Yesterday…

- Properties of MLE
  - Why do we prefer MLE to other estimators?
- Logistic regression
- Likelihood-Ratio test
  - For nested models

Today

- Interval estimation
  - Confidence interval
  - Confidence region
  - Normal approximation

Confidence interval estimation

- We are able to get point estimates for many examples
- Not enough to publish your results
- Need to include 95% CI when you quote your estimates
- There are many ways to calculate confidence interval, and some of them can be directly obtained from the log-likelihood function

- How about the 95% confidence interval? Say, in the coin tossing example, does the CI of p cover 0.5 (fair coin)?
- Remember the likelihood-ratio test?
- We can set up an LRT to test for H_0: $p = 0.5$
- So M1 for coin tossing is just putting $p = 0.5$ into the log-likelihood equation. Compute D and see if it is smaller than the critical value $\chi^2_{df=1,0.95}$=3.84

- $l(\hat{p}) = l(0.7) = -1.32115$
- $l(0.5) = -2.14398$
- $D = -2 * (-1.32115 + 2.14398) = 1.645 < \chi^2_{df=1,0.95}$
- According to LRT, H_0: $p = 0.5$ is not rejected

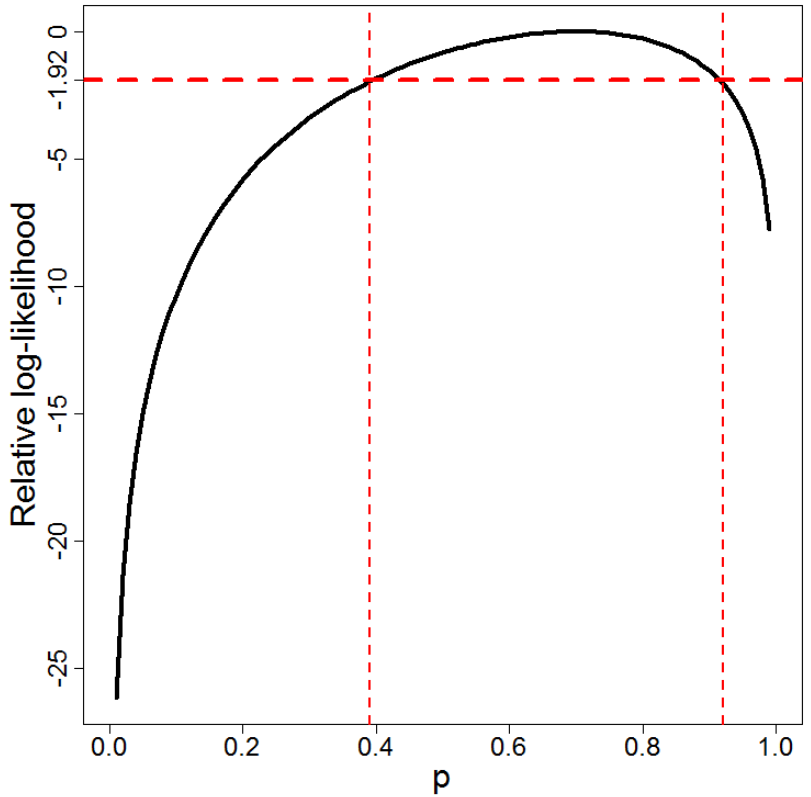
Confidence interval: coin tossing

- Instead of performing the test, we can find a range of p, such that D remains within the 'acceptance region', i.e. range of p in which $D < 3.84$.

- $D = 2 * (\ln(L2) - \ln(L1)) < 3.84$, so the difference of the log-likelihood value between M2 and M1 has to be smaller than 3.84/2=1.92

- In most cases, if we want to find the confidence interval for a single parameter, we look at the range of the parameter such that the log-likelihood is within 1.92 units from its maximum

- Rule of thumb: -1.92, or -2



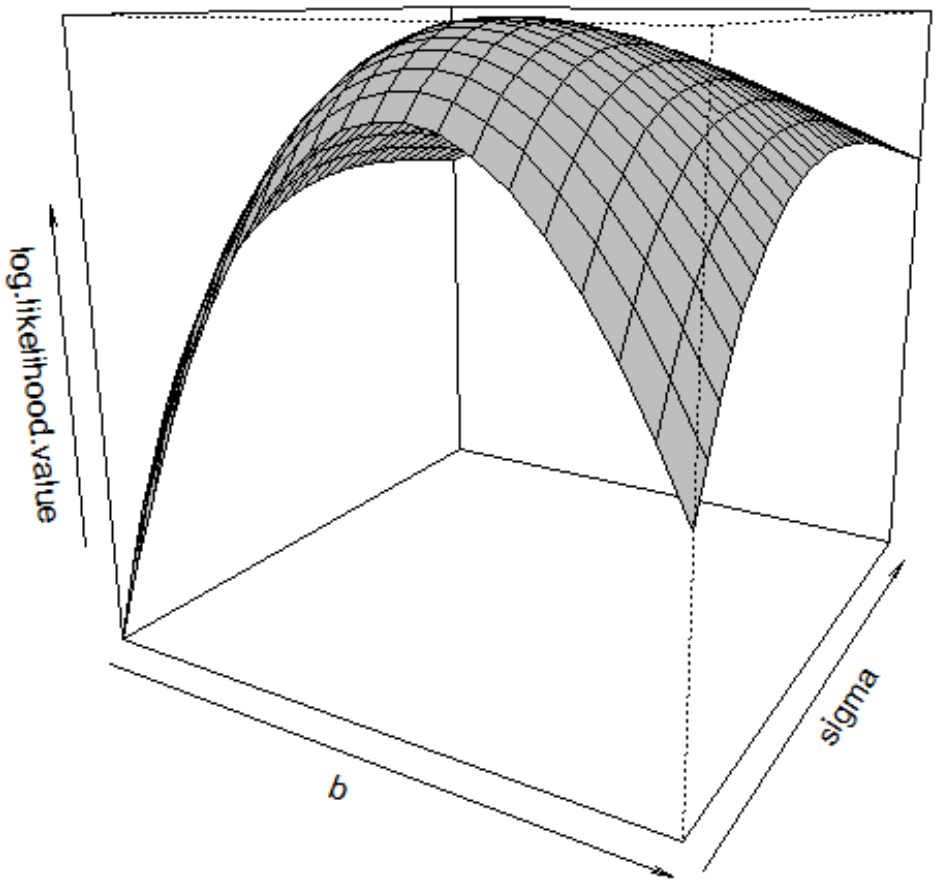
- If we observe 7 heads out of 10 tosses, the 95% CI for p is [0.39, 0.92]. Since 0.5 is within the 95% CI, we do not reject the 'fair coin' hypothesis.

Confidence interval: linear regression

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

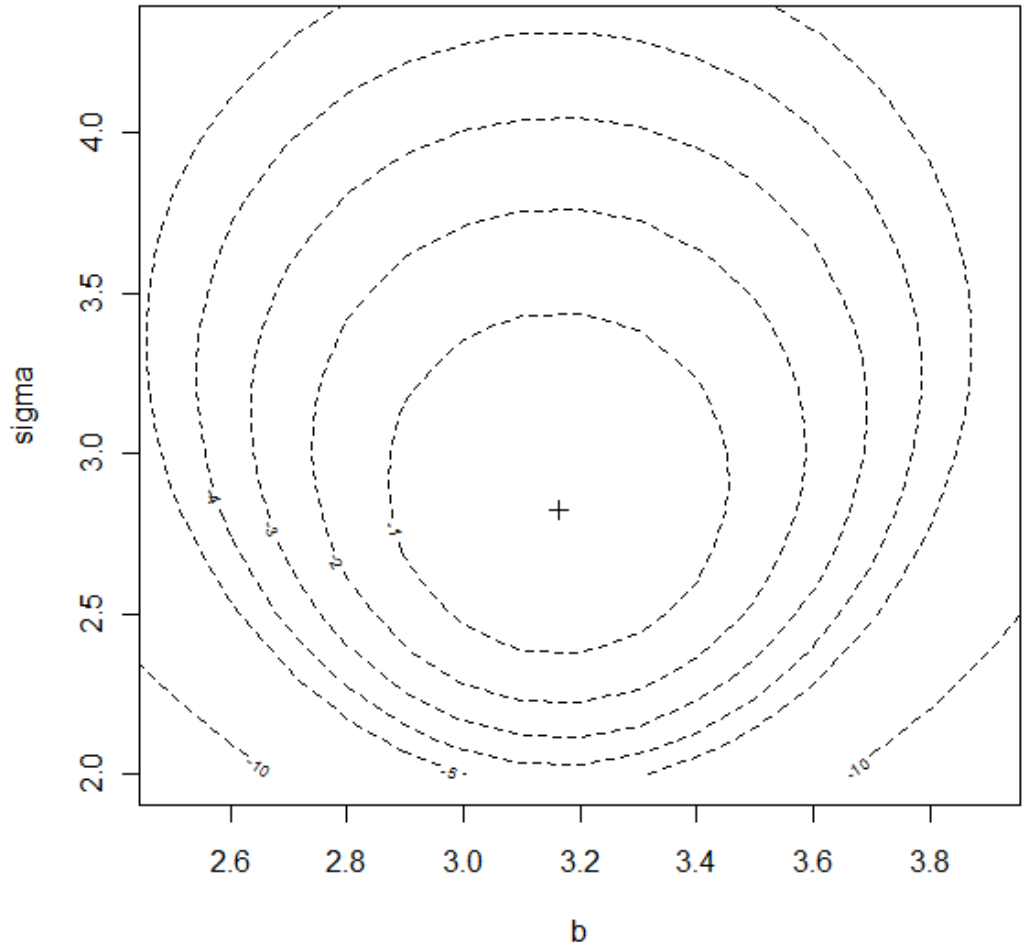
- Let's plot it out

```
# 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 VALUE RELATIVE TO THE MAXIMA
log.likelihood.value<-log.likelihood.value-M1$value

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


log.likelihood.value
b
sigma

How about a contour plot?

```
# CONTOUR PLOT
contour(b, sigma, 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)
```

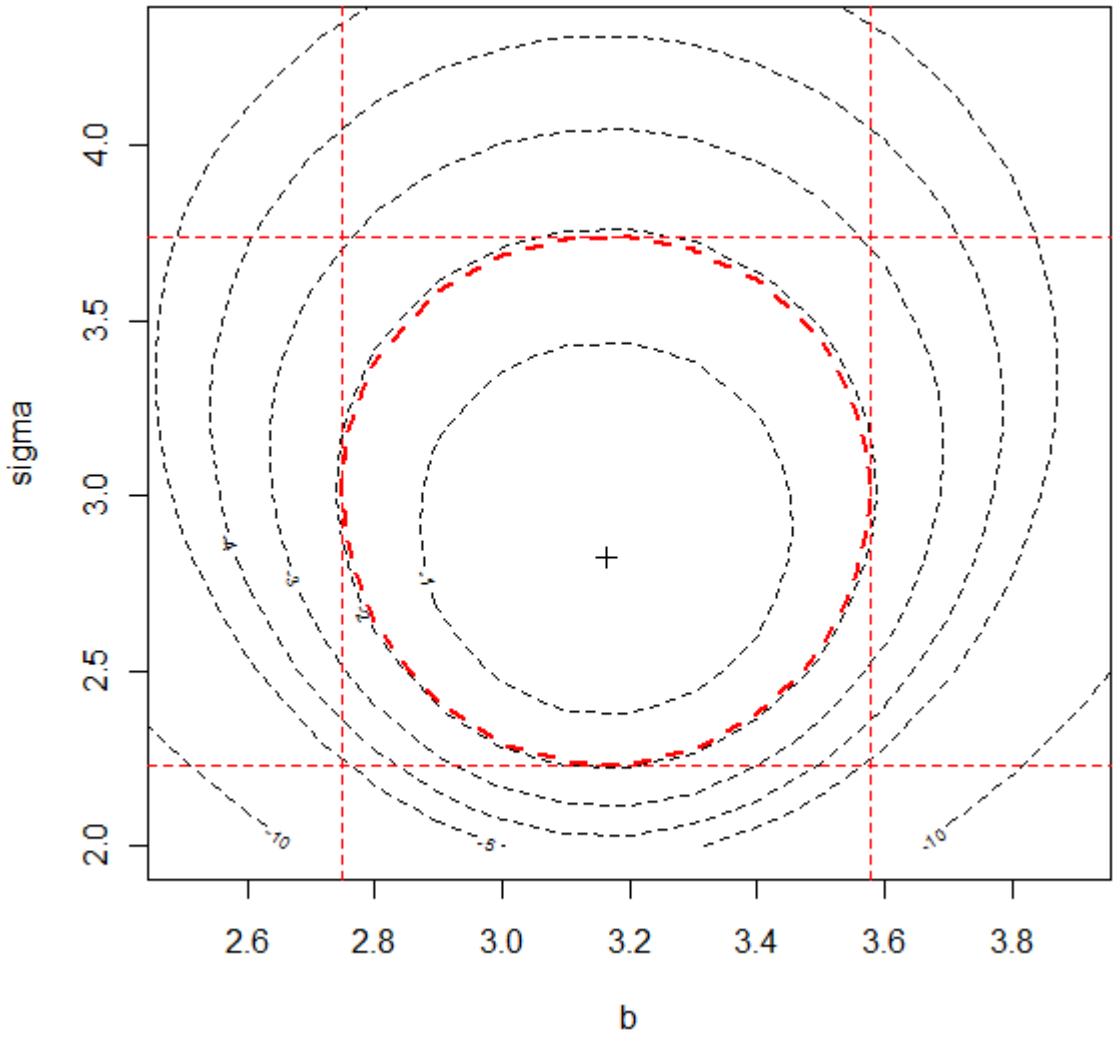


We can, again, draw the -1.92 line (circle) on the contour map

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

IF WE JUST LOOK AT ONE PARAMETER AT A TIME

95% CI for σ is $[2.23, 3.74]$



95% CI for b is $[2.75, 3.57]$

Joint confidence interval

- We know the 95% CI for b, and the 95% CI for σ
- But it does not mean we know the joint 95% confidence *region* for (b, σ)
- We need to consider the correlation between the two estimators
- Adjust for the overall α level
- THEY ARE NOT THE SAME THING!

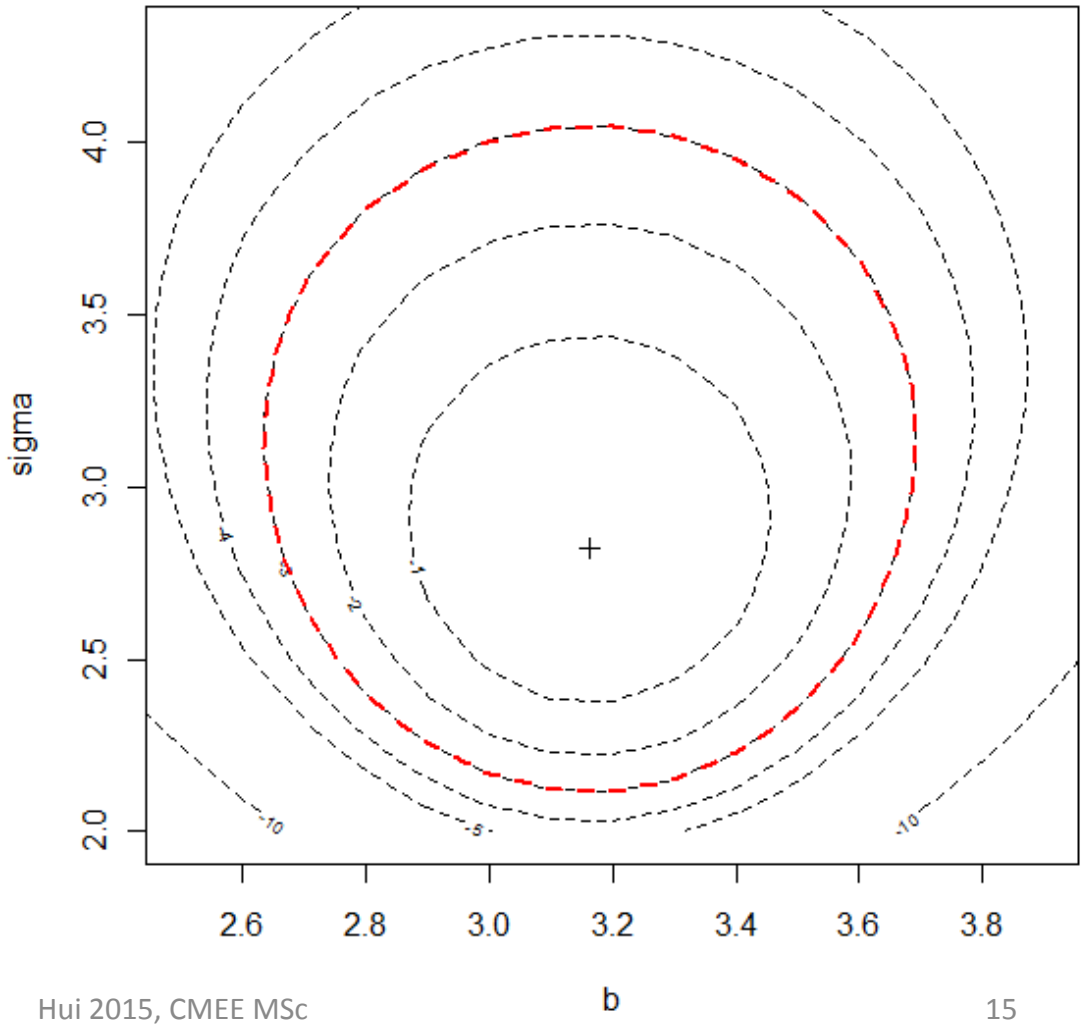
Joint confidence interval

- The general rule: the joint confidence interval (region) for k parameters is the collection of the parameter values for which the log-likelihood decreases by no more than half of $\chi^2_{k,1-\alpha}$ from its maximum.

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

- On the contour plot, we can circle the region in which the log-likelihood value is 2.99 units from the maximum.

The joint 95% CI for (b, σ) is all the points within the red circle



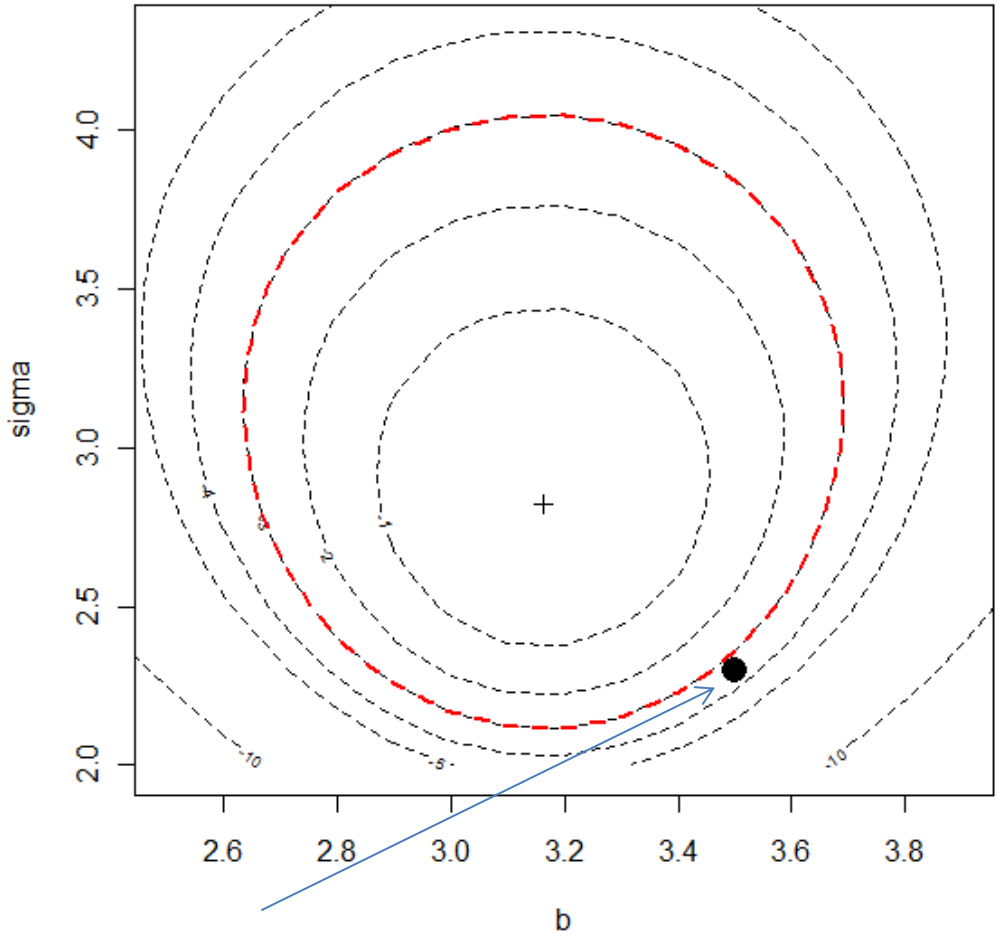
- Consider the parameter value $(b, \sigma) = (3.5, 2.3)$, does this value fall into the 95% confidence region?

No, it is outside the joint 95% confidence region, but… wait?

3.5 is within the 95% CI for b, and 2.3 is also within the 95% CI for σ.

Why is this the case???

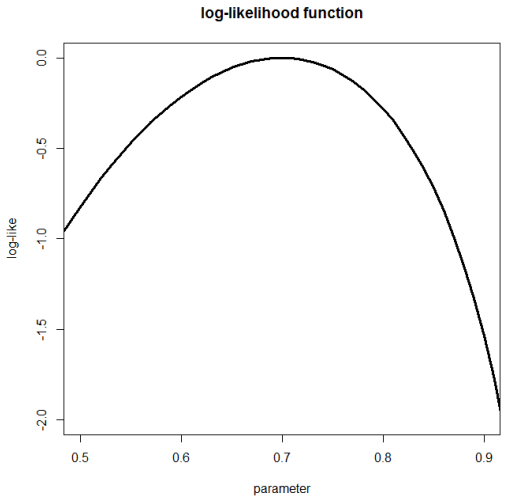
Multiple comparison!

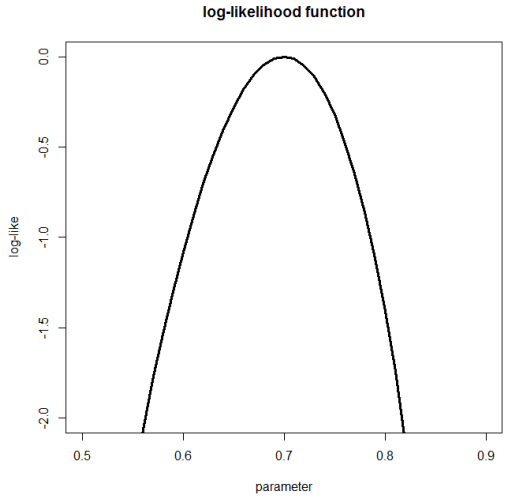


$(b, \sigma) = (3.5, 2.3)$

CI: Approximate normality of MLE

- Remember MLE is asymptotically normal?
- For one-parameter case, the 95% CI will be approximately $\hat{\theta} \pm 1.96\sqrt{var(\theta)}$
- But what is $var(\theta)$?
- The curvature of the log-likelihood function



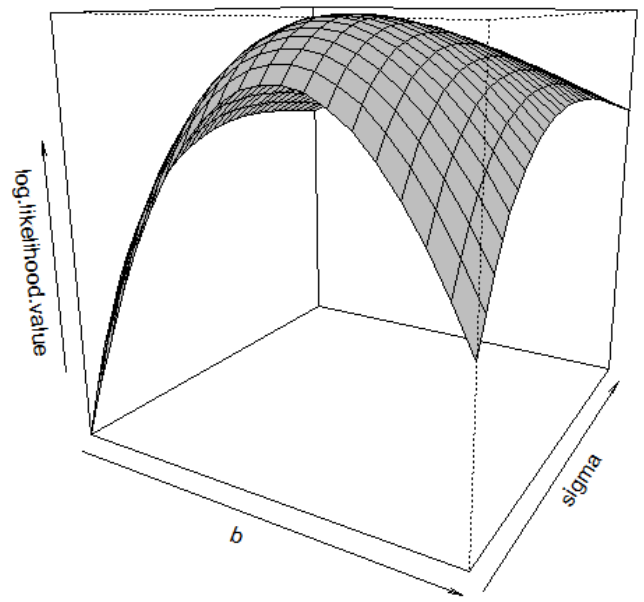


The rate of change of slope at the peak!

- Two log-likelihood curves: RHS is "steeper"
- The second order derivative of the log-likelihood function
- more concave downwards -> narrower CI
- $var(\theta) \approx -\frac{1}{\frac{\partial^2 l}{\partial \theta^2}}$

- For more than one parameter, the MLE follows (approximately) a multivariate normal distribution.

- The variance $V(\underline{\theta})$ this time is a variance-covariance matrix



- Luckily, we can obtain it in R!

Univariate case:

$$var(\theta) \approx -\frac{1}{\frac{\partial^2 l}{\partial \theta^2}}$$

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

Matrix inverse!

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

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

- Back to the rabbit data

```
# optim() GENERALISES TO MULTI-DIMENSIONAL CASES
# 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)
# GET BACK 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
variance.matrix<-(-1)*solve(result$hessian)
variance.matrix
```

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

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

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

- $(\hat{b}, \hat{\sigma})$ forms a bivariate normal distribution
- For example, 95% CI for b alone is $3.1629 \pm 1.96\sqrt{0.04227}$
- Can apply multivariate testing to $(\hat{b}, \hat{\sigma})$ (but the contents are out of this course)
- Look for "multivariate analysis" if interested
- Remember the bivariate normal distribution on day 2?

Notes on confidence interval

- Many more methods to calculate CI
  - Likelihood-ratio method
  - Wald CI: assumes (joint-)normality of MLE. Uses the magical 1.96 rule or Hessian matrix.
  - Profile CI: Partial maximisation of log-likelihood for high dimensional parameter space.

- "In Author's experience, the Wald and likelihood can give quite different results when used to test joint hypothesis (multiple parameters)… The likelihood method can require more effort to compute, but is generally prefered." (Millar, 2011)