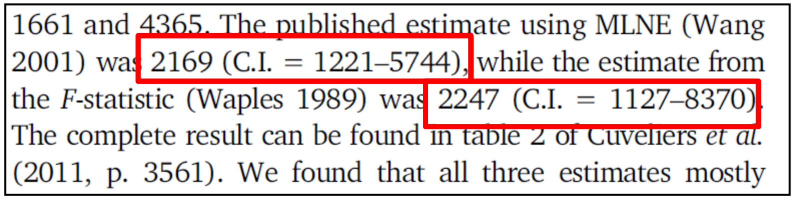
- Interval estimation
  - Confidence interval
  - Joint Confidence region
  - Profile likelihood
  - Normal approximation

Confidence interval estimation

- We are now able to find ML estimates by maximising the log-likelihood function

- Usually confidence intervals (C.I.) are also required while quoting them

- There are many ways to calculate C.I., some of which can be directly obtained from the log-likelihood function



1661 and 4365. The published estimate using MLNE (Wang 2001) was 2169 (C.I. = 1221–5744), while the estimate from the *F*-statistic (Waples 1989) was 2247 (C.I. = 1127–8370). The complete result can be found in table 2 of Cuveliers *et al.* (2011, p. 3561). We found that all three estimates mostly

- Say, in the coin tossing example, We can perform a likelihood-ratio test to test for $H_0\text{: } p = 0.5$ (fair coin hypothesis)
- $n = 10, y = 7$

- For the simplified model M1, there is no free parameter as p is fixed to $0.5. l(0.5) = -2.14398$
- The full model M2 has one free parameter p, $0 \leq p \leq 1$, and l is maximised at $p = 0.7. l(0.7) = -1.32115$

- The critical value of this test is $\chi^2_{0.95, df=1} = 3.84$
- $D = 2 * [-1.32115 - (-2.14398)] = 1.64566 < 3.84$

- According to LRT, $H_0\text{: } p = 0.5$ is not rejected

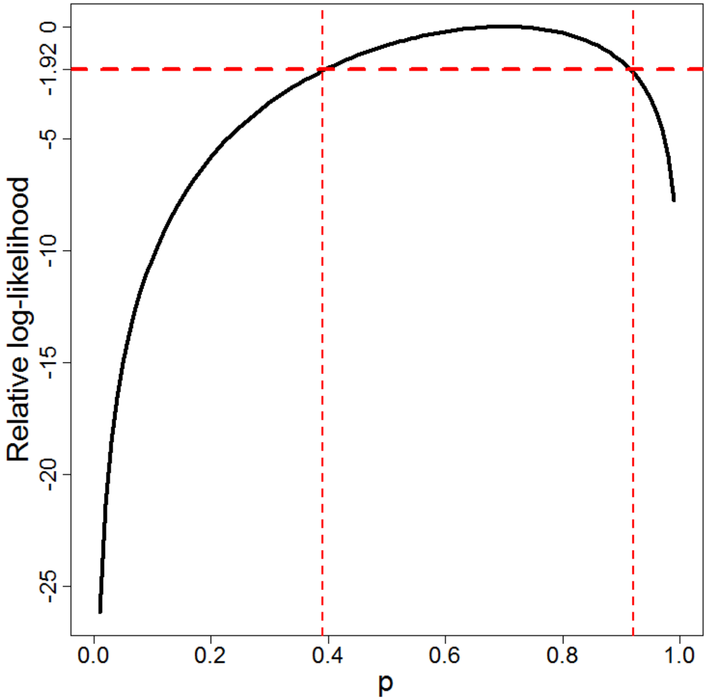
C.I example: coin tossing

- Instead of performing a LRT, or multiple LRT against different values of p, we can find a range of p, such that D remains within the “acceptance region”.

- In the same coin tossing example the maximised log-likelihood is $l(0.7) = -1.32115$
- The critical value for LRT is $\chi^2_{0.95,df=1} = 3.84$
- $D = 2 * (\ln(L2) - \ln(L1)) < 3.84$
- $\ln(L2) - \ln(L1) < 1.92$
- For a given $\tilde{p}$, as long as $l(\tilde{p}) > -1.32115 - 1.92 = -3.24$ we will not reject $H_0 : p = \tilde{p}$

- In most cases, if we want to find the 95% C.I. for a single parameter, we look at the range of parameter values such that the log-likelihood is within 1.92 units from its maximum

- Rule of thumb: -1.92, or -2



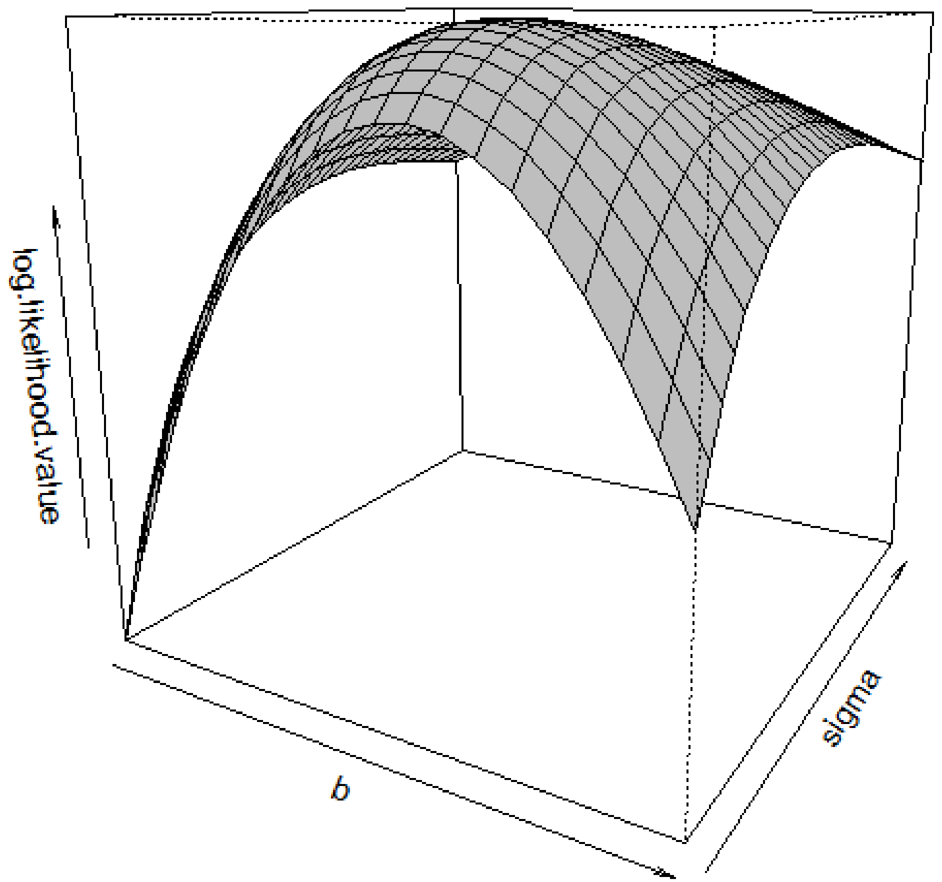
- If we observe 7 heads out of 10 tosses, the 95% C.I. for p is [0.39, 0.92].
- Since 0.5 lies inside the 95% C.I., we do not reject the “fair coin” hypothesis.

C.I. example: linear regression

- Back to our rabbit example, M1 has two parameters: b, σ
- For each pair of $\{b, \sigma\}$, there is an associated log-likelihood value
- Bivariate function → 3D plot

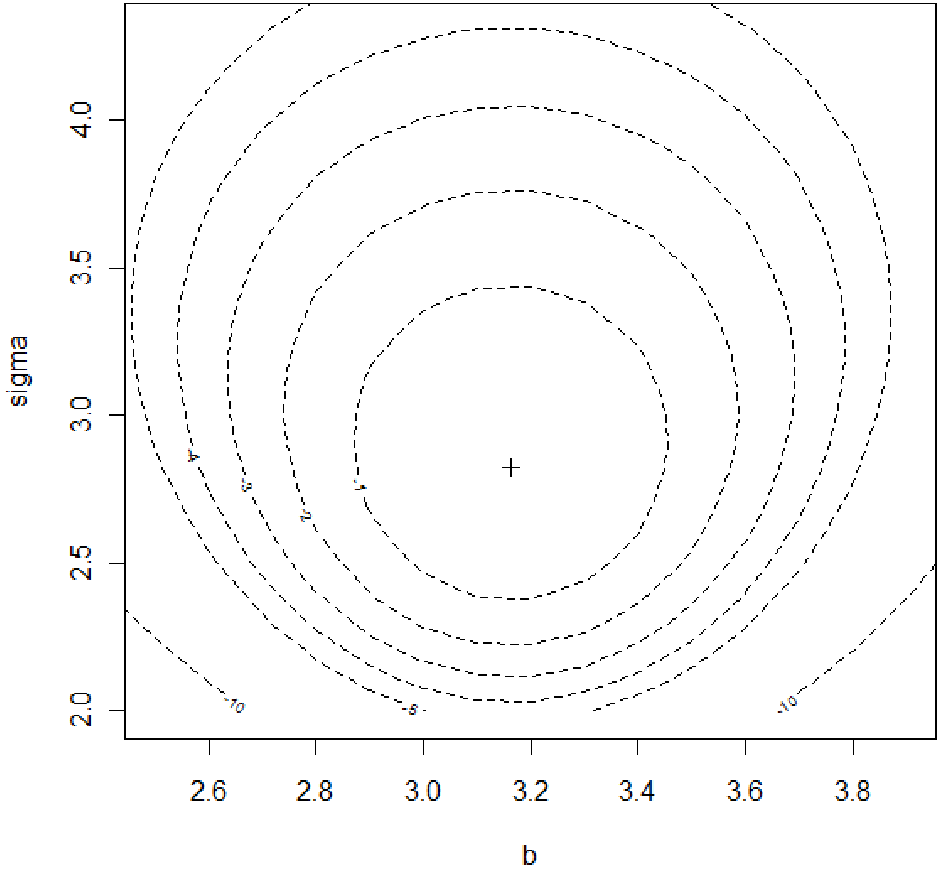
- 3D plot in R using `persp()`

```
# DEFINE THE RANGE OF PARAMETERS TO BE PLOTTED
b<-seq(2, 4, 0.1)
sigma<-seq(2, 5, 0.1)

# THE LOG-LIKELIHOOD VALUE IS STORED IN A MATRIX
log.likelihood.value<-matrix(nr=length(b), nc=length(sigma))

# COMPUTE THE LOG-LIKELIHOOD VALUE FOR EACH PAIR OF PARAMETERS
for (i in 1:length(b))
  {
   for (j in 1:length(sigma))
    {
    log.likelihood.value[i,j]<-
    regression.no.intercept.log.likelihood(parm=c(b[i],sigma[j]),
    dat=recapture.data)
    }
  }
# WE ARE INTERESTED IN KNOWING THE RELATIVE LOG-LIKELIHOOD VALUE
# RELATIVE TO THE PEAK (MAXIMUM)
rel.log.likelihood.value<-log.likelihood.value-M1$value

# FUNCTION FOR 3D PLOT
persp(b, sigma, rel.log.likelihood.value, theta=30, phi=20,
        xlab='b', ylab='sigma', zlab=‘rel.log.likelihood.value',
        col='grey')
```


log.likelihood.value
b
sigma

How about a contour plot?

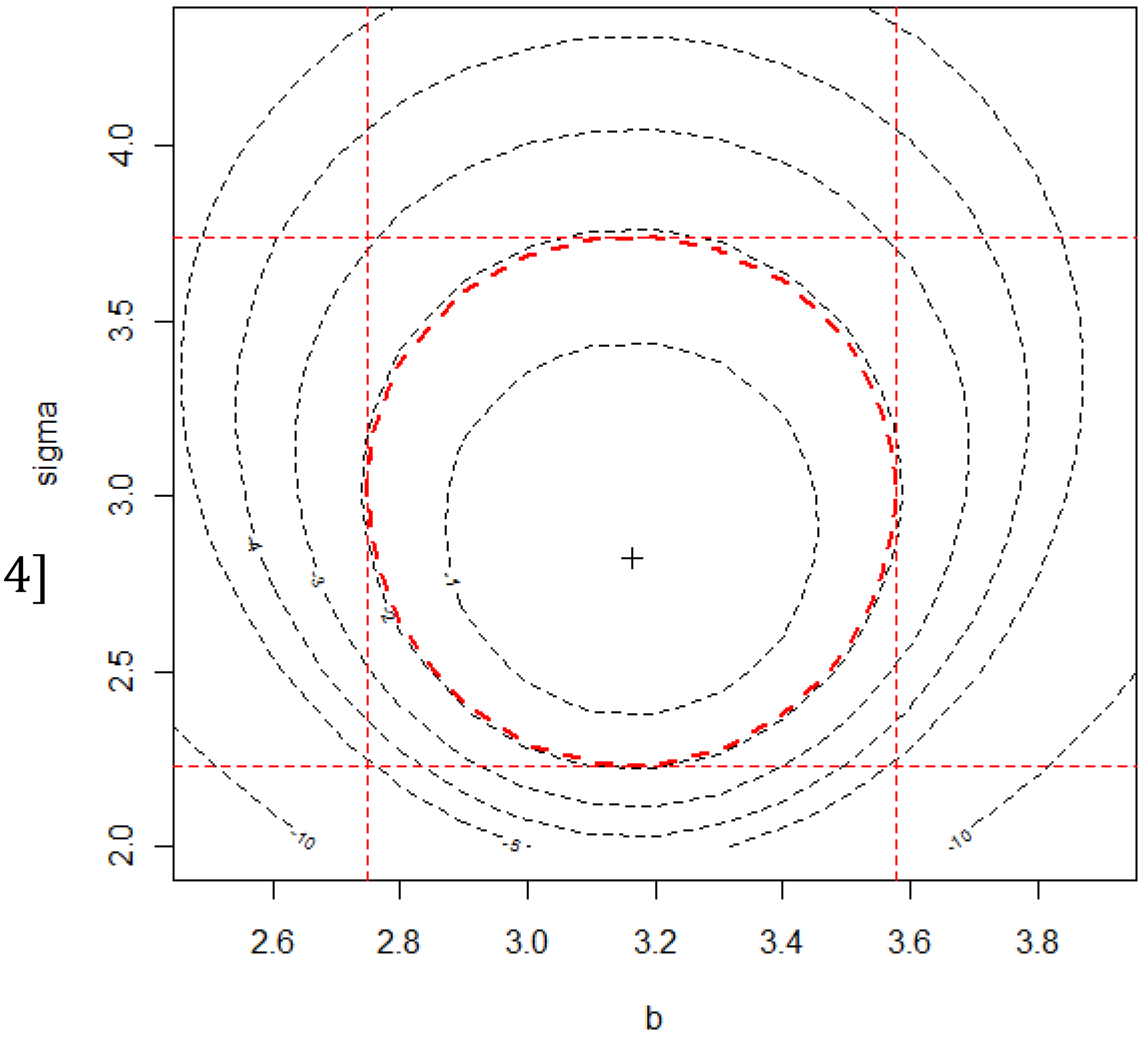
```
# CONTOUR PLOT
contour(b, sigma, rel.log.likelihood.value, xlab='b',
ylab='sigma',
      xlim=c(2.5, 3.9), ylim=c(2.0, 4.3),
      levels=c(-1:-5, -10), cex=2)
# DRAW A CROSS TO INDICATE THE MAXIMUM
points(M1$par[1], M1$par[2], pch=3)
```



We can, again, draw the -1.92 line (circle) on the contour map

```
contour.line<-contourLines(b, sigma,
rel.log.likelihood.value,  levels=-1.92)[[1]]
lines(contour.line$x, contour.line$y, col='red',
      lty=2, lwd=2)
```

IF WE LOOK AT ONE PARAMETER AT A TIME



95% C.I. for σ is $[2.23, 3.74]$

95% C.I. for b is $[2.75, 3.57]$

Joint confidence interval (region)

- We know the 95% C.I. for b, and the 95% C.I. for σ

- But it does not mean we know the joint 95% confidence region for the pair (b, σ)

- We need to consider the correlation between the two ML estimators

- Multiple comparisons?

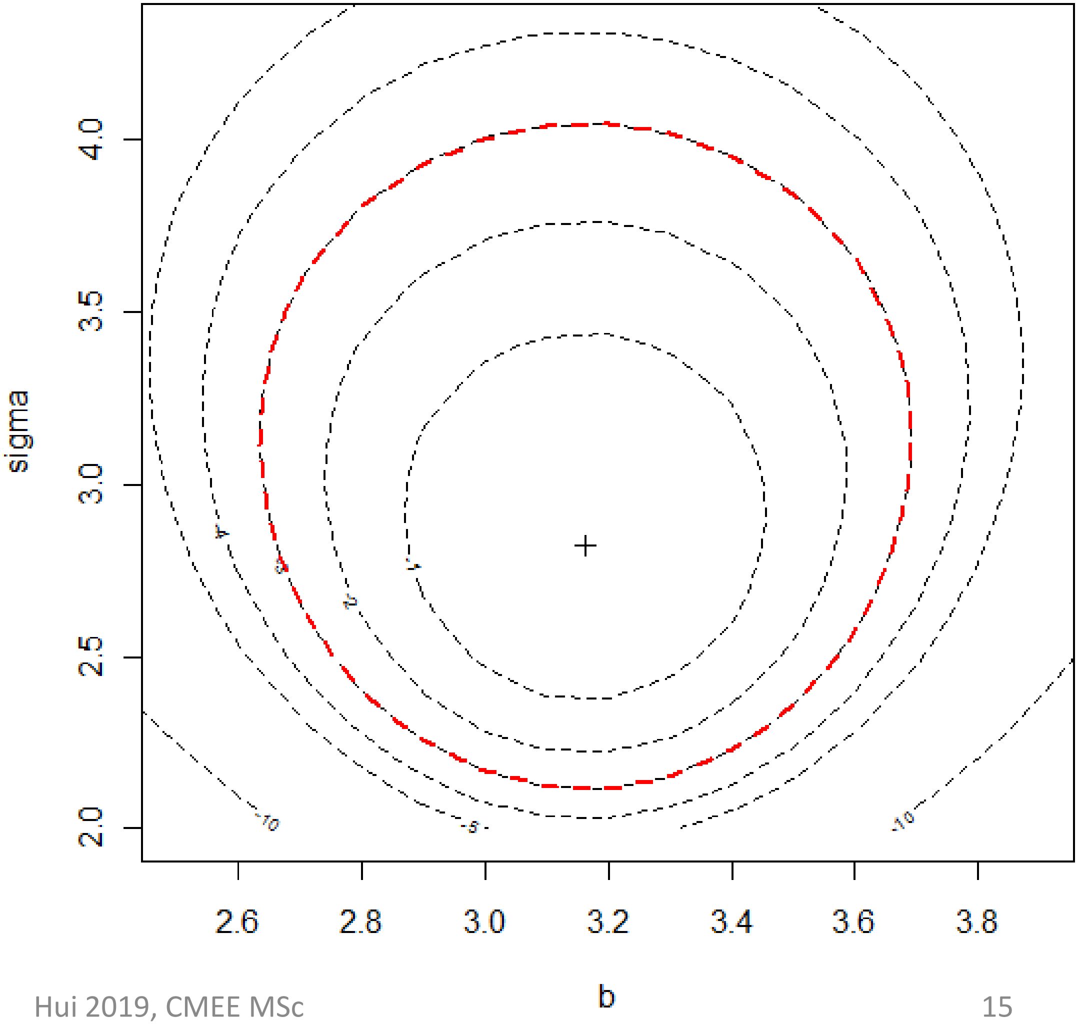
Joint confidence interval (region)

- The general rule: the 95% joint C.I. (region) for k parameters is the collection of parameter values for which the log-likelihood decreases by no more than half of $\chi^2_{0.95,df=k}$ from its maximum.

- 95% C.I. for one parameter: $0.5 * \chi^2_{0.95,df=1} = 1.92$
- Joint 95% C.I. for two parameters: $0.5 * \chi^2_{0.95,df=2} = 2.99$

- On the contour plot, we can circle the region where the log-likelihood value is 2.99 units below the maximum.

The joint 95% confidence region for (b, σ) are all the points within the red dotted circle



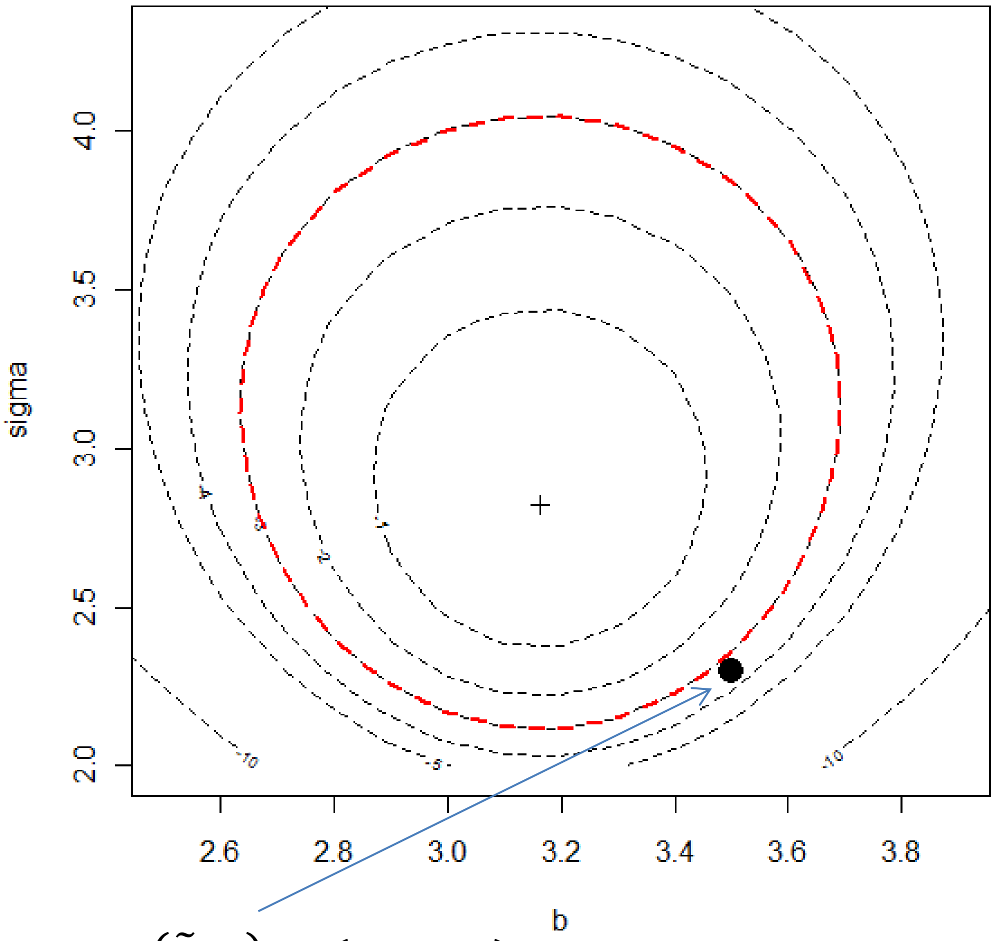
- Consider a set of parameter values $(\tilde{b}, \tilde{\sigma}) = (3.5, 2.3)$

$\tilde{b} = 3.5$ alone is within the 95% C.I. for b

$\tilde{\sigma} = 2.3$ alone is also within the 95% C.I. for σ

But it is possible for the pair $(\tilde{b}, \tilde{\sigma}) = (3.5, 2.3)$ to lie outside the joint 95% C.I.

Multiple comparisons!



$(\tilde{b}, \tilde{\sigma}) = (3.5, 2.3)$

Profile likelihood

- We wish to focus on a subset of parameter(s)
- Partition the parameters into two subsets: $\underline{\theta} = (\underline{\theta_1}, \underline{\theta_2})$, and aim to obtain the C.I. for $\underline{\theta_1}$ only
- We can perform profiling, partial maximisation of the original log-likelihood along $\underline{\theta_1}$
- $l^*\left(\underline{\theta_1}\right) = \max_{\underline{\theta_2}} l(\underline{\theta_1}, \underline{\theta_2}\,; \underline{x})$
  - Fix $\underline{\widetilde{\theta_1}}$, then vary $\underline{\theta_2}$ such that the log-likelihood is (partially) maximised
  - Record down the maximised log-likelihood, and this is your $l^*(\underline{\widetilde{\theta_1}})$
  - Repeat this for a range of of $\underline{\widetilde{\theta_1}}$, then you get the profile log-likelihood function for $\underline{\theta_1}$
- The LRT statistic (and also C.I.) can be calculated using this profile log-likelihood

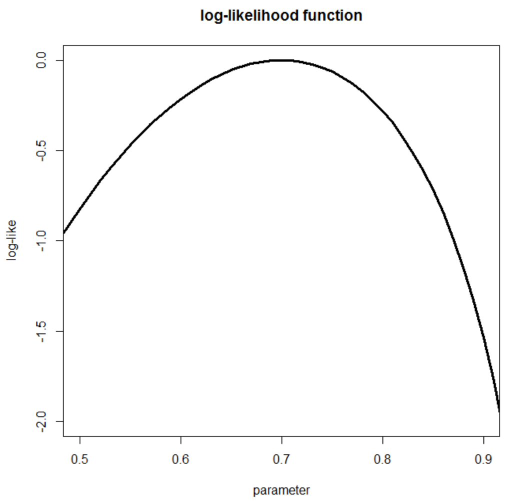
CI: Approximate normality of MLE

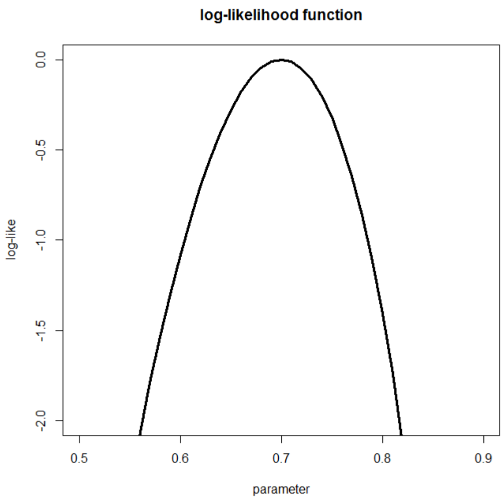
- One key property of ML estimators is asymptotic normality for reasonably large n

- For one-parameter case θ, the 95% C.I. for θ is approximately $\hat{\theta} \pm 1.96\sqrt{var(\hat{\theta})}$, where the magical number 1.96 comes from the 2.5- and 97.5-percentile of a standard normal distribution

- But what is $var(\hat{\theta})$?

- The curvature of the log-likelihood function





The rate of change of slope at the peak!

- Look at the two log-likelihood curves above: the right one is "steeper" around its peak

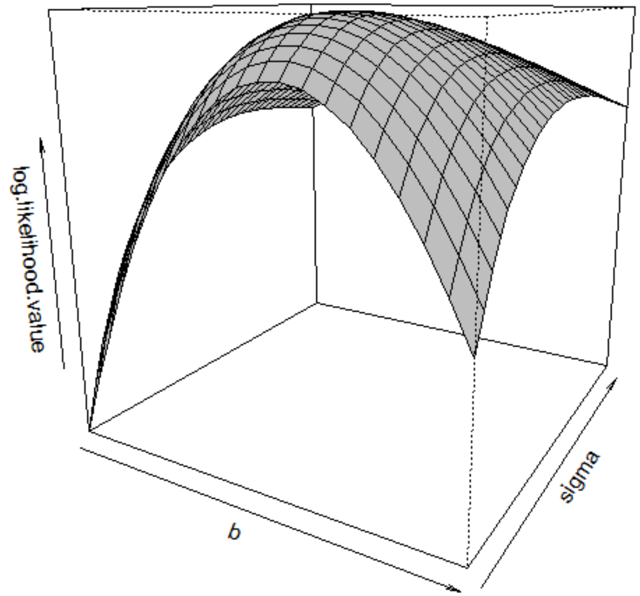
- Steepness = curvature = rate of change of the slope = the second derivative of the log-likelihood function

- more concave downwards -> narrower CI -> smaller variance

- $var(\hat{\theta}) \approx -\frac{1}{l''(\hat{\theta})}$
  - the second derivative of the log-likelihood function, evaluated at $\hat{\theta}$

- For multiple parameters, the ML estimators follow (asymptotically) a multivariate normal distribution.

- $V(\underline{\hat{\theta}})$ is a variance-covariance matrix



Univariate case:

$$var(\theta) \approx -\frac{1}{l''(\underline{\hat{\theta}})}$$

- Empirically $V(\underline{\hat{\theta}}) \approx -H(\underline{\hat{\theta}})^{-1}$

Matrix inverse!

- $H(\underline{\hat{\theta}})$ is called the Hessian matrix, the second derivative of the log-likelihood function evaluated at its peak $\underline{\hat{\theta}}$
- [OR] $-H(\underline{\hat{\theta}})$ plays a prominent role in likelihood theory. It is called the *observed Fisher information matrix*.

- It measures the amount of information a r.v. carries about an unknown parameter.
- $H(\underline{\hat{\theta}})$ is readily available in `optim()`

- Back to the rabbit data

```
# optim() FOR TWO-DIMENSIONAL PARAMETER SPACE, b AND sigma
# WITH HESSIAN MATRIX

result<-optim(par=c(1,1), regression.no.intercept.log.likelihood,
         method='L-BFGS-B',
         lower=c(-1000,0.0001), upper=c(1000,10000),
         control=list(fnscale=-1), dat=recapture.data, hessian=T)
# OBTAIN THE HESSIAN MATRIX
result$hessian
```

```
> result$hessian
              [,1]          [,2]
[1,] -2.365675e+01 -2.486900e-07
[2,] -2.486900e-07 -7.278254e+00
```

```
# THE VARIANCE-COVARIANCE MATRIX IS THE NEGATIVE OF
# THE INVERSE OF THE HESSIAN MATRIX.
# BY solve() FUNCTION
var.cov.matrix<-(-1)*solve(result$hessian)
var.cov.matrix
```

```
> var.cov.matrix
              [,1]          [,2]
[1,]  4.227123e-02 -1.444362e-09
[2,] -1.444362e-09  1.373956e-01
```

This is the variance-covariance structure of $(\hat{b}, \hat{\sigma})$

```
> var.cov.matrix
              [,1]          [,2]
[1,]  4.227123e-02 -1.444362e-09
[2,] -1.444362e-09  1.373956e-01
```

- $(\hat{b}, \hat{\sigma})$ follows (approximately) a bivariate normal distribution

- For example, 95% C.I. for b alone is $3.1629 \pm 1.96\sqrt{0.04227}$

- We can apply multivariate testing to test for $H_0: (b, \sigma) = (b_0, \sigma_0)$
  - Multivariate version of z-test
  - Multivariate analysis (beyond the scope of this course)

Note on confidence interval

- “In Author’s experience, the Wald (normality) and likelihood method can give quite different results when used to test joint hypotheses… The likelihood method can require more effort to compute, but is generally preferred.” (Millar, 2011)