BUILDING SCIENCE IN AFRICA

AIMS VIDEO COURSES

SUPPORTING BOOKLET

PROBABILITY & STATISTICS

WITH

PROF DAVID SPIEGELHALTER

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 SOUTH AFRICA

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 13

AIMS
SOUTH AFRICA

Conditional probability

Definition: Let A and B be two events with $P(A) > 0$.

Then the conditional probability of B given A, written $P(B|A)$, is defined to be

$$P(B|A) = \frac{P(B \cap A)}{P(A)}.$$

If $P(A) = 0$, $P(B|A)$ is undefined (pointless to condition on an impossible event)

Bayes theorem

Since

$$p(B \cap A) = p(A \cap B),$$

conditional probability implies that

$$p(A|B)p(B) = p(B|A)p(A),$$

or equivalently

$$p(A|B) = \frac{p(B|A)}{p(B)} \times p(A).$$

This is Bayes theorem

We can interpret this as a formal mechanism for learning from experience

An initial probability $p(A)$ is changed into a conditional probability $p(A|B)$ when taking into account the event A occurring

Joint probability and cross-tabulations

Consider the sample:

	B: like football	B^c: don't like	
A: female	40	10	50
A^c: male	20	30	50
	60	40	100

Suppose I pick a random person and they like football (observe B). What is the chance I have chosen a woman (A)?

We want $p(A|B)$, and we have $p(A) = 50/100 = 0.5$, $p(B|A) = 40/50 = 0.8$, $p(B) = 60/100 = 0.6$. So

$$p(A|B) = \frac{p(B|A)}{p(B)} \times p(A) = \frac{0.8}{0.6} \times 0.5 = 0.67$$

Can also read this directly off the table: $p(A|B) = 40/60 = 0.67$

Diagnosis: Bayes theorem in diagnostic testing

A new home HIV test is claimed to have "95% sensitivity and 95% specificity",

i.e. for people with HIV, 95% will get a positive test

for people without HIV, 95% will get a negative test

To be used in a population with an HIV prevalence of 1/100

If someone has a positive test result, what is the chance they actually have HIV?

Diagnosis: Bayes theorem in diagnostic testing

Expected status of 10000 tested individuals in population:

	HIV -	HIV +	
Test -	9802	5	9807
Test +	198	95	193
	9900	100	10000

Thus of 193 who test positive, only 95 are truly HIV positive

A 'predictive value positive' of only $95/193 = 49\%$.

The lie detector

Some students do not hand in their assignments, and they give the lecturer an excuse.

A lecturer claims that he can tell when someone is not telling the truth about their excuse.

1. If the student is lying, the lecturer always can tell
2. If the student is telling the truth, there is a 20% chance the lecturer will wrongly accuse him of lying

Suppose 10% of students do not tell the truth when they fail to hand in their assignments

If the lecturer accuses a student of lying, what is the chance he is right?

What is wrong with maximum likelihood?

Problems can occur

- With small samples (asymptotics don't work, background evidence ignored)
- When the observation is on the boundary of the parameter space
- When the parameter defines the parameter space
- When there are many parameters

The unlucky football team

Assume the number of goals a team scores has a Poisson distribution with mean θ

In the first 4 matches they do not score at all, i.e. $x_1 = x_2 = x_3 = x_4 = 0$

The likelihood is $e^{-4\theta}$

The MLE $\hat{\theta}$ is 0

So we predict they will never, ever score!

How many buses

Busses in a city have a number (not the route number) consecutively $1, 2, 3, , N$

The first bus we see is numbered 10

How many buses are there?

How many buses

The first number X has a uniform distribution

$$p(X = x|N) = \frac{1}{N}, \quad x = 1, 2, 3, 4, .., N$$

So having observed $x = 10$, the likelihood is $\propto \frac{1}{N}, \quad N > 10$

This is maximised at $\hat{N} = 10$

So we would estimate there are 10 busses.

Does not seem reasonable.

Bayesian inference

Treat θ as random variable

Express uncertainty about a parameter θ as a probability distribution

The probability distribution expresses our *ignorance*, not randomness

Epistemic uncertainty, rather than *aleatory* uncertainty

This is very different from what we have been learning!

No P-values, no confidence intervals!

Thomas Bayes

In 1763, Reverend Thomas Bayes of Tunbridge Wells invented the following 'thought experiment'

Imagine someone throws a ball at random on a Pool Table, and draws a line across then table where it lands.

Then he throws n more balls, and tells you the number y that lie to the left of the line.

Where is the line?

In modern language, given $y \sim \mathrm{Binomial}(\theta, n)$, what is the probability distribution for θ?

Animation

Bayes / Laplace

If something has happened y out of n times, then the probability it will happen next time is

$$p = \frac{y+1}{n+2}.$$

What if something has happened n out of n times?

The probability it will happen next time is

$$p = \frac{n+1}{n+2}$$

After $n = 0$ events, $p = 1/2$

After $n = 1$ events, $p = 2/3$

After $n = 1,000,000,000$ events, $p = 1,000,000,001/1,000,000,002$

Never reaches 1!

Bayesian inference

The Bayesian analyst (continuous parameters) needs to

- explicitly state a reasonable opinion concerning the plausibility of different values of the parameters *excluding* the evidence from the study (the **prior distribution**)
- provide the support for different values of the parameter effect based *solely* on data from the study (the **likelihood**),
- weight the likelihood from the study with the relative plausibilities defined by the prior distribution to produce
- a final opinion about the parameters (the **posterior distribution**)

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)} \propto p(y \mid \theta)\, p(\theta)$$

when considering $p(y \mid \theta)$ as a function of θ: ie the *likelihood*.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Putting probabilities on parameters

What is a reasonable form for a prior distribution for a proportion?

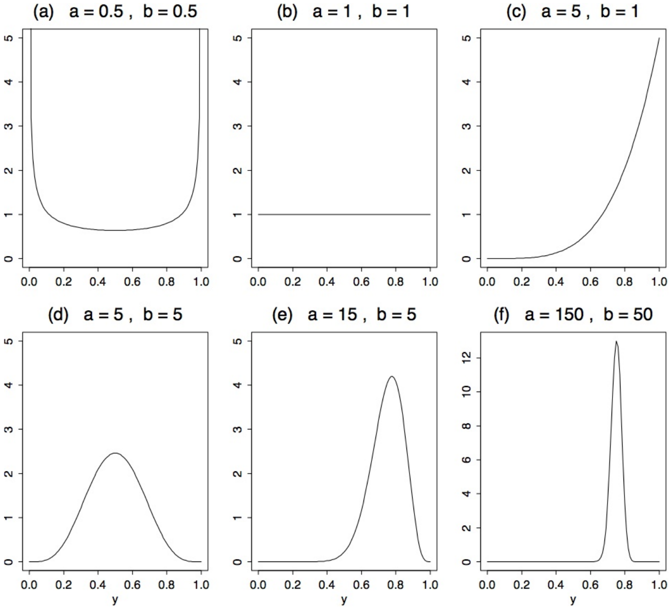
$\theta \sim \text{Beta}[a, b]$ represents a beta distribution with properties:

$$
\begin{aligned}
p(\theta|a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \theta^{a-1}\, (1-\theta)^{b-1}; \quad \theta \in (0, 1) \\
\mathrm{E}(\theta|a, b) &= \frac{a}{a+b} \\
\mathrm{V}(\theta|a, b) &= \frac{ab}{(a+b)^2(a+b+1)} :
\end{aligned}
$$

$$
\Rightarrow \int_0^1 \theta^{a-1}\, (1-\theta)^{b-1}\, d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.
$$

($\Gamma(a) = (a-1)!$ if a integer; $\Gamma(1) = 0! = 1$)

Beta distributions

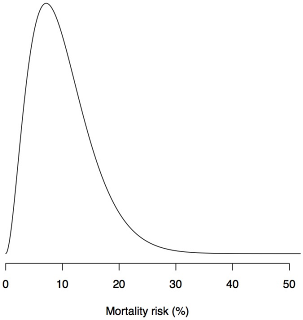


Putting probabilities on parameters

Example:

- Suppose a hospital is considering a new high-risk operation
- Experience in other hospitals indicate that the risk θ for each patient is expected to be around 10%
- it would be fairly surprising (all else being equal) to be less than 3% or more than 20%

Surgical example



- A Beta[3,27] proportional to $\theta^2(1-\theta)^{26}$
- Mean $= 3/(3+27) = 0.1$, standard deviation 0.054, variance 0.003, median 0.091, mode 0.071.
- An equi-tailed 90% interval is (0.03, 0.20) which has width 0.17, but a narrower 'Highest posterior density' interval is (0.02, 0.18) with width 0.16

Inference on proportions using a continuous prior

Suppose we observe y positive responses out of n Bernoulli trials.

Binomial sampling distribution:

$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \;\propto\; \theta^y (1-\theta)^{n-y}$$

Assume a $\text{Beta}(a, b)$ prior distribution for θ

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Combining this with the binomial likelihood gives a posterior distribution

$$\begin{aligned} p(\theta \mid y, n) &\propto p(y \mid \theta, n)p(\theta) \\ &\propto \theta^y (1-\theta)^{n-y} \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{n-y+b-1} \\ &\propto \text{Beta}(y+a,\; n-y+b) \end{aligned}$$

Inference on proportions using a continuous prior

$$\mathbf{E}(\theta|y,n) = (y+a)/(n+a+b) = w\frac{a}{a+b} + (1-w)\frac{y}{n}$$

where $w = (a+b)/(a+b+n)$

a weighted average of the prior mean and y/n, the standard maximum-likelihood estimator,

the weight w reflects the relative contribution of the prior 'effective sample size' $a+b$.

Hence the prior parameters a and b can be interpreted as equivalent to observing a events in $a+b$ trials

Surgery (continued)

Suppose we now operate on $n = 10$ patients and observe $y = 0$ deaths. What is the current posterior distribution?

We used a $\text{Beta}(3, 27)$ as a prior distribution for a mortality rate.

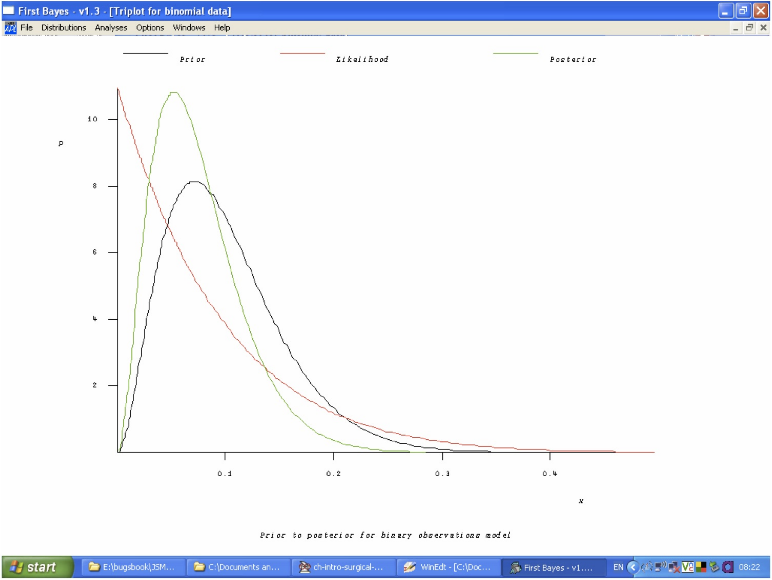
Plugging in the relevant values of $a = 3$, $b = 27$, $y = 0$ and $n = 10$

we obtain a posterior distribution for the mortality rate θ of $p(\theta|y, n) = \text{Beta}(3, 37)$

Can use *First Bayes* : `www.firstbayes.co.uk/`

Written by Tony O'Hagan: not nice to install but fun to use!

Surgery (continued)



Prior distributions

The prior distribution should reflect 'external evidence' available before the data

e.g. last year's performance for a football team

If necessary, could be based on judgement

'Subjective probability distribution'

Subjective probability distributions

Draw a table with three columns

Uncertain quantity	25% lower bound	75% upper bound
1		
2		
3		
3		
4		
5		
6		
7		
8		
9		
10		

For each question, give a range reflecting your uncertainty

This is a 50% interval (75% - 25%), so you should be 50:50 whether the true answer lies inside or outside the interval

Subjective probability distributions

Questions

1. What is the distance from Capetown to Cairo (straight line) in km?
2. What is the population of South Africa (2012 estimated)?
3. In what year was the railway to Muizenberg opened?
4. How many countries are there in Africa (member states in UN)?
5. How much is the Cape Times newspaper (rand)?
6. What percentage of the South African population is 'white' (Wikipedia)?
7. What is the average distance from the Earth to the moon (km)?
8. How many views has Shakira's Waka Waka 2010 World Cup video had on Youtube?
9. What is the height of Table Mountain above sea level (metres)?
10. What is the distance from Capetown to Cairo by road (Google maps) in km?