

CSE 312

Spring 2015

Maximum Likelihood Estimators
and the EM algorithm

Outline

MLE: Maximum Likelihood Estimators

EM: the Expectation Maximization Algorithm

Learning From Data: MLE

Maximum Likelihood Estimators

Parameter Estimation

Given: independent samples x_1, x_2, \dots, x_n from a parametric distribution $f(x|\theta)$

Goal: estimate θ .

E.g.: Given sample HHTTTTTHTTTTHH of (possibly biased) coin flips, estimate

$\theta =$ probability of Heads

$f(x|\theta)$ is the Bernoulli probability mass function with parameter θ

Likelihood

$P(x | \theta)$: Probability of event x given *model* θ

Viewed as a function of x (fixed θ), it's a *probability*

$$\text{E.g., } \sum_x P(x | \theta) = 1$$

Viewed as a function of θ (fixed x), it's called *likelihood*

E.g., $\sum_{\theta} P(x | \theta)$ can be anything; *relative* values of interest.

E.g., if θ = prob of heads in a sequence of coin flips then

$$P(\text{HHTHH} | .6) > P(\text{HHTHH} | .5),$$

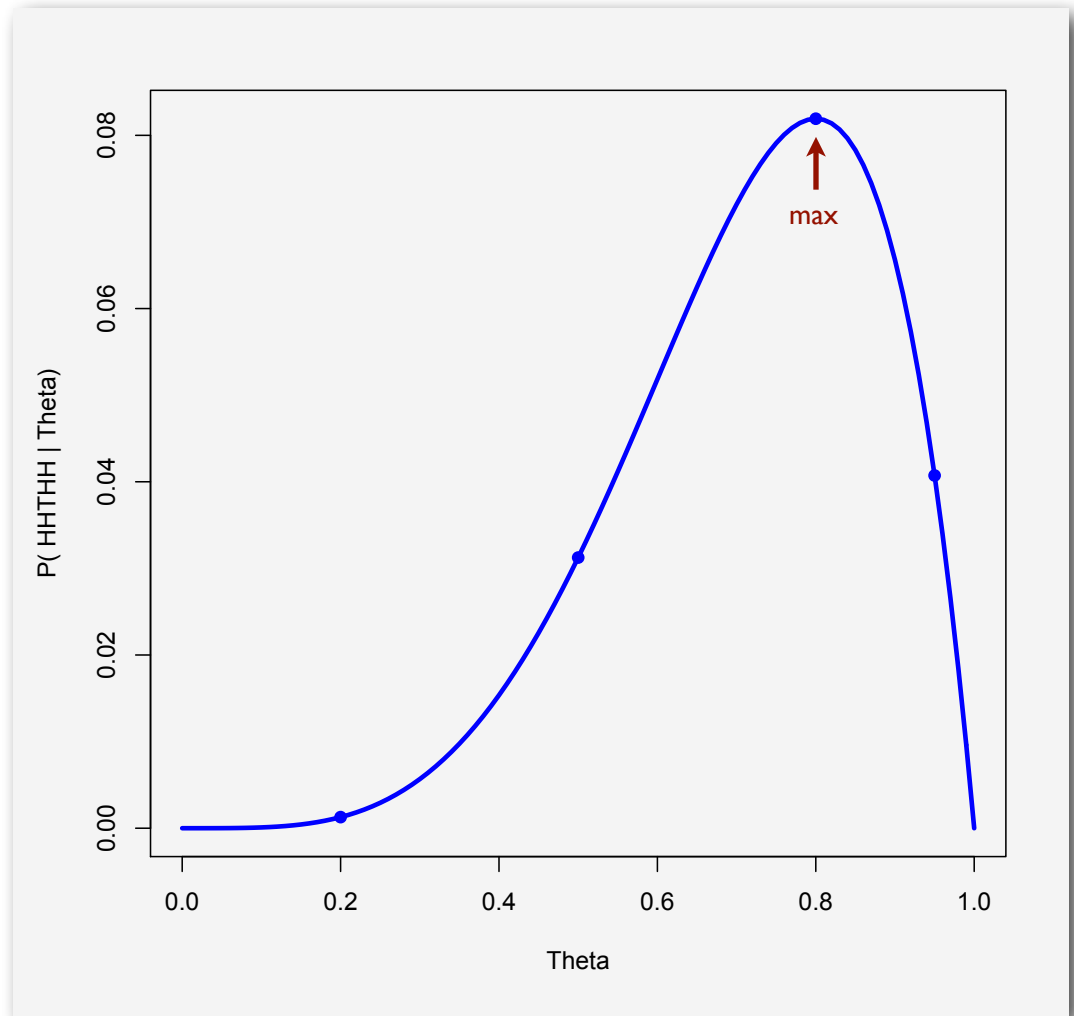
I.e., event HHTHH is *more likely* when $\theta = .6$ than $\theta = .5$

And **what θ make HHTHH *most likely*?**

Likelihood Function

$P(\text{HHTHH} \mid \theta)$:
Probability of HHTHH,
given $P(H) = \theta$:

θ	$\theta^4(1-\theta)$
0.2	0.0013
0.5	0.0313
0.8	0.0819
0.95	0.0407



Maximum Likelihood Parameter Estimation

One (of many) approaches to param. est.

Likelihood of (indp) observations x_1, x_2, \dots, x_n

$$L(x_1, x_2, \dots, x_n \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$$

As a function of θ , what θ maximizes the likelihood of the data actually observed

Typical approach: $\frac{\partial}{\partial \theta} L(\vec{x} \mid \theta) = 0$ or $\frac{\partial}{\partial \theta} \log L(\vec{x} \mid \theta) = 0$

Example 1

n independent coin flips, x_1, x_2, \dots, x_n ; n_0 tails, n_1 heads,
 $n_0 + n_1 = n$; $\theta =$ probability of heads

$$L(x_1, x_2, \dots, x_n \mid \theta) = (1 - \theta)^{n_0} \theta^{n_1}$$

$$\log L(x_1, x_2, \dots, x_n \mid \theta) = n_0 \log(1 - \theta) + n_1 \log \theta$$

$$\frac{\partial}{\partial \theta} \log L(x_1, x_2, \dots, x_n \mid \theta) = \frac{-n_0}{1 - \theta} + \frac{n_1}{\theta}$$

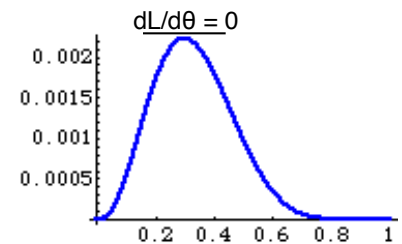
Setting to zero and solving:

$$\hat{\theta} = \frac{n_1}{n}$$

Observed fraction of
successes in *sample* is
MLE of success
probability in *population*

(Also verify it's max, not min, & not better on boundary)

NB: “ n choose n_1 ” term unneeded since outcome sequence is known, but even if unknown, it would drop out at the $d/d\theta$ step



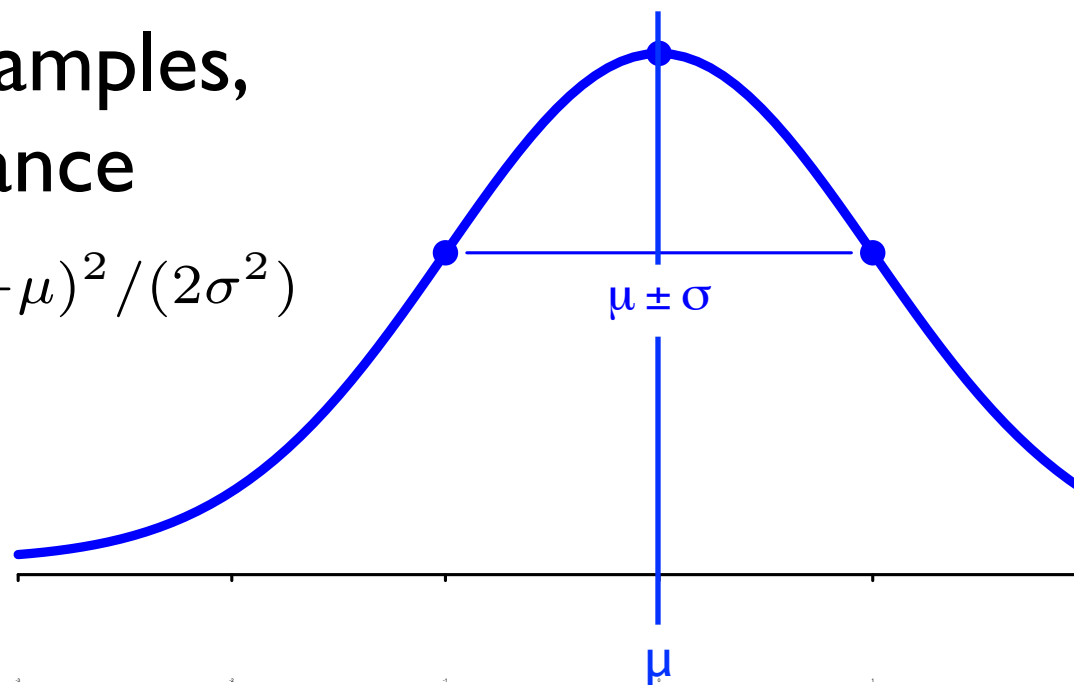
Parameter Estimation

Given: indep samples x_1, x_2, \dots, x_n from a parametric distribution $f(x|\theta)$, **estimate:** θ .

E.g.: Given n normal samples, estimate mean & variance

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$\theta = (\mu, \sigma^2)$$



Ex2: I got data; a little birdie tells me
it's normal, and promises $\sigma^2 = 1$

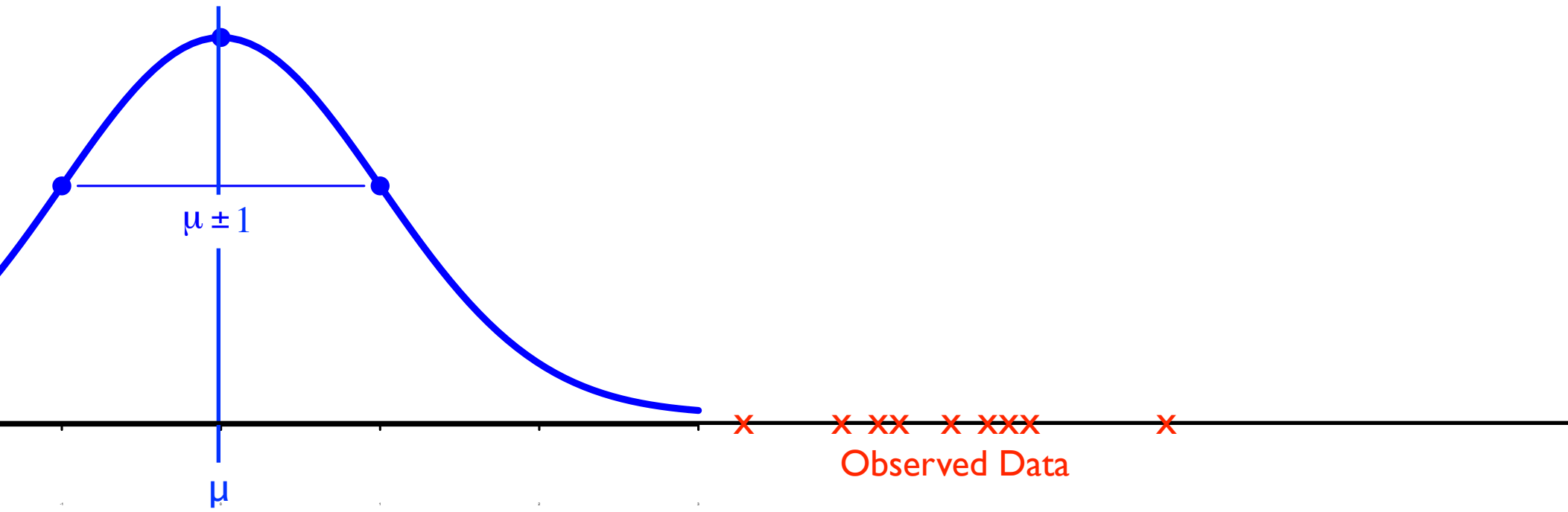


Observed Data

$x \rightarrow$

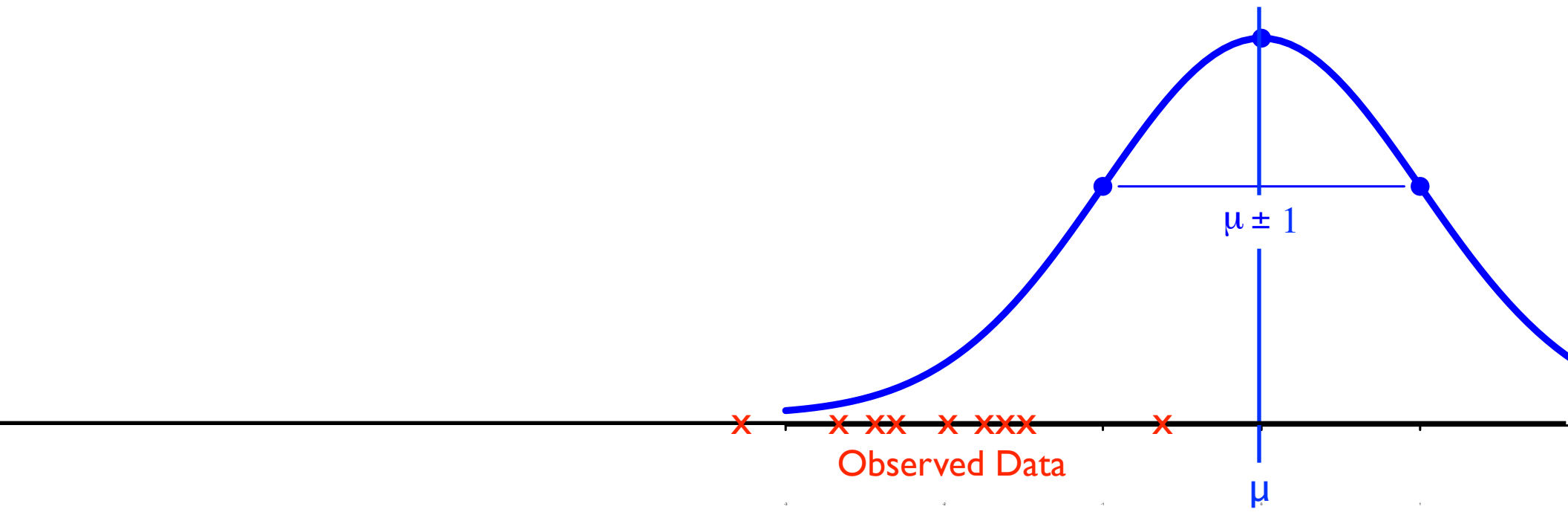
Which is more likely: (a) this?

μ unknown, $\sigma^2 = 1$



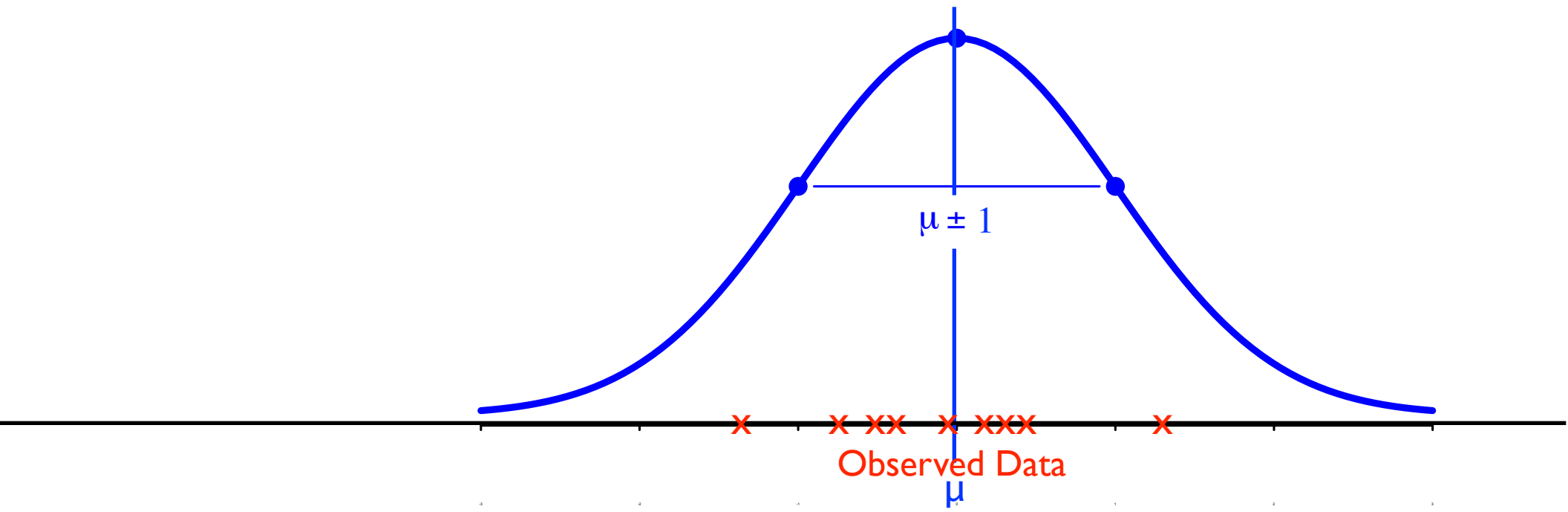
Which is more likely: (b) or this?

μ unknown, $\sigma^2 = 1$



Which is more likely: (c) or *this*?

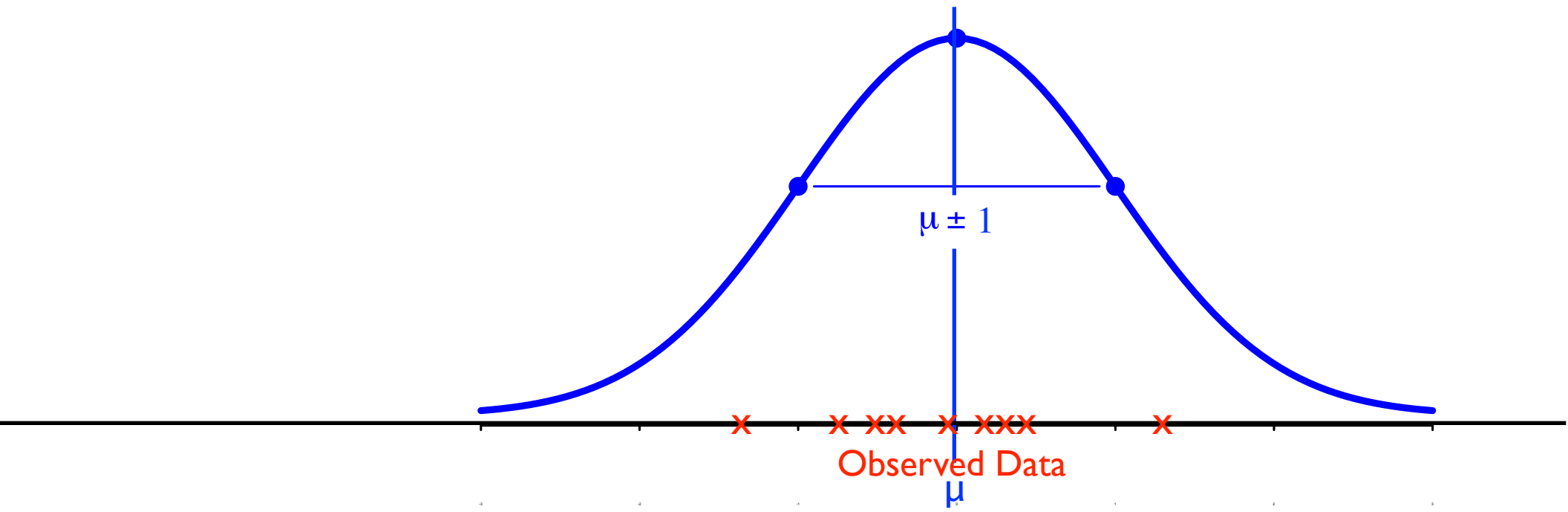
μ unknown, $\sigma^2 = 1$



Which is more likely: (c) or this?

μ unknown, $\sigma^2 = 1$

Looks good by eye, but how do I optimize my estimate of μ ?



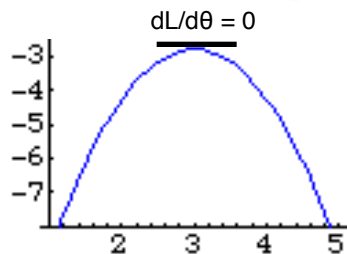
Ex. 2: $x_i \sim N(\mu, \sigma^2)$, $\sigma^2 = 1$, μ unknown

$$L(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i - \theta)^2 / 2}$$

$$\ln L(x_1, x_2, \dots, x_n | \theta) = \sum_{i=1}^n -\frac{1}{2} \ln(2\pi) - \frac{(x_i - \theta)^2}{2}$$

$$\frac{d}{d\theta} \ln L(x_1, x_2, \dots, x_n | \theta) = \sum_{i=1}^n (x_i - \theta)$$

And verify it's max,
not min & not better
on boundary



$$= \left(\sum_{i=1}^n x_i \right) - n\theta = 0$$

$$\hat{\theta} = \left(\sum_{i=1}^n x_i \right) / n = \bar{x}$$

Sample mean is MLE of
population mean

Hmm ..., density \neq probability

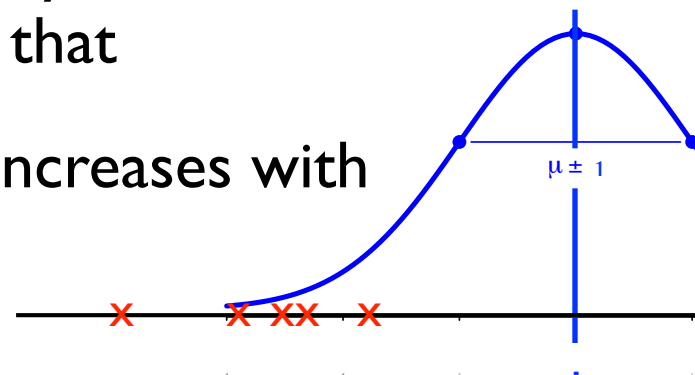
So why is “likelihood” function equal to product of *densities*?? (Prob of seeing any specific x_i is 0, right?)

a) for maximizing likelihood, we really only care about *relative* likelihoods, and density captures that

b) has desired property that likelihood increases with better fit to the model

and/or

c) if density at x is $f(x)$, for any small $\delta > 0$, the probability of a sample within $\pm \delta/2$ of x is $\approx \delta f(x)$, but δ is *constant* wrt θ , so it just drops out of $d/d\theta \log L(\dots) = 0$.



Ex3: I got data; a little birdie tells me it's normal (but does *not* tell me μ, σ^2)

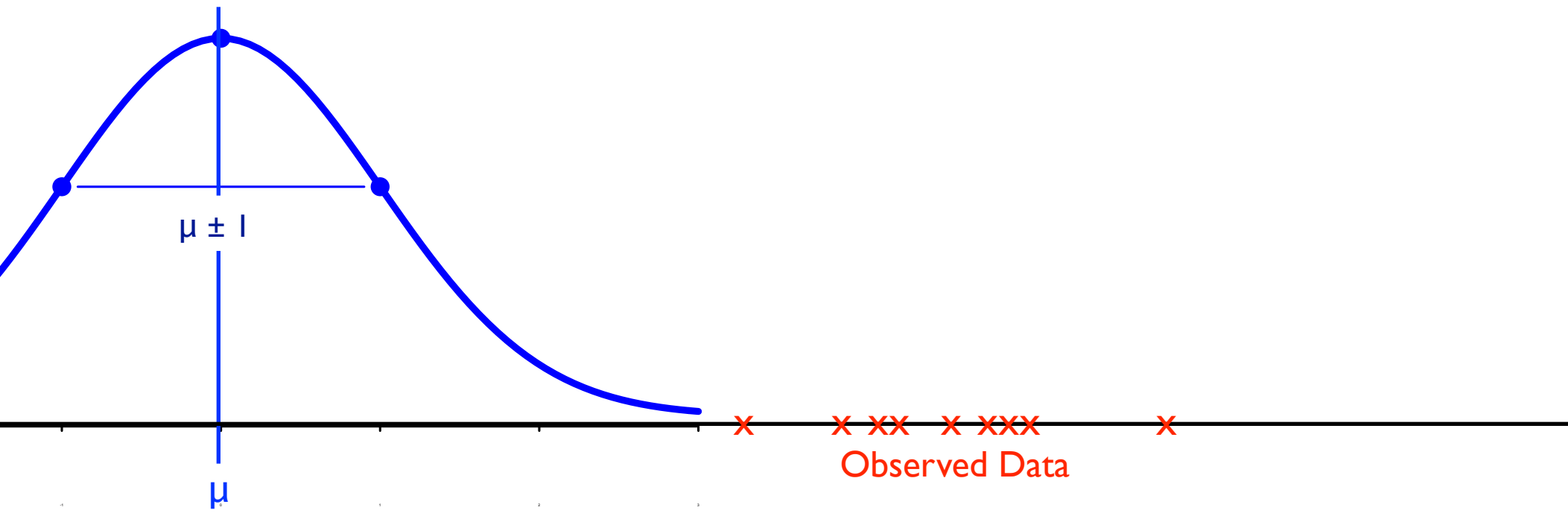


Observed Data

$x \rightarrow$

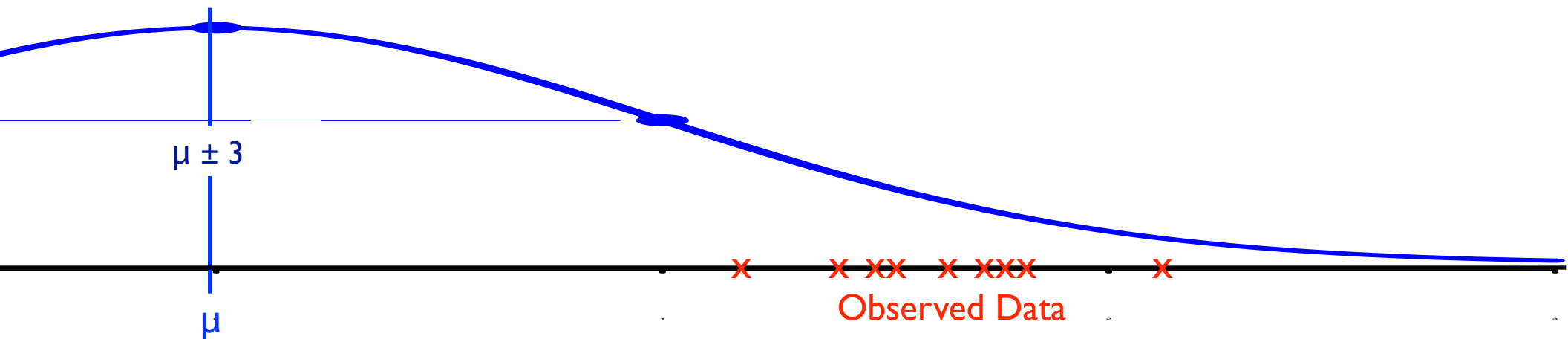
Which is more likely: (a) this?

μ, σ^2 both unknown



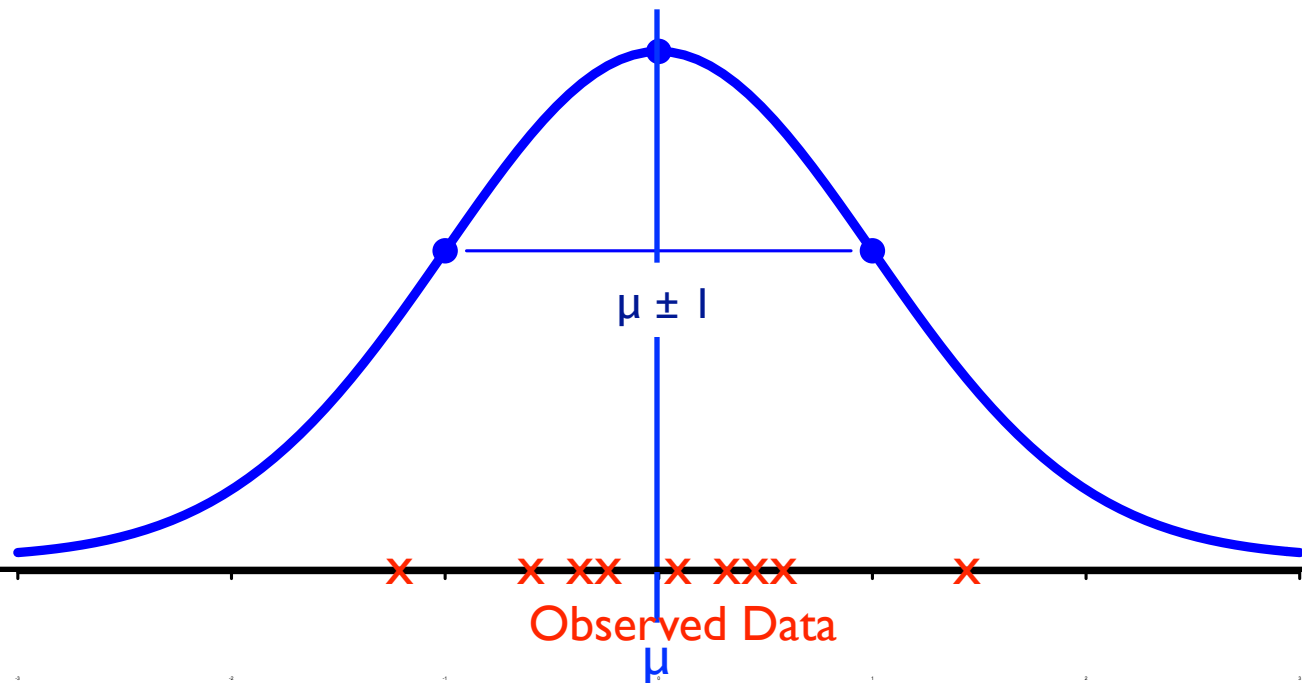
Which is more likely: (b) or this?

μ, σ^2 both unknown



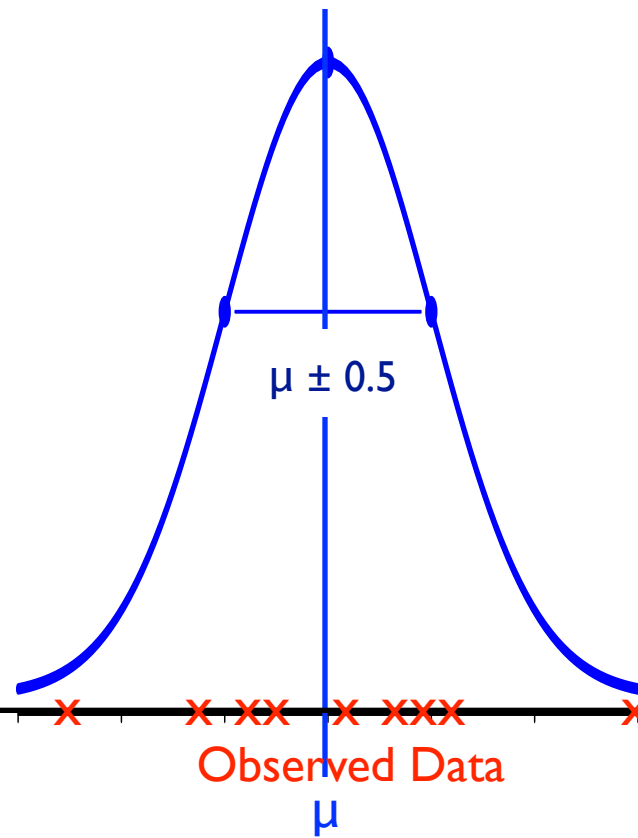
Which is more likely: (c) or this?

μ, σ^2 both unknown



Which is more likely: (d) or *this*?

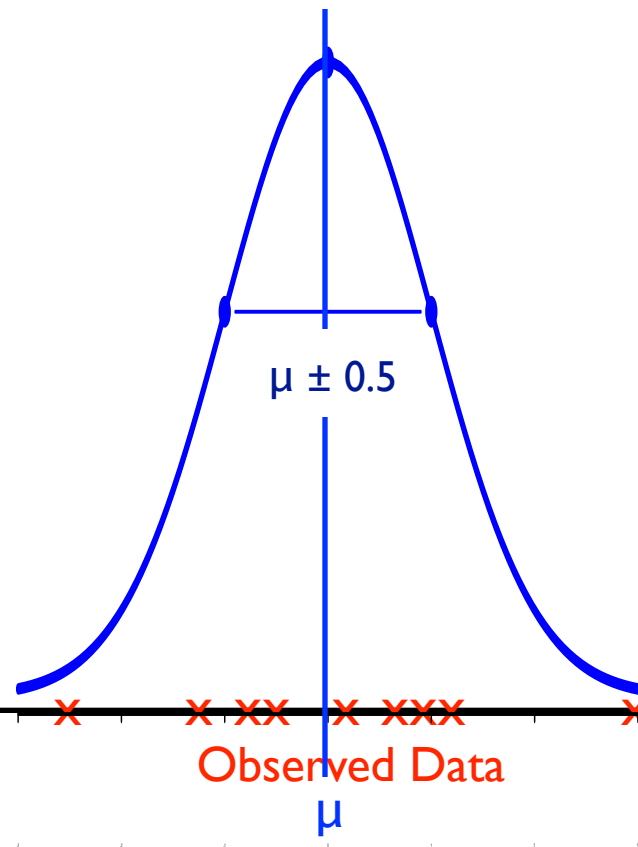
μ, σ^2 both unknown



Which is more likely: (d) or *this*?

μ, σ^2 both unknown

Looks good by eye, but how do I optimize my estimates of μ & $\underline{\underline{\sigma^2}}$?



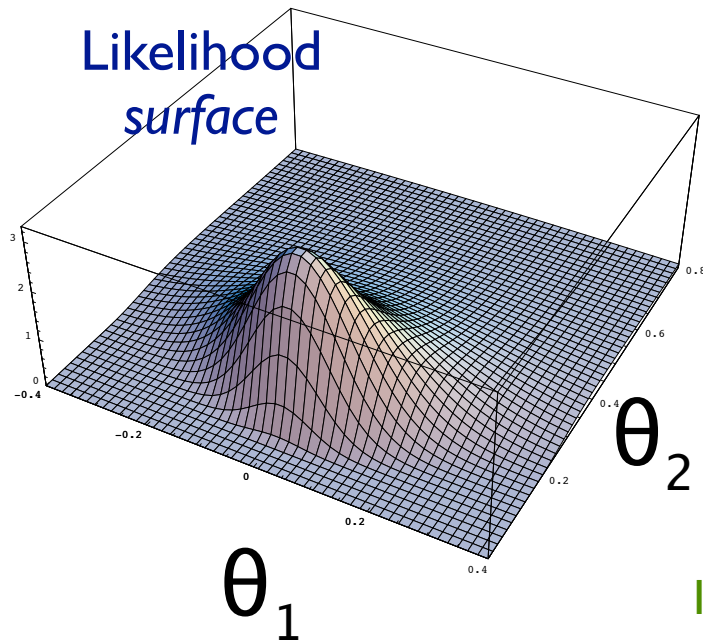
Ex 3: $x_i \sim N(\mu, \sigma^2)$, μ, σ^2 both unknown

$$\ln L(x_1, x_2, \dots, x_n | \theta_1, \theta_2) = \sum_{i=1}^n -\frac{1}{2} \ln(2\pi\theta_2) - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, x_2, \dots, x_n | \theta_1, \theta_2) = \sum_{i=1}^n \frac{(x_i - \theta_1)}{\theta_2} = 0$$

$$\hat{\theta}_1 = \left(\sum_{i=1}^n x_i \right) / n = \bar{x}$$

Likelihood
surface



Sample mean is MLE of
population mean, again

In general, a problem like this results in 2 equations in 2 unknowns.
Easy in this case, since θ_2 drops out of the $\partial/\partial\theta_1 = 0$ equation 23

Ex. 3, (cont.)

$$\ln L(x_1, x_2, \dots, x_n | \theta_1, \theta_2) = \sum_{i=1}^n -\frac{1}{2} \ln(2\pi\theta_2) - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_2} \ln L(x_1, x_2, \dots, x_n | \theta_1, \theta_2) = \sum_{i=1}^n -\frac{1}{2} \frac{2\pi}{2\pi\theta_2} + \frac{(x_i - \theta_1)^2}{2\theta_2^2} = 0$$

$$\hat{\theta}_2 = \left(\sum_{i=1}^n (x_i - \hat{\theta}_1)^2 \right) / n = \bar{s}^2$$

*Sample variance is MLE of
population variance*

Summary

MLE is *one* way to estimate *parameters* from *data*

You choose the *form* of the model (normal, binomial, ...)

Math chooses the *value(s)* of parameter(s)

Defining the “Likelihood Function” (based on the form of the model) is often the critical step; the math/algorithms to optimize it are generic

Often simply $(d/d\theta)(\log \text{Likelihood}) = 0$, check max vs min, boundaries, ...

Has the intuitively appealing property that the parameters maximize the *likelihood* of the observed data; basically just assumes your sample is “representative”

Of course, unusual samples will give bad estimates (estimate normal human heights from a sample of NBA stars?) but that is an unlikely event

Often, but not always, MLE has other desirable properties like being *unbiased*, or at least *consistent*