# Likelihood Review

Gavin Fay

09/07/2023

# Objectives

- Present Likelihood
- Statistical inference based on maximum likelihood

# Probability

Often dicsuss these in terms of PDFs, $f(x) = P(X = x)$

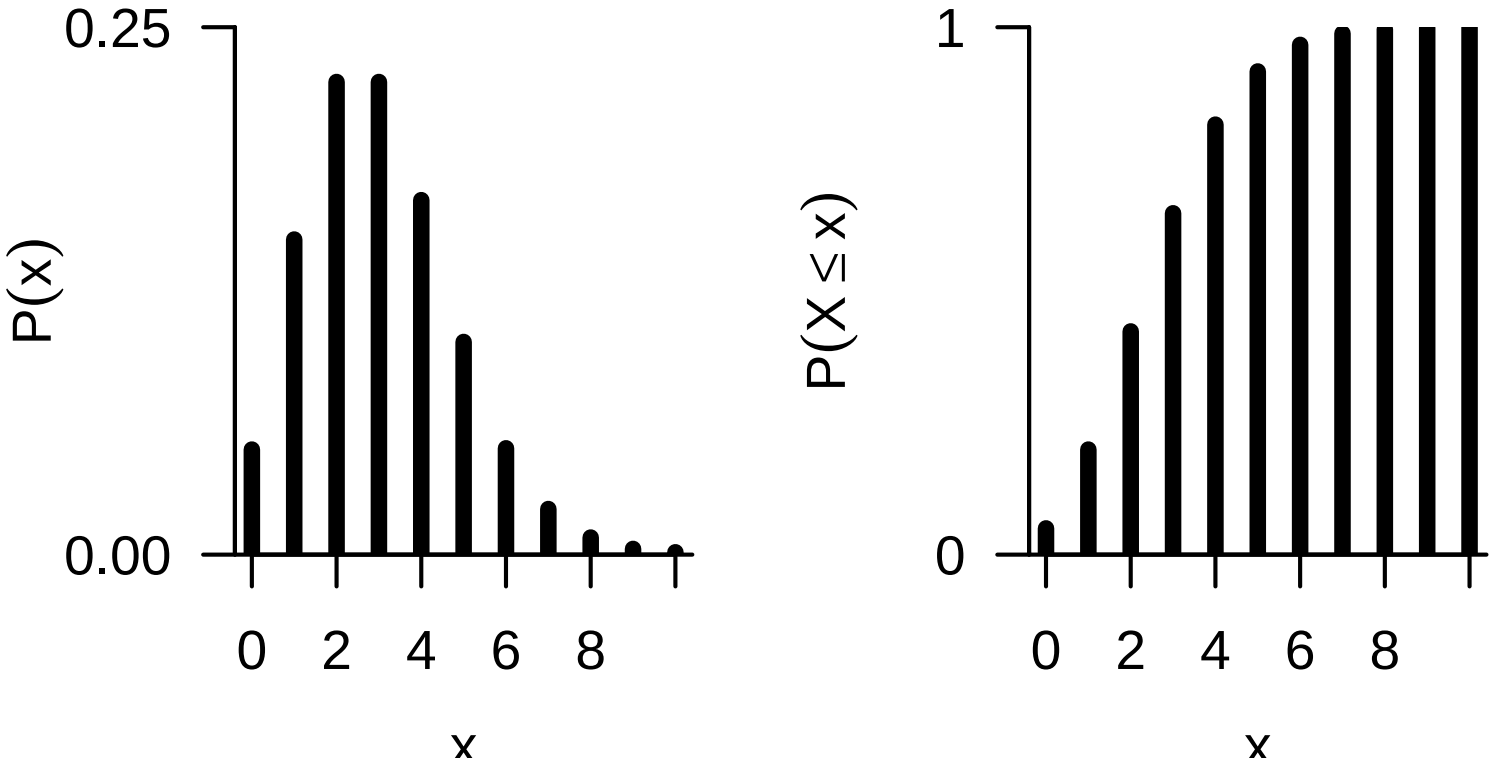
- Probability distribution function (discrete)
- Probability density function (continuous)

PDFs are a function of parameters.

We are frequently interested in estimating the values for parameters.

# Example: Poisson $\lambda = 3$

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$




Often used for count data

# Methods of estimation

- Method of Moments
- Least Squares
- Maximum Likelihood

Other techniques:
minimum chi-square, minimum mean-square error estimators, finite sampling theory, and Bayes procedures.

Method of Moments: equate sample results with their expected values under a sampling model.
Expected value(s) in turn is a function of the parameter(s) of interest in estimating.

# Likelihood function

Common problem:

Given some data, and a model of interest, find the one PDF among all the probability densities that the model prescribes, that is most likely to have produced the data.
**inverse problem**

i.e. Try to find values for parameters such that the model for $P(X)$ yields numbers that are as close to the data as possible.

We choose parameter estimates to **maximize** the likelihood function.

In linear regression problem, least squares approach is a special case of maximum likelihood.

Minimize distance between model and data.
e.g. sum of squares in linear regression problem.

# Likelihood function

Define the **likelihood function** by reversing roles of data vector and the parameter vector.

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{f}(\mathbf{y}|\boldsymbol{\theta}_\mathbf{i})$$

Likelihood functions is defined as the pdf of some given data over the alternative values for the parameters.

Sometimes referred to as: likelihood of a parameter given the data (because we know the data).

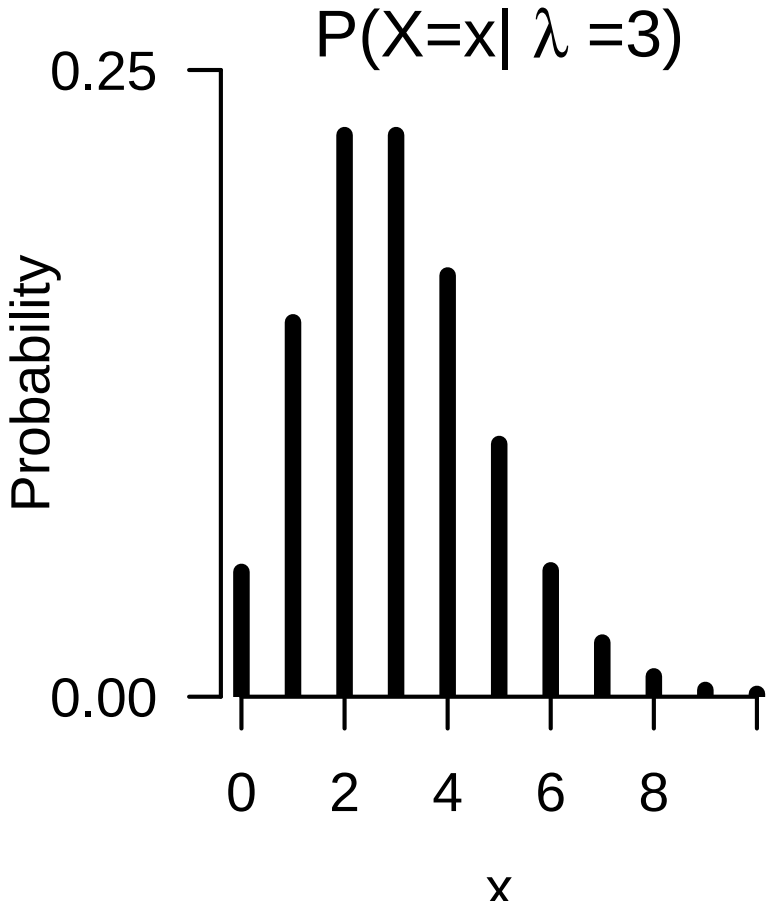
# Example: Likelihood function for poisson

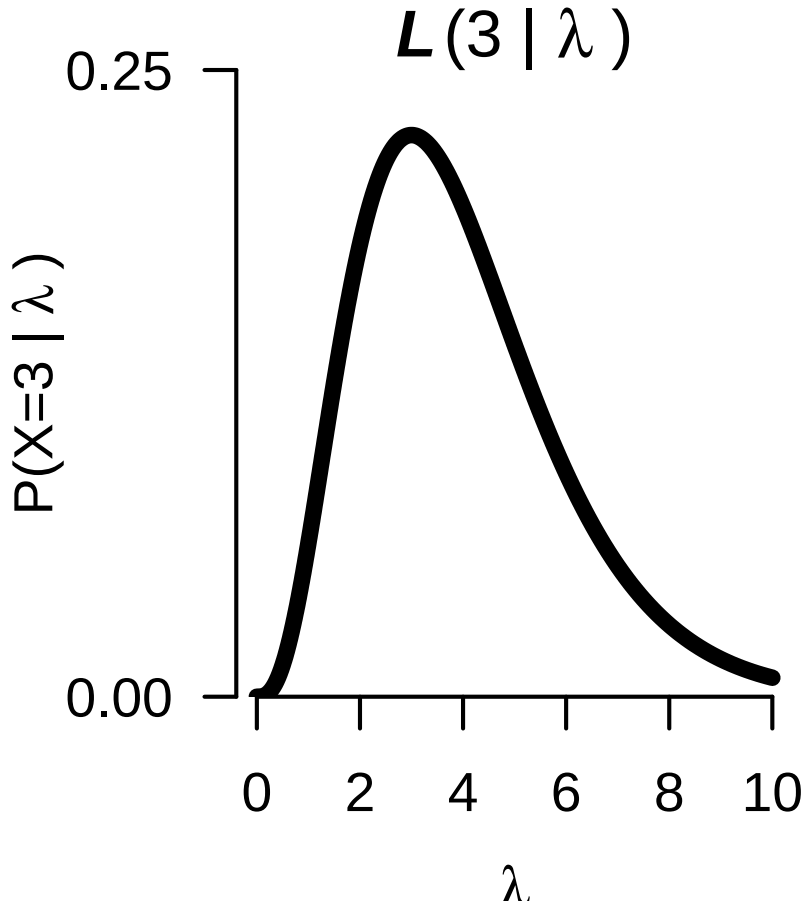
Counting river herring at Jenney Grist Mill, Plymouth MA.

We observe at the fish ladder for 1 hour and count 3 herring.

# Example: Likelihood function for poisson

To define the likelihood function we evaluate $P(X = 3|\lambda)$ for all values of $\lambda$.

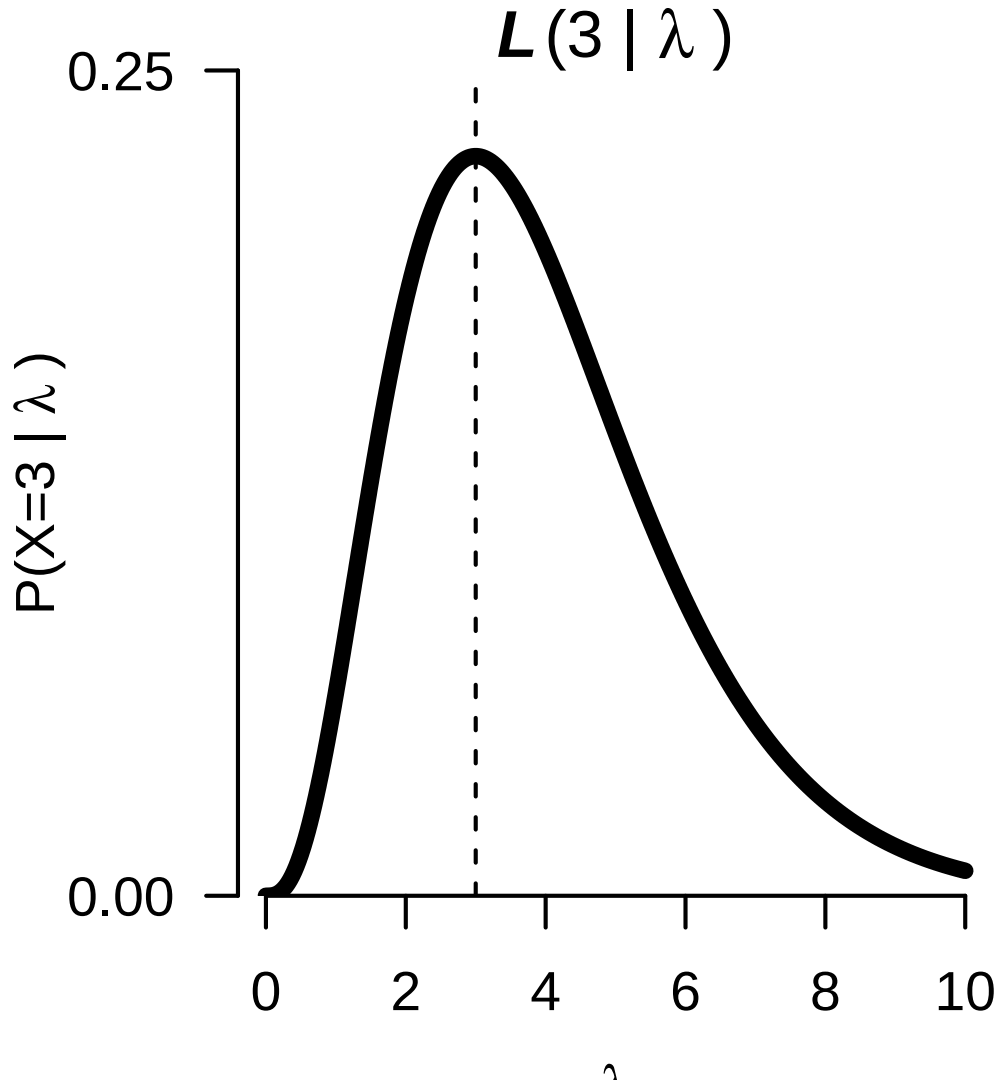


# Maximum likelihood estimation

What is the arrival rate of herring per hour?

We find the value of the parameter $\lambda$ that **maximizes** the likelihood function.



# Maximum likelihood estimation

Maximum of a function $f(x)$?
- 1st derivative is equal to 0
- Function is convex around maximum (2nd derivative is negative)

We often work with the log-likelihood ($\ln\mathcal{L}$).

Procedure:
- Differentiate log-likelihood with respect to the parameters,
- Set 1st derivative equal to zero,
- Solve for the parameters to find the MLEs.

# Maximum likelihood estimation: Poisson

$$\mathcal{L}(\lambda; y) = \frac{\mathrm{e}^{-\lambda}\lambda^{y}}{y!}$$

$$\ln(\mathcal{L}(\lambda; y)) =$$

$$\frac{\mathrm{d}\ln\mathcal{L}}{\mathrm{d}\lambda} =$$

# Maximum likelihood estimation: Poisson

$$\mathcal{L}(\lambda; y) = \frac{\mathrm{e}^{-\lambda}\lambda^{y}}{y!}$$

$$\ln(\mathcal{L}(\lambda; y)) = -\lambda + y\ln\lambda - \ln(y!)$$

$$\frac{\mathrm{d}\ln\mathcal{L}}{\mathrm{d}\lambda} =$$

# Maximum likelihood estimation: Poisson

$$\mathcal{L}(\lambda; y) = \frac{\mathrm{e}^{-\lambda}\lambda^{y}}{y!}$$

$$\ln(\mathcal{L}(\lambda; y)) = -\lambda + y\ln\lambda - \ln(y!)$$

$$\frac{\mathrm{d}\ln\mathcal{L}}{\mathrm{d}\lambda} = -1 + \frac{y}{\lambda}$$

# Maximum likelihood estimation

Our example for poisson:

Set $\frac{\mathrm{dln}\mathcal{L}}{\mathrm{d}\lambda} = 0$, and solve for $\lambda$ with $y = 3$.

$0 = -1 + \frac{y}{\lambda}$

MLE for $\lambda = 3$

# Maximum likelihood estimation

Our example for poisson:

Set $\frac{\mathrm{d}\ln\mathcal{L}}{\mathrm{d}\lambda} = 0$, and solve for $\lambda$ with $y = 3$.

MLE for $\lambda = 3$.

Say we had $n$ independent hourly observations of herring arrival:

Recall for independent events:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2) \cdot P(A_3) \ldots$$

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

$$\ln(\mathcal{L}(\lambda; \mathbf{y})) = -n\lambda + \ln\lambda \sum_{i=1}^{n} y - \sum_{i=1}^{n} \ln(y!)$$

# Maximum likelihood estimation

Principle developed by R.A. Fisher (1920s)

MLE estimates need not exist nor be unique.

The likelihood equation represents a necessary condition for the existence of an MLE estimate.

Not usually possible to obtain analytic form solutions for the MLE estimates.

Particularly when the model involves many parameters and PDF is highly non-linear.

In which case:
MLE estimate sought numerically using nonlinear optimization algorithms.

# Estimating variance for your parameters

Fisher showed that the negative inverse of the 2nd partial derivative of the log-likelihood function (the negative inverse of the Hessian), evaluated at the MLE, is the MLE of the variance of the parameter.

$$Var(\hat{\theta}) = \left[ - \left( \frac{\partial^2 \ln L(\theta; \mathbf{y})}{\partial \theta^2} \right) \right]^{-1}_{\theta = \hat{\theta}}$$

Returning to our herring example with 1 observation of 3 fish in an hour:

$$Var(\hat{\lambda}) = \left[ - \left( \frac{-y}{\lambda^2} \right) \right]^{-1}_{\lambda = \hat{\lambda}} = \left[ - \left( \frac{-3}{3^2} \right) \right]^{-1} = 3$$

So with only 1 observation, our variance is pretty high.
(indeed, the same as the variance of the Poisson process!)

# Relationship to least squares estimation

Least squares assumptions:

1. $y_i$'s are pair-wise uncorrelated
2. variance constant, $\sigma^2$
3. Expected value of $y_i$, $E[y_i] = \beta_0 + \beta_1 x_i$

No statistical or distributional assumptions required for parameter estimates.

Maximum Likelihood technique of least squares assumptions:

1. $y_i$'s follow the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
2. $\epsilon_i \sim N(0, \sigma^2)$
3. $\sigma^2$ known or unknown parameter

# Regression example, normal Likelihood

$$\mathcal{L}(\mathbf{y}|\beta_0, \beta_1, \sigma^2, \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2\sigma^2}(y_i-(\beta_0+\beta_1 x_i))^2}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2}$$

$$\ln\mathcal{L} = n\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2$$

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\beta_0} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))$$

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\beta_1} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))x_i$$

# Regression example, normal Likelihood

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\beta_0} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))$$

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\beta_1} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))x_i$$

Let $\bar{X} = \sum x_i/n$, $\bar{Y} = \sum x_i/n$, $\overline{X^2} = \sum x_i^2/n$, $\overline{XY} = \sum x_i y_i/n$

Then

$\beta_0 + \beta_1\bar{X} = \bar{Y}$ and $\beta_0\bar{X} + \beta_1\overline{X^2} = \overline{XY}$

Solving for $\beta_1$ & $\beta_0$ gives

$\beta_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}$, and $\beta_0 = \bar{Y} - \beta_1\bar{X}$

which are the same as the least square estimates

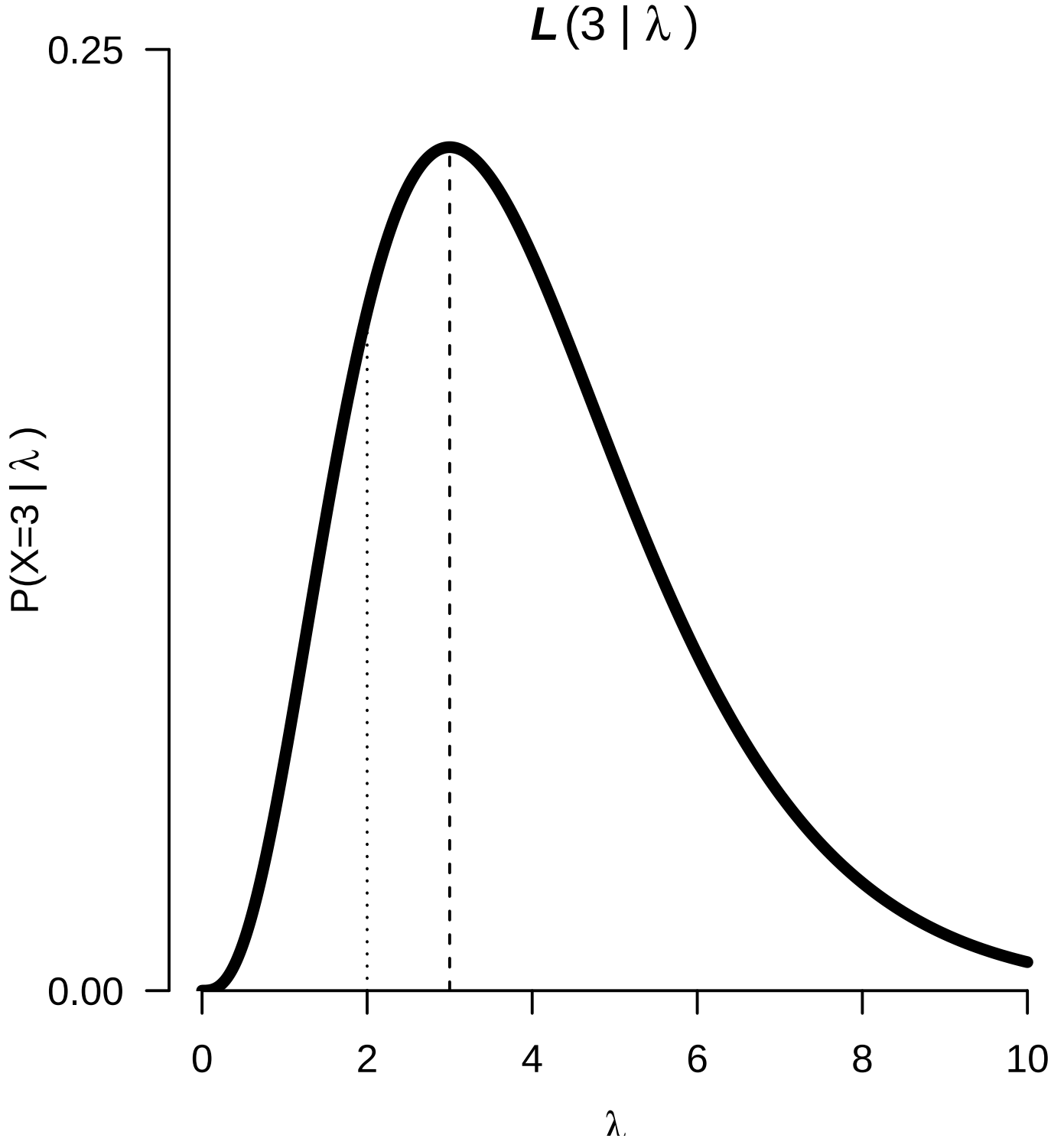
# Maximum likelihood and statistical testing

So far, we calculated the 'best' estimate for the parameters (MLEs).

We are often also interested in evidence for 1 hypothesis over another.
i.e. to compare the fits of different models.

For nested models, the **likelihood ratio** gives the weight of evidence for one model over another.

herring example: How less likely is it that we counted 3 fish when the mean arrival rate is 2?


L(3 | λ)
0.25
0.00
P(X=3 | λ)
0
2
4
6
8
10
λ

# Likelihood ratios

Herring example: How less likely is it that we counted 3 fish when the mean arrival rate is 2?

$$\begin{aligned}\Lambda(x) &= \frac{L(y|\lambda = 2)}{L(y|\lambda = 3)} \\ &= \frac{P(y = 3|\lambda = 2)}{P(y = 3|\lambda = 3)} \\ &= \frac{0.180}{0.224} = 0.805\end{aligned}$$

# Likelihood ratios

*Statistical testing*

Wilks showed that as $n \to \infty$, $-2\ln(\Lambda)$ is asymptotically chi-squared distributed ($\chi^2$) with degrees of freedom equal to difference in dimensionality between the models.

$$\begin{aligned} -2\ln(\Lambda) &= -2(\ln L(y|\lambda = 2) - \ln L(y|\lambda = 3)) \\ &= -2(\ln(0.180) - \ln(0.224)) \\ &= 0.433 \end{aligned}$$

Critical value of $\chi^2$ with 1 d.f. at $\alpha = 0.05$ is 3.84

$0.433 < 3.84$ so we would not reject the hypothesis that the data were generated by a poisson with $\lambda$ of 2, even though it is less likely than the MLE of 3.

# Likelihood profiles

We can compute a 95% confidence interval for MLEs using the **likelihood profile**

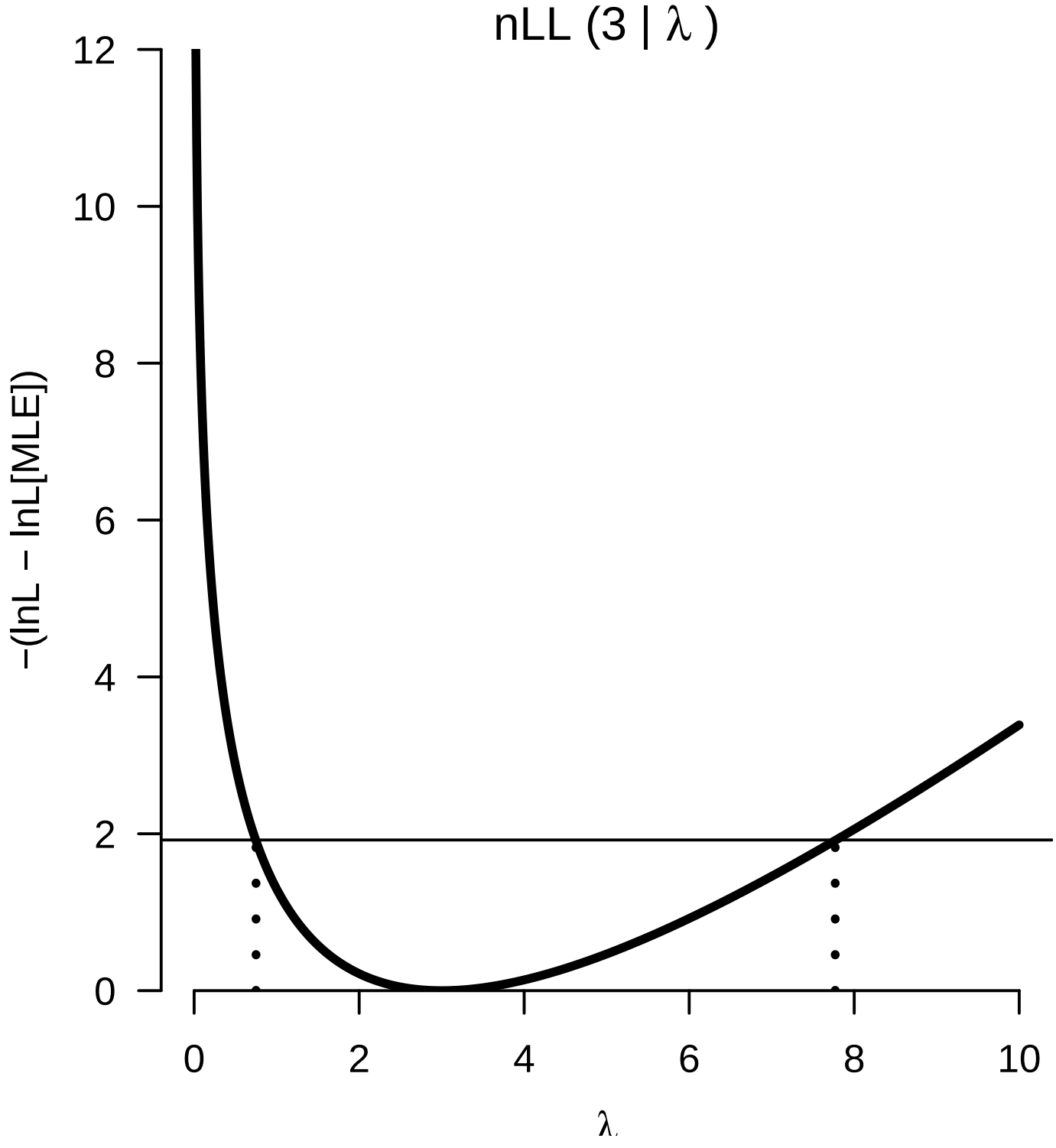
An $100 - x\%$ confidence interval for $p$ parameters is determined by finding the values for the parameter(s) for which:

$$-(\ln L - \ln L_{MLE}) = \frac{1}{2}\chi^2(p, x/100)$$

where $\ln L_{MLE}$ is the log-likelihood corresponding to the maximum likelihood estimates.

If the number of parameters is $> 1$ then we profile the likelihood over a parameter using the MLEs for the other parameters given the profiled values for the parameter of interest.

i.e. re-fit the model to the data for each of the profiled values for the parameter of interest.


nLL (3 | λ )
−(lnL − lnL[MLE])
12
10
8
6
4
2
0
0
2
4
6
8
10
λ

# Properties of maximum likelihood estimators

**1.** For the estimator of $\theta$, the

$$MLE \sim N\left(\theta, \frac{-1}{E\left[\frac{\delta^2 \ln L}{\delta \theta^2}\right]}\right)$$

**2.** MLEs are solutions to the equations

$$\frac{\delta \mathcal{L}(\theta)}{\delta \theta} \stackrel{set}{=} 0$$

which satisfy the relationship

$$\frac{\delta^2 \mathcal{L}(x|\theta)}{\delta^2 \theta} < 0$$

**3.** Invariance property of MLE.
Let $\hat{\theta}$ be the MLE of $\theta$, then the MLE of the function $\tau(\theta)$ is $\tau(\hat{\theta})$.

**4.** The MLE of $\theta$ is asymptotically a Unique Minimum Variance Unbiased Estimator (UMVUE).
(asymptotically efficient with minimum variances)

$$\sigma^2_{\hat{\theta}} = \frac{1}{E\left[\left(\frac{\delta \ln L}{\delta \theta}\right)^2\right]} = \frac{-1}{E\left[\frac{\delta^2 \ln L}{\delta \theta^2}\right]}$$