Imperial College London

AIMS African Institute for Mathematical Sciences NEXT EINSTEIN INITIATIVE

STATISTICAL MODELLING OF INFECTIOUS DISEASES: Estimating parameters from data using likelihood

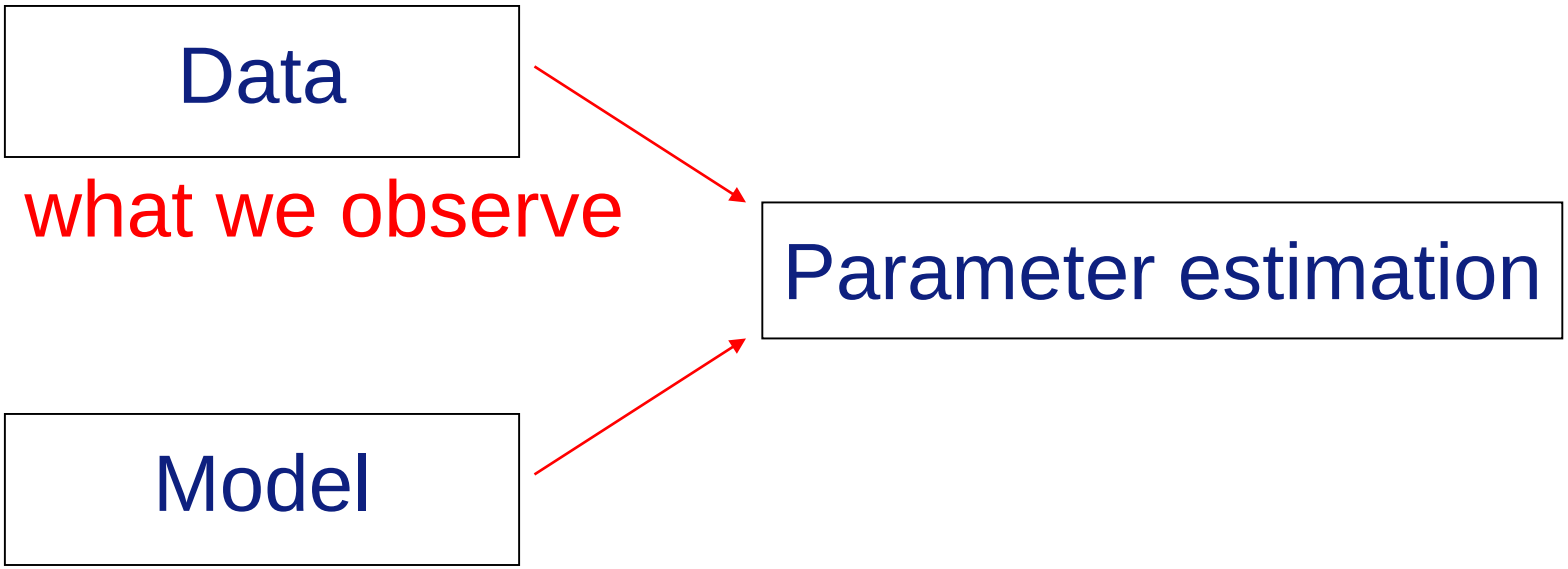
Nathan Green
(kindly provided by Christl Donnelly)

4th February 2020

Overview

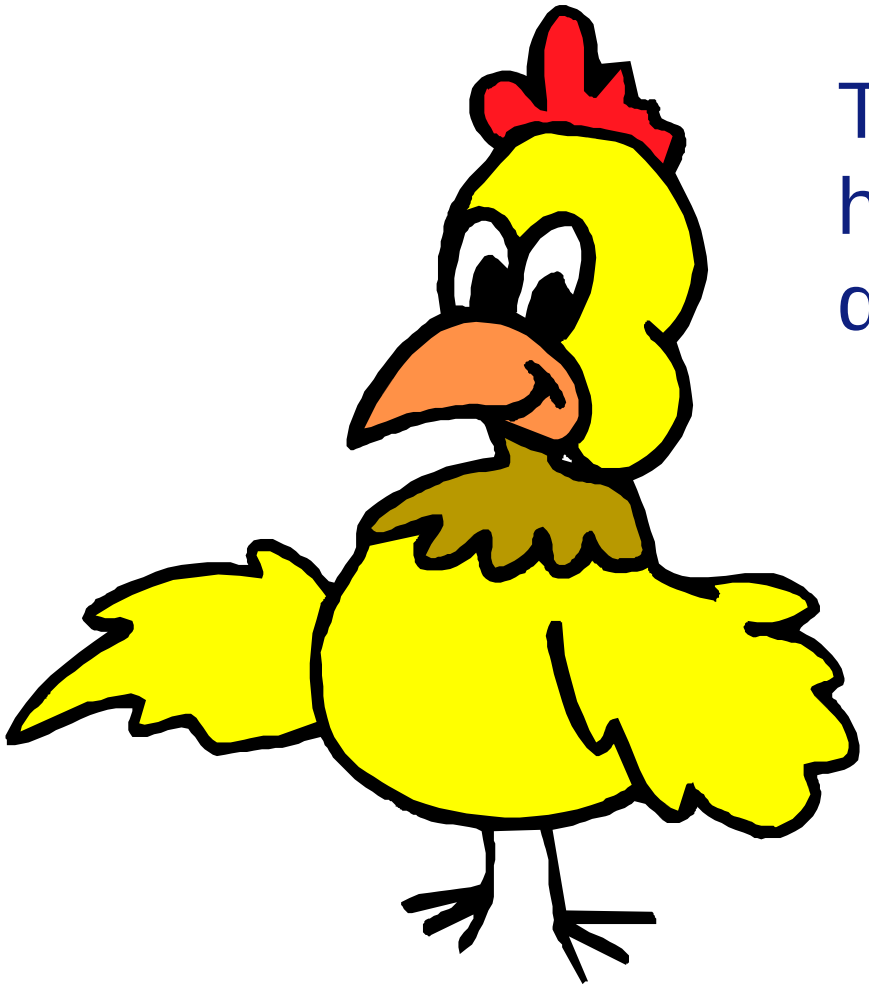
- Models and data
- Writing down the model
- Focussing on the parameter(s) of interest
- Writing down the likelihood of the observed data
- Maximizing the likelihood to obtain the parameter estimate
- Evaluating the uncertainty associated with the parameter estimate

Parameter estimation
– the basic requirements

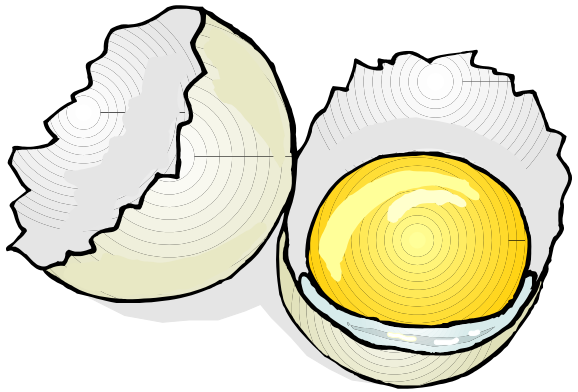


describes the process we believe generated the data and contains parameters

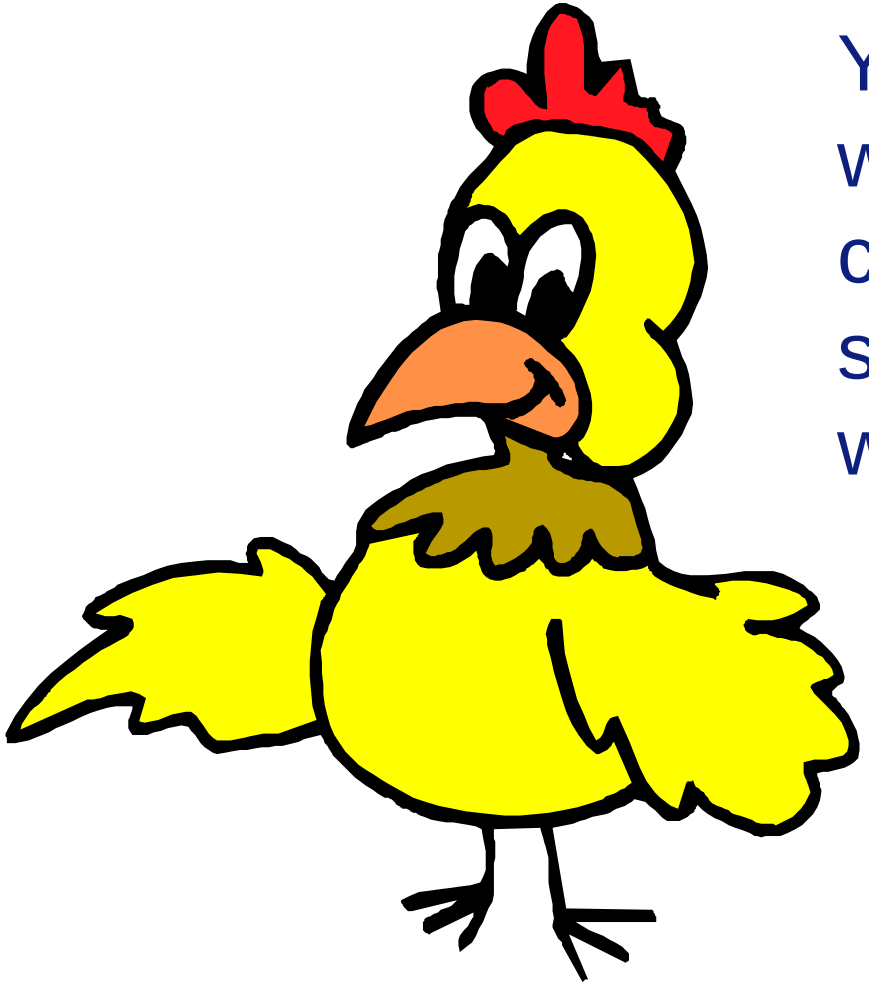
The chicken and the egg



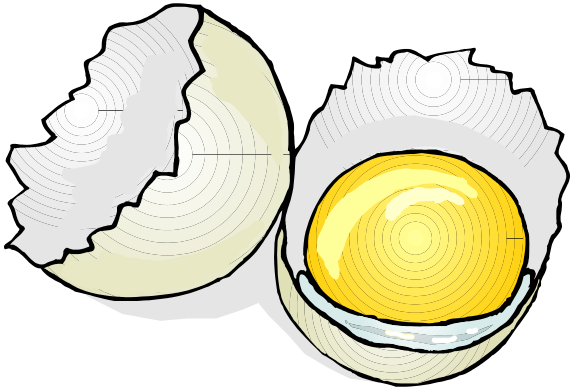
The model is believed to have generated the data.



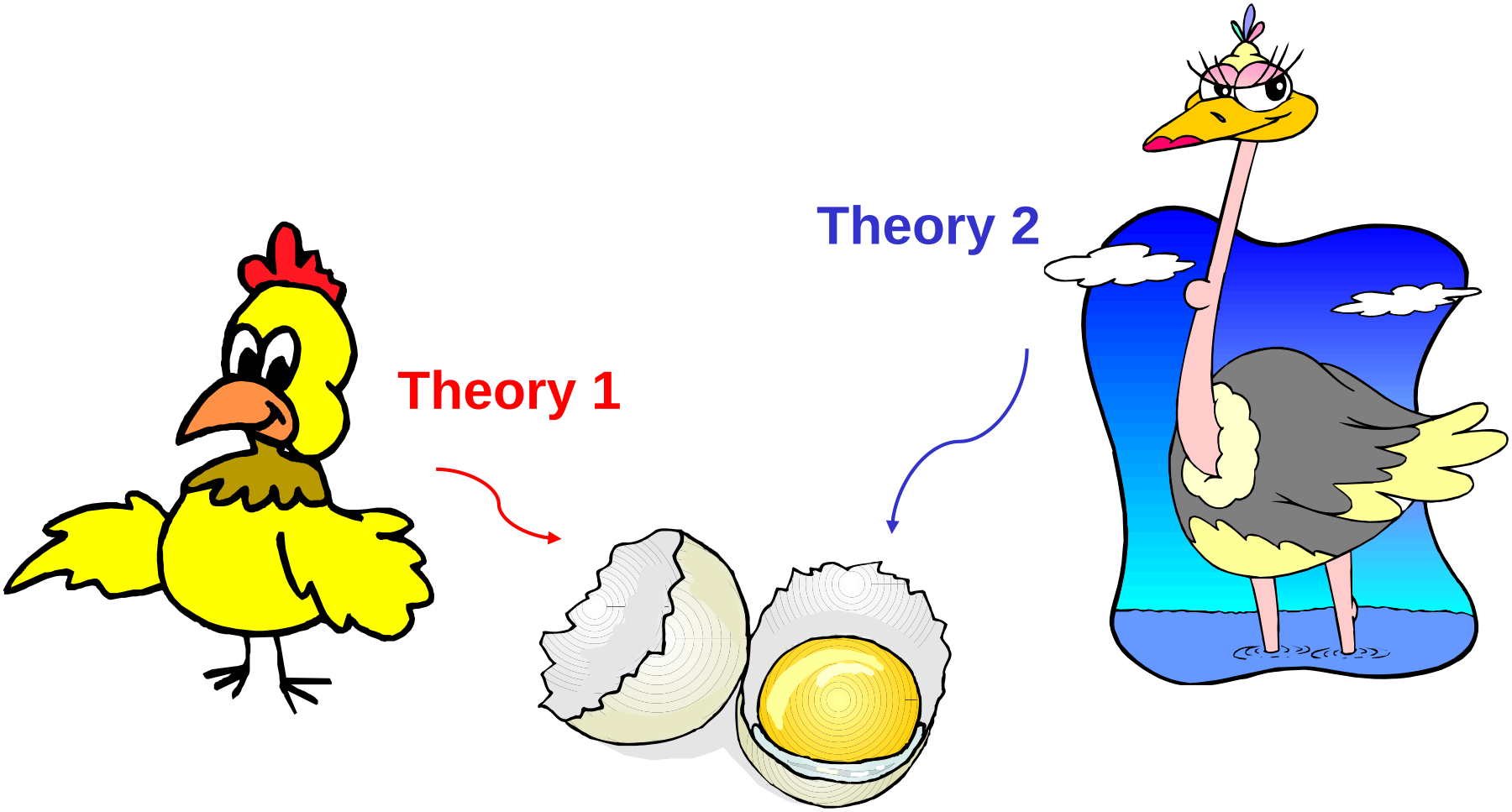
Which came first?
The chicken or the egg?



You may have started with the model and then collected the data or started with the data and written down the model.



Competing theories generate different models for the observed data



What experiment could be designed to distinguish between the two theories?

In parallel ask:

What parameter could be estimated to distinguish between the two theories?

We have a sample of eggs – an obvious feature to examine to decide between the chicken/ostrich theories is to measure the size of the eggs. This is the data.

Data collection

Imagine that our scientist here measures the eggs and obtains the dimensions.

We'll consider such data later…

Common statistical distributions

Discrete Data

- Binomial (counts of successes - out of n)
- Multinomial (counts of categories - out of n)
- Poisson (counts of events - no limiting n)

Continuous Data

- Normal
- Gamma, Weibull (e.g. survival times)

Likelihood (probability) of **Binomial** data

The data: x successes out of n trials

Let π = the probability of success

Likelihood $$L = \pi^{x}(1-\pi)^{n-x}$$

Log Likelihood $$l = x\ln(\pi) + (n-x)\ln(1-\pi)$$

Remember these **measures of disease frequency**?

Prevalence:

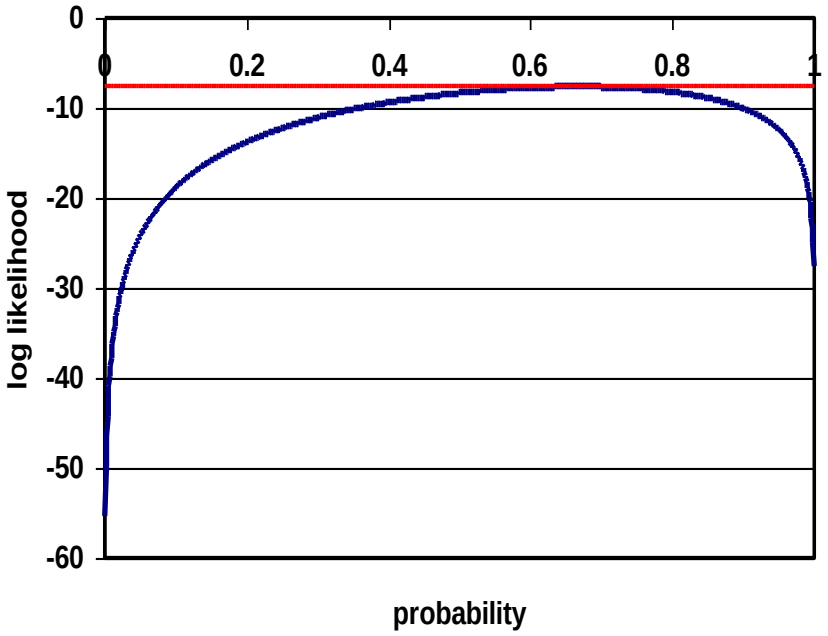
$$\frac{\text{number of existing cases of disease}}{\text{total population}}$$

Incidence:

$$\frac{\text{number of new cases of disease over a given time}}{\text{total population at risk}}$$

Binomial data set

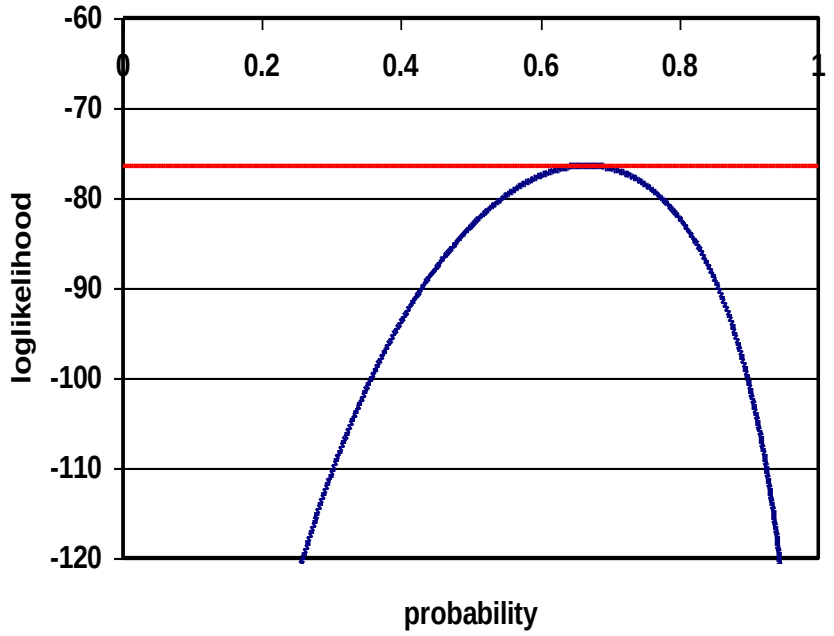
Of 12 people surveyed, 8 were infected with Ascaris. (n=12 x=8)



MLE = 8/12
= 0.667

A larger **Binomial** data set

Of 120 people surveyed, 80 were infected with Ascaris. (n=120 x=80)



MLE = 80/120
= 0.667

Likelihood of **Multinomial** data

k mutually exclusive outcomes: $C_1, C_2, \ldots, C_k$

The data: x_1 observations of C_1,
x_2 observations of C_2
...
where $x_1+x_2+ \ldots + x_k=n$

$$L = \pi_1^{x_1} \pi_2^{x_2} \ldots \pi_k^{x_k}$$

$$l = x_1 \ln(\pi_1) + x_2 \ln(\pi_2) + \ldots + x_k \ln(\pi_k)$$

$$(\pi_1 + \pi_2 + \ldots + \pi_k = 1)$$

Likelihood of **Poisson** data

The data: q counts recorded $x_1, x_2, \ldots, x_q$

$$L = \prod_i \frac{\mu^{x_i} e^{-\mu}}{x_i!} = \frac{\mu^{\sum_i x_i} e^{-q\mu}}{\prod_i x_i!}$$

$$l = \left(\sum_i x_i\right)\ln(\mu) - q\mu - \ln\left(\prod_i x_i!\right)$$

A data set of counts

Of 20 cities surveyed, 134 infected people were recorded. ($n=20 \quad \sum x_i=134$)

MLE = 134/20
= 6.7

Likelihood-based confidence intervals

You can think of the confidence interval containing all those parameters values which fit the data not significantly worse than the maximum likelihood estimates of the parameters.

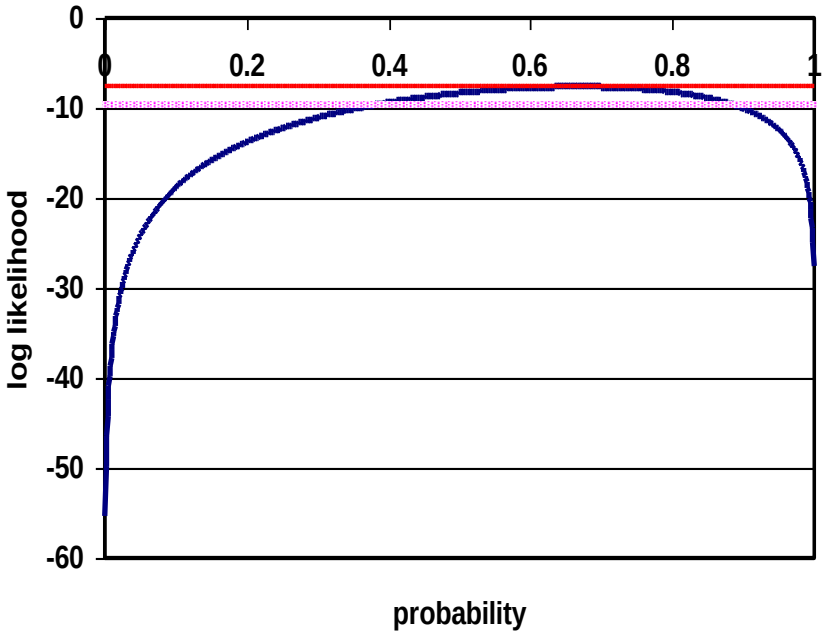
Likelihood-based confidence intervals

Mathematically, the 95% CI contains all those parameter values with log likelihood values within $\left(\chi^2_{df=p} / 2\right)$ of the maximum log likelihood

(where *p* is the number of parameters being estimated)

Binomial data set

Of 12 people surveyed, 8 were infected with Ascaris. (n=12 x=8)

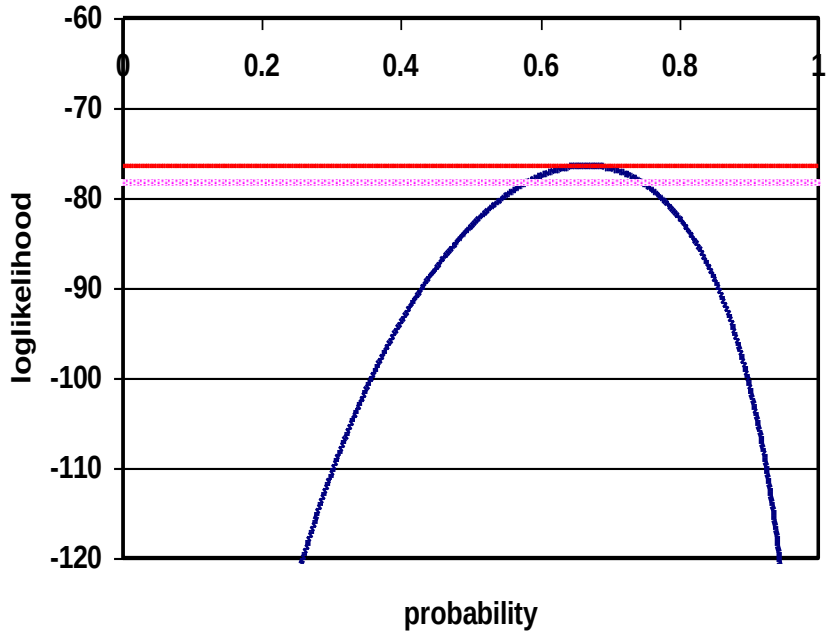


MLE = 8/12
= 0.667

95% CI:
(0.387,0.882)

A larger **Binomial** data set

Of 120 people surveyed, 80 were infected with Ascaris. (n=120 x=80)

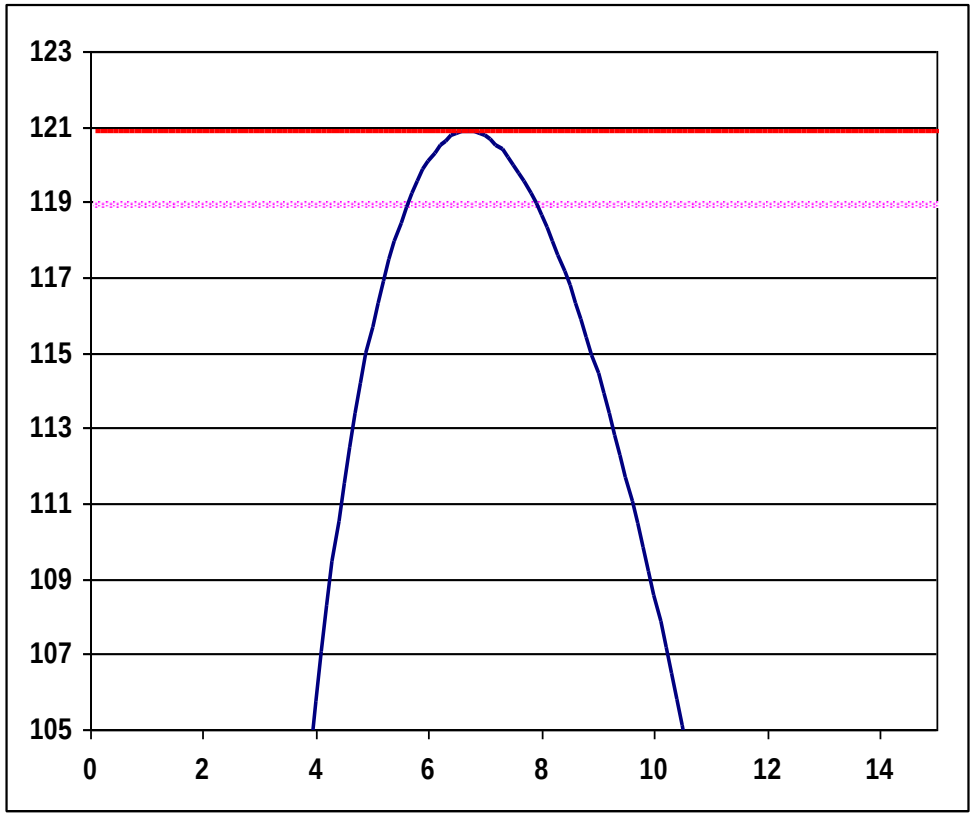


MLE = 80/120
= 0.667

95% CI:
(0.580,0.746)

A data set of counts

Of 20 cities surveyed, 134 infected people were recorded. ($n=20 \quad \sum x_i=134$)



MLE = 134/20
= 6.7

95% CI: (5.7,7.9)

Confidence Intervals and Hypothesis Testing

Another way to think of 95% confidence interval:

The 95% confidence interval contains all values that would be judged consistent with the data at the 5% significance level.

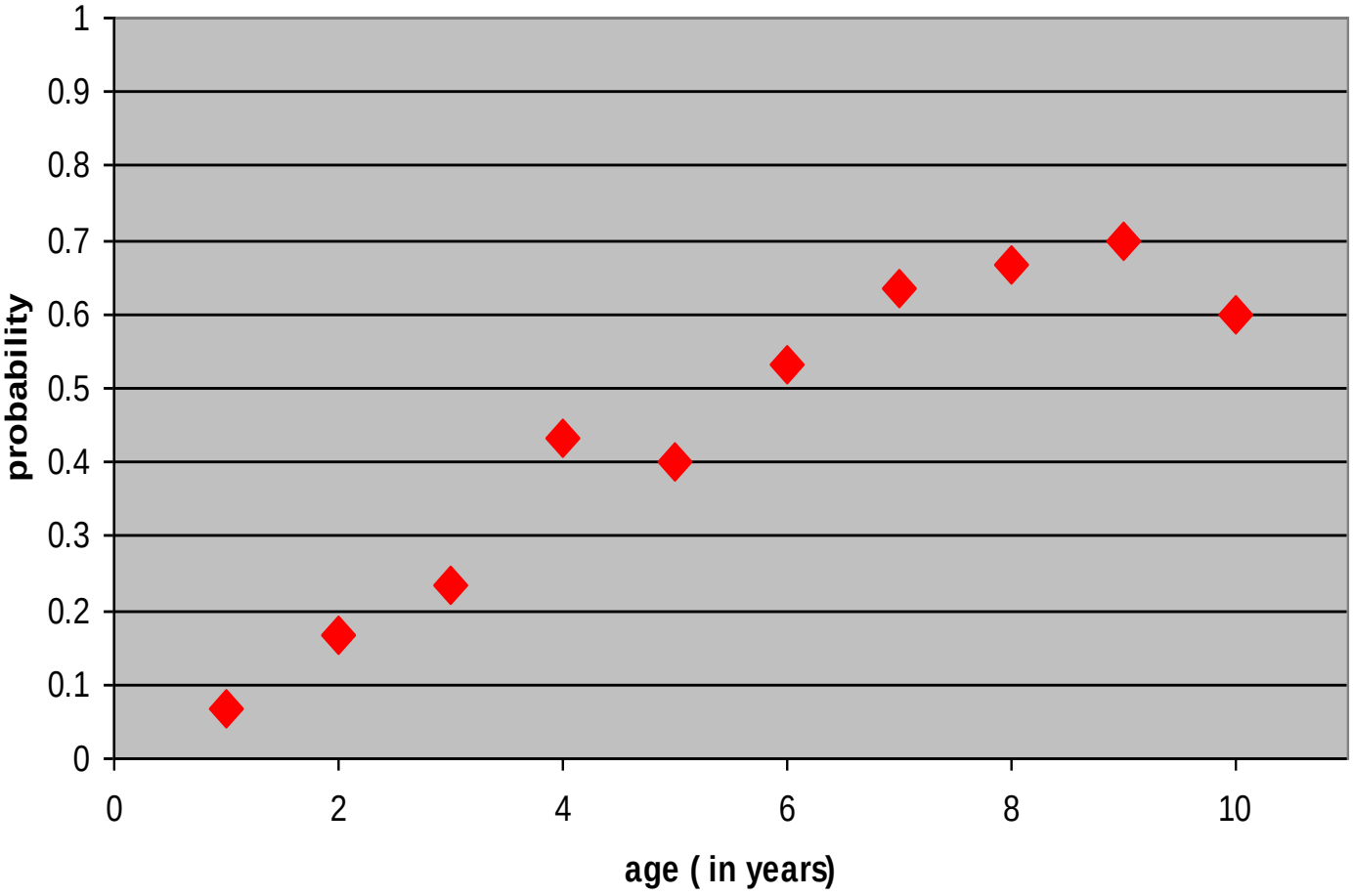
So if a value is outside the 95% CI, we can reject it.

A Maximum Likelihood Case Study: Prevalence & Logistic Regression

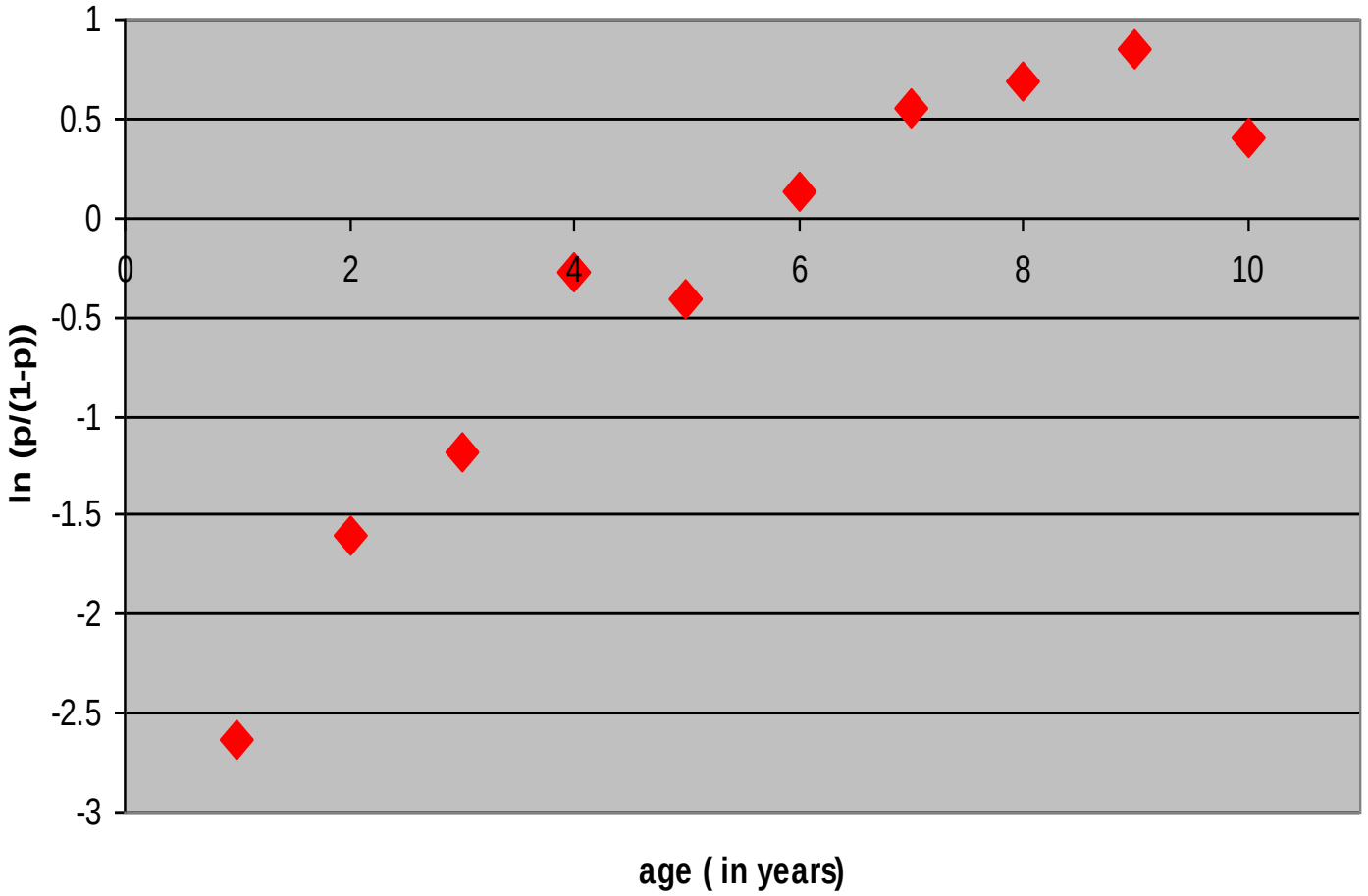
Our data arise from a survey of children (ages 1 to 10). 30 children were surveyed for each year of age. The following numbers were found positive for antibodies:

Age	1	2	3	4	5	6	7	8	9	10
Number positive	2	5	7	13	12	16	19	20	21	18

Age-Specific Prevalence Data



“Logit” of the Age-Specific Data



Logistic Regression Model

A model of association between a covariate a and a probability π. Thus, it is appropriate for the analysis of binary (binomial) data.

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta a \qquad \pi = \frac{\exp(\alpha + \beta a)}{1 + \exp(\alpha + \beta a)}$$

Logistic Regression Model (Linear Age)

Allowing for a linear age effect, we estimate parameters by maximising this likelihood:

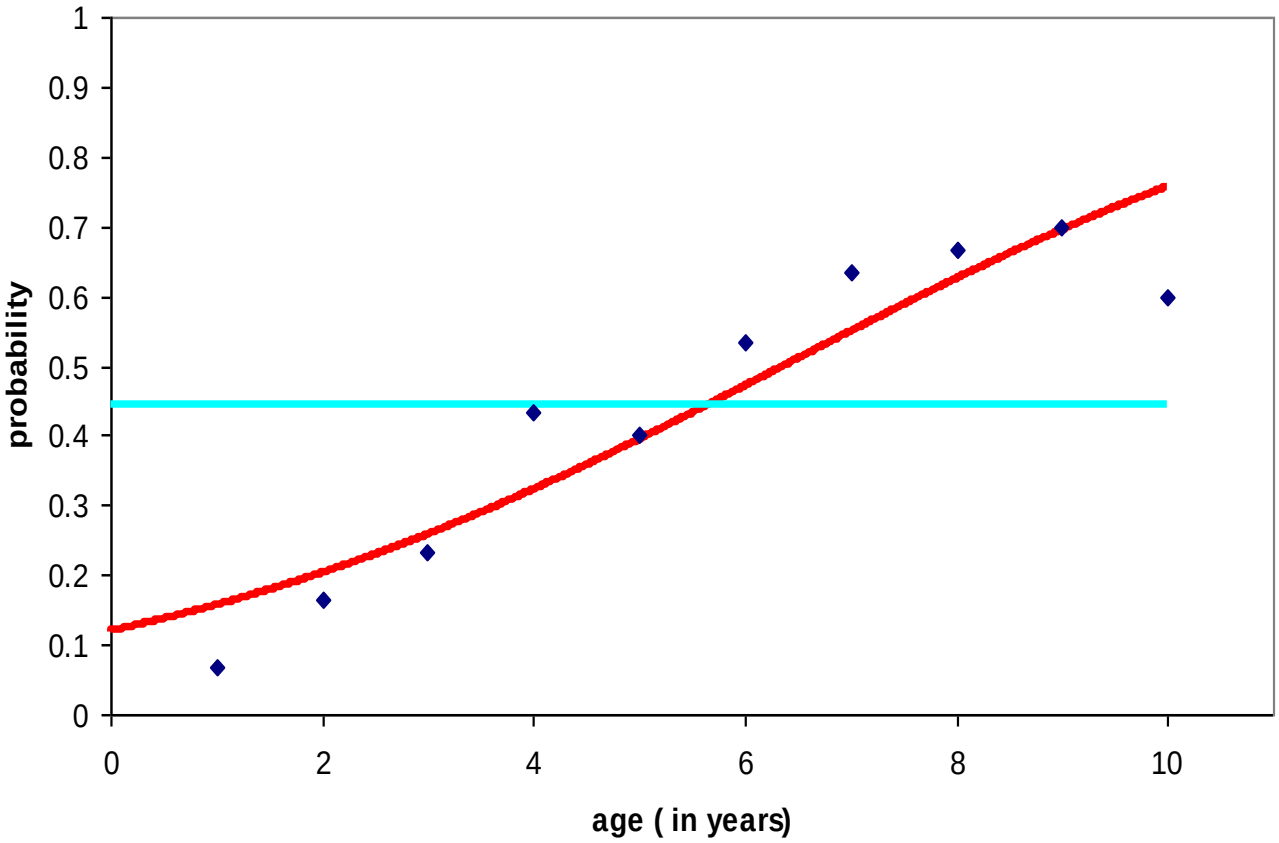
$$l = \sum x \ln\left(\frac{\exp(\alpha+\beta a)}{1+\exp(\alpha+\beta a)}\right) + (n-x)\ln\left(1-\frac{\exp(\alpha+\beta a)}{1+\exp(\alpha+\beta a)}\right)$$

Logistic Regression Results

We will allow for both linear and quadratic age effects and estimate parameters by maximising this likelihood:

Model	Estimates $\hat{\alpha}$	$\hat{\beta}$	Log Likelihood
Linear	-1.99	0.31	-180.635
No age effect	-0.23	0	-206.013

Logistic Regression Results (Linear Age)



Logistic Regression Results

To test the null hypothesis that the effect of age is linear (i.e. that $\beta=0$) we compare:

$$2*(-180.635 - -206.013) = 50.76$$

with a χ^2 with 1 degree of freedom
(p-value << 0.001)

The effect of age is highly significant!

The Saturated Log Likelihood

In some cases it is possible to calculated an upper limit on the log likelihood, known as the saturated log likelihood.

In the binomial case, for example, the upper limit for the log likelihood is reached when the fitted values match each observed proportion:

$$l = \sum x \ln\left(\frac{x}{n}\right) + (n-x)\ln\left(\frac{n-x}{n}\right)$$

Models galore!

Of course, the logistic regression model with a linear age effect is an extremely simple model for prevalence data and yields no insights into possible transmission mechanisms underlying this disease system.

The best (and most robust) insights are obtained with a transmission model (based for example on ordinary differential equations) is used to obtain the likelihood for the observed data.

Chain Models

We consider analysis based on SIR (susceptible, infected/infectious, recovered) in discrete time.

For some diseases (chicken pox, measles and mumps, for example) the latent period is **long** relative to the infectious period and **neither period varies much** between individuals.

In such cases, it may be possible to identify the 'generation' in which an individual has been infected.

Generations of infection?

- By **generation** we mean the number of predecessors in the transmission chain tracing back to the introductory case (or cases).

- In real applications, it is only possible to distinguish the first few generations (after which the data are too ‘messy’).

- However, under some conditions, chain models may be used to derive the distribution for the total size of an outbreak (so that individual generations don’t have to be distinguished).

Chain Binomial Models

- These are models for spread within small groups.

- The infectious period is reduced to a single time point.
- The discrete time unit equals the latency period.

- Suggested
  - by Reed & Frost in the US in 1928 and similarly
  - by Greenwood in England in 1931

- The assume that the chance of >1 person in any household/group is independently infected from outside is negligible.

Household of two individuals

One is infectious at time 0.

At time 1, the other is

- Infectious (with probability p)
- Or not (with probability q, where p+q=1)

So two chains are possible:

Chain	Probability	Chain notation
1	q	{1}
1→ 1	p	$\{1^2\}$

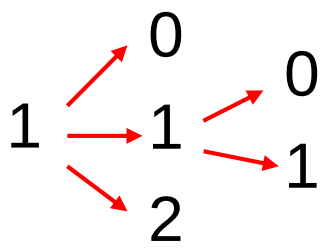
Household of three individuals

One is infectious at time 0.

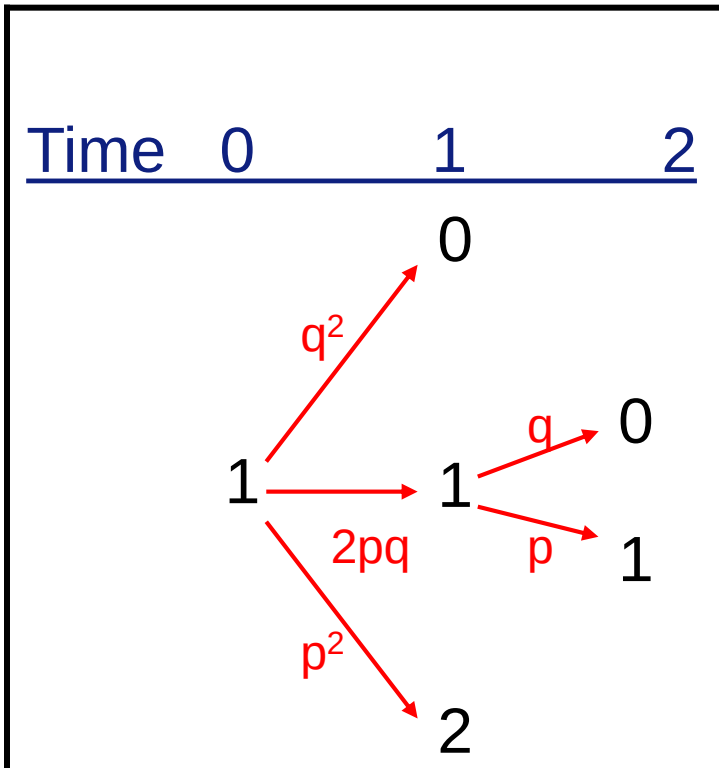
At time 1, there may be 0, 1 or 2 infectious people with probabilities q^2, $2pq$ and p^2.

[since the two susceptibles at time 0 are infected – or not – independently]

At time 2, there may be 0 or 1 infectious people.

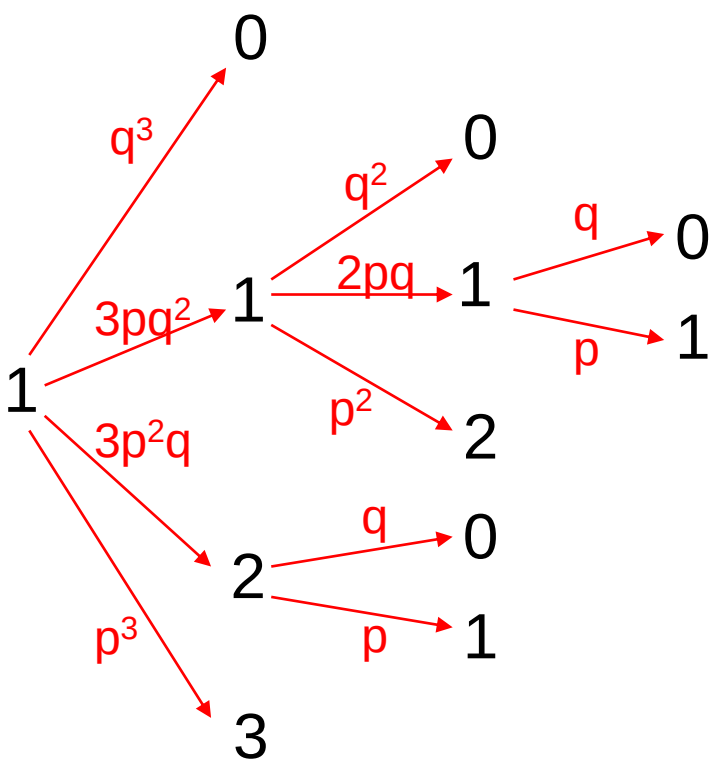


Household of three individuals



Probability	Chain notation	Total infections
q^2	{1}	1
$2pq^2$	$\{1^2\}$	2
$2p^2q$	$\{1^3\}$	3
p^2	{12}	

Household of four individuals

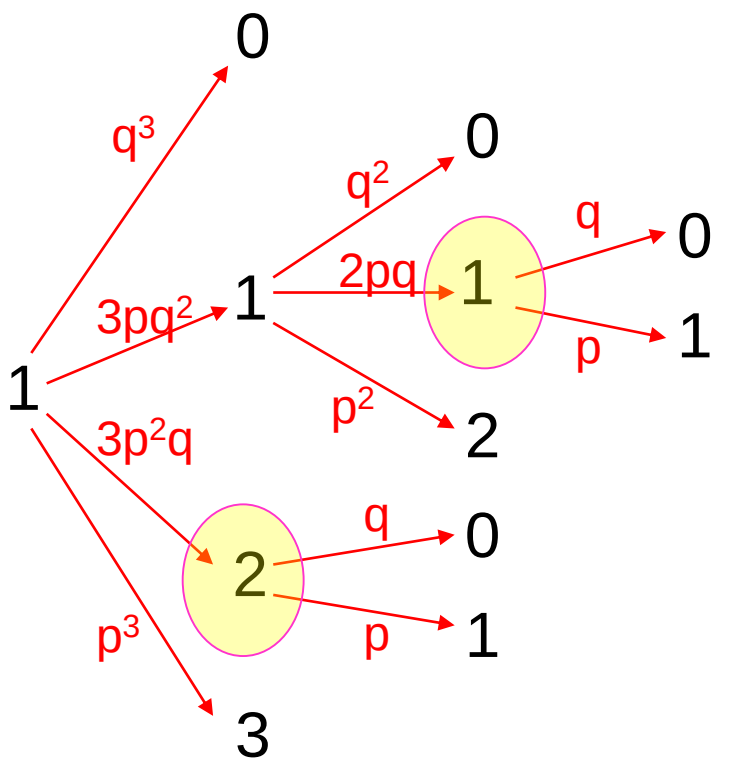


Probability	Chain notation	Total infections
q^3	{1}	1
$3pq^4$	$\{1^2\}$	2
$6p^2q^4$	$\{1^3\}$	3
$6p^3q^3$	$\{1^4\}$	4
$3p^3q^2$	$\{1^22\}$	4
$3p^2q^2$	{12}	3
$3p^3q$	{121}	4
p^3	{13}	4

Household of four individuals

Time

0 1 2 3



Basic assumption

S_t = num who become infectious at time t

R_t = num who remain susceptible at time t

$S_t \sim \text{Binomial}(\pi, R_{t-1})$

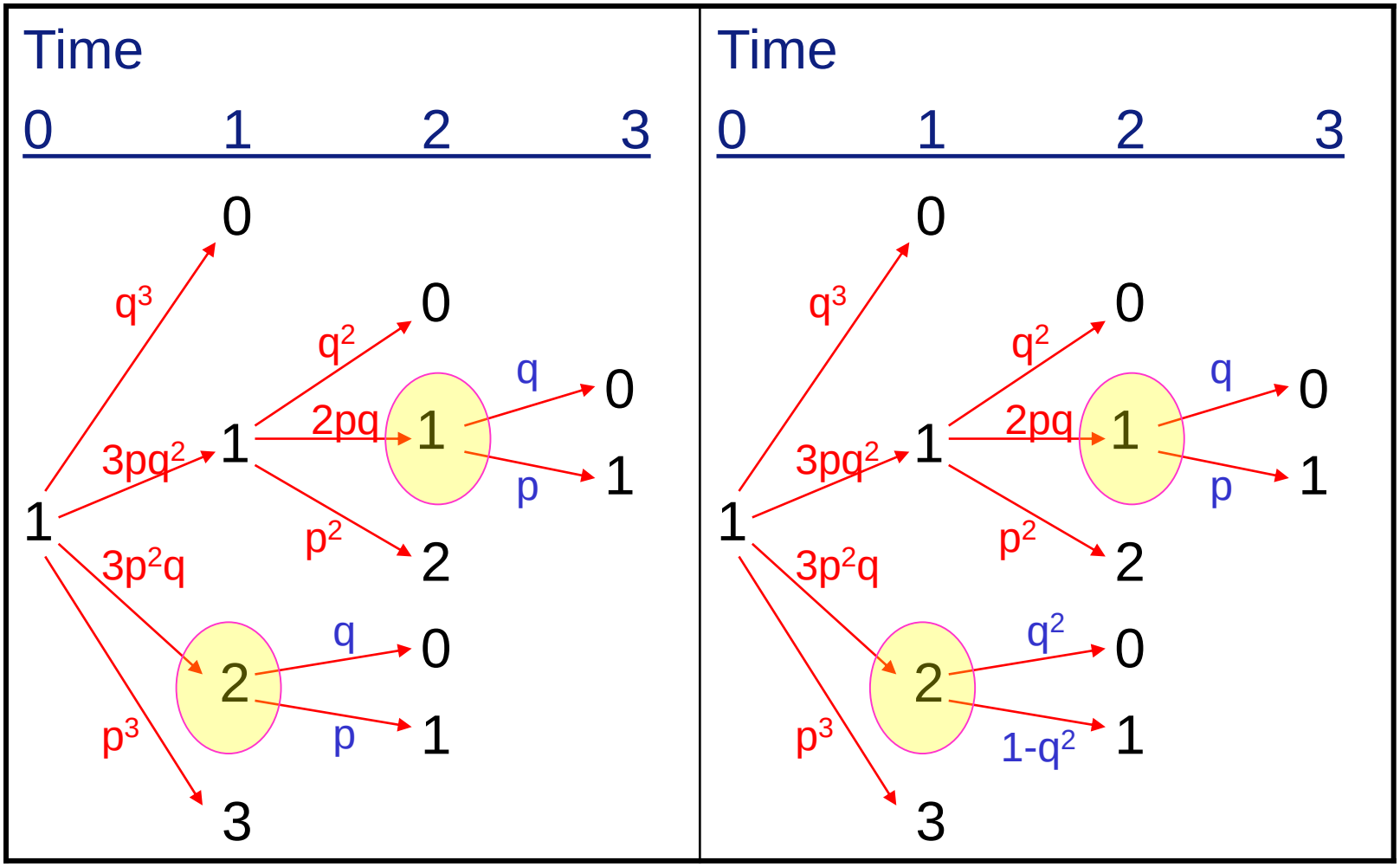
Greenwood model (1931) = π constant with respect to multiple infectious people

Reed and Frost

- They had a different model
- S_t ~ Binomial (π,R_{t-1}) with $\pi=1-q^{(S_{t-1})}$
- So each infectious individual has a (1-q) probability of transmitting infection (with infection probabilities being independent).
- Won't affect things when household has 2 or 3 individuals but will affect if 4 or more!

Greenwood

Reed and Frost



Can, of course, compare the fit of these two competing models if you have sufficiently detailed data.

Possible elaborations

- Immunity developing within the family/household
- Introduction of more than one infectious individual
- Measurement (counting) errors
- Variation of π in individuals or households

  For example, you could regard q as arising from a beta distribution. The probability of each chain is obtained by integration over the distribution of q.

Providence measles data

Type of chain	Expect number of households	Observed in Providence data	Fitted values
{1}	nq^2	34	14.9
$\{1^2\}$	$n2pq^2$	25	23.5
$\{1^3\}$	$n2p^2q$	36	87.7
{12}	np^2	239	207.9
Total	n	334	334.0