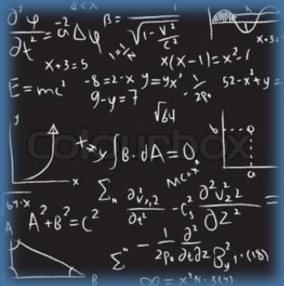


Models and Data

Introduction to Model Fitting

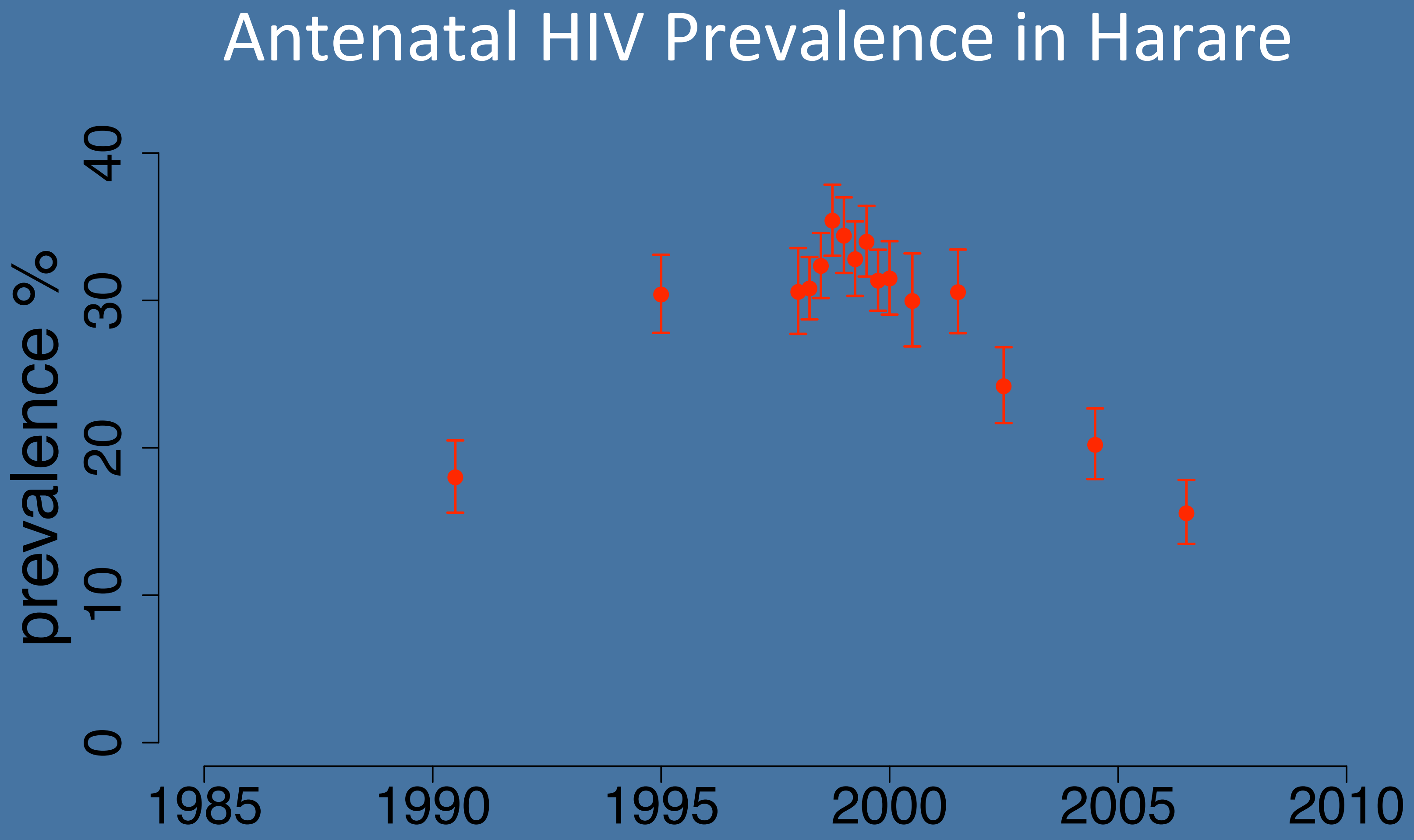
Clinic on Dynamical Approaches to Infectious Disease Data

December 16, 2014

Steve Bellan, MPH, PhD

Post-doctoral Researcher, University of Texas at Austin

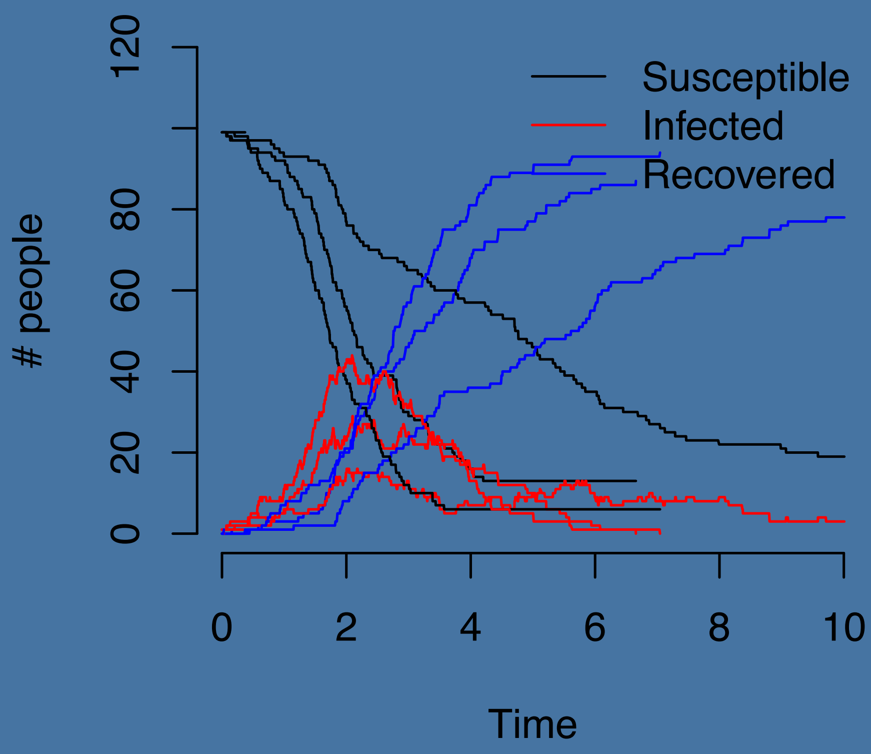
What happened?



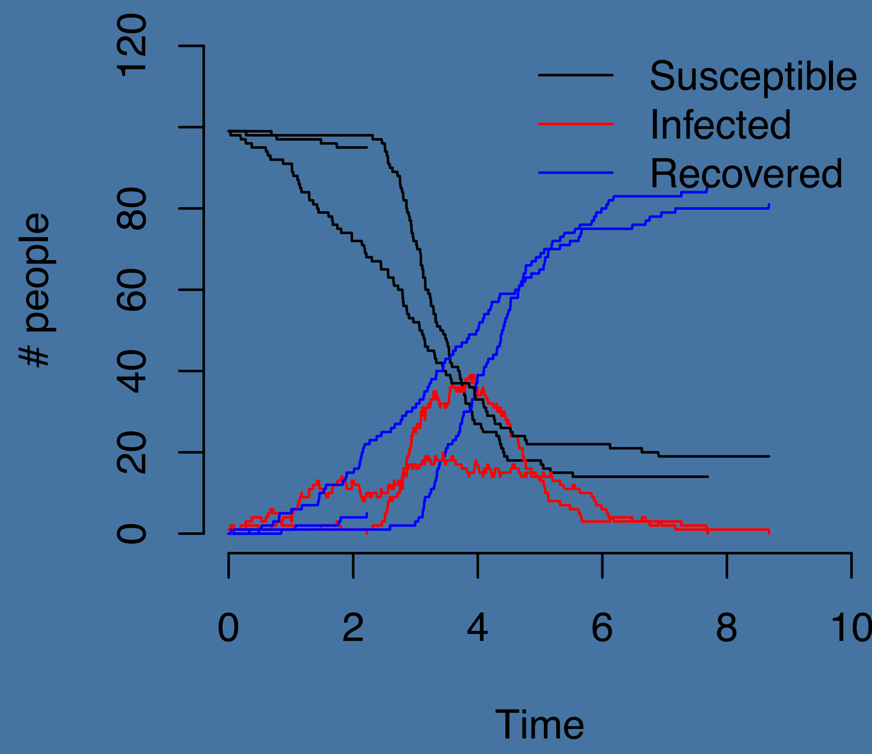
Are these different?

Measles Outbreaks

5 Urban Villages



5 Rural Villages



Why fit models to data?

- Estimate quantities/parameters of interest
- Inference: Test hypotheses
- Model assessment:

  Assess plausibility or model comparison
- End goal: explain observed patterns or predict

Statistical Models

- A familiar starting point

- Analogous to fitting dynamical models

- Abstraction of real relationships

- Explaining variation in data through correlational relationships (hopefully causal)

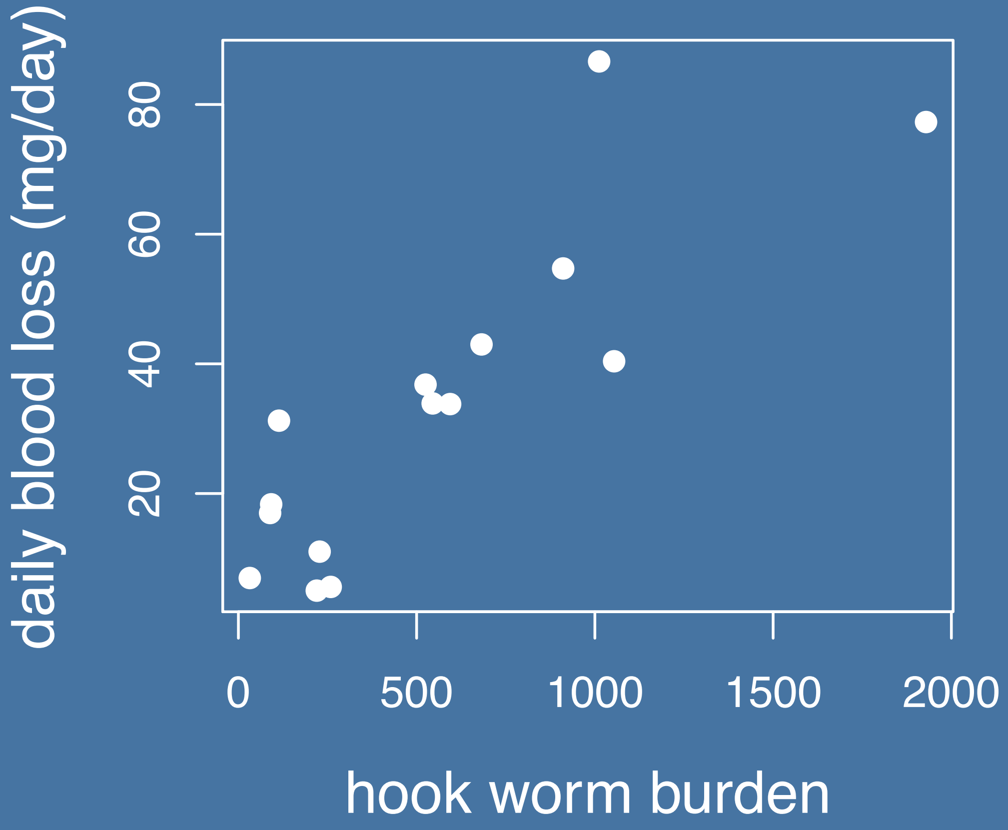
Dynamic Models and Time Series Data

- Dynamic models evolve through time
- and simulate time series
- Informally compare observed time series & simulated time series
- Fitting models to data formally compares them

Linear Regression

How does hook worm burden affect blood loss?

Is there any relationship?



Data in Epicalc R Library taken from Areekul et al. (1970).

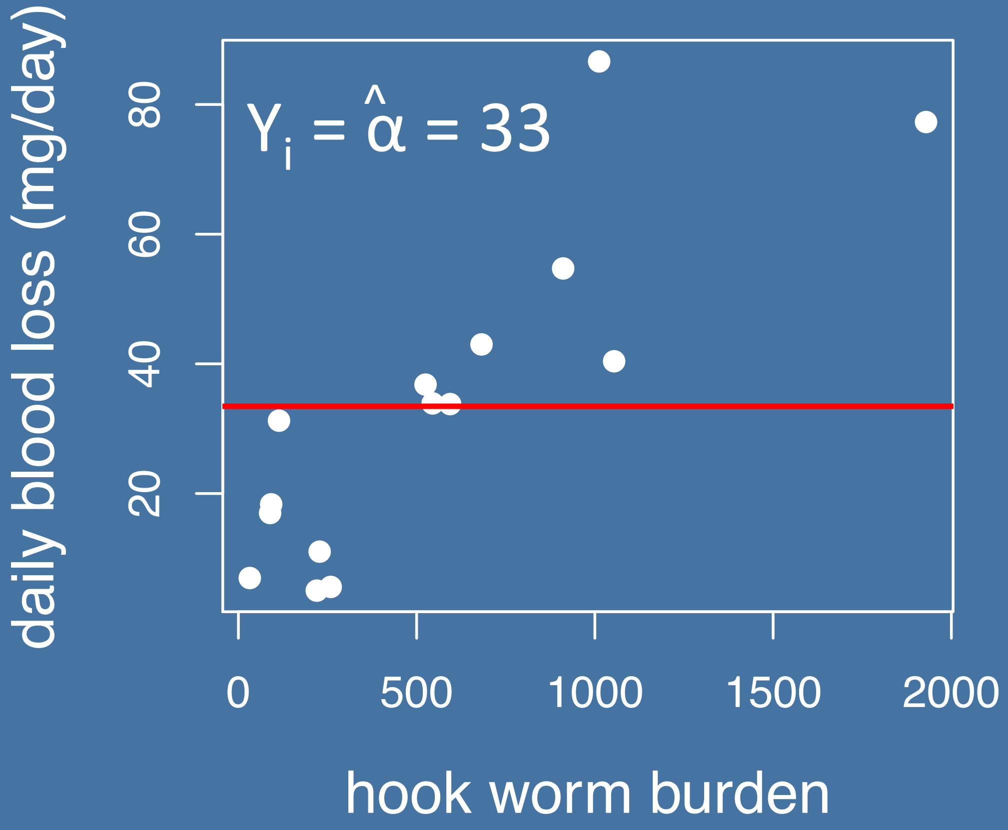
Linear Regression

Null hypothesis: No relationship

$Y = \alpha$

Is this a good fit?

How can we get a better fit, or the best fit?



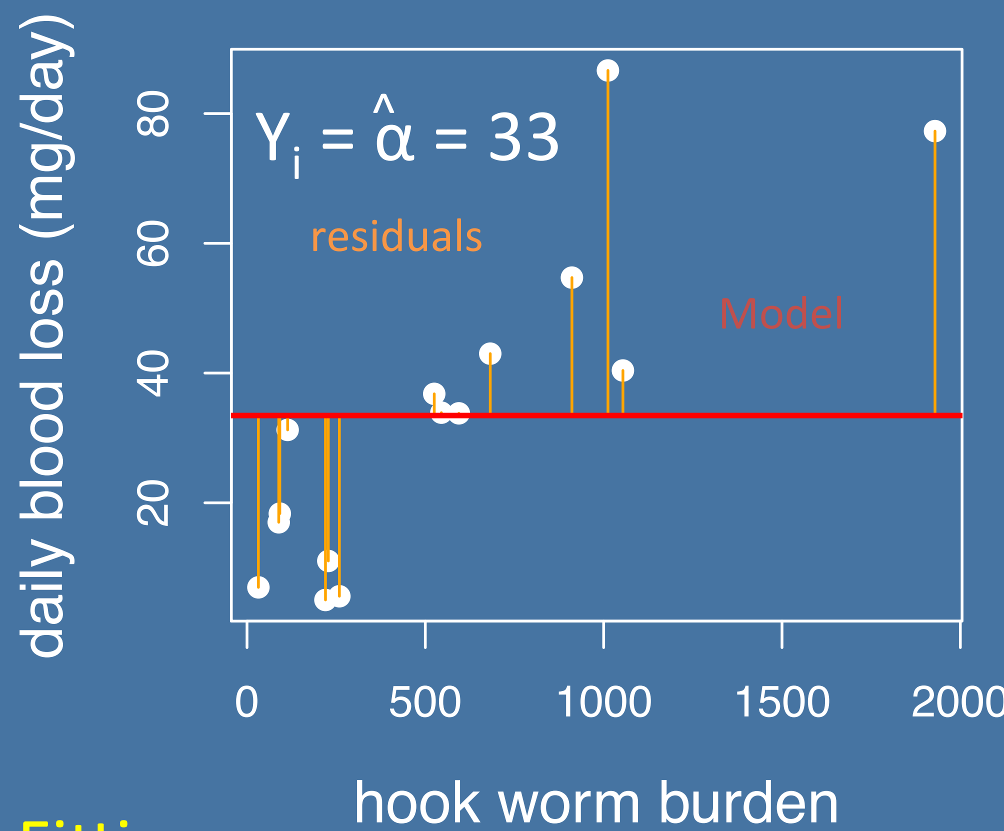
Linear Regression

Null hypothesis: No relationship

$$Y_i = \alpha + \varepsilon_i$$

Is this a good fit?

How can we get a better fit, or the best fit?



One option is Least Squares Fitting

Choose a line $Y = \hat{\alpha} + \hat{\beta}X$ to minimize $\Sigma(\text{residuals})^2$

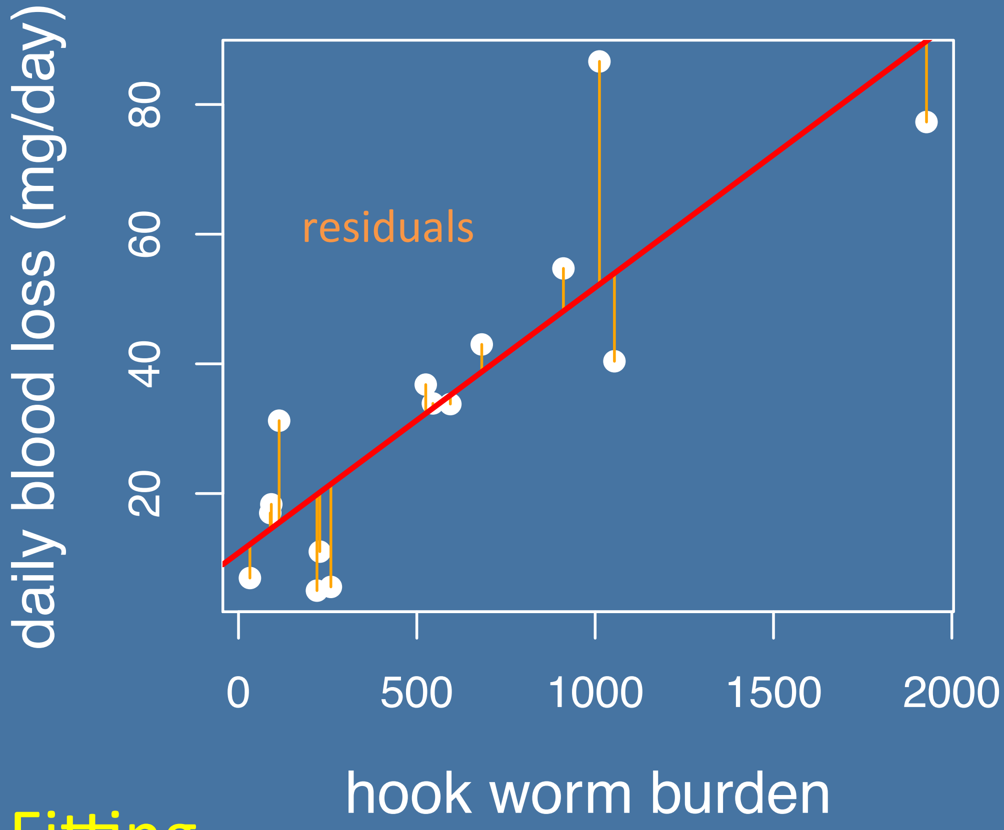
Linear Regression

Null hypothesis: No relationship

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Is this a good fit?

How can we get a better fit, or the best fit?



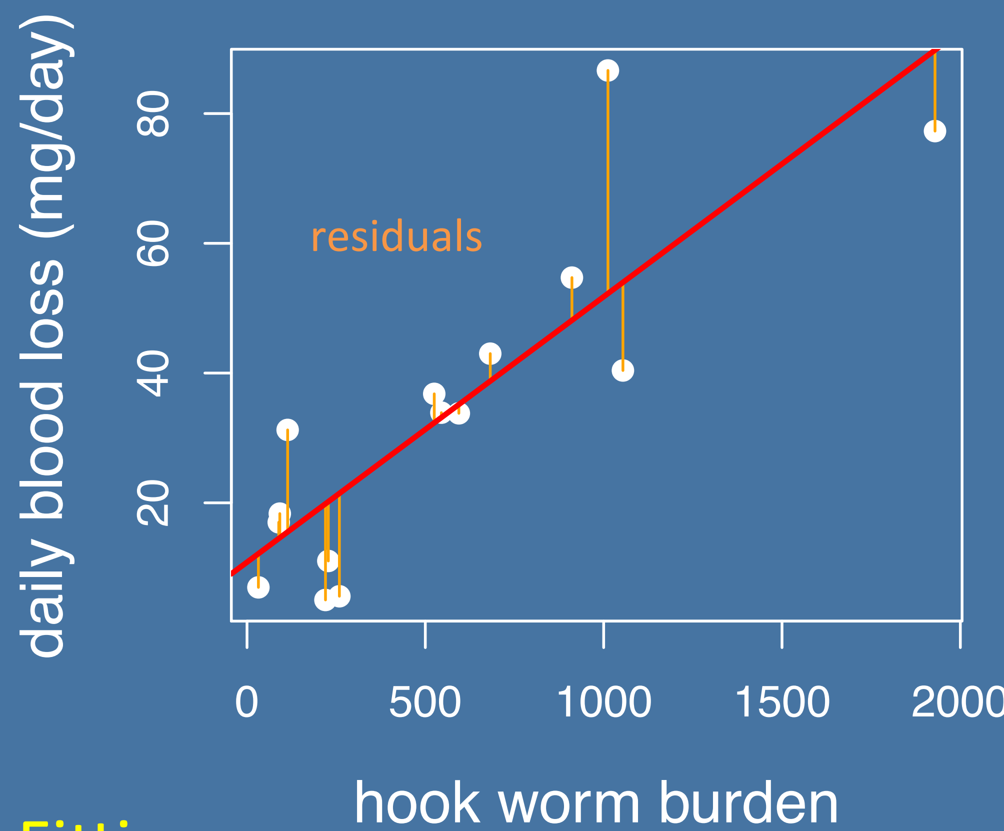
One option is Least Squares Fitting

Choose a line $Y = \hat{\alpha} + \hat{\beta}X$ to minimize $\Sigma(\text{residuals})^2$

Linear Regression

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- Y_i: expected daily blood loss
- α: intercept
- βX_i: effect of hook worm burden (X_i: hook worm burden)
- ε_i: error



One option is Least Squares Fitting

Choose a line $Y = \hat{\alpha} + \hat{\beta}X$ to minimize $\Sigma(\varepsilon_i)^2$

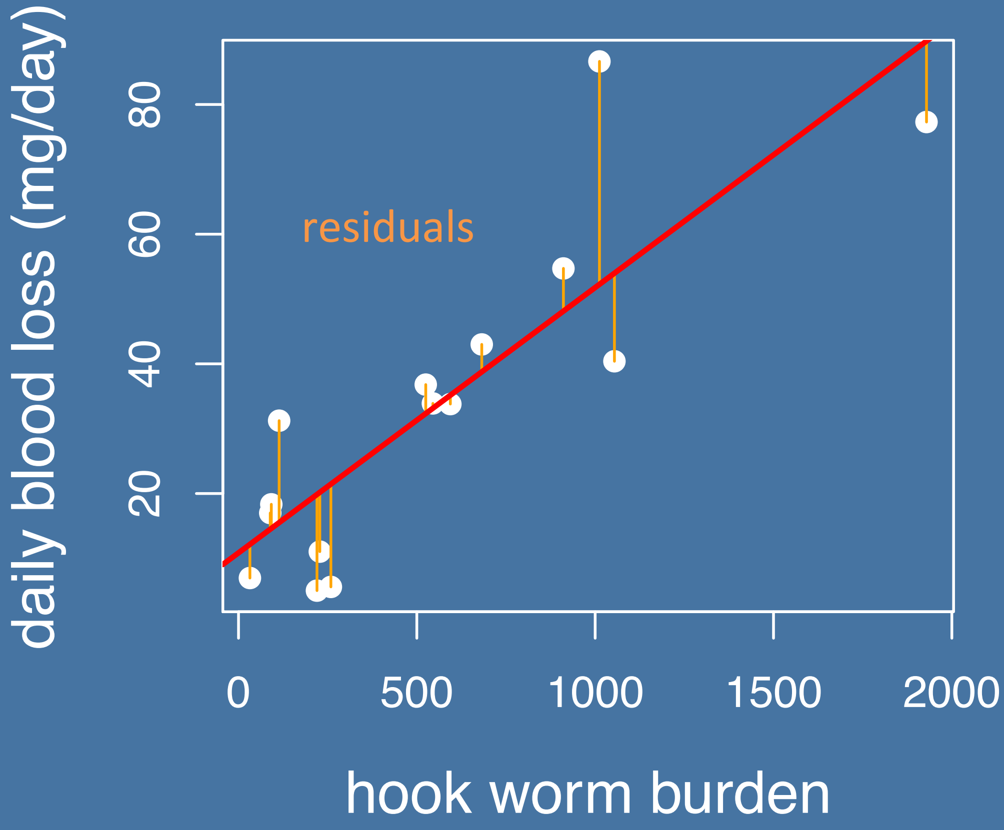
Linear Regression

Another option is

Maximum Likelihood

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$



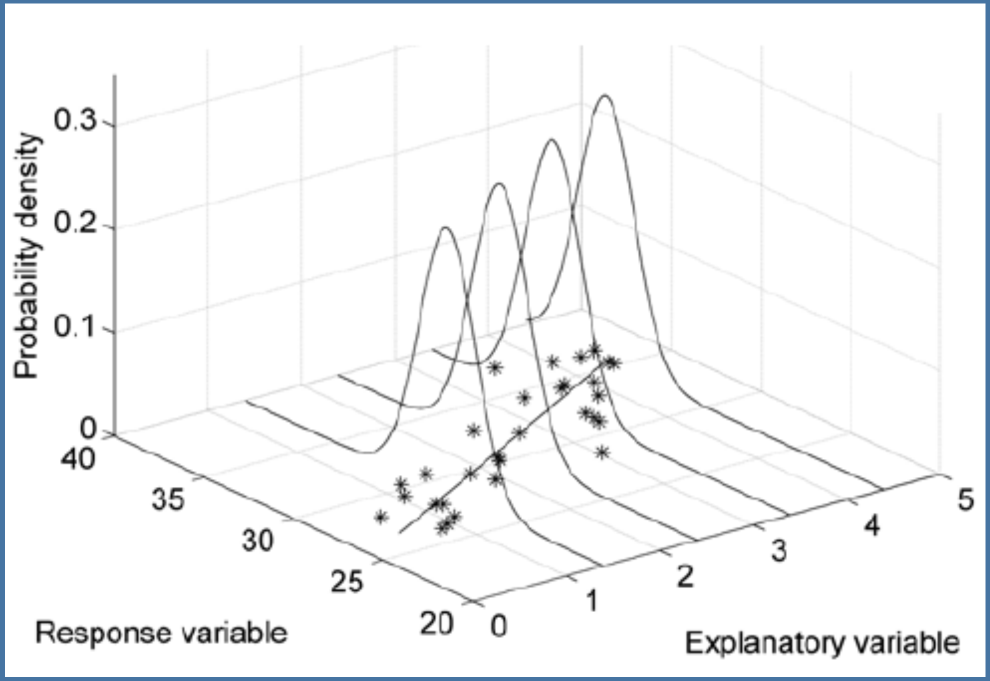
Choose $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$ to maximize the likelihood

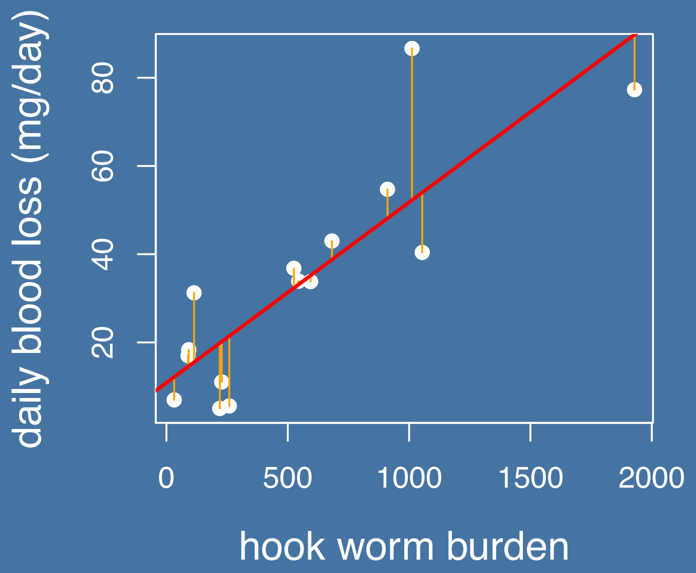
i.e. probability of observed data given a model

Linear Regression

Maximum Likelihood

$$Y_i \sim N(\alpha + \beta X_i,\ \sigma^2)$$





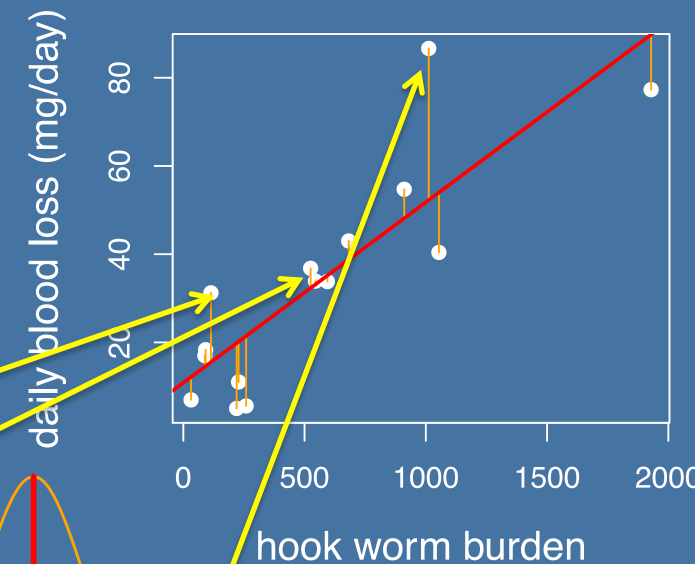
Choose $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$ to maximize the likelihood

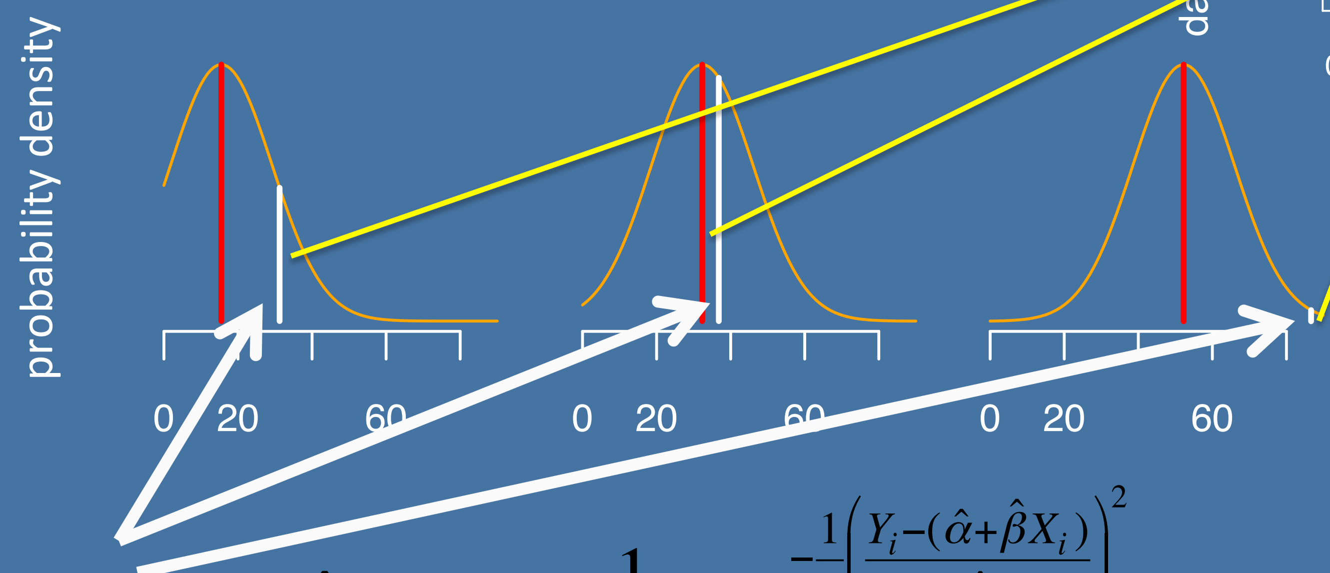
i.e. probability of observed data given a model

Linear Regression

Maximum Likelihood

$$Y_i \sim N(\alpha + \beta X_i,\ \sigma^2)$$



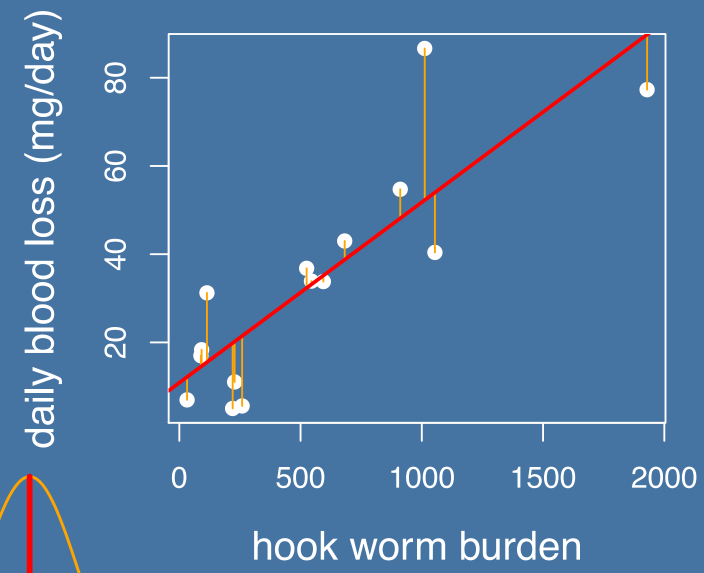


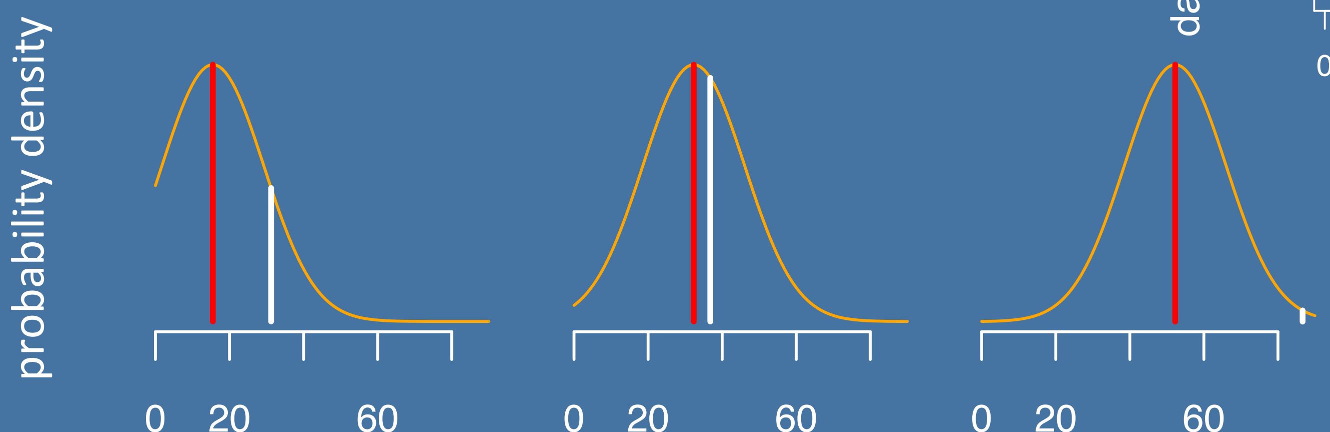
$$P(Y_i \mid \hat{\alpha}, \hat{\beta}, \hat{\sigma}) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i - (\hat{\alpha} + \hat{\beta} X_i)}{\hat{\sigma}}\right)^2}$$

Linear Regression

Maximum Likelihood

$$Y_i \sim N(\alpha + \beta X_i,\ \sigma^2)$$

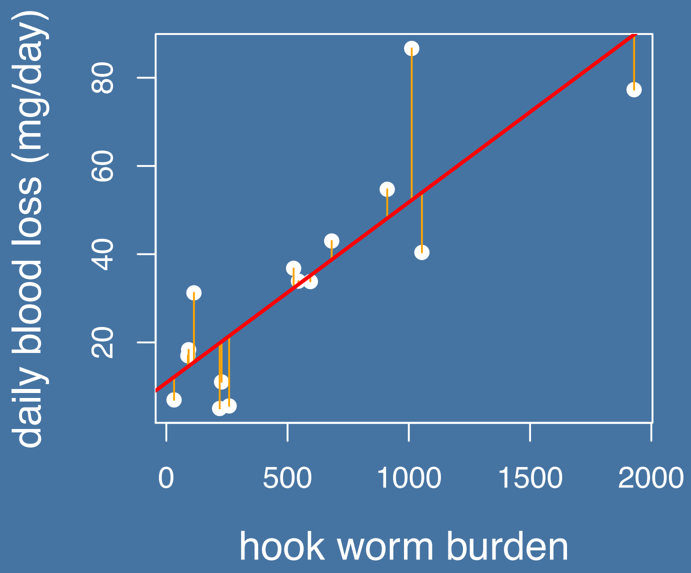




$$P(Y_1,\ldots,Y_n \mid \hat{\alpha},\hat{\beta},\hat{\sigma}) = \prod_{i=1}^{n} P(Y_i \mid \hat{\alpha},\hat{\beta},\hat{\sigma})$$

Linear Regression

Maximum Likelihood



function of data

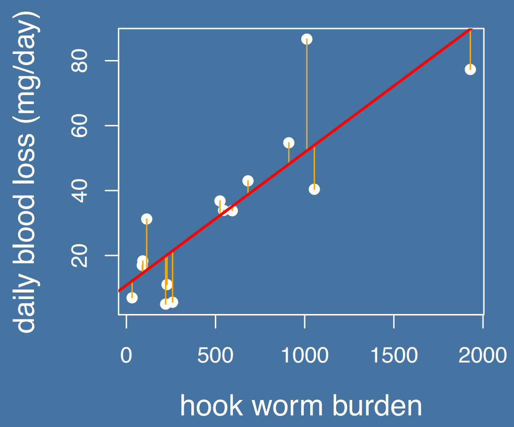
PDF: $$P(Y_1,\ldots,Y_n \mid \hat{\alpha},\hat{\beta},\hat{\sigma}) = \prod_{i=1}^{n} P(Y_i \mid \hat{\alpha},\hat{\beta},\hat{\sigma})$$

LIKELIHOOD: $$L(\hat{\alpha},\hat{\beta},\hat{\sigma} \mid Y_1,\ldots,Y_n) = \prod_{i=1}^{n} P(Y_i \mid \hat{\alpha},\hat{\beta},\hat{\sigma})$$

function of parameters

Linear Regression

Parameter Estimation & Inference



Null hypothesis: $\beta = 0$

$\hat{\beta} = 0.04$

P(estimating a β this extreme |null)

P = 6.99e-05 < 0.05,

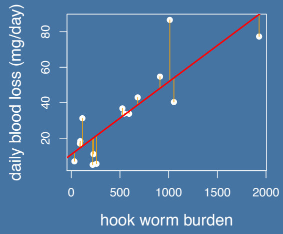
so we reject the null hypothesis.

Confidence intervals

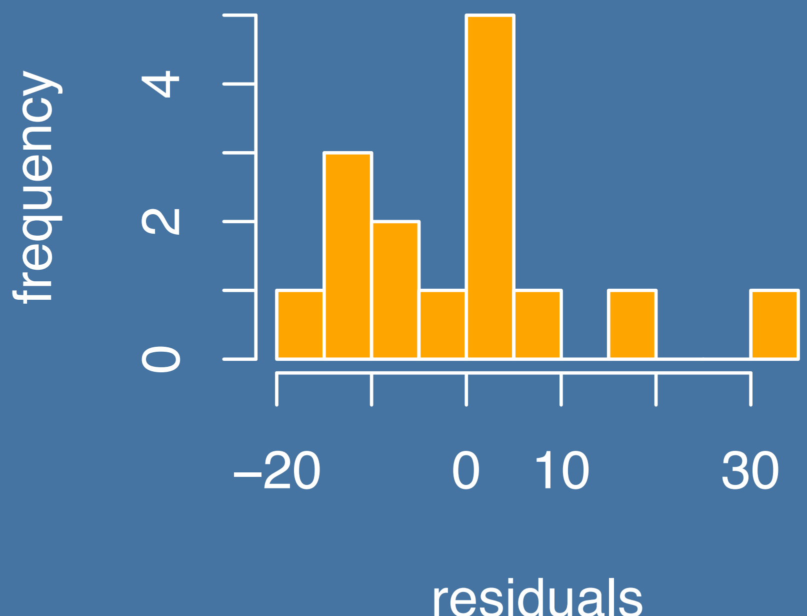
Collection of
non-rejectable null hypotheses

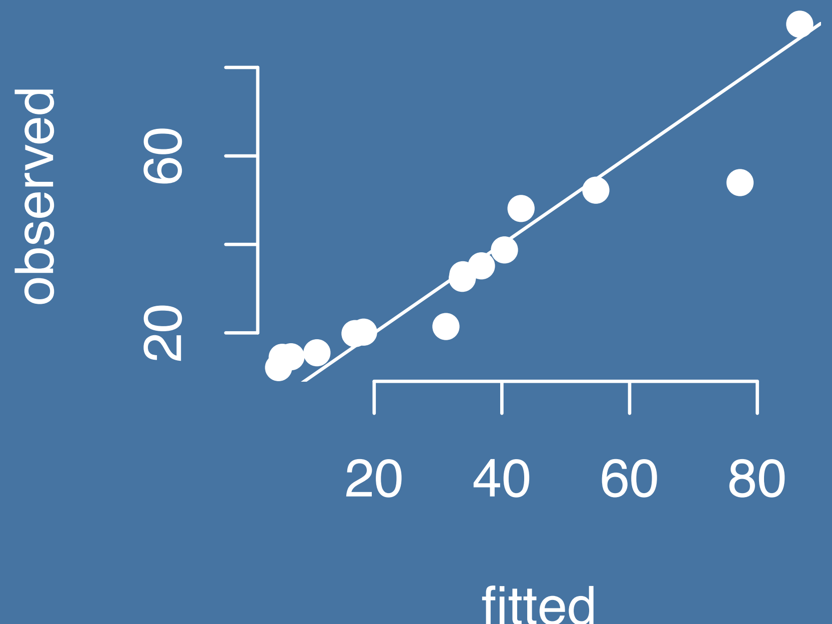
$\hat{\beta} = 0.04\ (0.025, 0.056)$

Is it a good model:
Checking Assumptions

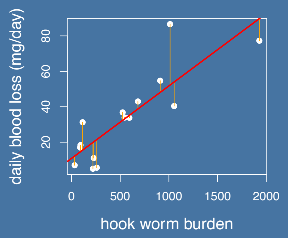


Normality

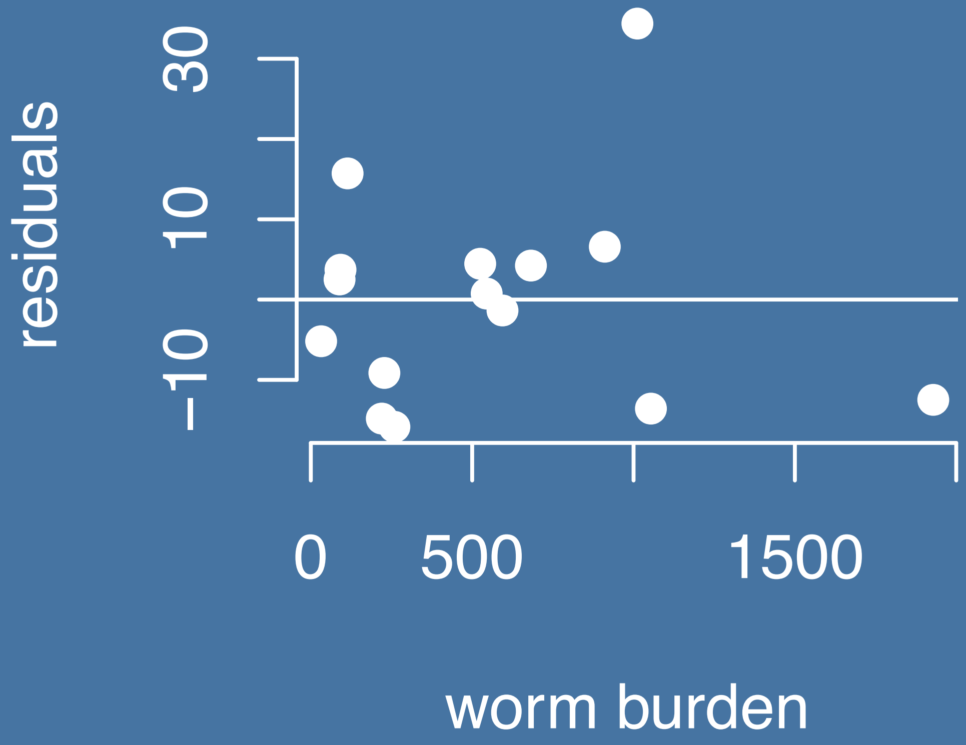




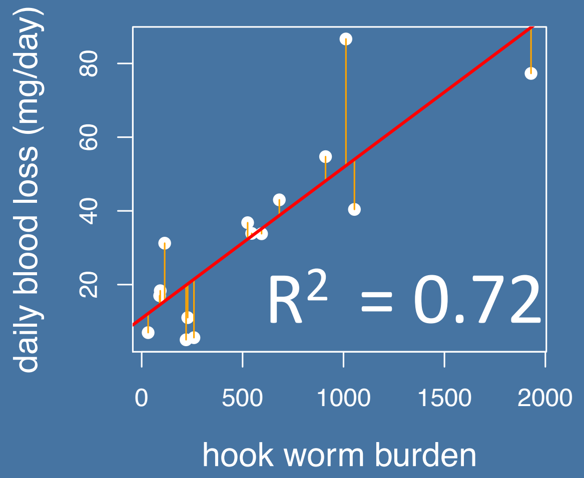
Is it a good model:
Checking Assumptions



Linearity Independence Constant Variance



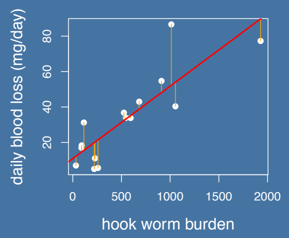
Is it a good model: Goodness of Fit



R^2 = (correlation coefficient)^2

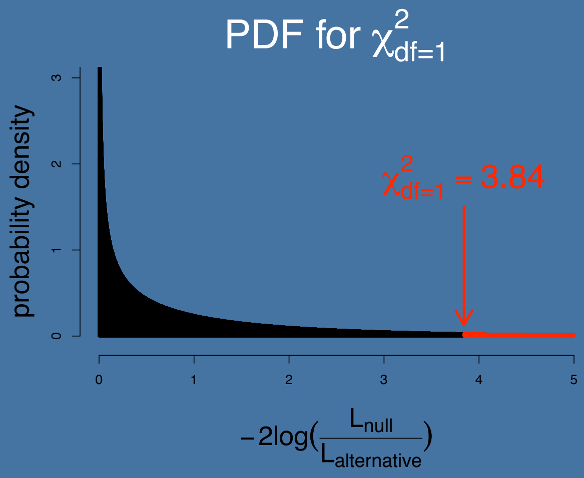
How much of the variation in Y is explained by the model?

Is it a good model:
Goodness of Fit



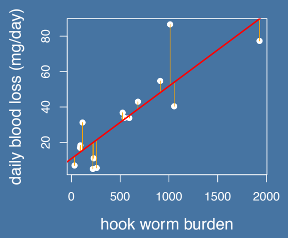
Chi Squared
Goodness of Fit Test

$$\chi^2 = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(Observed_i - Expected_i)^2}{\sigma^2}$$

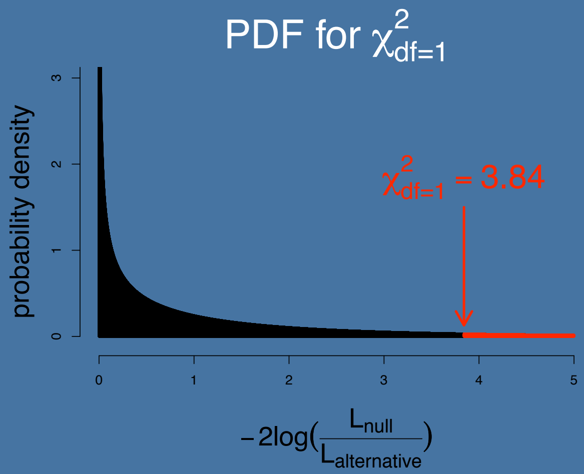


- Does the observed data differ significantly from our model?
- If not, then we cannot reject our model as a bad model.
- But we cannot accept our model (the null hypothesis) !

Is it a good model:
Goodness of Fit



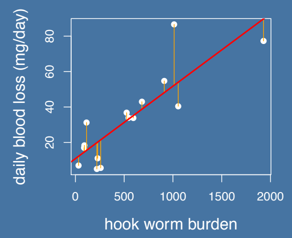
Likelihood Ratio Test (G test, Analysis of Deviance, ANOVA)



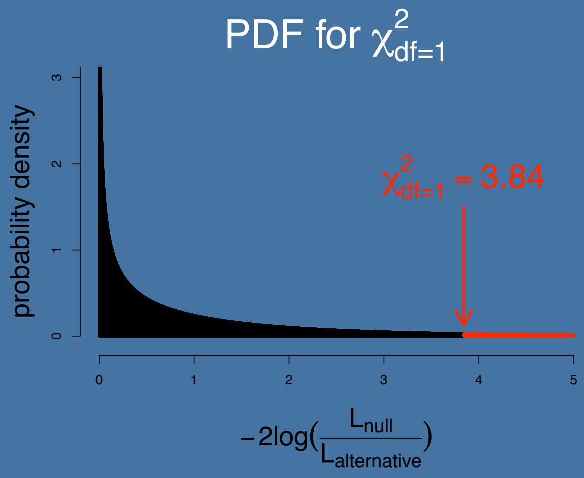
Under the null hypothesis:

$$2\log\frac{L_{MLE}}{L_{Null}} \sim \chi^2_{\text{df} = \text{difference in \# of parameters}}$$

Is it a good model:
Model Selection



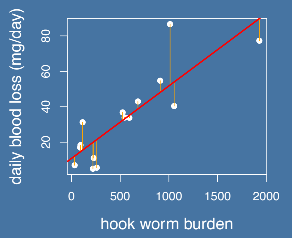
Likelihood Ratio Test (G test, Analysis of Deviance, ANOVA)



Under the null hypothesis:

$$2\log\frac{L_{\text{more parameters}}}{L_{\text{less parameters}}} \sim \chi^2_{\text{df = difference in \# of parameters}}$$

Is it a good model:
Model Selection



Akaike's Information Criterion (AIC)

$$AIC = -2\log(L) + 2(\# \text{ of parameters})$$

penalty for adding parameters

Rank proposed models by AIC: lowest is best.

All models within 2 of lowest should be considered.

Overfitting

- You can always fit N data points with N parameters.
- How many is too many?
- Bias/Variance Tradeoff
- AIC, Cross-validation

Collinearity

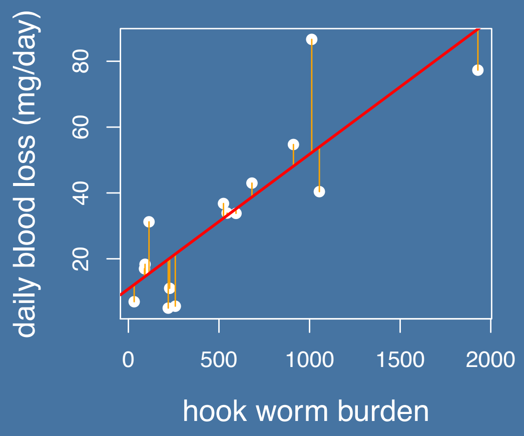
- Independent variables that vary with each other

Non-Identifiability

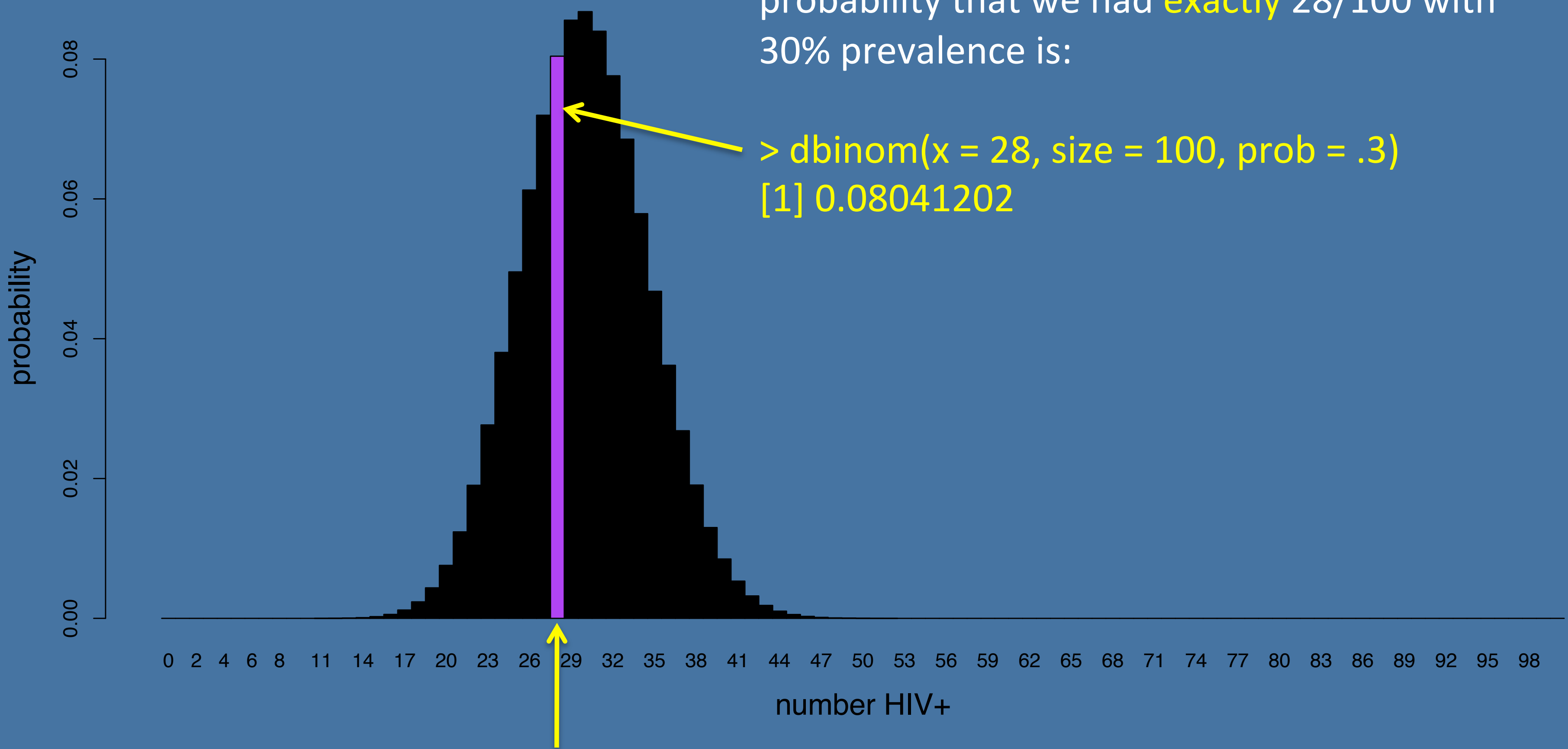
- Multiple parameter sets fit about equally well

What did we just do?

- Asked a question about a relationship
- Made some observations (data)
- Formulated the relationship into a model
- Fitted the model to data
- Assessed model fit/quality (model selection)
- Inference/parameter estimation
- Improved our understanding of the world



Introduction to Likelihood



We don't know the true prevalence, but the probability that we had exactly 28/100 with 30% prevalence is:

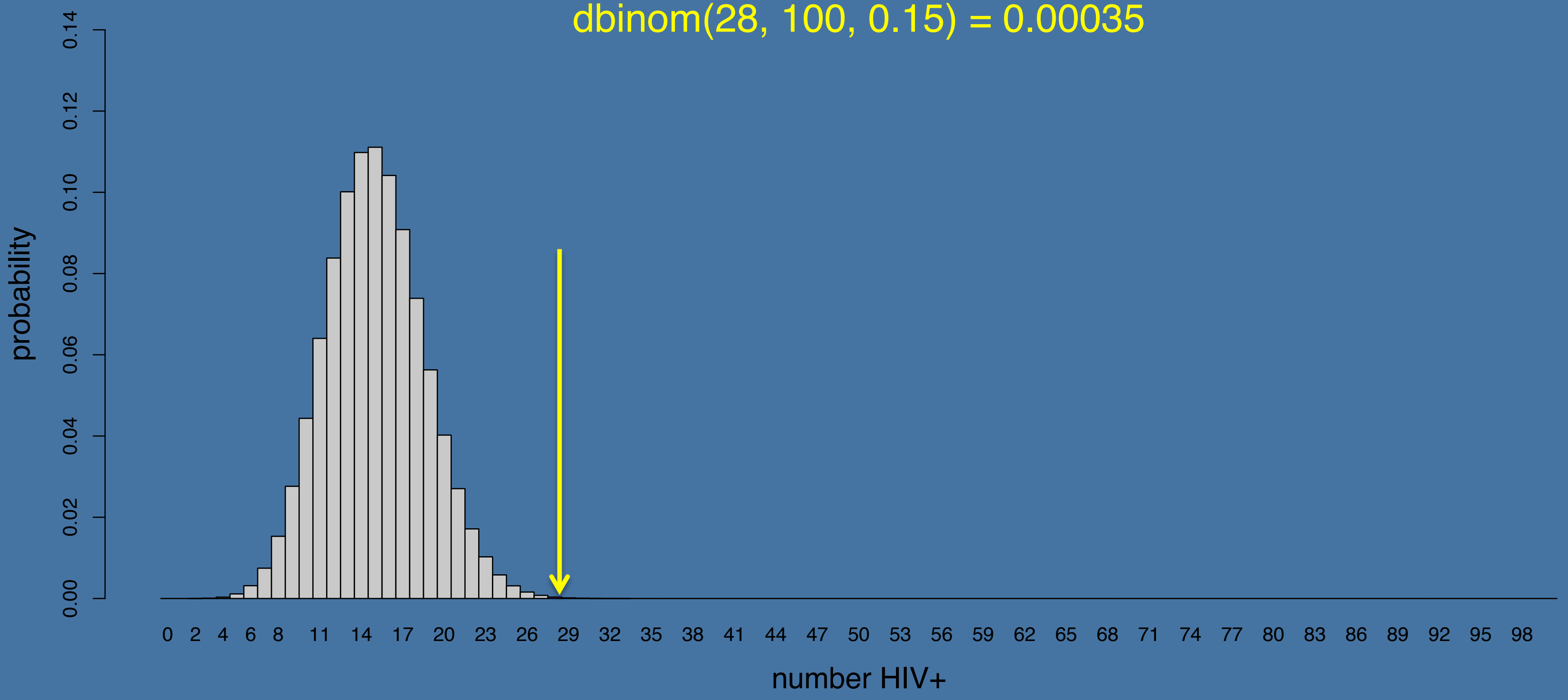
```
> dbinom(x = 28, size = 100, prob = .3)
[1] 0.08041202
```

We sample 100 people once and 28 are positive:

```
> rbinom(n = 1, size = 100, prob = .3)
[1] 28
```

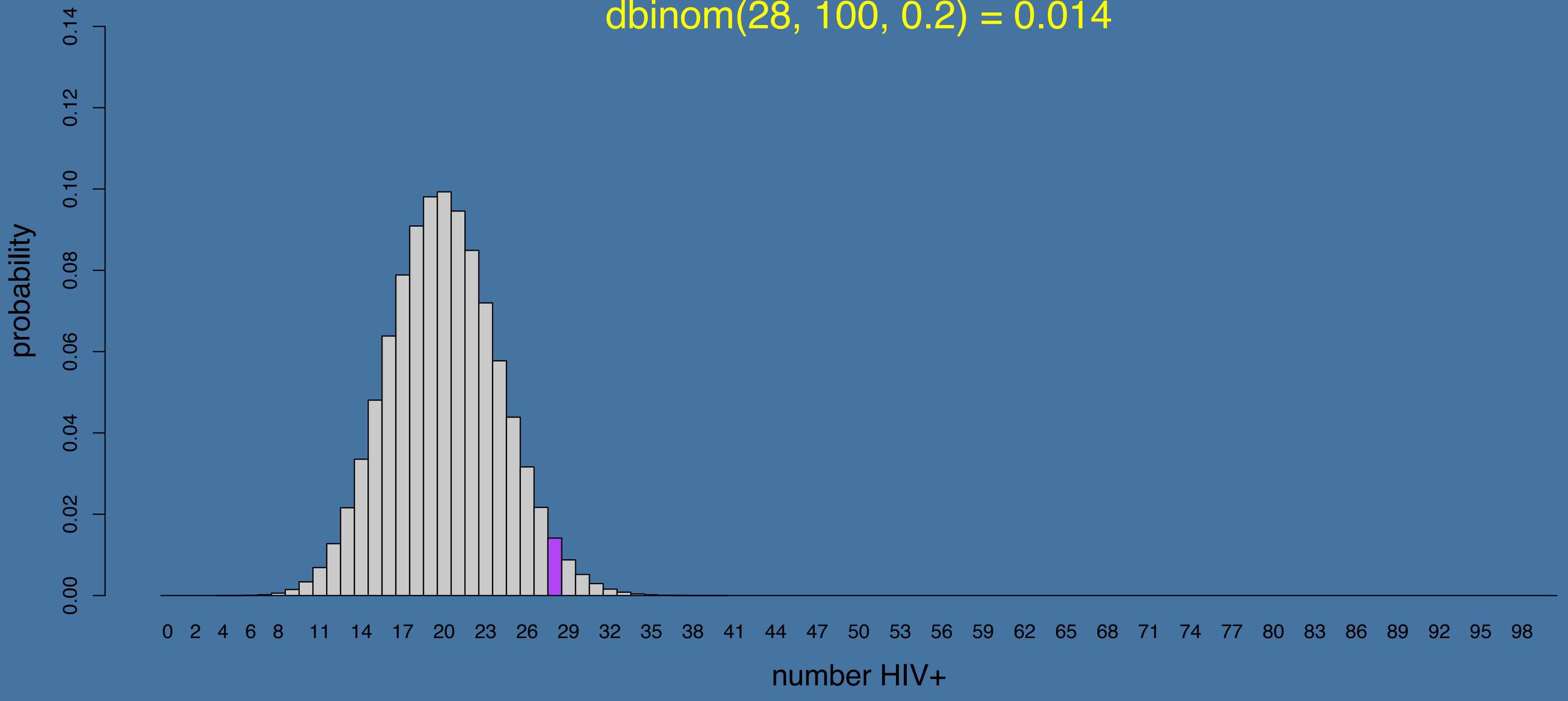
hypothetical prevalence: 15 %

dbinom(28, 100, 0.15) = 0.00035



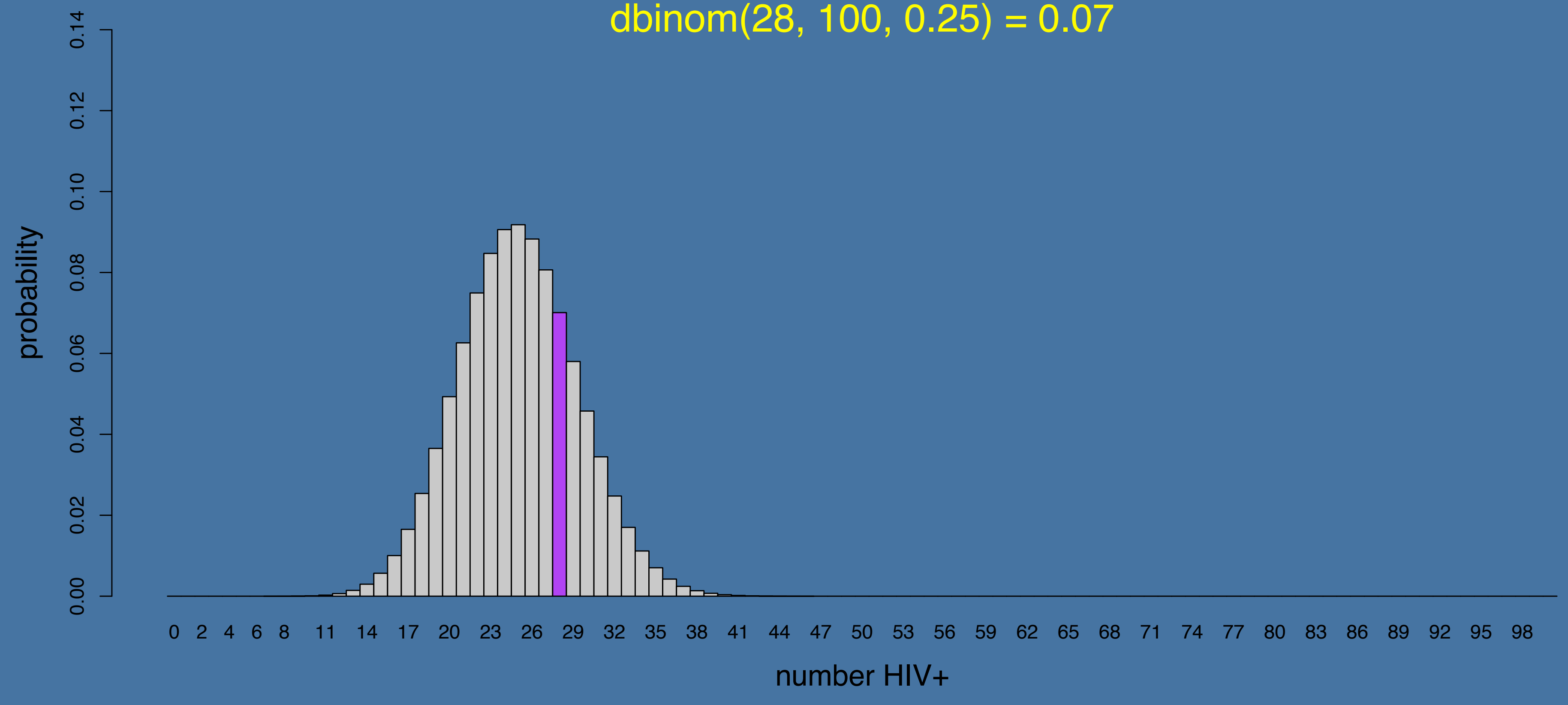
hypothetical prevalence: 20 %

dbinom(28, 100, 0.2) = 0.014



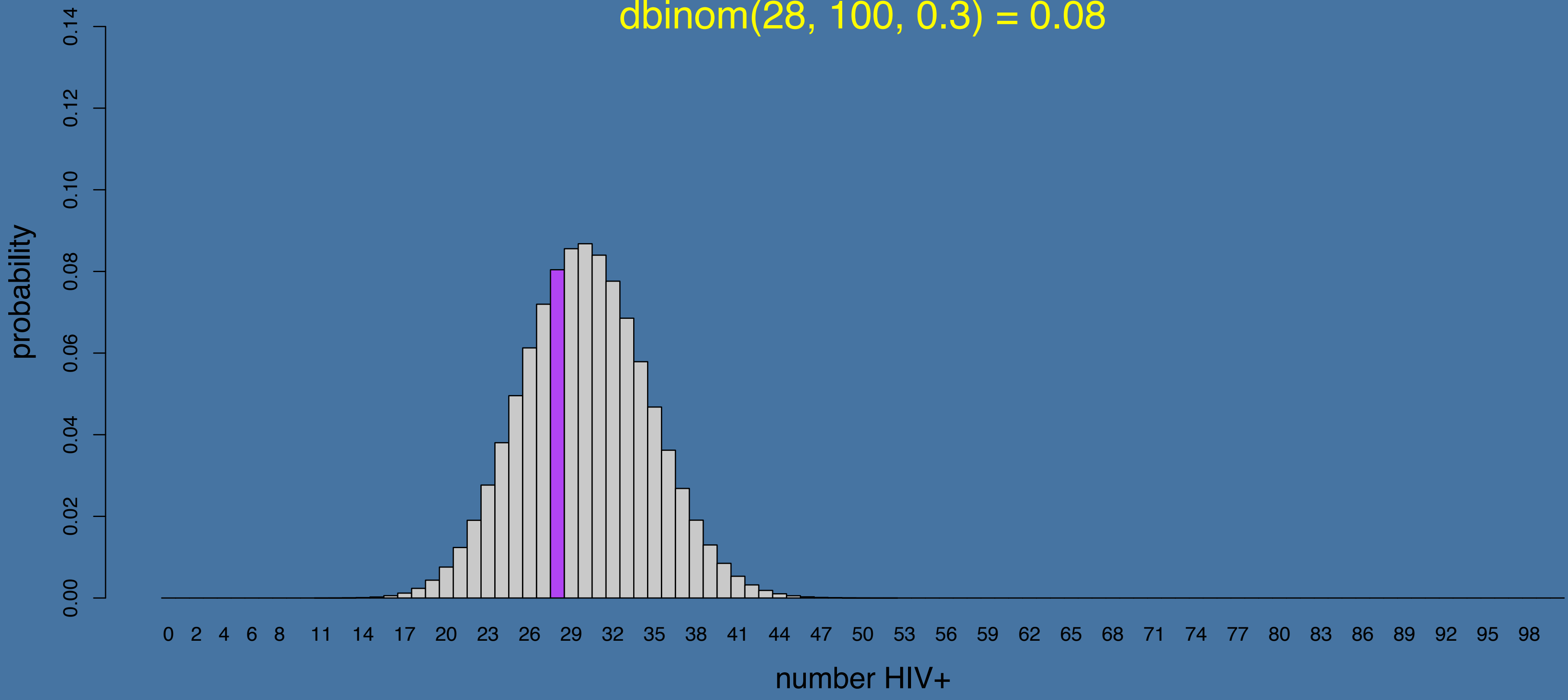
hypothetical prevalence: 25 %

dbinom(28, 100, 0.25) = 0.07



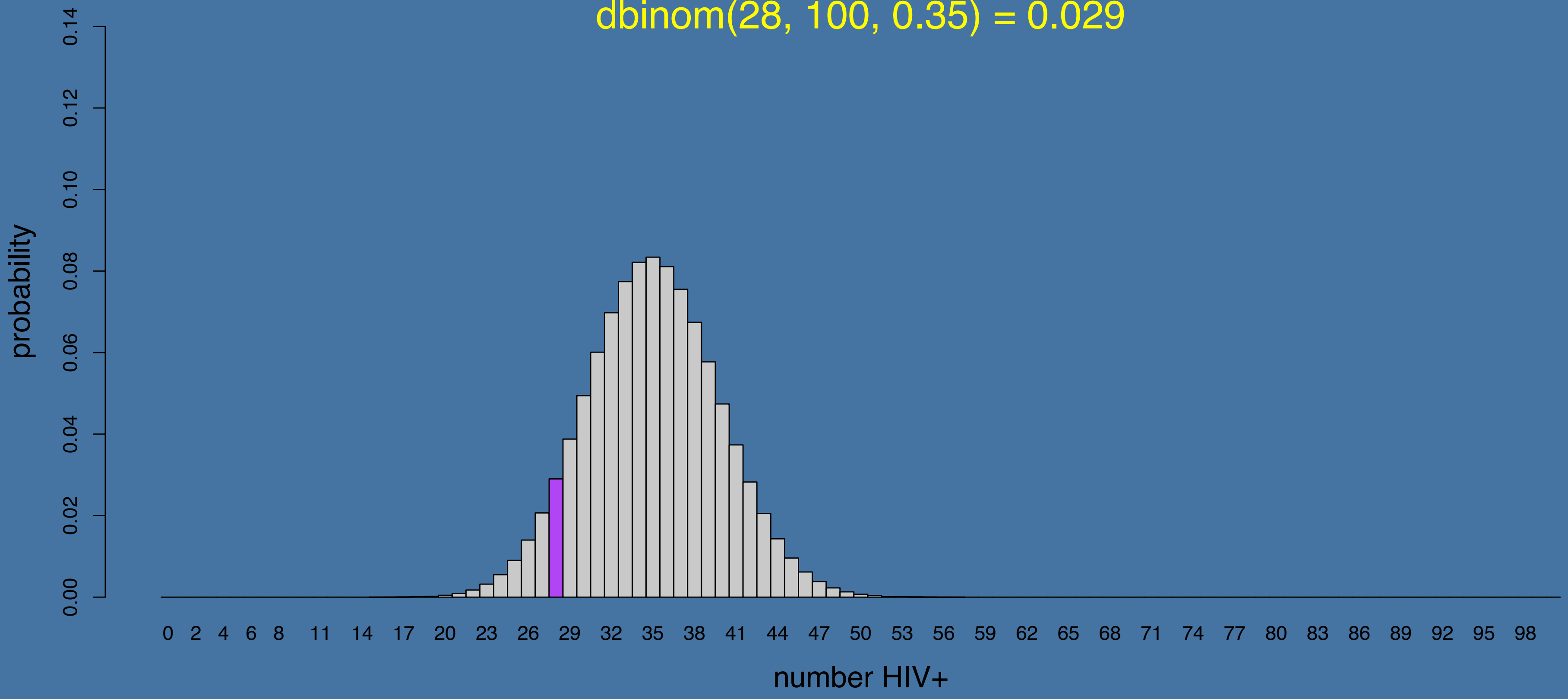
hypothetical prevalence: 30 %

dbinom(28, 100, 0.3) = 0.08



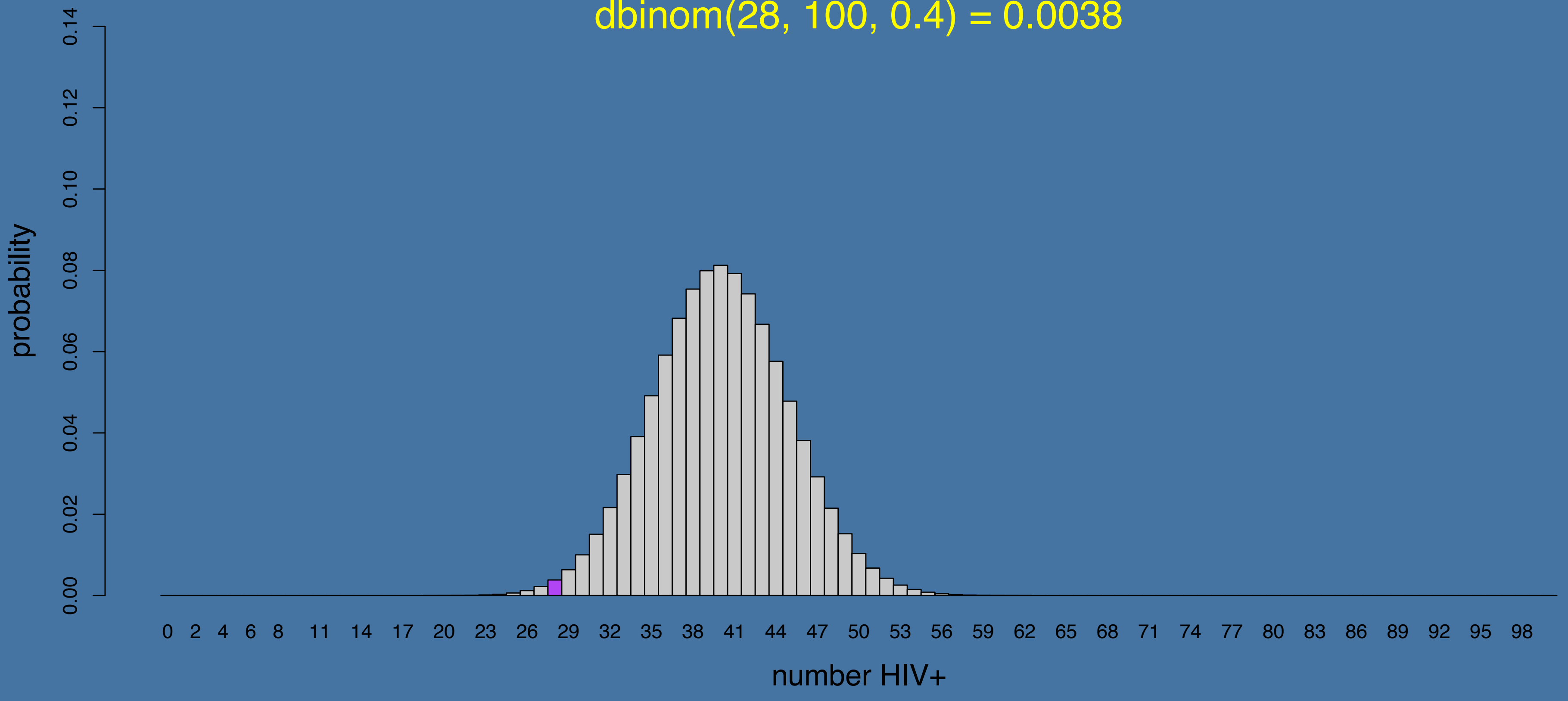
hypothetical prevalence: 35 %

dbinom(28, 100, 0.35) = 0.029

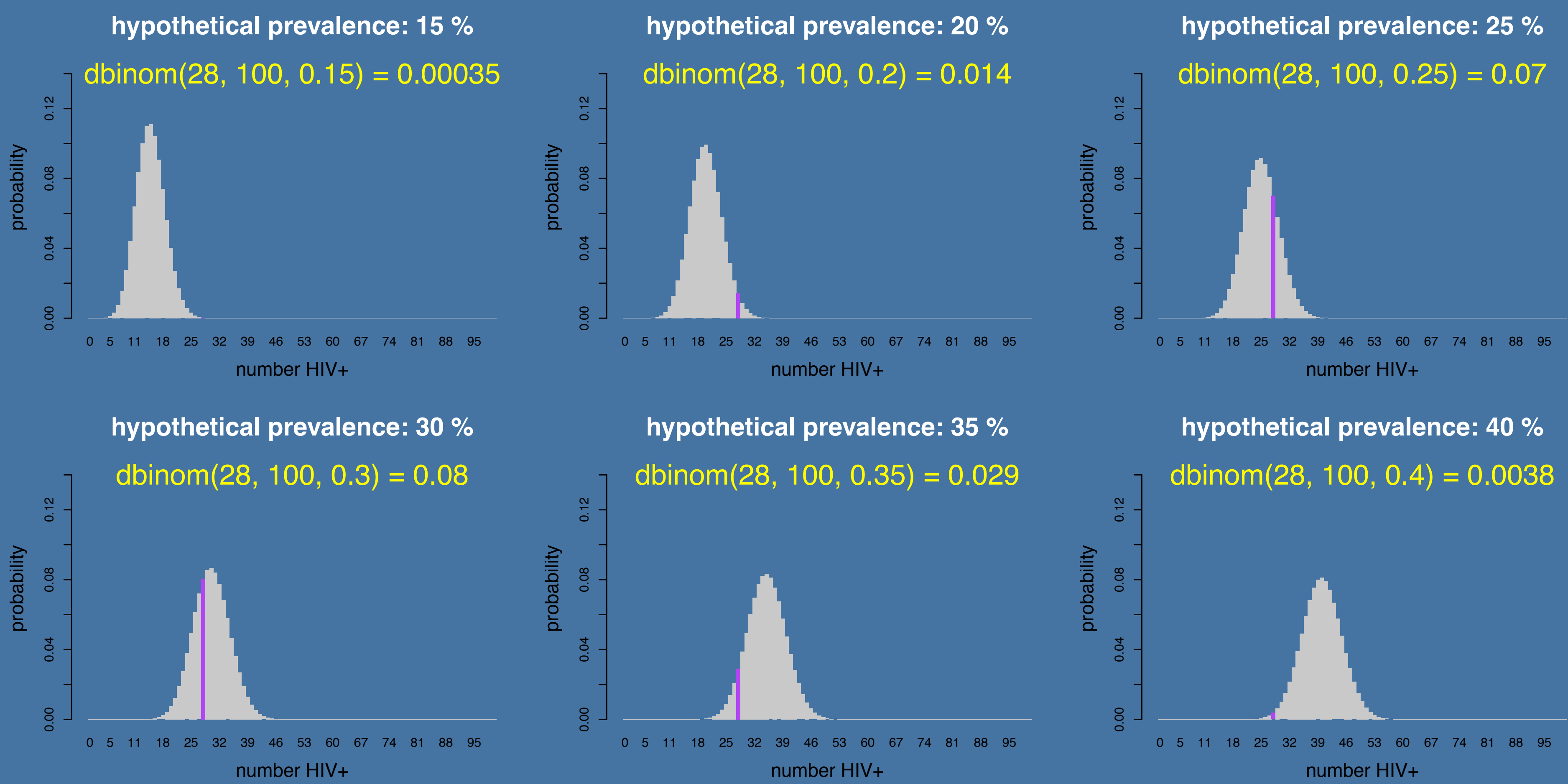


hypothetical prevalence: 40 %

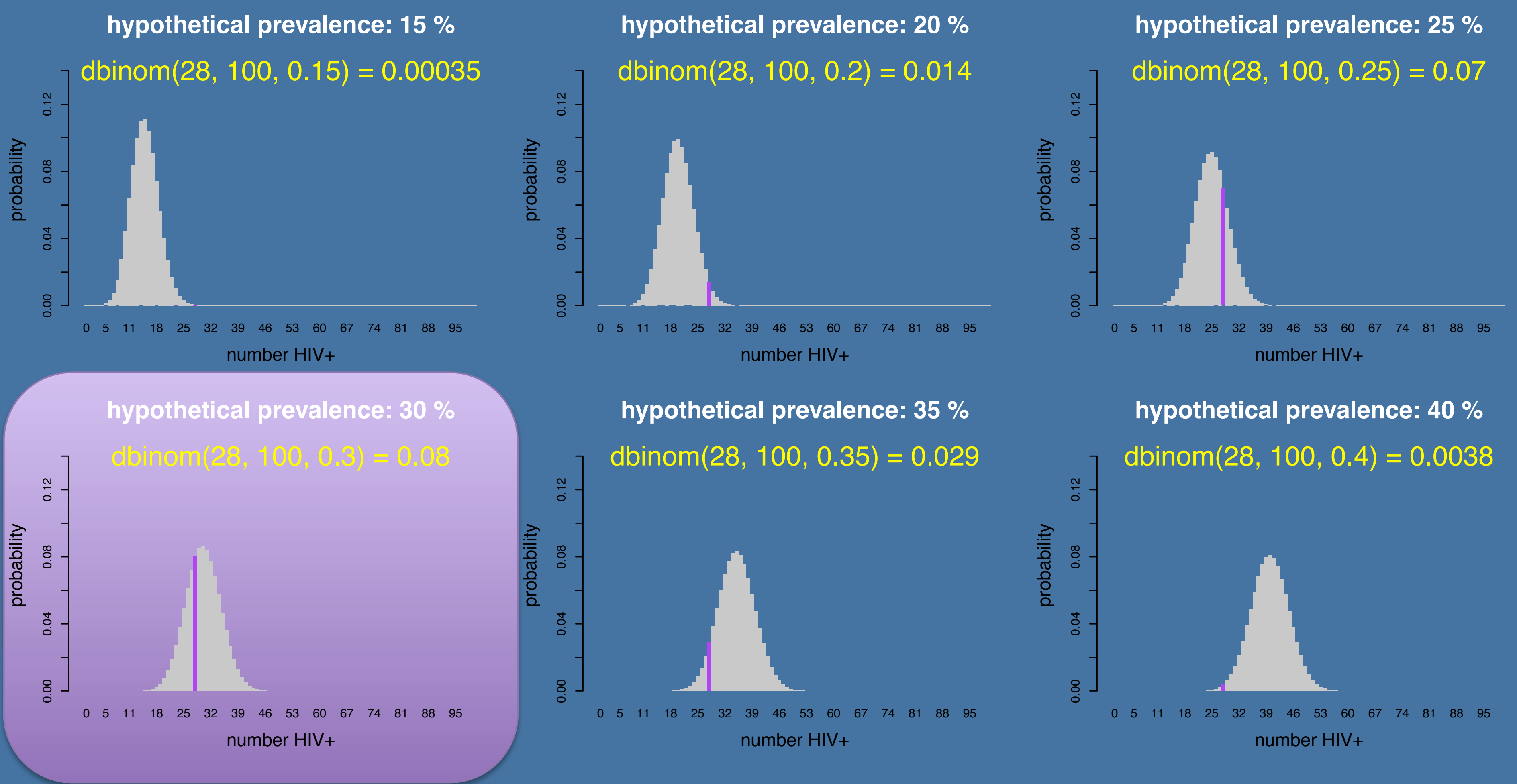
dbinom(28, 100, 0.4) = 0.0038

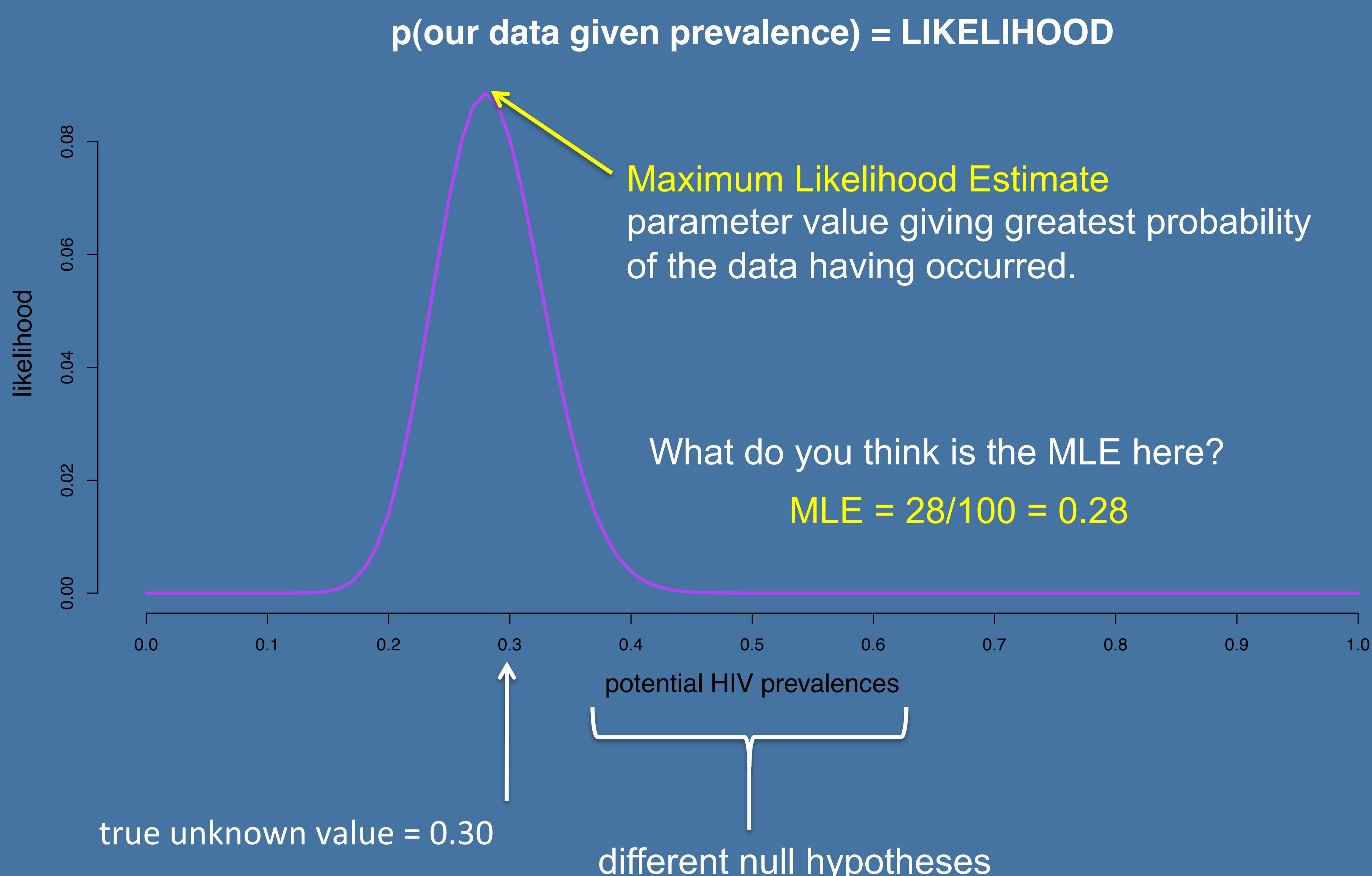


Which prevalence gives the greatest probability of observing exactly 28/100?



Which of these prevalence values is most likely given our data?




p(our data given prevalence) = LIKELIHOOD
Maximum Likelihood Estimate
parameter value giving greatest probability
of the data having occurred.
What do you think is the MLE here?
MLE = 28/100 = 0.28
likelihood
0.00
0.02
0.04
0.06
0.08
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
potential HIV prevalences
true unknown value = 0.30
different null hypotheses

Defining Likelihood

- L(parameter | data) = p(data | parameter)

- Not a probability distribution.

- Probabilities taken from many different distributions.

function of x

PDF: $$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

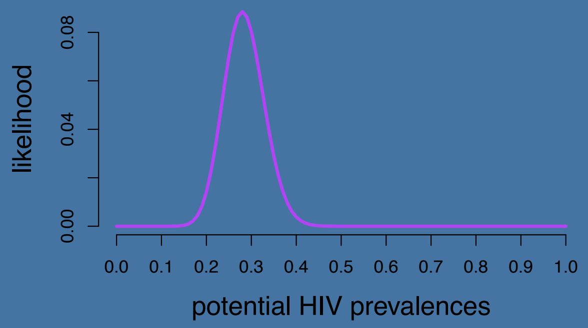
LIKELIHOOD: $$L(p \mid x) = \binom{n}{x} p^x (1-p)^{n-x}$$

function of p

Deriving the Maximum Likelihood Estimate

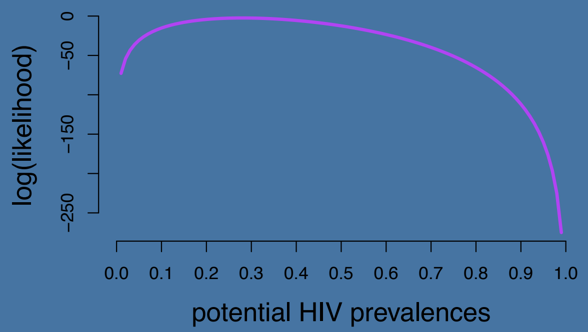
maximize

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$



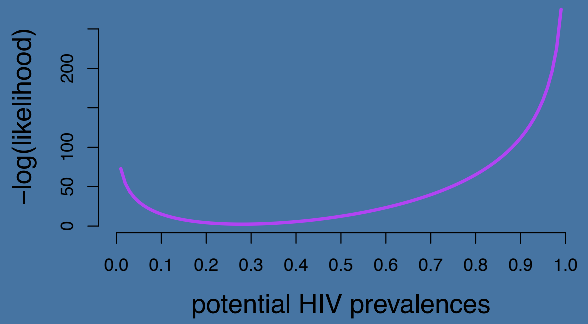
maximize

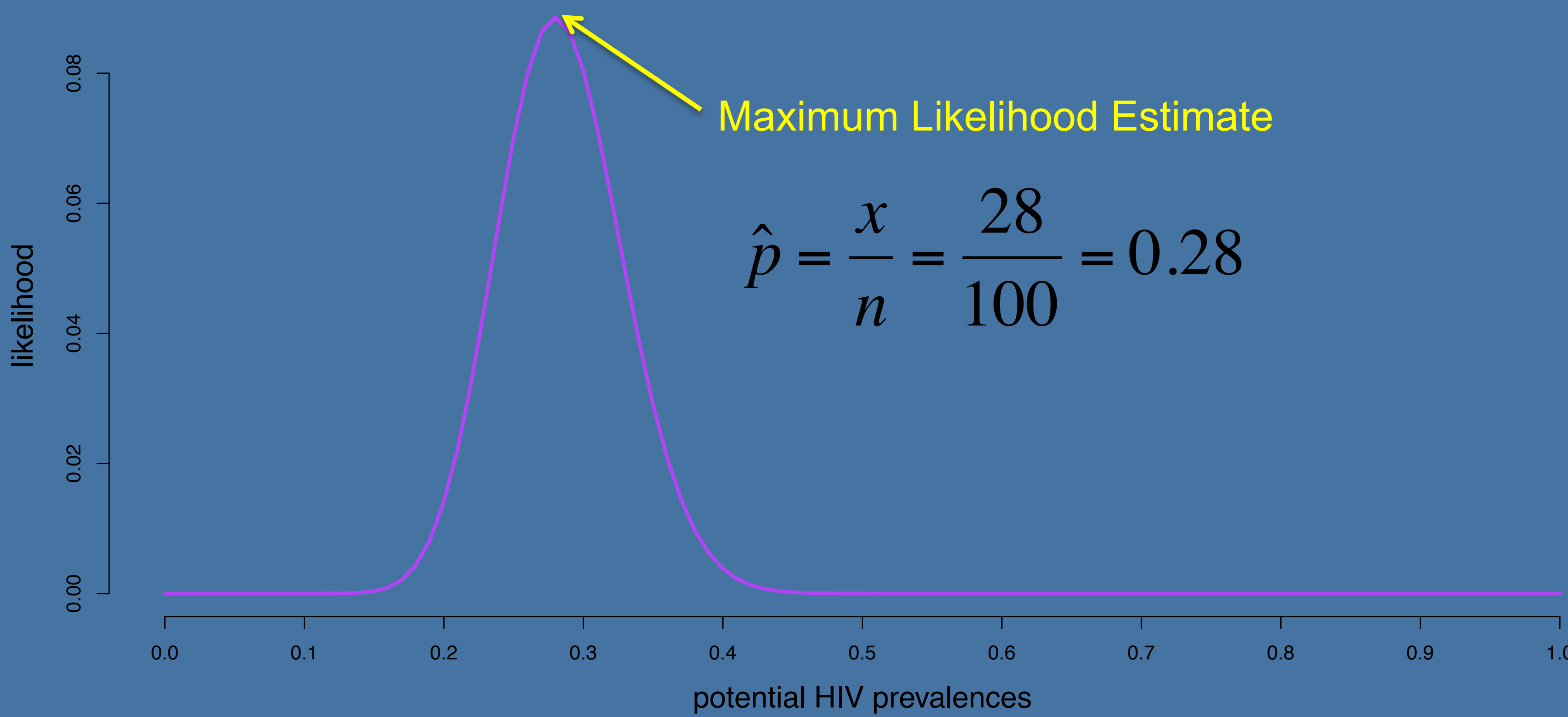
$$\log(L(p)) = \log\left[\binom{n}{x} p^x (1-p)^{n-x}\right]$$



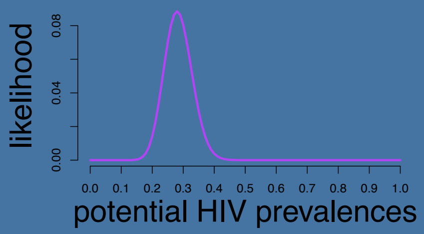
minimize

$$l(p) = -\log\left[\binom{n}{x} p^x (1-p)^{n-x}\right]$$

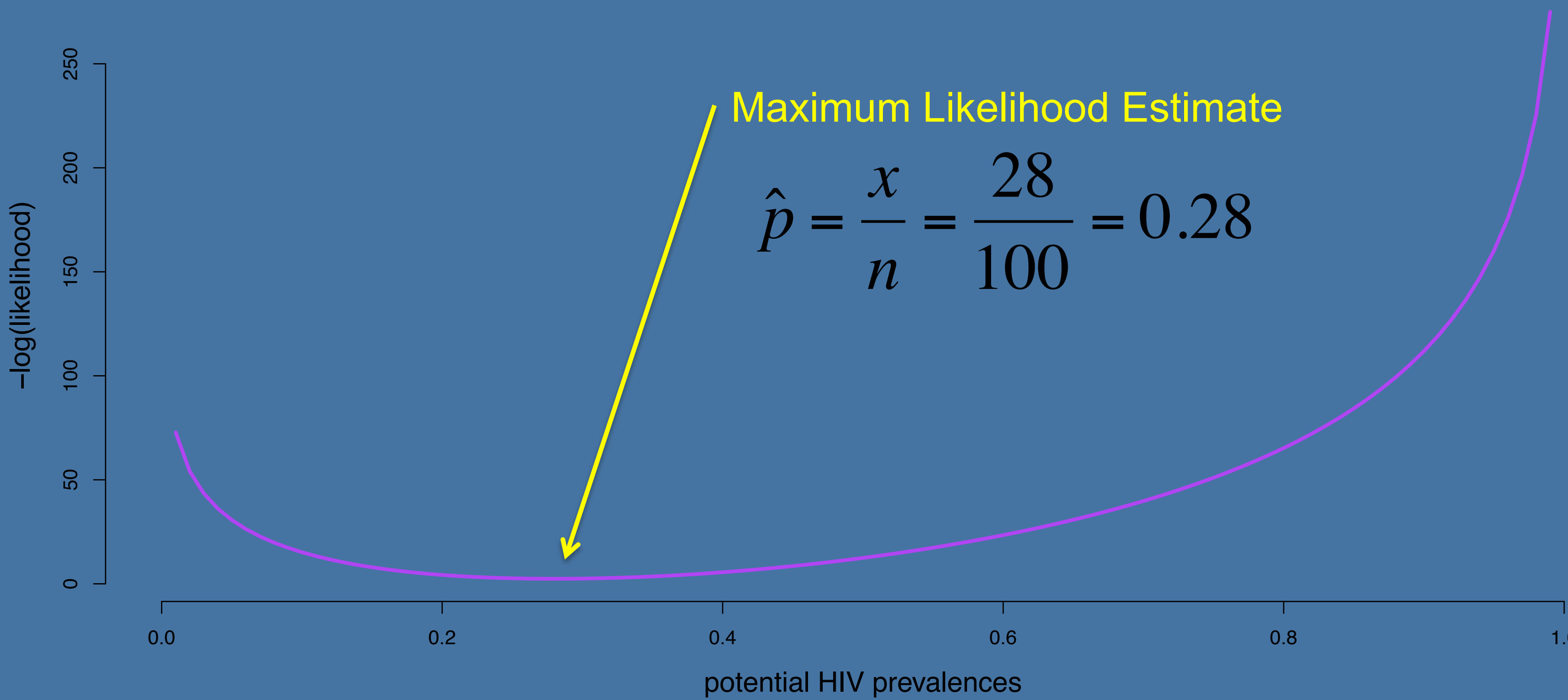




$$\hat{p} = \frac{x}{n} = \frac{28}{100} = 0.28$$



we usually minimize the –log(likelihood)



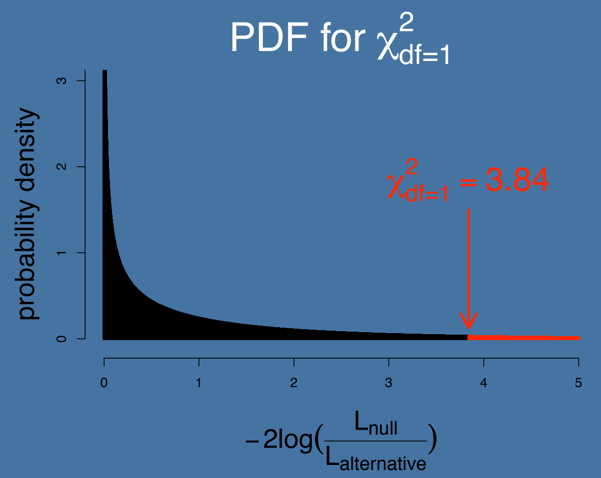
Building Confidence Intervals

Likelihood Ratio Test

If the null hypothesis were true then

$$-2\log\left(\frac{L(\text{null hypothesis})}{L(\text{alternative hypothesis})}\right) \sim \chi^2_{df=1}$$

$$2l_{alternative} - 2l_{null} \sim \chi^2_{df=1}$$



So if our α = .05, then we reject any null hypothesis for which

$$2l_{MLE} - 2l_{null} > \chi^2_{df=1,\ \alpha=0.05} = 3.84$$

```
> qchisq(p = .95, df = 1)
[1] 3.841459
```

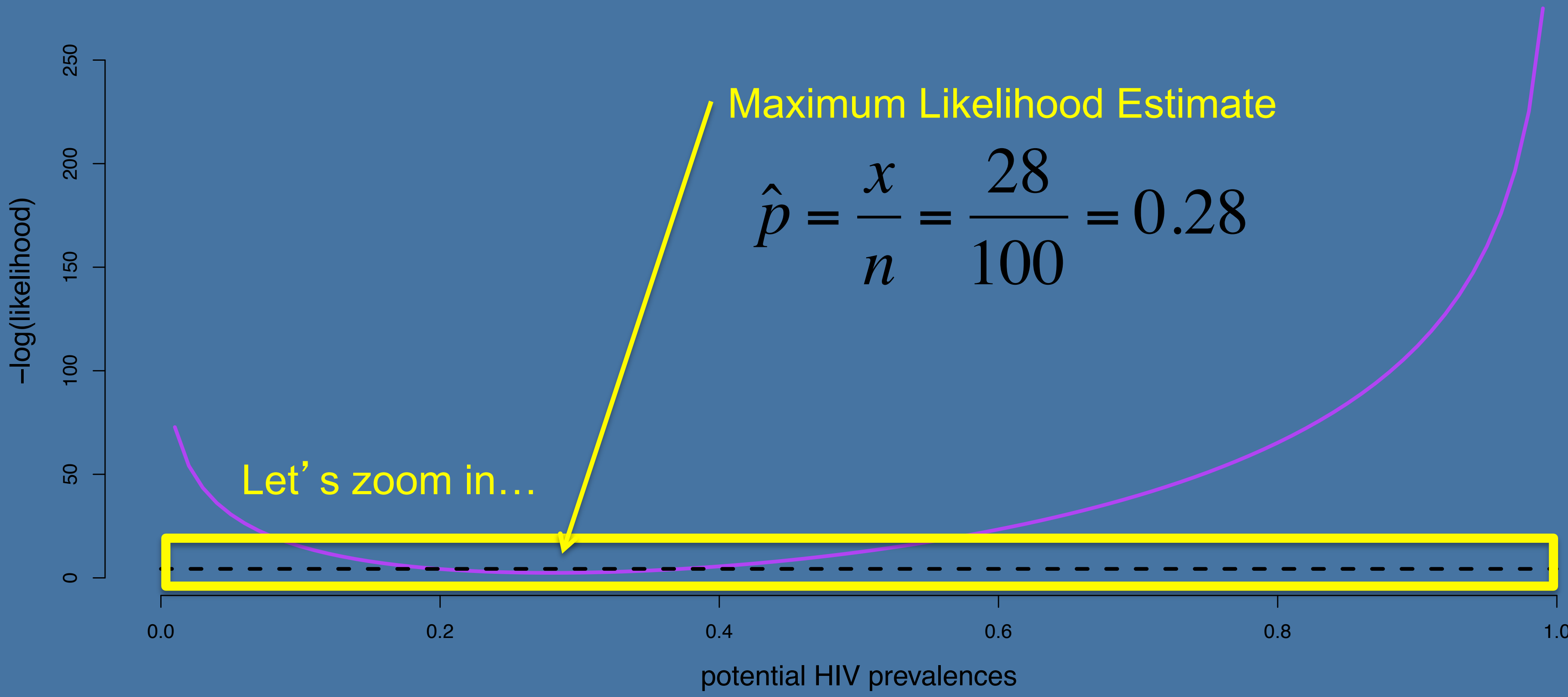
$$2l_{MLE} - 2l_{null} > 3.84$$

$$l_{MLE} - l_{null} > 1.92$$

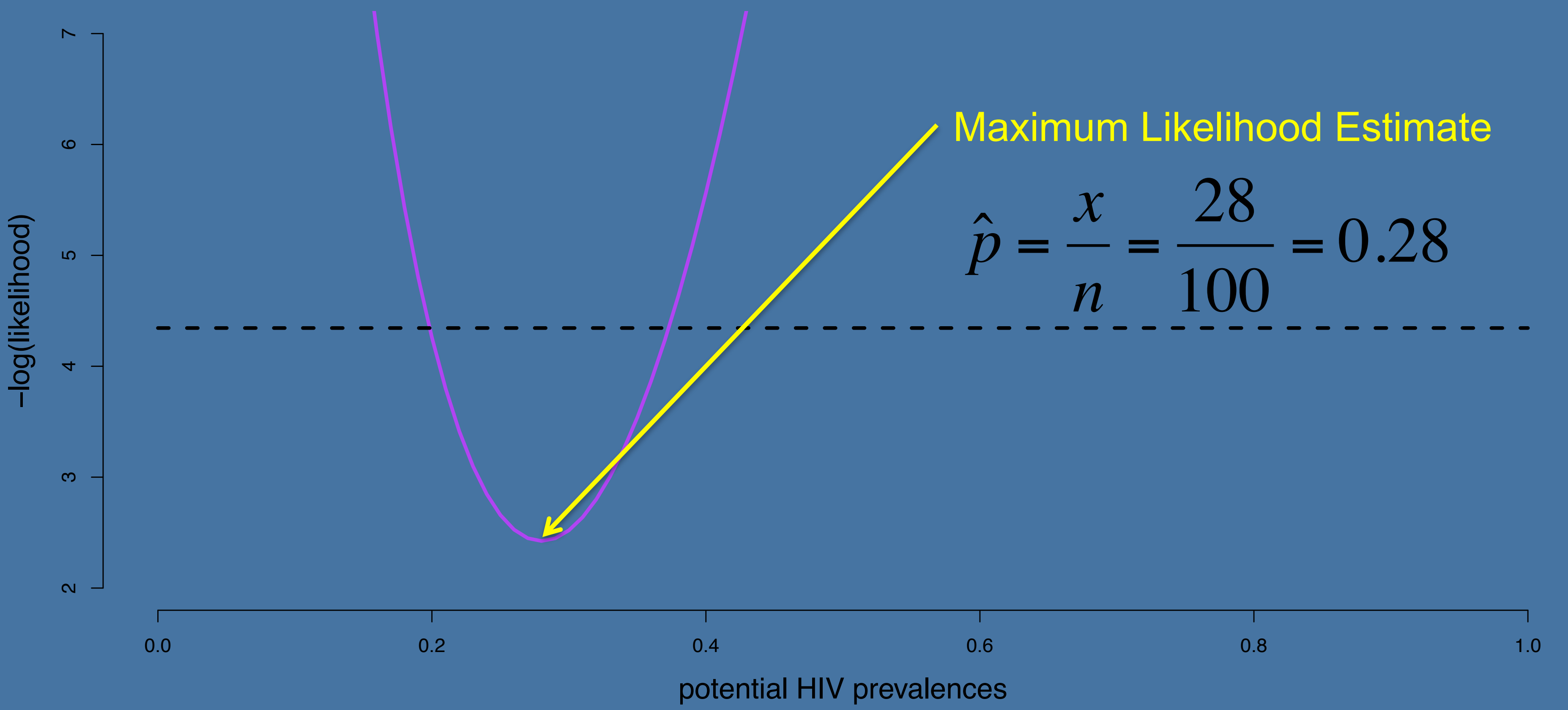
When l_{MLE} - l_{null} > 1.92,
we reject that null hypothesis.

Building Confidence Intervals

Likelihood Ratio Test

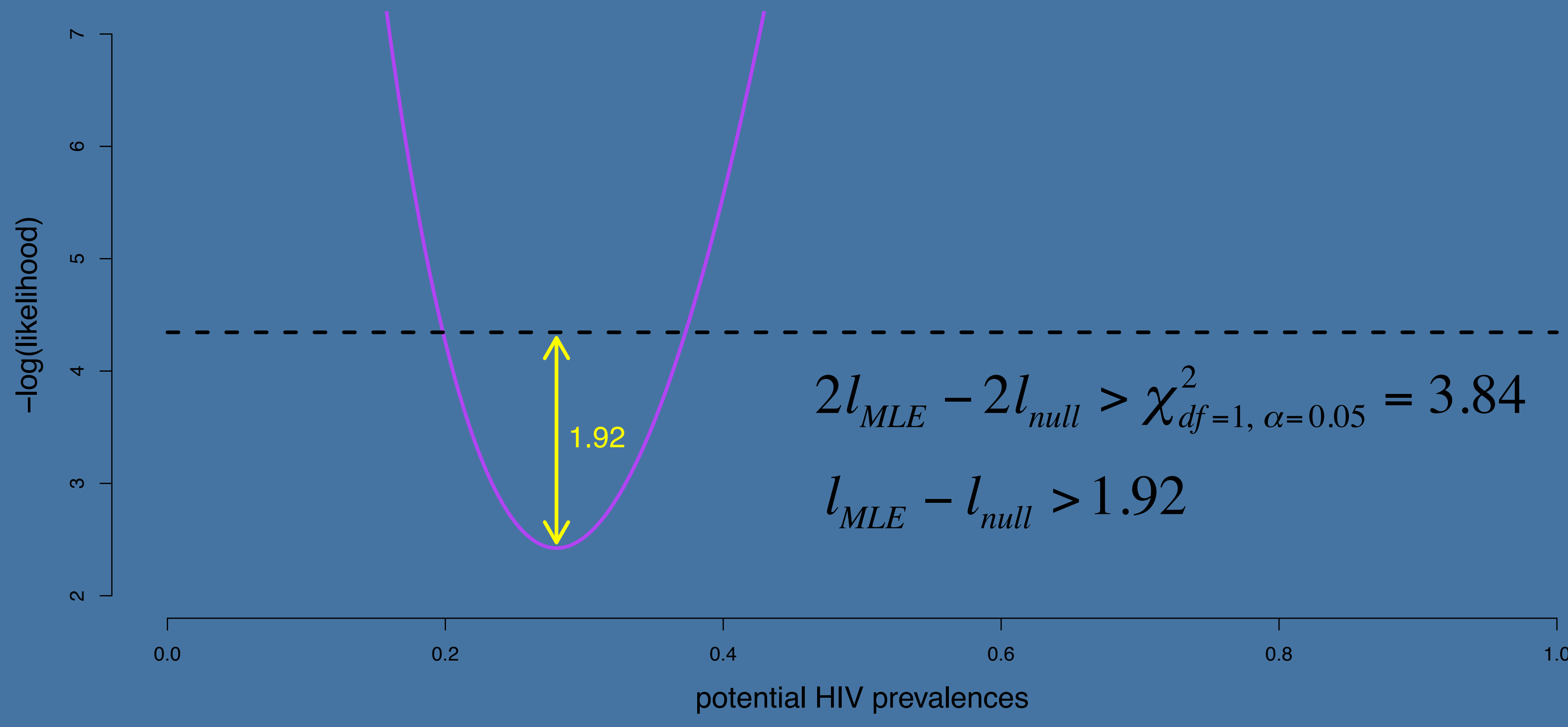


Building Confidence Intervals
Likelihood Ratio Test



Building Confidence Intervals

Likelihood Ratio Test

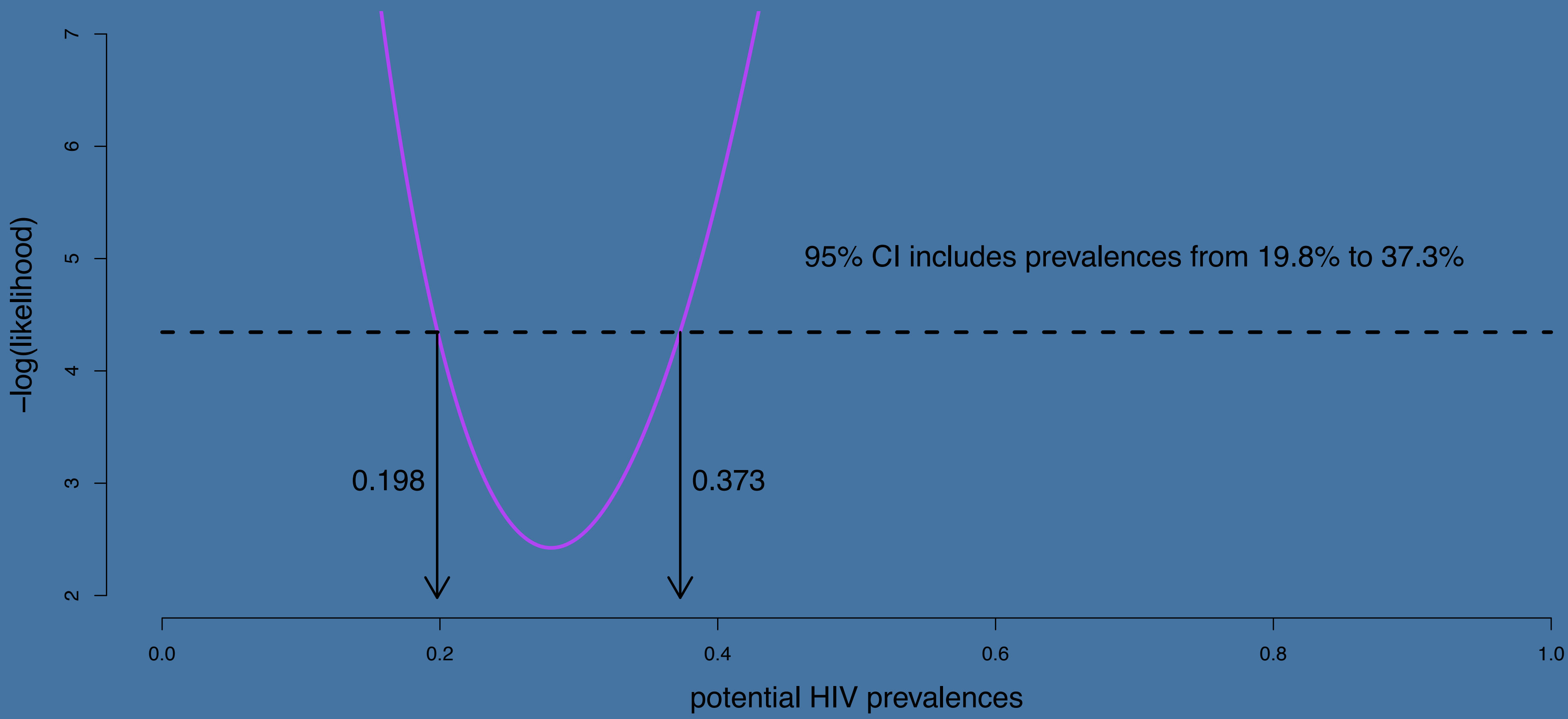


$$2l_{MLE} - 2l_{null} > \chi^2_{df=1,\ \alpha=0.05} = 3.84$$

$$l_{MLE} - l_{null} > 1.92$$

Building Confidence Intervals

Likelihood Ratio Test



Statistical Models & Dynamic Models

Statistical Models

- Account for bias and random error to find correlations that may imply causality.
- Often the first step to assessing relationships.
- Assume independence of individuals (at some scale).

Dynamic Models

- Systems Approach: Explicitly model multiple mechanisms to understand their interactions.
- Links observed relationships at different scales.
- Explicitly focuses on dependence of individuals

By developing dynamic models in a probabilistic framework we can account for dependence, random error, and bias while linking patterns at multiple scales.

Fitting Dynamic Models to Data

Adapt our dynamic models in a probabilistic framework so we can ask:

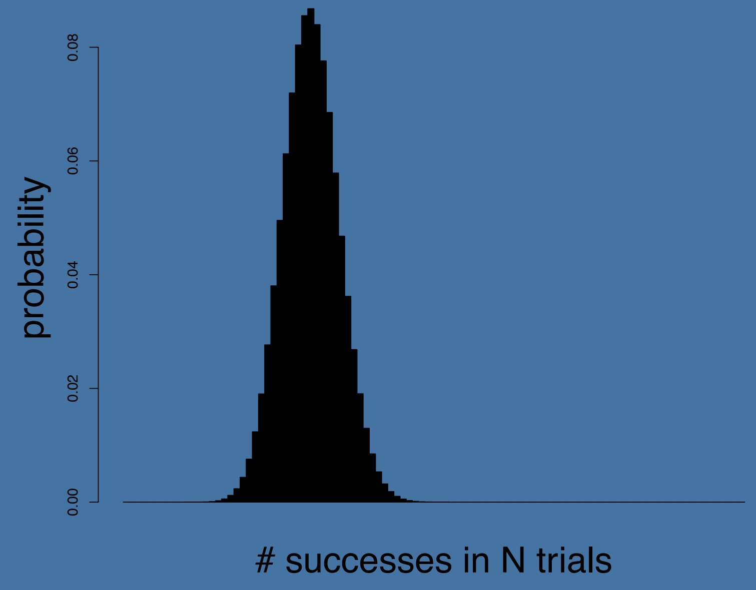
What is the probability that a model would have generated the observed data?

↓

What is the likelihood of a model given the data?

Likelihood of parameters
(given data)

Binomial Distribution

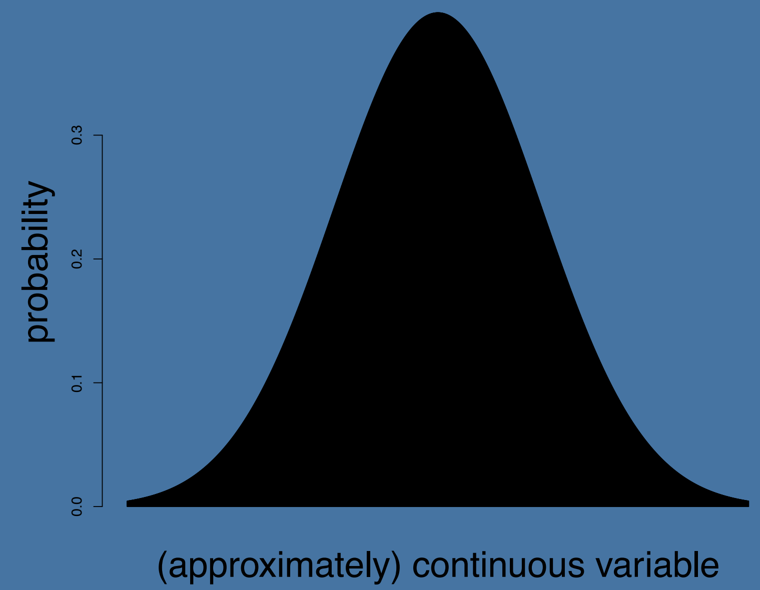


Distribution

↓

Likelihood of parameters
(given data)

Normal Distribution

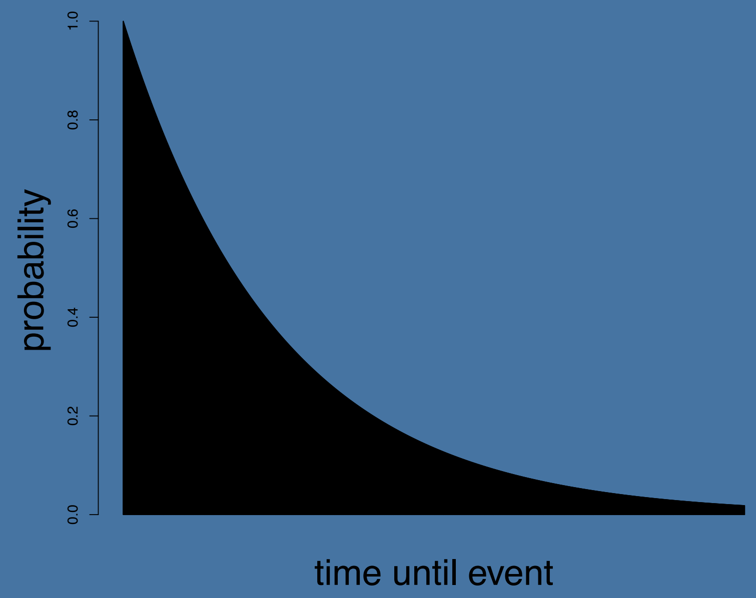


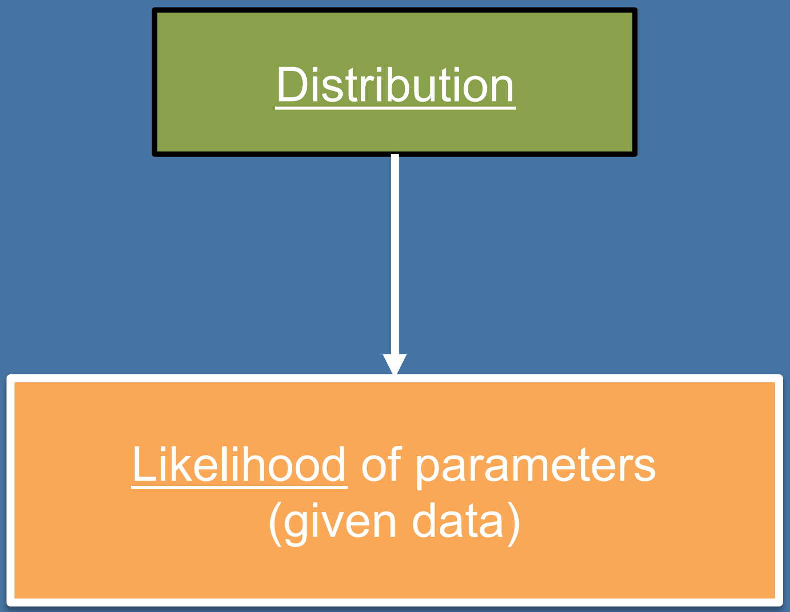
Distribution

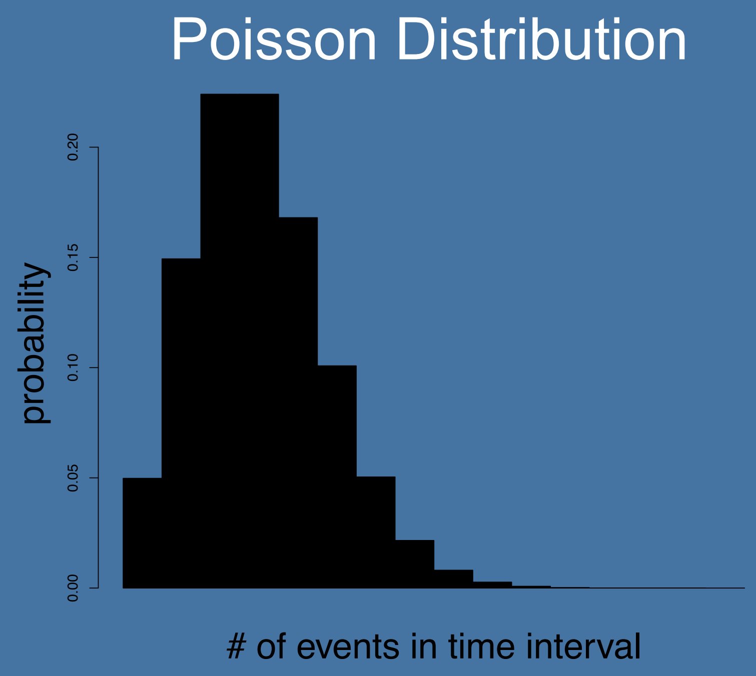
↓

Likelihood of parameters
(given data)

Exponential Distribution



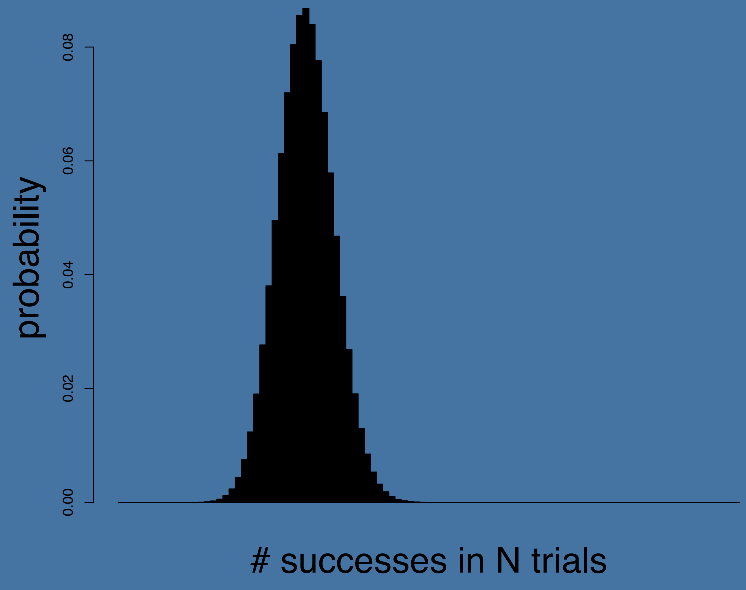




Distribution

Likelihood of parameters
(given data)

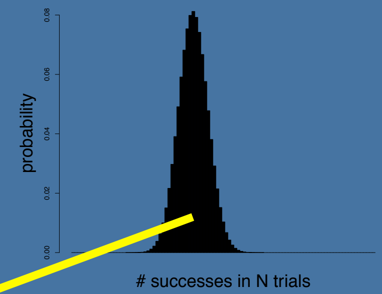
Binomial Distribution



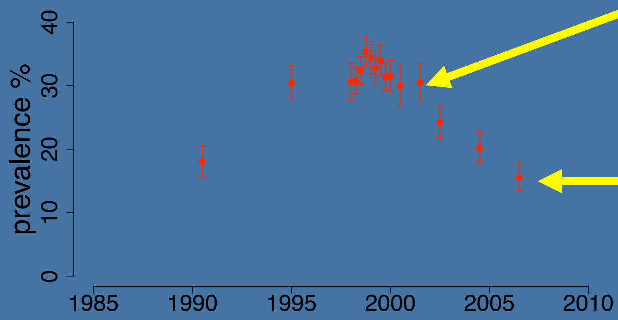
Stochastic Component of Model

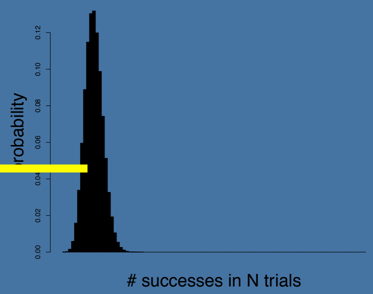
Distribution

Likelihood of parameters
(given data)



HIV in Harare

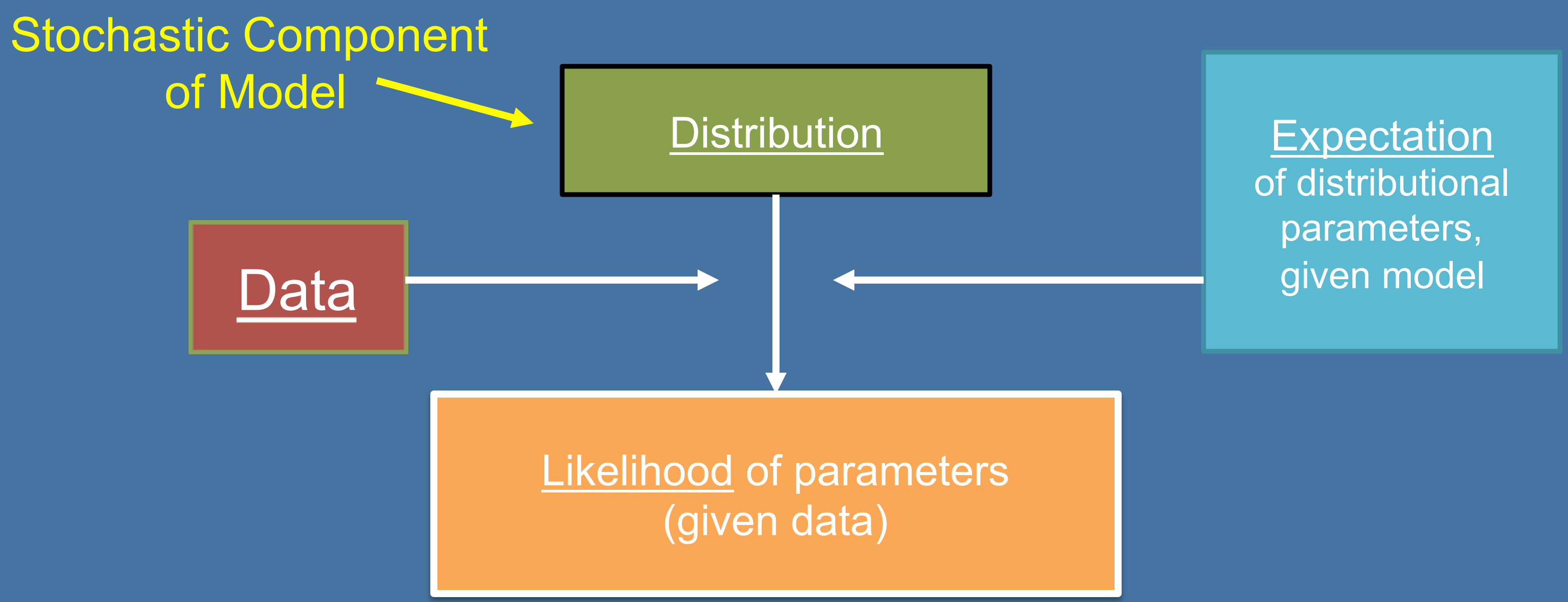


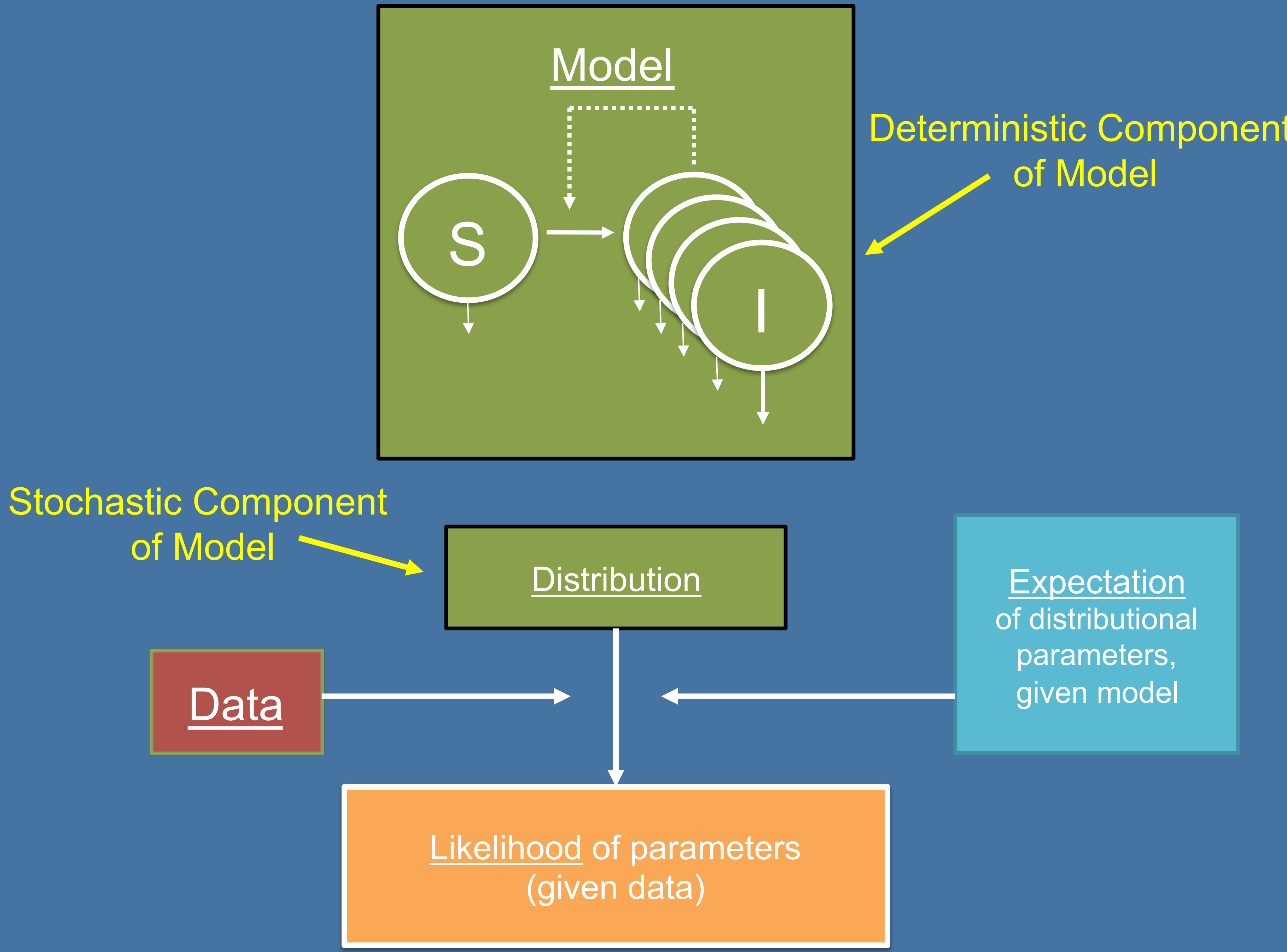


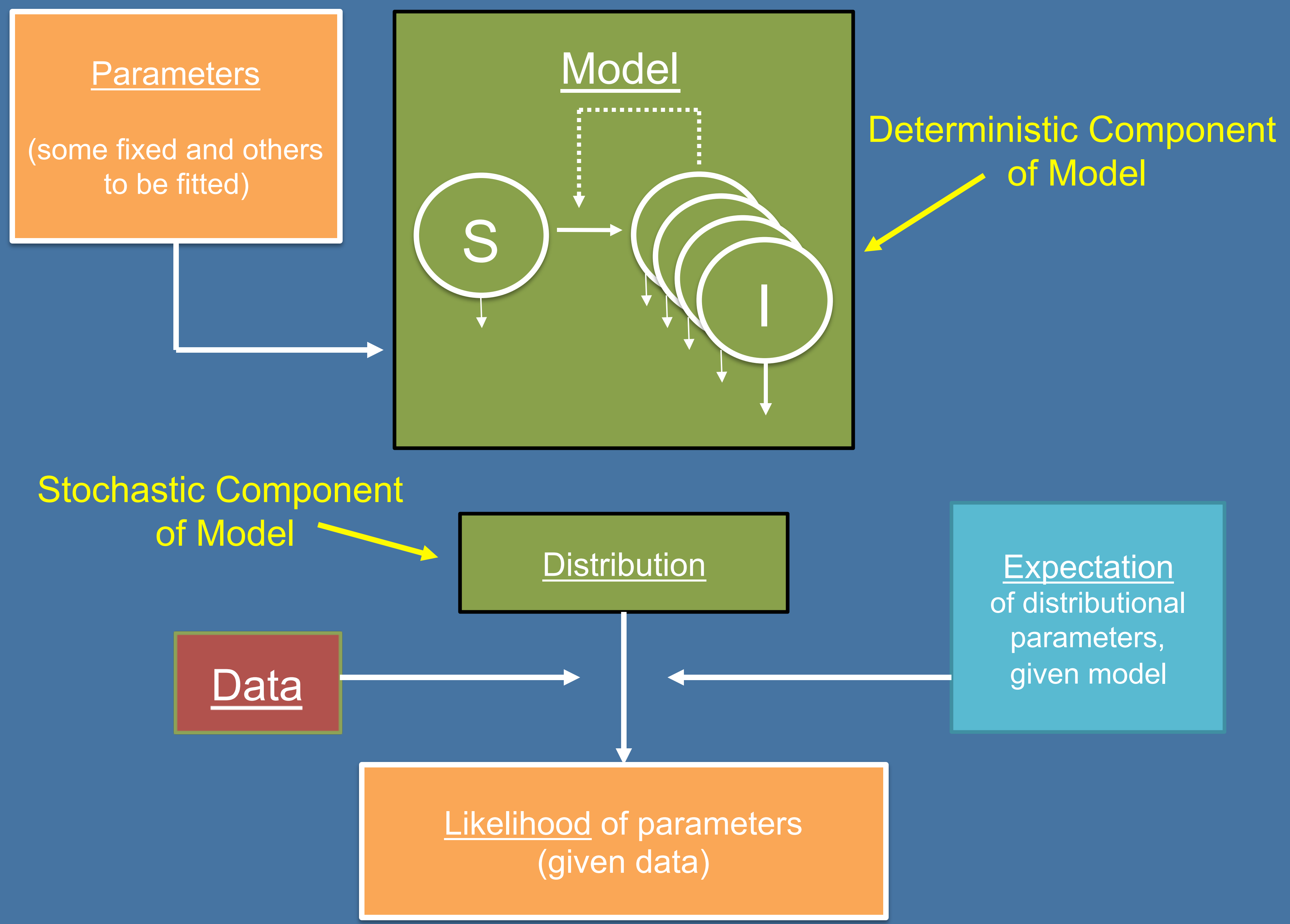
Distribution

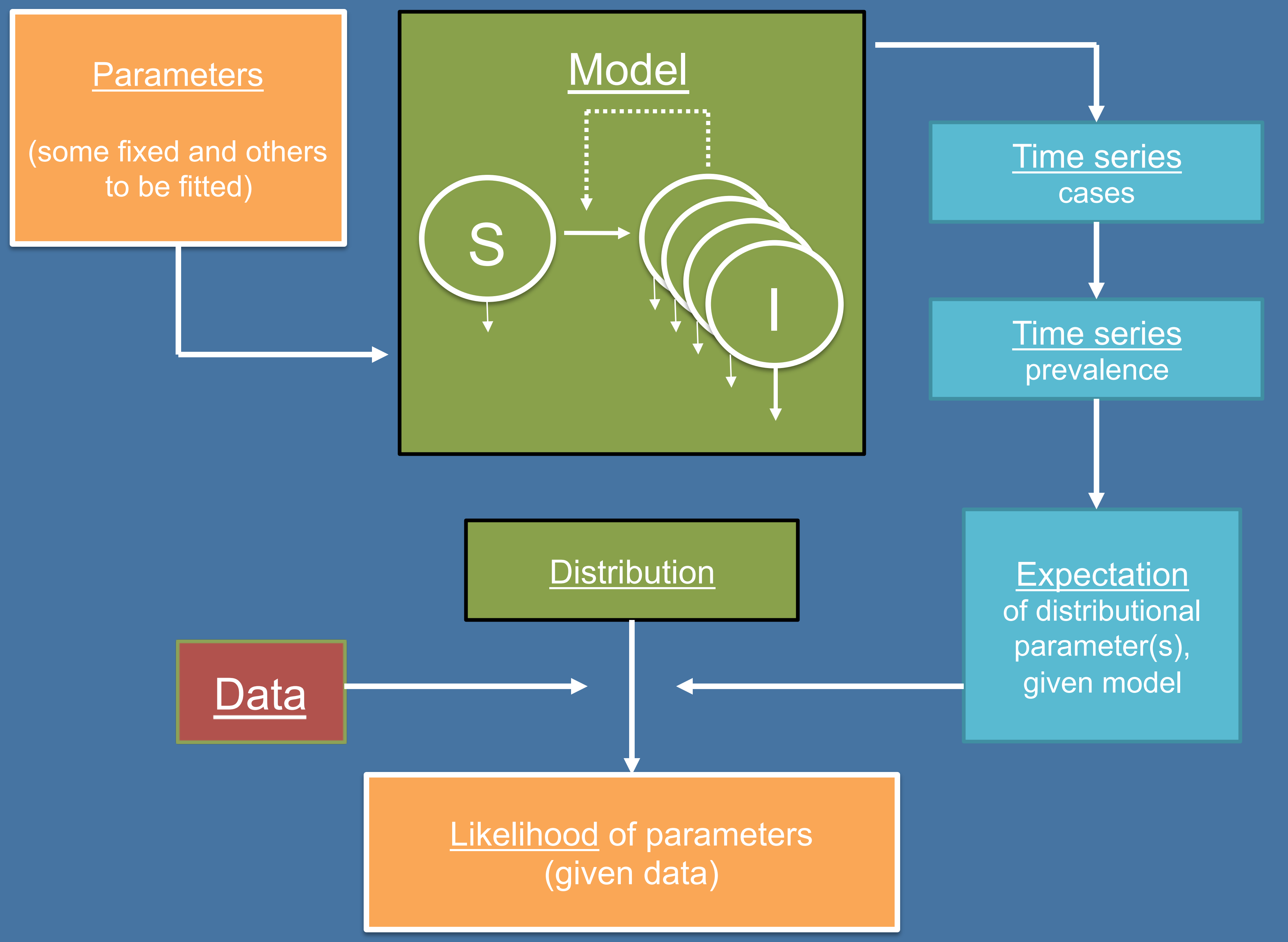
Data

Likelihood of parameters
(given data)


Stochastic Component of Model
Distribution
Data
Expectation of distributional parameters, given model
Likelihood of parameters (given data)


Model
S
I
Deterministic Component of Model
Stochastic Component of Model
Distribution
Data
Expectation
of distributional parameters, given model
Likelihood of parameters (given data)


Parameters
(some fixed and others to be fitted)
Model
S
I
Deterministic Component of Model
Stochastic Component of Model
Distribution
Data
Expectation of distributional parameters, given model
Likelihood of parameters (given data)



Parameters
(some fixed and others to be fitted)
Model
S
I
Time series
cases
Time series
prevalence
Expectation
of distributional parameter(s), given model
Distribution
Data
Likelihood of parameters (given data)

Collinearity

- Independent variables that vary with each other

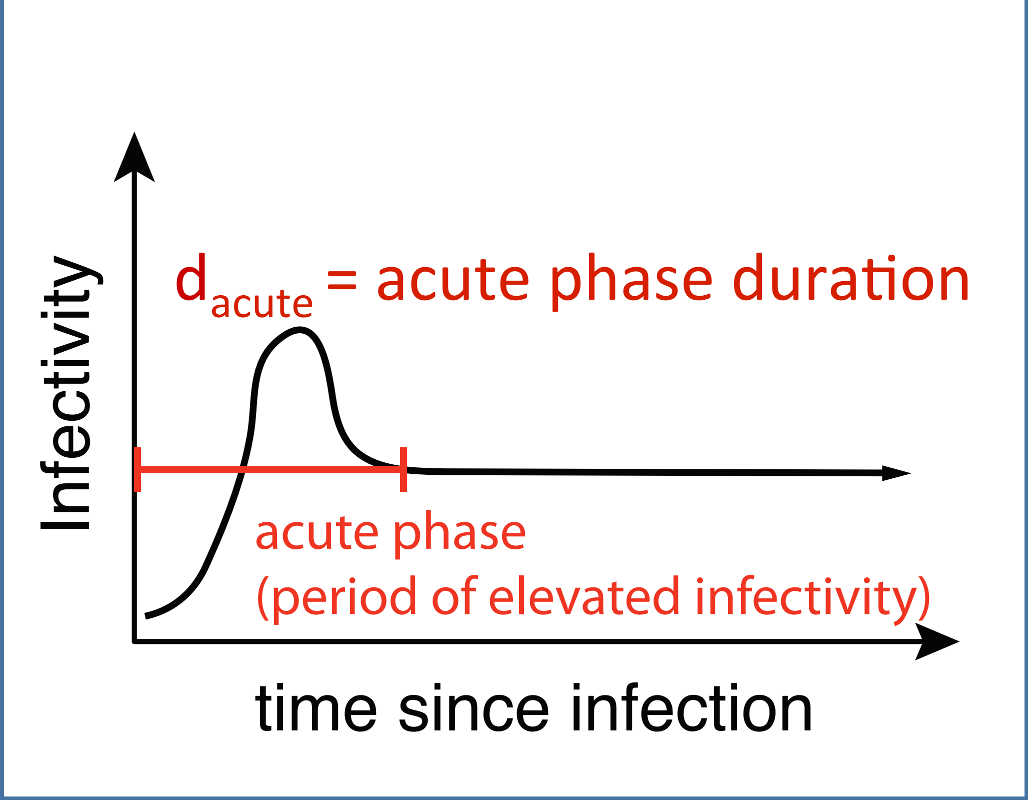
Non-Identifiability

- Multiple parameter sets fit about equally well

- Can be informative in dynamic models

Acute HIV Infection

- Thought to be extremely infectious
- Epidemiological evidence from a Ugandan couples cohort

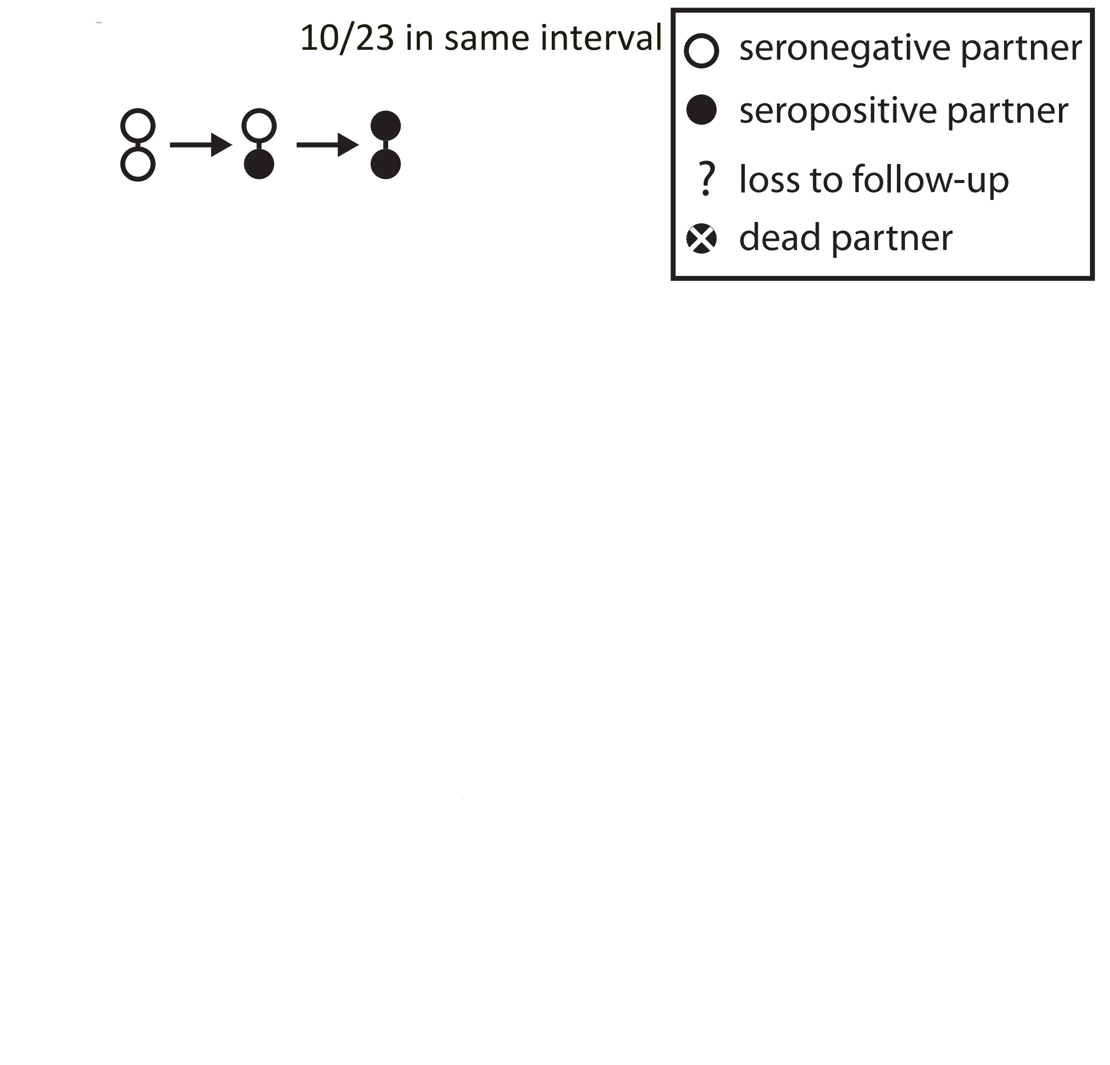


The Rakai Retrospective Cohort Study

acute

chronic

late



Wawer et al. (2005). *Journal of Infectious Diseases*.

Mechanistic Transmission Model

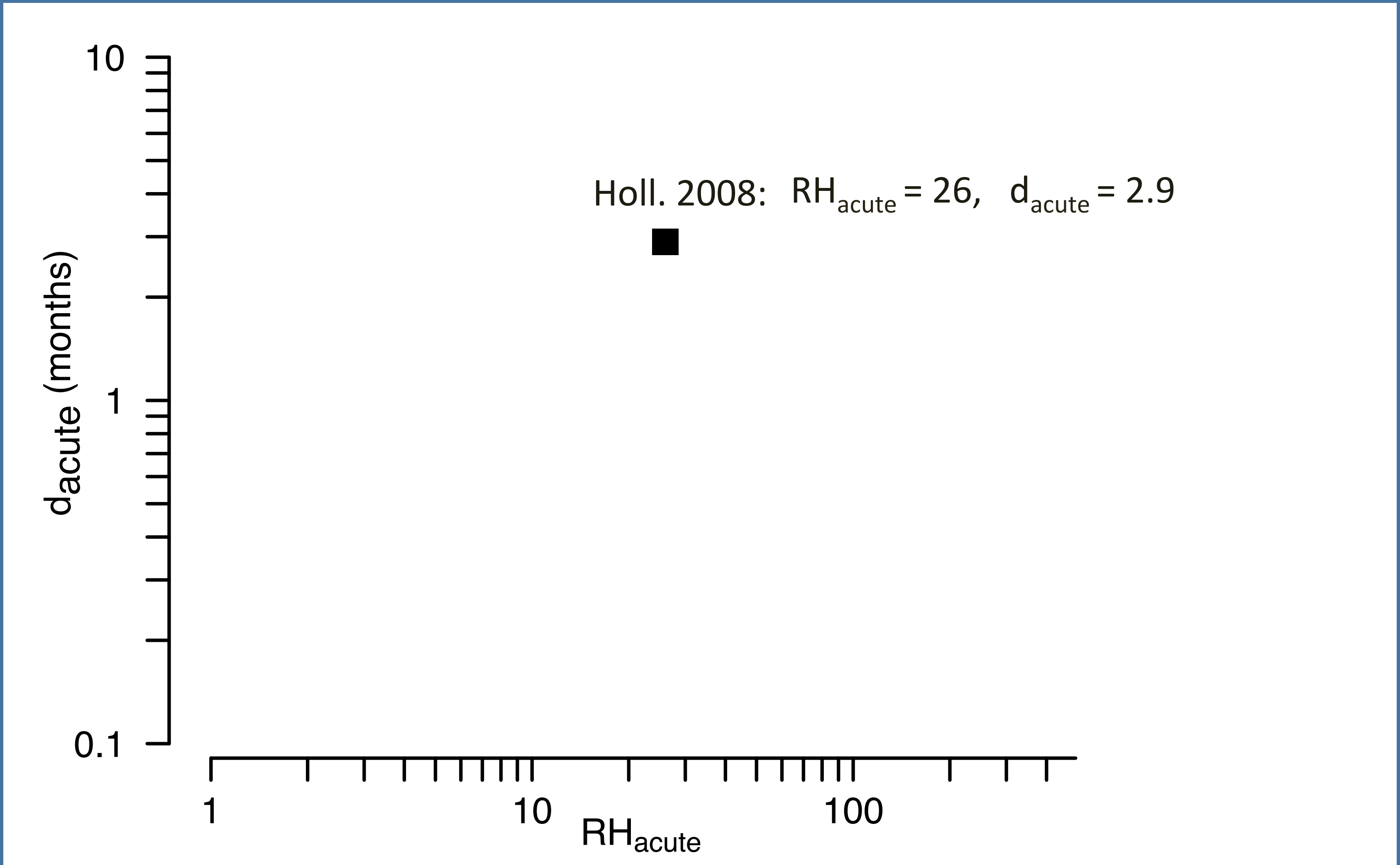
Parameter	Description	Value (95% CI)
β_{acute}	Transmission rate / 100 person-years	276 (131-509)
d_{acute}	Acute phase duration	2.90 (1.23-6.00)
$\beta_{chronic}$	Transmission rate / 100 person-years	10.6 (7.61 – 13.3)

$RH_{acute} = 276/10.6 = 26$

But what about the wide confidence intervals?

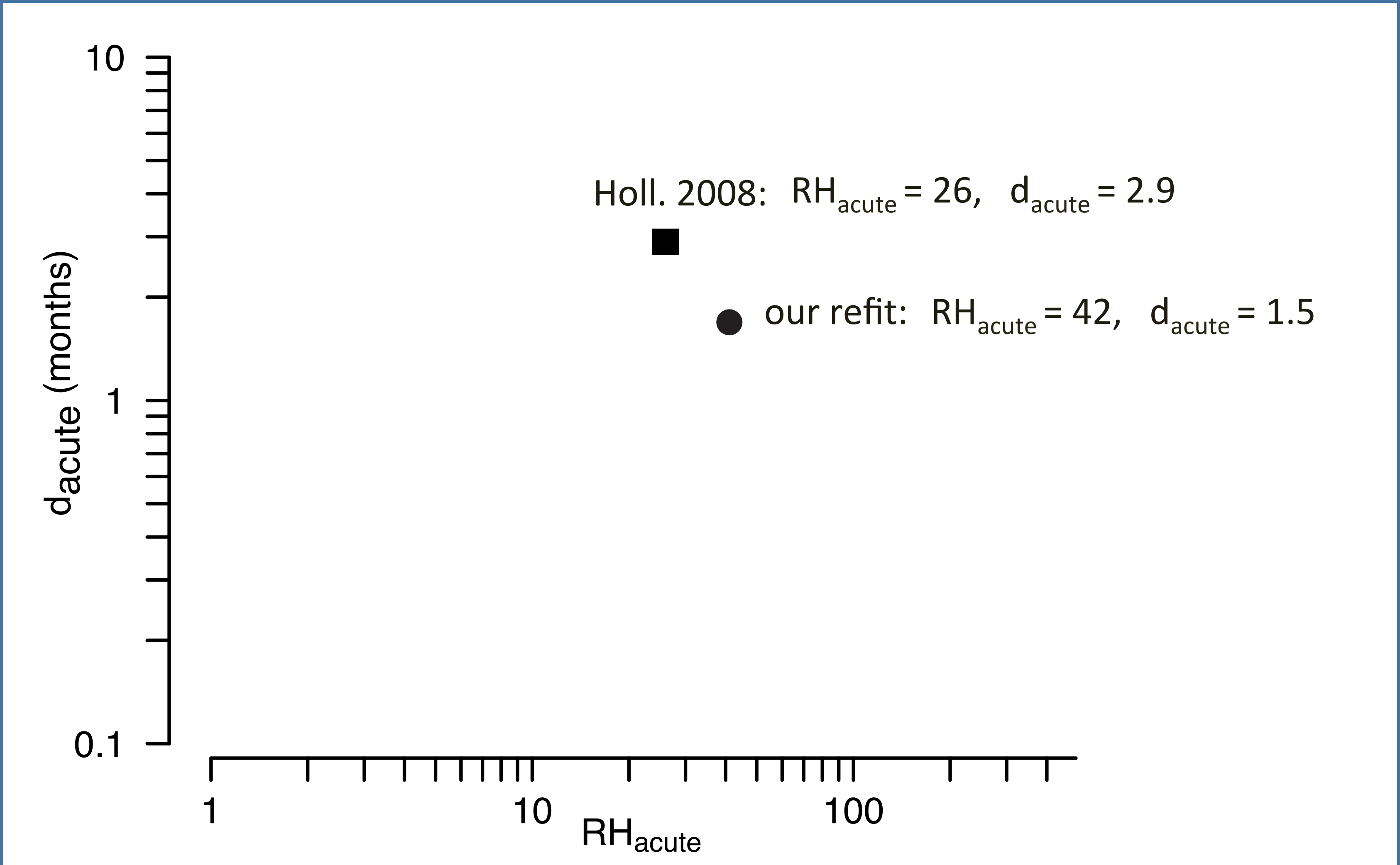
Hollingsworth et al. (2008). *Journal of Infectious Diseases*.

Collinearity in Fitted Parameters



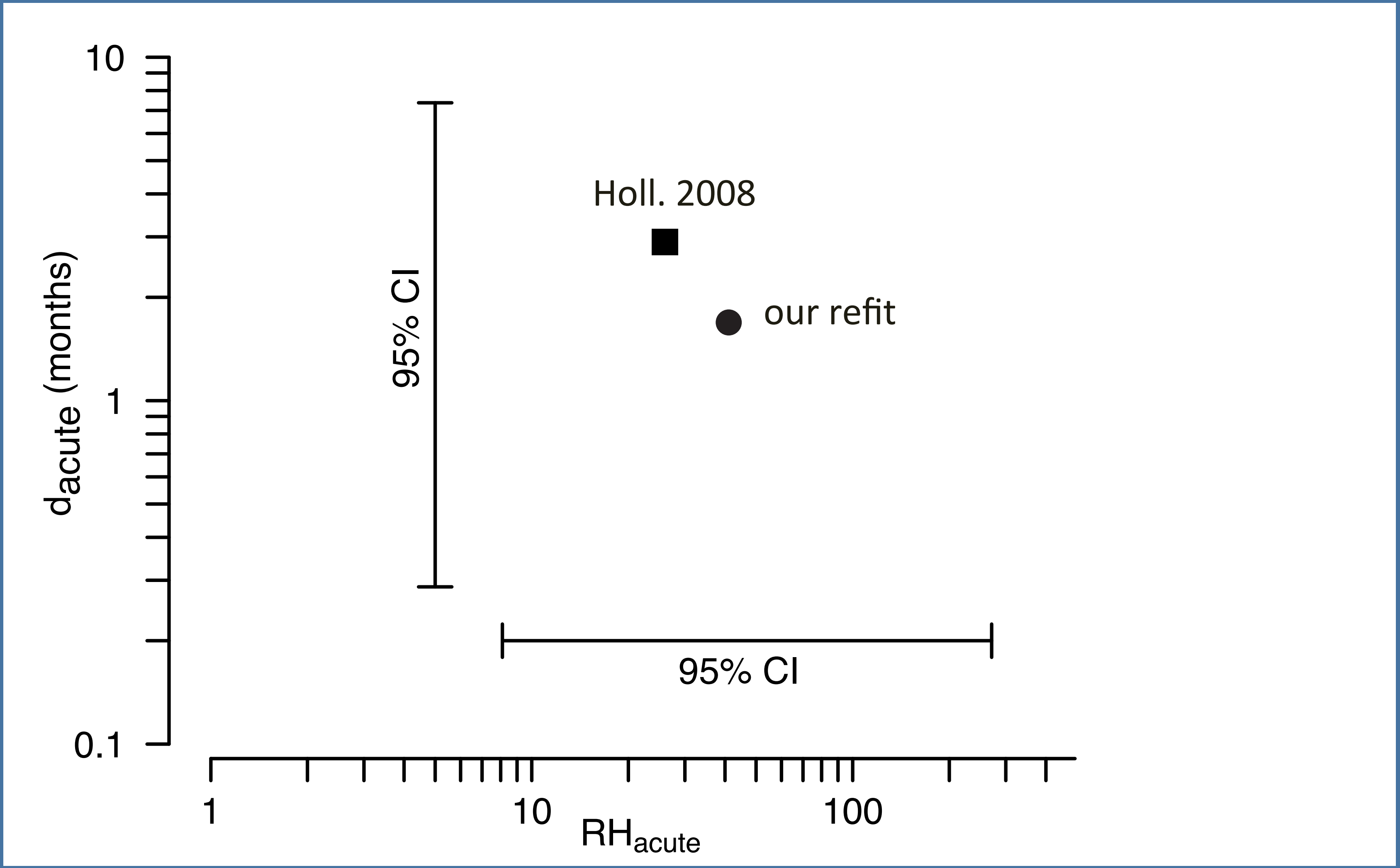
Revisit original data & method.

Collinearity in Fitted Parameters



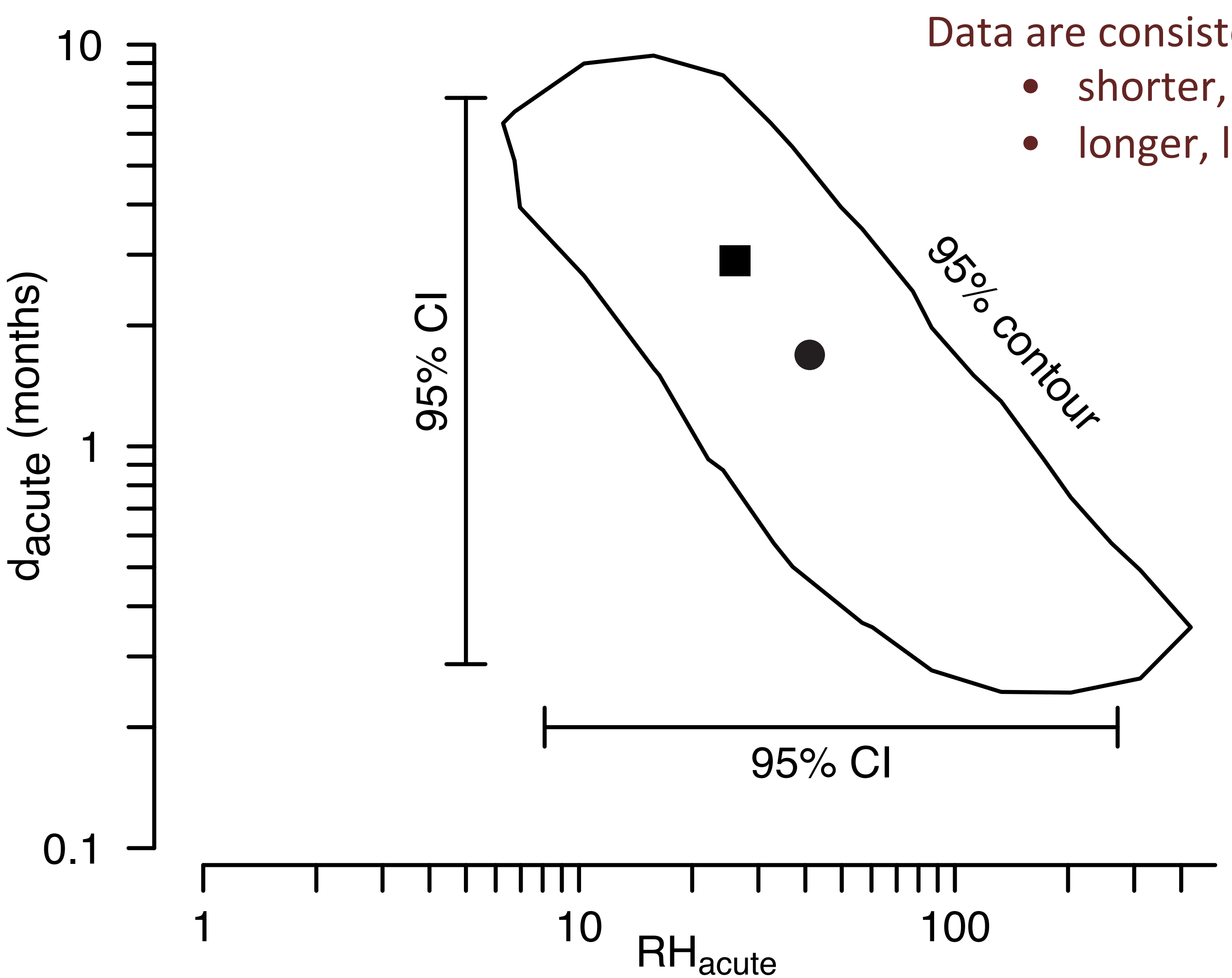
Refit the same model using Bayesian MCMC

Collinearity in Fitted Parameters



Refit the same model using Bayesian MCMC

Collinearity in Fitted Parameters

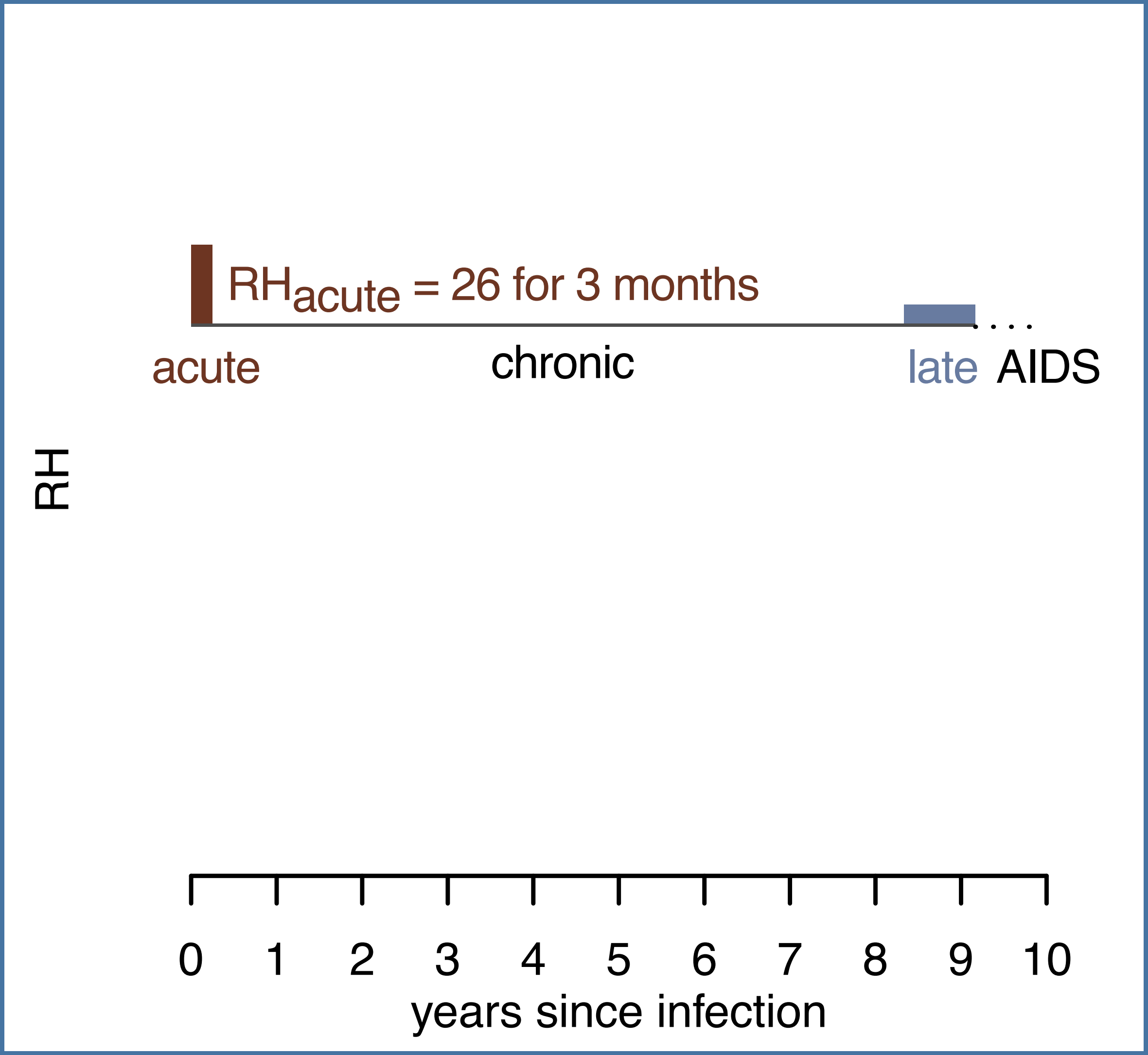


Data are consistent with both
- shorter, highly infectious
- longer, less infectious

acute phases

Refit the same model using Bayesian MCMC

Collinearity in Fitted Parameters



What is actually Identifiable?

Excess Hazard-Months due to acute phase

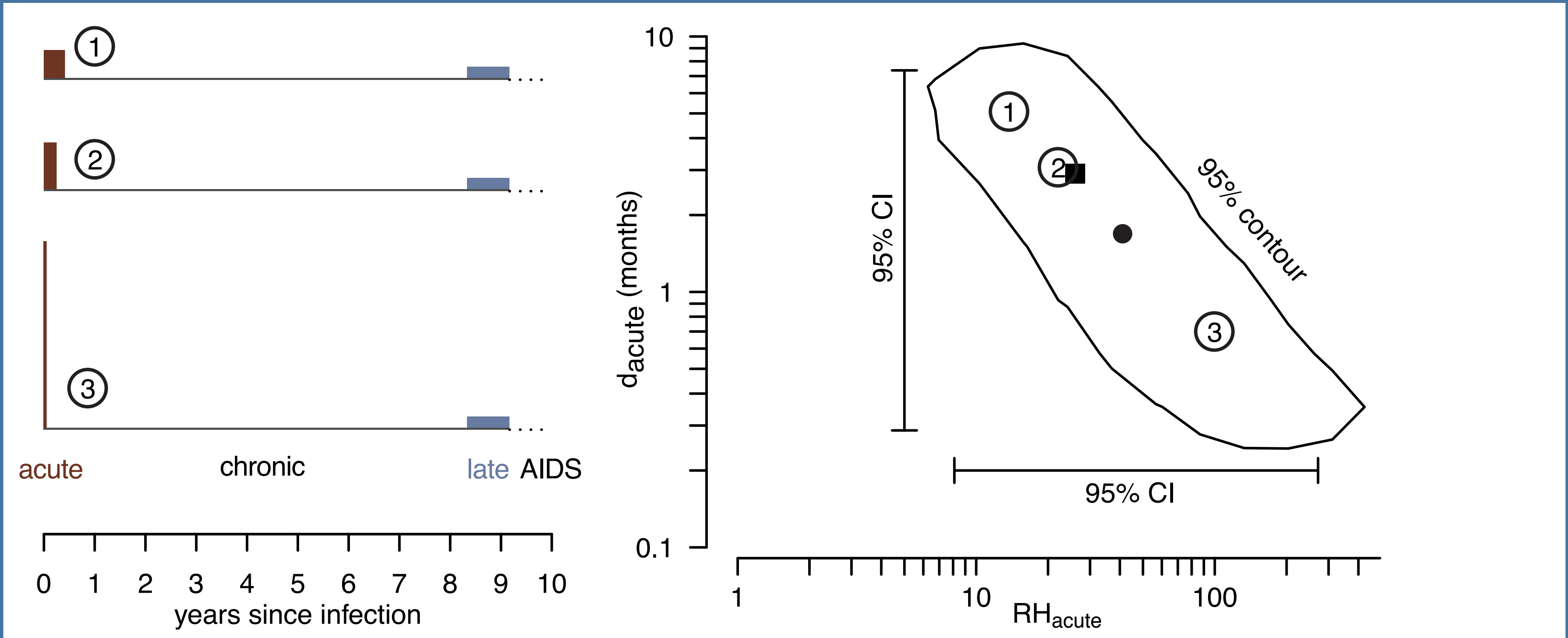
$EHM_{acute} = (RH_{acute}-1)d_{acute}$

$EHM_{acute} = 25*3 = 75$

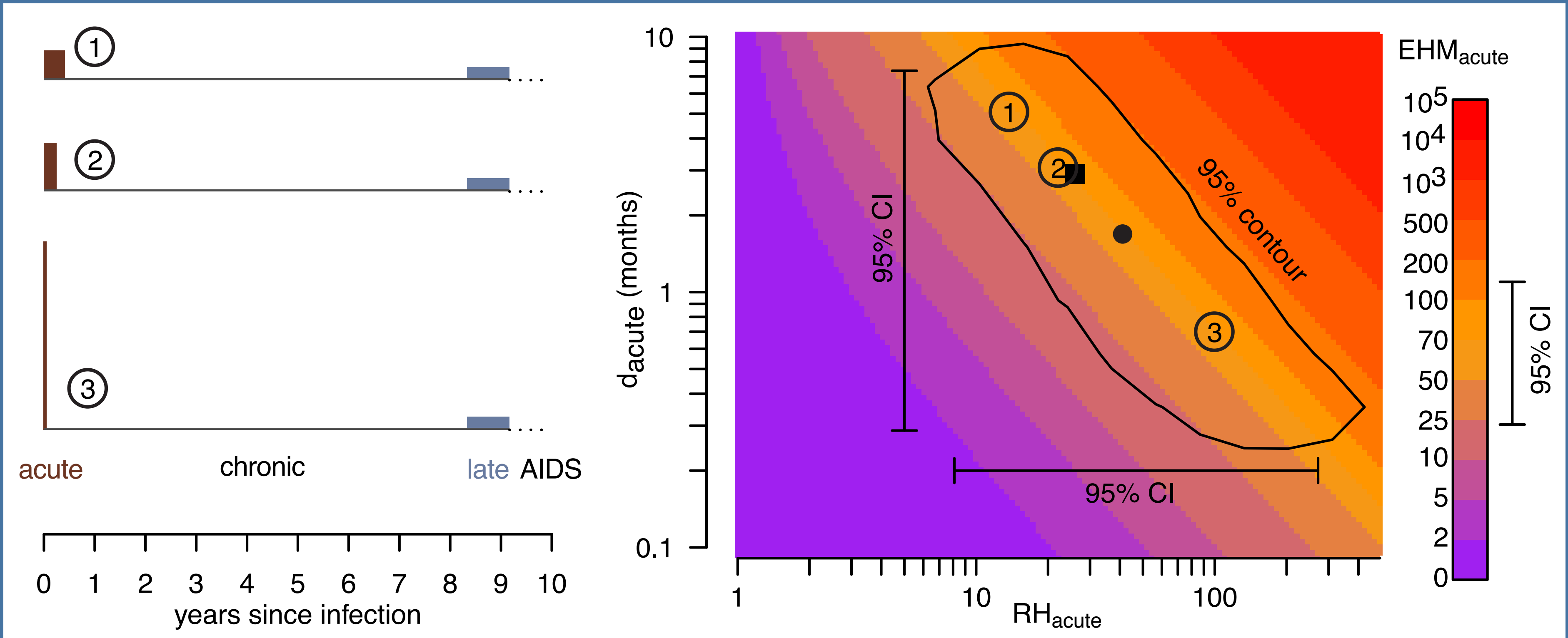
$EHM_{acute} = 15*5 = 75$

$EHM_{acute} = 100*3/4 = 75$

Excess Hazard Months (EHM_{acute})



Excess Hazard Months (EHM_{acute})



RH_{acute} and d_{acute} are not identifiable from 10-month interval cohorts

We should focus on EHM_{acute}

Formally vs Informally Fitting

- Most modeling studies do not fit data formally
- Unnecessary for demonstration of qualitative dynamics
- Necessary for

    parameter estimation

    inference

    formal model comparison

Learning More: Methods for Fitting

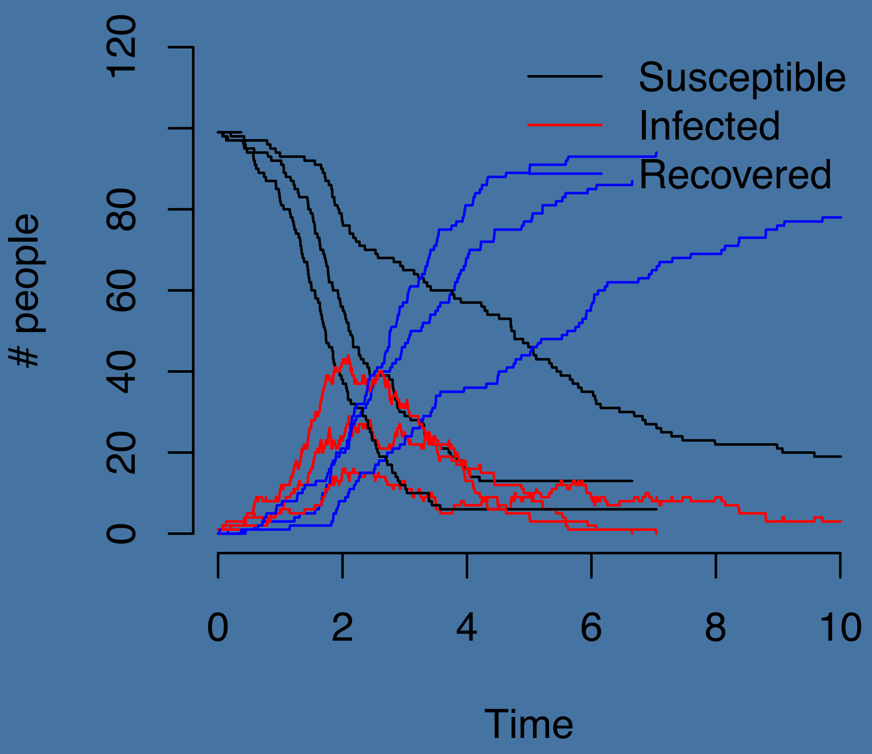
- Least Squares
- Frequentist Maximum Likelihood Fitting
- Bayesian Posterior Estimation (usually MCMC)

Simulating to test methods

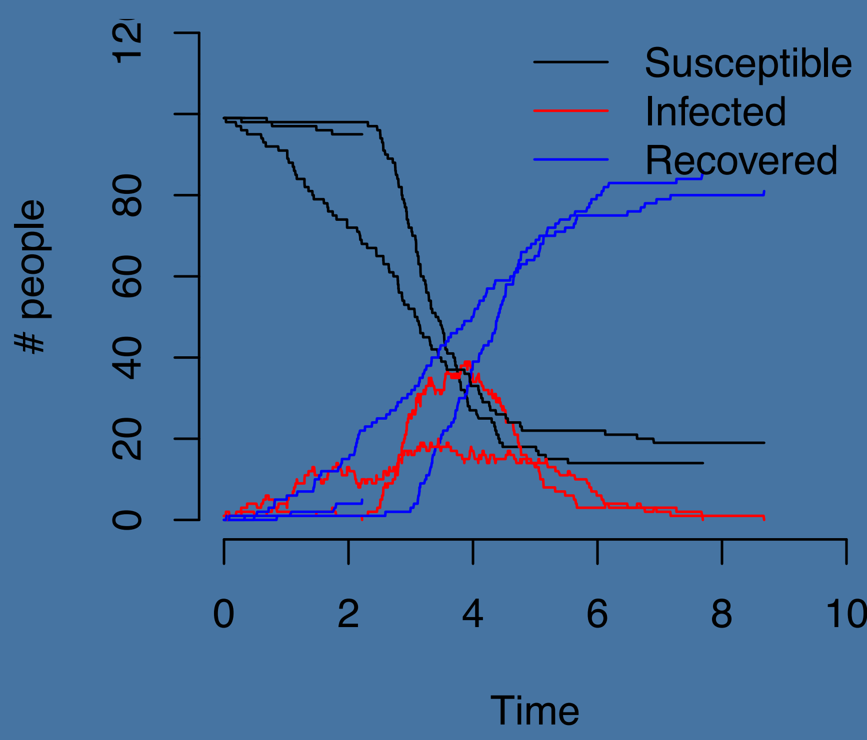
- Create model
- Simulate data
- Can you estimate the inputted parameters for the simulation by fitting?

Simulating to test methods

5 Urban Villages



5 Rural Villages



Summary

- Why we fit
  - parameter estimation
  - inference
  - formal model comparison

- How we fit
  - Create a probabilistic framework that links our model to data—ie, write a likelihood

- What to consider when fitting
  - Assumptions
  - Overfitting
  - Goodness of fit
  - Identifiability