CSEP 527

Spring 2016

4. Maximum Likelihood Estimation and the E-M Algorithm

Outline

HW#2 Discussion

MLE: Maximum Likelihood Estimators

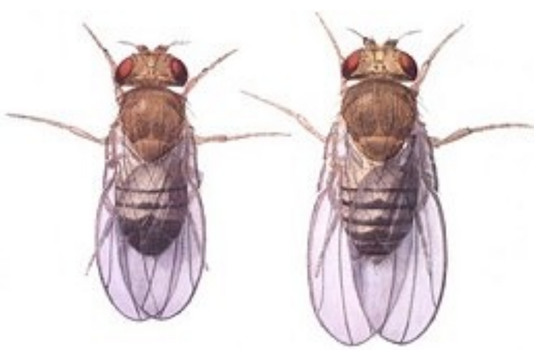
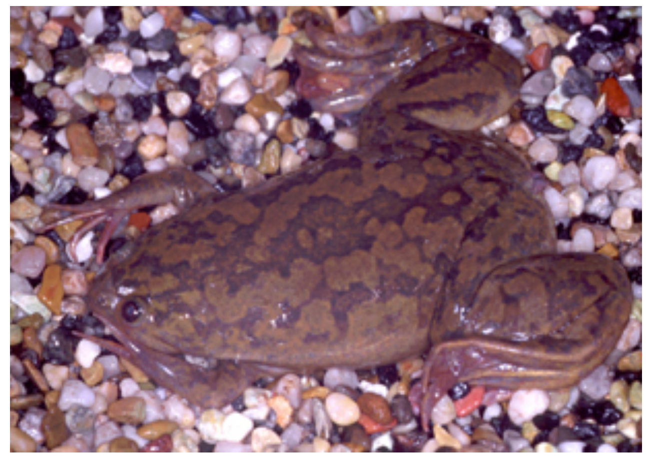
EM: the Expectation Maximization Algorithm

Next: Motif description & discovery

HW # 2 Discussion

	Species	Name	Description	Access -ion	score to #1
1	Homo sapiens (Human)	MYOD1_HUMAN	Myoblast determination protein 1	P15172	1709
2	Homo sapiens (Human)	TAL1_HUMAN	T-cell acute lymphocytic leukemia protein 1 (TAL-1)	P17542	143
3	Mus musculus (Mouse)	MYOD1_MOUSE	Myoblast determination protein 1	P10085	1494
4	Gallus gallus (Chicken)	MYOD1_CHICK	Myoblast determination protein 1 homolog (MYOD1 homolog)	P16075	1020
5	Xenopus laevis (African clawed frog)	MYODA_XENLA	Myoblast determination protein 1 homolog A (Myogenic factor 1)	P13904	978
6	Danio rerio (Zebrafish)	MYOD1_DANRE	Myoblast determination protein 1 homolog (Myogenic factor 1)	Q90477	893
7	Branchiostoma belcheri (Amphioxus)	Q8IU24_BRABE	MyoD-related	Q8IU24	428
8	Drosophila melanogaster (Fruit fly)	MYOD_DROME	Myogenic-determination protein (Protein nautilus) (dMyd)	P22816	368
9	Caenorhabditis elegans	LIN32_CAEEL	Protein lin-32 (Abnormal cell lineage protein 32)	Q10574	118
10	Homo sapiens (Human)	SYFM_HUMAN	Phenylalanyl-tRNA synthetase, mitochondrial	O95363	56



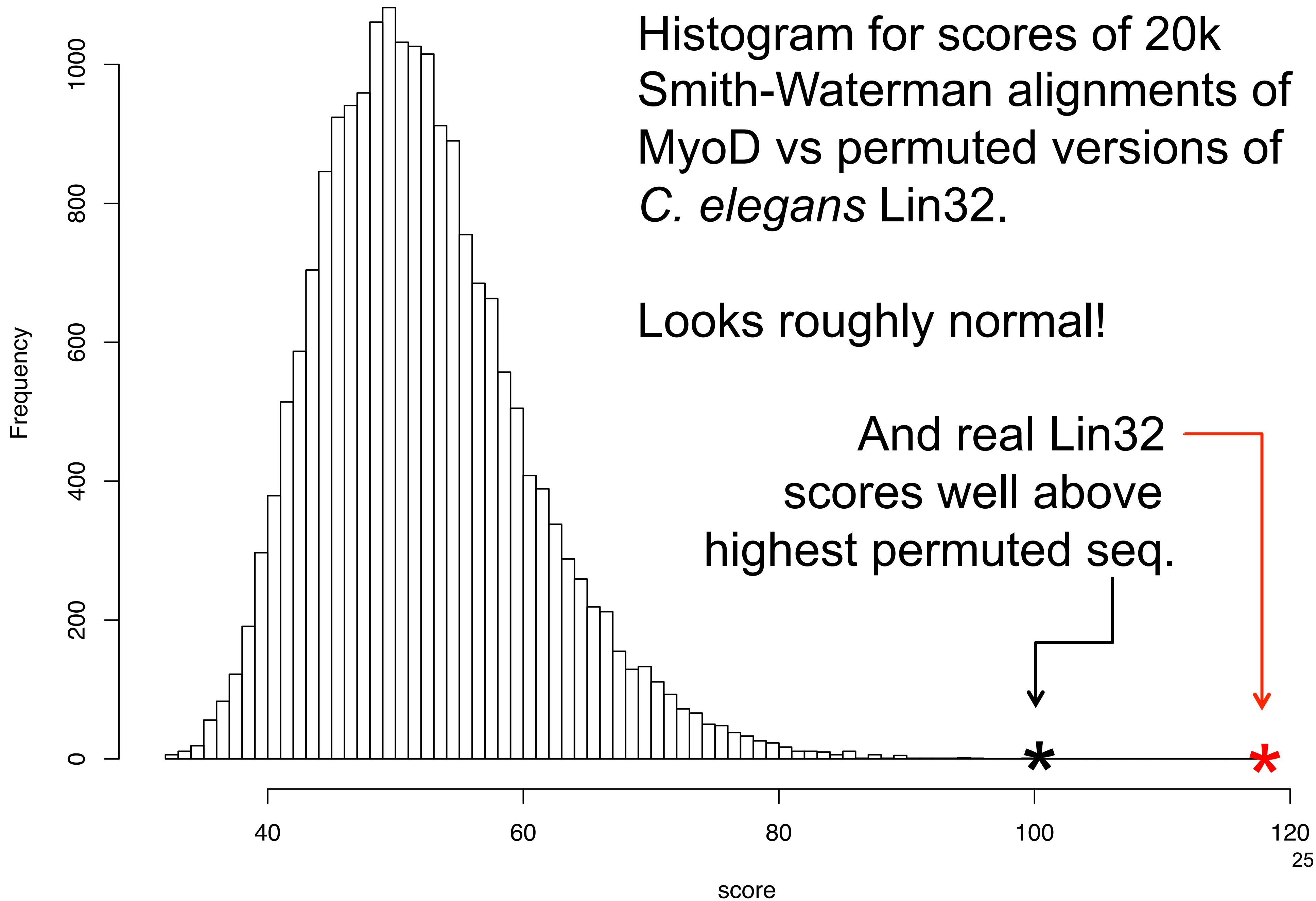



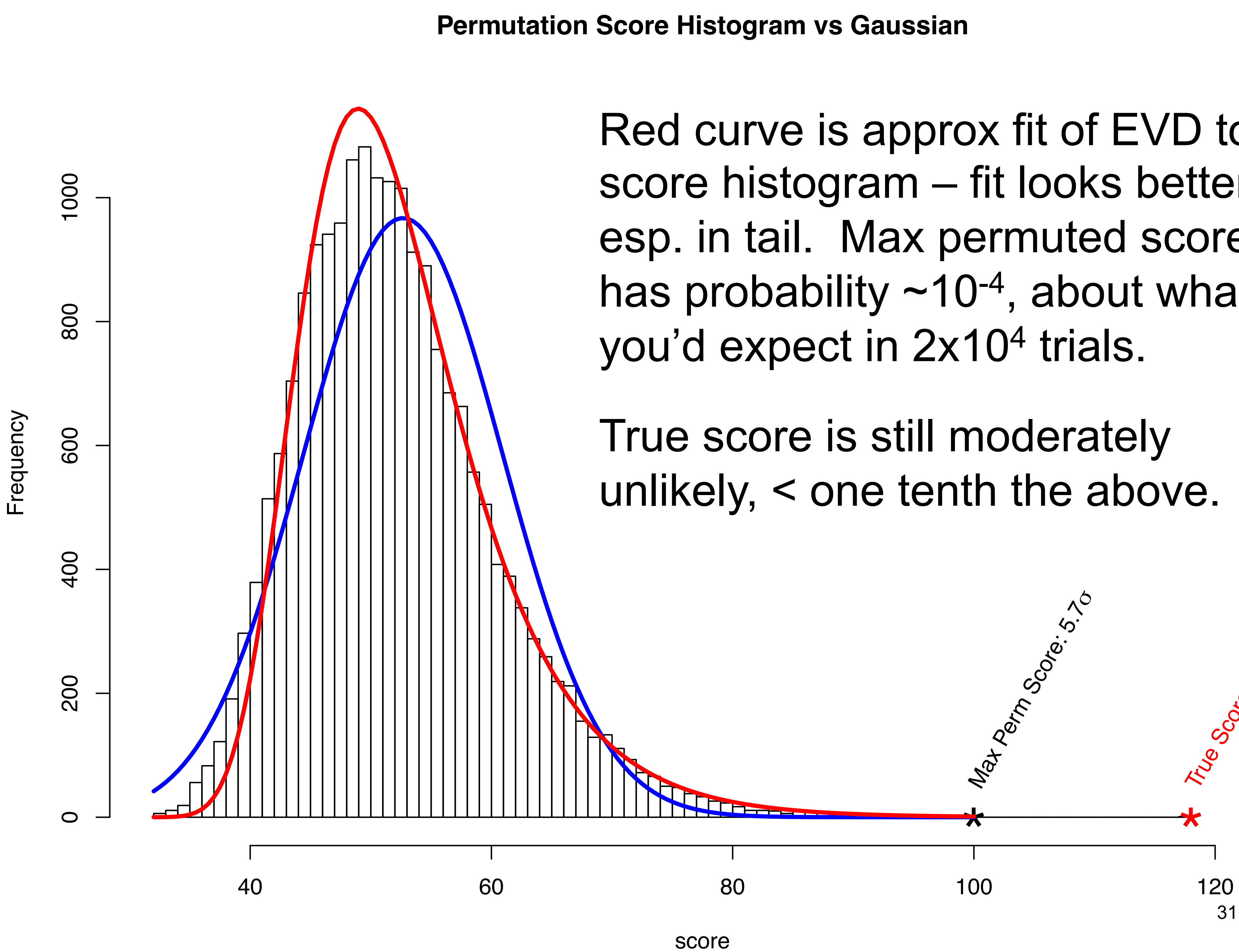

Courtesy of Ralf Sommer

MyoD

http://www.rcsb.org/pdb/explore/jmol.do?structureId=1MDY&bionumber=1

Permutation Score Histogram vs Gaussian





Red curve is approx fit of EVD to score histogram – fit looks better, esp. in tail. Max permuted score has probability ~10^{-4}, about what you'd expect in 2×10^4 trials.

True score is still moderately unlikely, < one tenth the above.

Learning From Data: MLE

Maximum Likelihood Estimators

Parameter Estimation

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

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

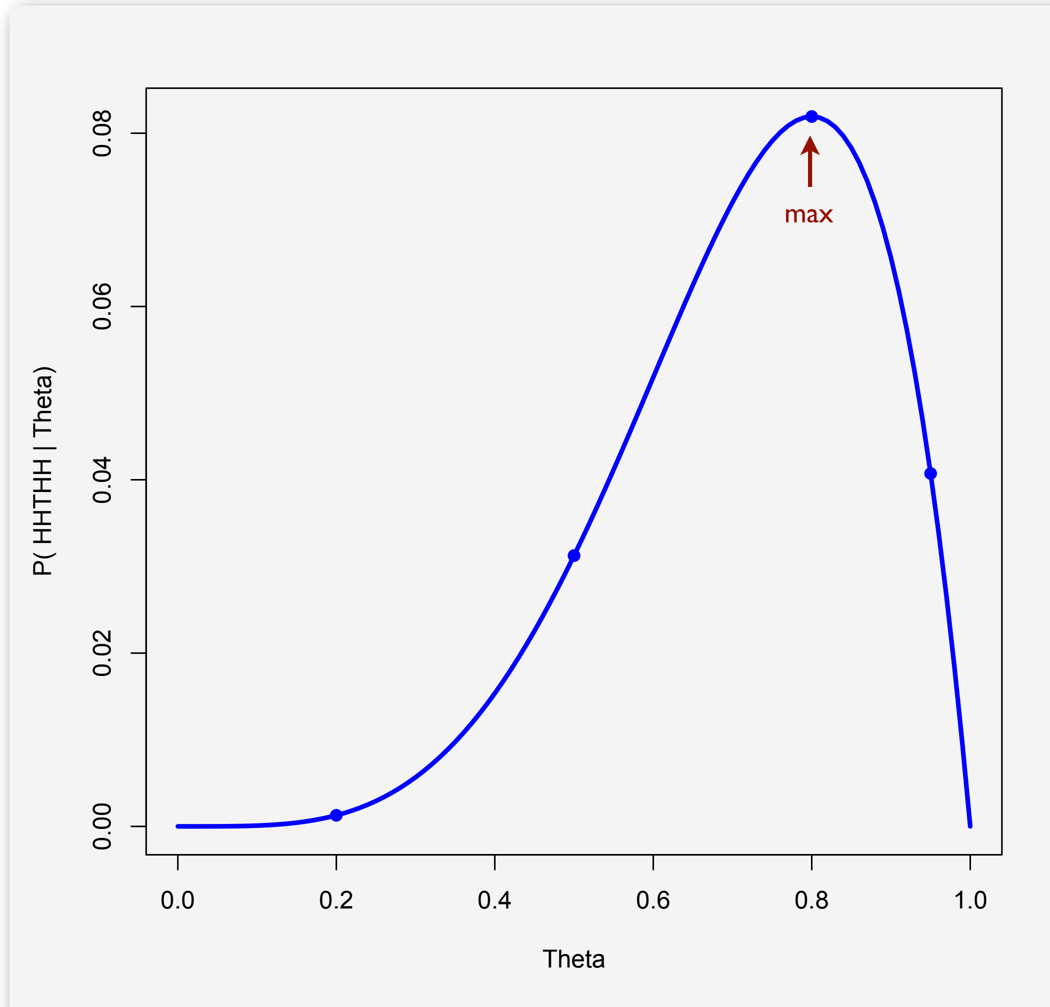
I.e., event HHTHH is *more likely* when $\theta = .6$ than $\theta = .5$

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 $x_1, x_2, ..., x_n$

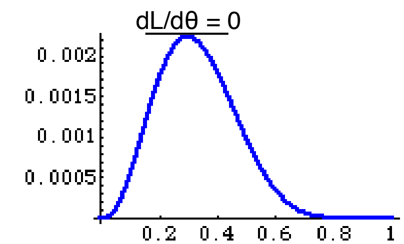
$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

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)

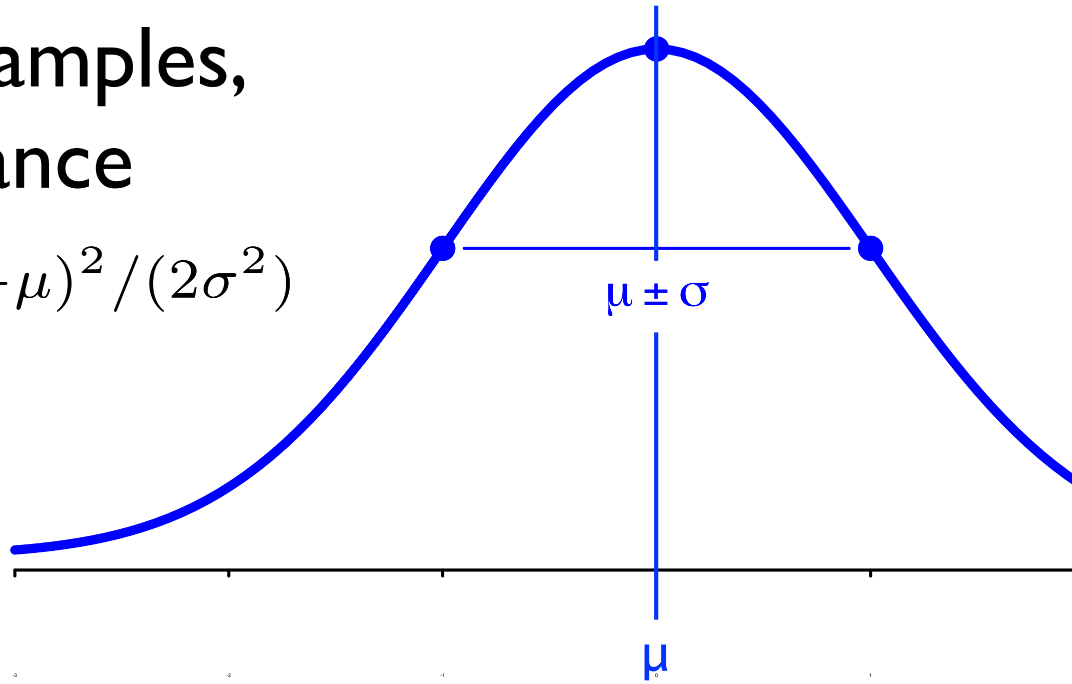
Parameter Estimation

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

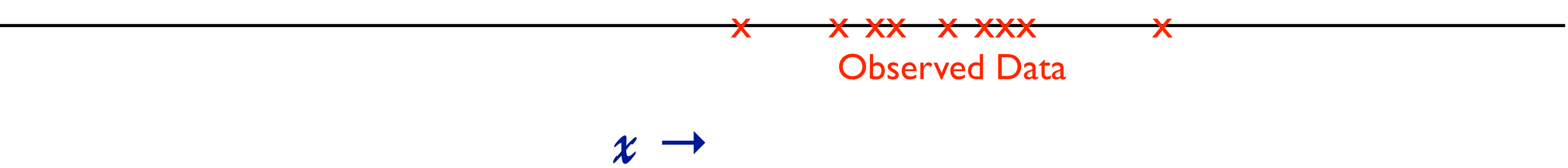
E.g.: Given *n normal* samples, estimate mean & variance

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

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

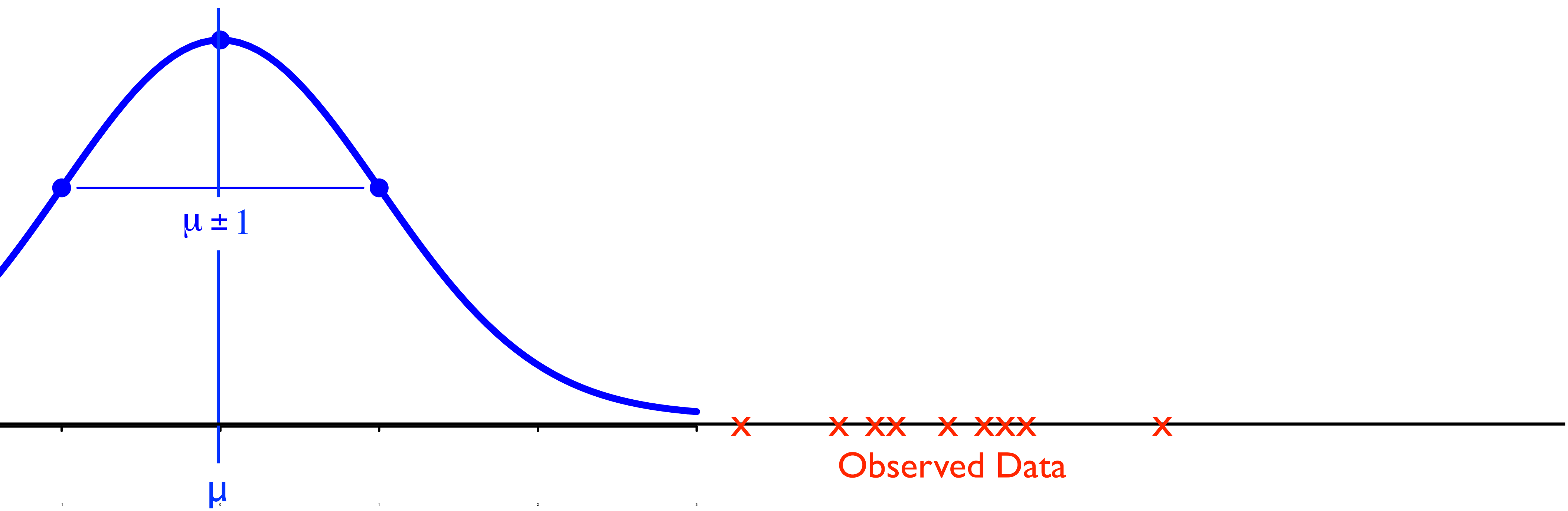


Ex2: I got data; a little birdie tells me it's normal, and promises $\sigma^2 = 1$



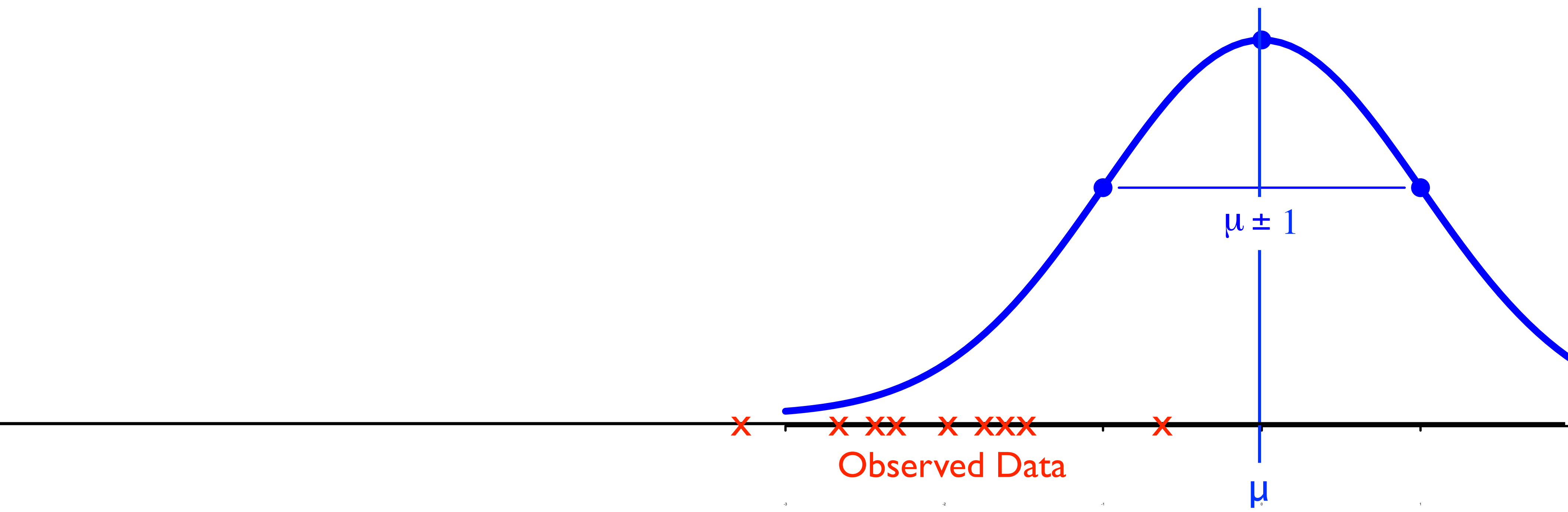
Which is more likely: (a) this?

μ unknown, $\sigma^2 = 1$



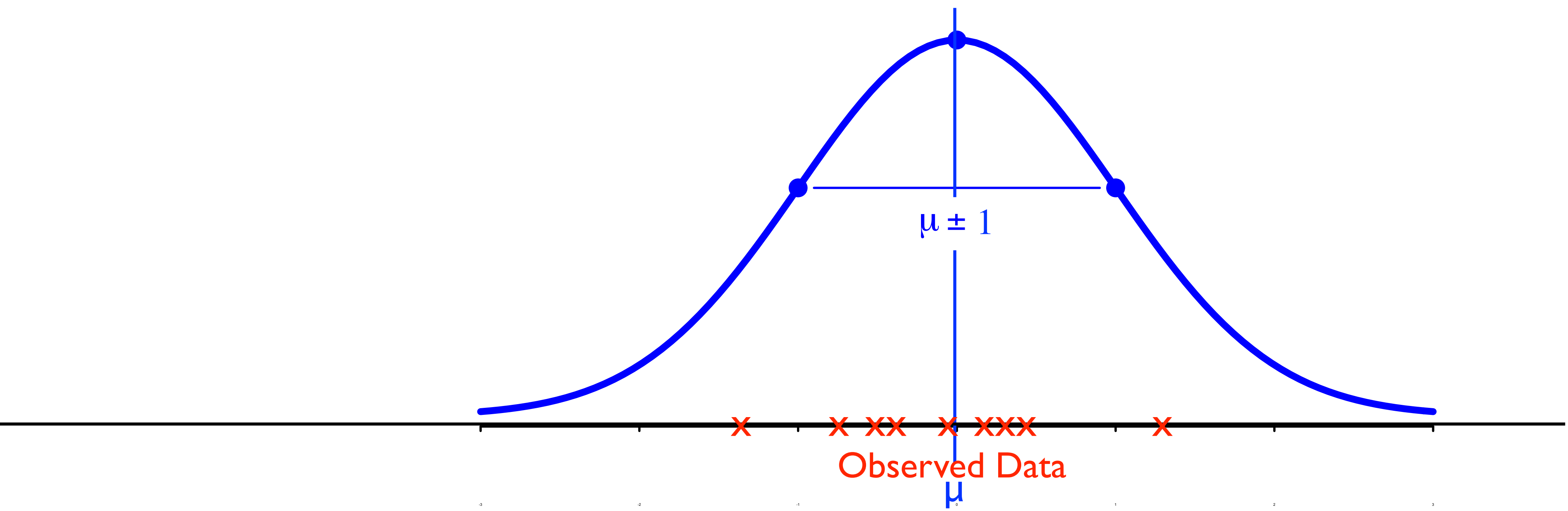
Which is more likely: (b) or this?

μ unknown, $\sigma^2 = 1$



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

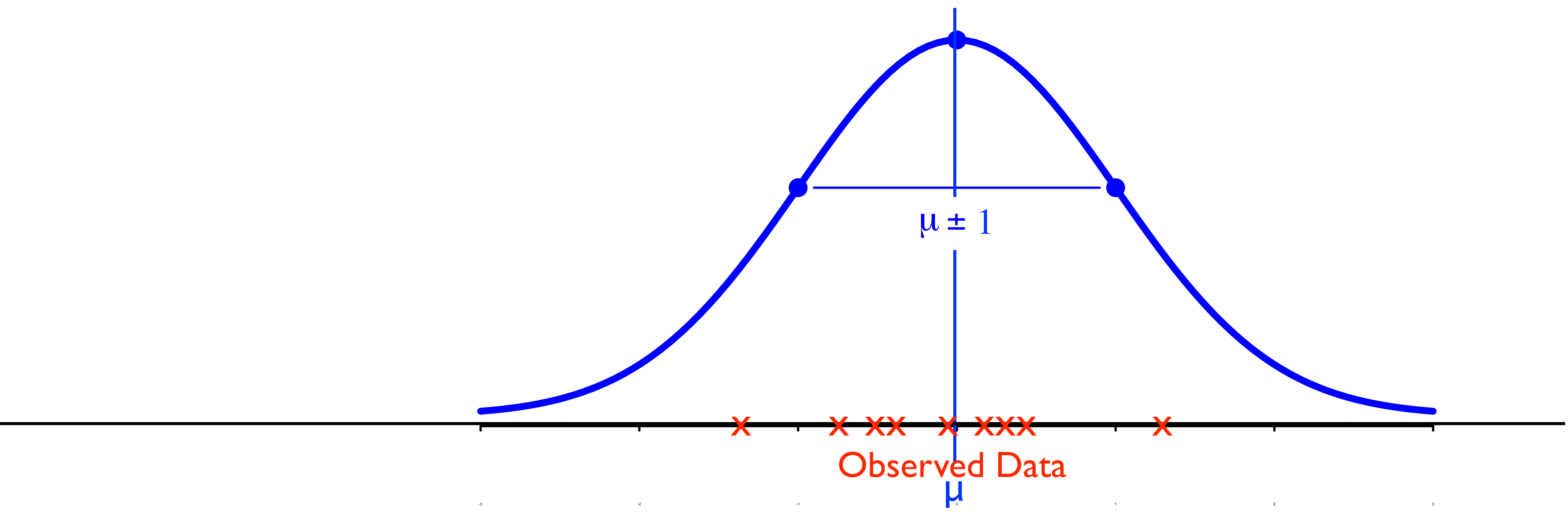
μ unknown, $\sigma^2 = 1$



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

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

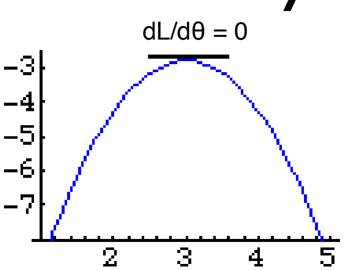
$$\ln L(x_1, x_2, \ldots, x_n|\theta) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \frac{(x_i-\theta)^2}{2}$$

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

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

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

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



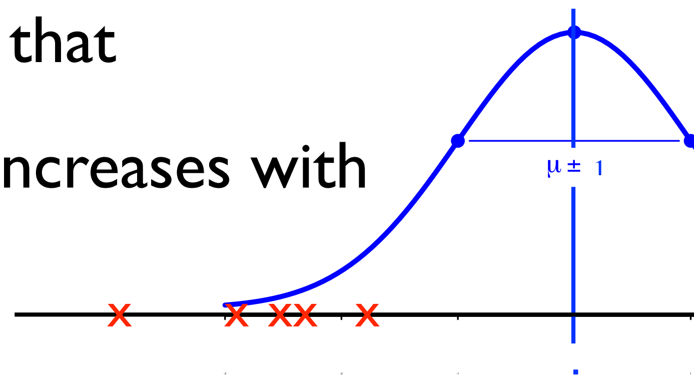
Sample mean is MLE of population mean

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

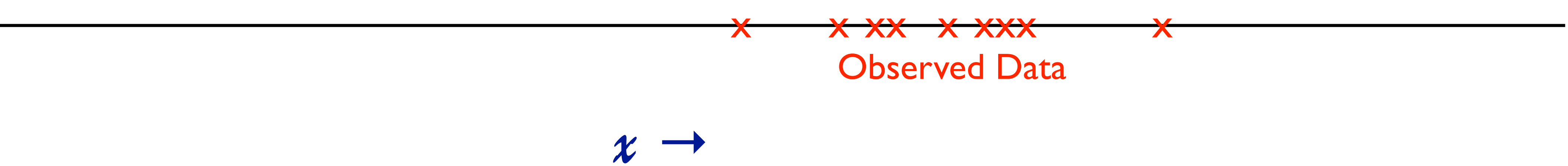
b) has desired property that likelihood increases with better fit to the model



and/or

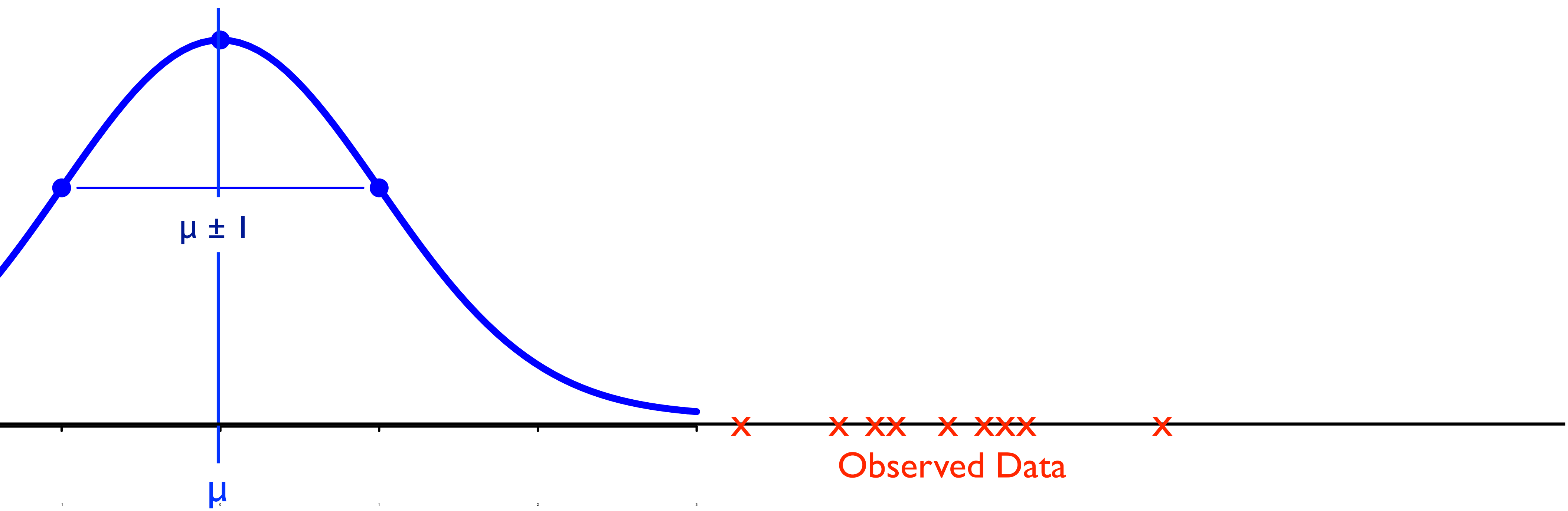
c) if density at x is $f(x)$, for any small $\delta>0$, the probability of a sample within $\pm\delta/2$ of x is $\approx \delta f(x)$, but δ is *constant* wrt θ, so it just drops out of $d/d\theta \log L(\ldots) = 0$.

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



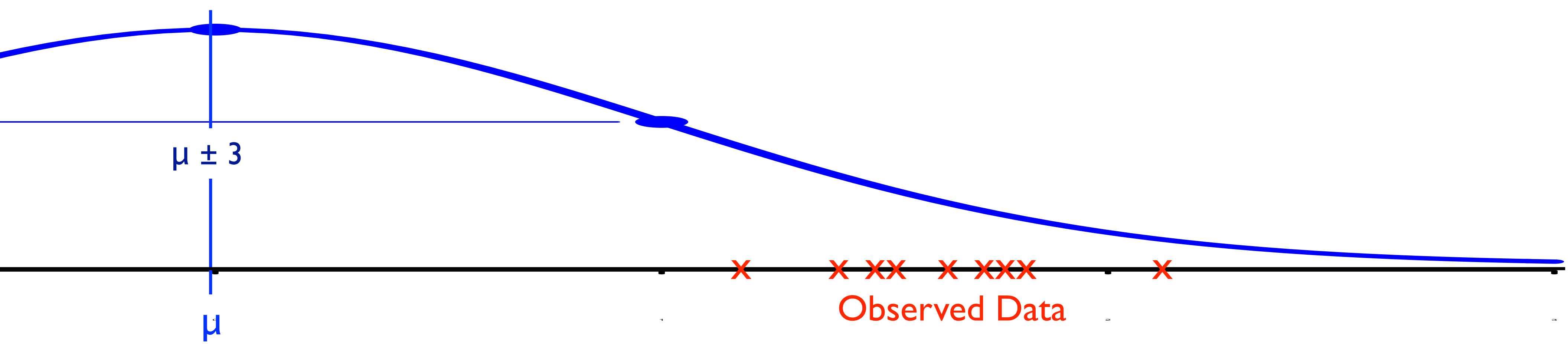
Which is more likely: (a) this?

μ, σ^2 both unknown



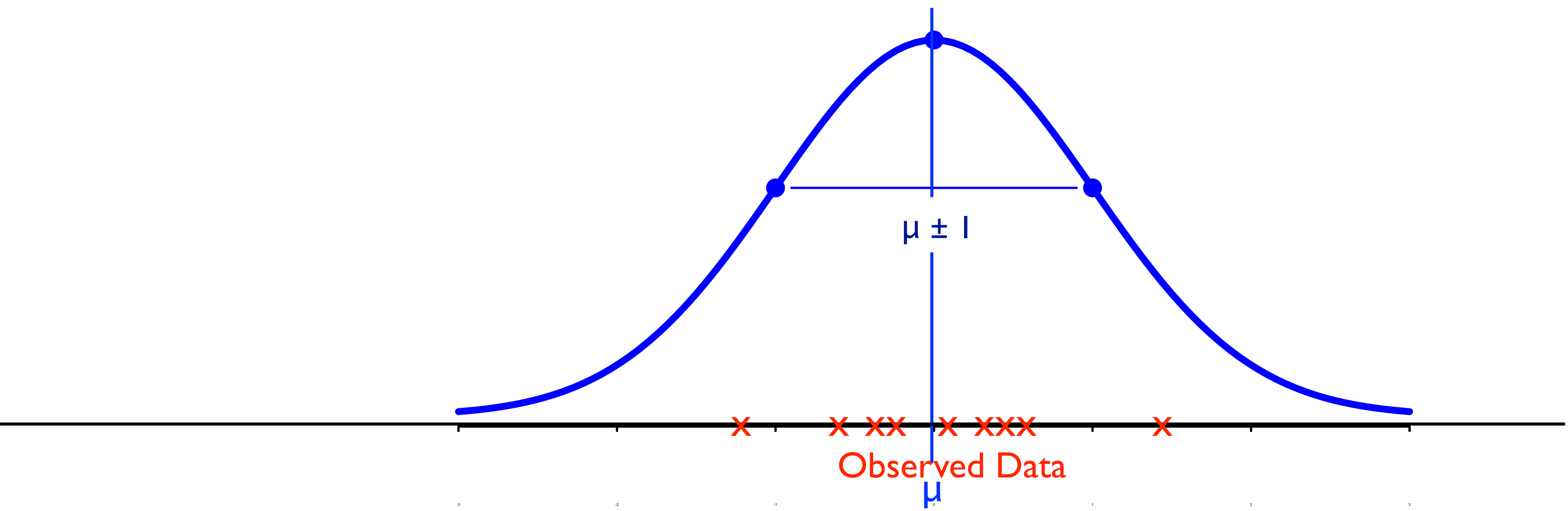
Which is more likely: (b) or this?

μ, σ^2 both unknown



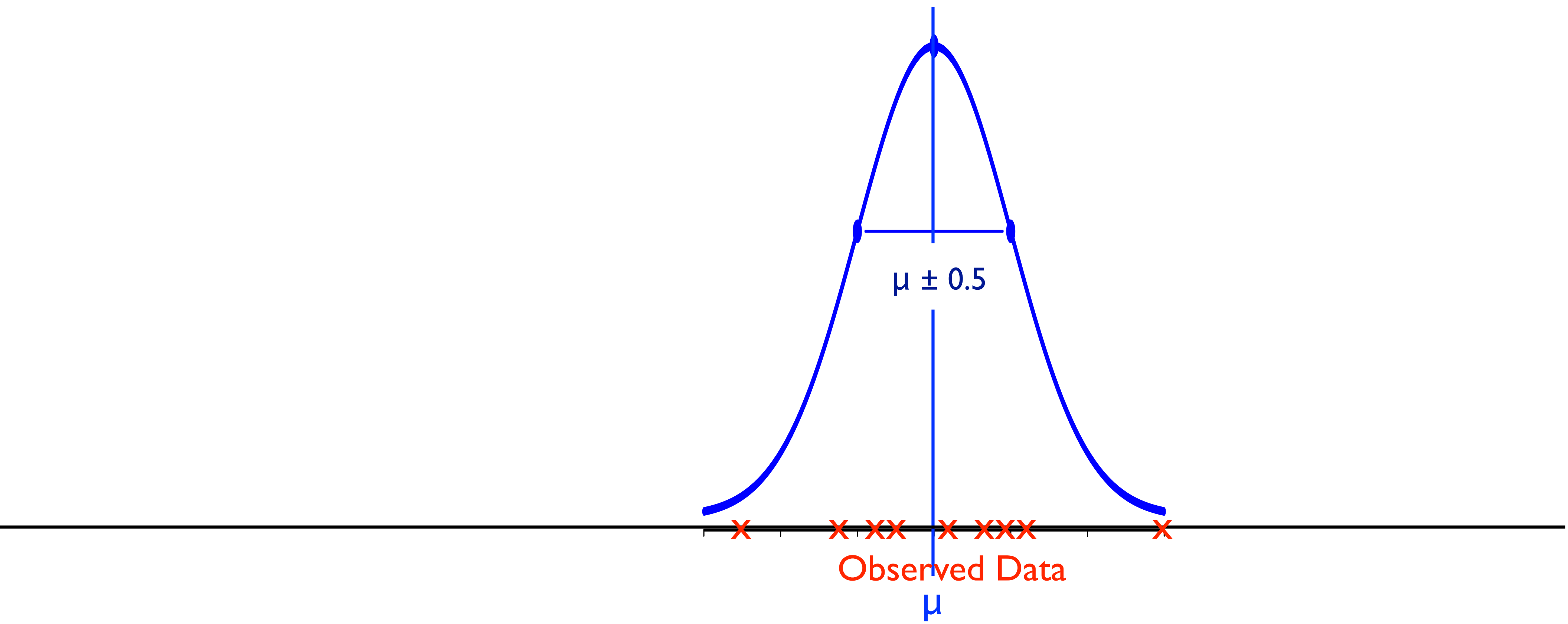
Which is more likely: (c) or this?

μ, σ^2 both unknown



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

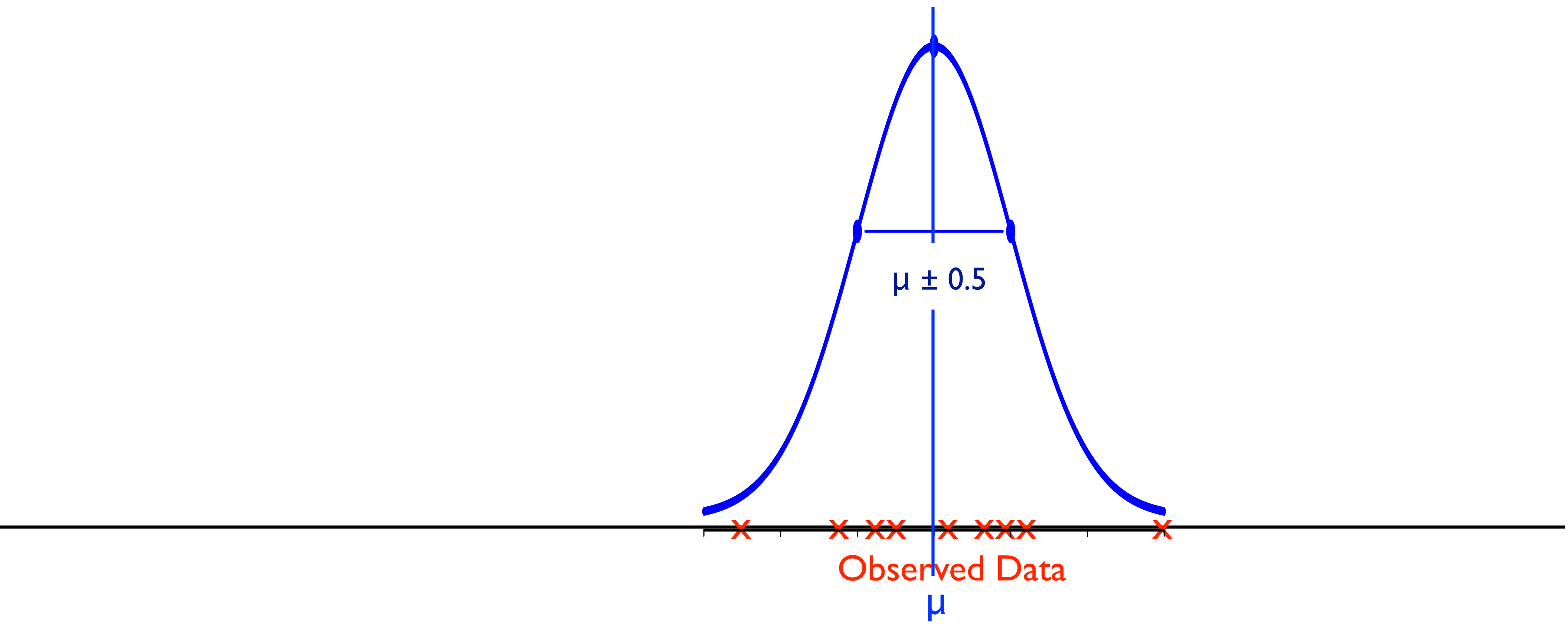
μ, σ^2 both unknown



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

μ, σ^2 both unknown

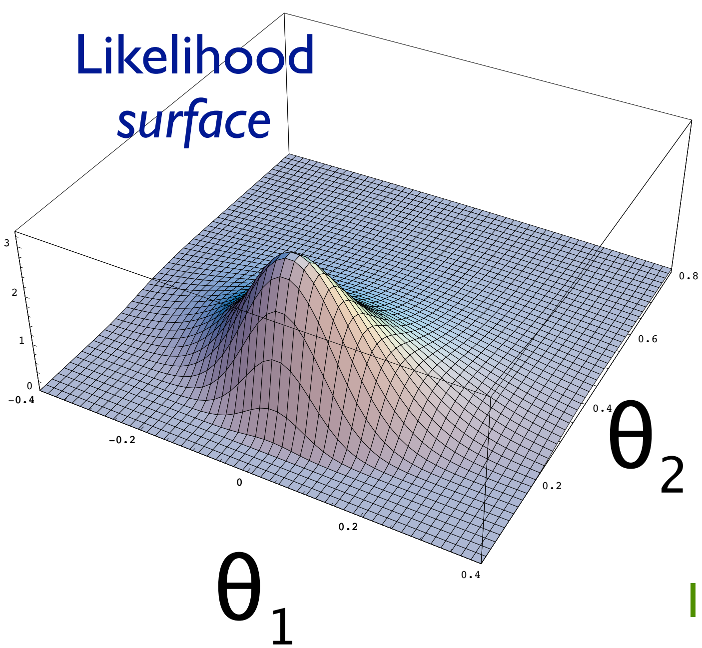
Looks good by eye, but how do I optimize my estimates of μ & σ^2 ?



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

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

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



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

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.)

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

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

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

Sample variance is MLE of *population* variance

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:

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

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, consider 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 (probability 0, in fact) 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)

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 $(d/d\theta)(\log \text{Likelihood}) = 0$

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*

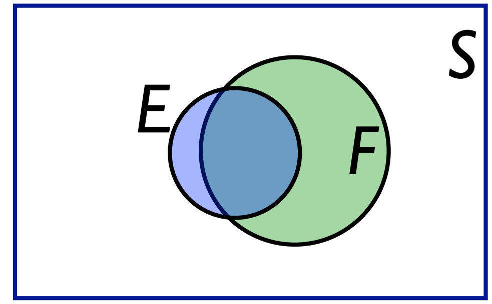
Conditional Probability
&
Bayes Rule

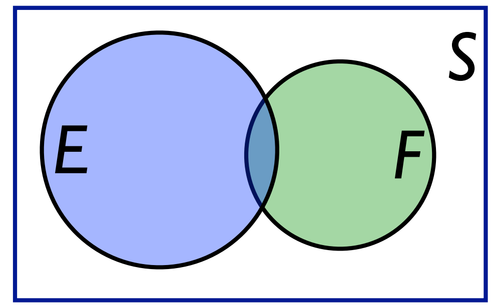
Conditional probability of E given F: probability that E occurs *given* that F has occurred.

"Conditioning on F"

Written as P(E|F)

Means "P(E has happened, given F observed)"



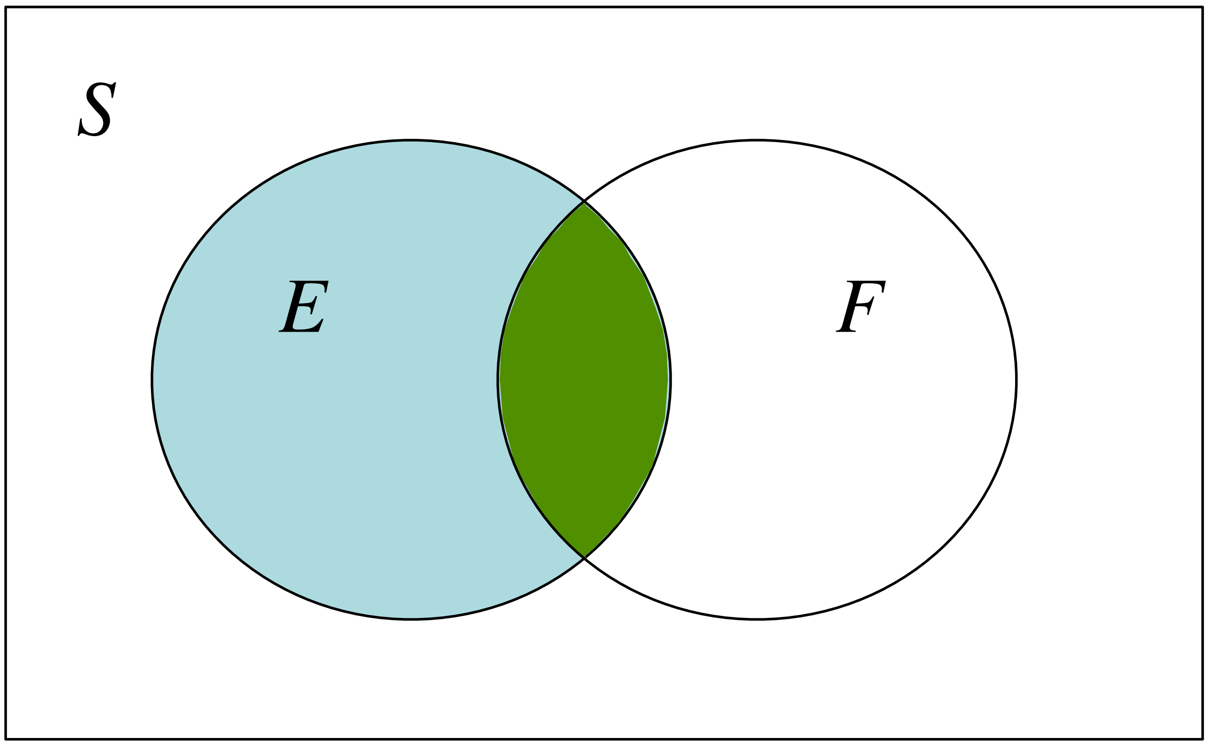


$$P(E \mid F) = \frac{P(EF)}{P(F)}$$

where P(F) > 0

E and F are events in the sample space S

$$E = EF \cup EF^c$$



$$EF \cap EF^c = \emptyset$$

$$\Rightarrow P(E) = P(EF) + P(EF^c)$$

Most common form:

$$P(F \mid E) = \frac{P(E \mid F)P(F)}{P(E)}$$

Expanded form (using law of total probability):

$$P(F \mid E) = \frac{P(E \mid F)P(F)}{P(E \mid F)P(F) + P(E \mid F^c)P(F^c)}$$

Proof:

$$P(F \mid E) = \frac{P(EF)}{P(E)} = \frac{P(E \mid F)P(F)}{P(E)}$$

EM

The Expectation-Maximization Algorithm
(for a Two-Component Gaussian Mixture)

A Hat Trick

Two slips of paper in a hat:

Pink: $\mu = 3$, and

Blue: $\mu = 7$.

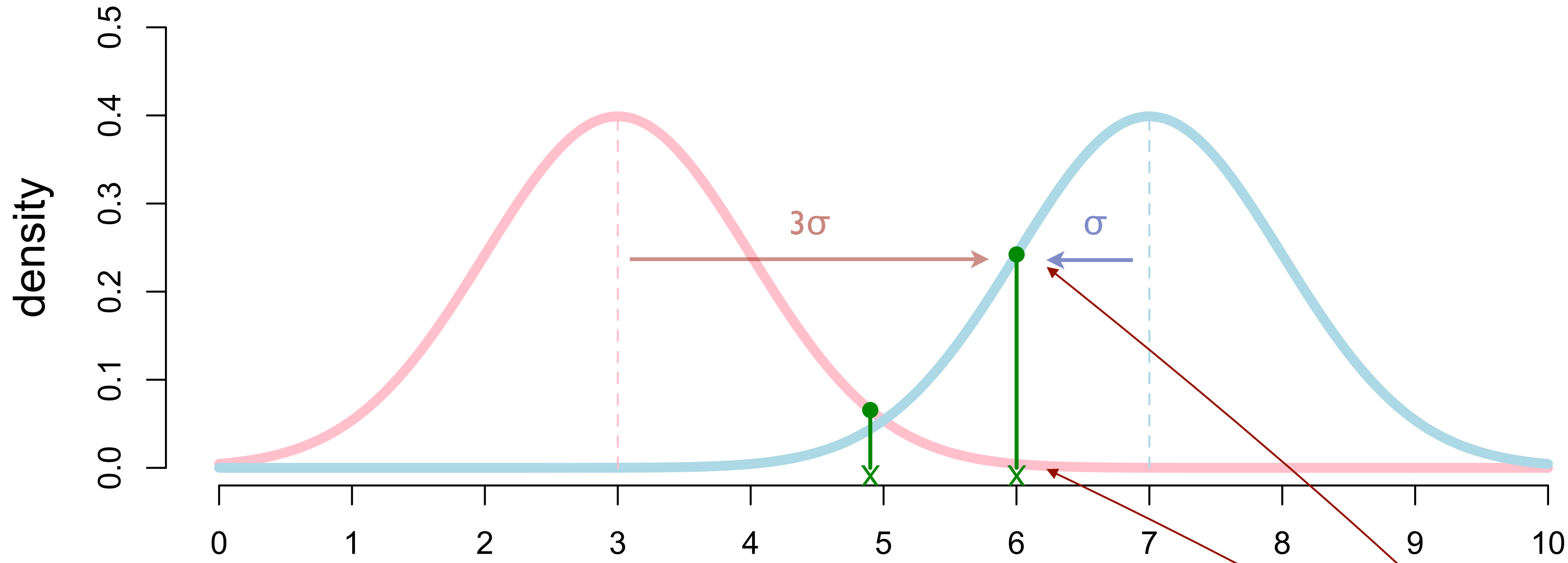
You draw one, then (without revealing color or μ) reveal a single sample $X \sim \text{Normal}(\text{mean } \mu, \sigma^2 = 1)$.

You happen to draw $X = 6.001$.

Dr. Mean says “your slip = 7.” What is P(correct)?

What if X had been 4.9?

A Hat Trick



Let "$X \approx 6$" be a shorthand for $6.001 - \delta/2 < X < 6.001 + \delta/2$

$$P(\mu = 7 | X = 6) = \lim_{\delta \to 0} P(\mu = 7 | X \approx 6)$$

$$P(\mu = 7 | X \approx 6) = \frac{P(X \approx 6 | \mu = 7) P(\mu = 7)}{P(X \approx 6)}$$

Bayes rule

$$= \frac{0.5 P(X \approx 6 | \mu = 7)}{0.5 P(X \approx 6 | \mu = 3) + 0.5 P(X \approx 6 | \mu = 7)}$$

$$\approx \frac{f(X = 6 | \mu = 7)\delta}{f(X = 6 | \mu = 3)\delta + f(X = 6) | \mu = 7)\delta}, \text{ so}$$

$$P(\mu = 7 | X = 6) = \frac{f(X = 6 | \mu = 7)}{f(X = 6 | \mu = 3) + f(X = 6) | \mu = 7)} \approx 0.982$$

Another Hat Trick

Two secret numbers, μ_{pink} and μ_{blue}

On pink slips, many samples of Normal(μ_{pink}, $\sigma^2 = 1$),

Ditto on blue slips, from Normal(μ_{blue}, $\sigma^2 = 1$).

Based on 16 of each, how would you "guess" the secrets (where "success" means your guess is within ±0.5 of each secret)?

Roughly how likely is it that you will succeed?

Another Hat Trick (cont.)

Pink/blue = red herrings; separate & independent

Given $X_1, \ldots, X_{16} \sim N(\mu, \sigma^2)$, $\sigma^2 = 1$

Calculate $Y = (X_1 + \ldots + X_{16})/16 \sim N(\ ?\ ,\ ?\ )$

$E[Y] = \mu$

$Var(Y) = 16\sigma^2/16^2 = \sigma^2/16 = 1/16$

I.e., X_i's are all $\sim N(\mu, 1)$; Y is $\sim N(\mu, 1/16)$

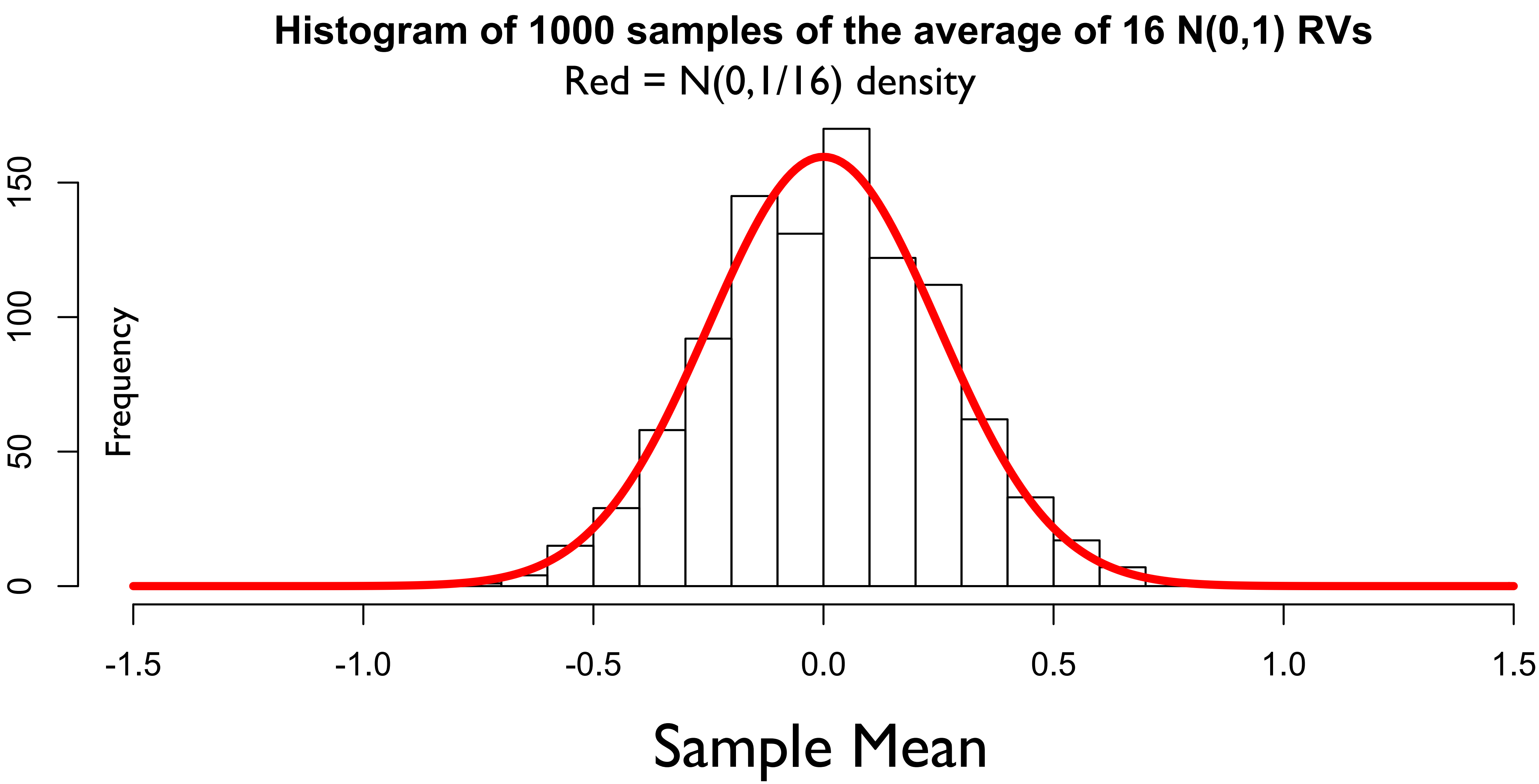
and since 0.5 = 2 sqrt(1/16), we have:

"Y within ±.5 of μ" = "Y within ±2 σ of μ" ≈ 95% prob

Note 1: Y is a *point estimate* for μ;

Y ± 2 σ is a *95% confidence interval* for μ

(More on this topic later)



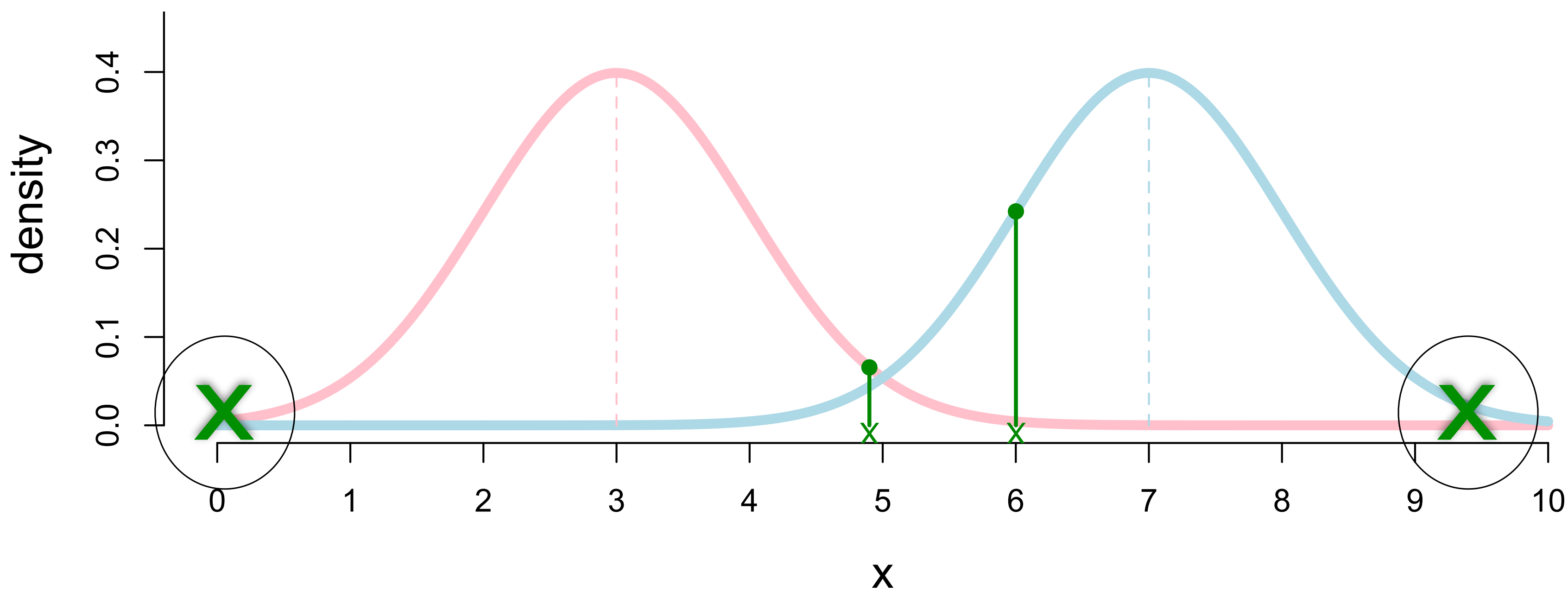
Histogram of 1000 samples of the average of 16 N(0,1) RVs
Red = N(0,1/16) density
Frequency
0
50
100
150
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Sample Mean

Hat Trick 2 (cont.)

Note 2:

What would you do if some of the slips you pulled had coffee spilled on them, obscuring color?

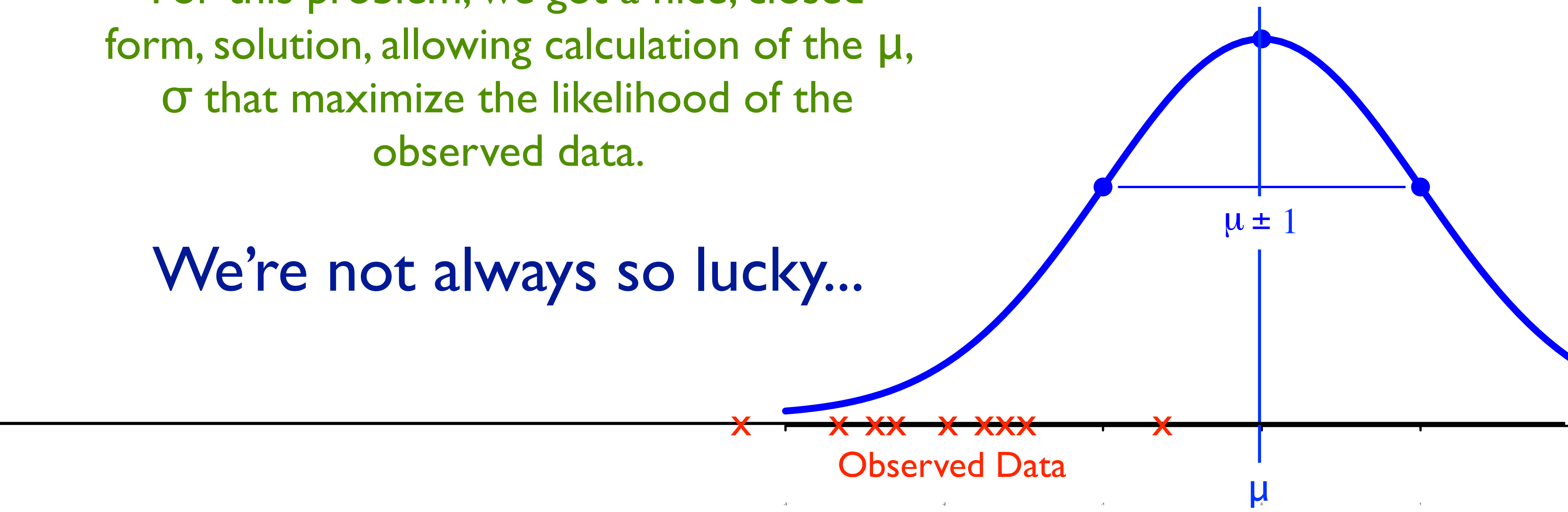
- If they were half way between means of the others?
- If they were on opposite sides of the means of the others



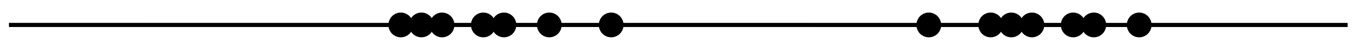
Previously:
How to estimate μ given data

For this problem, we got a nice, closed form, solution, allowing calculation of the μ, σ that maximize the likelihood of the observed data.

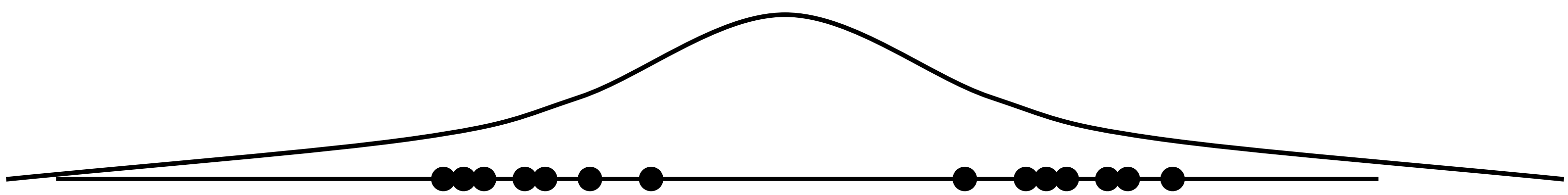
We're not always so lucky...



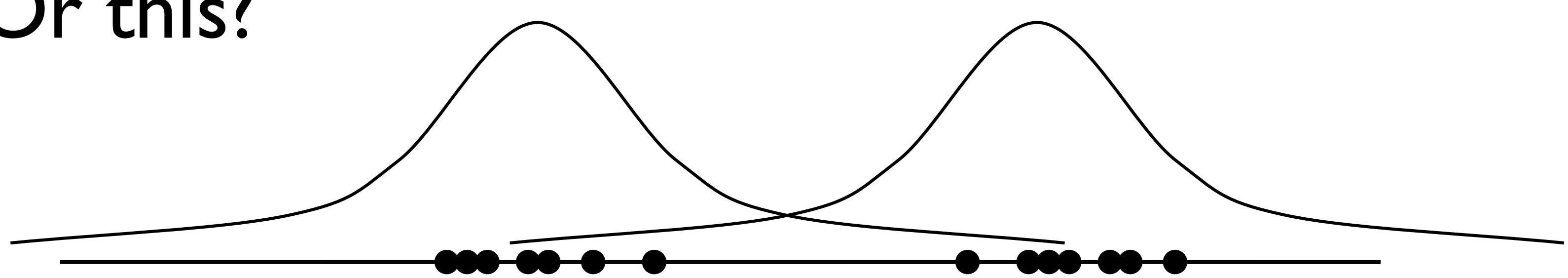
More Complex Example



This?

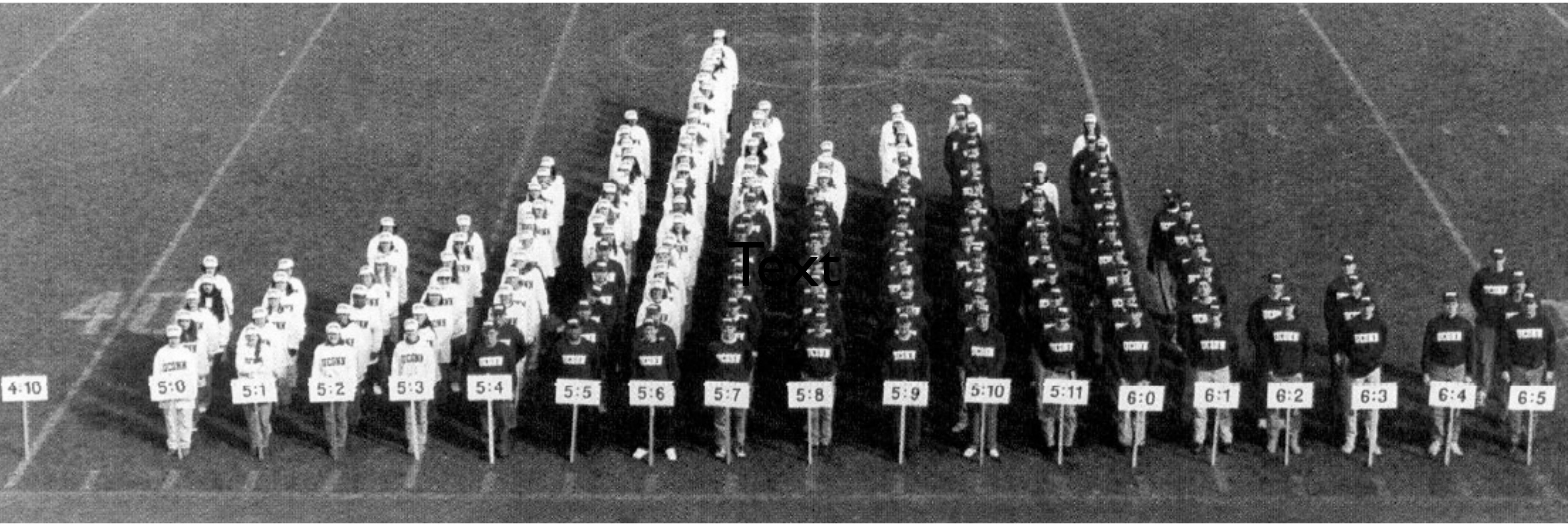


Or this?



(A modeling decision, not a math problem...,
but if the later, what math?)

A Living Histogram

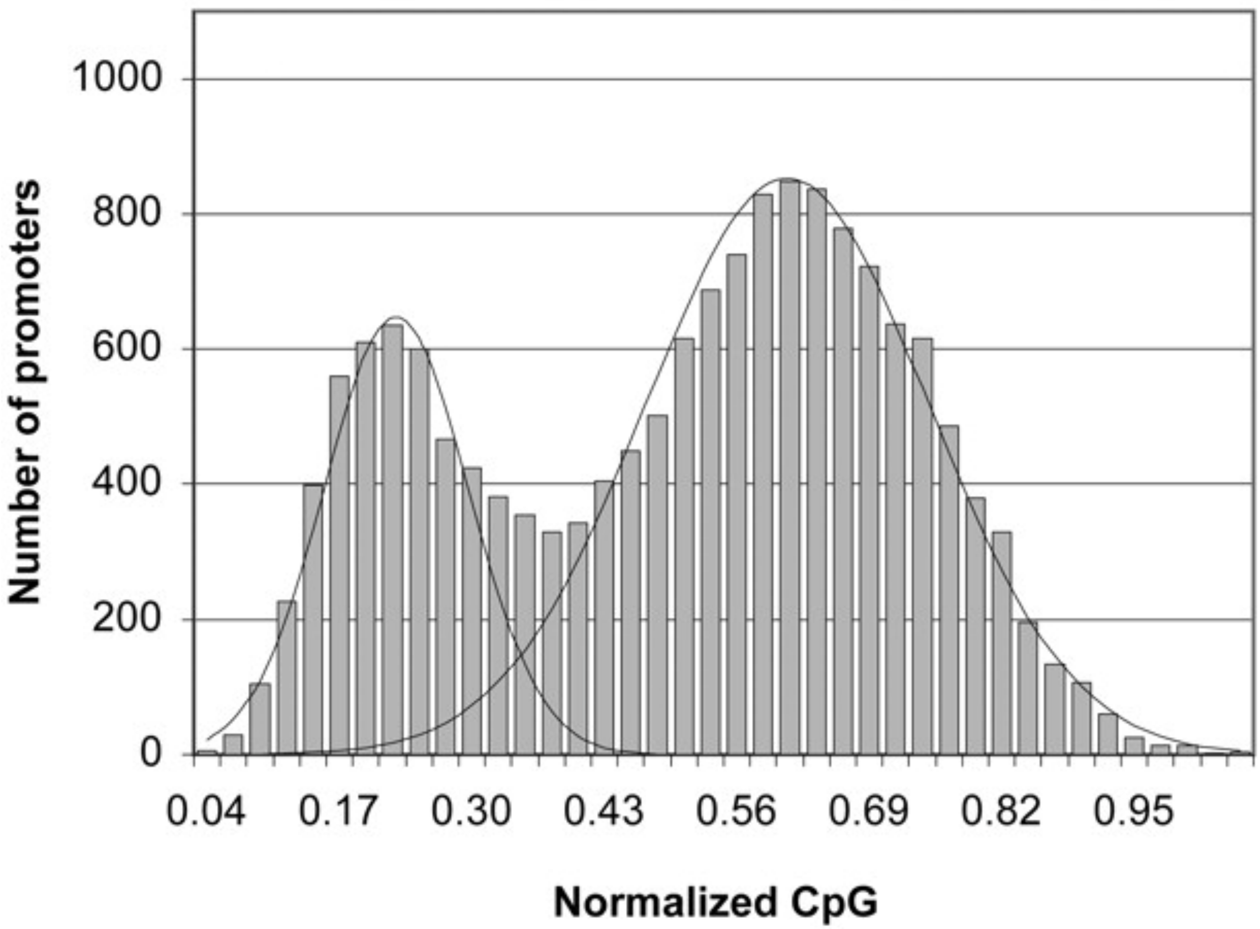


male and female genetics students, University of Connecticut in 1996

http://mindprod.com/jgloss/histogram.html

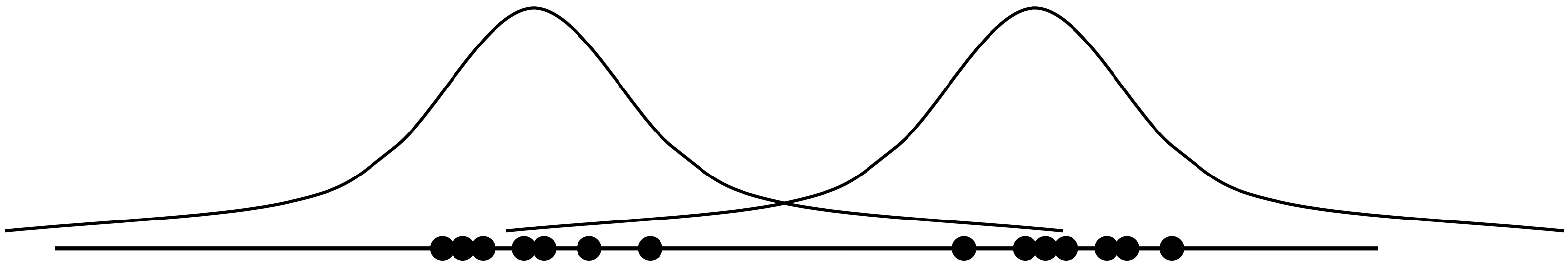
Another Real Example:

CpG content of human gene promoters



"A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters" Saxonov, Berg, and Brutlag, PNAS 2006;103:1412-1417

Gaussian Mixture Models / Model-based Clustering



Parameters θ

means	μ_1	μ_2
variances	σ_1^2	σ_2^2
mixing parameters	τ_1	$\tau_2 = 1 - \tau_1$

P.D.F. —separately→ $f(x|\mu_1, \sigma_1^2)$ $f(x|\mu_2, \sigma_2^2)$

—together→

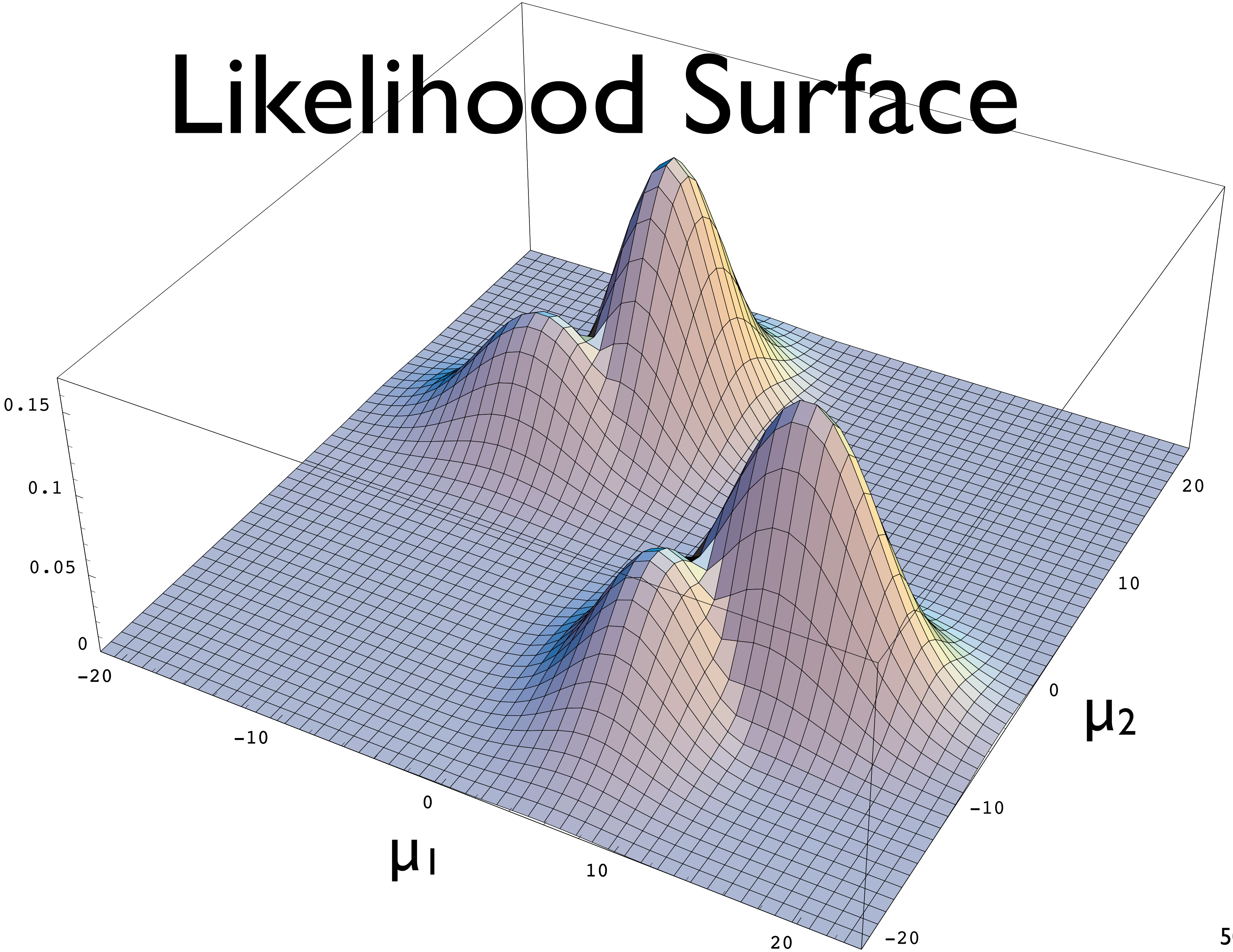
Likelihood $\tau_1 f(x|\mu_1, \sigma_1^2) + \tau_2 f(x|\mu_2, \sigma_2^2)$

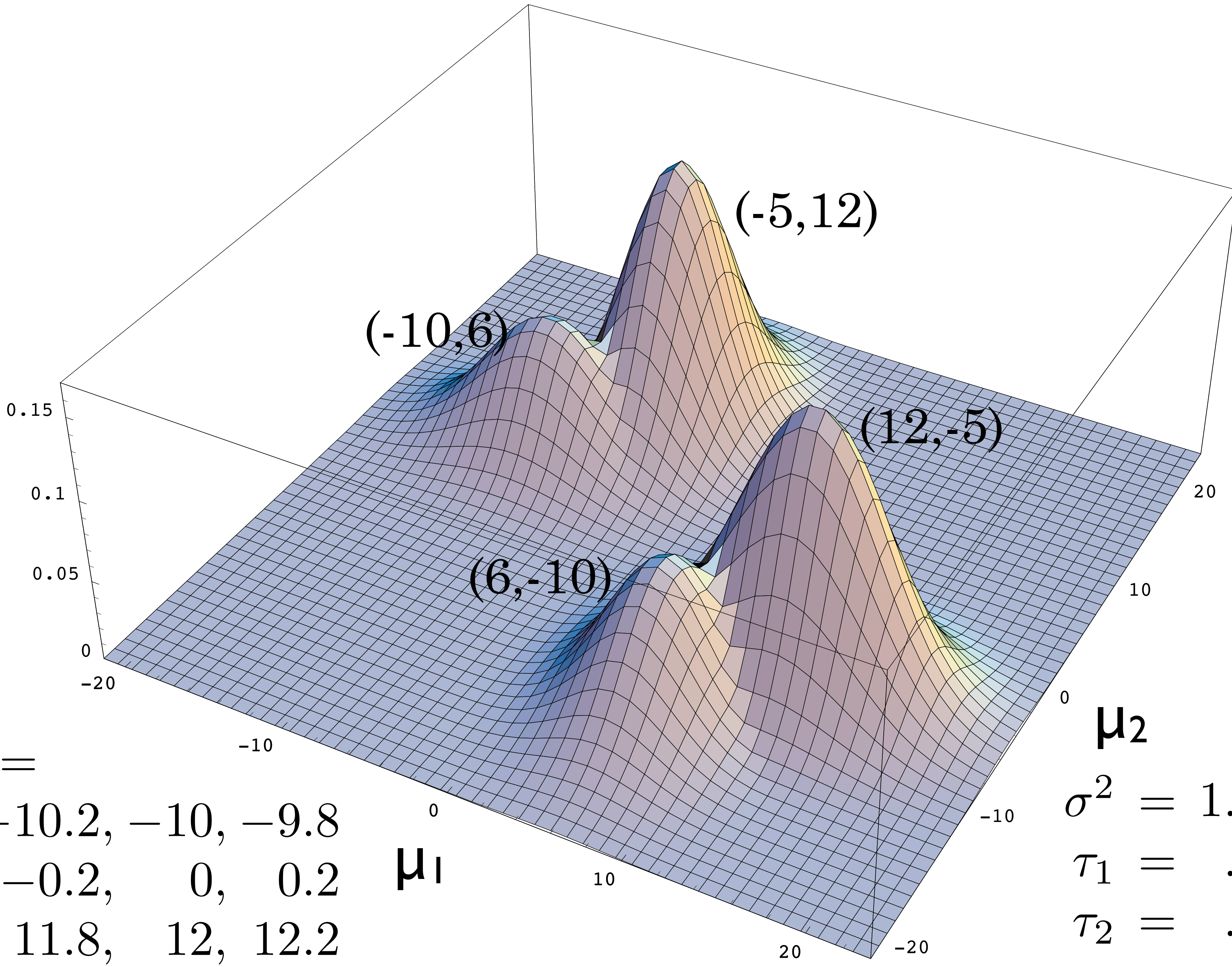
$$L(x_1, x_2, \ldots, x_n | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

No closed-form max

Likelihood Surface





$$x_i = \begin{array}{rrr} -10.2, & -10, & -9.8 \\ -0.2, & 0, & 0.2 \\ 11.8, & 12, & 12.2 \end{array}$$

$$\sigma^2 = 1.0 \quad \tau_1 = .5 \quad \tau_2 = .5$$

A What-If Puzzle

Likelihood

$$L(x_1, x_2, \ldots, x_n | \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

Messy: no closed form solution known for finding θ maximizing L

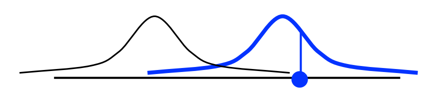
But *what if* we knew the *hidden data*?

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

EM as Egg vs Chicken

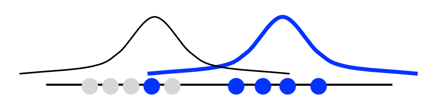
Hat Trick 1

- *IF* parameters θ known, could estimate z_{ij}
  - E.g., $|x_i - \mu_1|/\sigma_1 \gg |x_i - \mu_2|/\sigma_2 \Rightarrow P[z_{i1}=1] \ll P[z_{i2}=1]$



Hat Trick 2

- *IF* z_{ij} known, could estimate parameters θ
  - E.g., only points in cluster 2 influence μ_2, σ_2



- But we know neither; (optimistically) iterate:

Hat Trick 1

  - E-step: calculate expected z_{ij}, given parameters

Hat Trick 2

  - M-step: calculate "MLE" of parameters, given $E(z_{ij})$

- Overall, a clever "hill-climbing" strategy

Not what's needed for homework, but may help clarify concepts

Simple Version: "Classification EM"

If $E[z_{ij}] < .5$, pretend $z_{ij} = 0$; $E[z_{ij}] > .5$, pretend it's 1

I.e., *classify* points as component 1 or 2

Now recalc θ, assuming that partition (standard MLE)

Then recalc $E[z_{ij}]$, assuming that θ

Then re-recalc θ, assuming new $E[z_{ij}]$, etc., etc.

"K-means clustering," essentially

"Full EM" is slightly more involved, (to account for uncertainty in classification) but this is the crux.

Another contrast: HMM parameter estimation via "Viterbi" vs "Baum-Welch" training. In both, "hidden data" is "which state was it in at each step?" Viterbi is like E-step in classification EM: it makes a single state prediction. B-W is full EM: it captures the uncertainty in state prediction, too. For either, M-step maximizes HMM emission/transition probabilities, assuming those fixed states (Viterbi) / uncertain states (B-W).

Full EM

x_i's are known; θ unknown. Goal is to find MLE θ of:

$$L(x_1, \ldots, x_n \mid \theta)$$ (hidden data likelihood)

Would be easy *if* z_{ij}'s were known, i.e., consider:

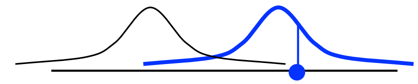
$$L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)$$ (complete data likelihood)

But z_{ij}'s aren't known.

Instead, maximize *expected* likelihood of visible data

$$E(L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)),$$

where expectation is over distribution of hidden data (z_{ij}'s)



The E-step:

Find $E(z_{ij})$, i.e., $P(z_{ij}=1)$

Assume θ known & fixed

A (B): the event that x_i was drawn from f_1 (f_2)

D: the observed datum x_i

Expected value of z_{i1} is P(A|D) — $E = 0 \cdot P(0) + 1 \cdot P(1)$

$$E[z_{i1}] = P(A|D) = \frac{P(D|A)P(A)}{P(D)}$$

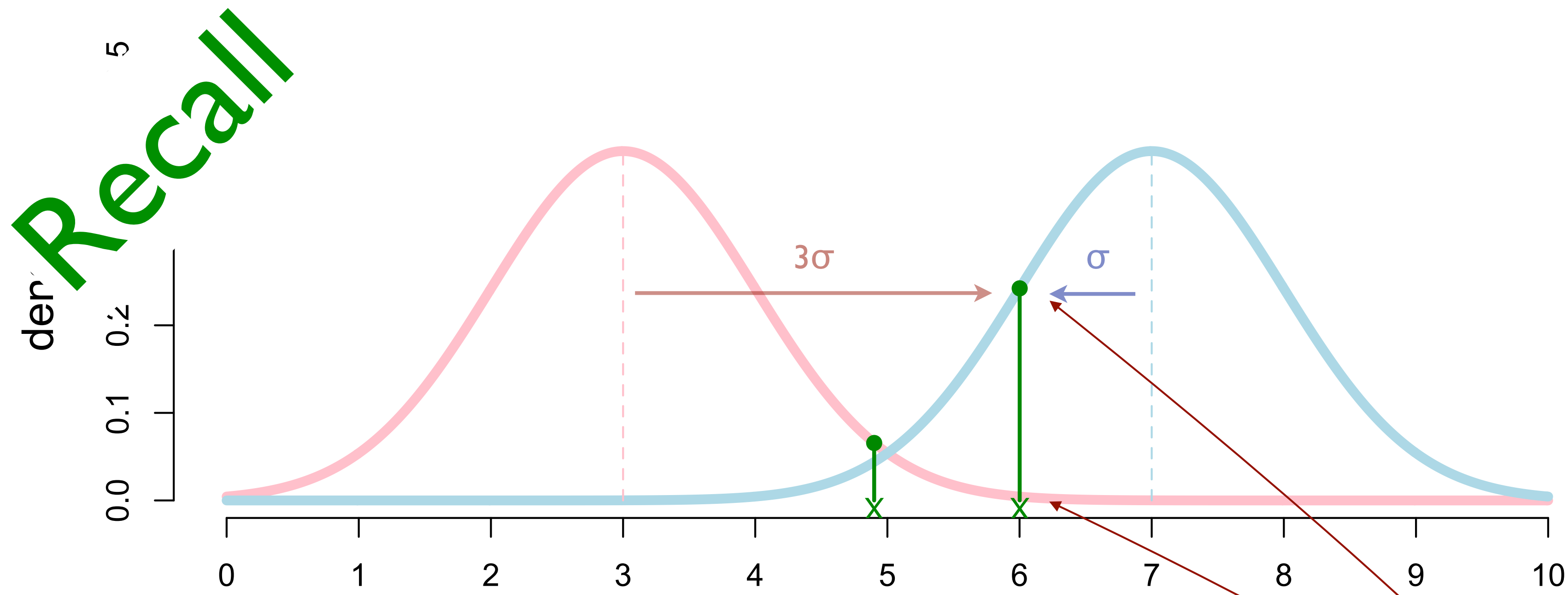
$$P(D) = P(D|A)P(A) + P(D|B)P(B)$$

$$= f_1(x_i|\theta_1)\,\tau_1 + f_2(x_i|\theta_2)\,\tau_2$$

Repeat for each x_i

Note: denominator = sum of numerators – i.e. that which normalizes sum to 1 (typical Bayes)

A Hat Trick



Let "$X \approx 6$" be a shorthand for $6.001 - \delta/2 < X < 6.001 + \delta/2$

$$P(\mu = 7|X = 6) = \lim_{\delta \to 0} P(\mu = 7|X \approx 6)$$

$$P(\mu = 7|X \approx 6) = \frac{P(X \approx 6|\mu = 7)P(\mu = 7)}{P(X \approx 6)}$$

$$= \frac{0.5P(X \approx 6|\mu = 7)}{0.5P(X \approx 6|\mu = 3) + 0.5P(X \approx 6|\mu = 7)}$$

$$\approx \frac{f(X = 6|\mu = 7)\delta}{f(X = 6|\mu = 3)\delta + f(X = 6)|\mu = 7)\delta}, \text{ so}$$

$$P(\mu = 7|X = 6) = \frac{f(X = 6|\mu = 7)}{f(X = 6|\mu = 3) + f(X = 6)|\mu = 7)} \approx 0.982$$

f = normal density

Complete Data Likelihood

Recall:

$$z_{1j} = \begin{cases} 1 & \text{if } x_1 \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

so, correspondingly,

$$L(x_1, z_{1j} \mid \theta) = \begin{cases} \tau_1 f_1(x_1 \mid \theta) & \text{if } z_{11} = 1 \\ \tau_2 f_2(x_1 \mid \theta) & \text{otherwise} \end{cases}$$

equal, if z_{ij} are 0/1

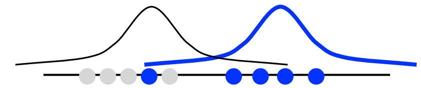
Formulas with "if's" are messy; can we blend more smoothly?
Yes, many possibilities. Idea 1:

$$L(x_1, z_{1j} \mid \theta) = z_{11} \cdot \tau_1 f_1(x_1 \mid \theta) + z_{12} \cdot \tau_2 f_2(x_1 \mid \theta)$$

Idea 2 (Better):

$$L(x_1, z_{1j} \mid \theta) = (\tau_1 f_1(x_1 \mid \theta))^{z_{11}} \cdot (\tau_2 f_2(x_1 \mid \theta))^{z_{12}}$$

M-step:



Find θ maximizing E(log(Likelihood))

(For simplicity, assume $\sigma_1 = \sigma_2 = \sigma; \tau_1 = \tau_2 = \tau = 0.5$)

$$L(\vec{x}, \vec{z} \mid \theta) = \prod_{i=1}^{n} \frac{\tau}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{j=1}^{2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$E[\log L(\vec{x}, \vec{z} \mid \theta)] = E\left[\sum_{i=1}^{n} \left(\log \tau - \frac{1}{2} \log(2\pi\sigma^2) - \sum_{j=1}^{2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right) \right]$$

wrt dist of z_{ij}

$$= \sum_{i=1}^{n} \left(\log \tau - \frac{1}{2} \log(2\pi\sigma^2) - \sum_{j=1}^{2} E[z_{ij}] \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

Find θ maximizing this as before, using $E[z_{ij}]$ found in E-step. Result:

$\mu_j = \sum_{i=1}^{n} E[z_{ij}] x_i / \sum_{i=1}^{n} E[z_{ij}]$ (intuit: avg, weighted by subpop prob)

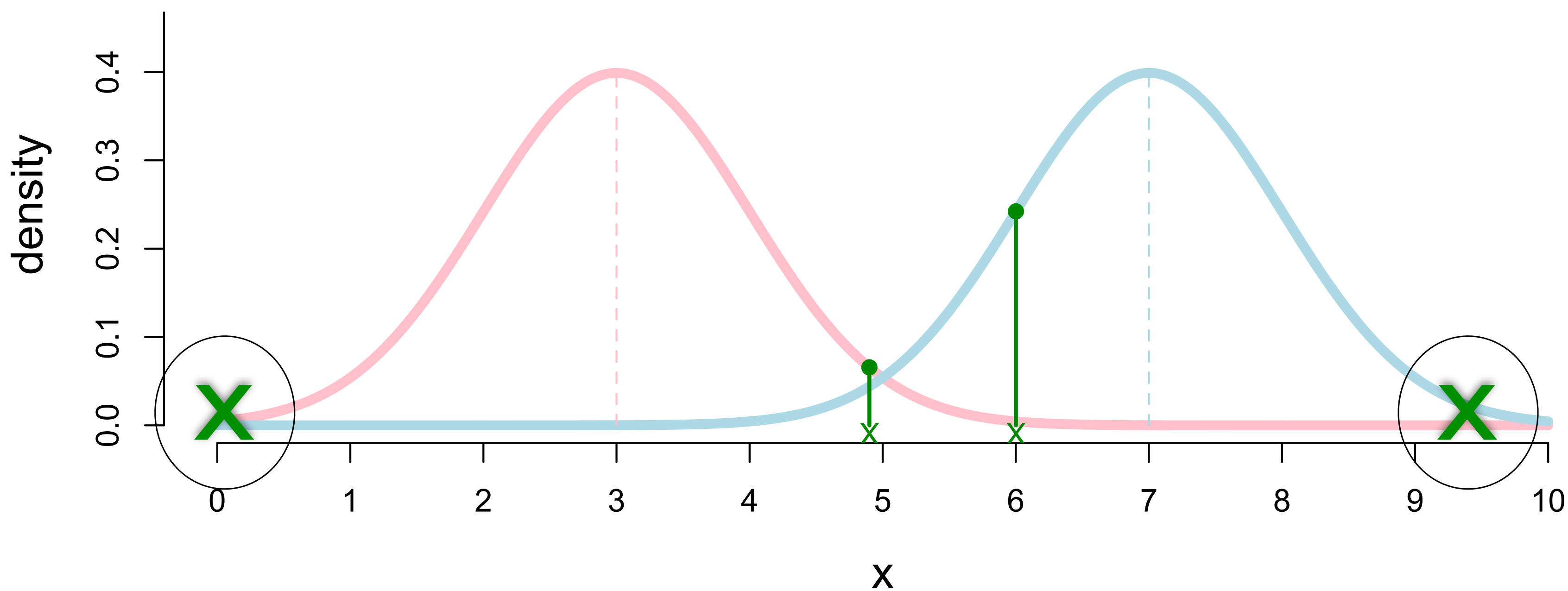
Recall

Hat Trick 2 (cont.)

Note 2: red/blue separation is just like the M-step of EM *if values of the hidden variables (z_{ij}) were known.*

What if they're not? E.g., what would you do if some of the slips you pulled had coffee spilled on them, obscuring color?

If they were half way between means of the others?
If they were on opposite sides of the means of the others



M-step:calculating mu's

$$\mu_j = \sum_{i=1}^{n} E[z_{ij}]x_i / \sum_{i=1}^{n} E[z_{ij}]$$

In words: μ_j is the average of the observed x_i's, weighted by the probability that x_i was sampled from component j.

old E's

							row sum	avg
$E[z_{i1}]$	0.99	0.98	0.7	0.2	0.03	0.01	2.91	
$E[z_{i2}]$	0.01	0.02	0.3	0.8	0.97	0.99	3.09	
x_i	9	10	11	19	20	21	90	15
$E[z_{i1}]x_i$	8.9	9.8	7.7	3.8	0.6	0.2	31.0	10.66
$E[z_{i1}]x_i$	0.1	0.2	3.3	15.2	19.4	20.8	59.0	19.09

new μ's

2 Component Mixture

$$\sigma_1 = \sigma_2 = 1; \ \tau = 0.5$$

			mu1	-20.00		-6.00		-5.00		-4.99
			mu2	6.00		0.00		3.75		3.75
x1	**-6**		**z11**		5.11E-12		1.00E+00		1.00E+00	
x2	**-5**		**z21**		2.61E-23		1.00E+00		1.00E+00	
x3	**-4**		**z31**		1.33E-34		9.98E-01		1.00E+00	
x4	**0**		**z41**		9.09E-80		1.52E-08		4.11E-03	
x5	**4**		**z51**		6.19E-125		5.75E-19		2.64E-18	
x6	**5**		**z61**		3.16E-136		1.43E-21		4.20E-22	
x7	**6**		**z71**		1.62E-147		3.53E-24		6.69E-26	

Essentially converged in 2 iterations

(Excel spreadsheet on course web)

EM Summary

Fundamentally a maximum likelihood parameter estimation problem; broader than just Gaussian

Useful if 0/1 hidden data, and if analysis would be more tractable if hidden data z were known

Iterate:

E-step: estimate E(z) for each z, given θ Bayes

M-step: estimate θ maximizing E[log likelihood] given E[z] [where "E[logL]" is wrt random z ~ E[z] = p(z=1)] MLE

EM Issues

Under mild assumptions (DEKM sect 11.6), EM is guaranteed to increase likelihood with every E-M iteration, hence will *converge*.

But it may converge to a *local*, not global, max. (Recall the 4-bump surface...)

Issue is intrinsic (probably), since EM is often applied to *NP-hard* problems (including clustering, above and motif-discovery, soon)

Nevertheless, widely used, often effective

Applications

Clustering is a remarkably successful exploratory data analysis tool

- Web-search, information retrieval, gene-expression, ...
- Model-based approach above is one of the leading ways to do it

Gaussian mixture models widely used

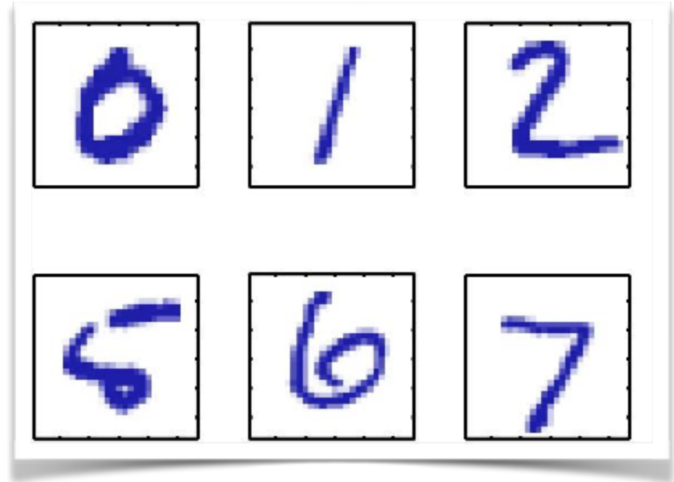
- With many components, empirically match arbitrary distribution
- Often well-justified, due to “hidden parameters” driving the visible data

EM is extremely widely used for “hidden-data” problems

- Hidden Markov Models – speech recognition, DNA analysis, ...

A "Machine Learning" Example

Handwritten Digit Recognition



Given: 10^4 unlabeled, scanned images of handwritten digits, say 25 x 25 pixels,

Goal: automatically classify new examples

Possible Method:

Each image is a point in $\mathbb{R}^{625}$; the "ideal" 7, say, is one such point; model other 7's as a Gaussian cloud around it

Do EM, as above, but 10 components in 625 dimensions instead of 2 components in 1 dimension

"Recognize" a new digit by best fit to those 10 models, i.e., basically max E-step probability

Relative entropy

Relative Entropy

- AKA Kullback-Liebler Distance/Divergence, AKA Information Content
- Given distributions P, Q

$$H(P||Q) = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)}$$

Notes:

$$\text{Let } P(x) \log \frac{P(x)}{Q(x)} = 0 \text{ if } P(x) = 0 \text{ [since } \lim_{y \to 0} y \log y = 0]$$

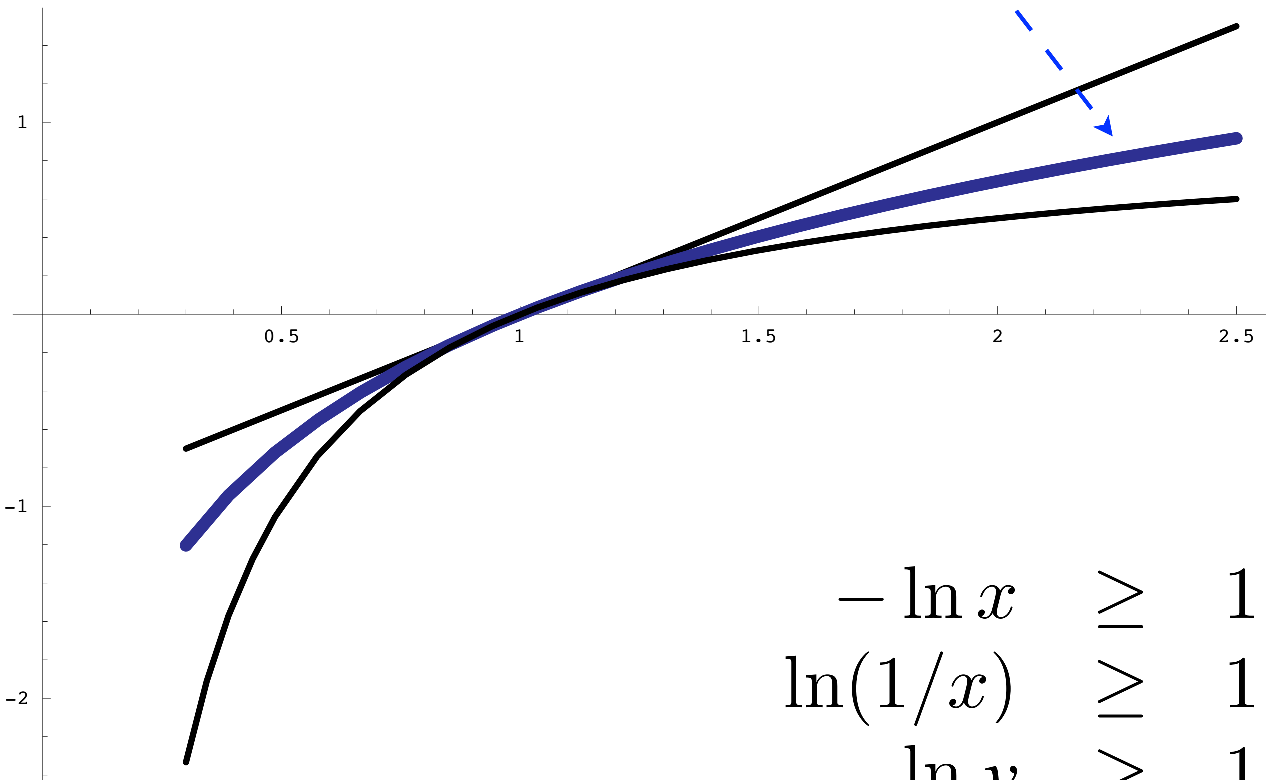
$$\text{Undefined if } 0 = Q(x) < P(x)$$

Relative Entropy

$$H(P||Q) = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)}$$

- Intuition: A quantitative measure of how much P "diverges" from Q. (Think "distance," but note it's not symmetric.)
  - If P ≈ Q everywhere, then log(P/Q) ≈ 0, so H(P||Q) ≈ 0
  - But as they differ more, sum is pulled above 0 (next 2 slides)
- What it means quantitatively: Suppose you sample x, but aren't sure whether you're sampling from P (call it the "null model") or from Q (the "alternate model"). Then log(P(x)/Q(x)) is the log likelihood ratio of the two models given that datum. H(P||Q) is the *expected per sample contribution to the log likelihood ratio* for discriminating between those two models.
- Exercise: if H(P||Q) = 0.1, say. Assuming Q is the correct model, how many samples would you need to confidently (say, with 1000:1 odds) reject P?

$$\ln x \quad \leq \quad x - 1$$



$$\begin{aligned} -\ln x &\geq 1 - x \\ \ln(1/x) &\geq 1 - x \\ \ln y &\geq 1 - 1/y \end{aligned}$$

$$(y = 1/x)$$

Theorem: $H(P||Q) \geq 0$

$$
\begin{aligned}
H(P||Q) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\
&\geq \sum_x P(x) \left(1 - \frac{Q(x)}{P(x)}\right) \\
&= \sum_x (P(x) - Q(x)) \\
&= \sum_x P(x) - \sum_x Q(x) \\
&= 1 - 1 \\
&= 0
\end{aligned}
$$

Idea: if P ≠ Q, then

P(x)>Q(x) ⇒ log(P(x)/Q(x))>0

and

P(y)<Q(y) ⇒ log(P(y)/Q(y))<0

Q: Can this pull H(P||Q) < 0?
A: No, as theorem shows.
Intuitive reason: sum is *weighted* by P(x), which is bigger at the positive log ratios vs the negative ones.

Furthermore: H(P||Q) = 0 if and only if P = Q
Bottom line: "bigger" means "more different"