CSE 312
Autumn 2013
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 $x_1, x_2, \ldots, x_n$ from a parametric distribution $f(x|\theta)$

Goal: estimate θ.

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

θ = probability of Heads

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

Likelihood

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

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

E.g., $\Sigma_x P(x \mid \theta) = 1$

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

E.g., $\Sigma_\theta P(x \mid \theta)$ can be anything; *relative* values of interest.

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

$P(HHTHH \mid .6) > P(HHTHH \mid .5)$,

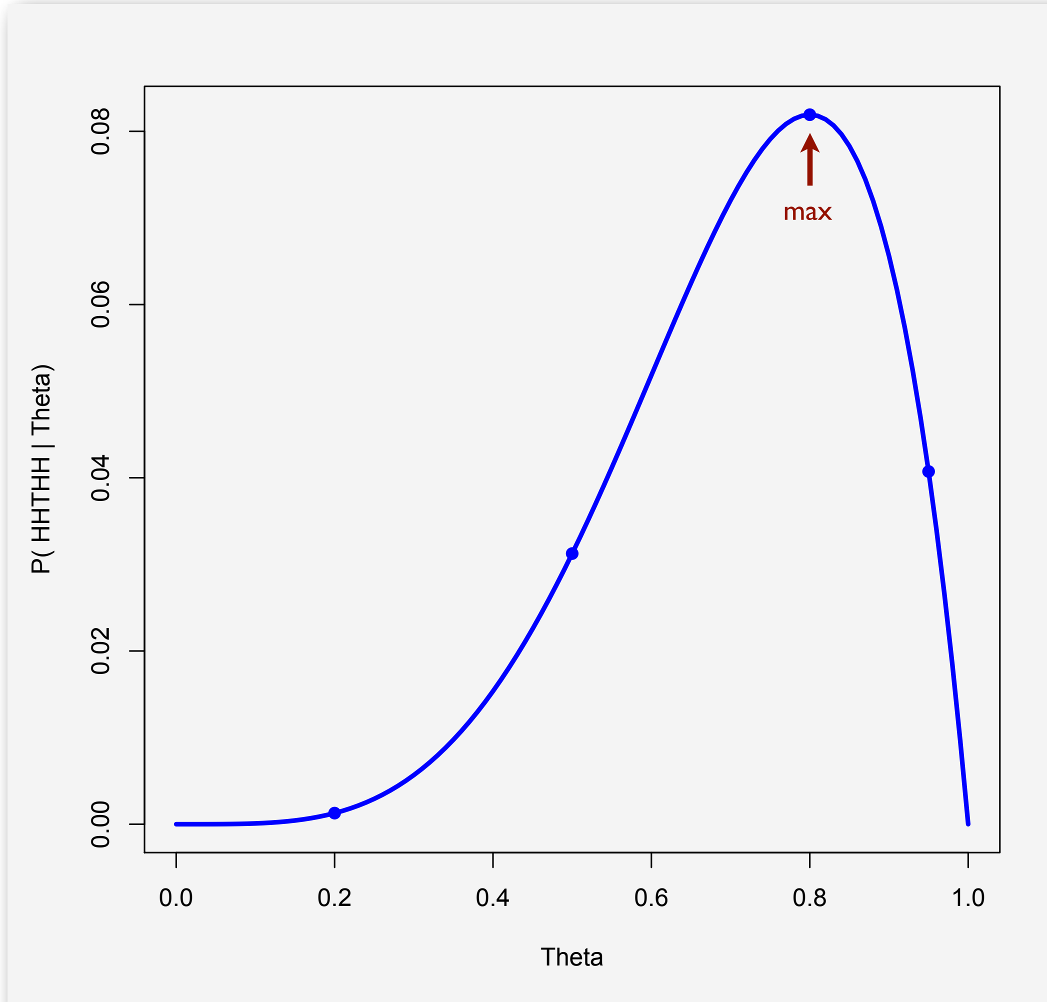
I.e., event HHTHH is *more likely* when $\theta = .6$ than $\theta = .5$

And what θ make HHTHH *most* likely?

Likelihood Function

P(HHTHH | θ):
Probability of HHTHH,
given P(H) = θ:

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 $x_1, x_2, ..., x_n$

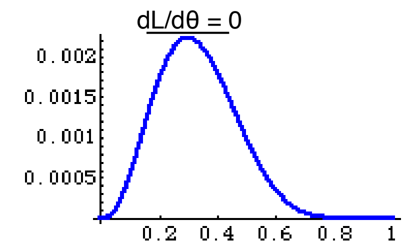
$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

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, $x_1, x_2, ..., x_n$; n_0 tails, n_1 heads, $n_0 + n_1 = n$; θ = probability of heads



$$\begin{aligned} L(x_1, x_2, \ldots, x_n \mid \theta) &= (1-\theta)^{n_0}\theta^{n_1} \\ \log L(x_1, x_2, \ldots, x_n \mid \theta) &= n_0 \log(1-\theta) + n_1 \log\theta \\ \frac{\partial}{\partial\theta} \log L(x_1, x_2, \ldots, x_n \mid \theta) &= \frac{-n_0}{1-\theta} + \frac{n_1}{\theta} \end{aligned}$$

Setting to zero and solving:

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

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

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

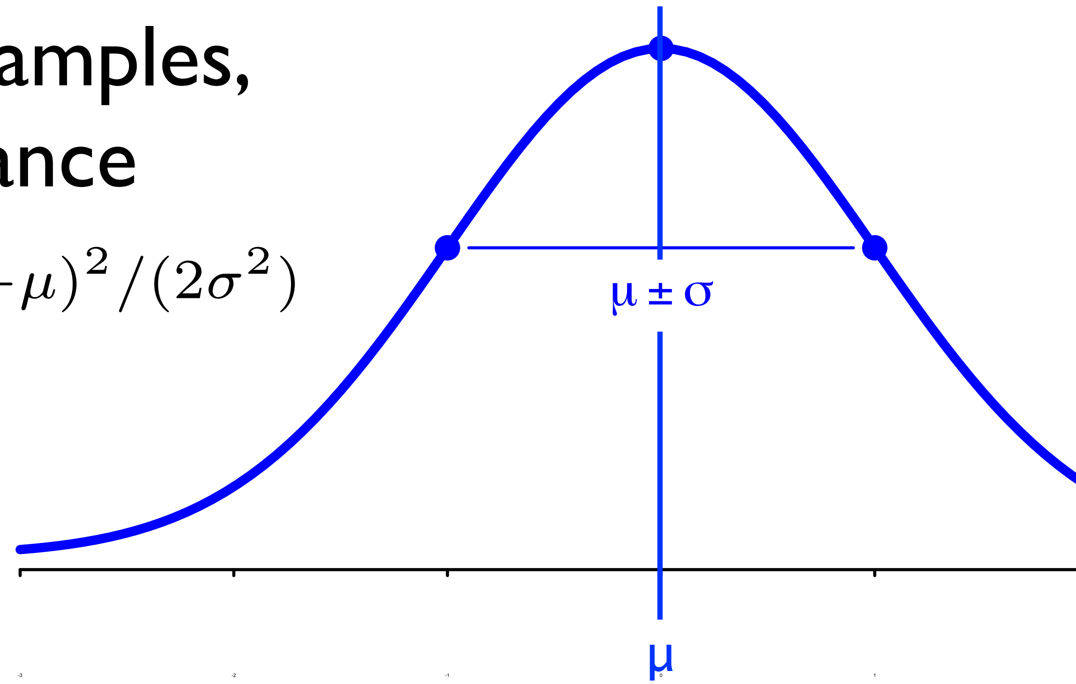
Parameter Estimation

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

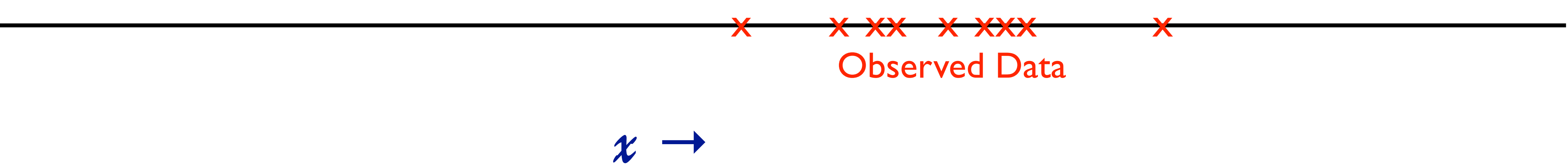
E.g.: Given *n normal* samples, estimate mean & variance

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

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

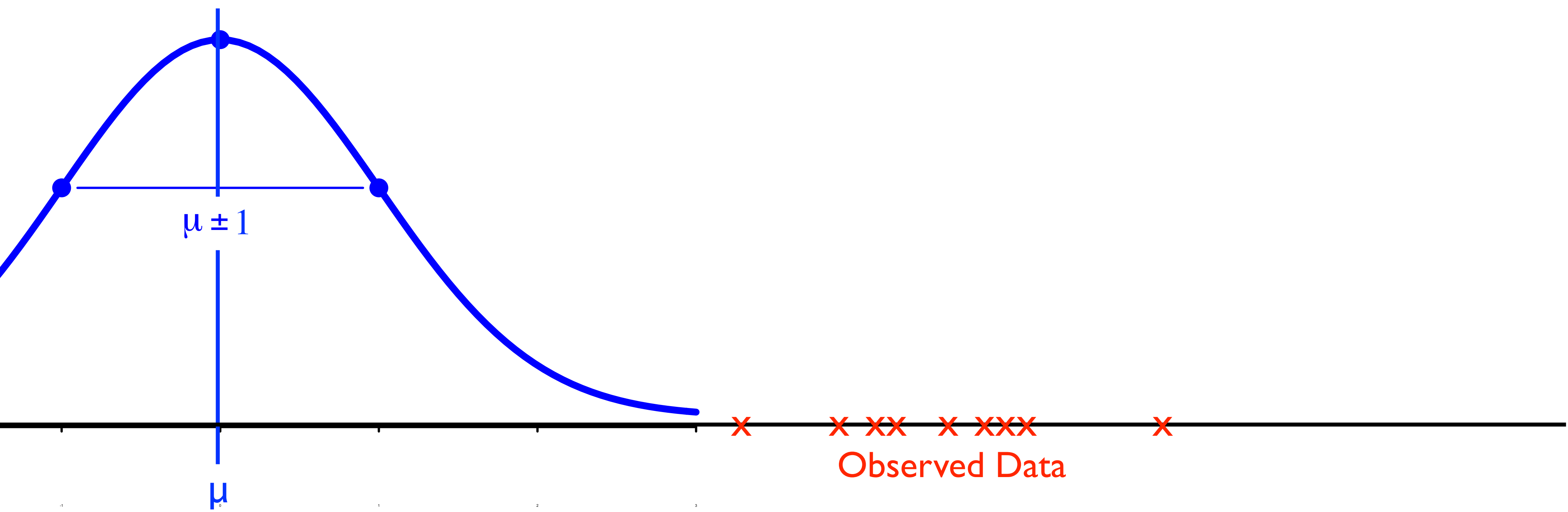


Ex2: I got data; a little birdie tells me it's normal, and promises $\sigma^2 = 1$



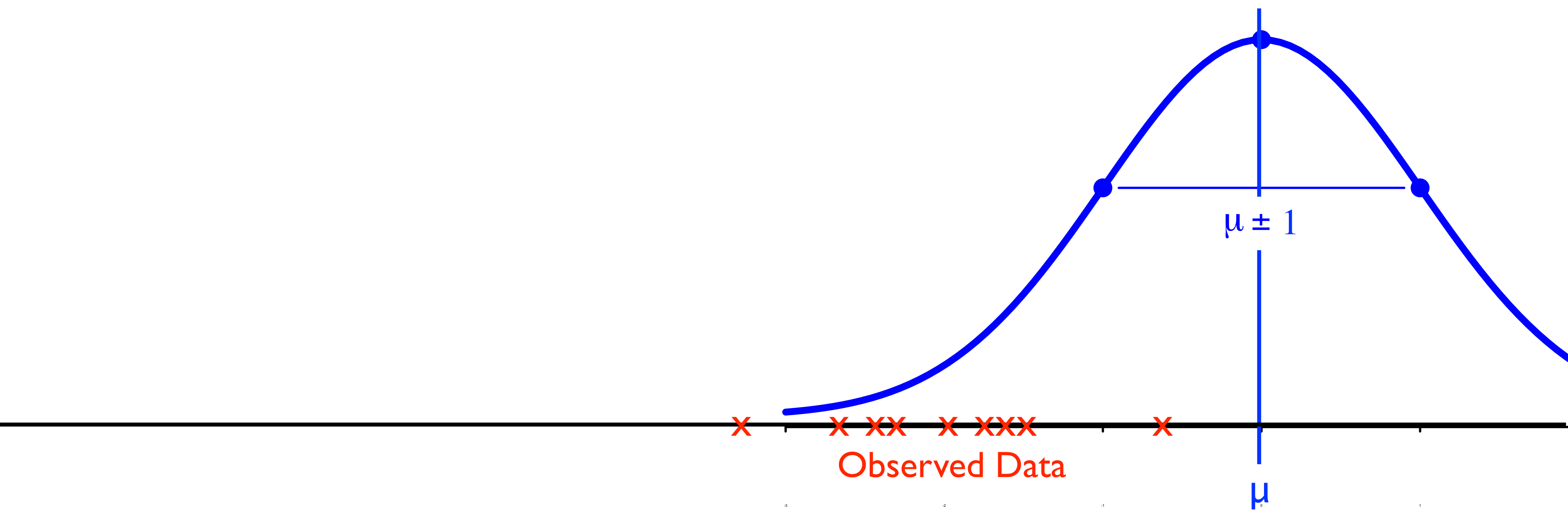
Which is more likely: (a) this?

μ unknown, σ^2 = 1



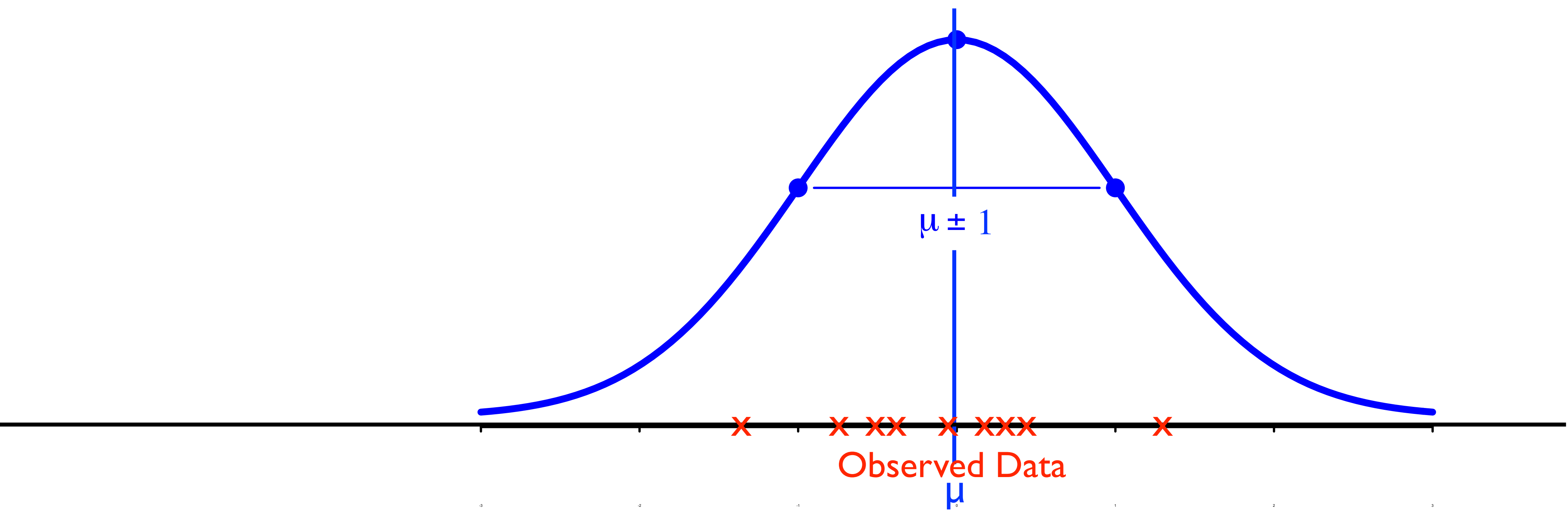
Which is more likely: (b) or this?

μ unknown, $\sigma^2 = 1$



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

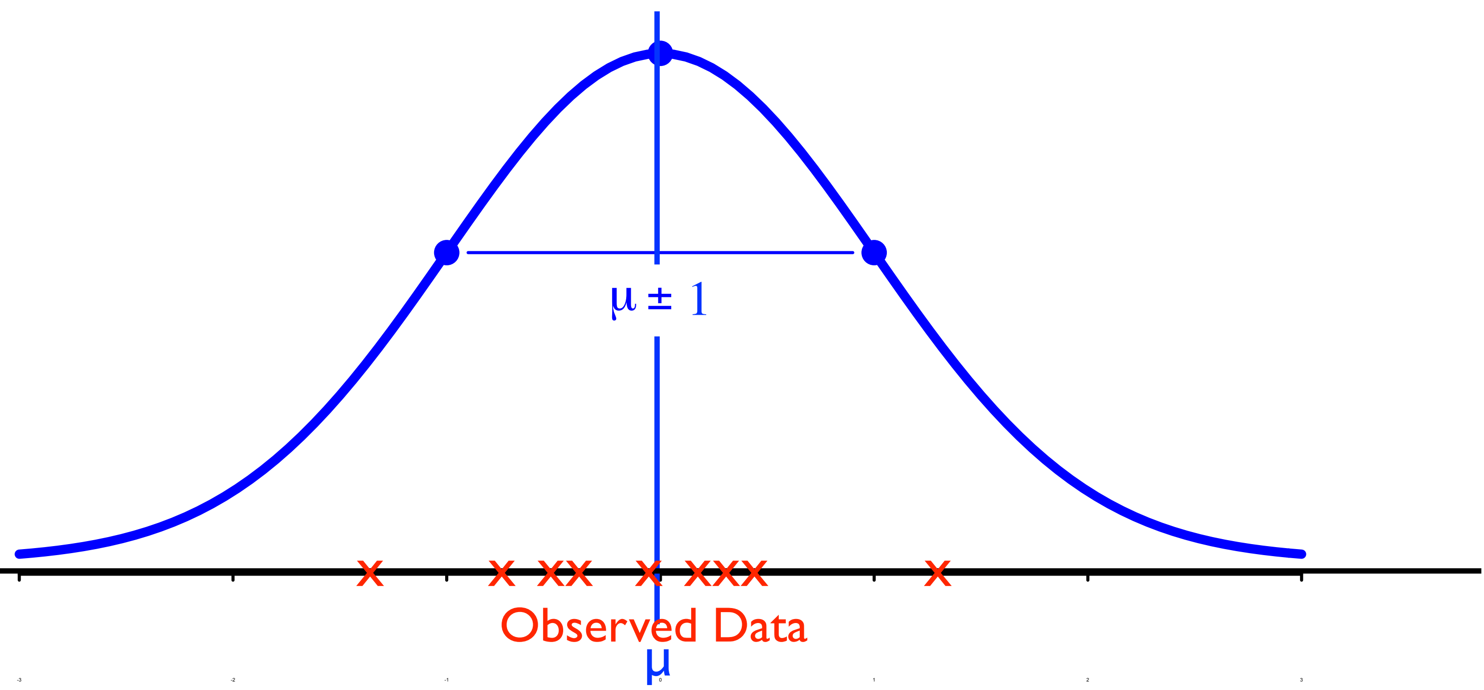
μ unknown, $\sigma^2 = 1$



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

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

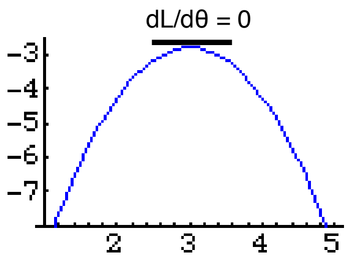
$$\ln L(x_1, x_2, \ldots, x_n|\theta) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \frac{(x_i-\theta)^2}{2}$$

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

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

$$\boxed{\widehat{\theta} = \left(\sum_{i=1}^{n} x_i\right) / n = \overline{x}}$$

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



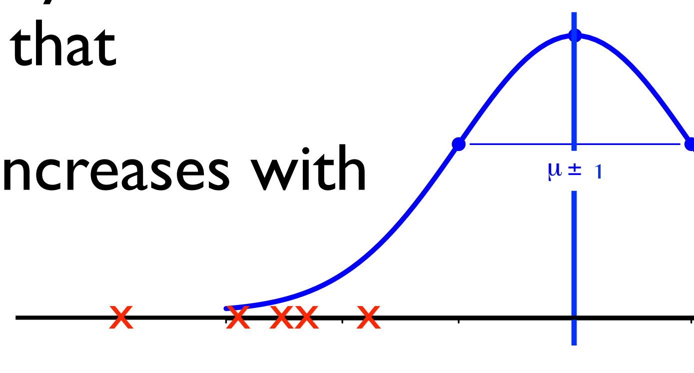
Sample mean is MLE of population mean

Hmm …, density ≠ 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

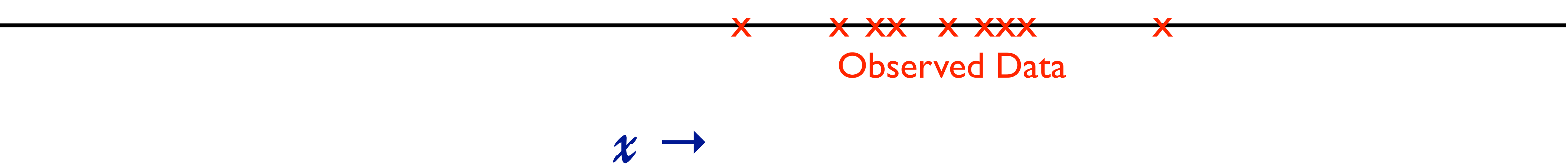
b) has desired property that likelihood increases with better fit to the model



and/or

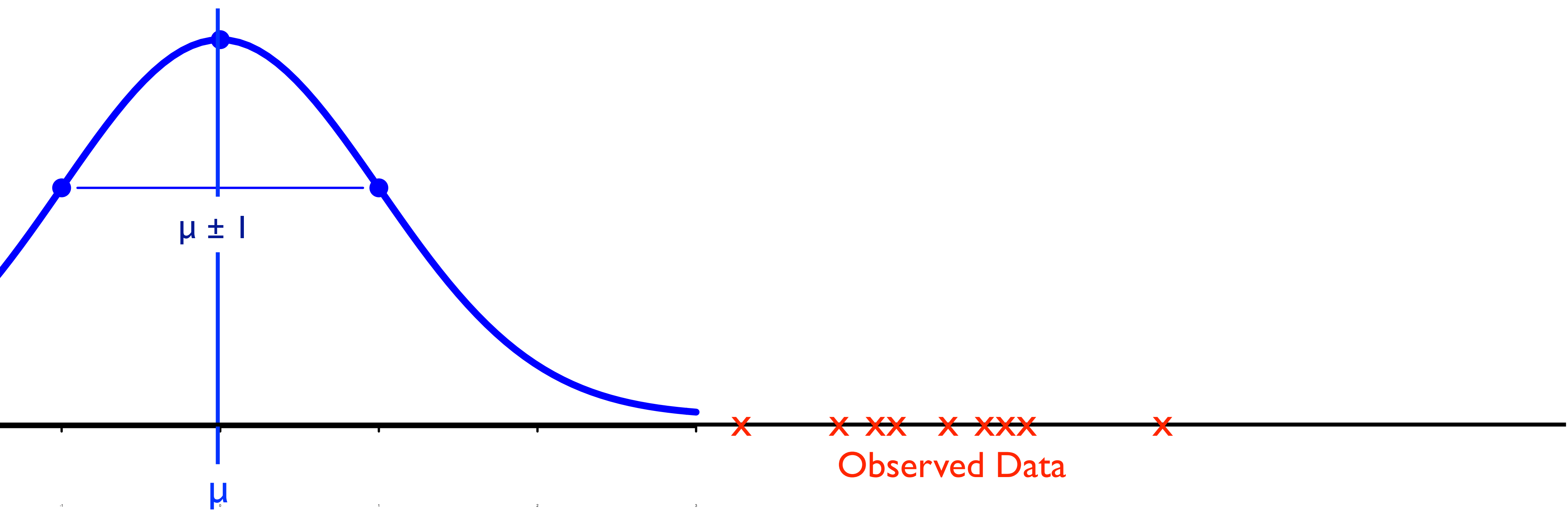
c) if density at x is $f(x)$, for any small $\delta>0$, the probability of a sample within $\pm\delta/2$ of x is $\approx \delta f(x)$, but δ is *constant* wrt θ, so it just drops out of $d/d\theta \log L(\ldots) = 0$.

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



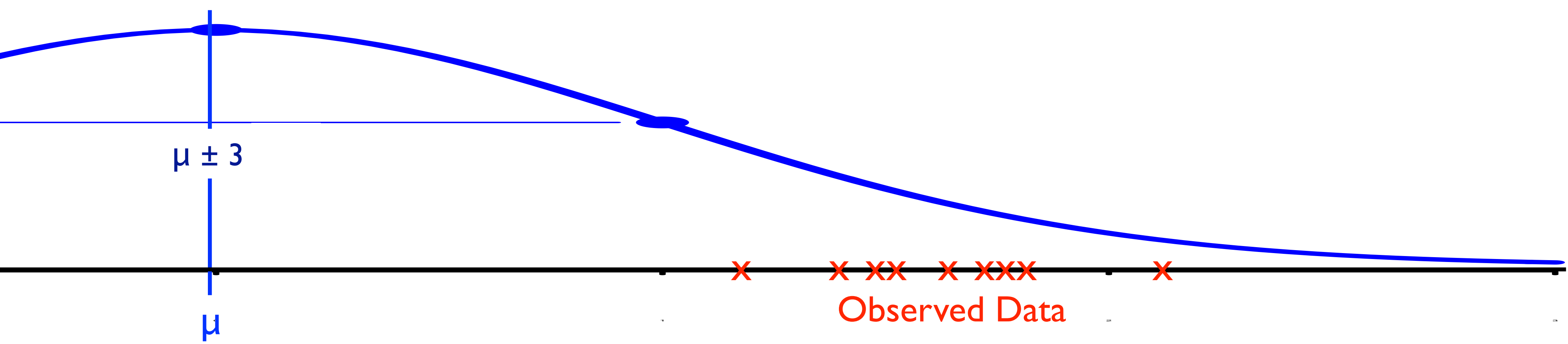
Which is more likely: (a) this?

μ, σ^2 both unknown



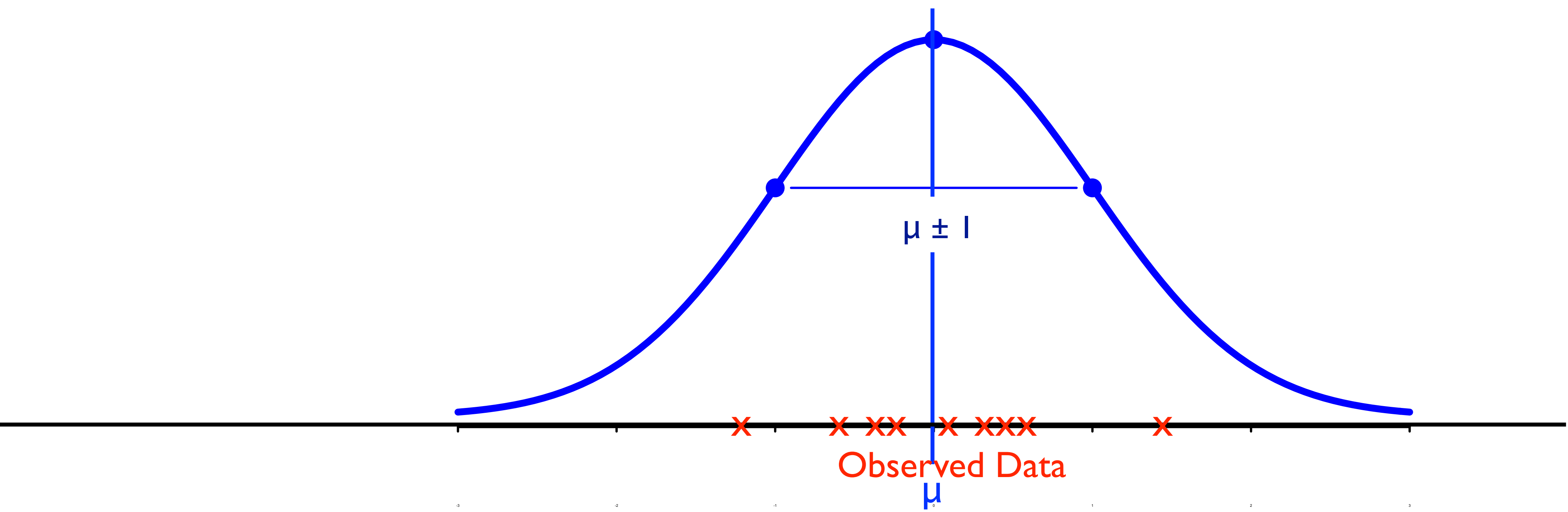
Which is more likely: (b) or this?

μ, σ^2 both unknown



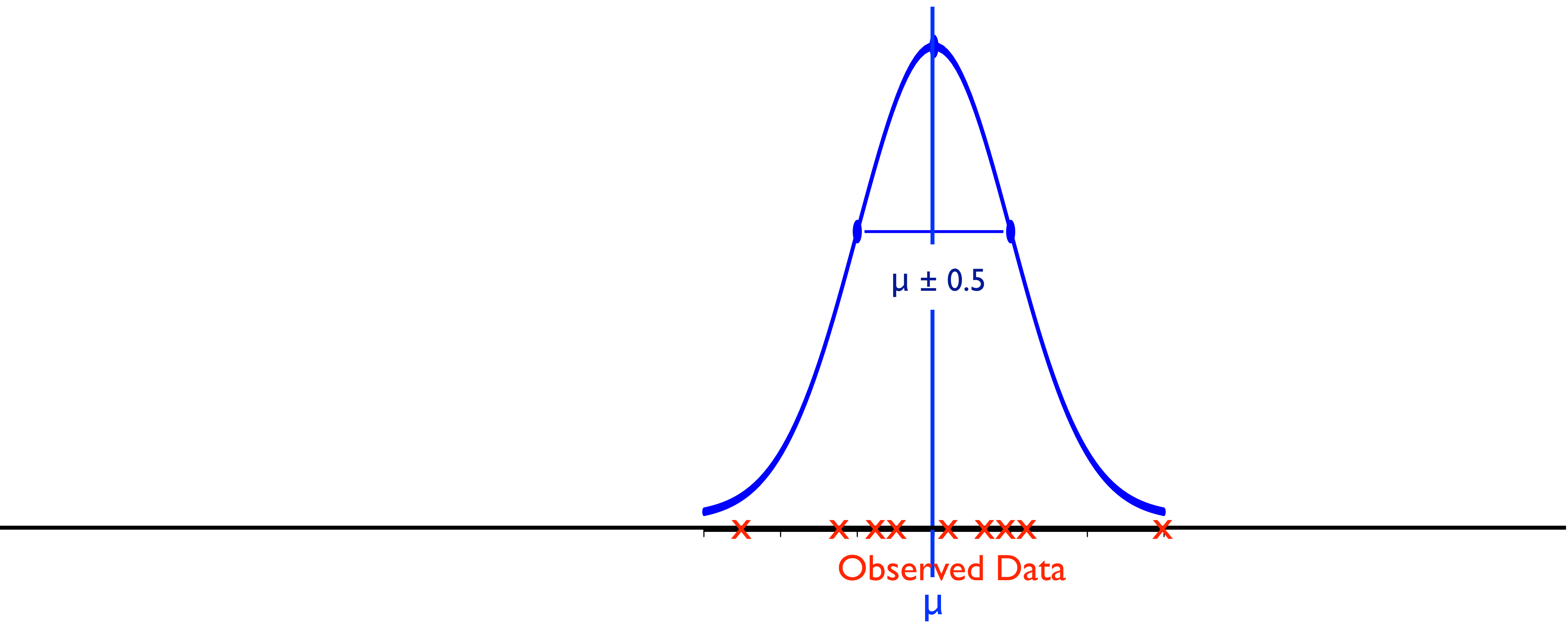
Which is more likely: (c) or this?

μ, σ^2 both unknown



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

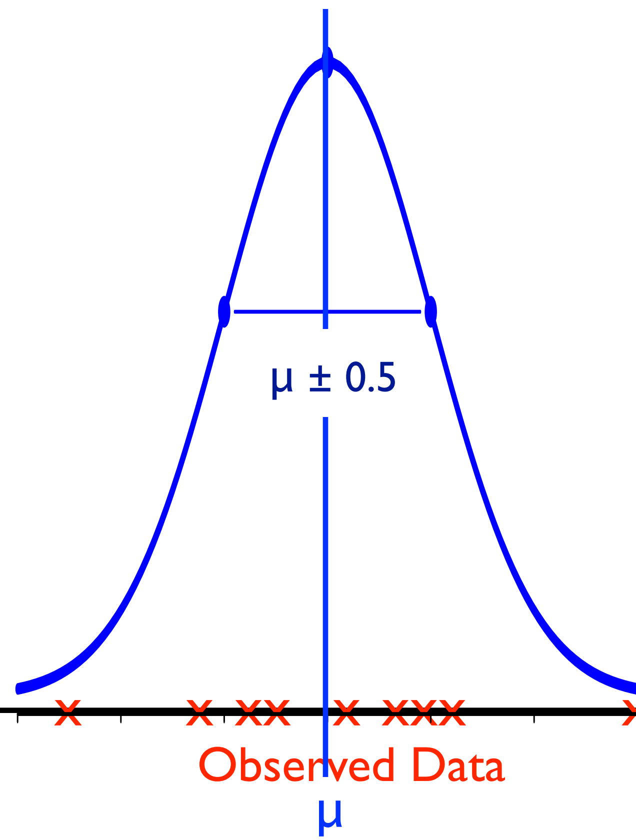
μ, σ^2 both unknown



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 $\underline{\underline{\& \ \sigma^2}}$?

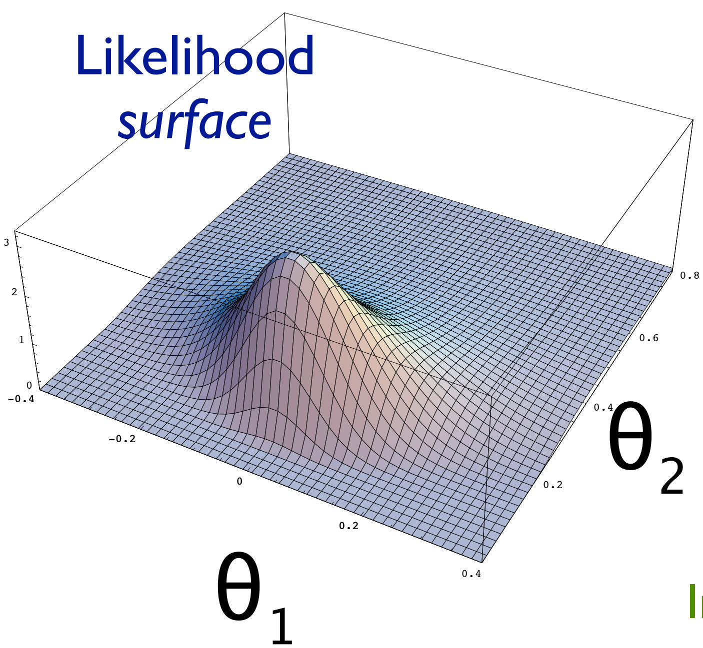


Ex 3: $x_i \sim N(\mu, \sigma^2),\ \mu, \sigma^2$ both unknown

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

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

$$\boxed{\widehat{\theta}_1 = \left(\sum_{i=1}^{n} x_i\right)/n = \overline{x}}$$



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

Ex. 3, (cont.)

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

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

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

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)

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*