Maximum Likelihood

Patrick Lam

Maximum Likelihood

Maximum Likelihood

Suppose we have some data y and we want to find some parameters θ that generated y.

Maximum Likelihood

Suppose we have some data y and we want to find some parameters θ that generated y.

Maximum likelihood is a way to find θ.

Derived from Bayes' Rule

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
p(\theta|y) &= p(y|\theta)k(y)
\end{aligned}
$$

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ p(\theta|y) &= p(y|\theta)k(y) \\ p(\theta|y) &\propto p(y|\theta) \end{aligned}$$

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
p(\theta|y) &= p(y|\theta)k(y) \\
p(\theta|y) &\propto p(y|\theta)
\end{aligned}
$$

$$
L(\theta|y) = p(y|\theta)
$$

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
p(\theta|y) &= p(y|\theta)k(y) \\
p(\theta|y) &\propto p(y|\theta)
\end{aligned}
$$

$$
L(\theta|y) = p(y|\theta)
$$

The likelihood function is mathematically the same as the distribution for y.

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
p(\theta|y) &= p(y|\theta)k(y) \\
p(\theta|y) &\propto p(y|\theta)
\end{aligned}
$$

$$
L(\theta|y) = p(y|\theta)
$$

The likelihood function is mathematically the same as the distribution for y.

Our best estimate of θ (MLE) is the value of θ that maximizes the likelihood function.

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
p(\theta|y) &= p(y|\theta)k(y) \\
p(\theta|y) &\propto p(y|\theta)
\end{aligned}$$

$$L(\theta|y) = p(y|\theta)$$

The likelihood function is mathematically the same as the distribution for y.

Our best estimate of θ (MLE) is the value of θ that maximizes the likelihood function. Why?

Derived from Bayes' Rule

Assume that y follows some distribution $p(y|\theta)$.

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
p(\theta|y) &= p(y|\theta)k(y) \\
p(\theta|y) &\propto p(y|\theta)
\end{aligned}
$$

$$L(\theta|y) = p(y|\theta)$$

The likelihood function is mathematically the same as the distribution for y.

Our best estimate of θ (MLE) is the value of θ that maximizes the likelihood function. Why? and why?

Why Does $L(\theta|y) = p(y|\theta)$ work?

Why Does $L(\theta|y) = p(y|\theta)$ work?

Suppose we have one observation $y = 0.5$ (assumed to be) drawn from a Normal distribution with $\sigma^2 = 1$.

Why Does $L(\theta|y) = p(y|\theta)$ work?

Suppose we have one observation $y = 0.5$ (assumed to be) drawn from a Normal distribution with $\sigma^2 = 1$.

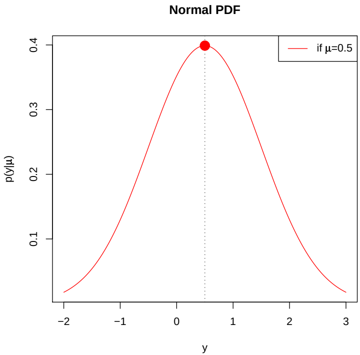
Estimate the mean μ of the distribution.

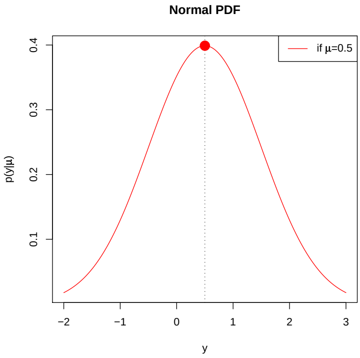
Why Does $L(\theta|y) = p(y|\theta)$ work?

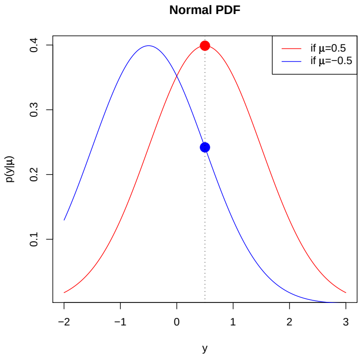
Suppose we have one observation $y = 0.5$ (assumed to be) drawn from a Normal distribution with $\sigma^2 = 1$.

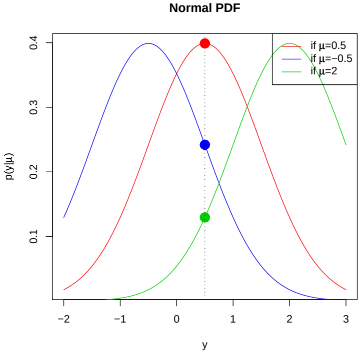
Estimate the mean μ of the distribution.

Obvious guess: $\mu = 0.5$

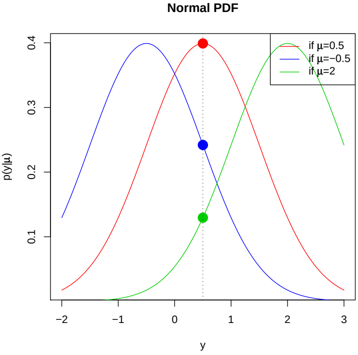

Normal PDF
if μ=0.5
p(y|μ)
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
y

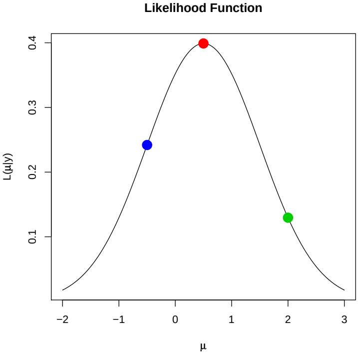

Normal PDF
if μ=0.5
p(y|μ)
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
y

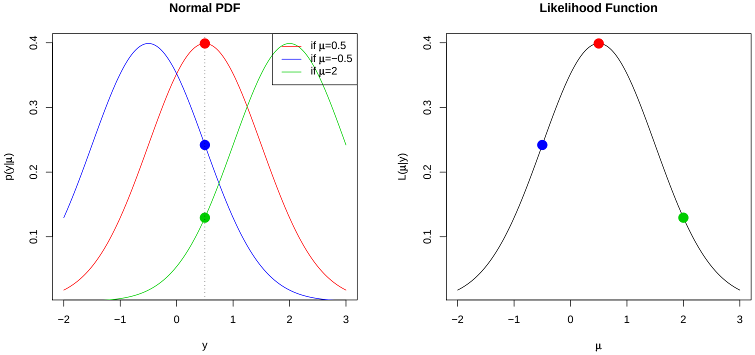

Normal PDF
if μ=0.5
if μ=−0.5
p(y|μ)
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
y



Normal PDF
if μ=0.5
if μ=−0.5
if μ=2
p(y|μ)
y
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3


Normal PDF
if μ=0.5
if μ=−0.5
if μ=2
p(y|μ)
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
y


Likelihood Function
L(μ|y)
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
μ



Our best estimate of θ is the value of θ that maximizes $L(\theta|y)$ (MLE).

Example: Poisson Distribution

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

Example: Poisson Distribution

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).

Example: Poisson Distribution

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).

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

Example: Poisson Distribution

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).

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

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

Example: Poisson Distribution

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).

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

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

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

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.

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.

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

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).

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.

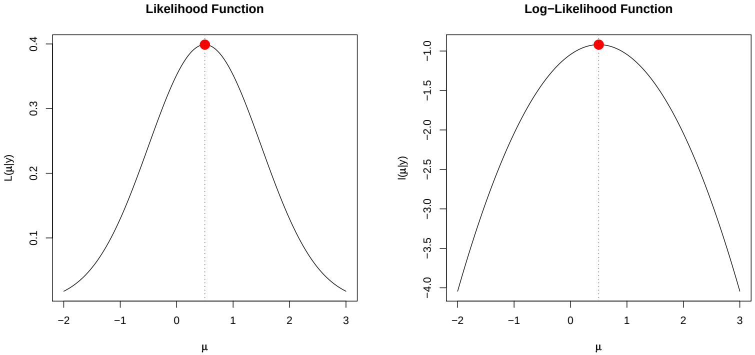
$$l(\lambda|\mathbf{y}) = \sum_{i=1}^{n} (y_i \ln\lambda - \lambda - \ln y_i!)$$

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?

Why Can We Use the Log-likelihood?



Finding the Maximum Likelihood Estimate (MLE)

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.

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.

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

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.

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

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

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.

$$\frac{\partial l(\lambda|\mathbf{y})}{\partial \lambda} = S(\theta) = \frac{\sum_{i=1}^{n} y_i}{\lambda} - n$$

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.

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

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

Maximum Likelihood In R

Write our log-likelihood function:

Maximum Likelihood In R

Write our log-likelihood function:

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

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)
+ }
```

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)

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

[1] 4.954001
```

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)

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

[1] 4.954001
```

What the exp()?

Reparameterization

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

We also need to reparameterize the output from `optim()`.

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

We also need to reparameterize the output from `optim()`.

Likelihood is invariant to reparameterization.

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

We also need to reparameterize the output from `optim()`.

Likelihood is invariant to reparameterization.

Common reparameterizations:

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

We also need to reparameterize the output from `optim()`.

Likelihood is invariant to reparameterization.

Common reparameterizations:

- To constrain to positive space:

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

We also need to reparameterize the output from `optim()`.

Likelihood is invariant to reparameterization.

Common reparameterizations:

- To constrain to positive space: `exp()`

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

We also need to reparameterize the output from `optim()`.

Likelihood is invariant to reparameterization.

Common reparameterizations:

- To constrain to positive space: `exp()`
- To constrain to $[0, 1]$,

Reparameterization

We need to reparameterize the parameters in our function to constrain the search space.

We also need to reparameterize the output from `optim()`.

Likelihood is invariant to reparameterization.

Common reparameterizations:

- To constrain to positive space: `exp()`
- To constrain to $[0, 1]$, use a cdf

Cumulative Distribution Function

Cumulative Distribution Function

We've learned how to define a distribution via is PDF or PMF f(x).

Cumulative Distribution Function

We've learned how to define a distribution via is PDF or PMF f(x).

Every distribution also has a CDF F(x):

Cumulative Distribution Function

We've learned how to define a distribution via is PDF or PMF f(x).

Every distribution also has a CDF F(x): $P(X \leq x)$

Cumulative Distribution Function

We've learned how to define a distribution via is PDF or PMF f(x).

Every distribution also has a CDF F(x): $P(X \leq x)$

For discrete case:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$$

Cumulative Distribution Function

We've learned how to define a distribution via is PDF or PMF f(x).

Every distribution also has a CDF F(x): $P(X \leq x)$

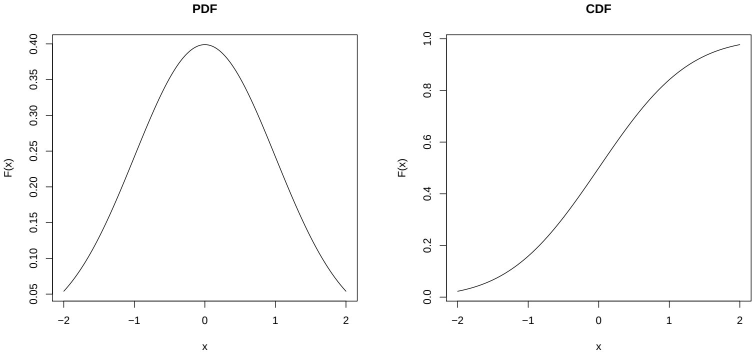
For discrete case:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$$

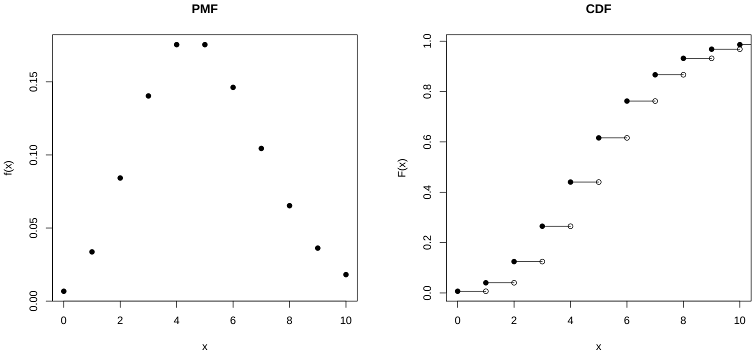
For continuous case:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Standard Normal PDF and CDF



Poisson(5) PMF and CDF



We can use a CDF to transform numbers from $(-\infty, \infty)$ to $[0, 1]$.

We can use a CDF to transform numbers from $(-\infty, \infty)$ to $[0, 1]$.

Commonly used:

We can use a CDF to transform numbers from $(-\infty, \infty)$ to $[0, 1]$.

Commonly used:

- Standard Normal CDF: $\Phi(\cdot)$

We can use a CDF to transform numbers from $(-\infty, \infty)$ to $[0, 1]$.

Commonly used:

- Standard Normal CDF: $\Phi(\cdot)$
- Logistic CDF: $\frac{1}{1+e^{-x}}$

We can use a CDF to transform numbers from $(-\infty, \infty)$ to $[0, 1]$.

Commonly used:

- Standard Normal CDF: $\Phi(\cdot)$
- Logistic CDF: $\frac{1}{1+e^{-x}}$

Why shouldn't we use a CDF of a discrete variable?

We can use a CDF to transform numbers from $(-\infty, \infty)$ to $[0, 1]$.

Commonly used:

- Standard Normal CDF: $\Phi(\cdot)$
- Logistic CDF: $\frac{1}{1+e^{-x}}$

Why shouldn't we use a CDF of a discrete variable?

What about the CDF of a bounded variable?

More Complicated Likelihoods

More Complicated Likelihoods

Up to this point, our distributions have been relatively simple (i.e. all observations are i.i.d. from a distribution with the same parameter).

More Complicated Likelihoods

Up to this point, our distributions have been relatively simple (i.e. all observations are i.i.d. from a distribution with the same parameter).

Suppose our observations come from different distributions, or some observations are not fully observed.

More Complicated Likelihoods

Up to this point, our distributions have been relatively simple (i.e. all observations are i.i.d. from a distribution with the same parameter).

Suppose our observations come from different distributions, or some observations are not fully observed.

How do we define the joint distribution for our data (and thus our likelihood)?

More Complicated Likelihoods

Up to this point, our distributions have been relatively simple (i.e. all observations are i.i.d. from a distribution with the same parameter).

Suppose our observations come from different distributions, or some observations are not fully observed.

How do we define the joint distribution for our data (and thus our likelihood)?

Use an indicator variable.

Indicator Variable

Let D be an indicator variable such that $d_i = 1$ if i follows one distribution, and $d_i = 0$ if i follows another distribution.

Indicator Variable

Let D be an indicator variable such that $d_i = 1$ if i follows one distribution, and $d_i = 0$ if i follows another distribution.

We can incorporate the indicator variable into our likelihood.

Example 1

Example 1

Suppose that we have 5 observations. The first 2 are assumed to be from distribution 1. The last 3 are assumed to be from distribution 2.

Example 1

Suppose that we have 5 observations. The first 2 are assumed to be from distribution 1. The last 3 are assumed to be from distribution 2.

$$d = c(1, 1, 0, 0, 0)$$

Example 1

Suppose that we have 5 observations. The first 2 are assumed to be from distribution 1. The last 3 are assumed to be from distribution 2.

$$
\begin{aligned}
d &= c(1,1,0,0,0) \\
p(\theta_1, \theta_2 | \mathbf{y}) &= \prod_{i=1}^{n} [p(y_i|\theta_1)]^{d_i} [p(y_i|\theta_2)]^{1-d_i}
\end{aligned}
$$

Example 2

Example 2

Suppose that we have 5 observations. We observe the first 2 observations completely, but we only observe that the last 3 are greater than some number z.

Example 2

Suppose that we have 5 observations. We observe the first 2 observations completely, but we only observe that the last 3 are greater than some number z.

$$d = c(1,1,0,0,0)$$

Example 2

Suppose that we have 5 observations. We observe the first 2 observations completely, but we only observe that the last 3 are greater than some number z.

$$\begin{aligned} d &= c(1,1,0,0,0) \\ p(\theta|\mathbf{y}) &= \prod_{i=1}^{n} [p(y_i|\theta)]^{d_i} [1 - F(z)]^{1-d_i} \end{aligned}$$

where $F(y)$ is the CDF for $p(y|\theta)$.

Example 2

Suppose that we have 5 observations. We observe the first 2 observations completely, but we only observe that the last 3 are greater than some number z.

$$
\begin{aligned}
d &= c(1,1,0,0,0) \\
p(\theta|\mathbf{y}) &= \prod_{i=1}^{n} [p(y_i|\theta)]^{d_i} [1 - F(z)]^{1-d_i}
\end{aligned}
$$

where $F(y)$ is the CDF for $p(y|\theta)$. Remember

$$F(z) = P(Y \leq z)$$