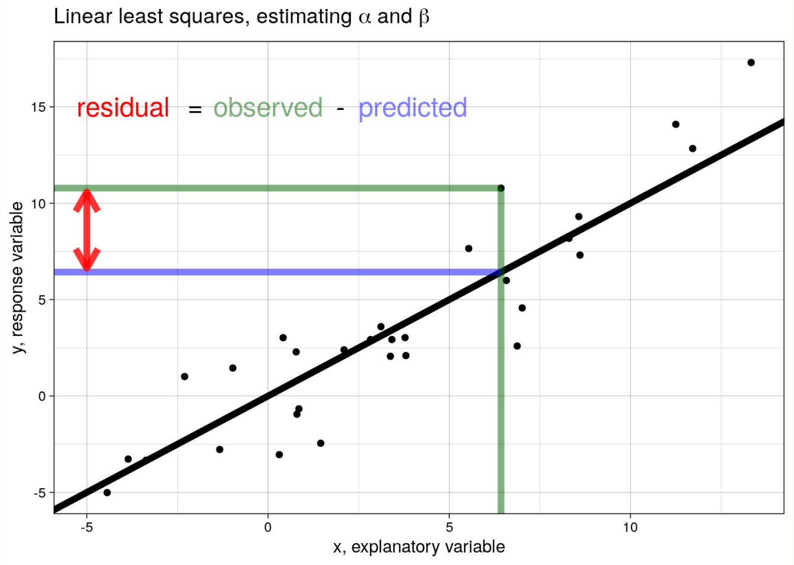
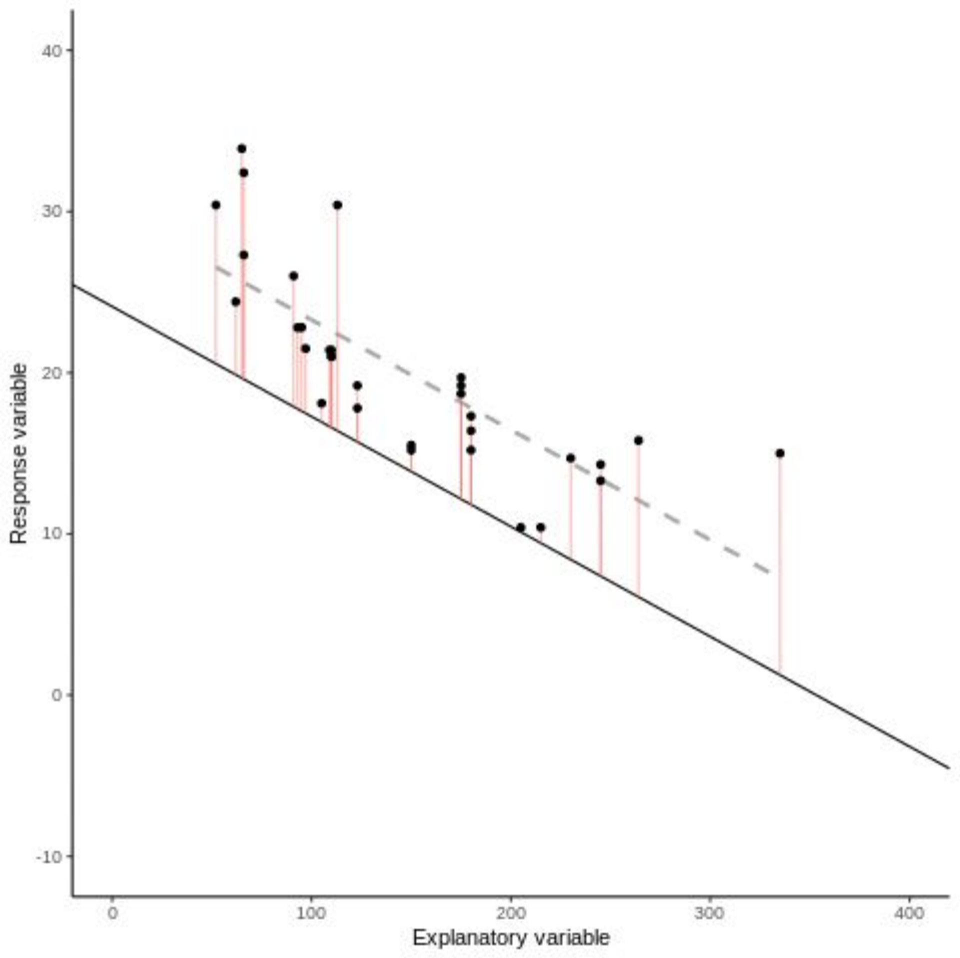
Maximum likelihood estimation

Least Squares Estimation

works by minimising the squared ‘distances’ between each observation and the line of best fit



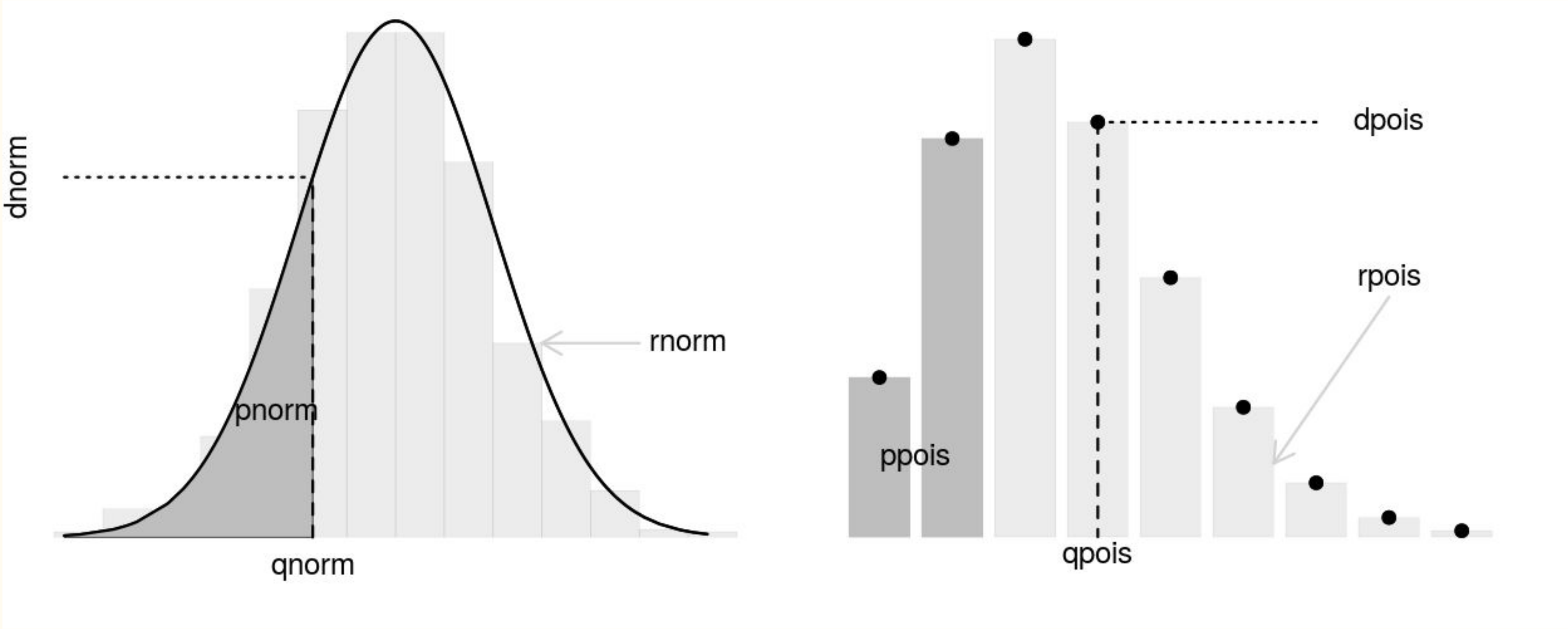


Probability distribution (& mass) functions

- A **probability mass function (PMF)** is a function that gives the probability that a **discrete** random variable is exactly equal to some value.
- A **probability density function (PDF)** is, roughly speaking, the alternative for a **continuous** random variable. We use it all the time to calculate probabilities and to gain an intuitive feel for the shape and nature of the distribution. Important: $P(X = x) = 0$ for a continuous random variable.

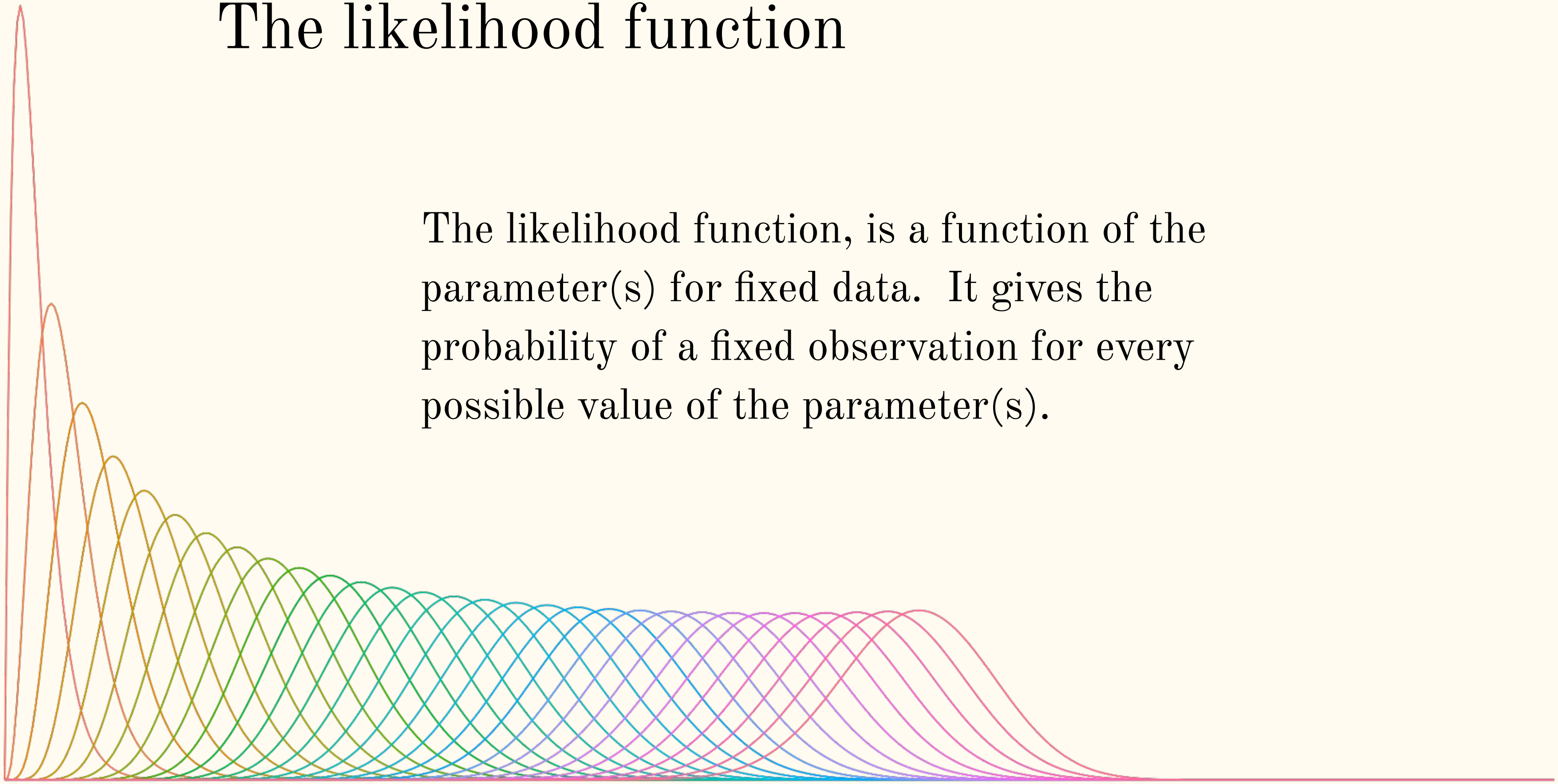
[https://cmjt.shinyapps.io/probable/]

R distribution functions and what they tell you

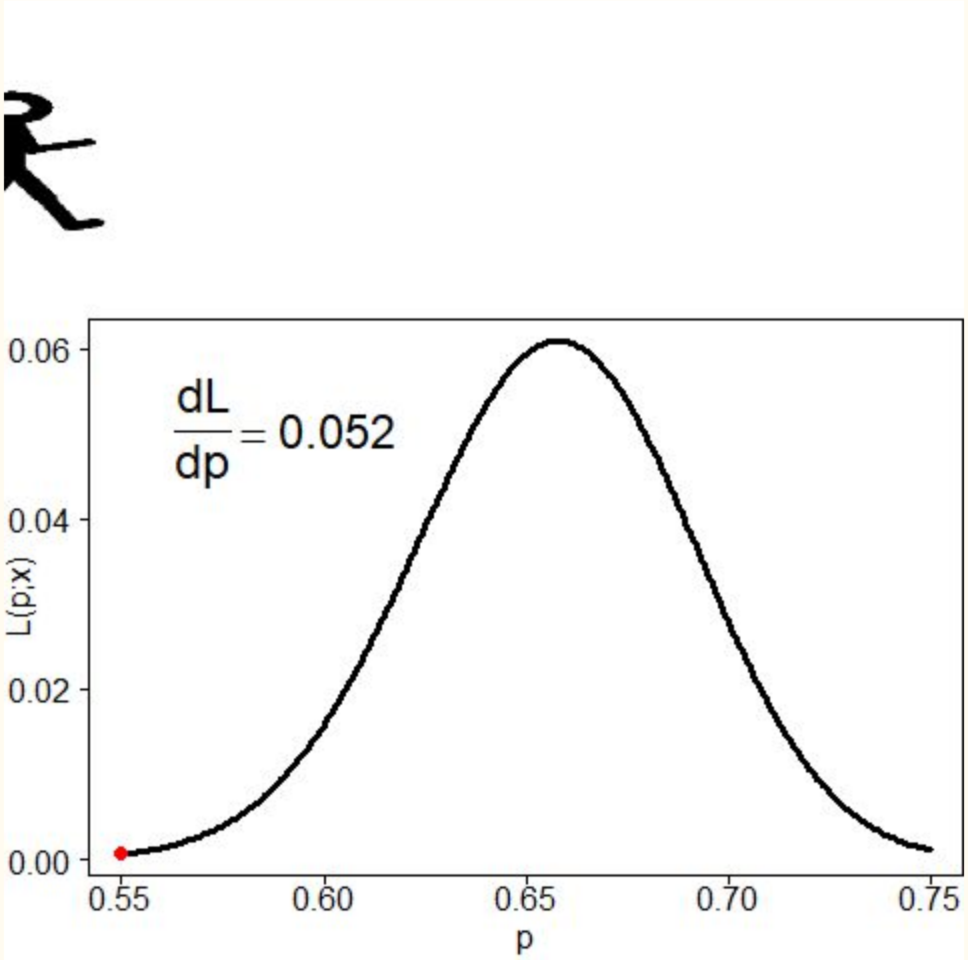


The likelihood function

The likelihood function, is a function of the parameter(s) for fixed data. It gives the probability of a fixed observation for every possible value of the parameter(s).



What is a derivative?



The binomial likelihood function (finding the MLE)

The likelihood function, $L(\theta; x)$ is a function of the parameter(s) θ for fixed data x and it gives the probability of a fixed observation x for every possible value of the parameter(s) θ, $P(X = x)$.

The binomial likelihood function (finding the MLE)

$$L(\theta; s) = P(S = s) = \binom{n}{s}\theta^s(1-\theta)^{n-s}.$$

Setting $\frac{\delta L}{\delta \theta} = 0$ and solve for θ:

$$\frac{\delta L}{\delta \theta} = (1-\theta)^{n-s-1}\theta^{s-1}\{s - n\theta\} = 0.$$

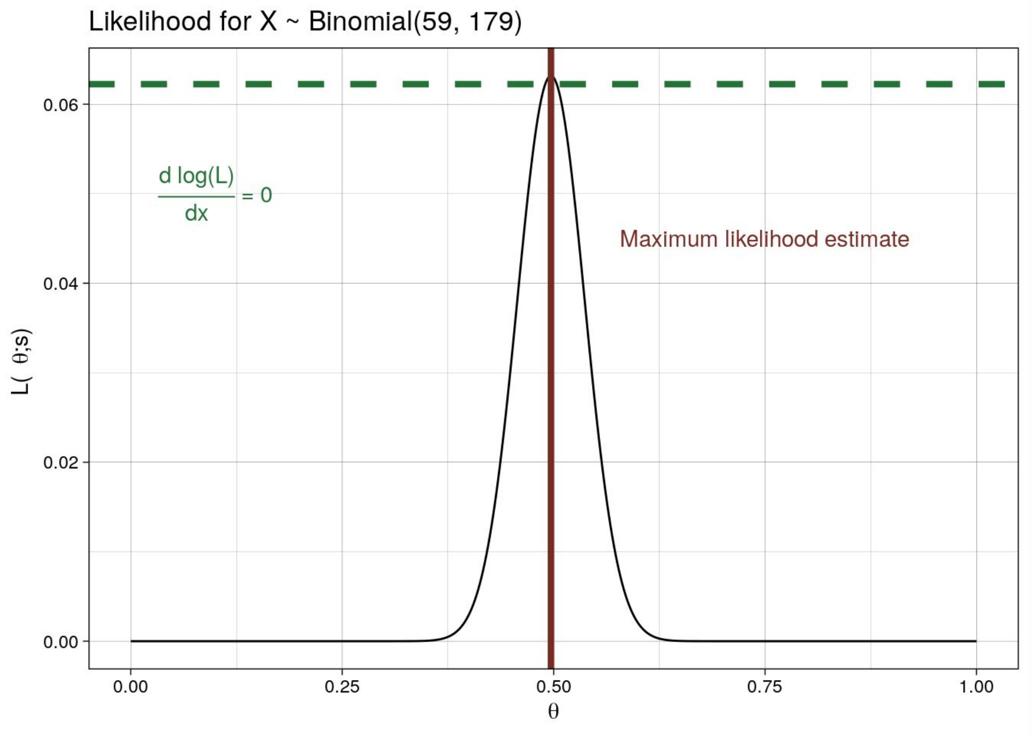
There are, technically, three possible solutions to this:

1. when $\theta^{s-1} = 0 \rightarrow \theta = 0$,
2. when $s - n\theta = 0 \rightarrow \theta = \frac{s}{n}$, or
3. when $(1-\theta)^{n-s-1} \rightarrow \theta = 1$.

The binomial likelihood function (finding the MLE)

```
likelihood <- function(theta) dbinom(x = 79, size = 159, prob = theta)
```

```
optimise(likelihood, c(0,1), maximum = TRUE)
```



The binomial log-likelihood function (finding the MLE)

$$\begin{aligned}\log(L(\theta;s)) &= \log(\tbinom{n}{s}) + \log(\theta^s) + \log((1-\theta)^{n-s}) \\ &= \log(\tbinom{n}{s}) + s\log(\theta) + (n-s)\log(1-\theta).\end{aligned}$$

Differentiating this:

$$\begin{aligned}\frac{\delta \log(L(\theta;s))}{\delta\theta} &= 0 + \tfrac{s}{\theta} \times 1 + \tfrac{n-s}{1-\theta} \times (-1) \\ &= \tfrac{s}{\theta} - \tfrac{n-s}{1-\theta}\end{aligned}$$

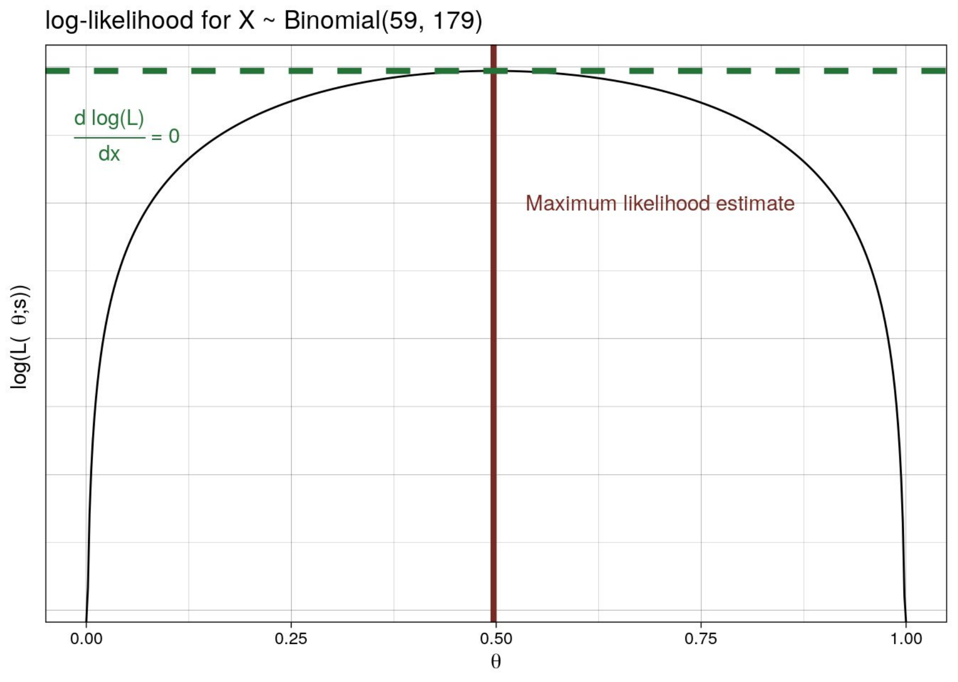
Setting this to zero we get
$\frac{s}{\theta} = \frac{n-s}{1-\theta} \rightarrow s(1-\theta) = \theta(n-s) \rightarrow s - s\theta = \theta n - s\theta \rightarrow s + (s\theta - s\theta) = \theta n \rightarrow \theta = \frac{s}{n}$.
Therefore, as above

$$\hat{\theta} = \frac{s}{n}.$$

The binomial log-likelihood function (finding the MLE)

```r
log_likelihood <- function(theta) dbinom(x = 79, size = 159, prob = theta, log = TRUE)
```

```r
optimise(log_likelihood, c(0,1), maximum = TRUE)
```



The Poisson likelihood function (finding the MLE)

Suppose that $x_1, \ldots, x_n$ are iid observations from a Poisson distribution with unknown parameter λ:

$$L(\lambda\,; x_1, \ldots, x_n) = K e^{-n\lambda} \lambda^{n\bar{x}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $K = \prod_{i=1}^{n} \frac{1}{x_i\,!}$ is a constant that doesn't depend on λ.

We differentiate $L(\lambda\,; x_1, \ldots, x_n)$ and set to 0 to find the MLE:

$$\begin{aligned} 0 &= \frac{\delta}{\delta\lambda} L(\lambda\,; x_1, \ldots, x_n) \\ &= K\left(-n e^{-n\lambda} \lambda^{n\bar{x}} + n\bar{x} e^{-n\lambda} \lambda^{n\bar{x}-1}\right) \\ &= K e^{-n\lambda} \lambda^{n\bar{x}-1} \left(-n\lambda + n\bar{x}\right) \end{aligned}$$

$\rightarrow \lambda = \infty, \lambda = 0, \text{or } \lambda = \bar{x}.$

The Poisson log-likelihood function (finding the MLE)

Suppose that $x_1, \ldots, x_n$ are iid observations from a Poisson distribution with unknown parameter λ:

$$L(\lambda\,; x_1, \ldots, x_n) = K e^{-n\lambda} \lambda^{n\bar{x}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $K = \prod_{i=1}^{n} \frac{1}{x_i\,!}$ is a constant that doesn't depend on λ.

$$\begin{aligned}
\log(L(\lambda\,; x_1, \ldots, x_n)) &= \textstyle\sum_{i=1}^{n} \log(\frac{\lambda^{x_i}}{x_i\,!} e^{-\lambda}) \\
&= \textstyle\sum_{i=1}^{n} \log(\frac{1}{x_i\,!}) + \log(\lambda^{x_i}) + \log(e^{-\lambda}) \\
&= \textstyle\sum_{i=1}^{n} \log(\frac{1}{x_i\,!}) + x_i \log(\lambda) + (-\lambda) \\
&= K' + \log(\lambda) \textstyle\sum_{i=1}^{n} x_i - n\lambda \quad \text{where } K' \text{ is a constant} \\
&= K' + \log(\lambda) n\bar{x} - n\lambda.
\end{aligned}$$

Differentiate and set to 0 for the MLE:

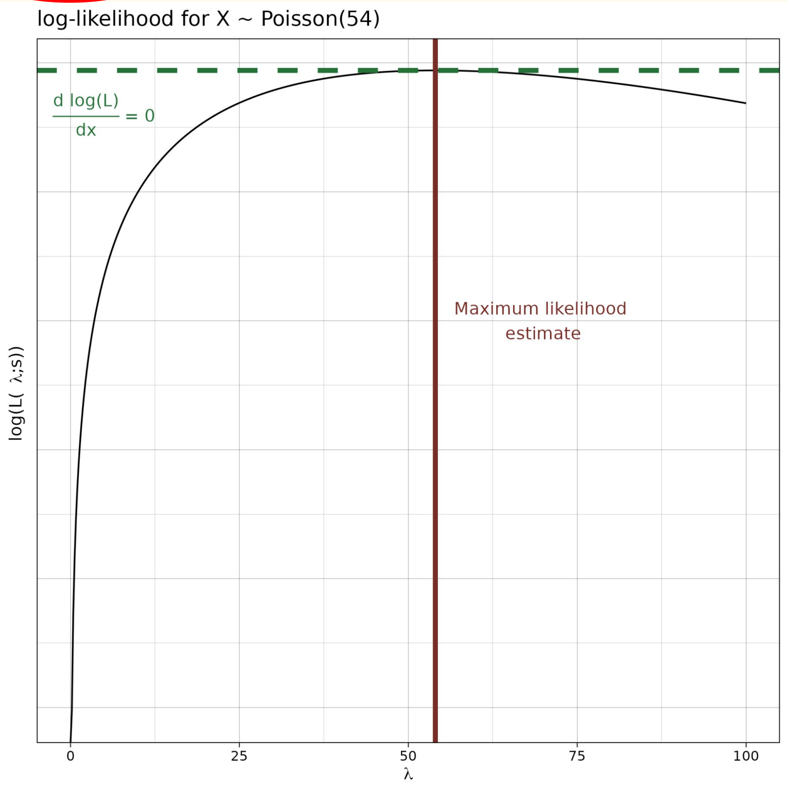
$$\begin{aligned}
0 &= \frac{\delta}{\delta\lambda} \log(L(\lambda\,; x_1, \ldots, x_n)) \\
&= \frac{\delta}{\delta\lambda} (K' + \log(\lambda) n\bar{x} - n\lambda) \\
&= \frac{n\bar{x}}{\lambda} - n
\end{aligned}$$

assuming a unique maximum in $0 < \lambda < \infty$ the MLE is $\hat{\lambda} = \bar{x}$ as before.

The Poisson log-likelihood function (finding the MLE)

```
log_likelihood <- function(lambda) dpois(x = 54, lambda, log = TRUE)

optimise(log_likelihood, c(0,100), maximum = TRUE)
```



In summary

1. The likelihood function tells us the relative probability that a given set of population parameters has generated the data
2. Maximum Likelihood Estimation is a common way to estimate parameters in GLM
3. It’s a very flexible technique, and can be applied to many different distributions

Typically, we also need the second derivative (*the rate of change of the rate of change*) of the likelihood.

- When it is positive, the likelihood is convex, so we reached a “valley” rather than a peak
- When it is negative, it confirms that the likelihood is concave, and we reached a maximum
- We use second derivatives to compute the standard errors:
    - The second derivative is a measure of the curvature of a function. The steeper the curve, the more certain we are about our estimates
    - The matrix of second derivatives is called “Hessian”
    - The inverse of the Hessian matrix is the variance-covariance matrix of the estimates
    - The standard errors of ML estimates are the square root of the diagonal entries of this matrix

BUT what if it's a relationship we're after?

Recall, with $L(\theta; s) = \binom{n}{s}\theta^s(1-\theta)^{n-s}$

$$\begin{aligned}\log(L(\theta; s)) &= \log(\tbinom{n}{s}) + \log(\theta^s) + \log((1-\theta)^{n-s}) \\ &= \log(\tbinom{n}{s}) + s\log(\theta) + (n-s)\log(1-\theta).\end{aligned}$$

Differentiating this:

$$\begin{aligned}\frac{\delta \log(L(\theta;s))}{\delta\theta} &= 0 + \tfrac{s}{\theta} \times 1 + \tfrac{n-s}{1-\theta} \times (-1) \\ &= \tfrac{s}{\theta} - \tfrac{n-s}{1-\theta}\end{aligned}$$

Setting this to zero we get

$$\hat{\theta} = \frac{s}{n}.$$

So what if our parameter θ were now a linear equation? Something like $\alpha + \beta\mathbf{x}$...

BUT what if it's a relationship we're after?

The parameters we want to estimate now are α and β. First remember that for a Binomial model θ is a probability and therefore $0 < \theta < 1$. To ensure that this holds we use a link function, typically the logit link function:

$$\text{logit}(\theta) = \log(\frac{\theta}{1-\theta}) = \alpha + \beta\mathbf{x}$$

Rearranging

$$\frac{\theta}{1-\theta} = \exp(\alpha + \beta\mathbf{x})$$

and

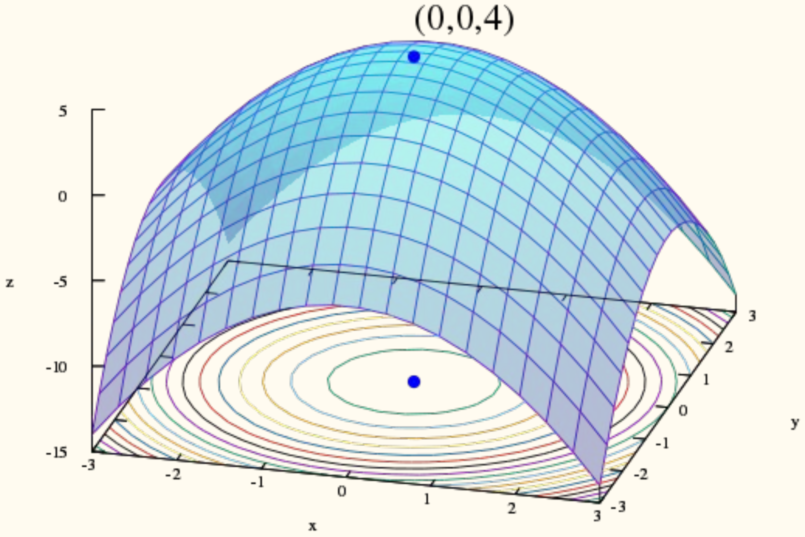
$$\theta = \frac{\exp(\alpha + \beta\mathbf{x})}{1 + \exp(\alpha + \beta\mathbf{x})}.$$

So, now our log-likelihood function is

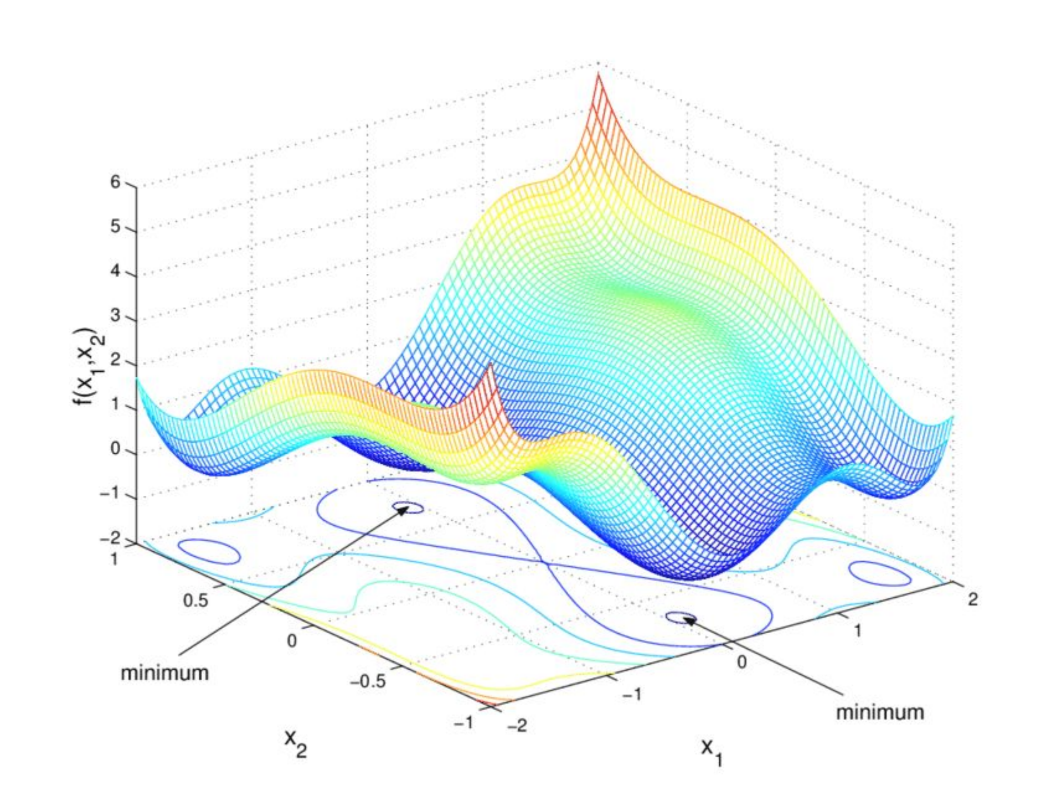
$$\begin{aligned}
\log(L(\alpha, \beta, ; s)) &= \log(\tbinom{n}{s}) + s\log(\theta) + (n-s)\log(1-\theta) \\
&= \log(\tbinom{n}{s}) + s\log\left(\tfrac{\exp(\alpha+\beta\mathbf{x})}{1+\exp(\alpha+\beta\mathbf{x})}\right) + (n-s)\log\left(1 - \tfrac{\exp(\alpha+\beta\mathbf{x})}{1+\exp(\alpha+\beta\mathbf{x})}\right) \\
&= \log(\tbinom{n}{s}) + s\log(\exp(\alpha+\beta x)) - s\log(1+\exp(\alpha+\beta x)) + (n-s)\log\left(\tfrac{1}{1+\exp(\alpha+\beta\mathbf{x})}\right) \\
&= \log(\tbinom{n}{s}) + s(\alpha+\beta x) - s\log(1+\exp(\alpha+\beta x)) + (s-n)\log\left(1+\exp(\alpha+\beta\mathbf{x})\right) \\
&= \log(\tbinom{n}{s}) + s(\alpha+\beta x) - n\log\left(1+\exp(\alpha+\beta x)\right)
\end{aligned}$$

- We can look for the maximum likelihood by taking the partial derivative of the equation in respect to, say, β, and setting it to 0
- Usually done iteratively:
  - Choose some arbitrary starting values of β
  - Evaluate the vector of partial derivatives of the log-likelihood function
  - Update the values of β using the information given by the partial derivatives
  - Stop when we reach values sufficiently close to 0
- And then we do this again for the intercept…
- There are several other optimization algorithms (not discussed in this course)
- The good news is, there's R!

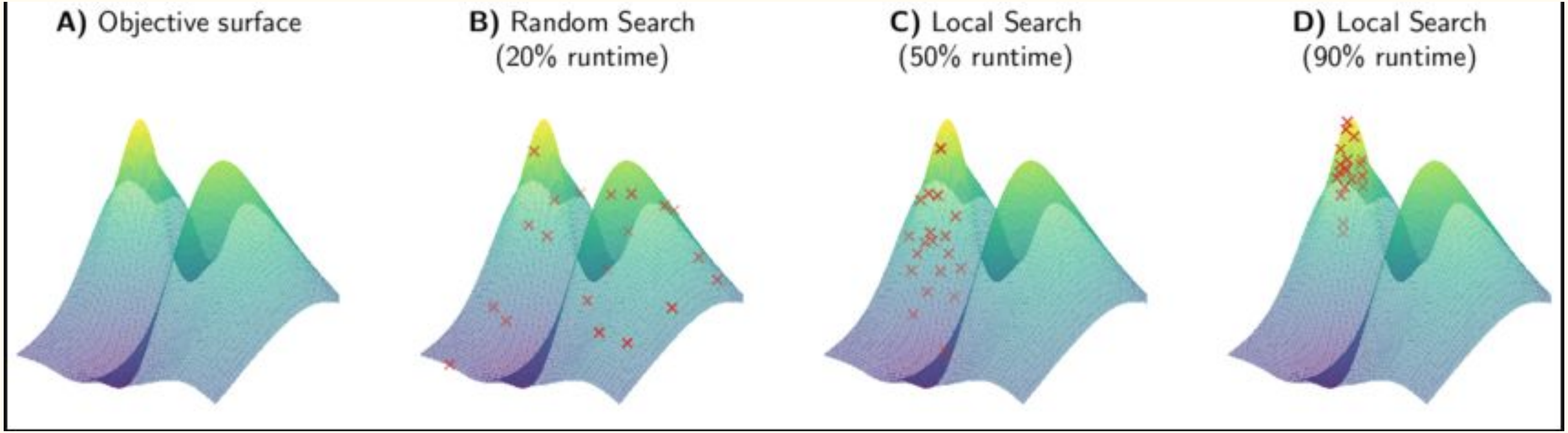
Issues in 2D



Issues in 2D



Issues in 2D


A) Objective surface
B) Random Search
(20% runtime)
C) Local Search
(50% runtime)
D) Local Search
(90% runtime)