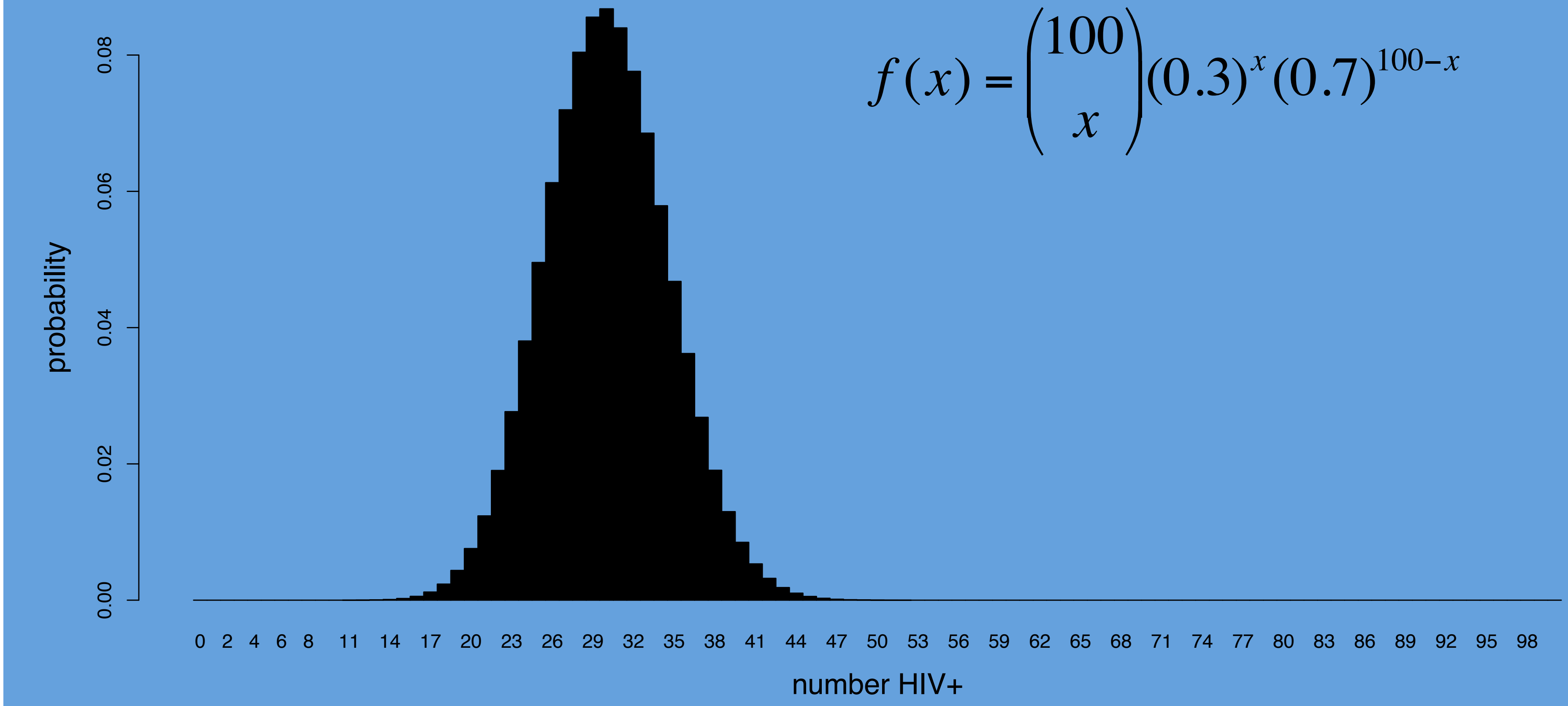
Introduction to Likelihood

Meaningful Modeling of Epidemiologic Data, 2011
AIMS, Muizenberg, South Africa

Steve Bellan
MPH Epidemiology
PhD Candidate
Department of Environmental Science, Policy & Management
University of California at Berkeley

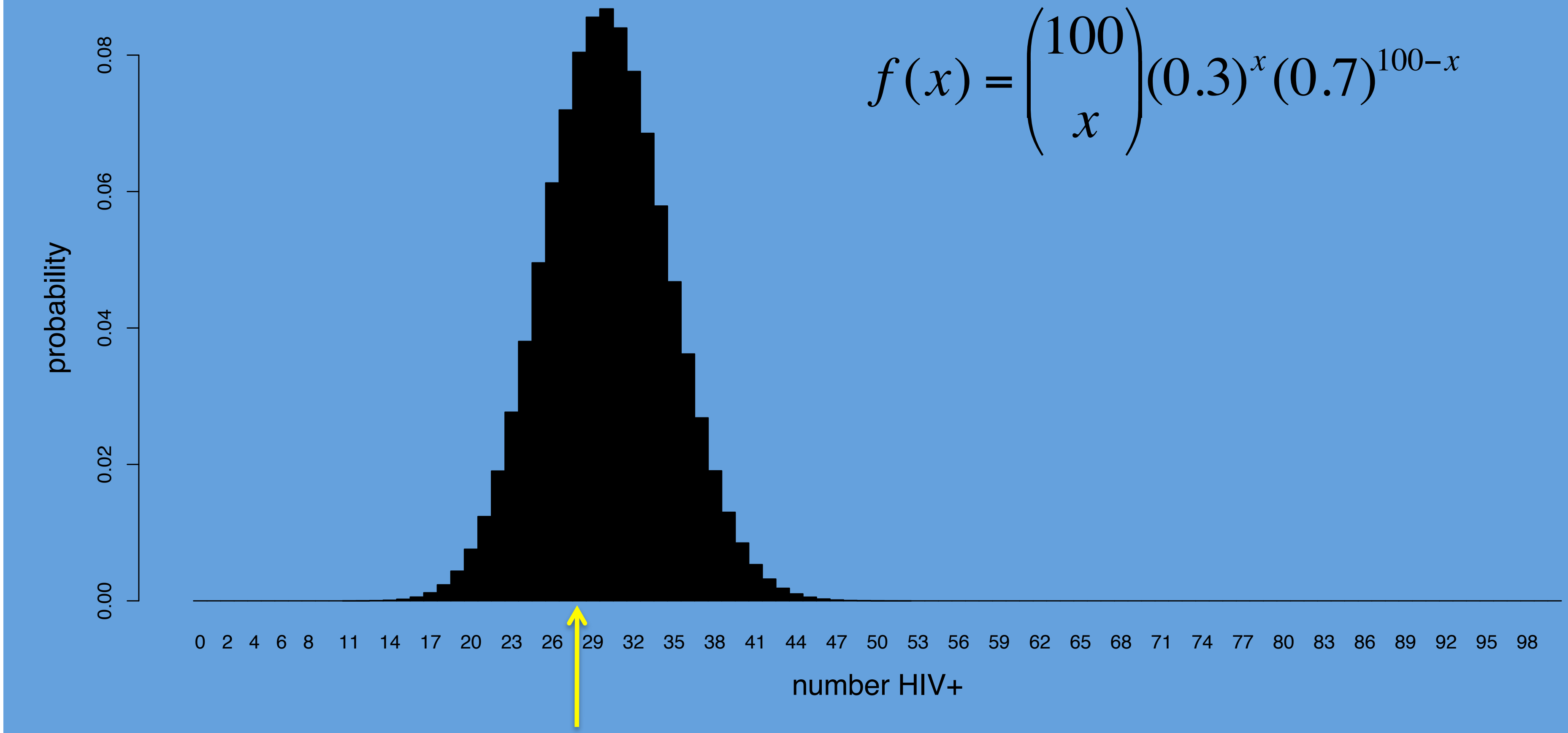
In a population of 1,000,000 people with a true prevalence of 30%, the probability distribution of number of positive individuals if 100 are sampled:



$$f(x) = \binom{100}{x}(0.3)^x(0.7)^{100-x}$$

```
barplot(dbinom(x = 0:100, size = 100, prob = .3), names.arg = 0:size)
```

In a population of 1,000,000 people with a true prevalence of 30%, the probability distribution of number of positive individuals if 100 are sampled:



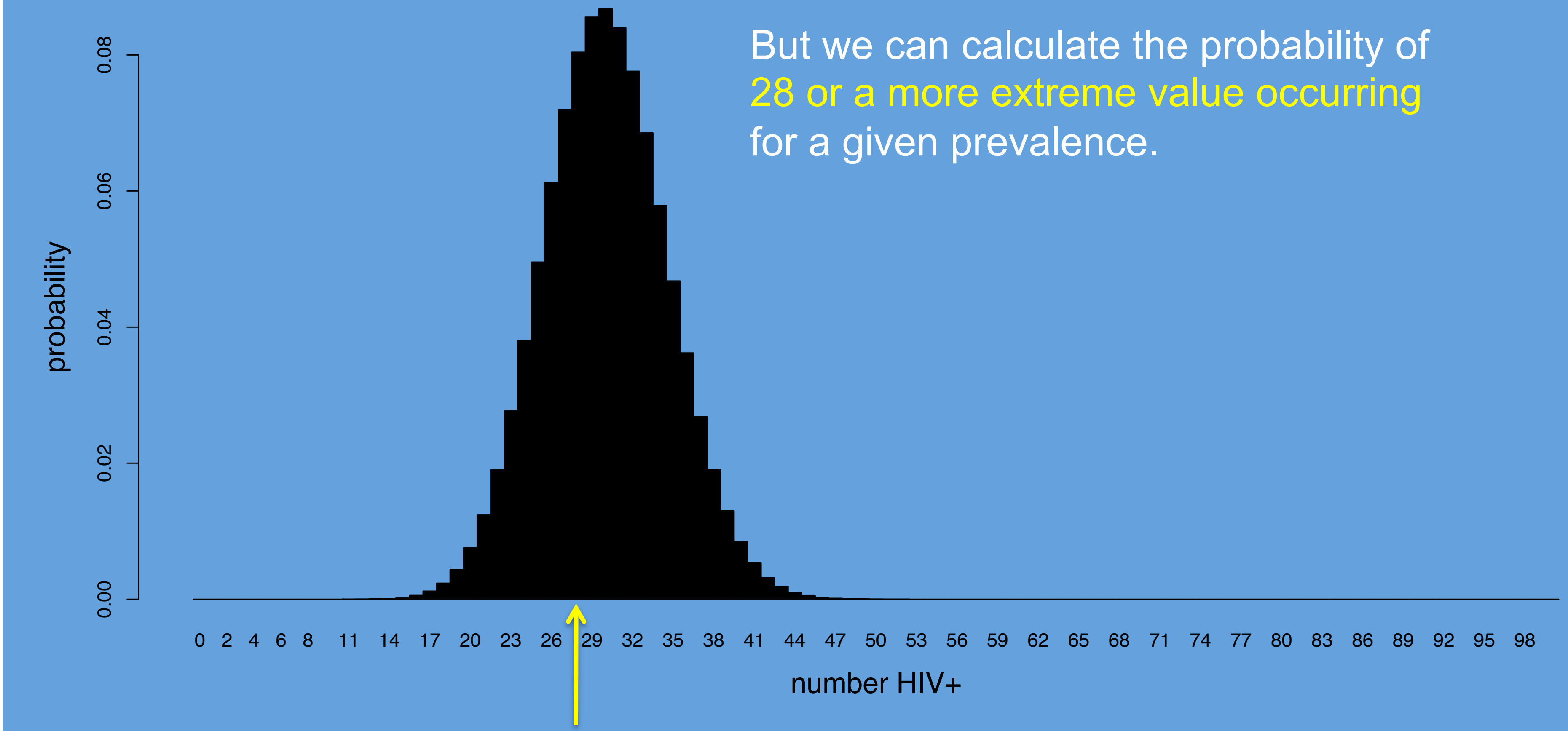
$$f(x) = \binom{100}{x}(0.3)^x(0.7)^{100-x}$$

We sample 100 people once and 28 are positive:

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

We don’t know the true prevalence!

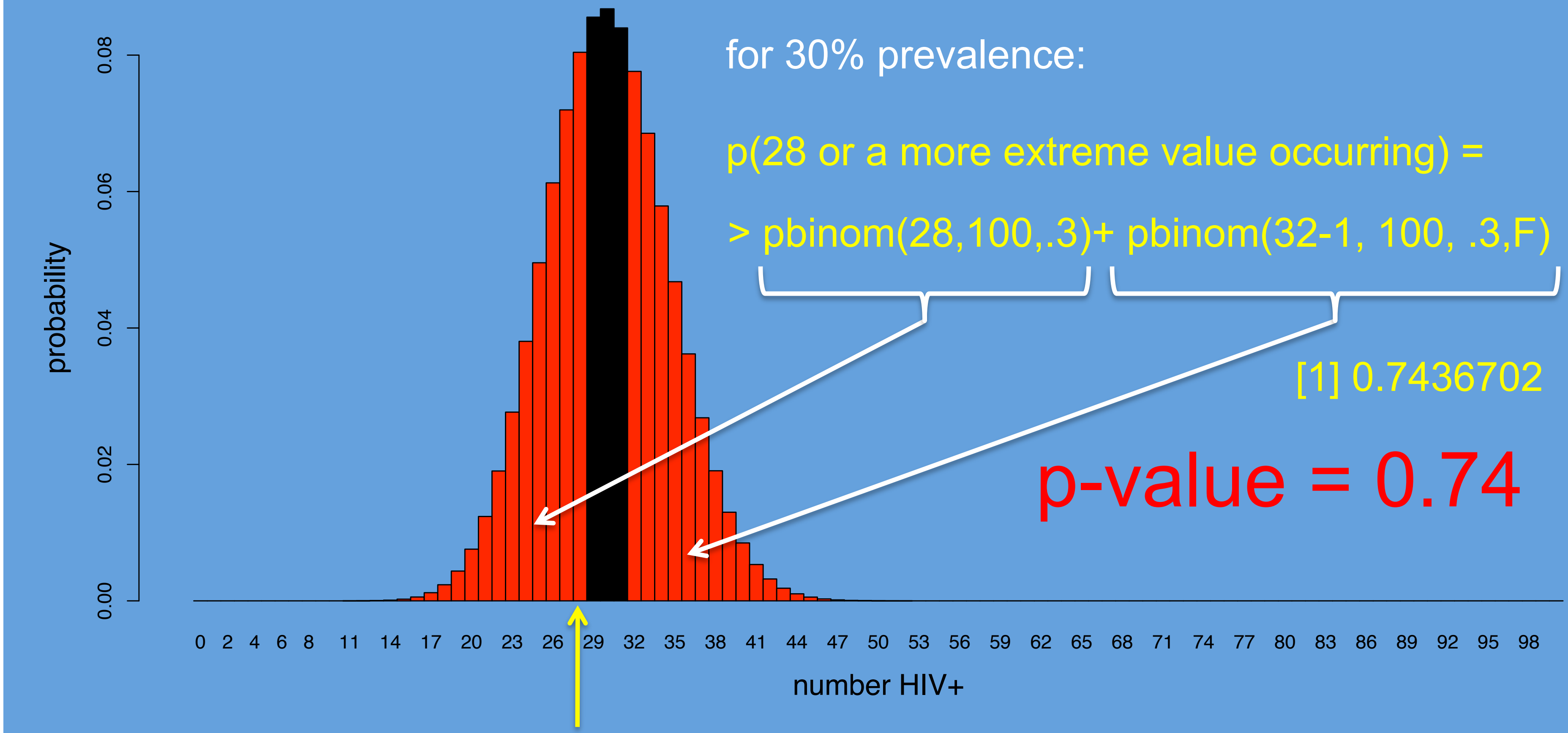
But we can calculate the probability of 28 or a more extreme value occurring for a given prevalence.



We sample 100 people once and 28 are positive:

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

Cumulative Probability & P Values



for 30% prevalence:

p(28 or a more extreme value occurring) =

```
> pbinom(28,100,.3)+ pbinom(32-1, 100, .3,F)
```

```
[1] 0.7436702
```

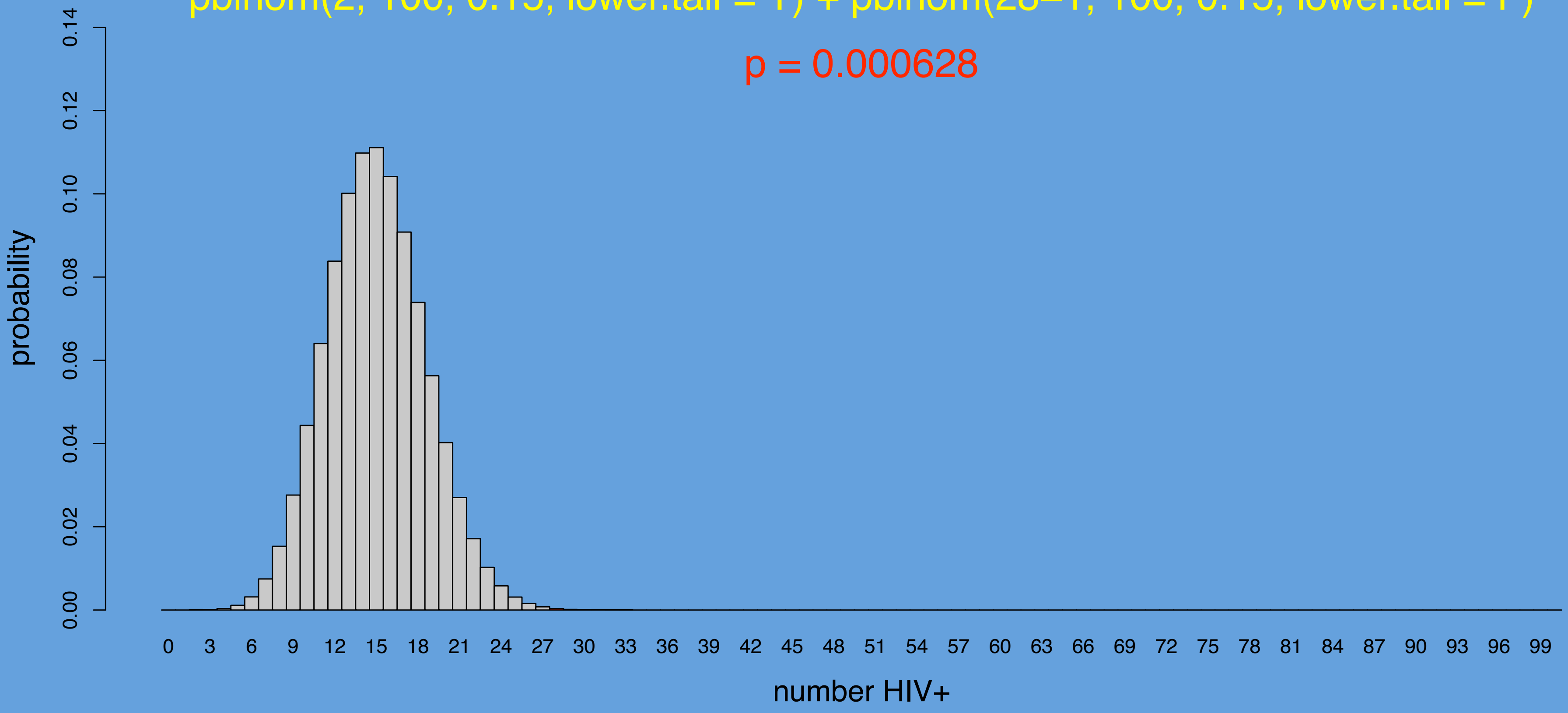
p-value = 0.74

We sample 100 people once and 28 are positive.

If true prevalence were 15%, then p(28 or more extreme) is

pbinom(2, 100, 0.15, lower.tail = T) + pbinom(28−1, 100, 0.15, lower.tail = F)

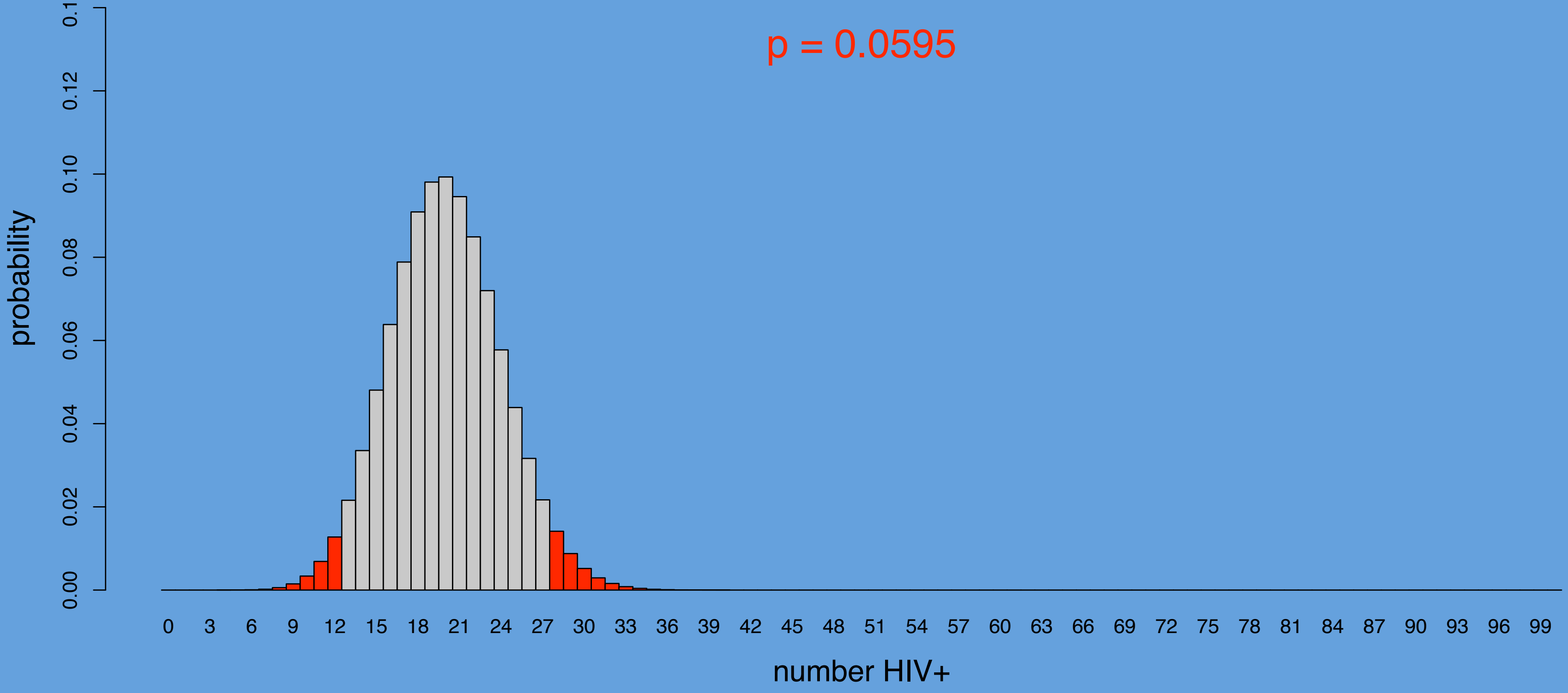
p = 0.000628



If true prevalence were 20%, then p(28 or more extreme) is

pbinom(12, 100, 0.2, lower.tail = T) + pbinom(28–1, 100, 0.2, lower.tail = F)

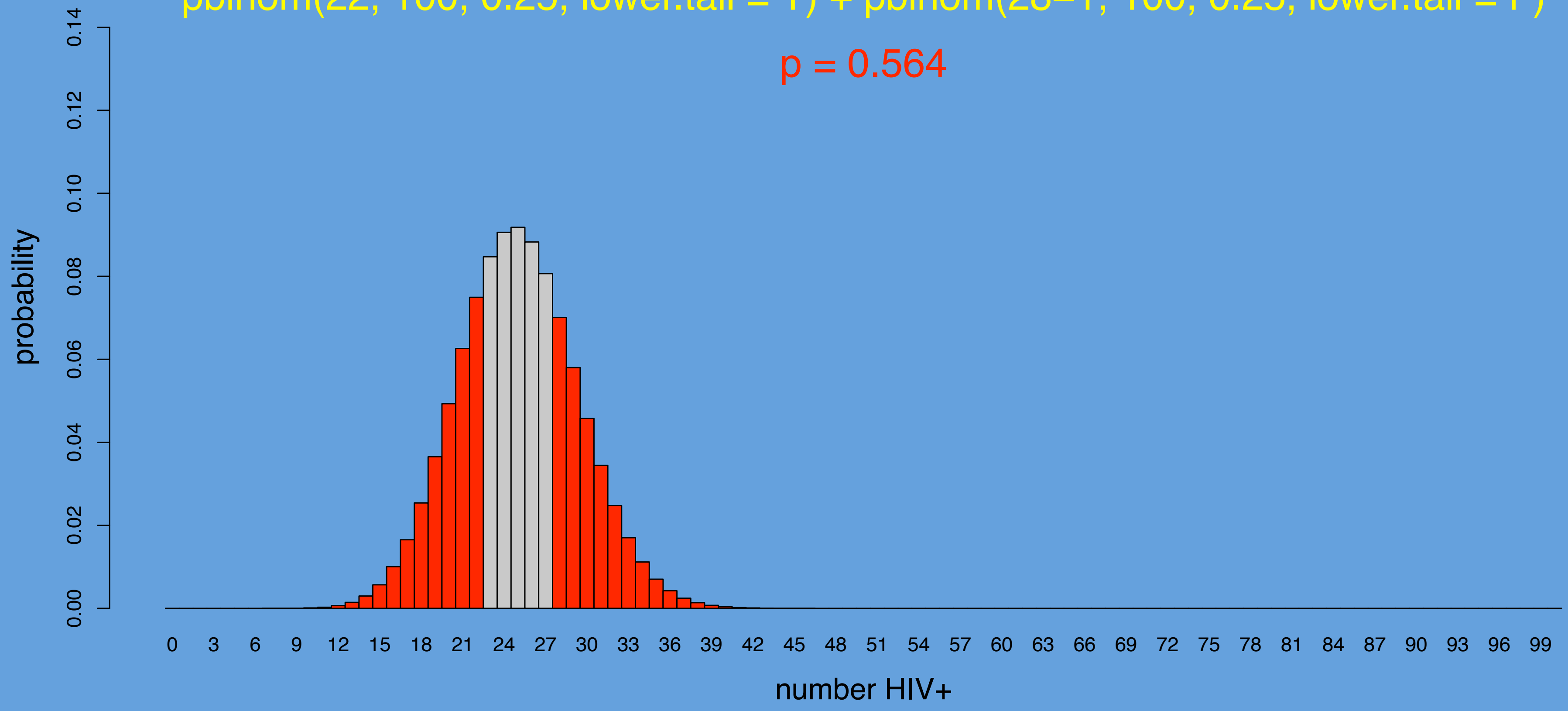
p = 0.0595



If true prevalence were 25%, then p(28 or more extreme) is

pbinom(22, 100, 0.25, lower.tail = T) + pbinom(28–1, 100, 0.25, lower.tail = F)

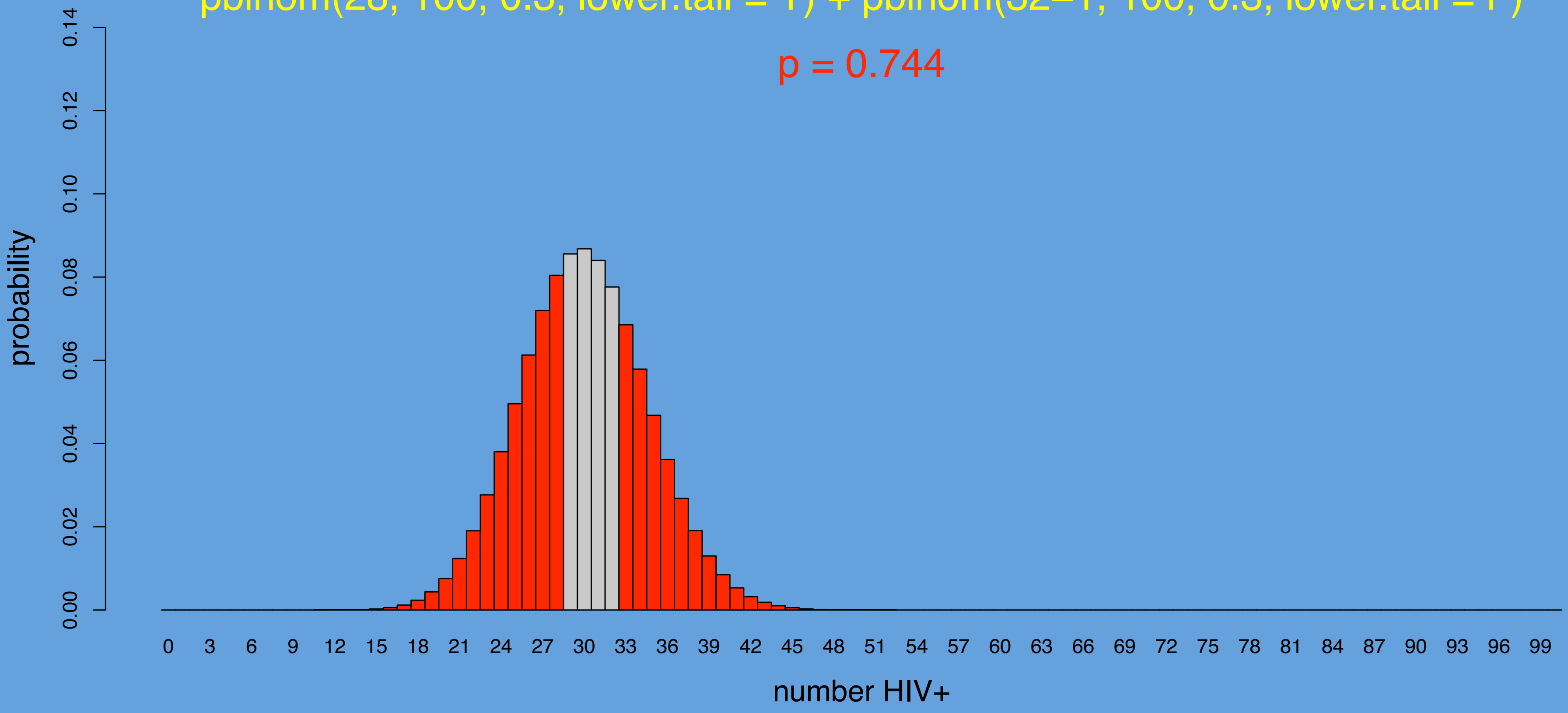
p = 0.564



If true prevalence were 30%, then p(28 or more extreme) is

pbinom(28, 100, 0.3, lower.tail = T) + pbinom(32–1, 100, 0.3, lower.tail = F)

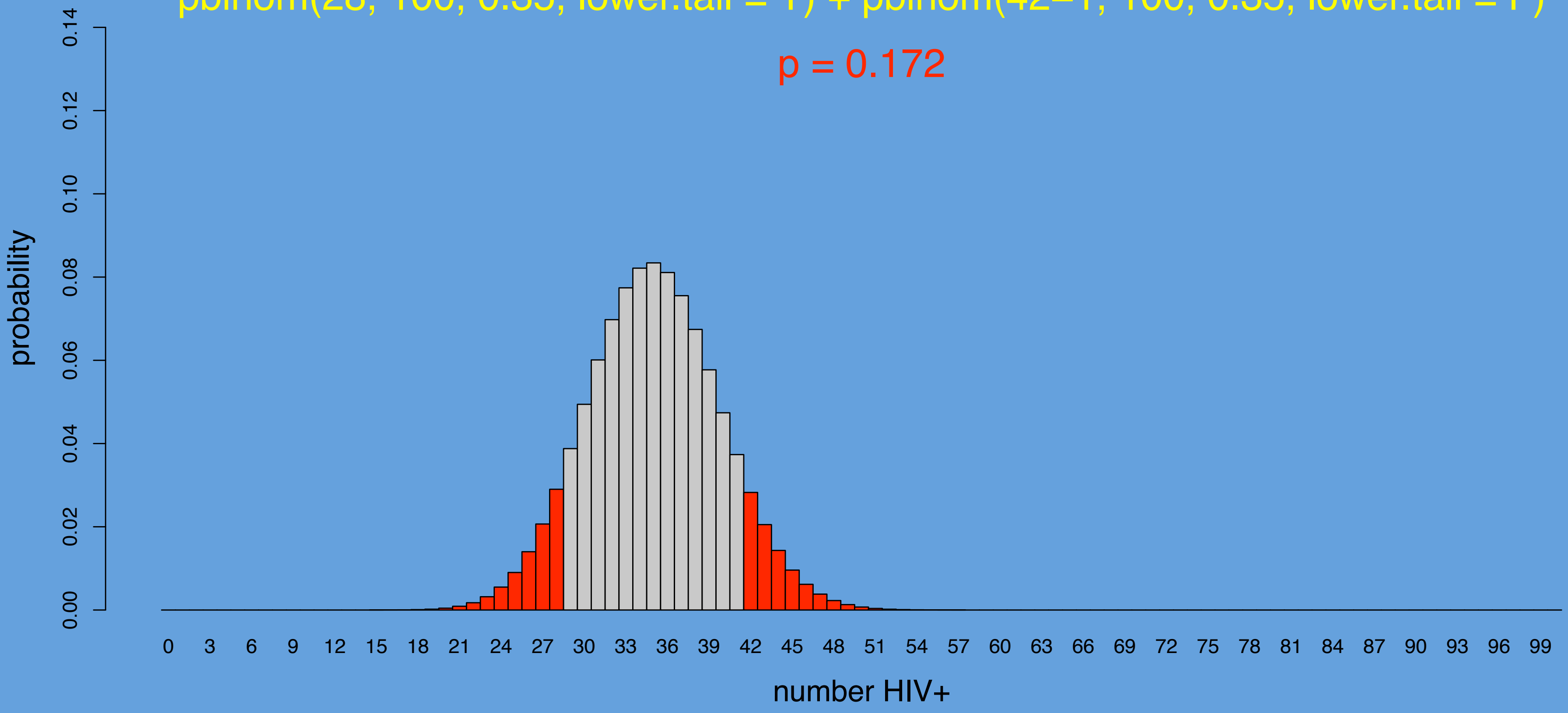
p = 0.744



If true prevalence were 35%, then p(28 or more extreme) is

pbinom(28, 100, 0.35, lower.tail = T) + pbinom(42−1, 100, 0.35, lower.tail = F)

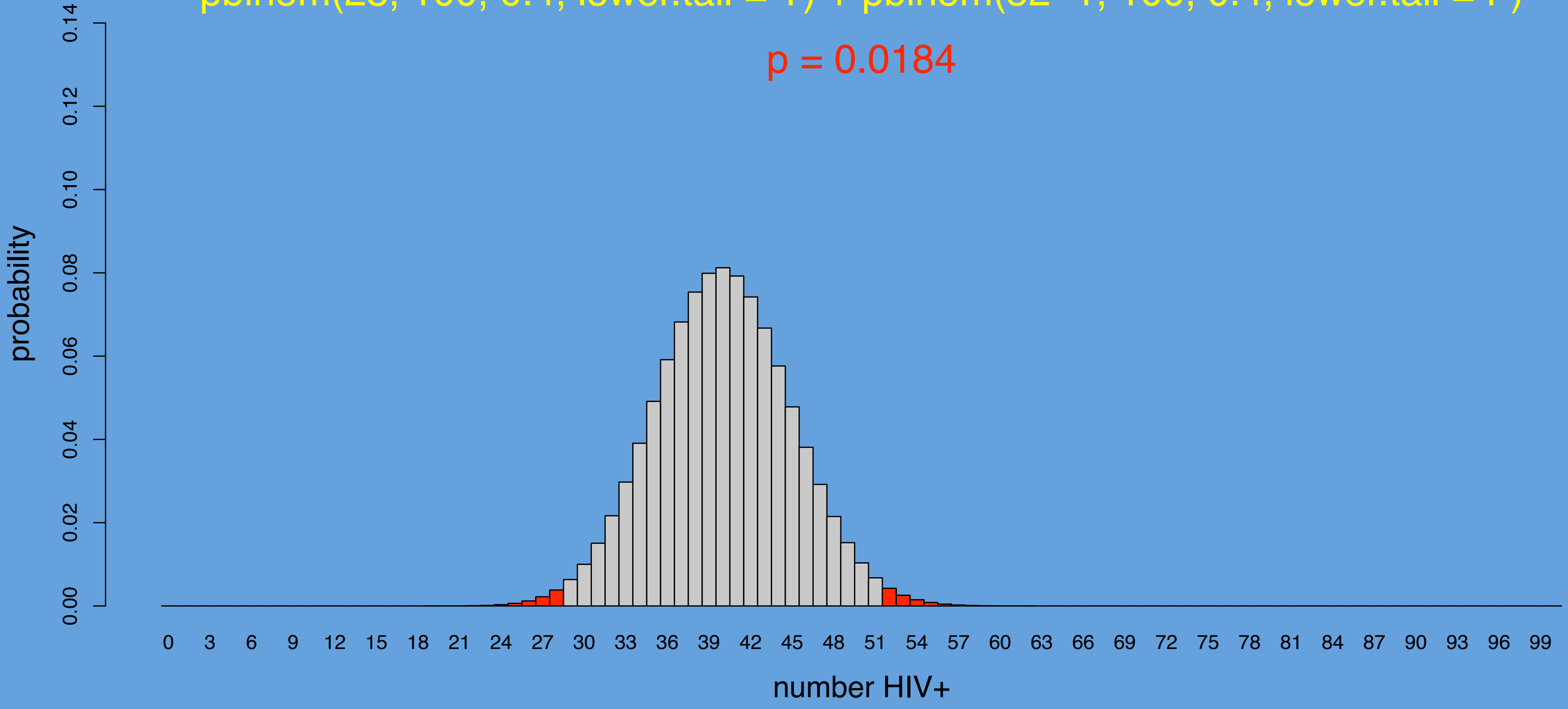
p = 0.172

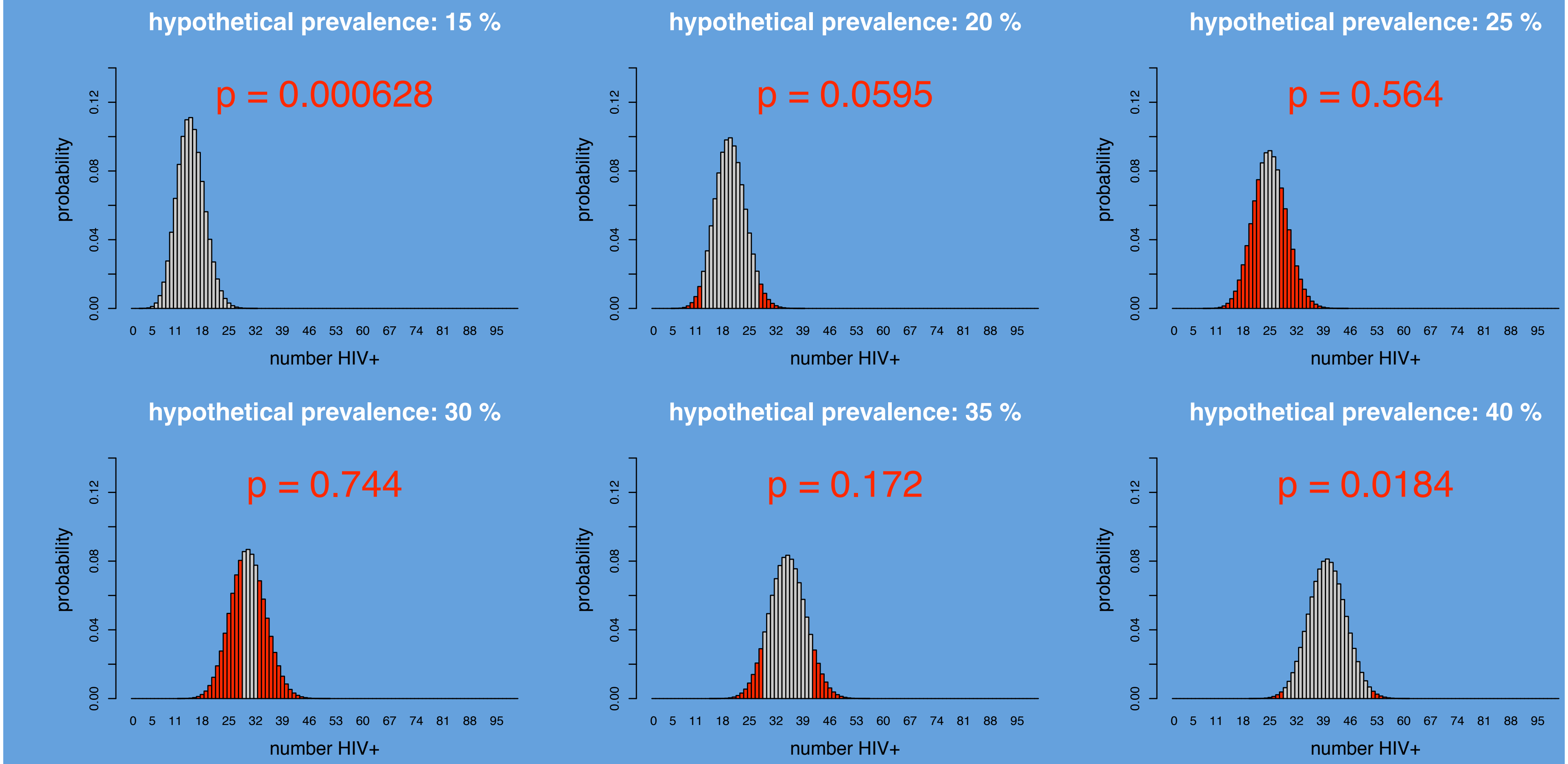


If true prevalence were 40%, then p(28 or more extreme) is

pbinom(28, 100, 0.4, lower.tail = T) + pbinom(52–1, 100, 0.4, lower.tail = F)

p = 0.0184





hypothetical prevalence: 15 %

p = 0.000628

hypothetical prevalence: 20 %

p = 0.0595

hypothetical prevalence: 25 %

p = 0.564

hypothetical prevalence: 30 %

p = 0.744

hypothetical prevalence: 35 %

p = 0.172

hypothetical prevalence: 40 %

p = 0.0184

Which hypotheses do we reject?

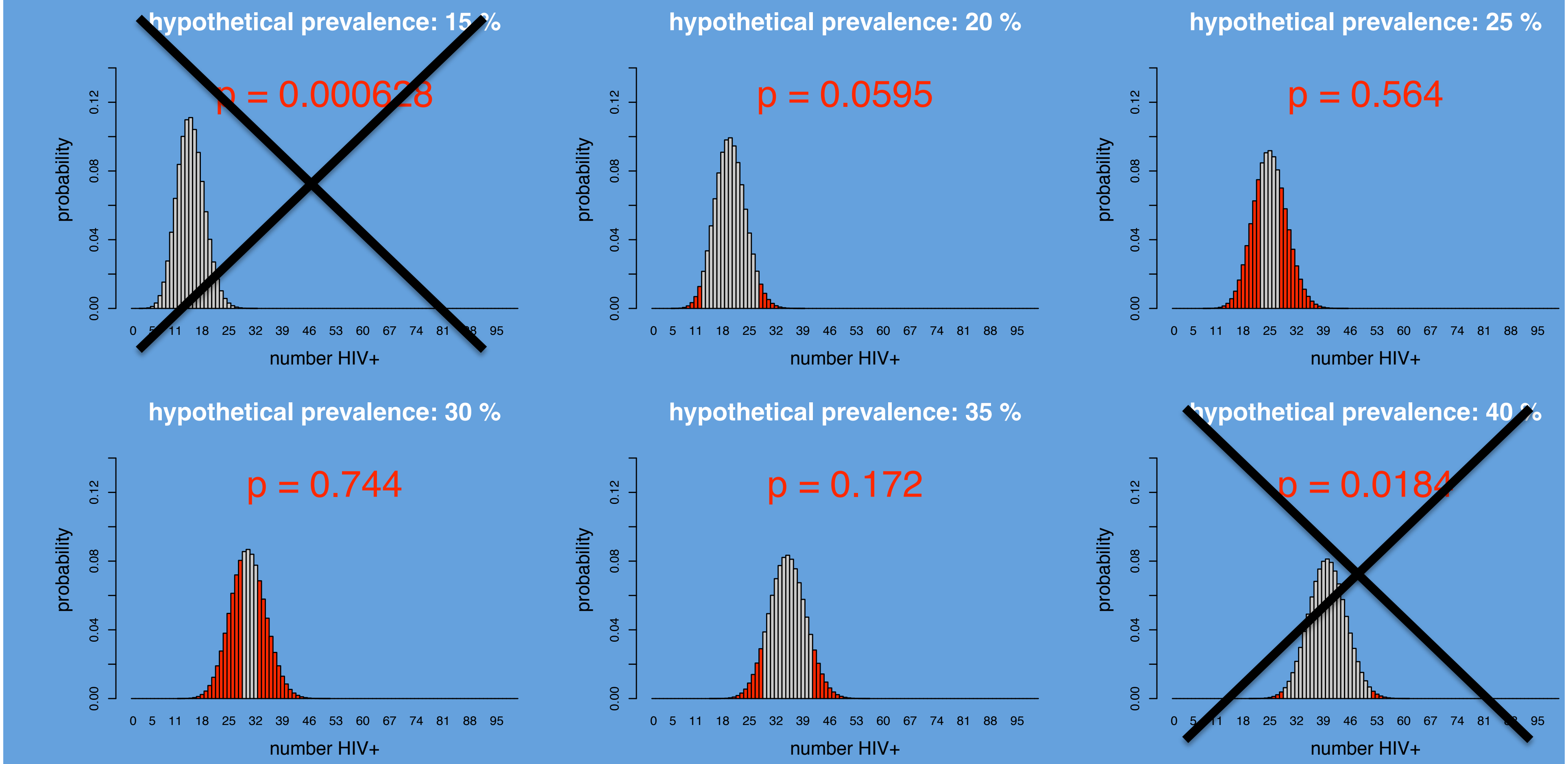
IF GIVEN THE HYPOTHESIS

p value < cutoff

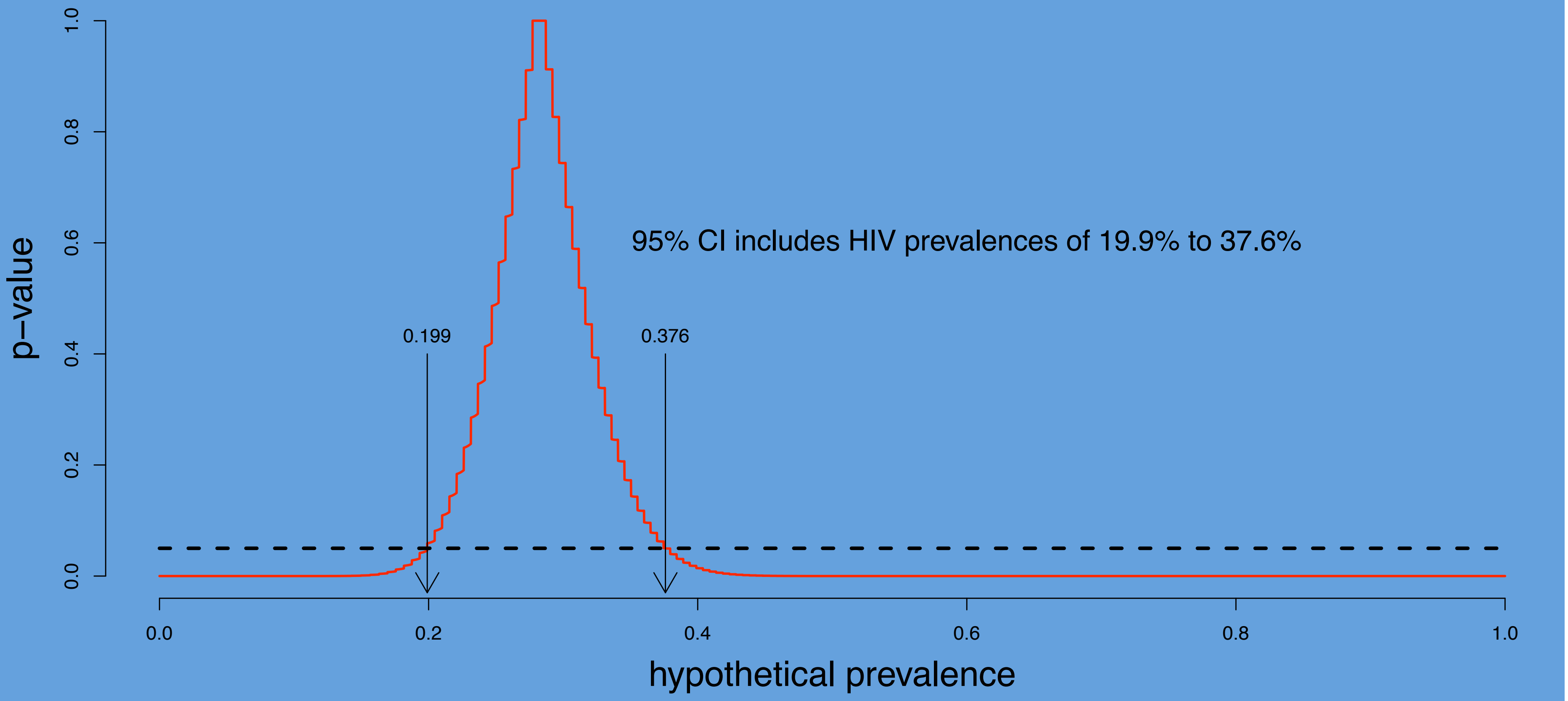
THEN REJECT HYPOTHESIS

Cutoff usually chosen as $\alpha = 0.05$

Which hypotheses do we reject?



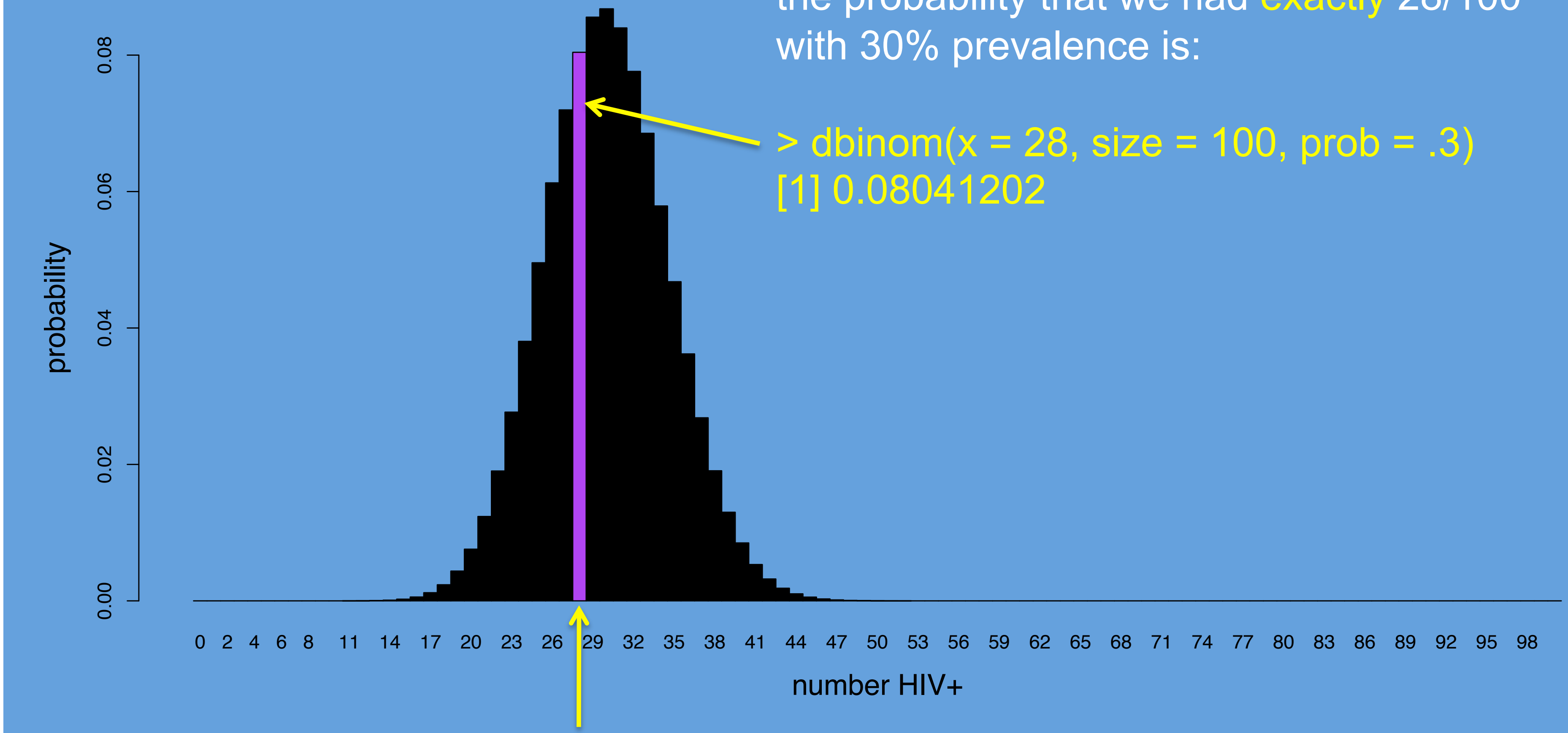
Which hypotheses do we NOT reject: CONFIDENCE INTERVAL



Let' s take another approach

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
```

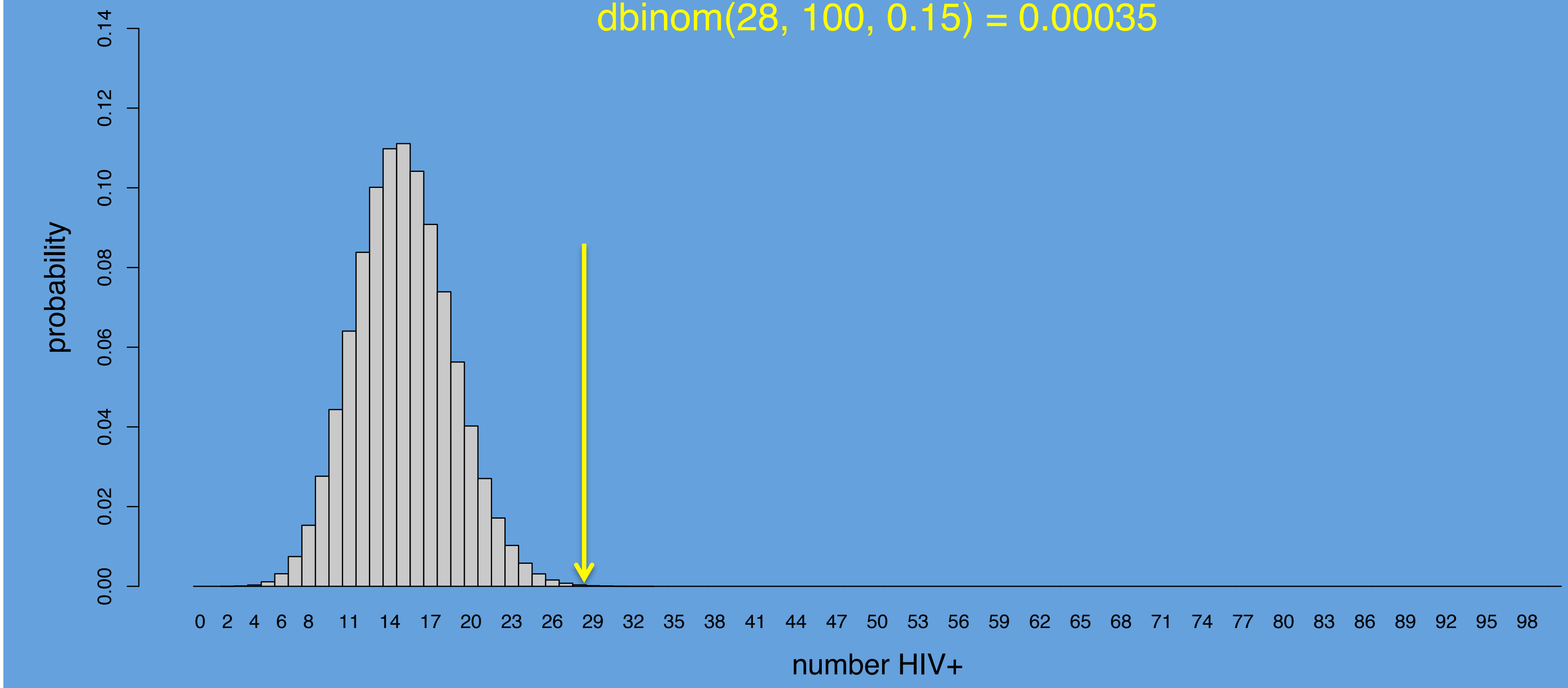


We sample 100 people once and 28 are positive:

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

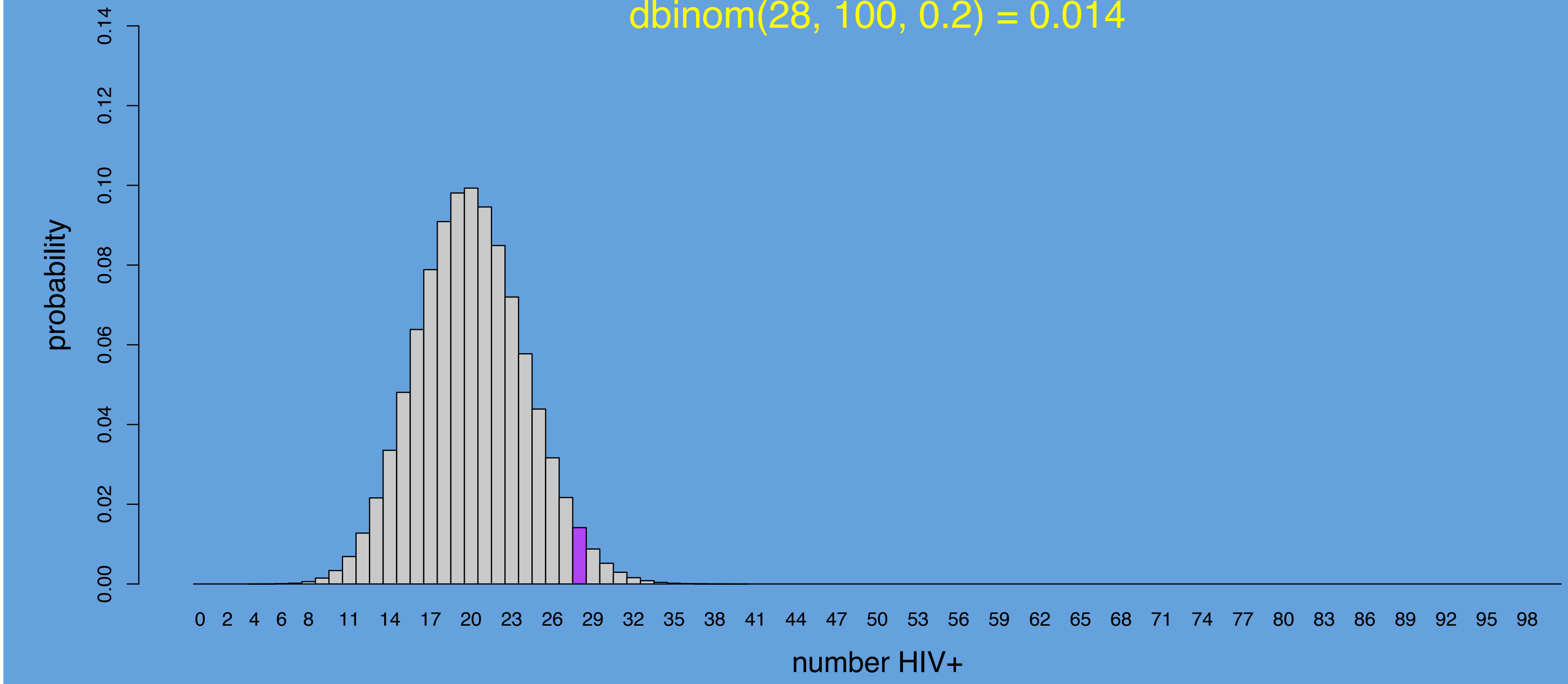
hypothetical prevalence: 15 %

dbinom(28, 100, 0.15) = 0.00035



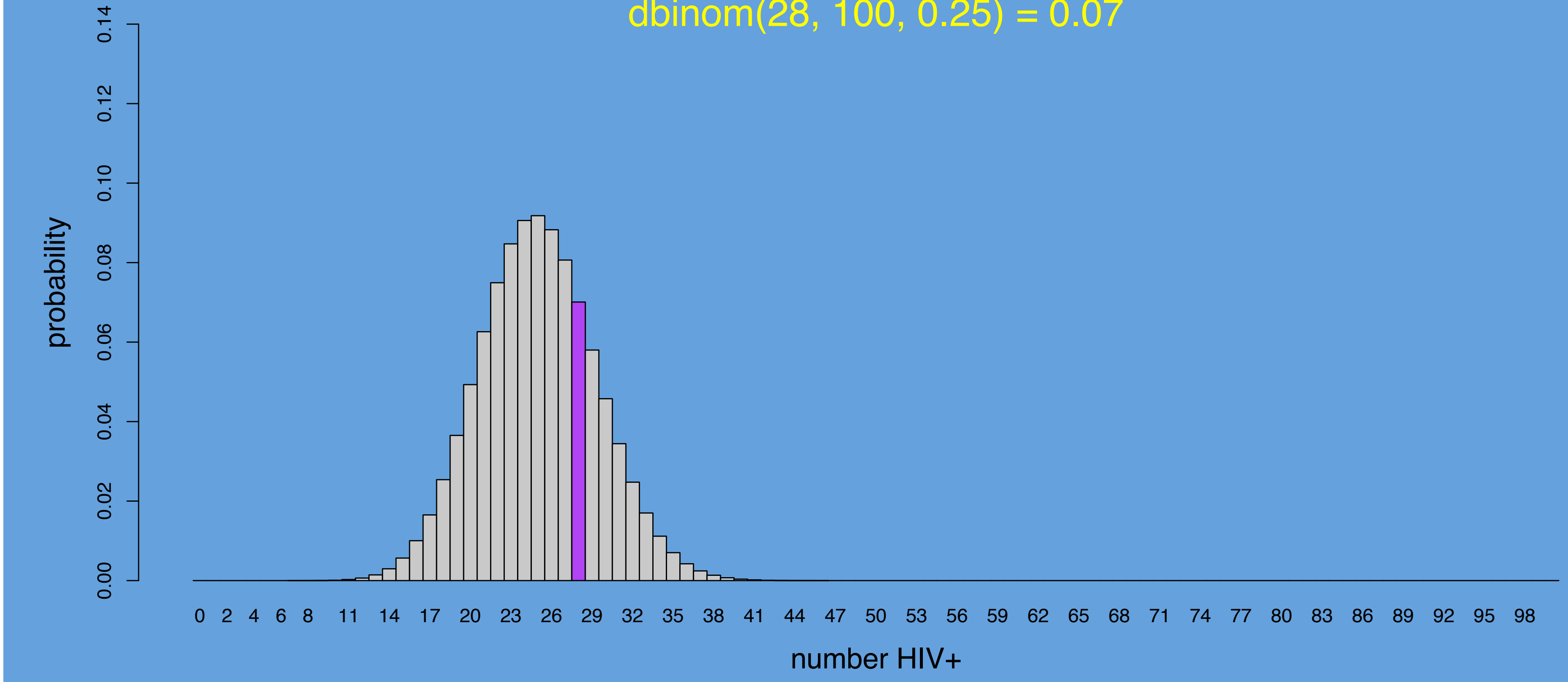
hypothetical prevalence: 20 %

dbinom(28, 100, 0.2) = 0.014



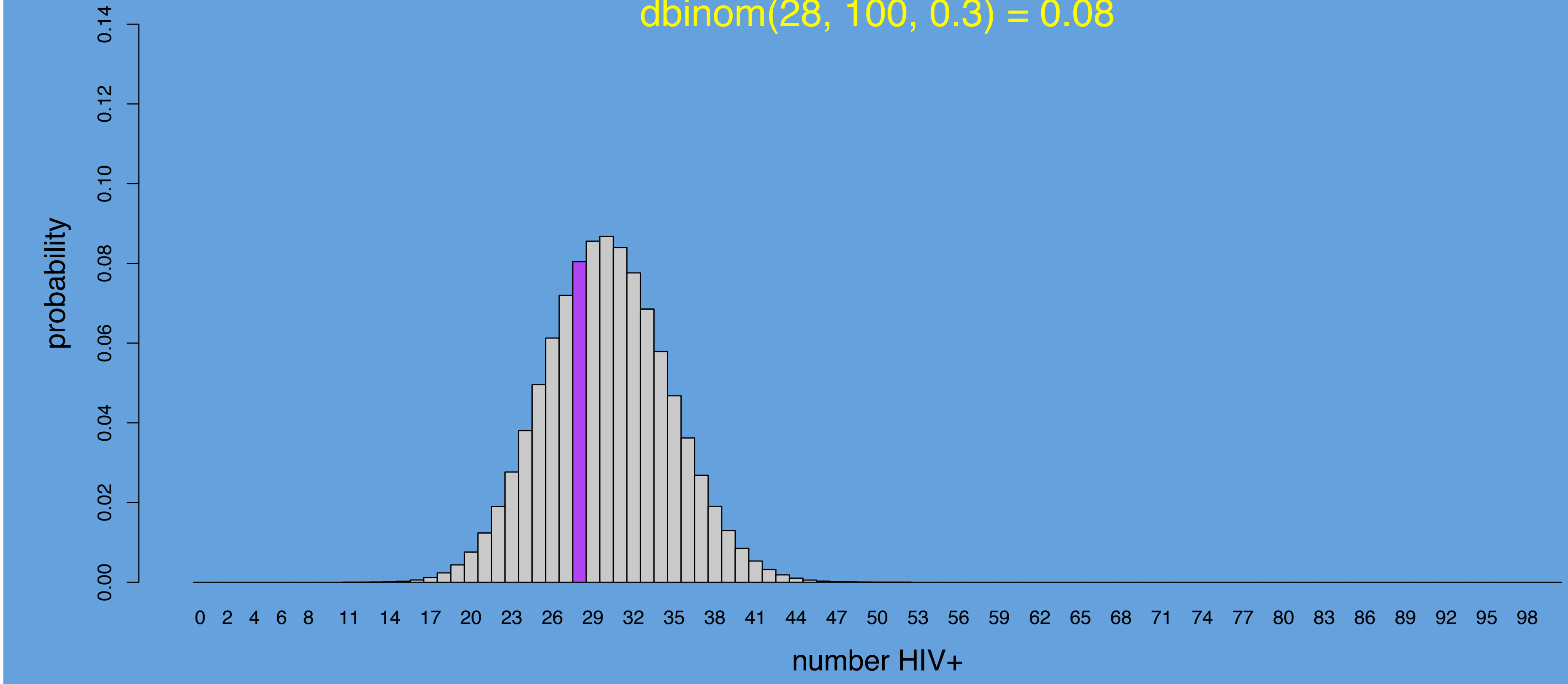
hypothetical prevalence: 25 %

dbinom(28, 100, 0.25) = 0.07



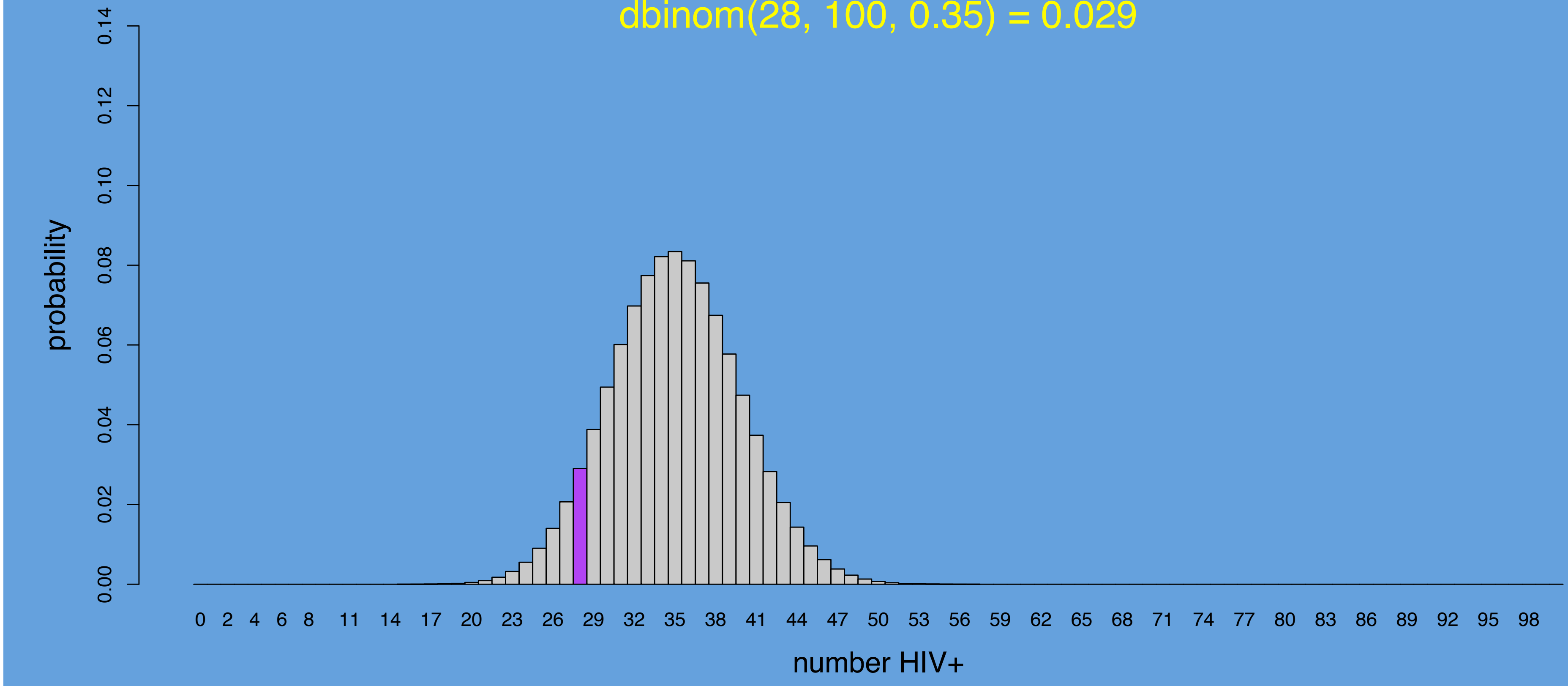
hypothetical prevalence: 30 %

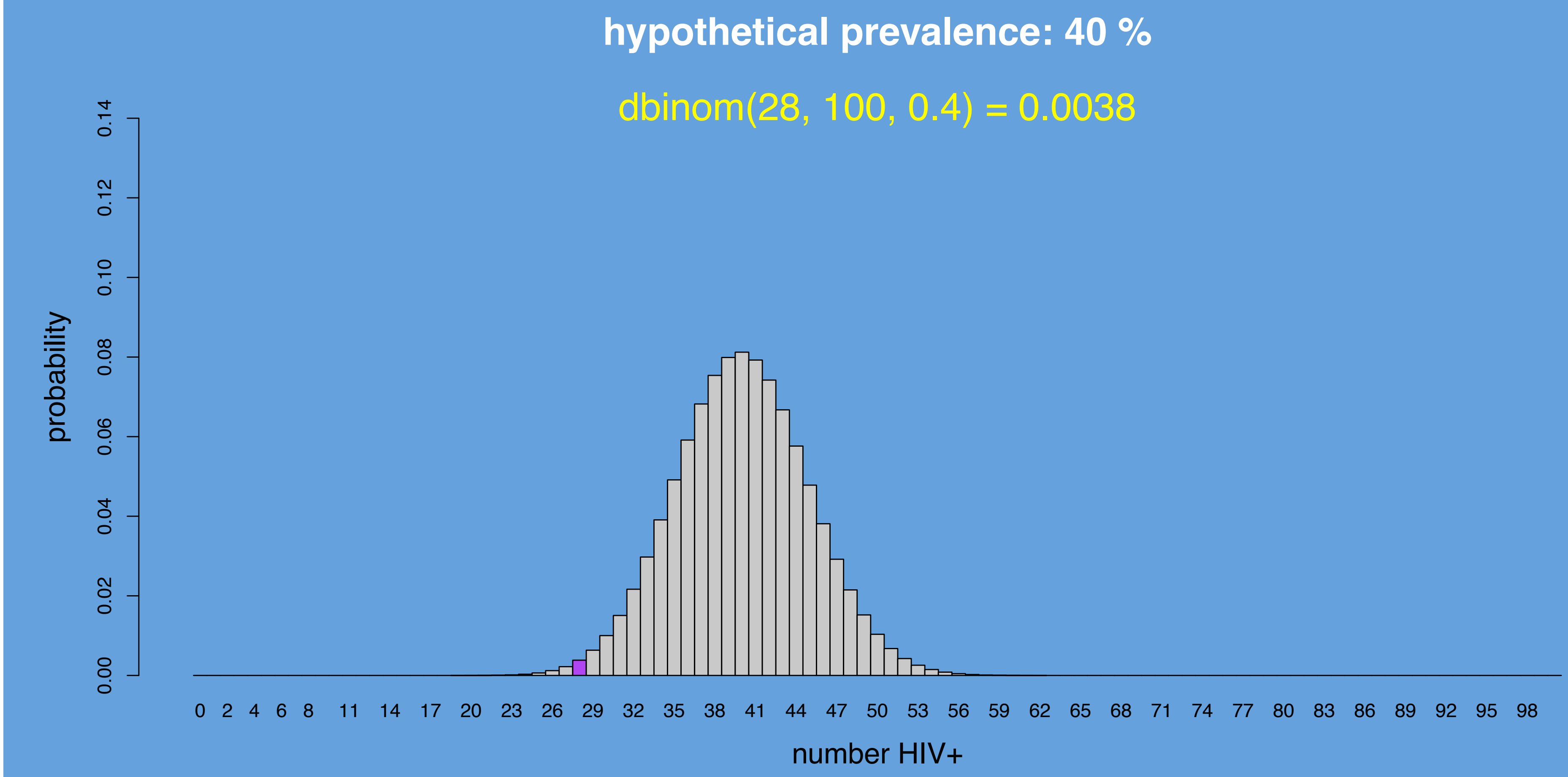
dbinom(28, 100, 0.3) = 0.08



hypothetical prevalence: 35 %

dbinom(28, 100, 0.35) = 0.029

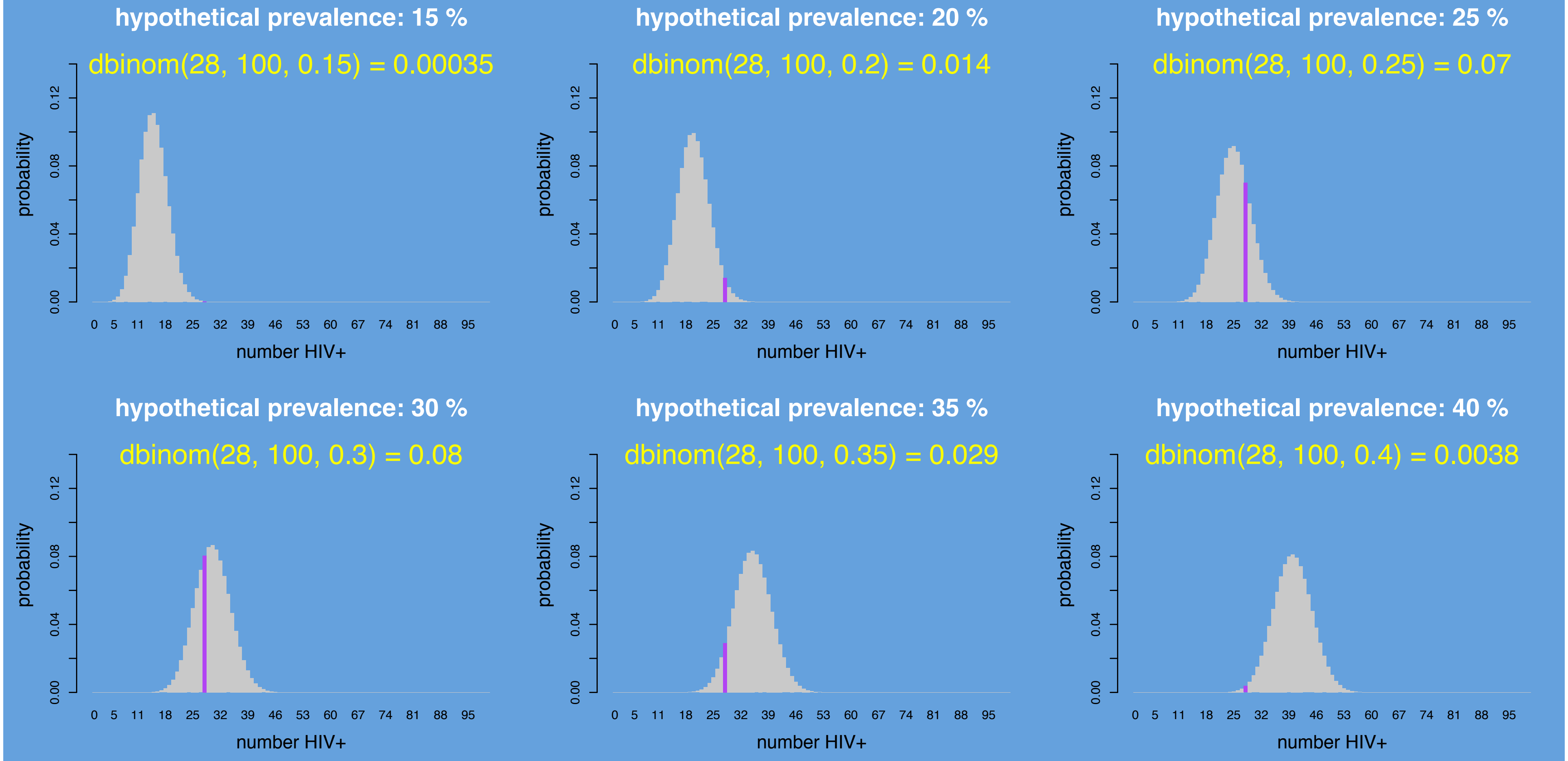




hypothetical prevalence: 40 %

dbinom(28, 100, 0.4) = 0.0038

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



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

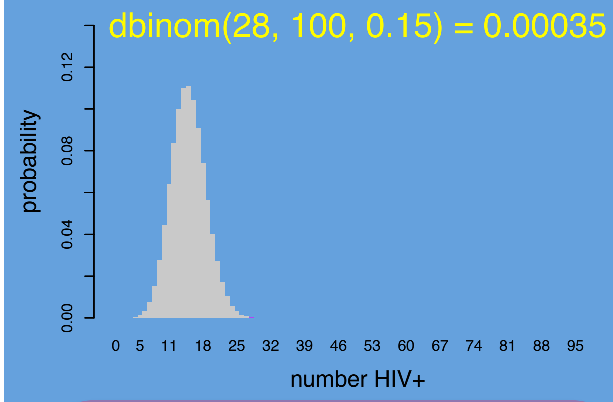
hypothetical prevalence: 40 %

dbinom(28, 100, 0.4) = 0.0038

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

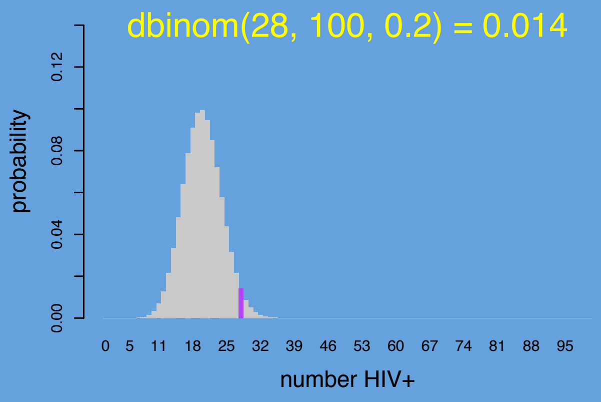
hypothetical prevalence: 15 %

dbinom(28, 100, 0.15) = 0.00035



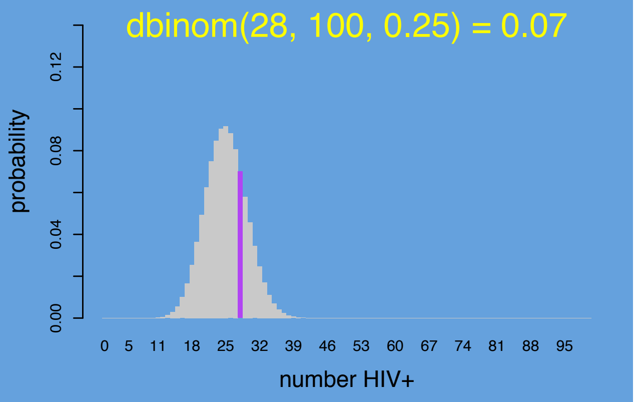
hypothetical prevalence: 20 %

dbinom(28, 100, 0.2) = 0.014



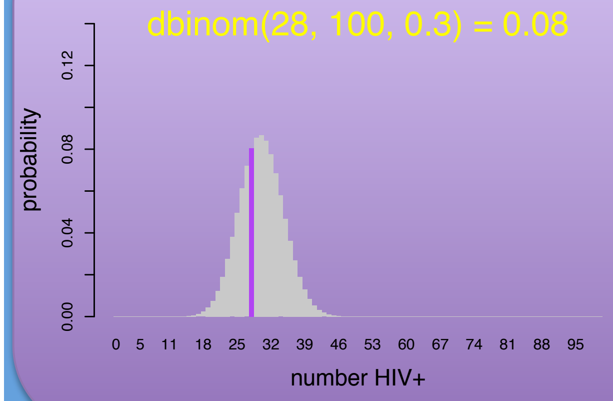
hypothetical prevalence: 25 %

dbinom(28, 100, 0.25) = 0.07



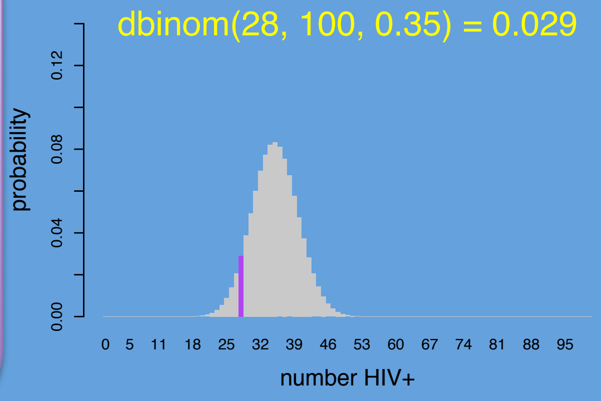
hypothetical prevalence: 30 %

dbinom(28, 100, 0.3) = 0.08



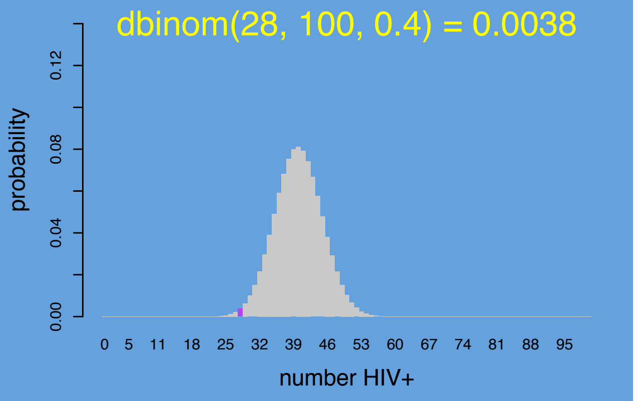
hypothetical prevalence: 35 %

dbinom(28, 100, 0.35) = 0.029

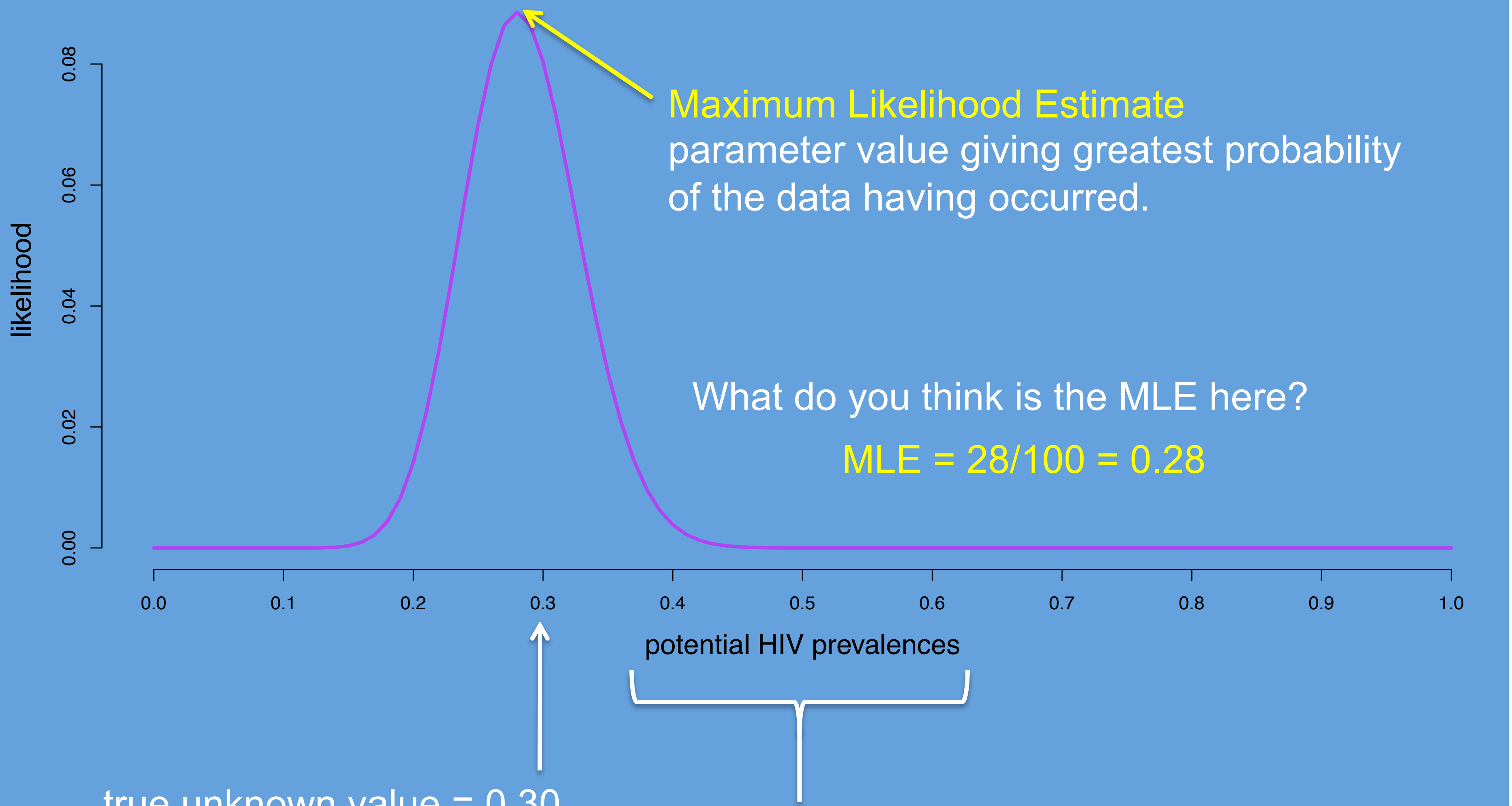


hypothetical prevalence: 40 %

dbinom(28, 100, 0.4) = 0.0038



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

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

$$\downarrow$$

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

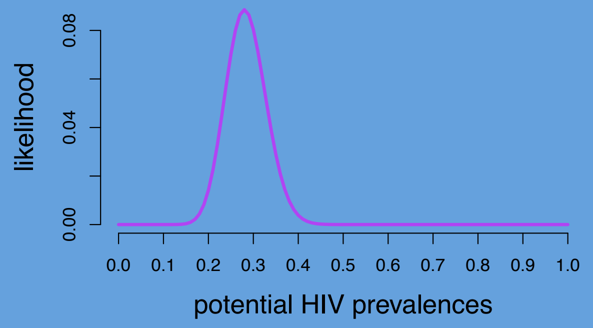
LIKELIHOOD: $L(p \mid x) = \binom{n}{x} p^{x} (1-p)^{n-x}$

$$\uparrow$$

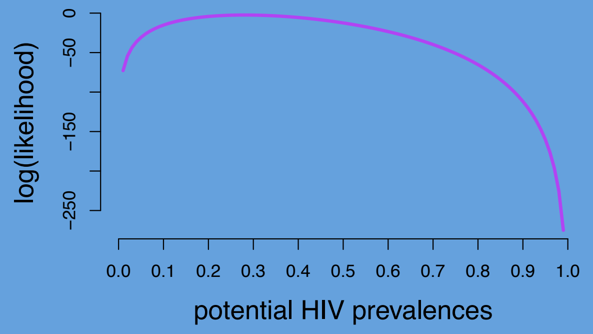
function of p

Deriving the Maximum Likelihood Estimate

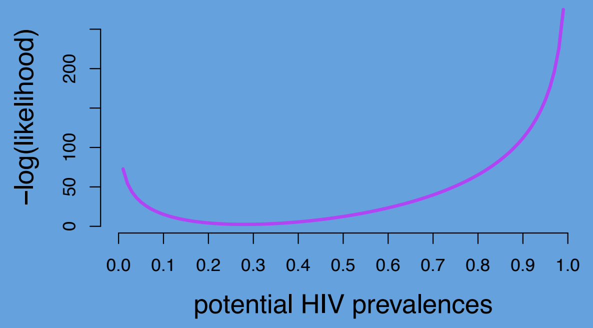
$$\overbrace{L(p)}^{\text{maximize}} = \binom{n}{x} p^x (1-p)^{n-x}$$



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



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



Deriving the Maximum Likelihood Estimate

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

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

$$l(p) = -\log\binom{n}{x} - x\log(p) - (n-x)\log(1-p)$$

Deriving the Maximum Likelihood Estimate

$$l(p) = -\log\begin{pmatrix} n \\ x \end{pmatrix} - x\log(p) - (n-x)\log(1-p)$$

$$\frac{dl(p)}{dp} = -\frac{x}{p} - \frac{-(n-x)}{1-p}$$

$$0 = -\frac{x}{\hat{p}} + \frac{n-x}{1-\hat{p}}$$

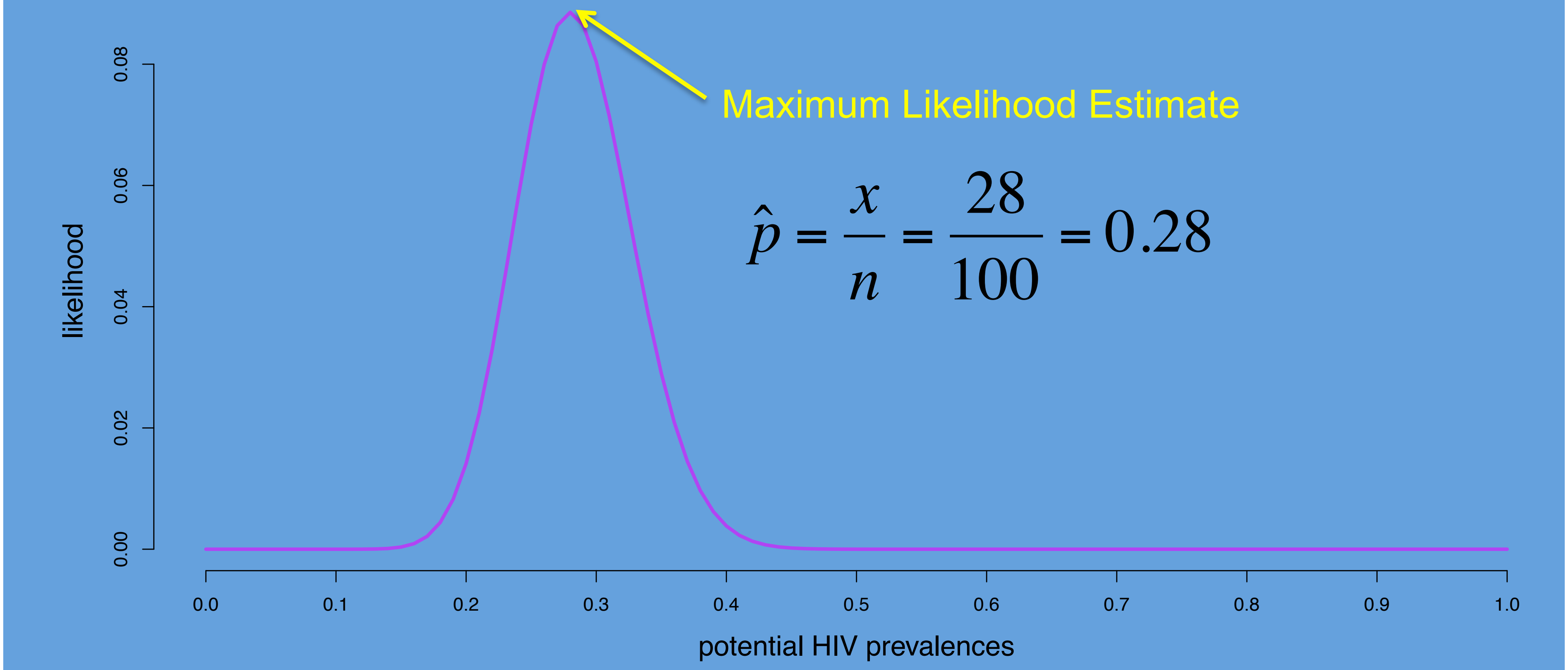
$$0 = \frac{-x(1-\hat{p}) + \hat{p}(n-x)}{\hat{p}(1-\hat{p})}$$

$$0 = -x + \cancel{\hat{p}x} + \hat{p}n - \cancel{\hat{p}x}$$

$$0 = -x + \hat{p}n$$

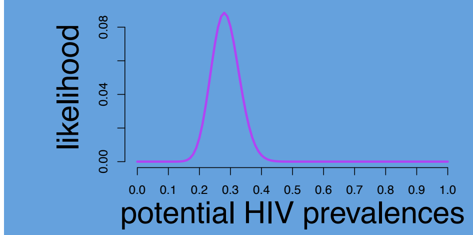
$$\hat{p} = \frac{x}{n}$$

The proportion of positives!

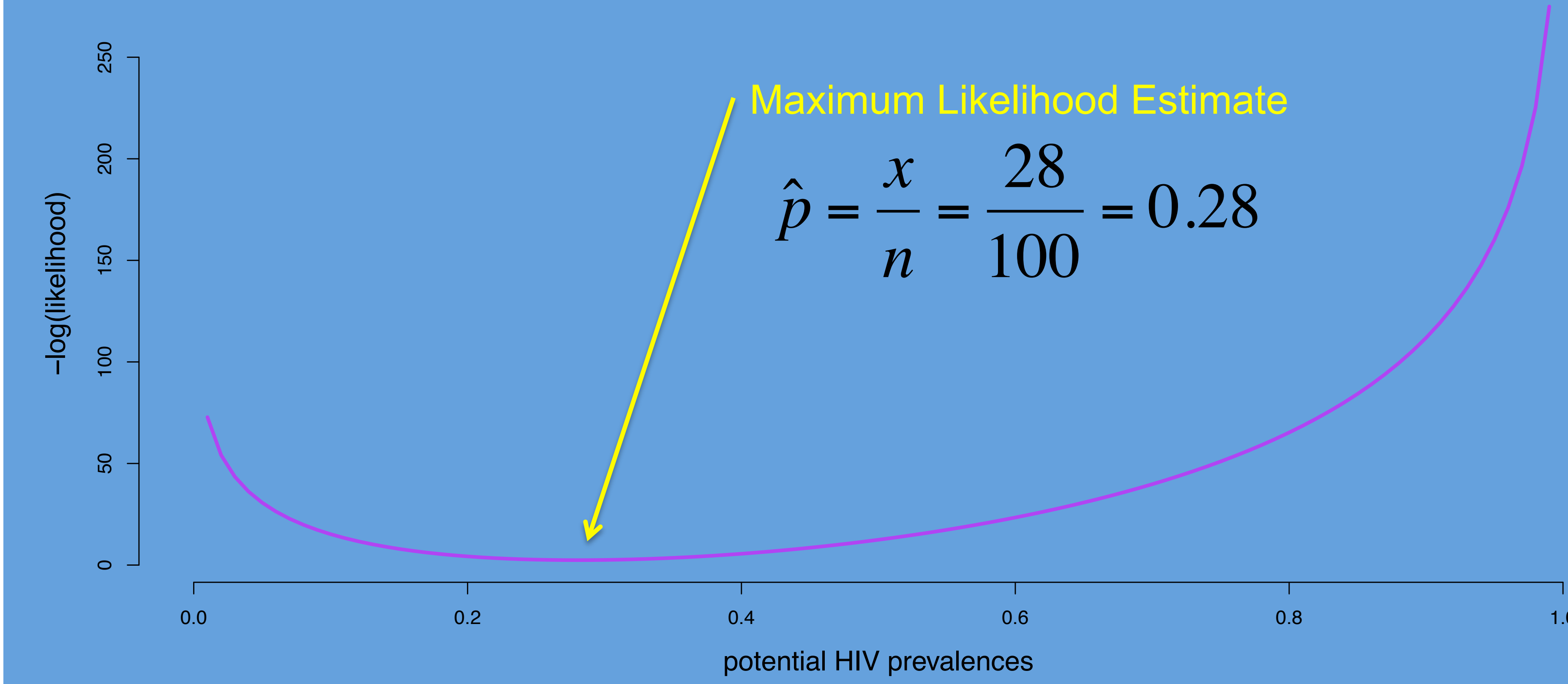


Maximum Likelihood Estimate

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



we usually minimize the –log(likelihood)



Maximum Likelihood Estimate

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

Building Confidence Intervals

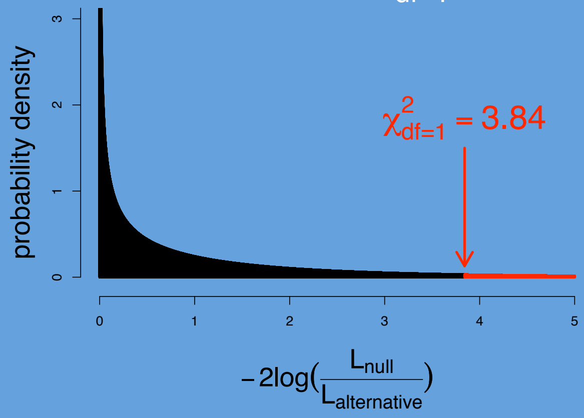
Likelihood Ratio Test

If the null hypothesis were true then

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

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

PDF for $\chi^2_{df=1}$



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
```

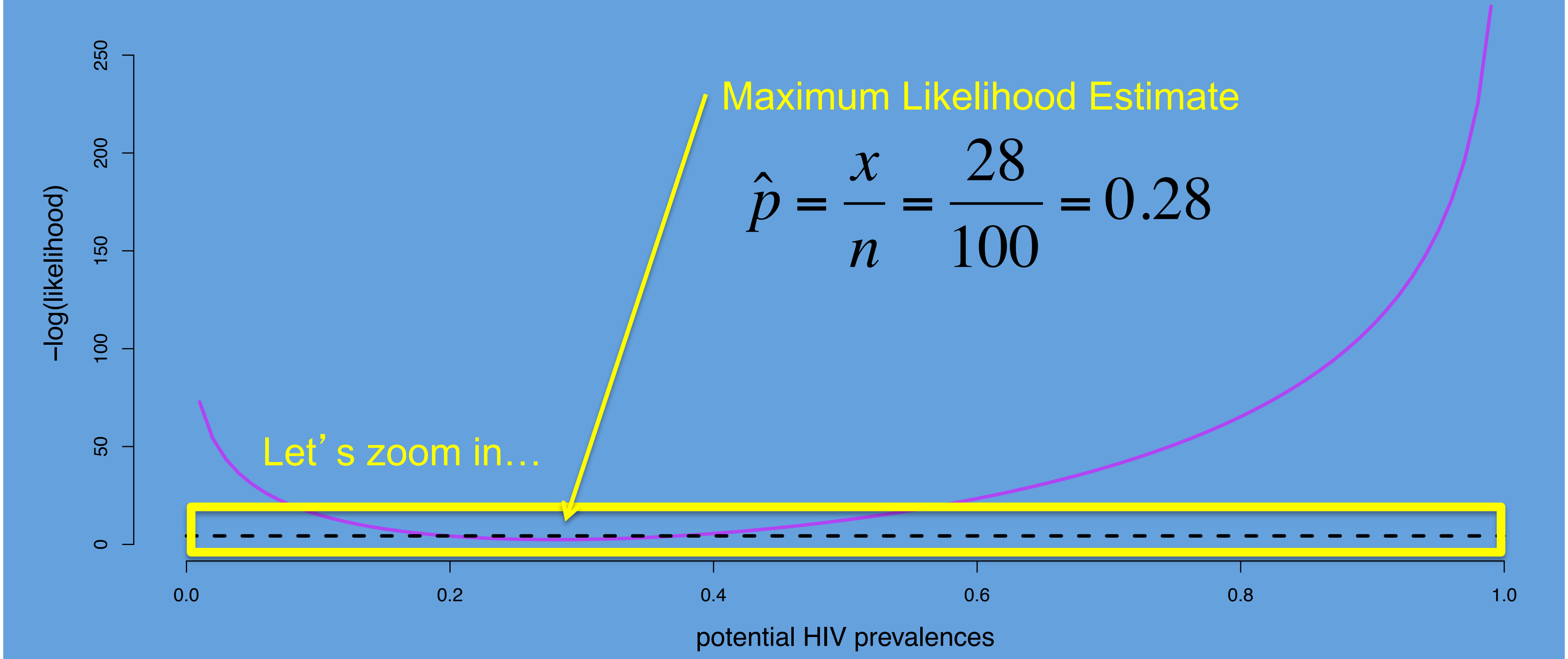
$$2l_{MLE} - 2l_{null} > 3.84$$

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

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

Building Confidence Intervals

Likelihood Ratio Test



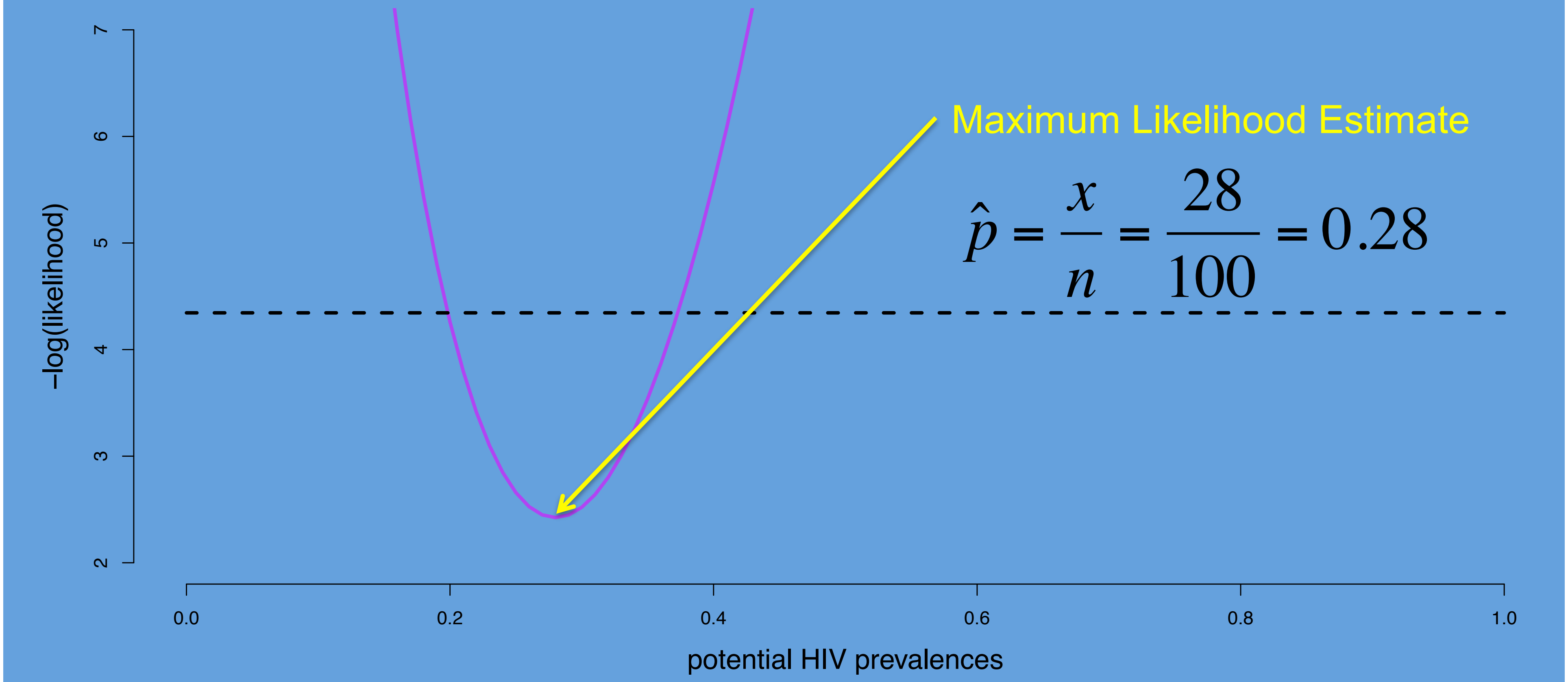
Maximum Likelihood Estimate

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

Let's zoom in…

Building Confidence Intervals

Likelihood Ratio Test

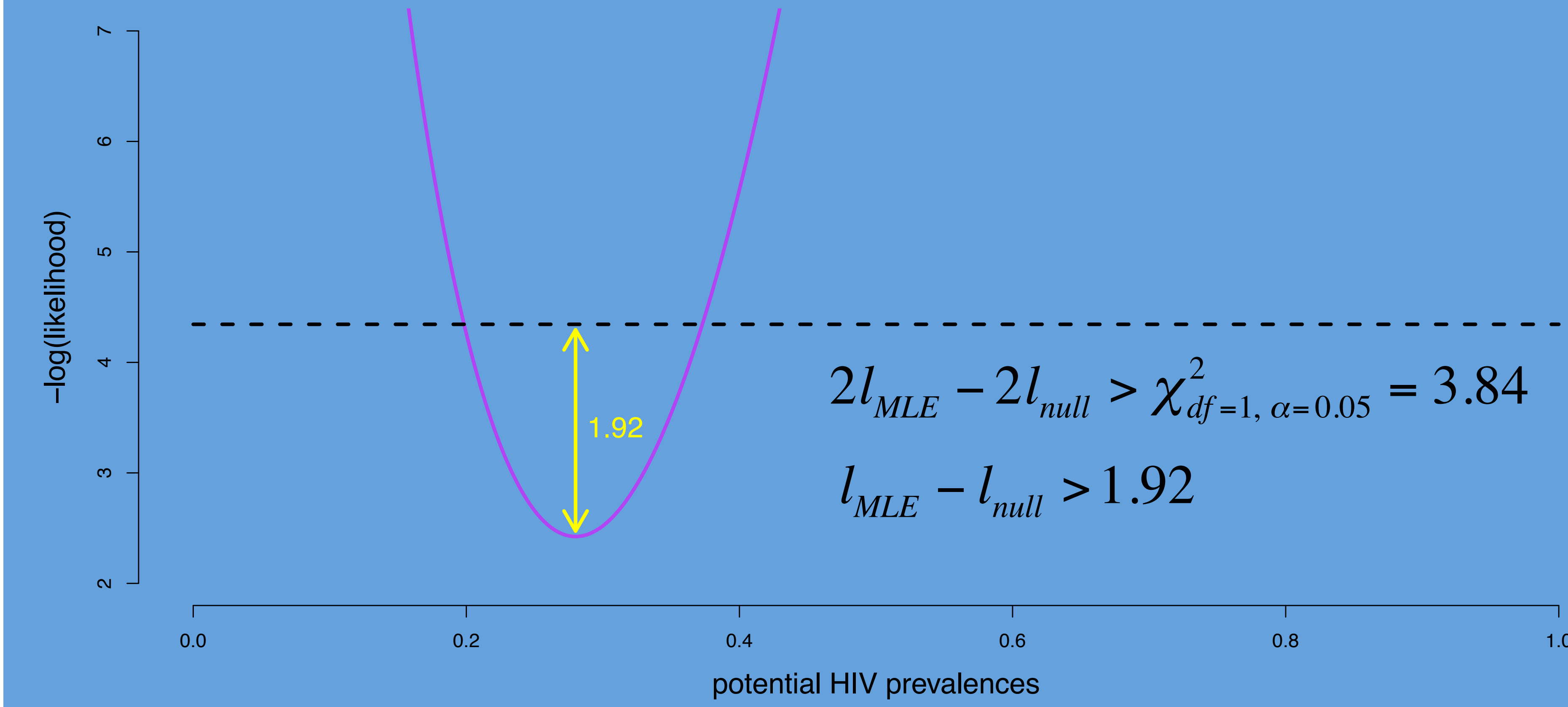


Maximum Likelihood Estimate

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

Building Confidence Intervals

Likelihood Ratio Test

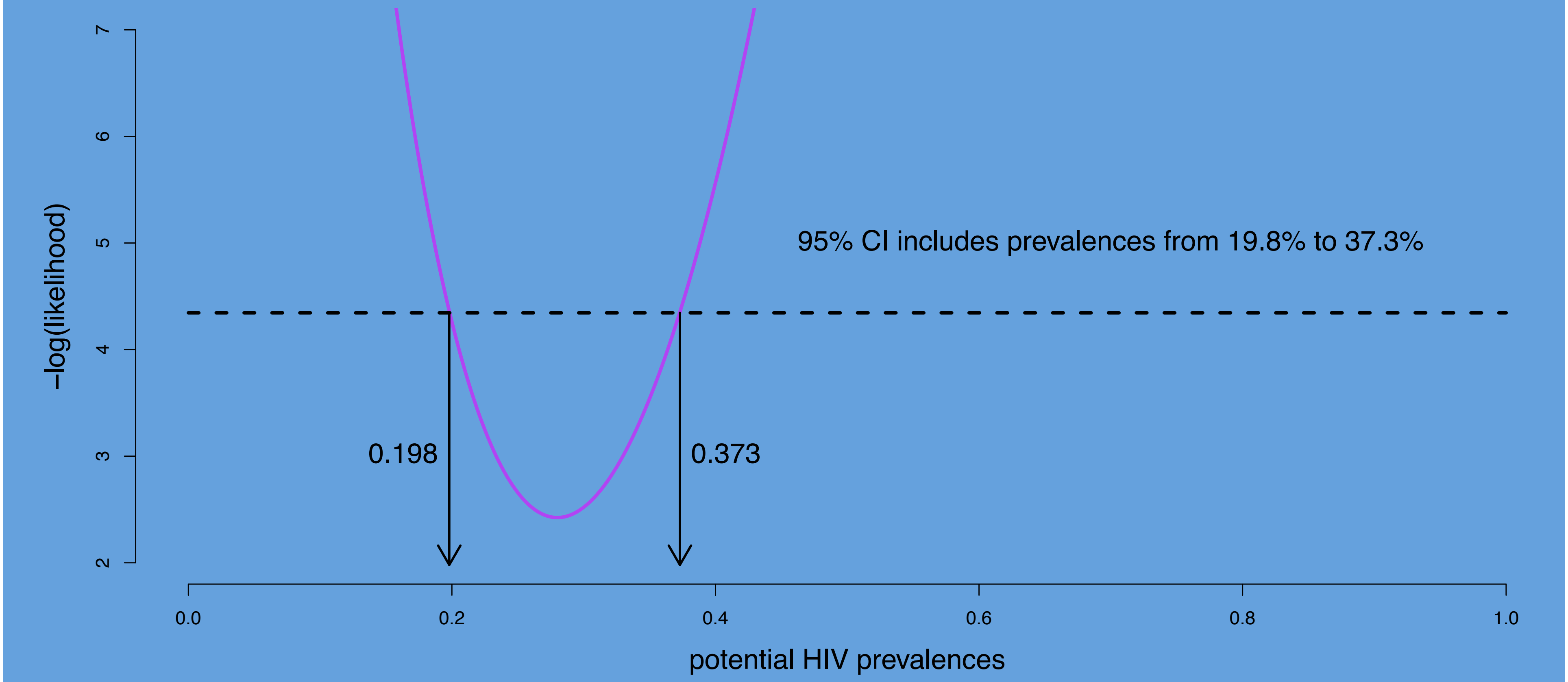


$$2l_{MLE} - 2l_{null} > \chi^2_{df=1,\ \alpha=0.05} = 3.84$$

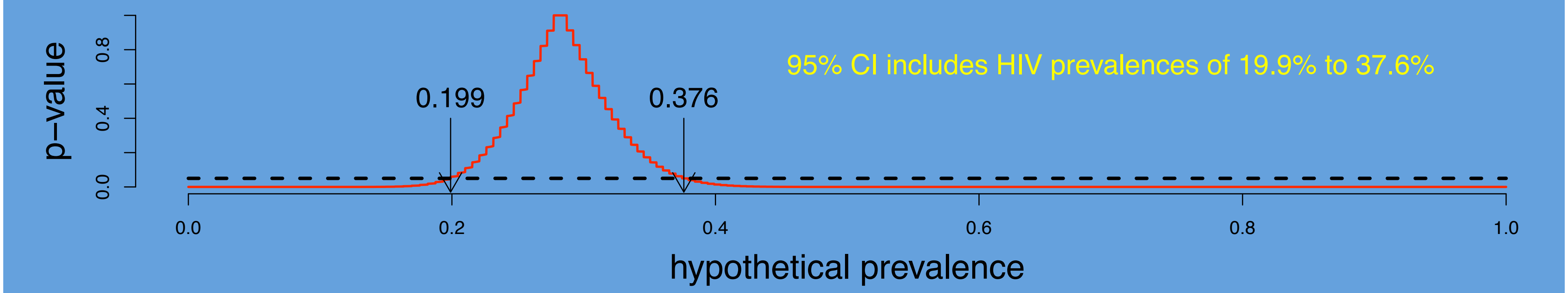
$$l_{MLE} - l_{null} > 1.92$$

Building Confidence Intervals

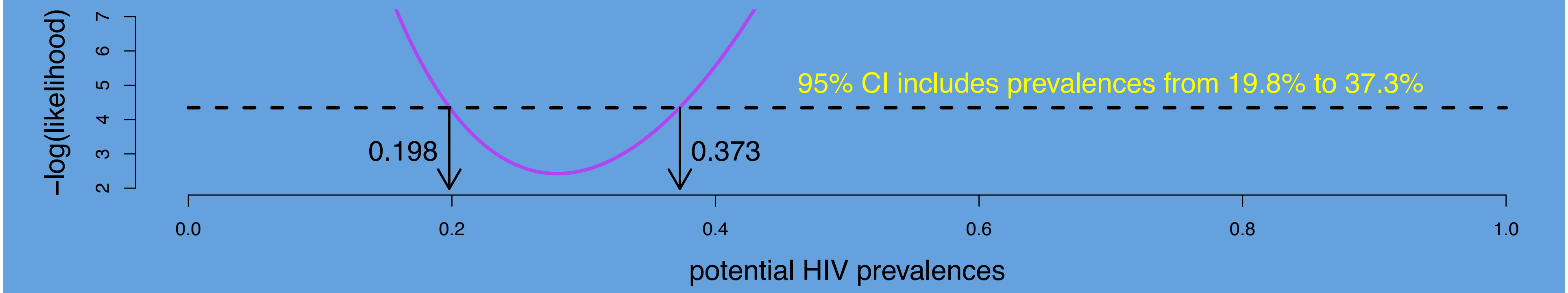
Likelihood Ratio Test



Comparing Confidence Intervals



95% CI includes HIV prevalences of 19.9% to 37.6%



95% CI includes prevalences from 19.8% to 37.3%

Advantages of Likelihood

- Practical method for
  estimating parameters
  estimating variance of our estimates

- Easily adaptable to different probability distributions & dynamic models