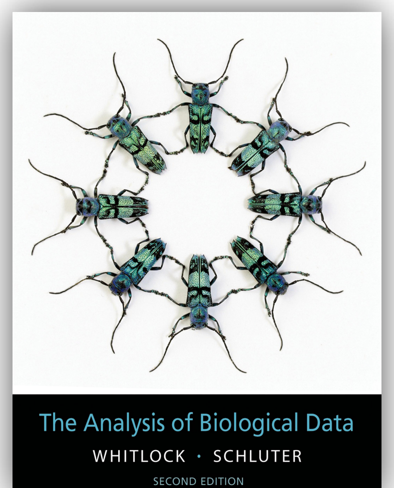
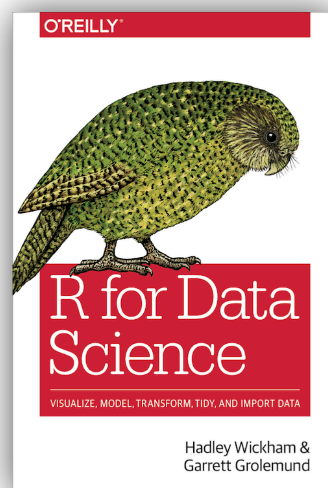
Data Science in Bioinformatics

Palle Villesen & Thomas Bataillon





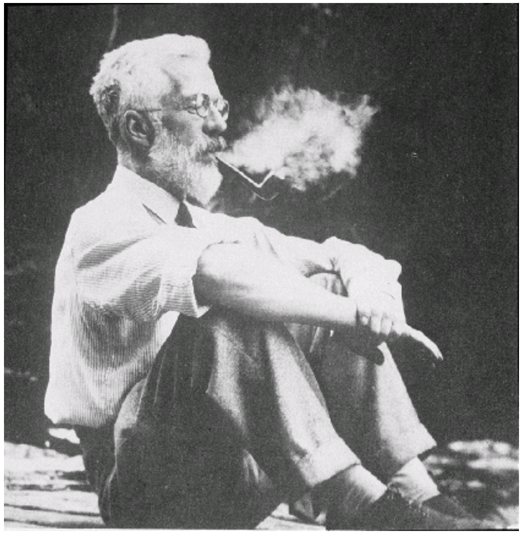
Outline for week 13

- Tuesday: Chapter 20 Likelihood
  - Wasp example (R code)
  - Human trios (R markdown)
  - Important generalizations
- Thursday session is
  - Exercises on likelihood
  - Final assignment (prepare / post Qs)

- We need your feedback

What is likelihood ? Why Bother?

- A general framework for parameter estimation and hypothesis testing
- An "old" idea (R.A. Fisher 1920s) that went a long way…
- Why likelihood is your "friend"
  - Think clearly about your data
  - Efficient way of extracting information from the data
  - Likelihood is often "hidden" behind tools you use…



Likelihood in a nutshell

- Choose a model M_θ for your data *D*
- Write down the probability of your data P(*D*)
- Figure out which parameter(s) value(s) of M_θ maximize P(*D*) (= make the data most likely)
- Distrust your model !
- "Wash, rinse and repeat …"

Which *model* behinds these data ?
=
Which probability distributions ?

- Height measurements of different individuals from a single population

---

- Allele frequencies in a sample taken from a (large) population
- Sex ratio in a progeny
- Presence/ absence of a species in a locality

The Binomial distribution

- "Thought Experiment":
  - n independent trials with probability of success p for each trial
  - X is a random variable
  - Formally X ~ B(n,p)

- What do we know about X ?
  - $E(X) = n\,p$ $V(X) = n\,p\,(1\text{-}p)$

  - $P(X\text{=}i) = n!/i!(n-i)!\ p^{i}\,(1-p)^{n-i}$

  - if n is large (np CONSTANT) , X becomes Poisson or even "normal"

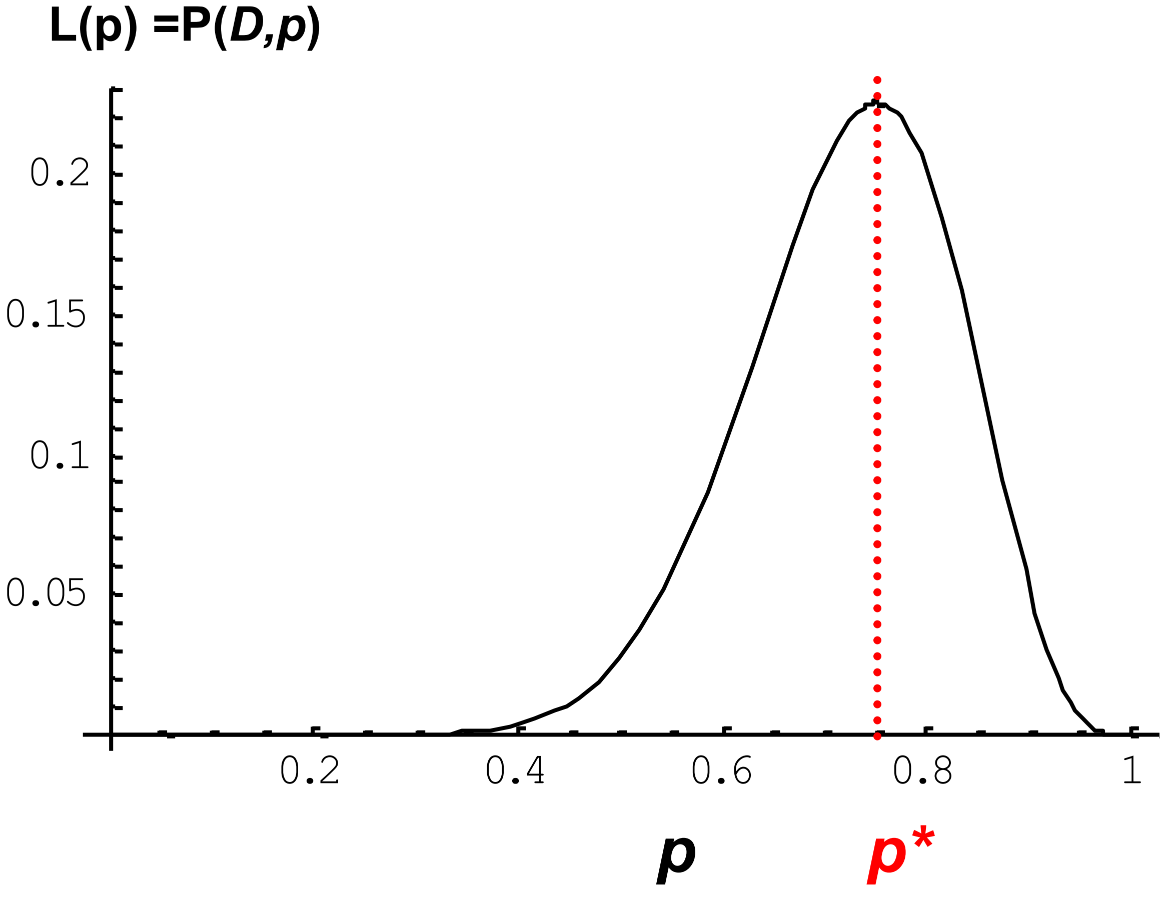
The Likelihood principle

- $L=P(D;\theta)$ is a function of data *D* and parameters θ
- Likelihood principle:
  - The Data D is fixed
  - Choose the parameter(s) value(s) θ^* that make the data D most likely

--> Maximize L (THAT's IT !)

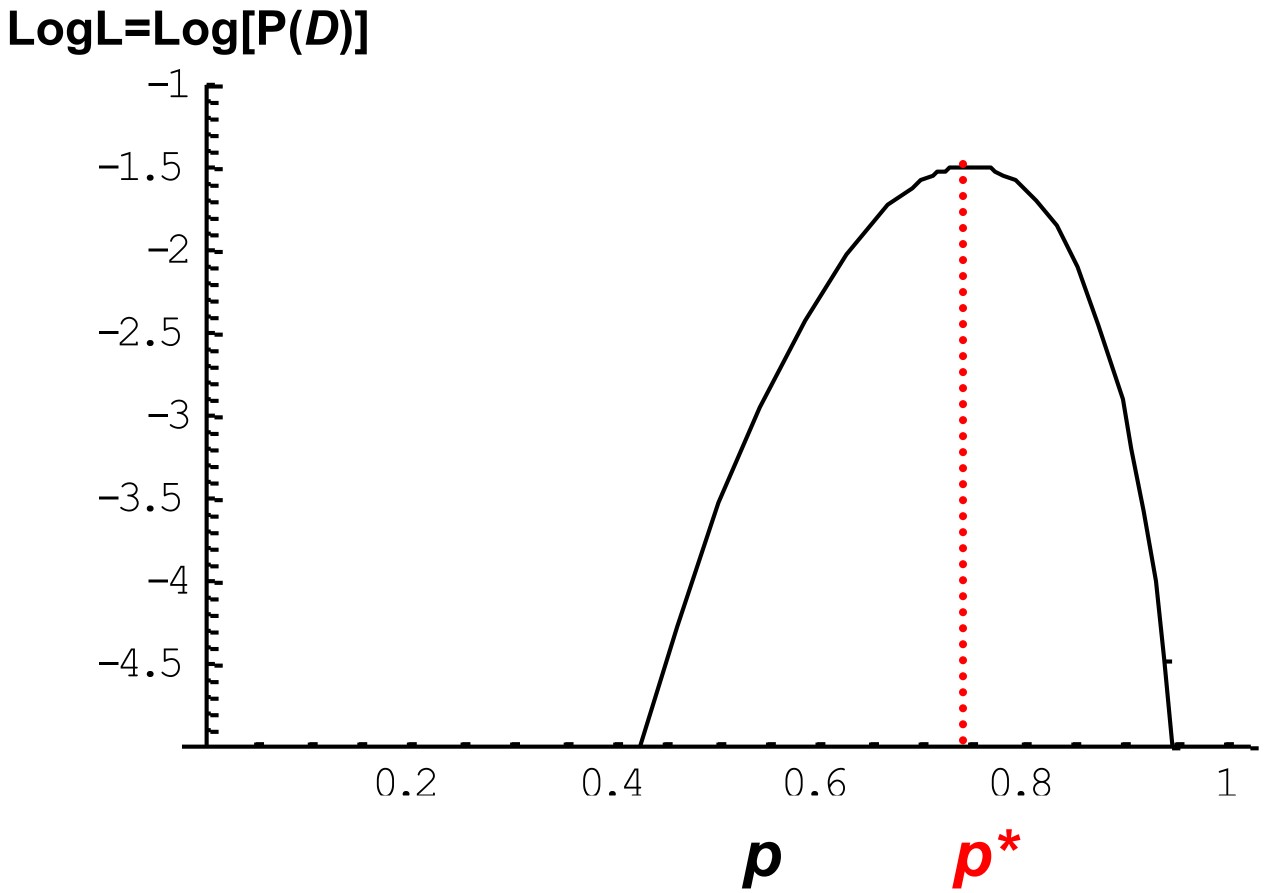
- The **invariance principle**: if alpha is a parameter and beta=f(alpha), with f continuous and monotonic, then $ML_{beta}=f(ML_{alpha})$

Visualizing the likelihood function



NB: L is a **continuous** function defined for p in [0,1]

Visualizing the (natural) log Likelihood function



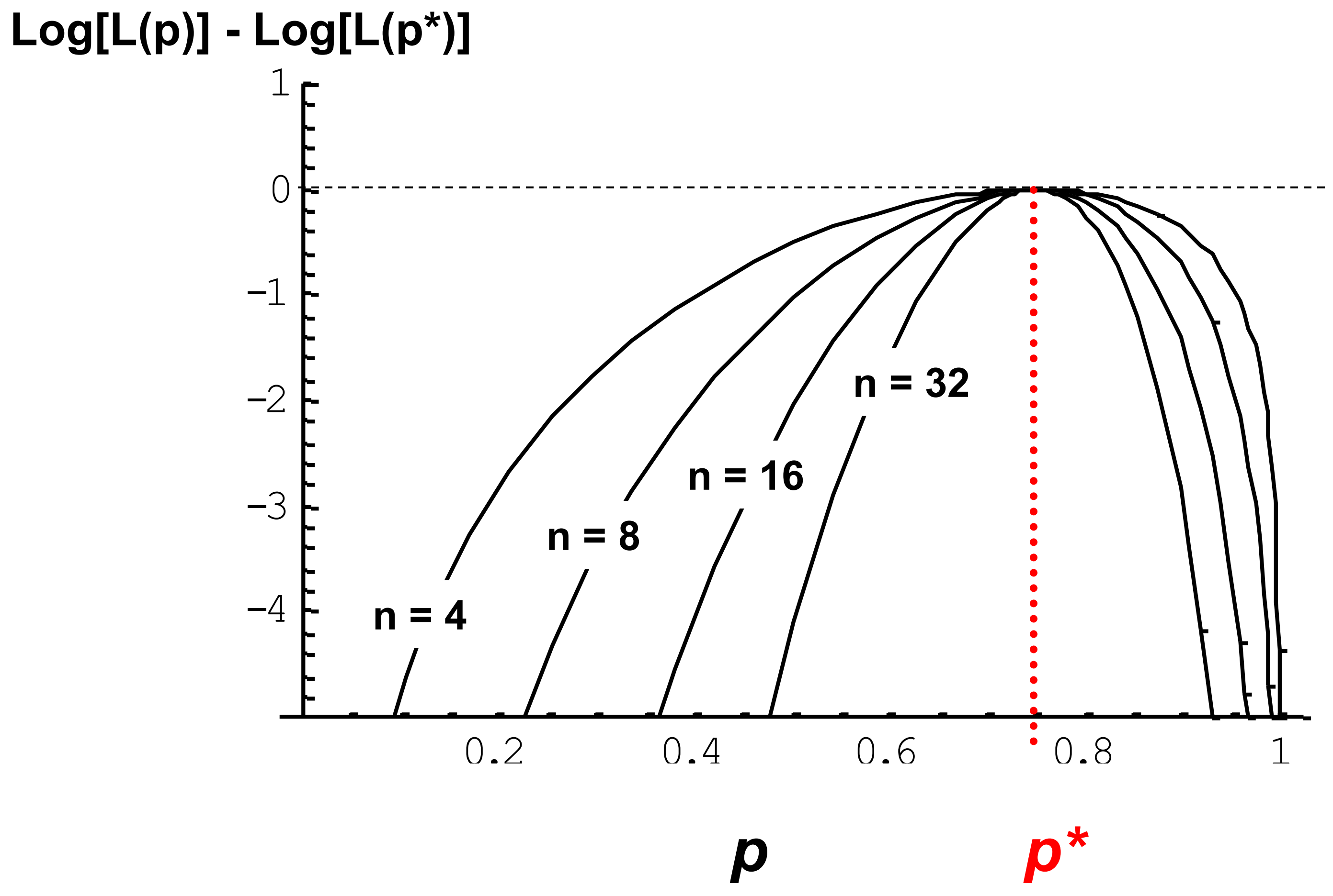
Maximizing the likelihood

- Binomial example (continued)

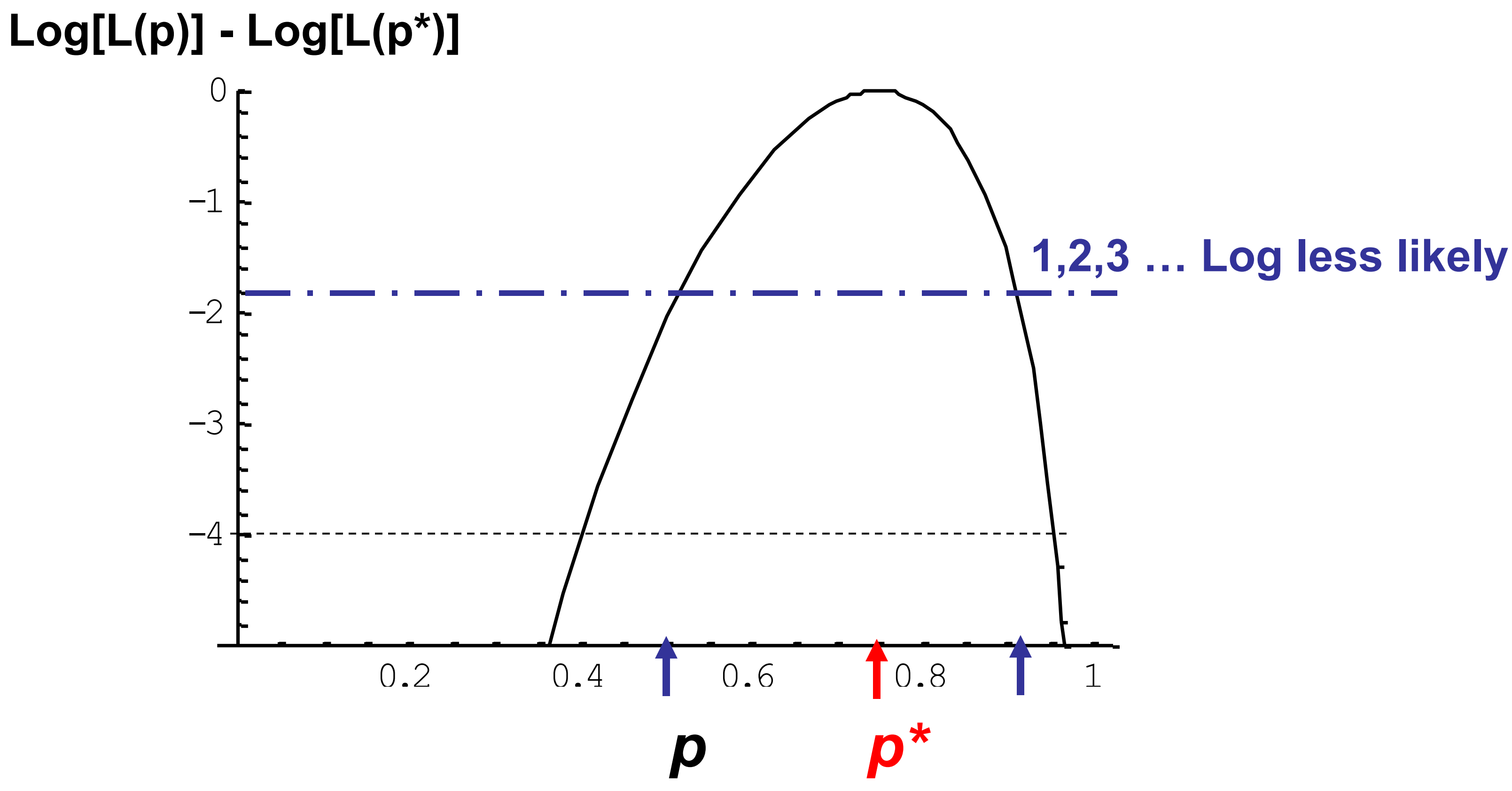
- Finding the maximum of a function
  1. set the derivative to zero,
  2. check for a maximum.

- Generalization MLEs for several variables…

The curvature of Likelihood function reflects the amount of info in your data



Visualizing the log Likelihood function rescaled



The likelihood profile (graph)

- Intuition: curvature says something about the precision on ML estimates
- General results
  - Asymptotic normality of MLE
  - Asymptotic unbiased of MLE
  - Known sampling variance …

  - Approximate 95% Confidence intervals:

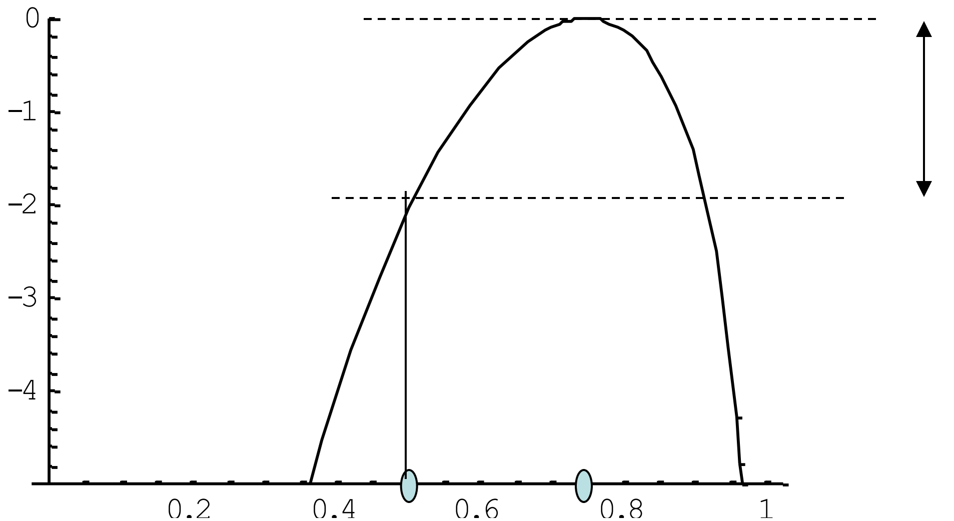
  Find the region of parameter values that yield up to ~2 log drop in likelihood …

Hypothesis testing in a likelihood framework

- Binomial example …continued

  Q: How to test against the Fisherian ½ sex ratio?

  A: compare maximum likelihood with likelihood under the "model" p=1/2.



Δ LogL ~ 2

The likelihood ratio test (LRT)

- Idea : comparing the fit of competing (nested) model to the data via their likelihood
- LRT statistic
  - 2 competing models M_a and M_b with M_b nested in M_a
  - $G = 2[\mathrm{LogL}(M_a) - \mathrm{LogL}(M_b)]$
- What is the "null" distribution of Λ ?

If M_a is correct $G \sim \chi^2_n$

n is the difference in the number of free parameters fitted in M_a and M_b

Comparing non nested models

- Akaike's Information Criteria (AIC)

- Collection of Models

  M_1 n1 free parameters
  M_2 n2 free parameters
  …
- Compute AIC for each fitted model:
  $AIC_i = - 2 \, Log[L(M_i)] + 2 \, n_i$

- Choose model with **lowest AIC**

- Perspective: robust estimation with model averaging …

Some good reads…

- **Lynch, M. and B. Walsh, *Genetics and Analysis of Quantitative Traits*. 1998, Sunderland: Sinauer Associates, Inc.**

Appendix 4 of this book is a good introduction to likelihood that reviews most of the general results that I find useful when analyzing real data. A bunch of worked simples examples. Connection between LT and te so called G test for goodness of fit

- Burnham, Kenneth P., Anderson, David R. 2nd ed. 2002 **Model Selection and Multi-Model Inference A Practical Information-Theoretic Approach** Springer Verlaag ISBN: 0-387-95364-7

Good introduction to AIC and model selection/averaging

- **Likelihood AWF 1992 Edwards J. Hopkins U Press 2nd edition**

Old fashioned especially in the style and notations but this book was for me the only gateway to the more technical likelihood literature. A must if you enjoy the British academic style !