Learning From Data: MLE

Maximum Likelihood Estimators

Parameter Estimation

$\boxed{\text{Ber}(p)}$

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

Goal: estimate θ.

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

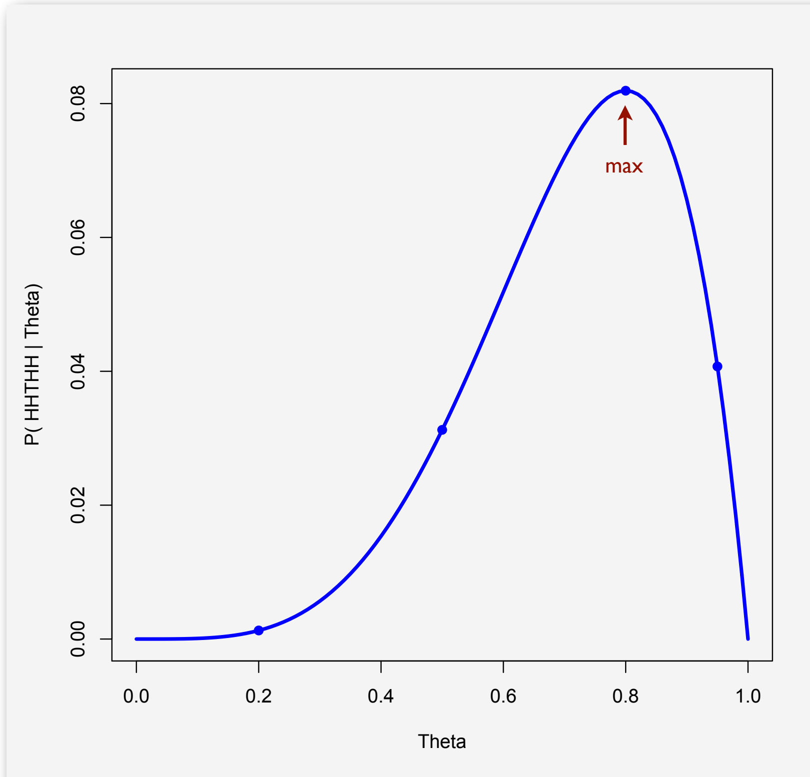
θ = probability of Heads

p(x|θ) is the Bernoulli probability mass function with parameter θ

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



Likelihood

outcome w/
x: $\boxed{12\ H\text{'s}\ ,\ 10\ T\text{'s}.}$

what value of θ maximizes this fn?

$\text{Likelihood}(x \mid \theta) = \theta^{12}(1-\theta)^{10}$

$P(x \mid \theta)$: 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?

Maximum Likelihood Parameter Estimation

One (of many) approaches to param. est.

Likelihood of (indp) observations $x_1, x_2, \ldots, x_n$

$$L(\underbrace{x_1, x_2, \ldots, x_n}_{\text{data}} \mid \underset{\uparrow}{\theta}) = \prod_{i=1}^{n} \underset{\text{Pr}}{f}(x_i \mid \theta)$$

param trying to est.

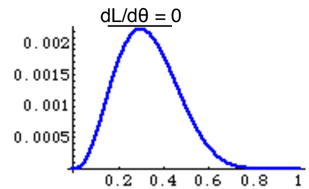
As a function of θ, what θ maximizes the likelihood of the data actually observed

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

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

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)

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

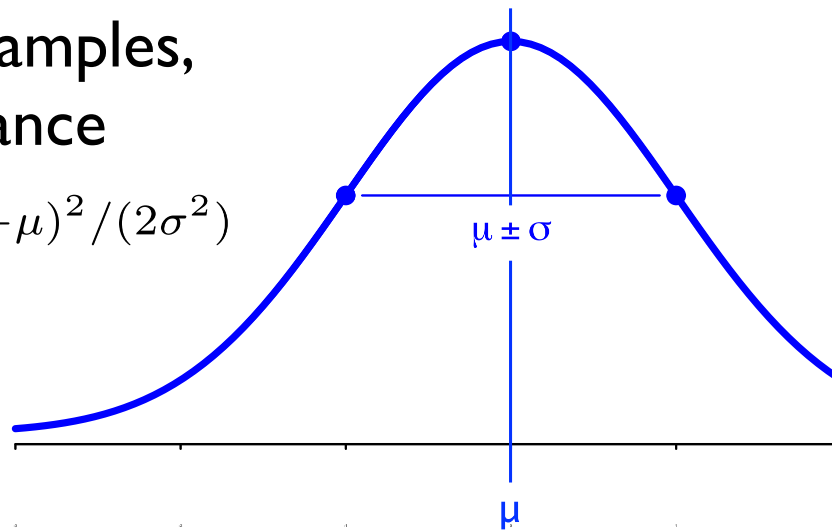
Parameter Estimation

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

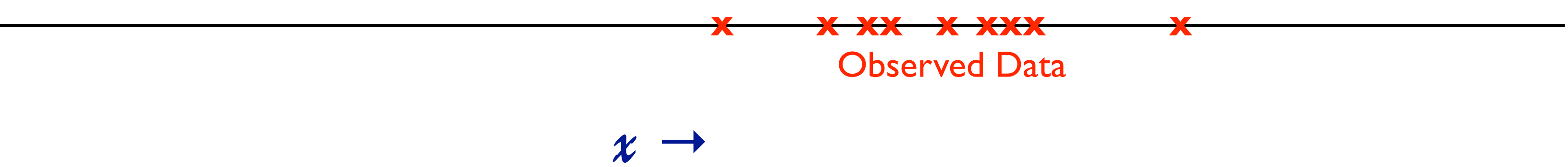
E.g.: Given *n normal* samples, estimate mean & variance

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

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



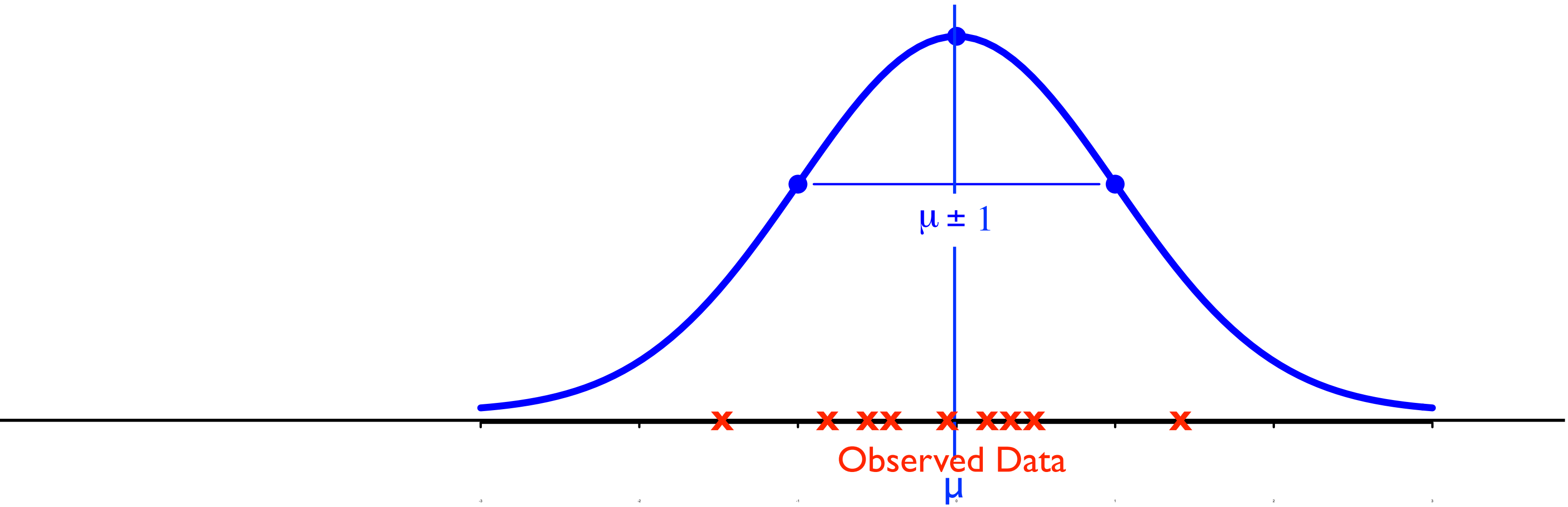
I got data; a little birdie tells me it's normal, and promises $\sigma^2 = 1$



Is the following likely?

μ unknown, σ^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

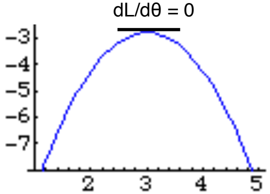
$$L(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2}$$

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

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

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

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



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

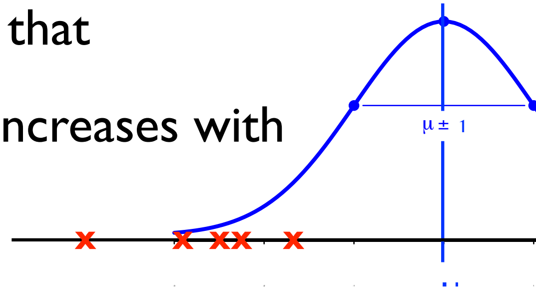
Sample mean is MLE of population mean

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

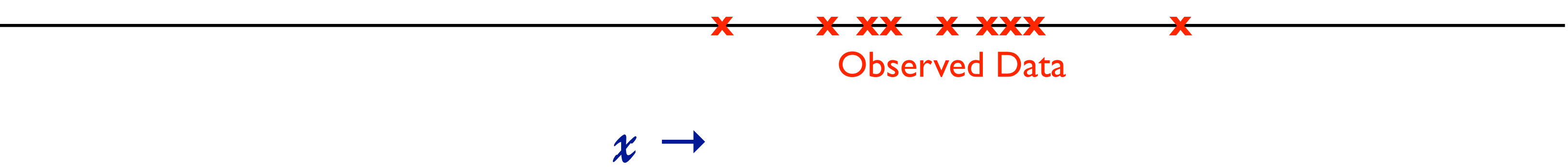
b) has desired property that likelihood increases with better fit to the model



and/or

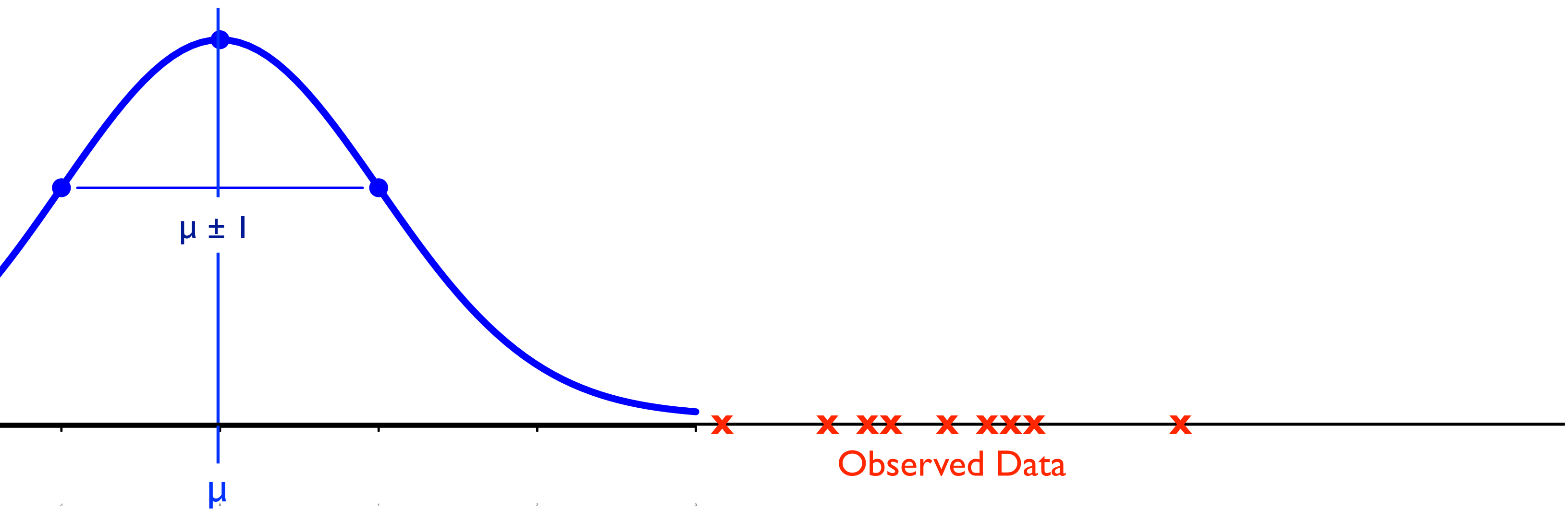
c) if density at x is $f(x)$, for any small $\delta>0$, the probability of a sample within $\pm\delta/2$ of x is $\approx \delta f(x)$, but δ is *constant* wrt θ, so it just drops out of $d/d\theta \log L(\ldots) = 0$.

Ex2: I got data; a little birdie tells me it's normal (but does *not* tell me μ, σ²)



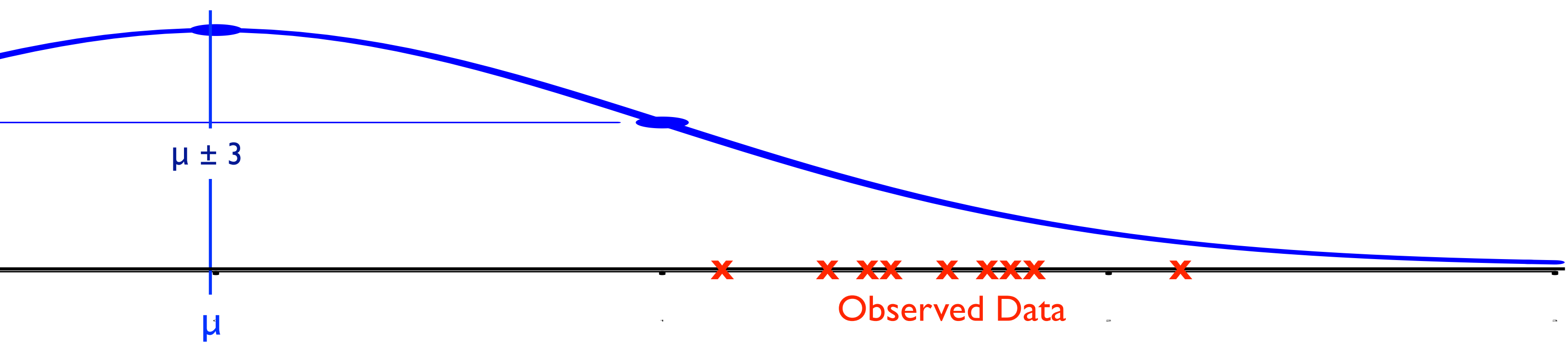
Which is more likely: (a) this?

μ, σ^2 both unknown



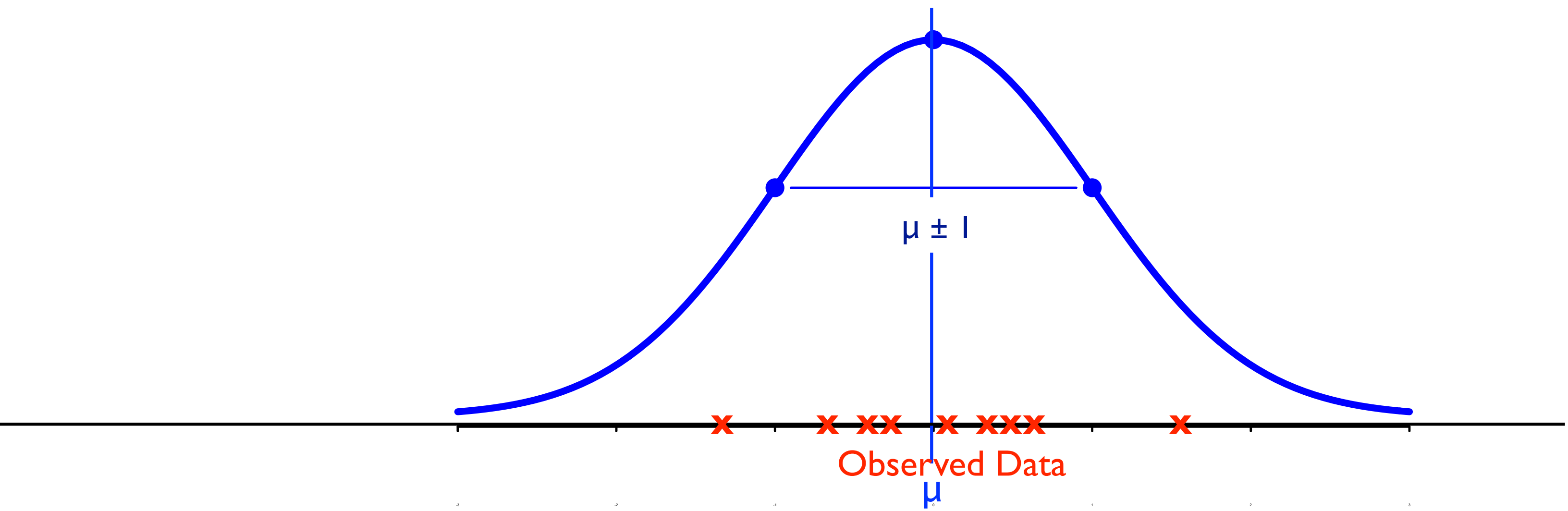
Which is more likely: (b) or this?

μ, σ^2 both unknown



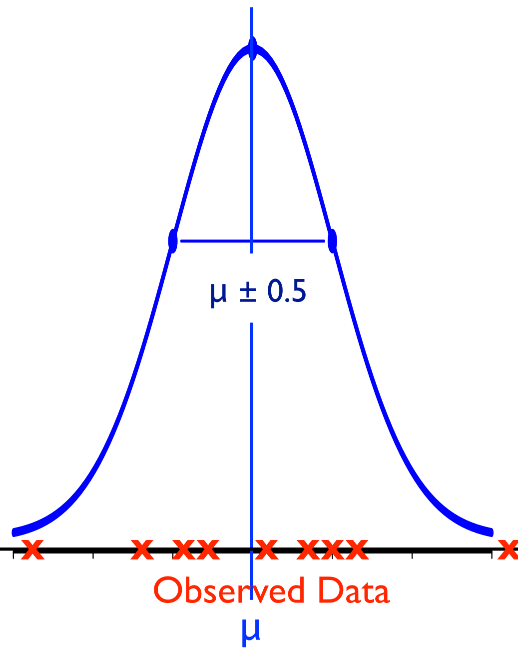
Which is more likely: (c) or this?

μ, σ^2 both unknown



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

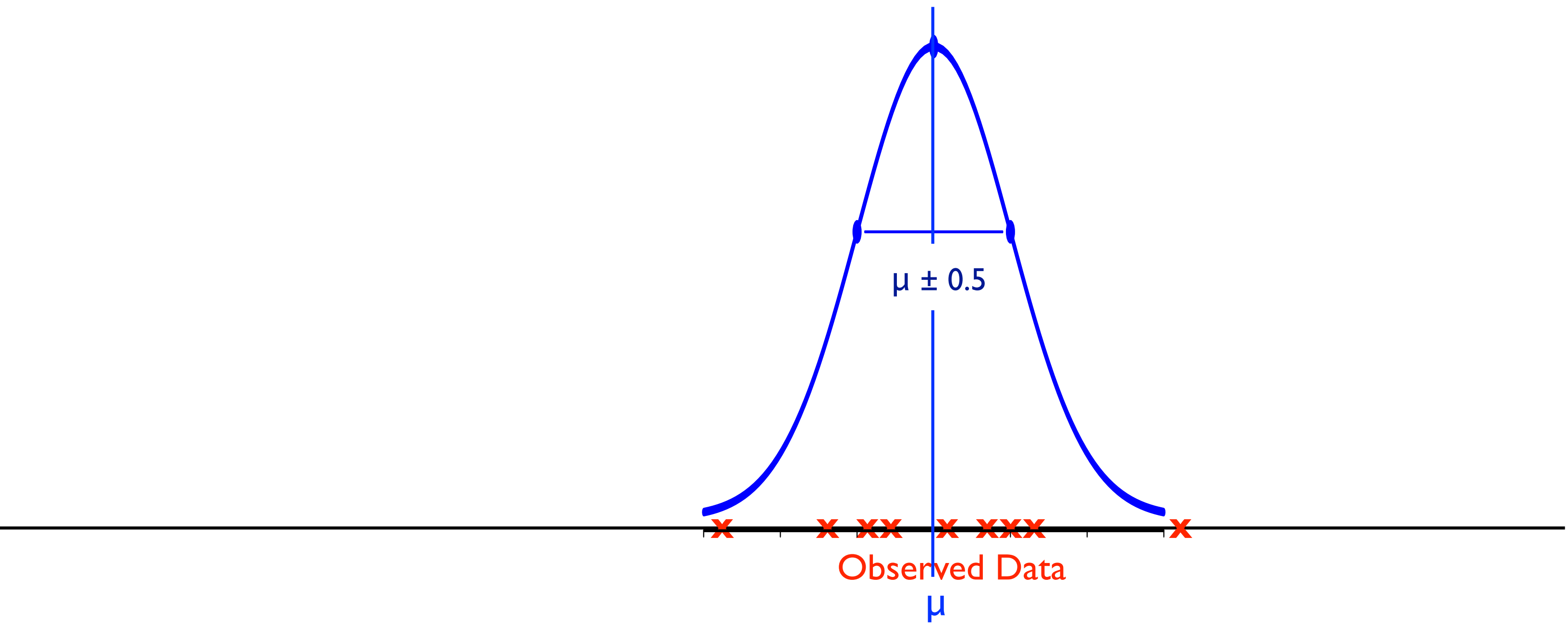
μ, σ^2 both unknown



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 ?

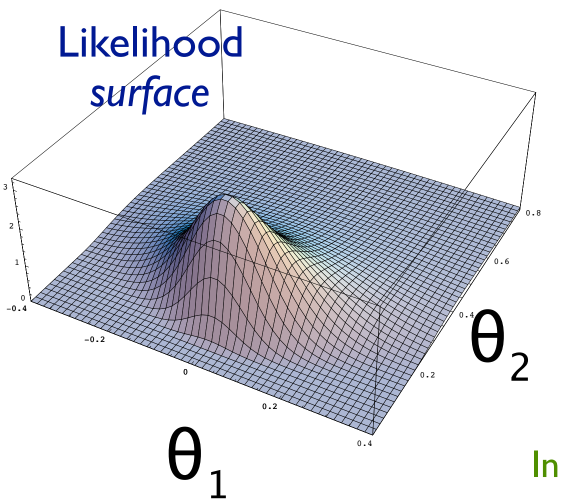


Ex 3: $x_i \sim N(\mu, \sigma^2)$, μ, σ^2 both unknown

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

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

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



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.)

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

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

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

Sample variance is MLE of *population* variance

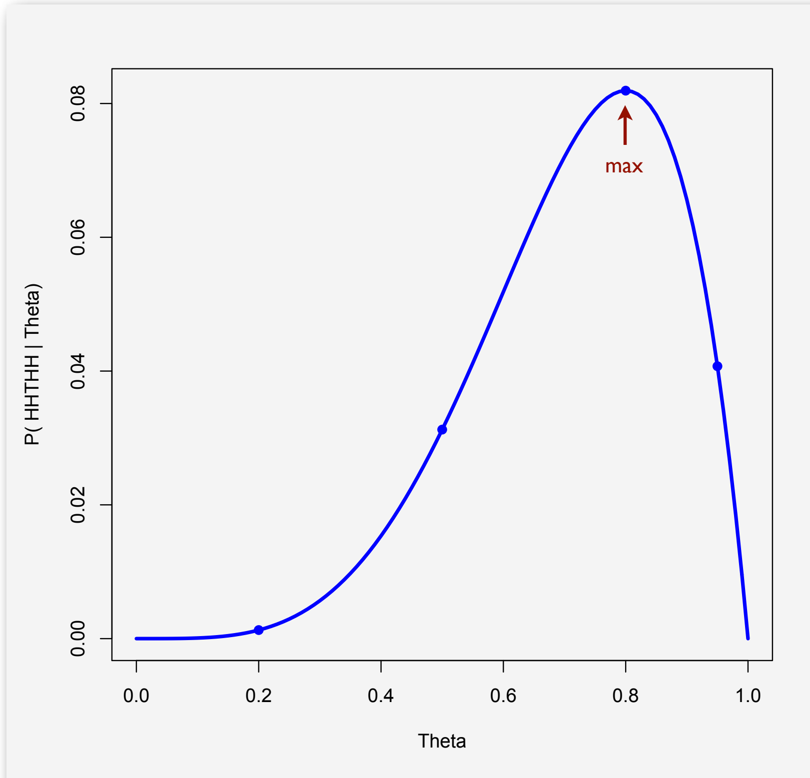
Bias

Recall

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

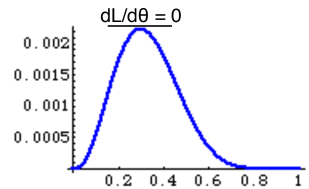


Recall

Example 1

n coin flips, $x_1, x_2, \ldots, x_n$; 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)

(un-) Bias

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

$$E[Y_n] = \theta$$

For coin ex. above, MLE is unbiased:

$$Y_n = \text{fraction of heads} = (\Sigma_{1\leq i\leq n} X_i)/n,$$

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

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

by linearity of expectation

Are all unbiased estimators equally good?

No!

E.g., “Ignore all but 1st flip; if it was H, let Y_n’ = 1; else Y_n’ = 0”

Exercise: show this is unbiased

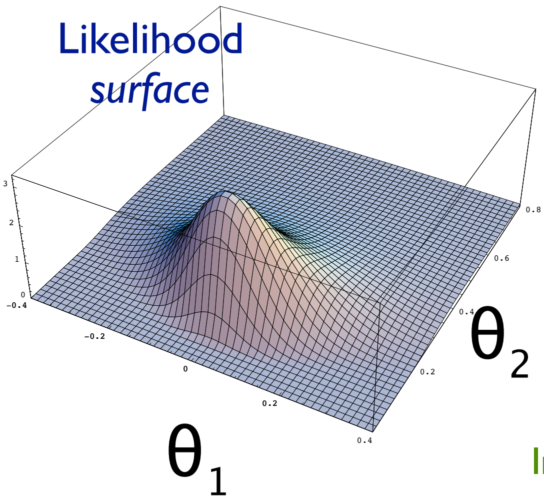
Exercise: if observed data has at least one H and at least one T, what is the likelihood of the data given the model with $\theta = Y_n$’ ?

Recall

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

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

Recall

Ex. 3, (cont.)

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

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

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

Sample variance is MLE of *population* variance

Ex. 3, (cont.)

Bias? if $Y_n = (\Sigma_{1 \le i \le n} X_i)/n$ is the sample mean then

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

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

known μ

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 (B&T p467):

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

Roughly, $\lim_{n \to \infty}$ = correct

One 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 that two sample points would be chosen exactly equidistant from, and on opposite sides of the mean (p=0, in fact), so the MLE $\hat{\theta}_2$ systematically *underestimates* θ_2, i.e., is biased.

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

Confidence Intervals

A Problem With Point Estimates

Reconsider: estimate the mean of a normal distribution.

Sample $X_1, X_2, \ldots, X_n$

Sample mean $Y_n = (\Sigma_{1 \le i \le n} X_i)/n$ is an unbiased estimator of the population mean.

But with probability 1, it's wrong!

Can we say anything about *how* wrong?

E.g., could I find a value Δ s.t. I'm 95% confident that the true mean is within $\pm\Delta$ of my estimate?

Confidence Intervals for a Normal Mean

Assume X_i*'s* are i.i.d. $\sim N(\mu, \sigma^2)$

Mean estimator $Y_n = (\Sigma_{1 \le i \le n} X_i)/n$ is a *random variable;* it has a distribution, a mean *and a variance.* Specifically,

$$Var(Y_n) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

So, $\boxed{Y_n \sim N(\mu, \sigma^2/n),}$ $\therefore$ $\boxed{\frac{Y_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)}$

Confidence Intervals for a Normal Mean

X_i's are i.i.d. ~ N(μ, σ²)

Y_n ~ N(μ, σ² /n) $\qquad \frac{Y_n - \mu}{\sigma/\sqrt{n}}$ ~ N(0,1)

$$P\left(-z < \frac{Y_n - \mu}{\sigma/\sqrt{n}} < z\right) = 1 - 2\Phi(-z)$$

$$P\left(-z < \frac{\mu - Y_n}{\sigma/\sqrt{n}} < z\right) = 1 - 2\Phi(-z)$$

$$P\left(-z\sigma/\sqrt{n} < \mu - Y_n < z\sigma/\sqrt{n}\right) = 1 - 2\Phi(-z)$$

$$P\left(Y_n - z\sigma/\sqrt{n} < \mu < Y_n + z\sigma/\sqrt{n}\right) = 1 - 2\Phi(-z)$$

E.g., true μ within $\pm 1.96\sigma/\sqrt{n}$ of estimate ~ 95% of time

N.B: μ is fixed, not random; Y_n is random