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 3

AIMS
SOUTH AFRICA

Random variable as a function

A discrete valued random variable X on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ is a function X from Ω to the real numbers R

The real-valued function f defined on R by $f_X(x) = P(X = x)$ is called the *discrete density function* (probability mass function)

$F_X(x) = P(X \leq x)$ is called the *discrete distribution function* : the event $X \leq x$ is interpreted as $\omega : X(\omega) \leq x$

Example: flip coin twice, let X be the total number of heads H

Then

$$
\begin{aligned}
f_X(0) &= P(X = 0) = P(TT) = 0.25 \\
f_X(1) &= P(X = 1) = P(HT \cup TH) = 0.50 \\
f_X(2) &= P(X = 2) = P(HH) = 0.25
\end{aligned}
$$

Expectation and variance

For a discrete random variable X with a finite set of possible values $x_1, .., x_r$, we define

- Expectation (mean): $\mathbb{E}_X[X] = \sum_{i=1}^{r} x_i f(x_i)$
- Variance: $\mathbb{V}_X[X] = \sum_{i=1}^{r}(x_i - \mathbb{E}_X[X])^2 f(x_i)$
- Standard deviation $= \sqrt{\text{Variance}}$

Generally easier to compute (dropping suffix)

$$\mathbb{V}[X] = \sum_{i=1}^{r} x_i^2 f(x_i) - (\mathbb{E}[X])^2 = \mathbb{E}_X[X^2] - \mathbb{E}[X]^2$$

since $\sum_{i=1}^{r}(x_i - \mathbb{E}[X])^2 f(x_i) = \sum_{i=1}^{r}[x_i^2 - 2\mathbb{E}[X]\, x_i + \mathbb{E}[X]^2] f(x_i) = \mathbb{E}[X^2] - 2\sum_{i=1}^{r} \mathbb{E}[X]\, x_i f(x_i) + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

When there is a parameter, say p, then conditioning can be explicitly represented as $f_X(x|p)$ etc

Bernoulli distribution

The distribution of an 'indicator' $0/1$ variable :

Bernoulli trial with probability p of 'success'

- Density: $f_X(0|p) = 1 - p; f_X(1|p) = p$
- Expectation (mean): $\mathbb{E}_X[X|p] = (1-p) \times 0 + p \times 1 = p$
- Variance:
  $\mathbb{V}_X[Xp] = \mathbb{E}_X[X^2] - \mathbb{E}_X[X]^2 = (1-p) \times 0^2 + p \times 1^2 - p^2 = p(1-p)$
- Standard deviation $= \sqrt{p(1-p)}$

Binomial distribution

The distribution of the sum of n Bernoulli trials

How many successes out of n trials, each with probability p of success?

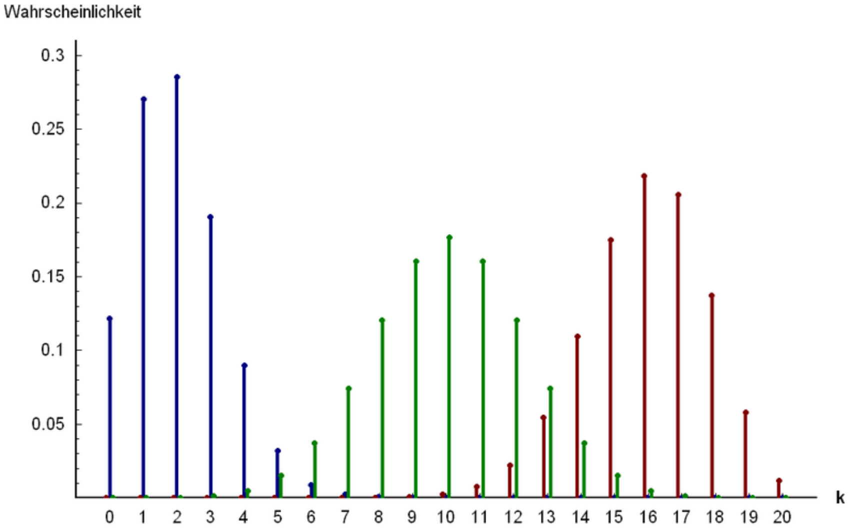
Probability of a particular sequence of x successes and $n-x$ failures is $p^x(1-p)^{n-x}$

But there are $\frac{n!}{x!(n-x)!} = \left(\begin{array}{c} n \\ x \end{array}\right)$ such sequences

- Denoted: Binomial(n, p)
- Density: $f_X(x|p) = \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x}; \;\; x = 0, 1, 2, 3, .., n$
- Expectation (mean): $\mathbb{E}_X[X|p] = np$
- Variance: $\mathbb{V}_X[X|p] = np(1-p)$
- Standard deviation $= \sqrt{np(1-p)}$

[will see how to get this mean and variance later]

Binomial distribution



Binomial distribution for $n = 20$; $p = 0.1$ (blue), $p = 0.5$ (green) and $p = 0.8$ (red)

Approximation to the Binomial distribution

Suppose each ticket in a lottery has a tiny probability p of winning

n tickets are sold: the total number of winning tickets is Binomial(n, p) with mean np, denoted μ, so that $p = \mu/n$

Then, as $n \to \infty$, the probability density function tends to

$$
\begin{aligned}
P(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\
&= \frac{\mu^x}{x!} \frac{n!}{n^x (n-r)!} \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-x} \\
&\to \frac{\mu^x}{x!} \times 1 \times e^{-\mu} \times 1 \\
&\to \frac{\mu^x}{x!} e^{-\mu}; \;\; x = 0, 1, 2, 3, \ldots
\end{aligned}
$$

This is known as the Poisson distribution

Poisson distribution

The distribution of the number of events X, when each has a very low chance of occurring, but there are many opportunities for an event to occur

- Density: $f_X(x|p) = e^{-\mu}\mu^x/x!;\ \ x = 0, 1, 2, 3, .., \infty$
- Expectation (mean): $\mathbb{E}_X[X|\mu] = \mu$
- Variance: $\mathbb{V}_X[X|\mu] = \mu$
- Standard deviation $= \sqrt{\mu}$

The Poisson is very widely used for *count* data, e.g.

- Annual cases of disease in a particular area
- Goals in a football match
- Crimes per day

Can we predict Premier League football results using a Poisson model?

‘Attack strength’ = goals for / average goals for

‘Defence weakness’ = goals against / average goals against

Team	Points	Goals for	'Attack strength'	Goals against	'Defence weakness'
Man United	87	67	1.46	24	0.52
Liverpool	83	74	1.61	26	0.57
Chelsea	80	65	1.41	22	0.48
Arsenal	69	64	1.39	36	0.78
Everton	60	53	1.15	37	0.80
Aston Villa	59	53	1.15	48	1.04
Fulham	53	39	0.85	32	0.70
Tottenham	51	44	0.96	42	0.91
West Ham	48	40	0.87	44	0.96
Man City	47	57	1.24	50	1.09
Stoke	45	37	0.80	51	1.11
Wigan	42	33	0.72	45	0.98
Bolton	41	41	0.89	52	1.13
Portsmouth	41	38	0.83	56	1.22
Blackburn	40	40	0.87	60	1.30
Sunderland	36	32	0.70	51	1.11
Hull	35	39	0.85	63	1.37
Newcastle	34	40	0.87	58	1.26
Middlesbrough	32	27	0.59	55	1.20
West Brom	31	36	0.78	67	1.46

Assessing expected goals

Hull City vs Manchester United: expected goals

Hull: = home-average x attack strength x defence weakness of opposition

= 1.36 x 0.85 x 0.52 = 0.60

Man U: = 1.06 x 1.46 x 1.37 = 2.12

Team	Expected goals	0	1	2	3	4	5
Hull City	0.60	55	33	10	2	0	0
Man U	2.12	12	25	27	19	10	4

Assume independent Poisson distributions to give probability of any result

Add to give win/draw/lose probabilities

BBC NEWS

LIVE BC NEWS CHANNEL

News s
Your ne
want it

ews Front Page

Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK

Page last updated at 14:11 GMT, Friday, 22 May 2009 15:11 UK

E-mail this to a friend

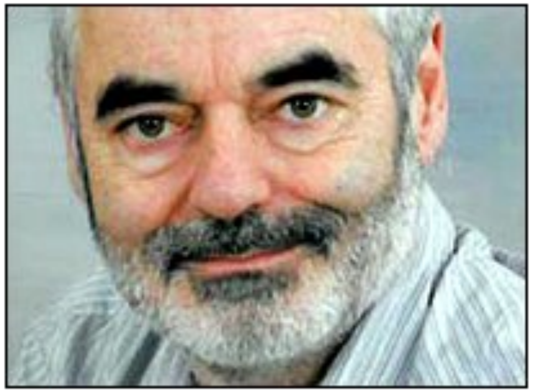
Printable version

The professor's Premiership probabilities

By Professor David Spiegelhalter
Professor for the Understanding of Risk, Cambridge University

Why should anyone take any notice of what a Cambridge professor of statistics, who knows little about football and does not even support ... bout this weekend's

... in the increasingly ... athematical models ... sed by sports betting ... t odds and identify ... bets.



Professor David Spiegelhalter analyses the football table

So I would not recommend anyone using these odds for betting.

You have been warned.

Understanding Uncertainty: Animated Premier League Statistics

PREMIER LEAGUE PROBABILITIES

Read how the professor did

ARSENAL V STOKE

Home win: 72%

Draw: 19%

Away win: 10%

V

Verdict: 2-0 (14%)

ASTON VILLA V NEWCASTLE

Home win: 62%

Draw: 21%

Away win: 17%

V

Verdict: 1-0 (10%)

Home	*Away*	*Predict*		*actual*
Arsenal	Stoke	2-0		***4-1***
Aston Villa	Newcastle	1-0		***1-0***
Blackburn	West Brom	1-1		***0-0***
Fulham	Everton	0-0		*0-2*
Hull	Man United	0-2		***0-1***
Liverpool	Tottenham	1-0		***3-1***
Man City	Bolton	2-1		***1-0***
Sunderland	Chelsea	0-1		***2-3***
West Ham	Middlesbro	1-0		***2-1***
Wigan	Portsmouth	1-0		***1-0***

Statistics: 9/10 win/draw/lose, 2 exact scores

BBC expert Mark Lawrenson: 7/10 win/draw/lose, 1 exact

Was something special about 10th July?

Page last updated at 22:22 GMT, Saturday, 12 July 2008 23:22 UK

E-mail this to a friend   Printable version

Victims in day of stabbings named

Four men stabbed to death in separate knife attacks in London in a single day have been named by police.

Melvin Bryan, 18, was fatally stabbed during a fight in a bedsit in Edmonton, north London, on Thursday.

Three hours later Adnan Patel, 20, died of stab wounds after a gang chased him in Leyton, east London.

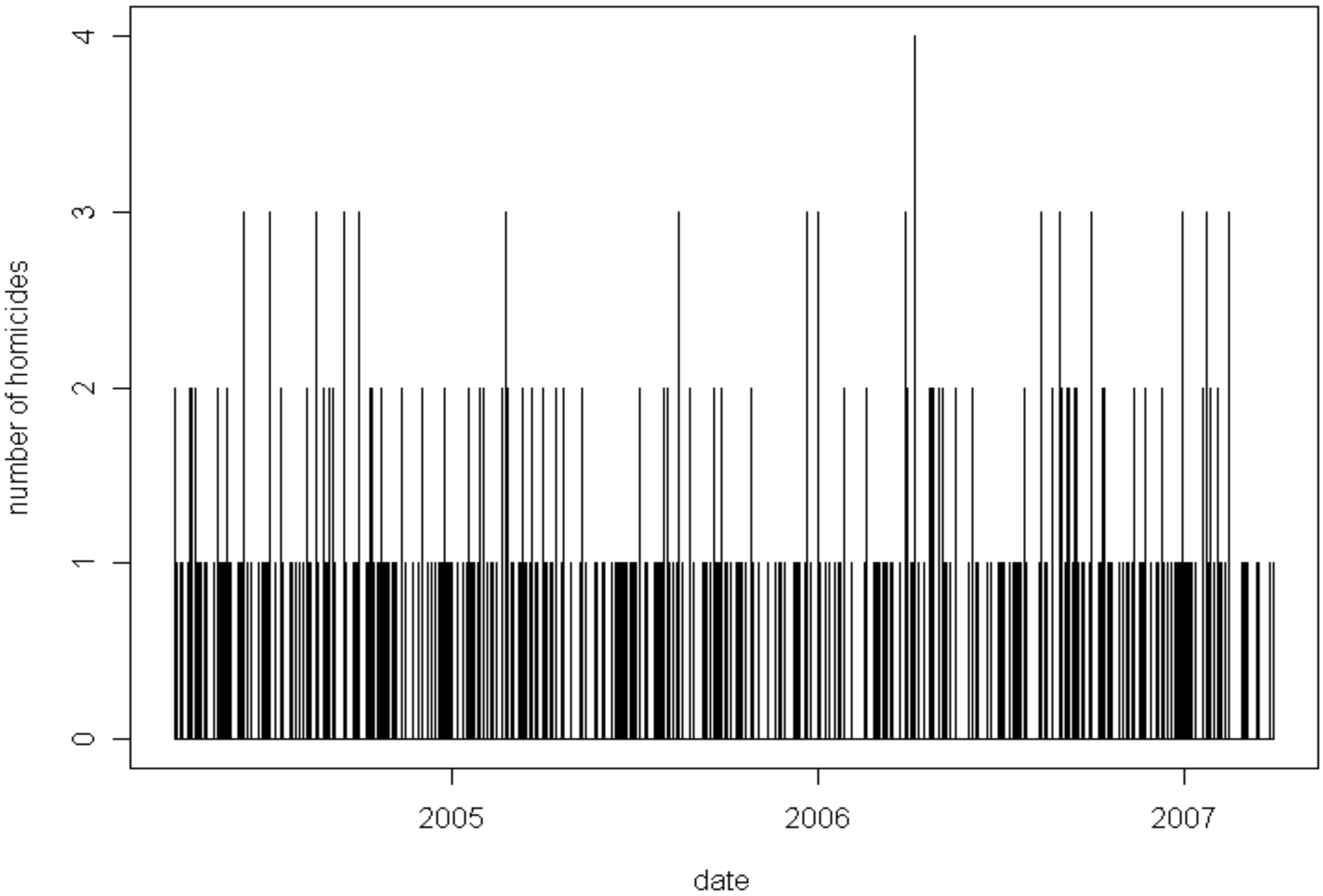
Police said Adnan Patel had been chased by a gang in Leyton

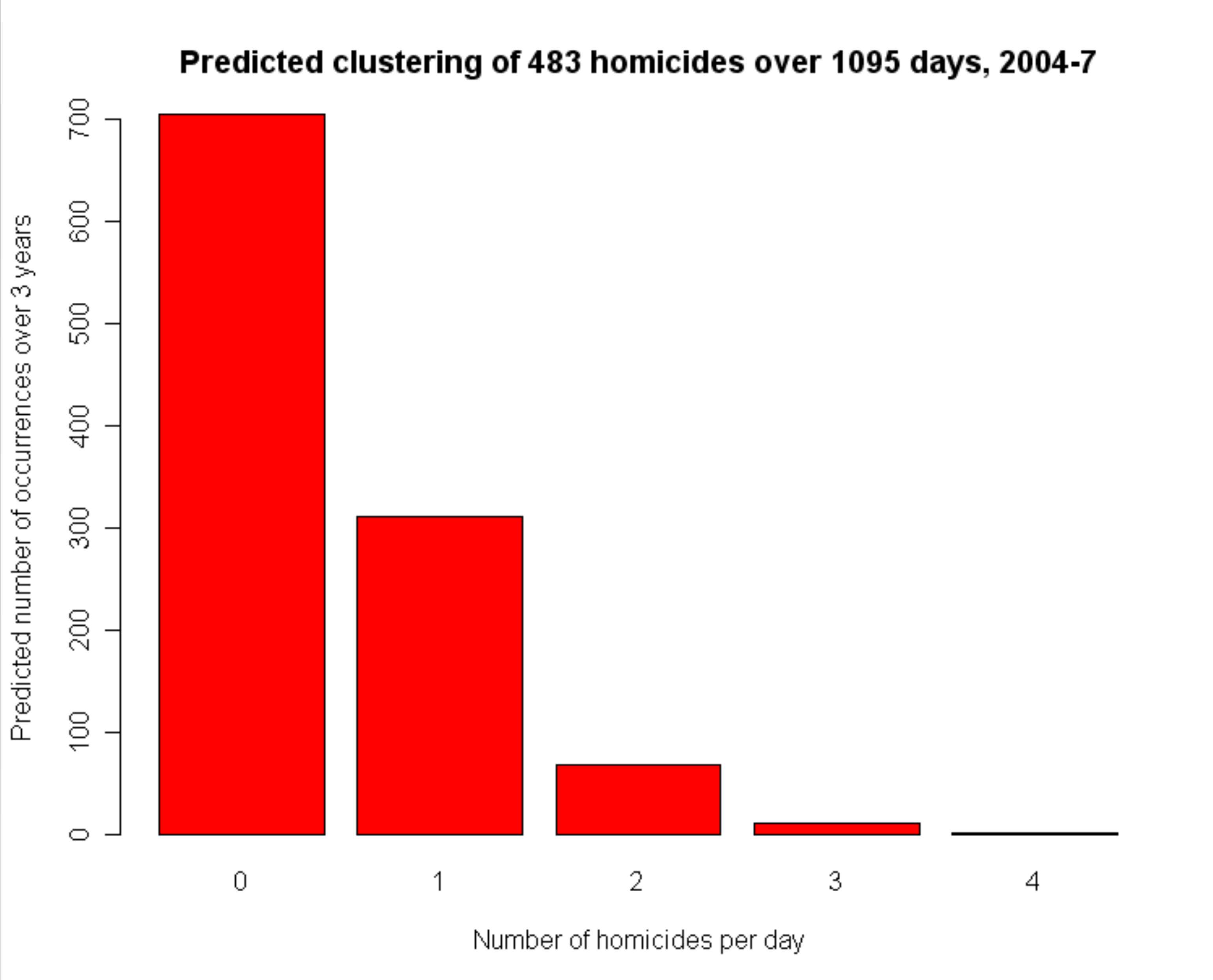
- *"Last year the Metropolitan Police recorded 160 homicides - about three every week. To have four fatal stabbings in one day could be a statistical freak, said BBC correspondent Andy Tighe"*

Homicides: Metropolitan Police, April 2004 – March 2007

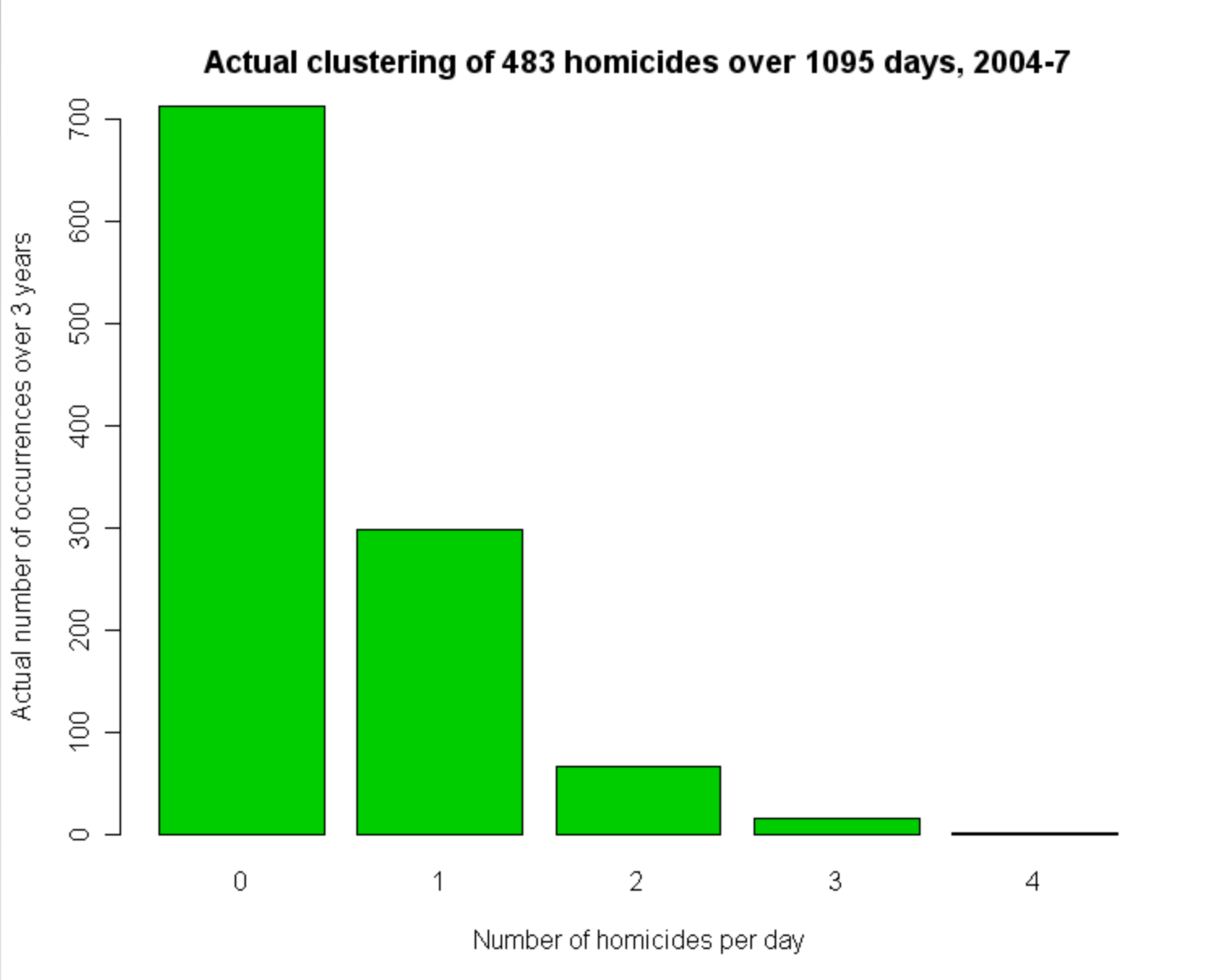
- 483 homicides in 3 years
- Removed 13 (7/7/05), 12 (unknown date)
- *On average*: 160 per year, 13 per month, 3 per week, 0.44 per day
- Just knowing this overall rate means we can predict how often ‘rare events’ will happen

Number of homicides each day, 2004 -2007





Predict 702 days with no homicides (64%), 10 days with 3 and 1 day with 4

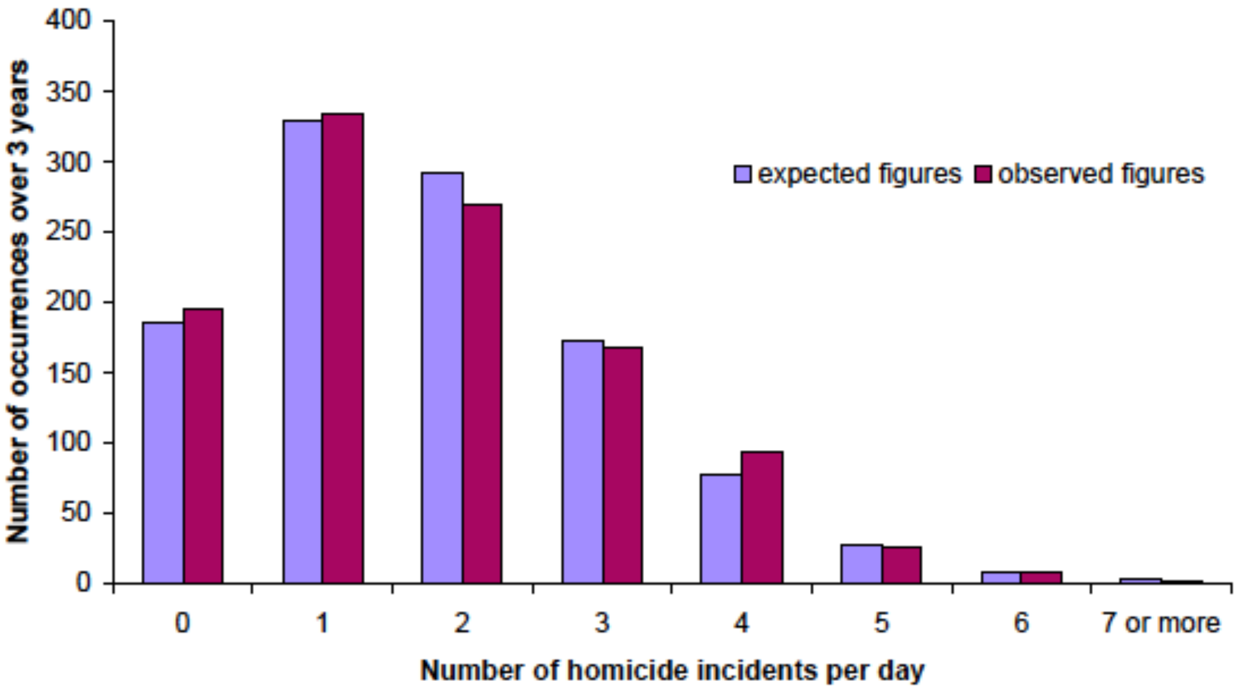


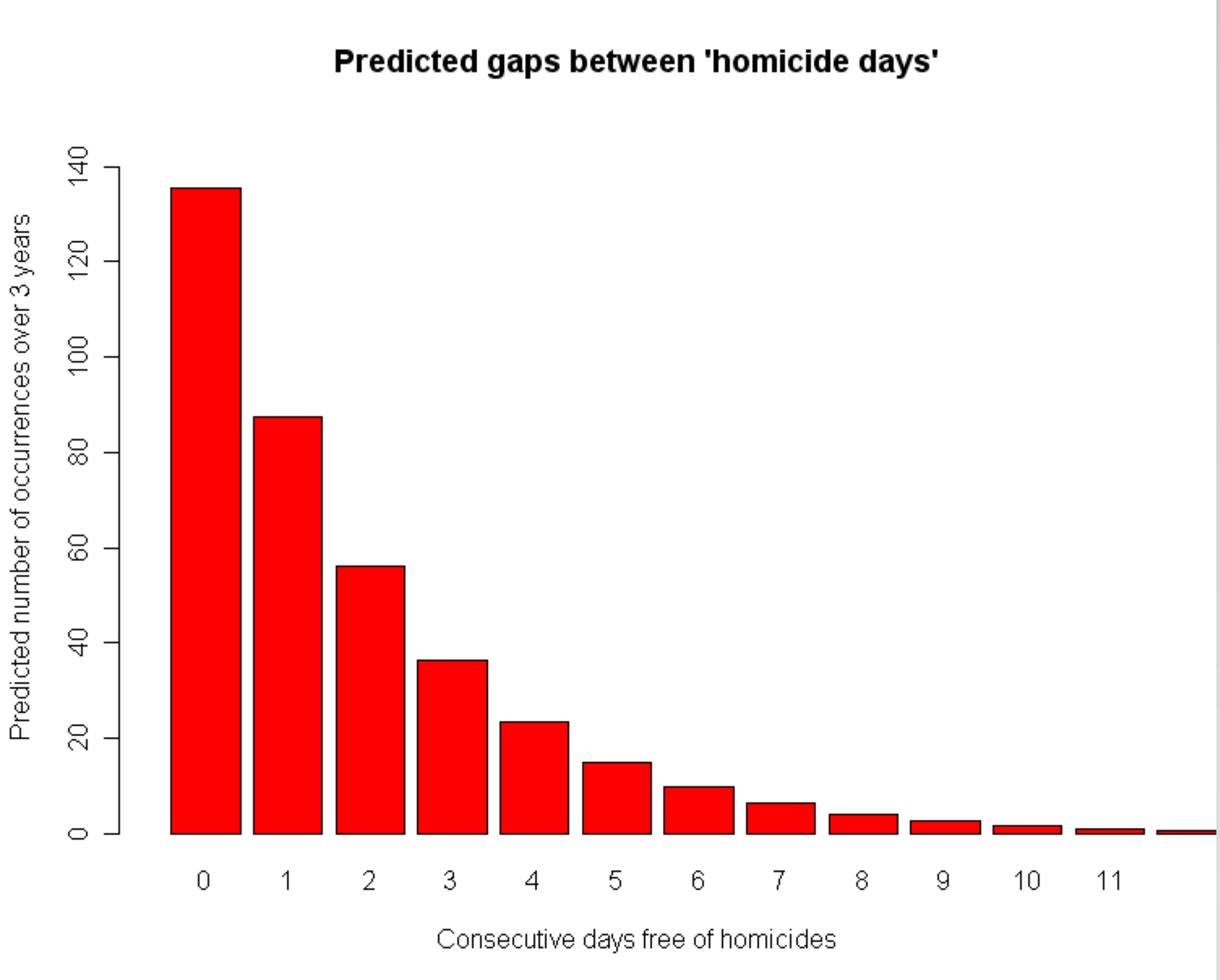
Observe 713 days with no homicides (65%), 16 days with 3 and 1 day with 4

Poisson method applied by government to national homicide figures in UK- average 1.78 incidents per day

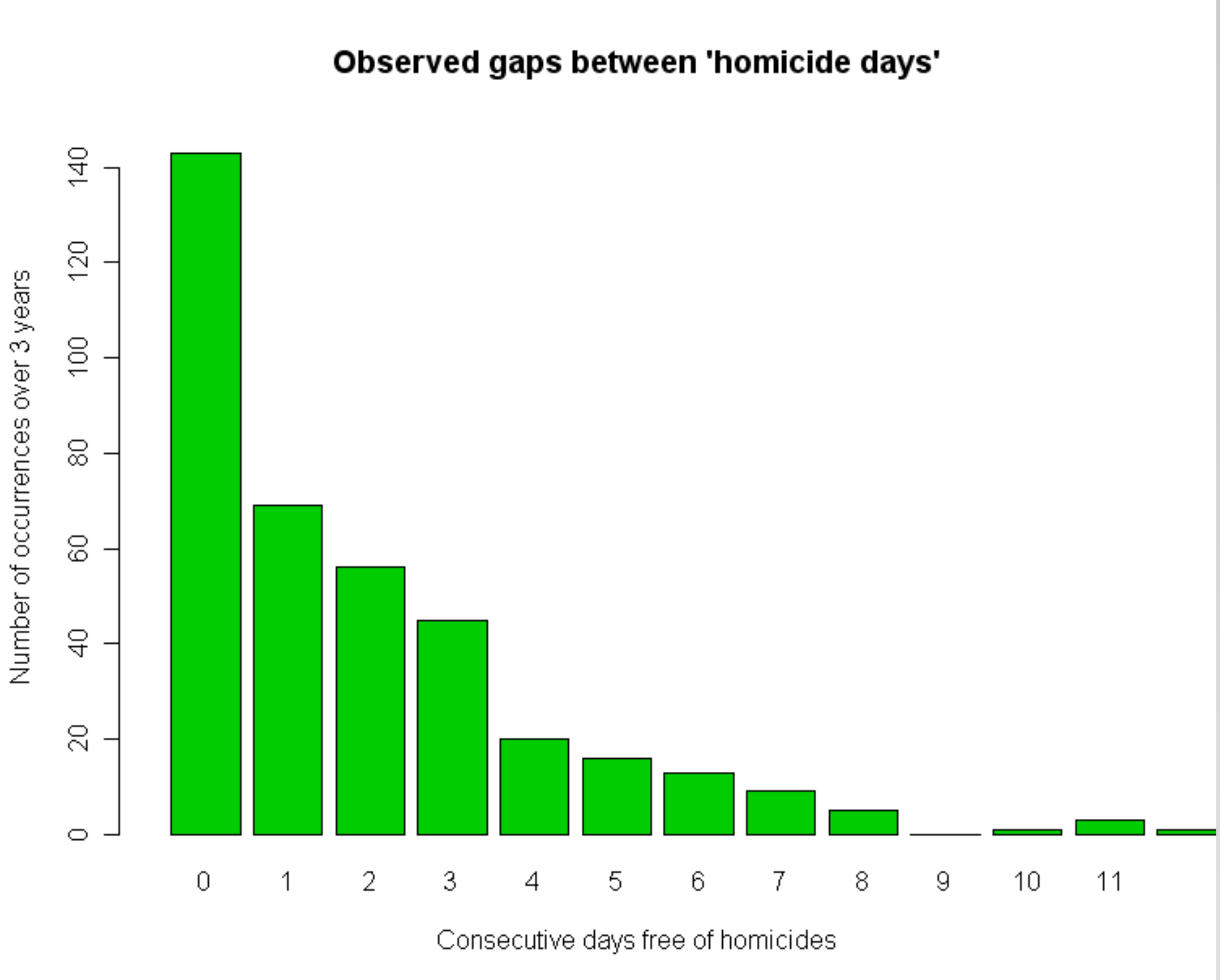
distribution. For example, from knowing there is an average of 1.78 incidents a day it was predicted that, over the period of 1,096 days, there would be 27 days on which there would be exactly five independent incidents. The observed number was 26, indicating that the occurrence of these apparent 'clusters' is not as surprising as one might anticipate. A statistical test (χ^2) shows no significant difference between the expected and observed figures. Thus, the observed figures are in fact Poisson distributed. This allows for calculation of the number of days on which it would be expected that no incidents or one incident occurs and so on.

Figure 1.7 Observed and expected number of homicide incidents recorded on a day, combined years 2007/08 to 2009/10





Predict 18 periods in which there are no homicides over 7 consecutive days



Observe 19 periods in which there are no homicides over 7 consecutive days

Mean of a discrete RV X with infinite support

Exists provided that $\sum_{i=1}^{\infty} |x_i| f_X(x_i) < \infty$

Counterexample: *St Petersburg paradox*

Suppose I flip a coin until I get a head for the first time, say on flip X, so $P(X = x) = \frac{1}{2^x}$.

Then I give you $Y = 2^X$ Rand, so Y can be $y_1 = 2, y_2 = 4, ..., y_i = 2^i$.

What is my expected loss $\mathbb{E}_Y[Y]$?

$$\mathbb{E}_Y[Y] = \sum_{i=1}^{\infty} y_i P(Y = y_i) = \sum_{i=1}^{\infty} 2^i P(X = i) = \sum_{i=1}^{\infty} 2^i 2^{-i} = 1+1+1... = \infty$$

So the expected amount I have to pay you is infinite. How much will you pay me to play this game?

Mean of a Poisson distribution

$$
\begin{aligned}
\mathbb{E}_X[X|\mu] &= \sum_{x=0}^{\infty} x \frac{\mu^x}{x!} \times e^{-\mu} \\
&= \mu \sum_{x=1}^{\infty} \frac{\mu^{x-1}}{(x-1)!} \times e^{-\mu} \\
&= \mu \sum_{i=0}^{\infty} \frac{\mu^i}{i!} \times e^{-\mu} \\
&= \mu
\end{aligned}
$$

Geometric distribution

The waiting time until the first success in a series of Bernoulli trials

- Density: $f_X(x|p) = (1-p)^{x-1}p; \;\; x = 1, 2, 3, .., \infty$
- Expectation (mean): $\mathbb{E}_X[X|p] = 1/p$
- Variance: $\mathbb{V}_X[X|p] = (1-p)/p^2$
- Standard deviation $= \sqrt{1-p}/p$

Simple transformations of random variables

Scalar addition:

- $\mathbb{E}[c + X] = \sum_x (c + x) f_X(x) = c + \mathbb{E}[X]$.
- $\mathbb{V}[c + X] = \mathbb{E}[(c + X - \mathbb{E}[c + X])^2] = \mathbb{V}[X]$.

Scalar multiplication:

- $\mathbb{E}[cX] = \sum_x (cx) f_X(x) = c\mathbb{E}[X]$.
- $\mathbb{V}[cX] = \mathbb{E}[((cX) - \mathbb{E}[cX])^2] = c^2\mathbb{V}[X]$.

Expectation of a function of a discrete random variable

[Best to put subscript for densities back in]

Other 1-1 functions $Y = g(X)$? Two ways to calculate:

1. Directly: $\mathbb{E}_X[g(X)] = \sum_x g(x) f_X(x)$
2. By finding distribution of Y; i.e.
   $f(y) = P(Y = y) = P(X = g^{-1}(y))$, and then calculating
   $\mathbb{E}_Y[y] = \sum_y y\ f(y)$

Two discrete random variables

Let X, Y be a pair of discrete random variables.

(Drop subscript for densities when obvious)

The joint density for X, Y is given by

$$f(x, y) = P(X = x, Y = y).$$

The conditional distribution for $X|Y$ is given by

$f(x|y) = f(x, y)/f(y) \quad$ if $f(y)$ exists, 0 elsewhere.

We obtain the *marginal* distribution for X by summing over the Y:

$$f(x) = \sum_y f(x, y) = \sum_y f(x|y) f(y)$$

'Extending the conversation'

They are said to be mutually independent if their joint density function is given by

$$f(x, y) = f(x)\ f(y).$$

Covariance and correlation

Two random variables X and Y (not necessarily independent)

$\mathbb{V}[X+Y] = \mathbb{E}\left[(X+Y-\mathbb{E}[X+Y])^2\right] =$

$\mathbb{E}\left[((X-\mathbb{E}[X])+(Y-\mathbb{E}[Y]))^2\right]$

$= \mathbb{V}[X] + \mathbb{V}[Y] + 2\mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right]$ (cross-multiply)

$= \mathbb{V}[X] + \mathbb{V}[Y] + 2\,\mathbb{COV}[X,Y]$

Covariance can also be written $\mathbb{COV}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Note that if X and Y are independent, then $\mathbb{COV}[X,Y] = 0$ and $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

The converse is NOT true: $\mathbb{COV}[X,Y]$ can be zero even if X, Y dependent.

Correlation: $\rho(X,Y) = \frac{\mathbb{COV}[X,Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$