Learning From Data: MLE

Maximum Likelihood Estimators

Parameter Estimation

Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, estimate θ.

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

$$\theta = \text{probability of Heads}$$

$f(x|\theta)$ 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 a *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

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

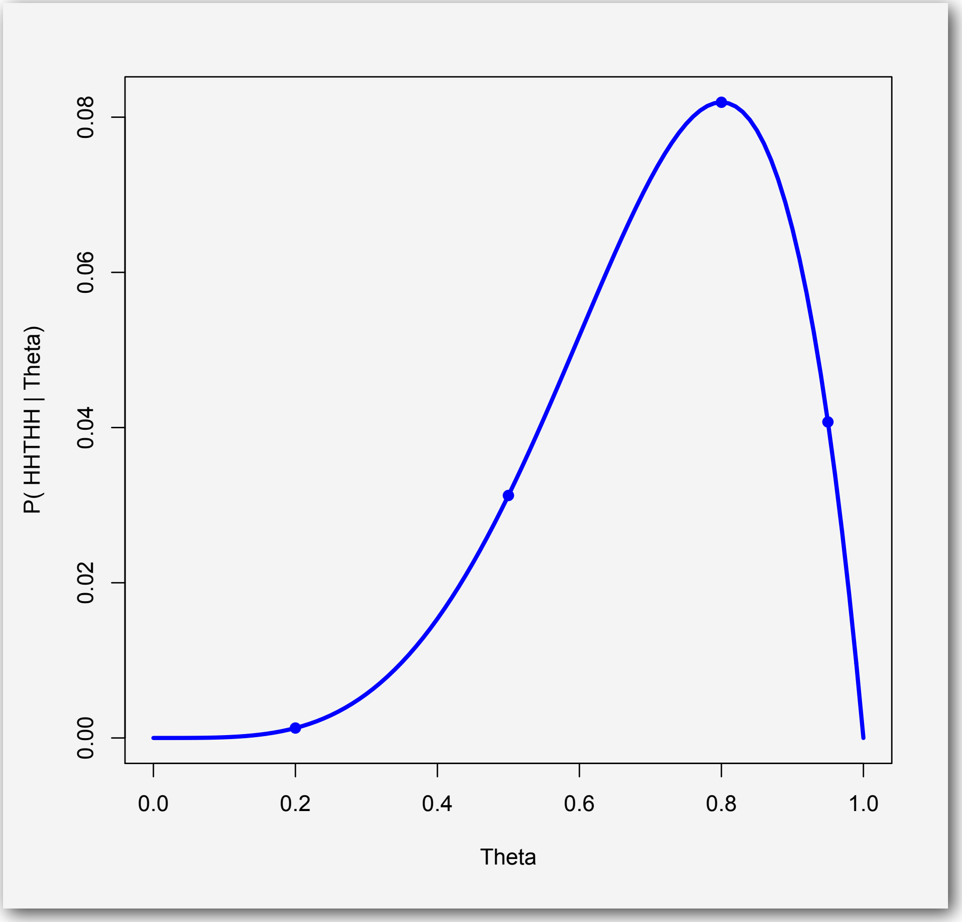
I.e., event HHTHH is *more likely* when $\theta = .6$ than $\theta = .5$

And what θ make HHTHH *most* likely?

Likelihood Function

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 $x_1, x_2, ..., x_n$

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

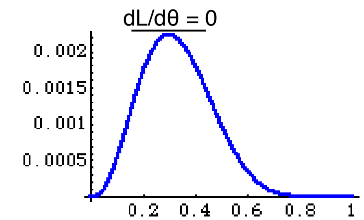
As a function of θ, what θ maximizes the likelihood of the data actually observed

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 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)

Bias

A desirable property: An estimator Y of a parameter θ is an *unbiased* estimator if

$$E[Y] = \theta$$

For coin ex. above, MLE is unbiased:

Y = fraction of heads = $(\Sigma_{1 \le i \le n} X_i)/n$,

(X_i = indicator for heads in i^{th} trial) so

$$E[Y] = (\Sigma_{1 \le i \le n} E[X_i])/n = n\,\theta/n = \theta$$

Aside: are all unbiased estimators equally good?

- No!
- E.g., "Ignore all but 1st flip; if it was H, let Y' = 1; else Y' = 0"
- Exercise: show this is unbiased
- Exercise: if observed data has at least one H and at least one T, what is the likelihood of the data given the model with $\theta = Y'$?

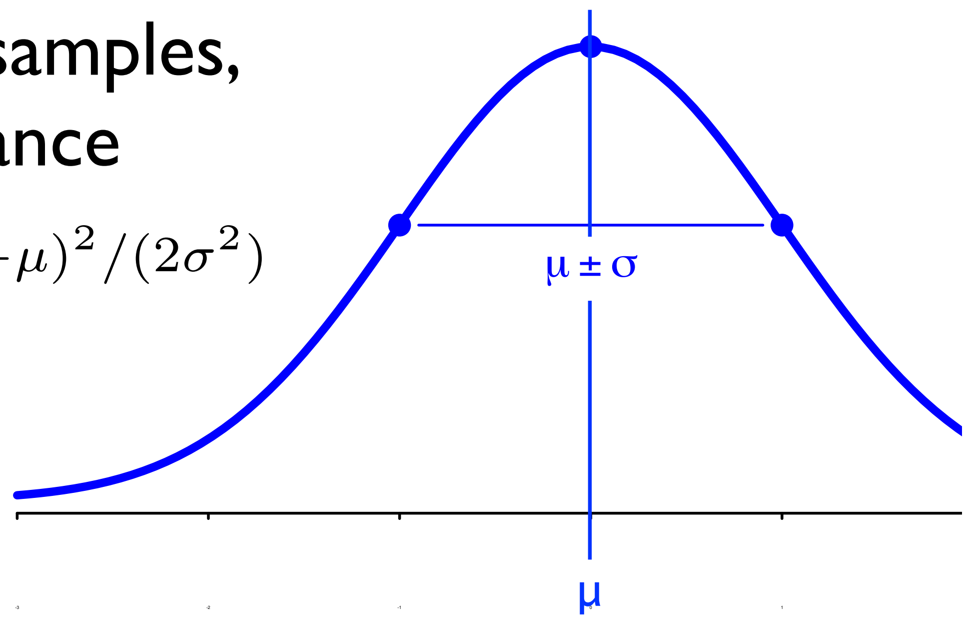
Parameter Estimation

Assuming sample $x_1, x_2, \ldots, x_n$ is from a parametric distribution $f(x|\theta)$, estimate θ.

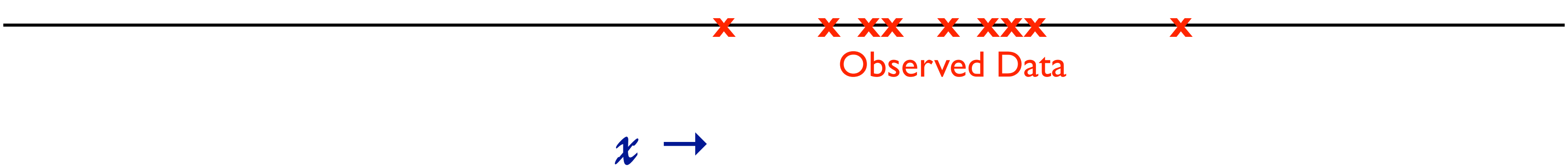
E.g.: Given n normal samples, estimate mean & variance

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

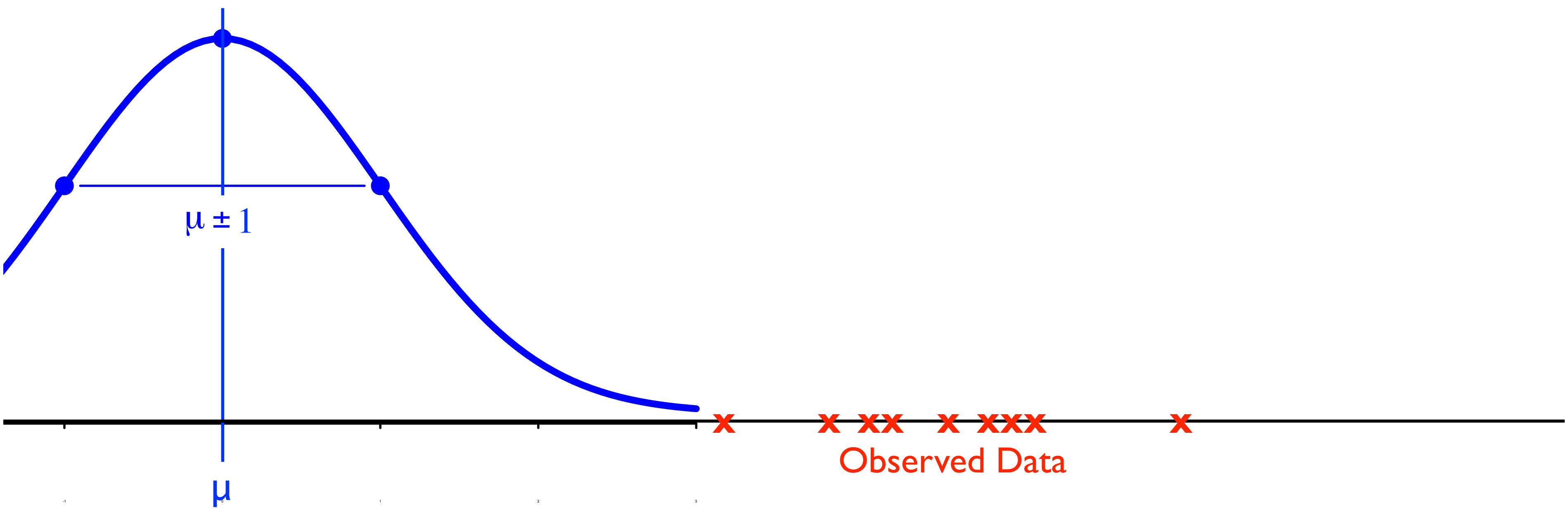
$$\theta = (\mu, \sigma^2)$$



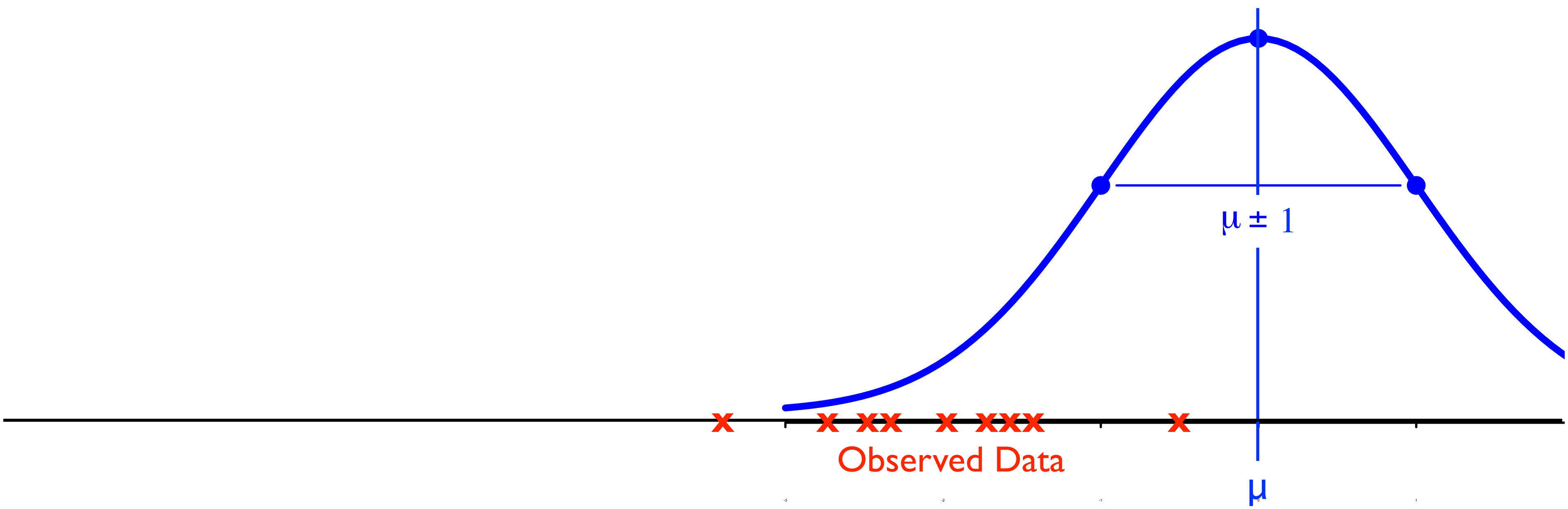
Ex2: I got data; a little birdie tells me it's normal, and promises $\sigma^2 = 1$



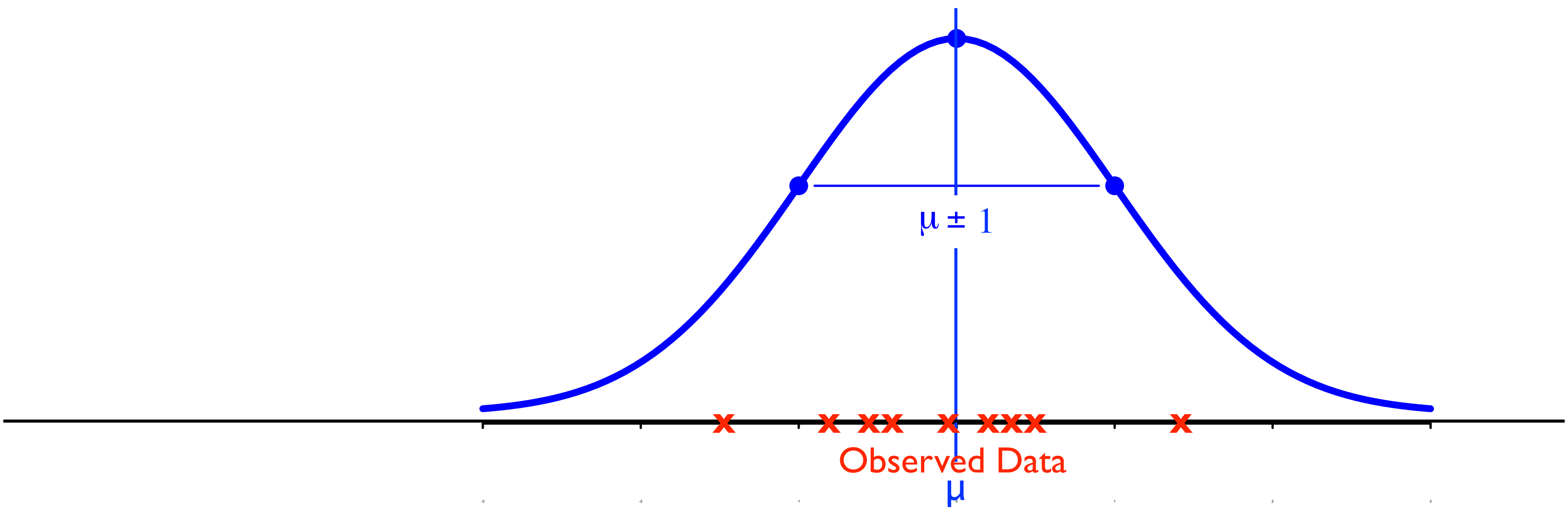
Which is more likely: (a) this?



Which is more likely: (b) or this?

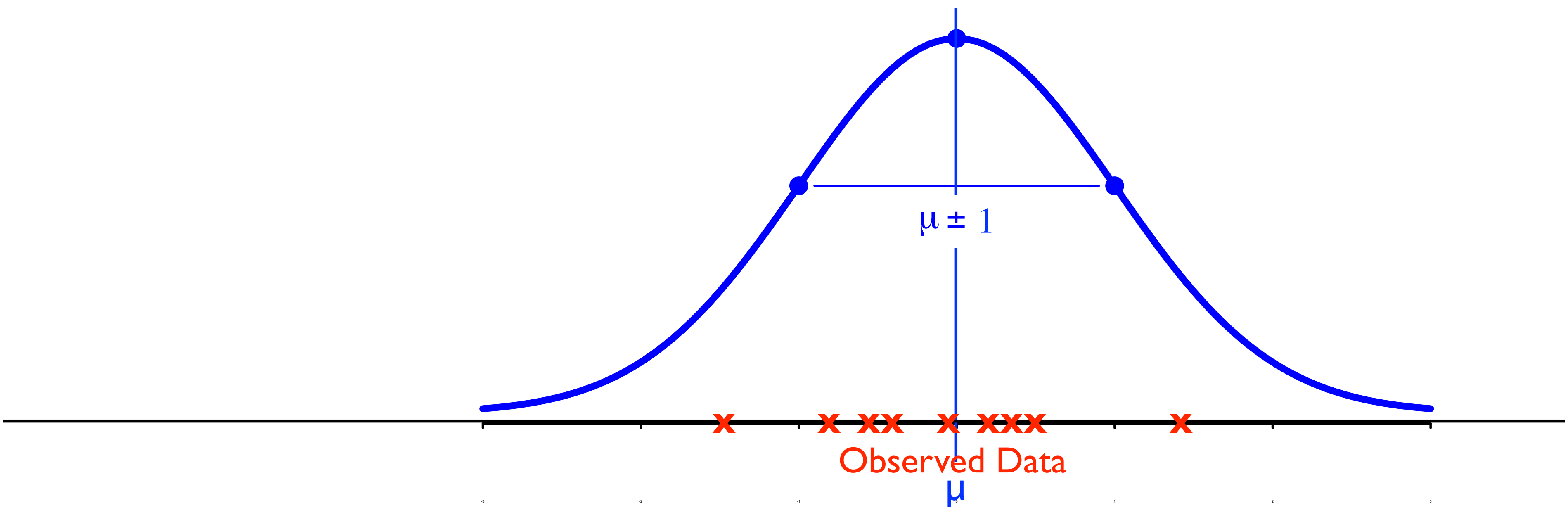


Which is more likely: (c) or *this*?



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

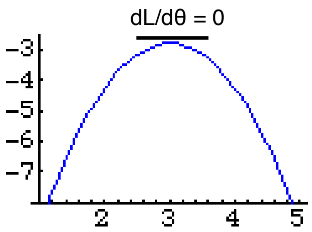
Looks good by eye, but how do I optimize my estimate of μ ?



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

$$
\begin{aligned}
L(x_1, x_2, \ldots, x_n|\theta) &= \prod_{1 \le i \le n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} \\
\ln L(x_1, x_2, \ldots, x_n|\theta) &= \sum_{1 \le i \le n} -\frac{1}{2} \ln 2\pi - \frac{(x_i-\theta)^2}{2} \\
\frac{d}{d\theta} \ln L(x_1, x_2, \ldots, x_n|\theta) &= \sum_{1 \le i \le n} (x_i - \theta) \\
&= \left(\sum\nolimits_{1 \le i \le n} x_i\right) - n\theta \ = \ 0
\end{aligned}
$$

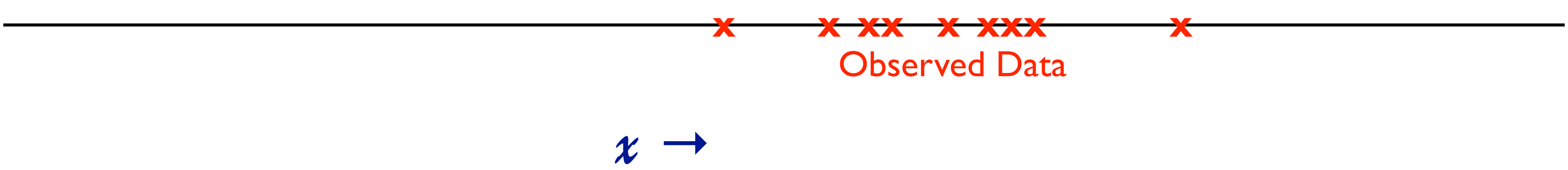
And verify it's max, not min & not better on boundary



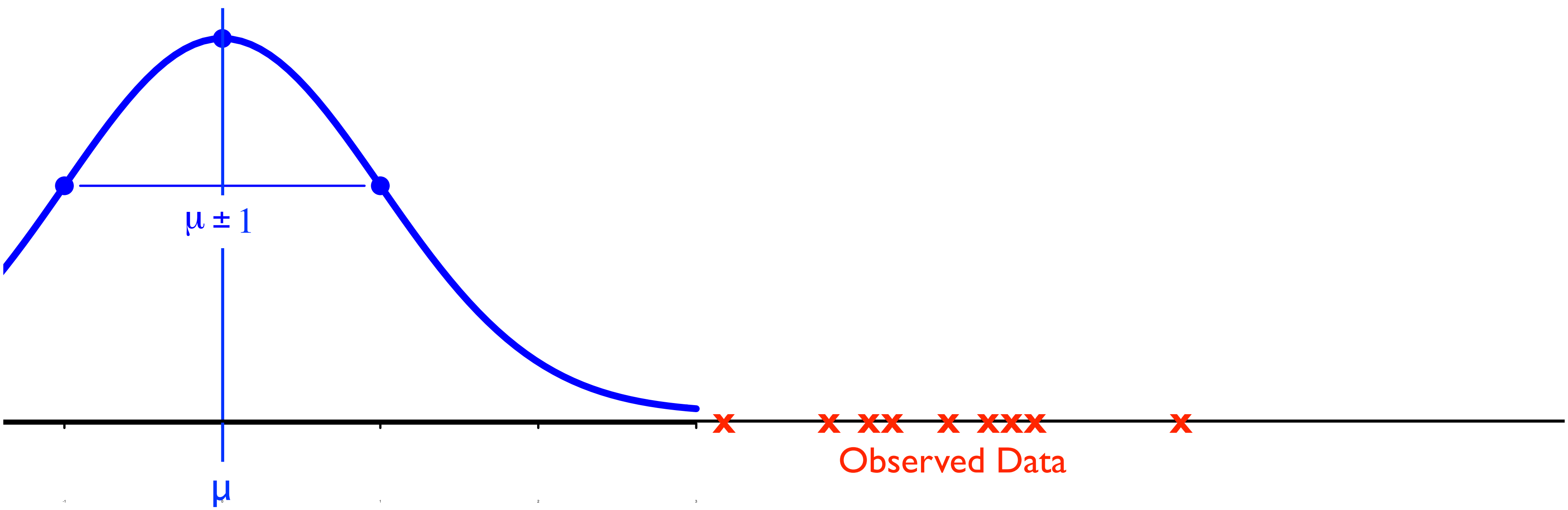
$$\boxed{\hat{\theta} \ = \ \left(\sum\nolimits_{1 \le i \le n} x_i\right)/n \ = \ \bar{x}}$$

Sample mean is MLE of population mean

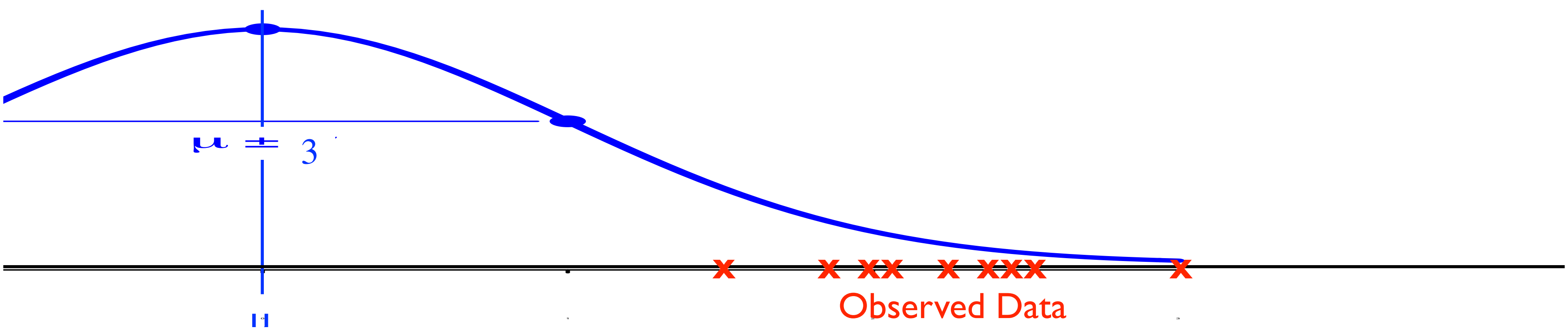
Ex3: I got data; a little birdie tells me it's normal (but does *not* tell me σ^2)



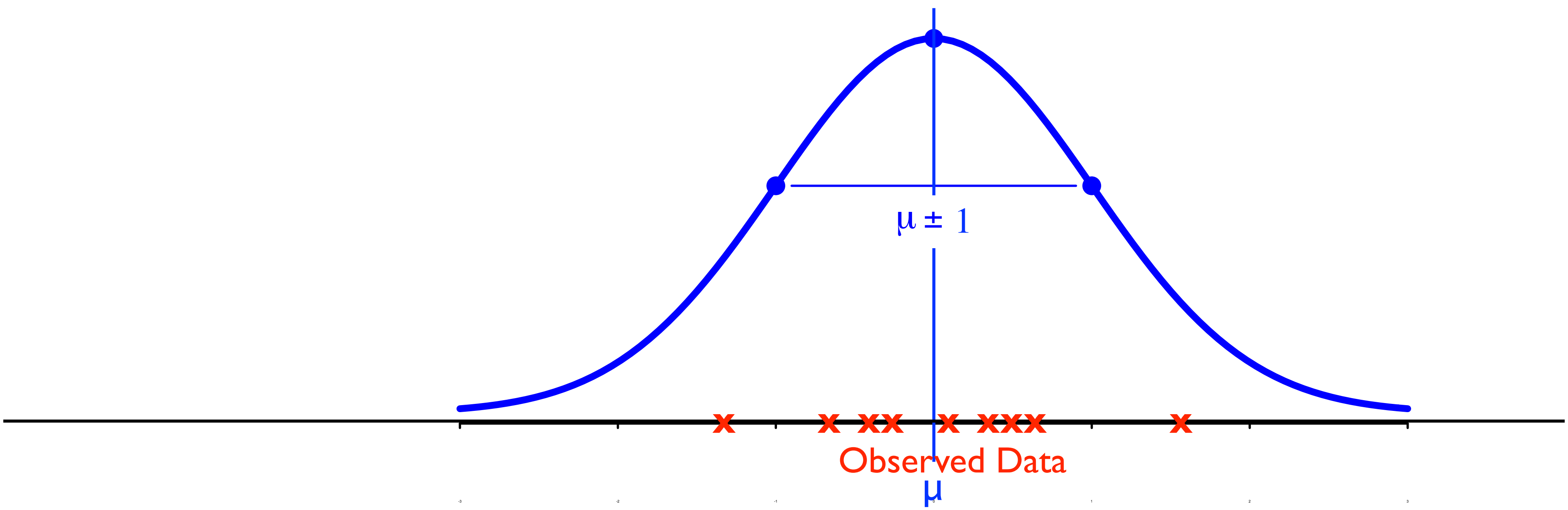
Which is more likely: (a) this?



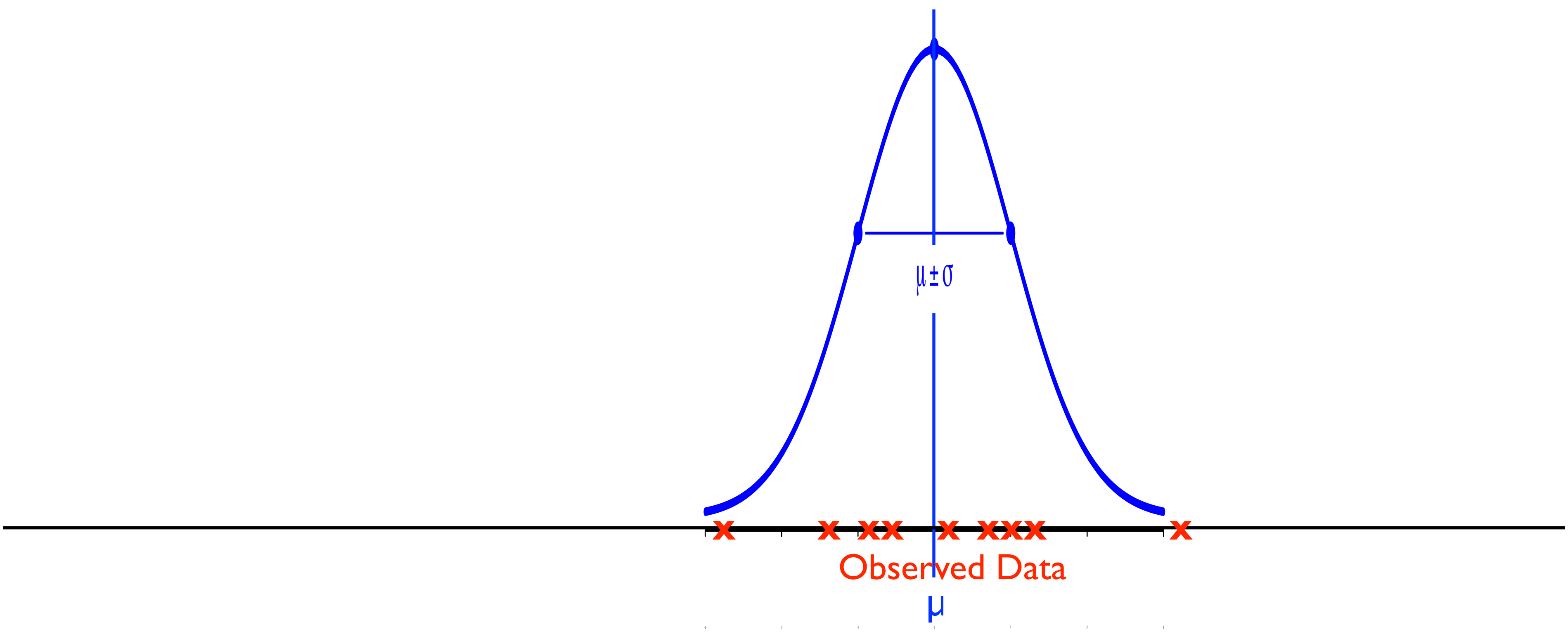
Which is more likely: (b) or this?



Which is more likely: (c) or this?

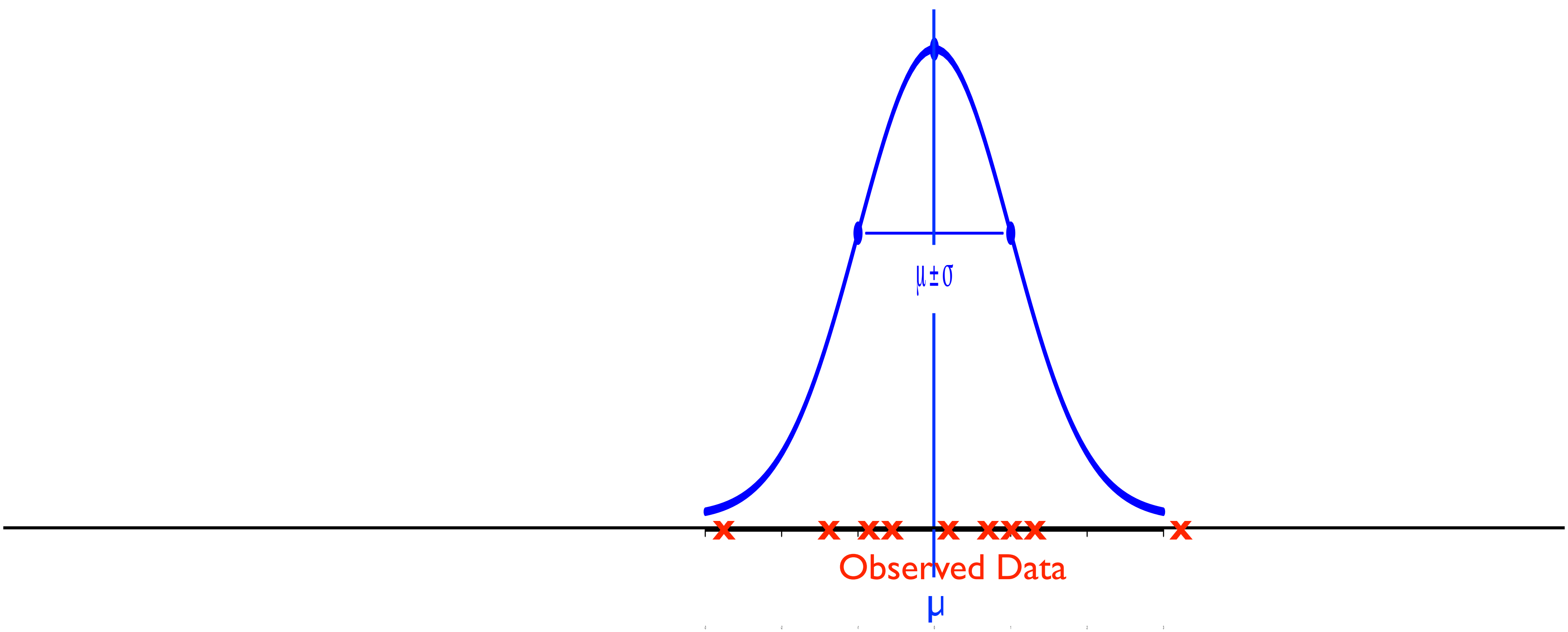


Which is more likely: (d) or *this*?



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

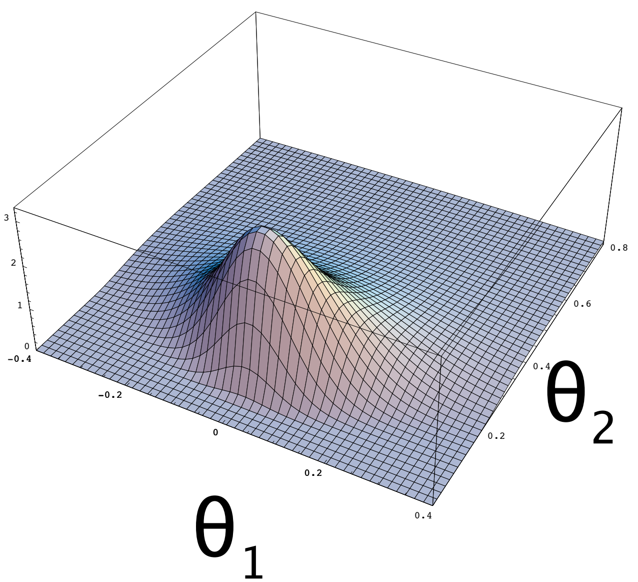
Looks good by eye, but how do I optimize my estimates of μ & σ ?



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

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

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



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

Sample mean is MLE of population mean, again

Ex. 3, (cont.)

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

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

Sample *variance* is MLE of population variance

Ex. 3, (cont.)

Bias? if Y is sample mean

$Y = (\Sigma_{1 \le i \le n} X_i)/n$

then

$E[Y] = (\Sigma_{1 \le i \le n} E[X_i])/n = n\ \mu/n = \mu$

so the MLE is an *unbiased* estimator of population mean

Similarly, $(\Sigma_{1 \le i \le n} (X_i-\mu)^2)/n$ is an unbiased estimator of σ^2. Unfortunately, if μ is unknown, estimated *from the same data*, as above, $\hat{\theta}_2 = \sum_{1 \le i \le n} \frac{(x_i - \hat{\theta}_1)^2}{n}$ is a consistent, but *biased* estimate of population variance. (An example of *overfitting.*) Unbiased estimate is:

$$\boxed{\hat{\theta}_2' = \sum_{1 \le i \le n} \frac{(x_i - \hat{\theta}_1)^2}{n-1}}$$

I.e., $\lim_{n \to \infty}$ = correct

Moral: MLE is a great idea, but not a magic bullet

More on Bias of $\hat{\theta}_2$

Biased? Yes. Why? As an extreme, think about n = 1. Then $\hat{\theta}_2 = 0$; probably an underestimate!

Also, think about n = 2. Then $\hat{\theta}_1$ is exactly between the two sample points, the position that exactly minimizes the expression for θ_2. Any other choices for θ_1, θ_2 make the likelihood of the observed data slightly *lower*. But it's actually pretty unlikely that two sample points would be chosen exactly equidistant from, and on opposite sides of the mean, so the MLE $\hat{\theta}_2$ systematically underestimates θ_2.

(But not by much, & bias shrinks with sample size.)

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*