BUILDING SCIENCE
IN AFRICA

AIMS VIDEO COURSES
SUPPORTING BOOKLET

PROBABILITY & STATISTICS

WITH
PROF DAVID SPIEGELHALTER

AIMS
SOUTH AFRICA

AIMS
SOUTH AFRICA

African Institute for Mathematical Sciences

6 MELROSE ROAD | MUIZENBERG | CAPE TOWN 7945 | SOUTH AFRICA
TEL: +27 (0)21 787 9320 | FAX: +27 (0)21 787 9321
EMAIL: info@aims.ac.za | WEB: www.aims.ac.za

AIMS Online Courses

The mission of the AIMS academic programme is to provide an excellent, advanced education in the mathematical sciences to talented African students in order to develop independent thinkers, researchers and problem solvers who will contribute to Africa's scientific development.

Teaching at AIMS is based on the principle of learning and understanding, rather than simply listening and writing, during classes, and on creating an atmosphere of increasing our knowledge through class discussions, through small group discussions, by formulating conjectures and assessing the evidence for them, and sometimes going down wrong paths and learning from the mistakes that led us there. The essential features of the classes at AIMS are that, in contrast to formal lecture courses, they are highly interactive, where the students engage with the lecturer throughout the class time, are encouraged to learn together in a journey of questioning and discovery, and where lecturers respond to the needs of the class rather than to a pre-determined syllabus. AIMS teaching philosophy is to promote critical and creative thinking, to experience the excitement of learning from true understanding, and to avoid rote learning directed only towards assessment.

Leading international and local experts offer the courses at AIMS, which are three weeks long (each module consisting of 30 hrs) and collectively form the coursework for a structured masters degree which also includes a research component. The advertised content is a guide, and the lecturers are encouraged, and indeed expected, to adapt daily to meet the current needs of the students.

Over the past ten years AIMS has achieved international recognition for this innovative and flexible approach. It has been the starting point for the remarkable success of our students and alumni and we all benefit from the support of many who have "witnessed the AIMS-magic and keep coming back for more."

This year we have decided to film selected courses and to make them available to a larger audience as an online facility. African universities may choose to use these courses to supplement and enhance their own postgraduate programmes. We believe this would be best achieved through engagement with AIMS. One way for this to happen, would be for AIMS to suggest or nominate a specialist tutor to spend time at the university, guiding students who follow the online programme. Where possible expert lecturers who have taught at AIMS may visit the university to give a short introduction to the course. We would welcome this interaction as well as the contribution our online courses will make to the growth of the mathematical sciences ecosystem in Africa.

Barry Green
Director & Professor of Mathematics
African Institute for Mathematical Sciences
January 2013

PROBABILITY & STATISTICS

2012

PROF DAVID SPIEGELHALTER

DAY 7

AIMS

SOUTH AFRICA

Statistical discovery

1. Important problem
2. Good survey / experimental design
3. Good quality data
4. Data exploration and visualisation
5. **Modelling data as random variables arising from some distribution**
6. Formal statistical inference about the truth

The concept of 'likelihood'

Assume we have a set of independent and identically distributed random variables $X_1, .., X_n$ each with probability density $f(x_i|\theta)$ for some unknown parameter θ.

Then the joint density of the observations $\underline{x}$ is $f(\underline{x}|\theta) = \prod_i f(x_i|\theta)$

The likelihood is a function $L(\theta|\underline{x})$ of θ which is proportional to $f(\underline{x}|\theta)$, i.e.

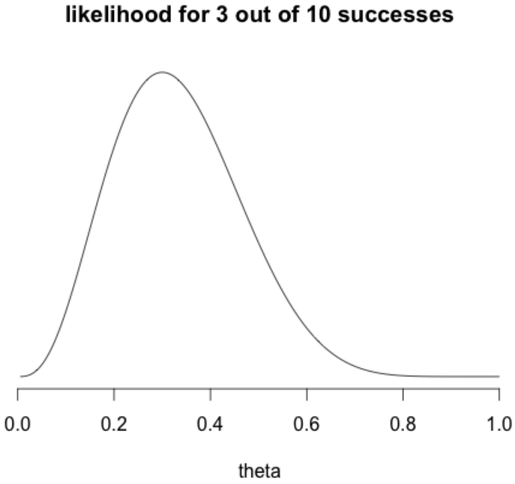
$$L(\theta|\underline{x}) \propto \prod_i f(x_i|\theta).$$

It contains the parts of $f(\underline{x}|\theta)$ that contain θ: it is only defined up to an arbitrary multiplicative constant.

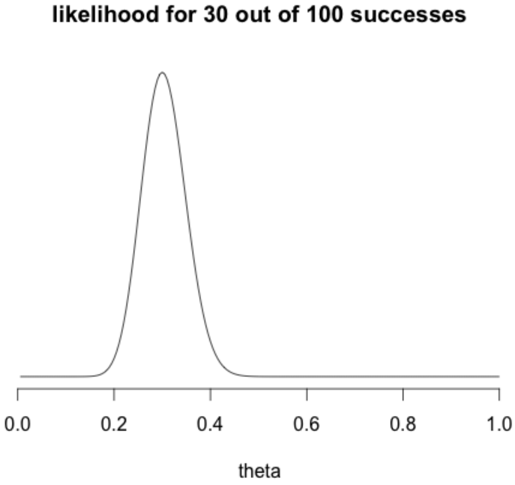
e.g. for the Binomial distribution ($x =$ sum of n Bernoulli variables) $f(x|\theta) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$, the likelihood is

$$L(\theta|x) \propto \theta^x (1-\theta)^{n-x}.$$

Plotting likelihoods



Plotting likelihoods



Maximum likelihood estimation

What parameter value to use an estimate?

Use the **mode** (the maximum value) of the likelihood - the 'maximum likelihood estimate' (MLE) $\hat{\theta}$

In most cases can find this value by differentiation (although not if maximum is on boundary of the parameter space)

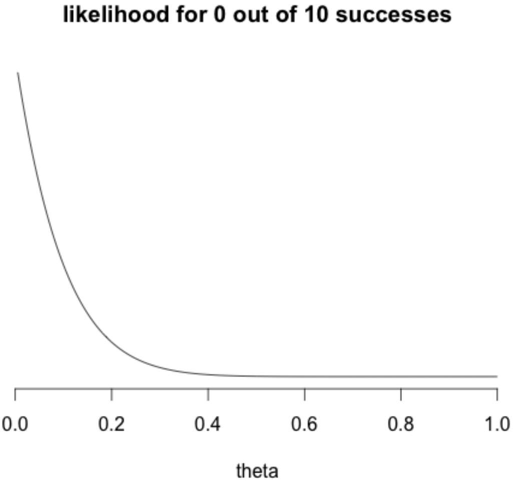
Easiest to take log-likelihood $\ell(\theta|\underline{x}) = \log L(\theta|\underline{x})$ and differentiate it (natural log)

These have the same maximum by chain rule)

$$\ell'(\theta|\underline{x}) = \frac{d}{d\theta} \log L(\theta|\underline{x}) = \frac{L'(\theta|\underline{x})}{L(\theta|\underline{x})}.$$

Works the same for a vector θ of parameters: solve for $\ell'(\theta|\underline{x}) = 0$.

Can't always differentiate to find maximum



Maximum likelihood estimation

Example: Binomial

$$\ell(\theta|x) = \text{constant} + x \log\theta + (n-x)\log(1-\theta).$$

[Note: the same likelihood as if considered as n Bernoulli trials]

$$\ell'(\theta|x) = \frac{x}{\theta} - \frac{(n-x)}{(1-\theta)}$$

Setting this to 0 and solving for θ gives

$$\hat{\theta} = \frac{x}{n},$$

which is what we would expect.

Maximum likelihood estimation: vector θ of parameters

Normal example: $f(\underline{x}|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, and so

$$\log f(x_1,..,x_n|\mu,\sigma) = \text{constant} - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\log\sigma^2 = \ell(\mu,\sigma|\underline{x})$$

For MLEs $(\hat{\mu},\hat{\sigma})$, need

$$\frac{\delta}{\delta\mu}\ell(\mu,\sigma|\underline{x})\bigg|_{\hat{\mu},\hat{\sigma}} = \sum_i \frac{(x_i-\hat{\mu})}{\hat{\sigma}^2} = 0,$$

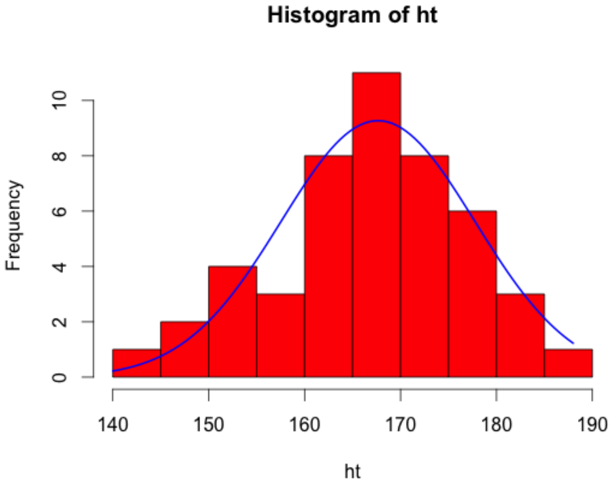
and so $\hat{\mu} = \overline{x} = \text{mean}(\underline{x})$.

$$\frac{\delta}{\delta\sigma}\ell(\mu,\sigma|\underline{x})\bigg|_{\hat{\mu},\hat{\sigma}} = \sum_i \frac{(x_i-\hat{\mu})^2}{\hat{\sigma}^3} - \frac{n}{\hat{\sigma}} = 0,$$

and so $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_i(x_i-\overline{x})^2} = \sqrt{\text{var}(\underline{x})} = \text{standard deviation}(\underline{x})$.

Fitting a distribution

Fitted normal, superimpose on histogram for heights



How accurate is the MLE?

We want to know how close the MLE is likely to be to the 'true' value θ_0

For this we need to know that, as the sample size gets bigger, the MLE

- tends to the true value ('consistency')
- has a known (or at least estimable) variance

Fisher Information

The Fisher Information $I(\theta)$ tells us how much information is in the likelihood for a single observation.

It is the expectation of the negative 2nd derivative (the curvature) of the log-likelihood.

$$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right].$$

Example for a Bernoulli trial with parameter θ

First derivative:

$$\frac{d}{d\theta}\log f(X|\theta) = \frac{d}{d\theta}\ell(\theta|X) = \frac{d}{d\theta}\left[X\log\theta + (1-X)\log(1-\theta)\right]$$

$$= \frac{X}{\theta} - \frac{(1-X)}{(1-\theta)}.$$

Second derivative:

$$\frac{d^2}{d\theta^2}\log f(X|\theta) = \frac{d}{d\theta}\ell'(\theta|X) = \frac{d}{d\theta}\left[\frac{X}{\theta} - \frac{(1-X)}{(1-\theta)}\right]$$

$$= -\frac{X}{\theta^2} - \frac{(1-X)}{(1-\theta)^2}.$$

Example for a Bernoulli trial with parameter θ

Take negative expectation to give:

$$I(\theta) = -\mathbb{E}\left[-\frac{X}{\theta^2} - \frac{(1-X)}{(1-\theta)^2}\right] = \frac{\theta}{\theta^2} + \frac{(1-\theta)}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

Asymptotic distribution of MLE

Under fairly general conditions, as $n \to \infty$, then

1. 'Consistency',: $\hat{\theta} \to \theta_0$; with enough data MLE will give the 'true' value
2. 'Efficiency': The variance of the MLE, $\mathbb{V}[\hat{\theta}]$ tends to $\frac{1}{nI(\theta)}$, which is (asymptotically) the minimum possible variance
3. 'Normality': The distribution of $\hat{\theta}$ tends to $\hat{\theta} \sim \text{Normal}\left(\theta_0, \frac{1}{nI(\theta)}\right)$

Binomial example: according to MLE theory, $\hat{\theta} = x/n$ has a Normal $\left(\theta_0, \frac{p(1-p)}{n}\right)$ distribution, just as we found earlier.

Invariant property

Really important!

If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ for 1-1 transformations g

Stages for MLEs

1. Write down $f(x|\theta)$ for a single observation
2. Take natural logs to give log-likelihood $\ell(\theta|x) = \log f(x|\theta)$, ignoring additive terms that do not contain θ
3. Get log-likelihood for n observations $\ell(\theta|\underline{x}) = \sum_i \log f(x_i|\theta)$
4. Differentiate, set to 0 and solve for MLE $\hat{\theta}$: ie $\frac{d}{d\theta}\ell(\theta|\underline{x})\big|_{\hat{\theta}} = 0$
5. Go back to log-likelihood for a single random variable X, i.e. $\ell(\theta|X) = \log f(X|\theta)$
6. Take first derivative $\ell'(\theta|X) = \frac{d}{d\theta}\log f(X|\theta)$
7. Take second derivative $\ell''(\theta|X) = \frac{d^2}{d\theta^2}\log f(X|\theta)$
8. Take negative expectation to give Fisher Information $I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right]$
9. MLE $\hat{\theta}$ has asymptotic variance $1/(nI(\theta))$.
10. Since θ is unknown, in practice we need to estimate the Fisher Information by substituting in the MLE to give $\hat{I}(\theta) = I(\theta)|_{\hat{\theta}}$

Stages for MLE for a Poisson

1. $f(x|\theta) = e^{-\theta}\theta^{x}/x!$
2. log-likelihood $\ell(\theta|x) = \log f(x|\theta) = -\theta + x\log\theta$+constant
3. log-likelihood for n observations $\ell(\theta|\underline{x}) = \sum_i \log f(x_i|\theta) = -n\theta + \sum_i x_i \log\theta$+constant
4. Solve for MLE $\frac{d}{d\theta}\ell(\theta|\underline{x})\big|_{\hat{\theta}} = -n + \frac{\sum_i x_i}{\theta} = 0$, so $\hat{\theta} = \overline{x}$.
5. Log-likelihood for X, i.e. $\ell(\theta|X) = -\theta + X\log\theta$+constant
6. First derivative $\ell'(\theta|X) = -1 + X/\theta$
7. Second derivative $\ell''(\theta|X) = -X/\theta^2$
8. Fisher Information $I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] = -\mathbb{E}[-X/\theta^2] = \theta/\theta^2 = 1/\theta$
9. MLE $\hat{\theta}$ has asymptotic variance $1/(nI(\theta)) = \theta/n$.
10. Estimated variance: $\hat{I}(\theta) = I(\theta)|_{\hat{\theta}} = \overline{x}/n$