Section 3: Optimization

February 9, 2012

Outline

Outline

A Running Example

Suppose we observe the following data from a dichotomous random variable Y: $\{1, 0, 0, 1, 1\}$.

Let's assume that Y is distributed Bernoulli with some constant probability of seeing a one across observations. We might also assume that observations are independent.

The model:

1. $Y_i \sim f_{\text{bern}}(y_i|\pi_i)$.
2. $\pi_i = \pi$.
3. Y_i and Y_j are independent for all $i \neq j$.

A Running Example

1. What's the PMF for a Bernoulli again?

$$Y_i = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1-\pi \end{cases}$$

2. So what's the joint distribution for our data conditional on π?

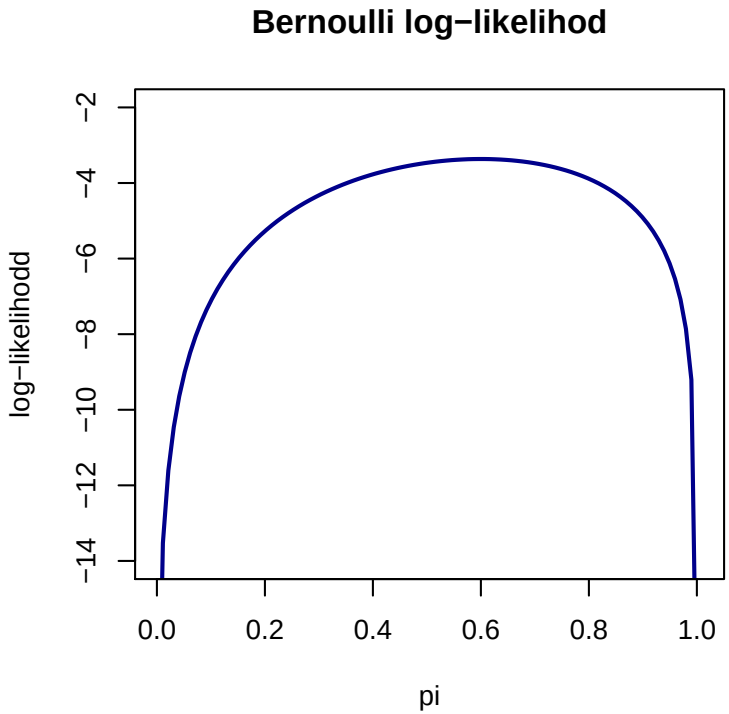
$$\begin{aligned} Pr(\mathbf{y}|\pi) &= Pr(Y_1 = 1, Y_2 = 0, ..., Y_5 = 1|\pi) \\ &= Pr(Y_1 = 1|\pi)Pr(Y_2 = 0|\pi)...Pr(Y_5 = 1|\pi) \\ &= \pi \cdot (1-\pi) \cdot (1-\pi) \cdot \pi \cdot \pi \\ &= \pi^3(1-\pi)^2 \end{aligned}$$

3. Recalling that $L(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)$, let's find our likelihood function.

$$\mathrm{L}(\pi|\mathbf{y}) \propto \pi^3(1-\pi)^2.$$

4. Let's take the natural logarithm, which is a relatively unimportant step now but will be essential for computational reasons later on.

$$\begin{aligned} \ln\left[\pi^3(1-\pi)^2\right] &= \ln(\pi^3) + \ln((1-\pi)^2) \\ &= 3\ln\pi + 2\ln(1-\pi). \end{aligned}$$



Where is the maximum, and how could we find it analytically using

$$\ln \mathrm{L}(\pi|\mathbf{y}) = 3 \ln \pi + 2 \ln(1 - \pi)?$$

5. Let's find the 1st derivative of $\ln \mathrm{L}(\pi|\mathbf{y})$, which is a function which tells us the slope of a line tangent to our function at any point. We want to know at which value of π the slope of the tangent line is zero.

$$\begin{aligned} \frac{\partial \ln \mathrm{L}(\pi|\mathbf{y})}{\partial \pi} &= \frac{\partial}{\partial \pi}[3 \ln \pi + 2 \ln(1-\pi)] \\ &= \frac{3}{\pi} + \frac{2}{1-\pi} \cdot (-1) \end{aligned}$$

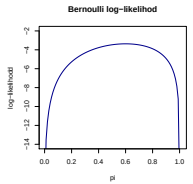
Note: this can also be done in R using `deriv()`:

```
> deriv(f(x) ~ 3*log(x) + 2*log(1-x), "x")
expression({
    .expr3 <- 1 - x
    ...
    .grad[, "x"] <- 3 * (1/x) - 2 * (1/.expr3)
    ...
```

5. Let's find the 1st derivative of $\ln \mathrm{L}(\pi|\mathbf{y})$, which is a function which tells us the slope of a line tangent to our function at any point. We want to know at which value of π the slope of the tangent line is zero.

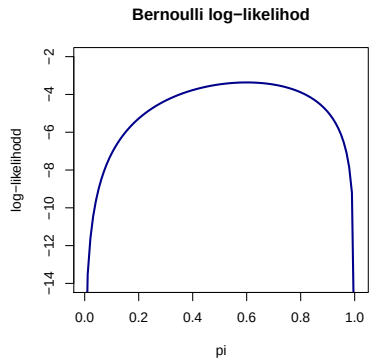
$$\begin{aligned} \frac{\partial \ln \mathrm{L}(\pi|\mathbf{y})}{\partial \pi} &= \frac{\partial}{\partial \pi}[3 \ln \pi + 2 \ln(1-\pi)] \\ &= \frac{3}{\pi} + \frac{2}{1-\pi} \cdot (-1) \end{aligned}$$

When this function is equal to zero, $\pi = \frac{3}{5}$. Q: Assuming we hadn't seen



how would we know if $\pi = \frac{3}{5}$ was a maximum or a minimum?

6. Check the 'critical value' (proposed maximum or minimum) by examining whether the slope of the function is decreasing or increasing at the critical value.



$$\begin{aligned}
\frac{\partial^2 \ln \mathrm{L}(\pi|\mathbf{y})}{\partial \pi^2} &= \frac{\partial}{\partial \pi}[\frac{3}{\pi} + \frac{2}{1-\pi} \cdot (-1)] \\
&= -\frac{3}{\pi^2} - \frac{2}{(1-\pi)^2} \\
&= -\frac{3}{.6^2} - \frac{2}{(1-.6)^2} \\
&= -20.83.
\end{aligned}$$

Outline

Outline

Optimization

The previous example we just developed is an example of **optimization**: the process of minimizing or maximizing a function by systematically choosing the values of variables from within an allowed set. For example:

$$\min_{x \in [-\infty, \infty]} f(x) = -\frac{1}{2}(3 - x)^2$$

- $f(x)$ is called the objective function
- x is the parameter (for us, β, π, σ, etc.)
- $x \in [-\infty, \infty]$ is the allowed set or the parameter space

Two ways to solve:

1. analytically (we just did this using derivatives)
2. numerically (we'll do this in a minute)

Outline

Analytical Optimization: Univariate case

Step One: Take the first derivative of the function and identify the critical value(s).

- The derivative of a function at a value x_0, denoted by $f'(x_0)$ or $\frac{\partial f}{\partial x}(x_0)$, is the instantaneous rate of change in $f(\cdot)$ at x_0.
- It partially describes the behavior of a function on an interval [a,b]
  - If $f^{'}(x) > 0$ for all $x \in [a, b]$, then f is increasing on the interval $[a, b]$
  - If $f^{'}(x) < 0$ for all $x \in [a, b]$, then f is decreasing on the interval $[a, b]$
  - If $f^{'}(x) = 0$ at some $x \in [a, b]$ then we say x is a critical value of f. Critical values may be maxima, minima, or saddle points.

Analytical Optimization: Univariate case

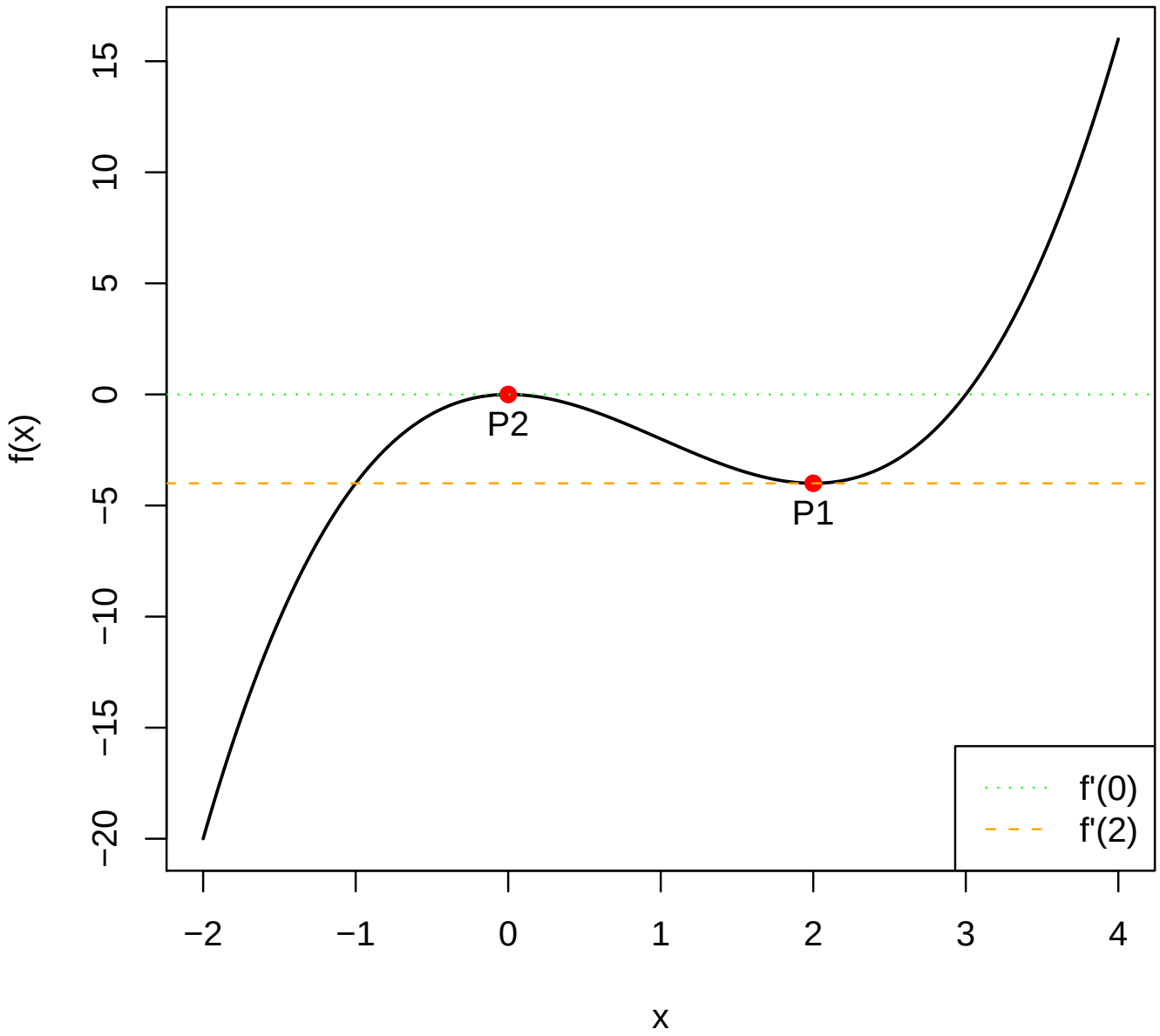
Step Two: Compute the second derivative of the function at the critical value(s) and evaluate.

- The second derivative of a function $f''(x)$ or $\frac{\partial^2 f}{\partial x \partial x}(x)$ is the derivative of the derivative, or the rate of change of the rate of change.
- Use the following to evaluate your critical value(s):
  - If $f'(x_0) = 0$, and $f''(x_0) < 0$, then x_0 is a maximum
  - If $f'(x_0) = 0$ and $f''(x_0) > 0$, then x_0 is a minimum
  - If $f'(x_0) = 0$ and $f''(x_0) = 0$, then x_0 may be a minimum, a maximum, or neither.

Analytical Optimization: Univariate case

For example:

- Find extreme value of: $f(x) = x^3 - 3x^2$
- $f'(x) = 3x^2 - 6x = 3x(x-2)$ so $f'(x) = 0$ for $x = 0$ and $x = 2$.
- Critical values are:
  - $P_1 = (0, f(0)) = (0, 0)$
  - $P_2 = (2, f(2)) = (2, -4)$
- Second derivative test: $f''(x) = 6x - 6 = 6(x-1)$
- Evaluate second derivative at critical values:
  - $f''(2) = 6$ so P_1 will be a minimum
  - $f''(0) = -6$ so P_2 will be a maximum


f(x)
15
10
5
0
−5
−10
−15
−20
−2
−1
0
1
2
3
4
x
P2
P1
f'(0)
f'(2)

Outline

Numerical Optimization

For problems of only slightly more complexity, using derivatives and solving for parameters in order to maximize may be not only impractical but impossible.

There are a number of functions in R which can be used to optimize functions, but the one we will use most heavily is `optim()`.

Numerical Optimization Example

`optim()` takes a starting value (`par`) and a function (`fn`) as its main arguments.

```
ll <- function(pie) return(3*log(pie) + 2*log(1-pie))

optim(par = .3, fn = ll, control = list(fnscale = -1),
      method = "BFGS", hessian = TRUE)
```

What are these three extra arguments?

1. `fnscale` multiplies the function by the given constant. As a default `optim()` finds the minimum so multiplying our function by -1 fools `optim()` into finding the maximum.
2. `method` is the variety of algorithm used to find the maximum. More on this in a second.
3. `hessian = TRUE` requests that `optim` return a matrix of second derivatives which in the above case will be 1x1.

Numerical Optimization Example Cont.

```
$par
[1] 0.6000034

$value
[1] -3.365058
...
$hessian
          [,1]
[1,] -20.83365

$Warning messages:
1: In log(1 - x) : NaNs produced
2: In log(1 - x) : NaNs produced
```

What is `optim` doing?

It depends on your choice in the `method` argument.

- `Nelder-Mead`: this is the default; it is slow but somewhat robust to non-differentiable functions.
- `BFGS`: a quasi-Newton Method; it is fast but needs a well behaved objective function.
- `L-BFGS-B`: similar to BFGS but allows box-constraints (i.e. upper and lower bounds on variables).
- `CG`: conjugate gradient method, may work for really large problems (we won't really use this).
- `SANN`: uses simulated annealing – a stochastic global optimization method; it is very robust but *very* slow.

Newton's Method

Newton's method: a pretty good approach for a continuous and twice-differentiable function. We'll look at a univariate function here.

Suppose we know our function $f(\cdot)$ and we have a starting value of x_0. Our goal is to find move from x_0 to x_1 such that $f'(x_1) = 0$ (or is at least closer to 0 then than at x_0).

This will be a sequential process of approximation and eventually $f'(x_n)$ will be close enough to zero to let us declare that x_n a critical value.

Newton's Method

Recall Taylor's Theorem:

$$g(x_1) \approx g(x_0) + g'(x_0)(x_1 - x_0) + \frac{g''(x_0)}{2!}(x_1 - x_0)^2 + \ldots + \frac{g^k(x_0)}{k!}(x_1 - x_0)^k.$$

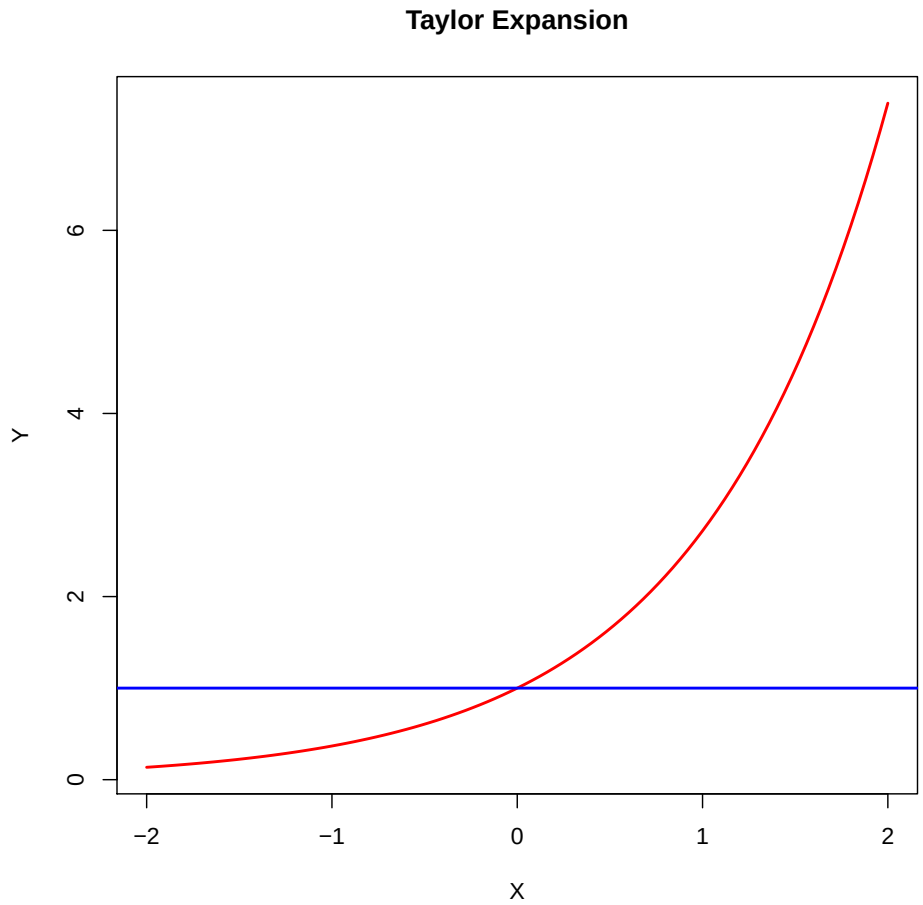


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_0(x_1) = 1$ (from Wikipedia)

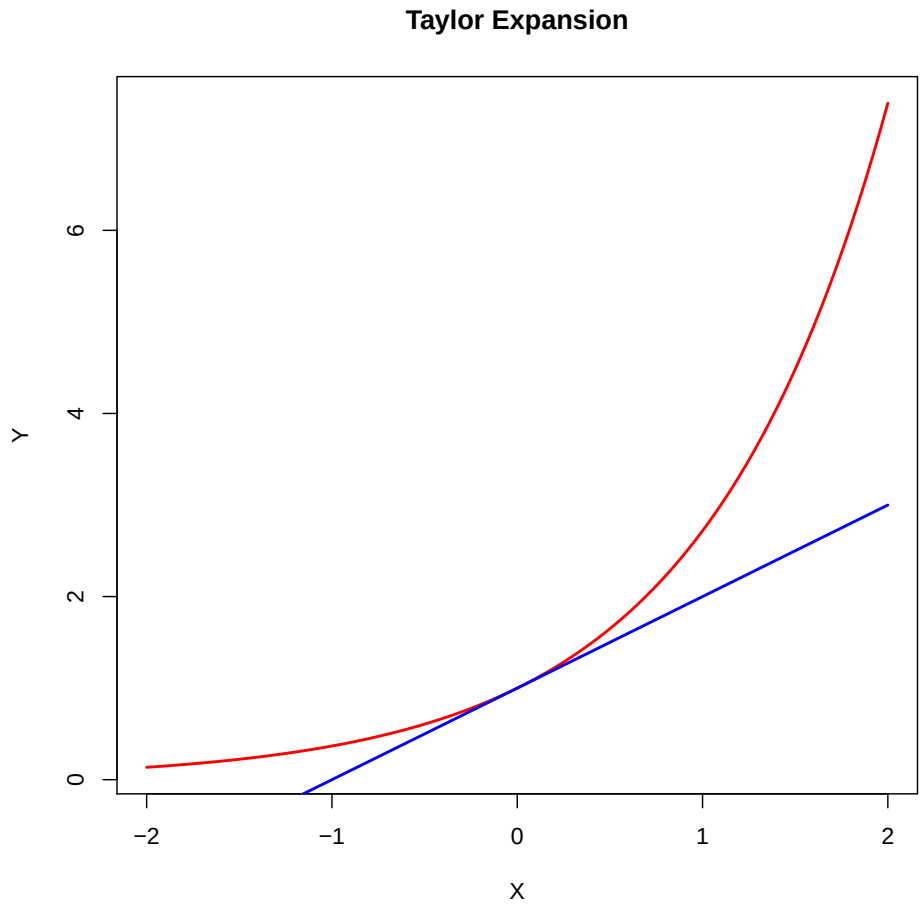


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_1(x_1) = 1 + x_1$ (from Wikipedia)

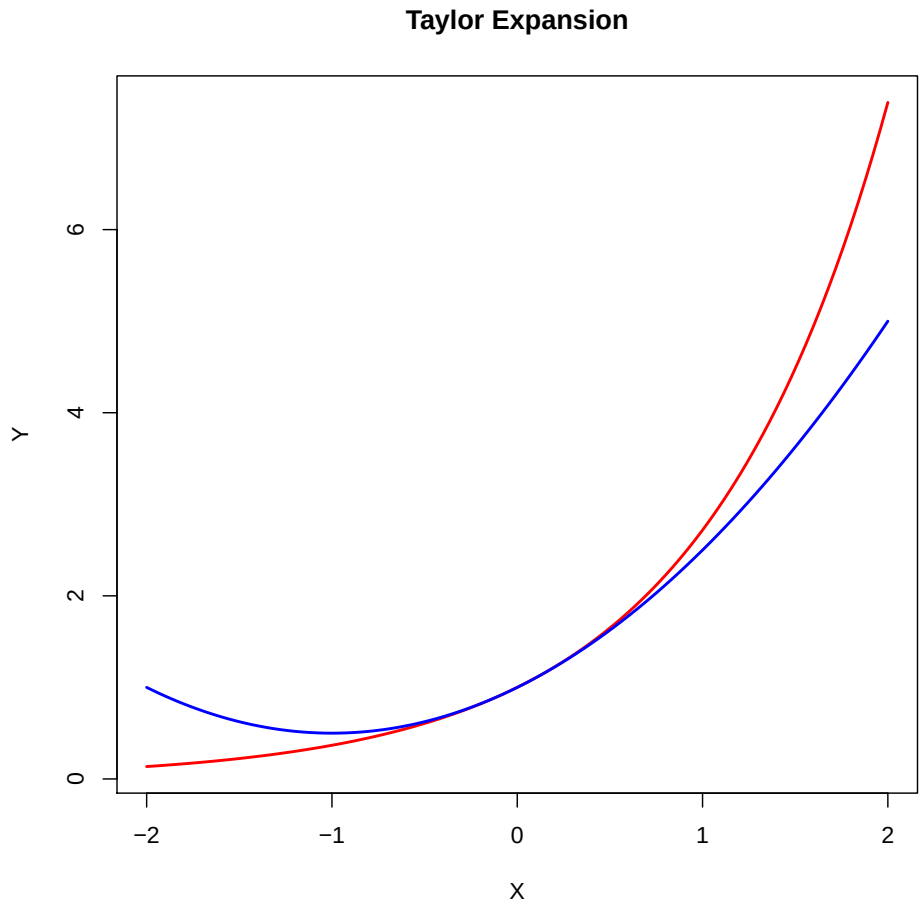


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_2(x_1) = 1 + x_1 + \frac{x^2}{2}$ (from Wikipedia)

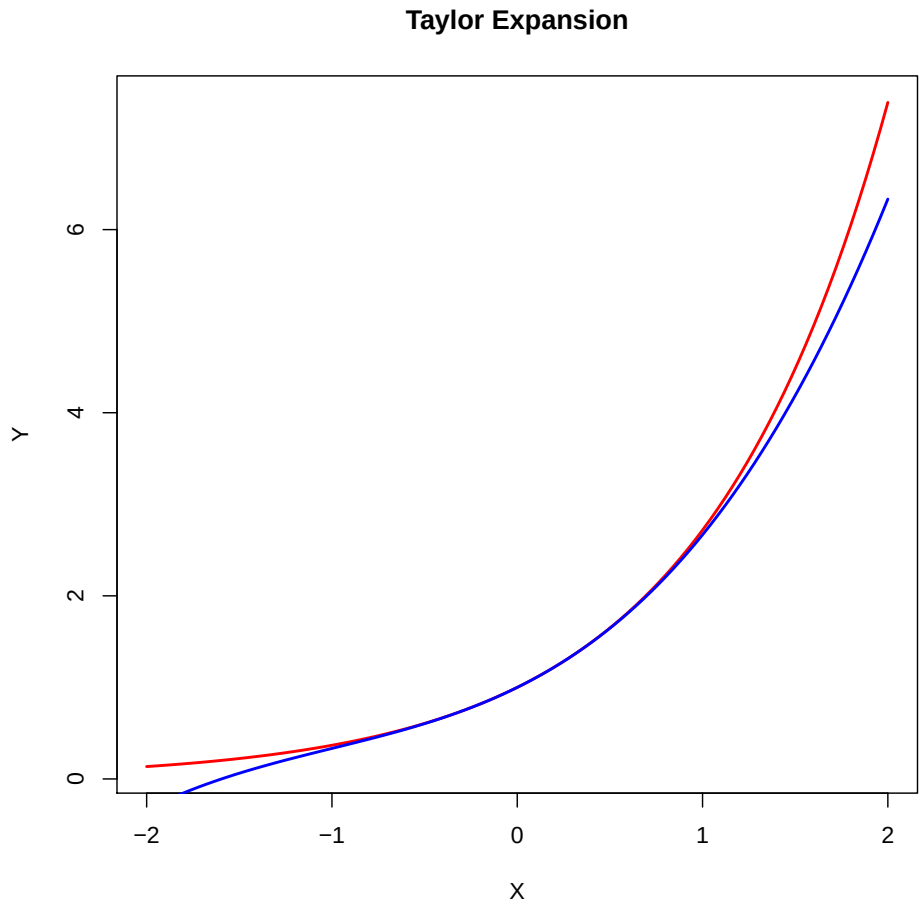


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_3(x_1) = 1 + x_1 + \frac{x^2}{2} + \frac{x^3}{6}$ (from Wikipedia)

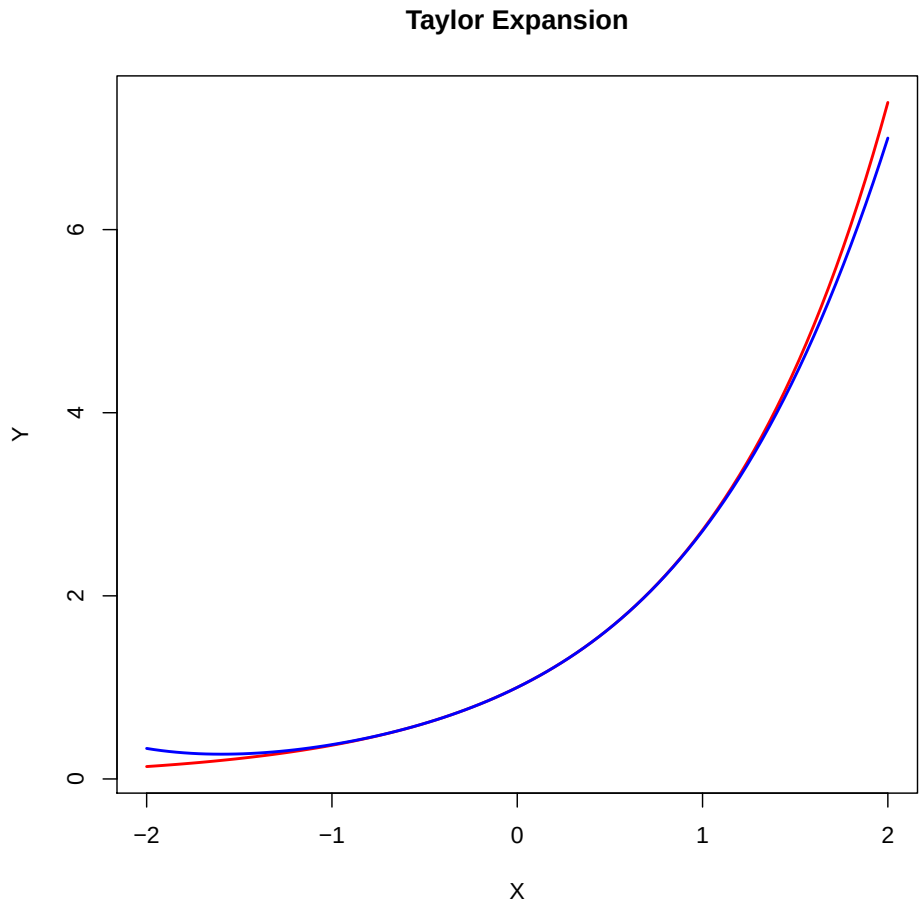


Figure: The exponential function, $g(x) = e^x$, and the Taylor Series approximation: $x_0 = 0$, $g_4(x_1) = 1 + x_1 + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}$ (from Wikipedia)

Newton's Method

Recall Taylor's Theorem:

$$g(x_1) \approx g(x_0)+g'(x_0)(x_1-x_0)+\frac{g''(x_0)}{2!}(x_1-x_0)^2+\ldots+\frac{g^k(x_0)}{k!}(x_1-x_0)^k.$$

We want to find x_1 such that $f'(x_1) = 0$ so our $g(\cdot)$ is $f'(\cdot)$. Let's *really* approximate: $f'(x_1) = f'(x_0) + f''(x_0)(x_1 - x_0)$.

Setting this to zero we get an updating formula to make $f'(x_1)$ approximately zero:

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

or for the nth iteration of our procedure

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}.$$

Newton in Action

Let's maximize our likelihood: $3\ln\pi + 2\ln(1-\pi)$.

Recall that $L'(\pi) = \frac{3}{\pi} - \frac{2}{1-\pi}$ and $L''(\pi) = -\frac{3}{\pi^2} - \frac{2}{(1-\pi)^2}$.

Starting at $\pi_0 = .3$, we use our updating formula:

$$\pi_1 = \pi_0 - \frac{L'(\pi_0)}{L''(\pi_0)} = .3 - \frac{L'(.3)}{L''(.3)} = 0.4909.$$

Now use $\pi_1 = .4909$ as a starting value.

$$\pi_2 = \pi_1 - \frac{L'(\pi_1)}{L''(\pi_1)} = .4909 - \frac{L'(.4909)}{L''(.4909)} = 0.5991.$$

So we're already there!

Properties of Newton-Raphson

- Converges quickly
- Can get stuck in local minima/maxima
- Can have troubles with root jumping
- Won't walk at all on a flat space

Outline

The Multivariate Normal

Suppose I give you two observations (x=2, y=4) and tell you that they are from a bivariate normal distribution. It's known that $\sigma_x = \sigma_y = 1$ and x and y are uncorrelated. The two unknown parameters are μ_x and μ_y. The pdf is:

$$f(x, y|\mu_x, \mu_y) = \frac{1}{2\pi} e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]}.$$

So what's the likelihood?

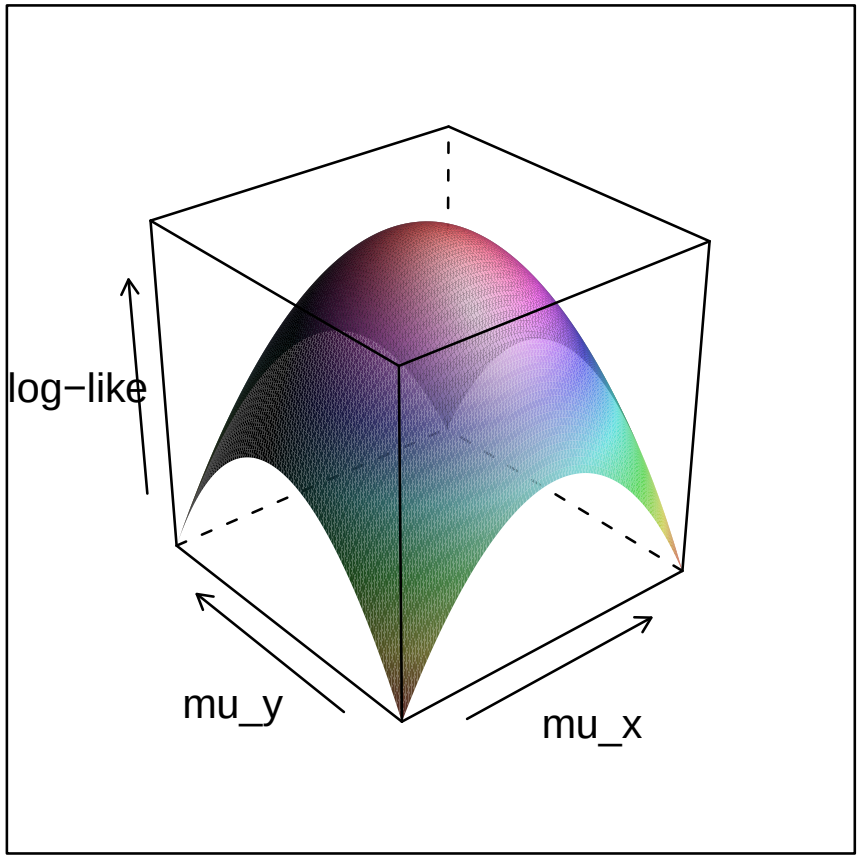
$$L(\mu_x, \mu_y|x, y) \propto \frac{1}{2\pi} e^{-\frac{1}{2}\left[(x-\mu_x)^2+(y-\mu_y)^2\right]}.$$

And the log-likelihood is especially nice:

$$\begin{aligned} \ln L(\mu_x, \mu_y|x, y) &\propto \ln \frac{1}{2\pi} - \frac{1}{2}\left[(x-\mu_x)^2 + (y-\mu_y)^2\right] \\ &\doteq -\frac{1}{2}\left[(x-\mu_x)^2 + (y-\mu_y)^2\right] \end{aligned}$$

Visualizing the Likelihood

We can plug in the datum and plot the likelihood:



Multivariable Case: Analytically

Find extrema of:

$$\begin{aligned} l(\mu_x, \mu_y | x, y) &\propto -\frac{1}{2}\left[(x-\mu_x)^2 + (y-\mu_y)^2\right] \\ &= -\frac{1}{2}\left[(2-\mu_x)^2 + (4-\mu_y)^2\right] \\ &= -10 + 2\mu_x - \frac{\mu_x^2}{2} + 4\mu_y - \frac{\mu_y^2}{2}. \end{aligned}$$

Step One: Find critical value(s) using the the partial derivatives of $f(x, y)$ with respect to x and y.

- $\frac{\partial l}{\partial \mu_x} = 2 - \mu_x$ is 0 if $\mu_x = 2$.
- $\frac{\partial l}{\partial \mu_y} = 4 - \mu_y$ is 0 if $\mu_y = 4$.
- So we have a critical value at $(\mu_x = 2, \mu_y = 4)$.

Multivariable Case: Analytically

Step Two: Evaluate critical value using the second derivative, but we have a wrinkle here because we have two variables. We're going to end up with a 2x2 Hessian matrix.

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 l}{\partial \mu_x \partial \mu_x}(2,4) = -1 & \frac{\partial^2 l}{\partial \mu_x \partial \mu_y}(2,4) = 0 \\ \frac{\partial^2 l}{\partial \mu_y \partial \mu_x}(2,4) = 0 & \frac{\partial^2 l}{\partial \mu_y \partial \mu_y}(2,4) = -1 \end{pmatrix}$$

It turns out that the key issue here for determining whether we have a maximum settles on whether the Hessian matrix is negative or positive definite (or something else). Define $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$.

1. negative definite: $\mathbf{a}'\mathbf{H}\mathbf{a} < 0$ for all $\mathbf{a} \neq 0 \rightarrow$ maximum.
2. positive definite: $\mathbf{a}'\mathbf{H}\mathbf{a} > 0$ for all $\mathbf{a} \neq 0 \rightarrow$ minimum.

Our example is easy: $\mathbf{a}'\mathbf{H}\mathbf{a} = \begin{pmatrix} -a_1 & -a_2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = -a_1^2 - a_2^2$. So we're **negative definite**.

Multivariate Case: Numerically

$$l(\mu_x, \mu_y | x, y) = -\frac{1}{2}\left[(2-\mu_x)^2 + (4-\mu_y)^2\right]$$

```
ll <- function(par){
  mux <- par[1]
  muy <- par[2]
  like <- -.5*((2-mux)^2+(4-muy)^2)
  return(like)
}

optim(par = c(0,0), fn = ll, method = "BFGS",
      control = list(fnscale = -1), hessian = TRUE)
```

```
$par
[1] 2 4

$value
[1] -4.930381e-31
...
$convergence
[1] 0
...
$hessian
                [,1]           [,2]
[1,] -1.000000e+00 -2.646978e-17
[2,] -2.646978e-17 -1.000000e+00
```

optim() Techniques in Practice

You can always test these techniques on some function.

```
fkt <- function(x){10*sin(0.3*x)*sin(1.3*x^2)
                   + 0.00001*x^4 + 0.2*x+80}
```

Figure: Plotting the Wild Function

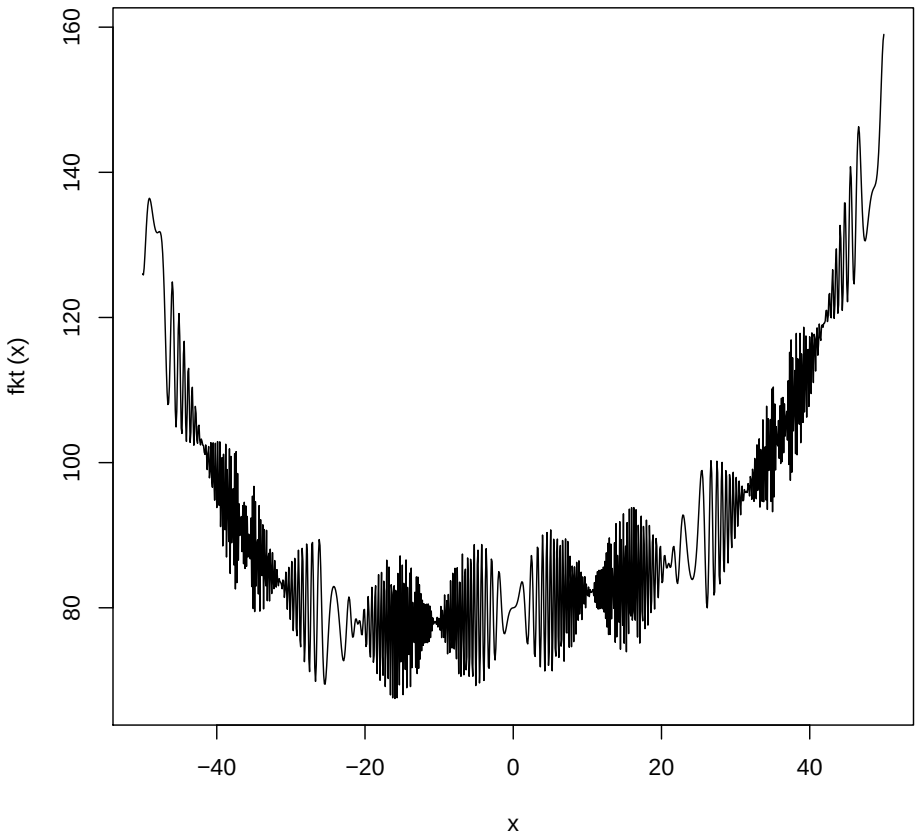
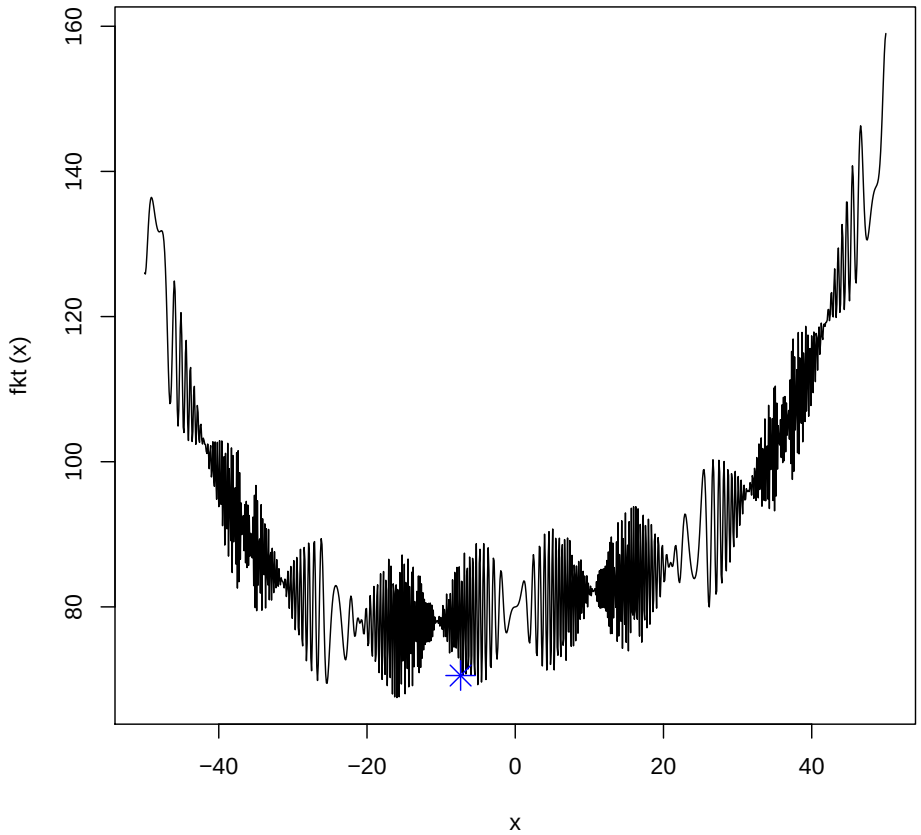
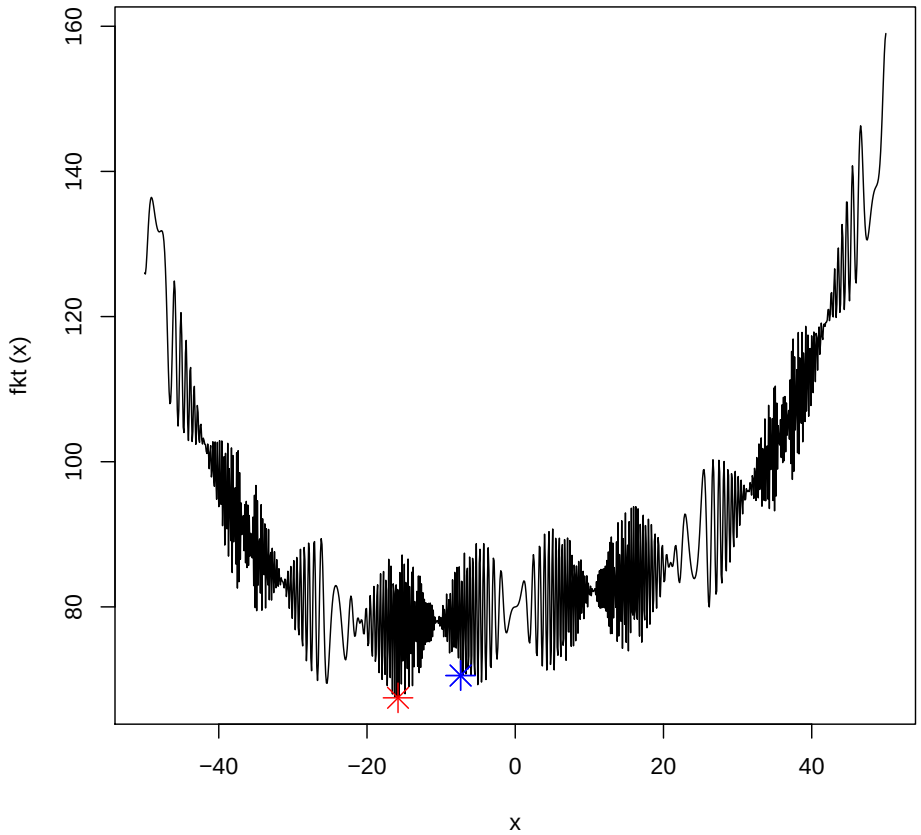


Figure: Using BFGS at x=50




```
out <- optim(par = 50, fkt, method="BFGS",
             control=list(maxit=20000, parscale=20))
points(out$par, out$val, pch = 8, col="blue", cex=2)
```

Figure: Using Simulated Annealing at x=50




```
out <- optim(par = 50, fkt, method="SANN",
             control=list(maxit=20000, temp=20, parscale=20))
points(out$par, out$val, pch = 8, col = "red", cex = 2)
```

Outline

Putting It All Together

Suppose we have some count data (number of coups in a year).

We can use a Poisson distribution to model the data (we will learn more about Poisson later).

$$Y_i \sim_{iid} \text{Poisson}(\lambda)$$

We want to find λ, which is the mean of the Poisson distribution.

The PMF (discrete) for the data is

$$p(\mathbf{y}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

Since $L(\theta|y) = p(y|\theta)$, we have

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

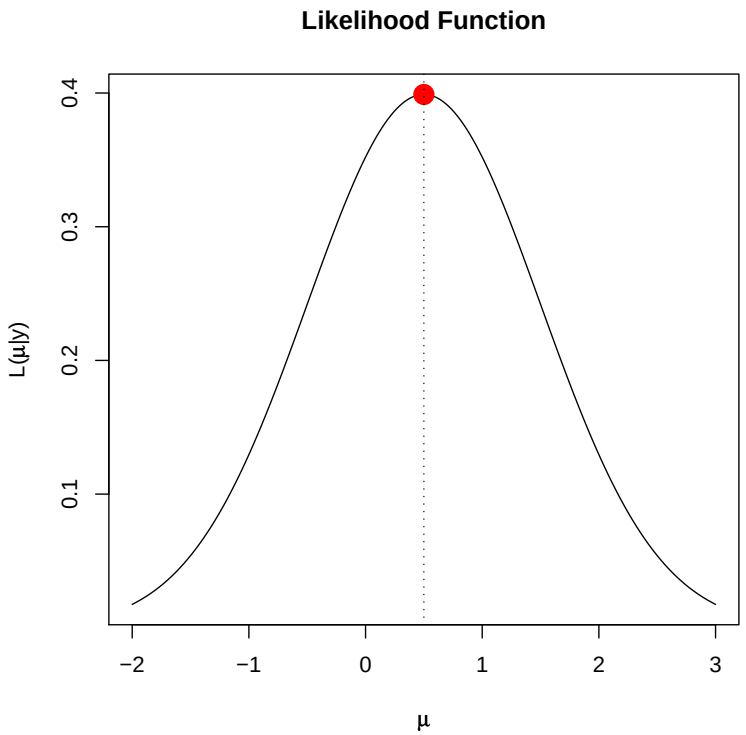
To make the math easier, we will take the log-likelihood.

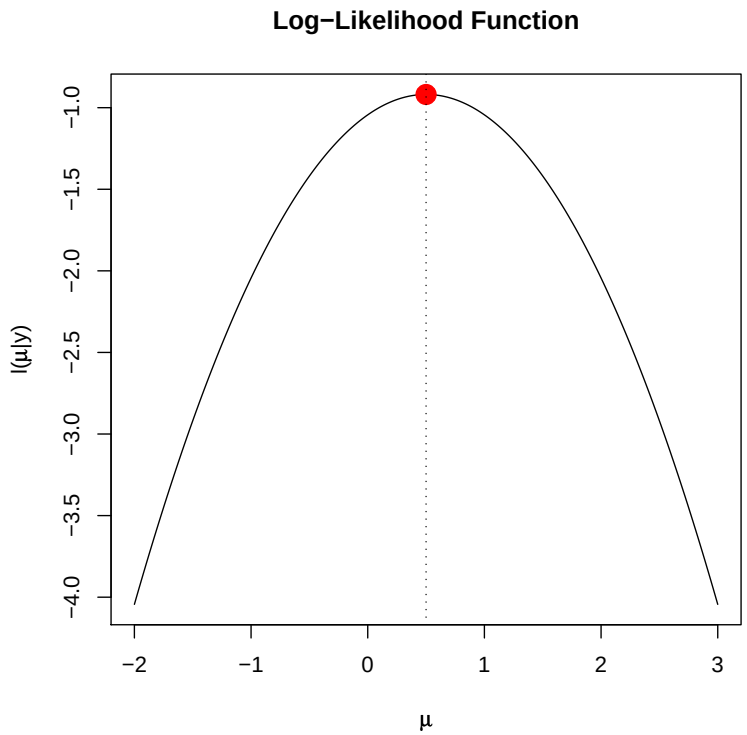
$$l(\lambda|\mathbf{y}) = \sum_{i=1}^{n} (y_i \ln \lambda - \lambda - \ln y_i!)$$

We can drop all terms that don't depend on λ (because likelihood is a relative concept and is invariant to shifts).

$$l(\lambda|\mathbf{y}) = \sum_{i=1}^{n} (y_i \ln \lambda) - n\lambda$$

Why Can We Use the Log-likelihood?





Finding the Maximum Likelihood Estimate (MLE)

Remember that to find our MLE, we want to find the value of the parameter(s) that maximizes our log-likelihood.

$$l(\lambda|\mathbf{y}) = \sum_{i=1}^{n}(y_i \ln \lambda) - n\lambda$$

We need to set the derivative (known as the score function) to zero and solve for λ.

$$\begin{aligned}
\frac{\partial l(\lambda|\mathbf{y})}{\partial \lambda} = S(\theta) &= \frac{\sum_{i=1}^{n} y_i}{\lambda} - n \\
0 &= \frac{\sum_{i=1}^{n} y_i}{\lambda} - n \\
\hat{\lambda} &= \frac{\sum_{i=1}^{n} y_i}{n}
\end{aligned}$$

Maximum Likelihood In R

Write our log-likelihood function:

$$l(\lambda|\mathbf{y}) = \sum_{i=1}^{n}(y_i \ln \lambda) - n\lambda$$

```
> ll.poisson <- function(par, y) {
+     lambda <- exp(par)
+     out <- sum(y * log(lambda)) - length(y) * lambda
+     return(out)
+ }
```

Find the maximum (with sample data)

```
> data <- rpois(1000, 5)
> opt <- optim(par = 2, fn = ll.poisson, method = "BFGS", control = list(fnscale = -1),
+     y = data)$par
> mle <- exp(opt)
> mle

[1] 4.996001
```