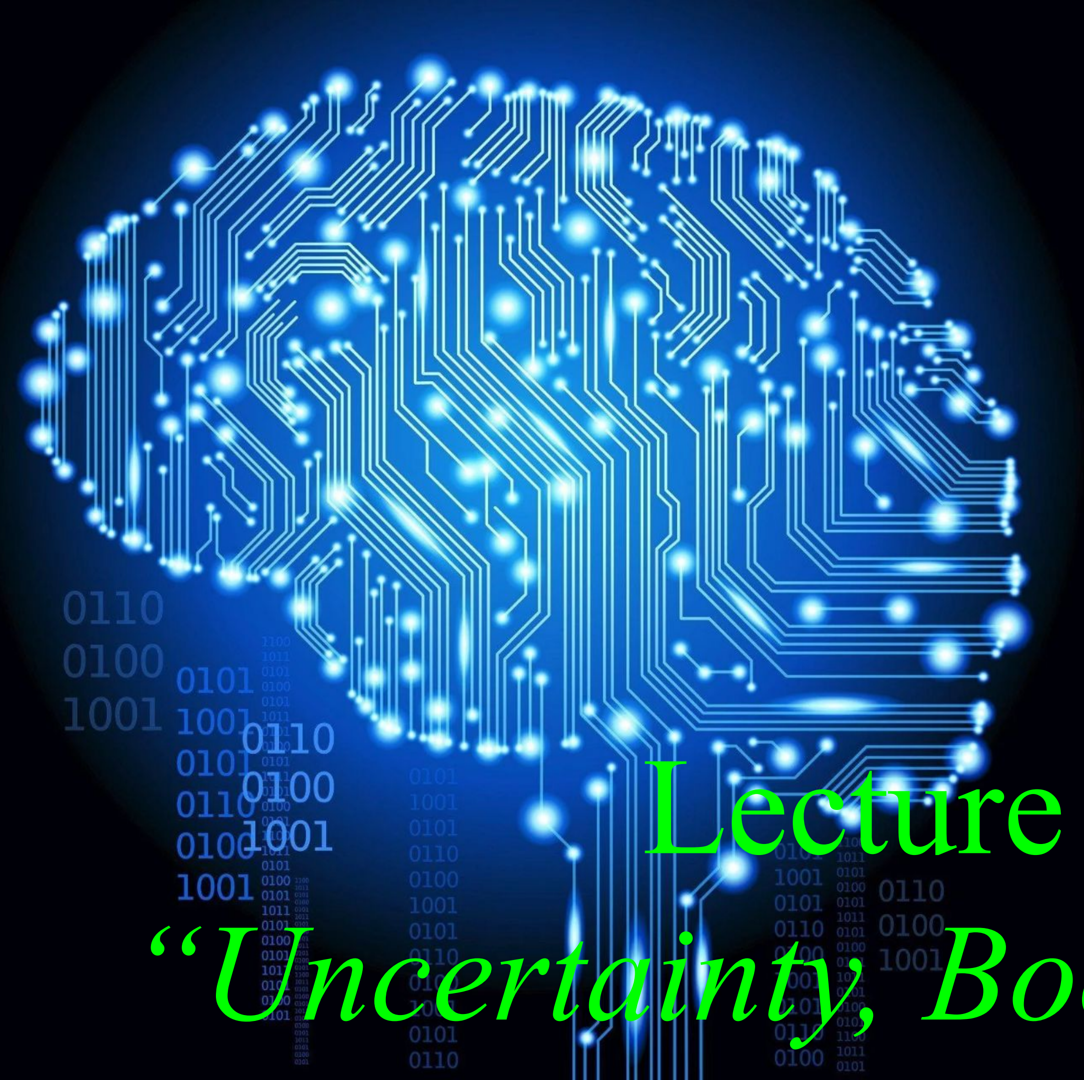
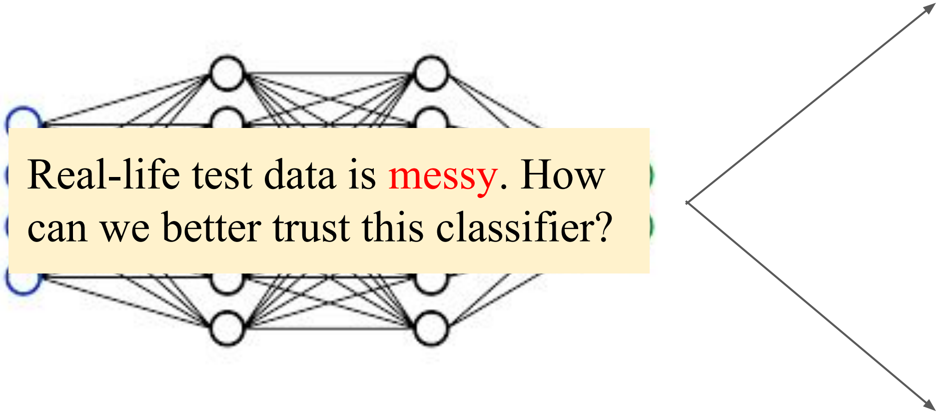


Lecture 04:
“Uncertainty, Bootstrapping”

An Uncertain World

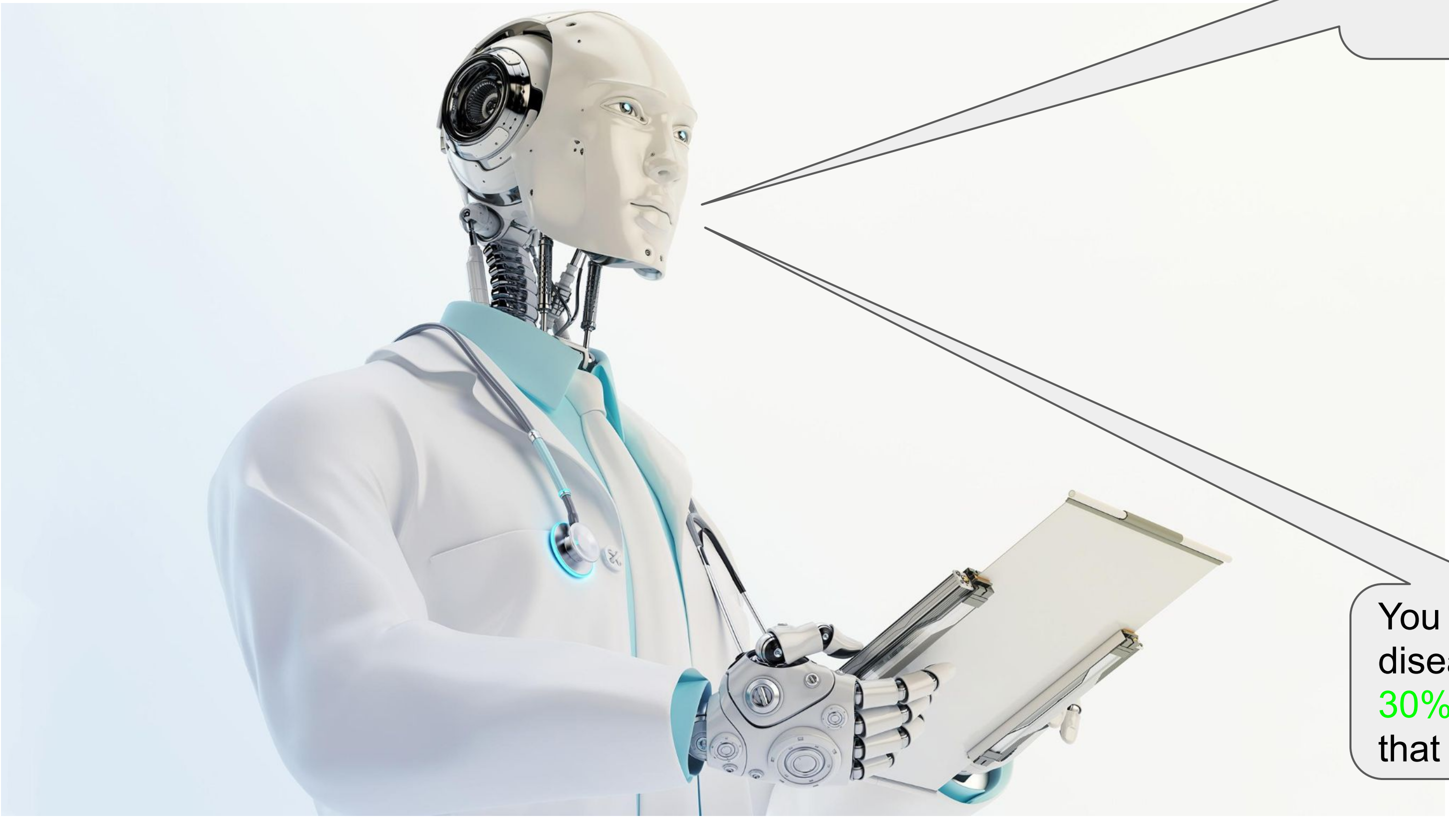


Real-life test data is messy. How can we better trust this classifier?

dog

non-dog

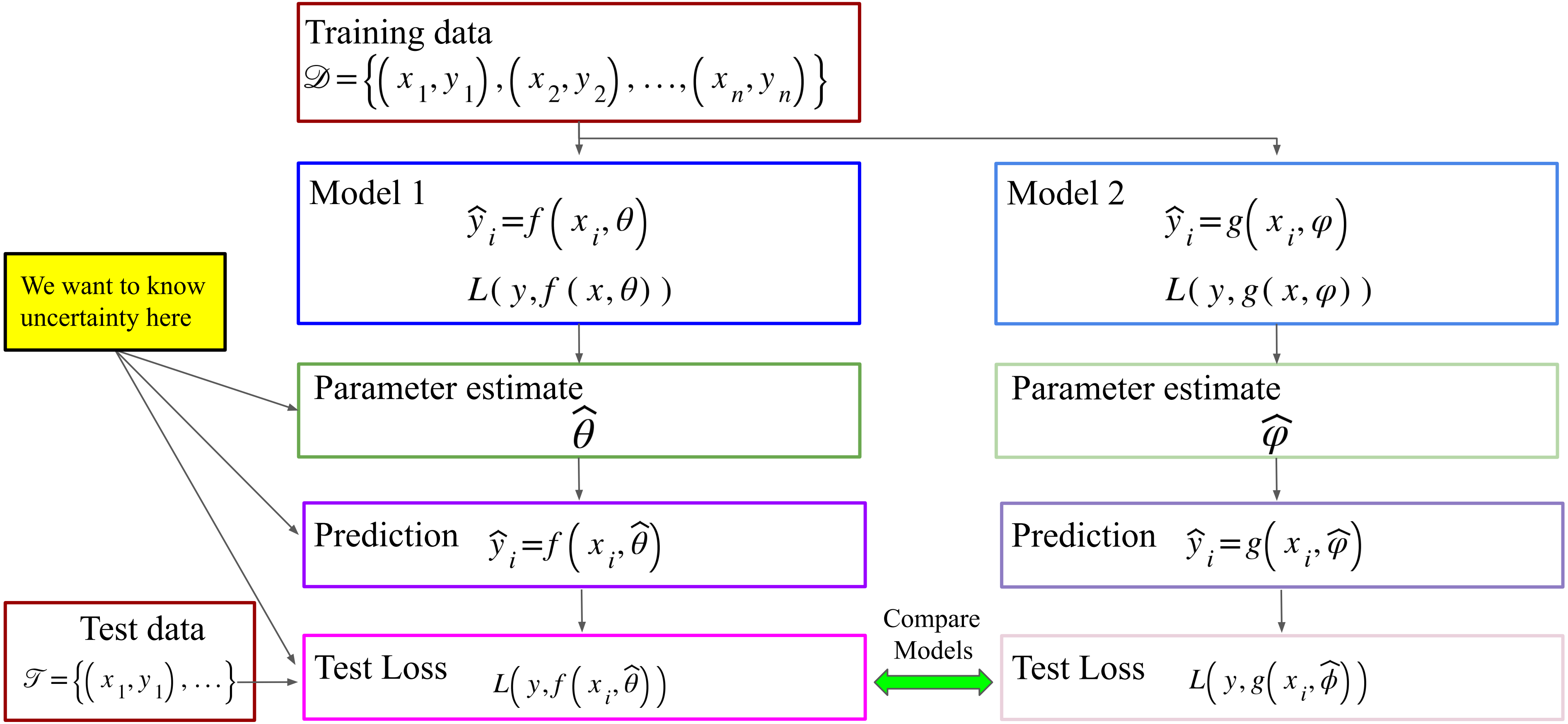
An Uncertain World



You have the disease.

You have the disease. I have 30% confidence that this is true.

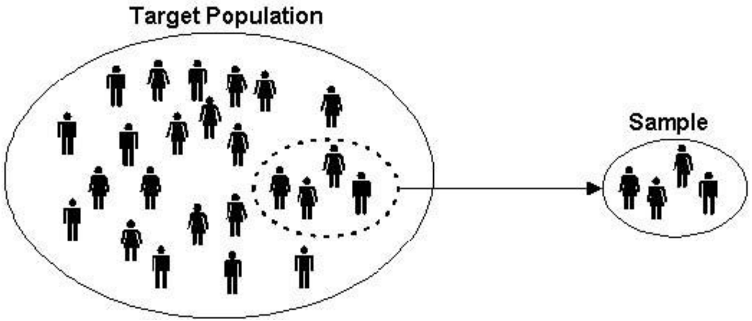
Uncertainty at 3 Points



Parameter Uncertainty

- Parameter = value which summarizes data for a population; these can be expectations (*mean*) or values which describe an input-output relationship (*slope of a linear model*)
- Statistic = value which summarizes data from a particular sample (*i.e. sample mean*).
- Estimation = use a *statistic* to estimate a *parameter* of the distribution of a random variable, where
  - Estimator ($\hat{\theta}$): function used to compute estimate
  - Estimand (θ): parameter of interest

μ_X = parameter

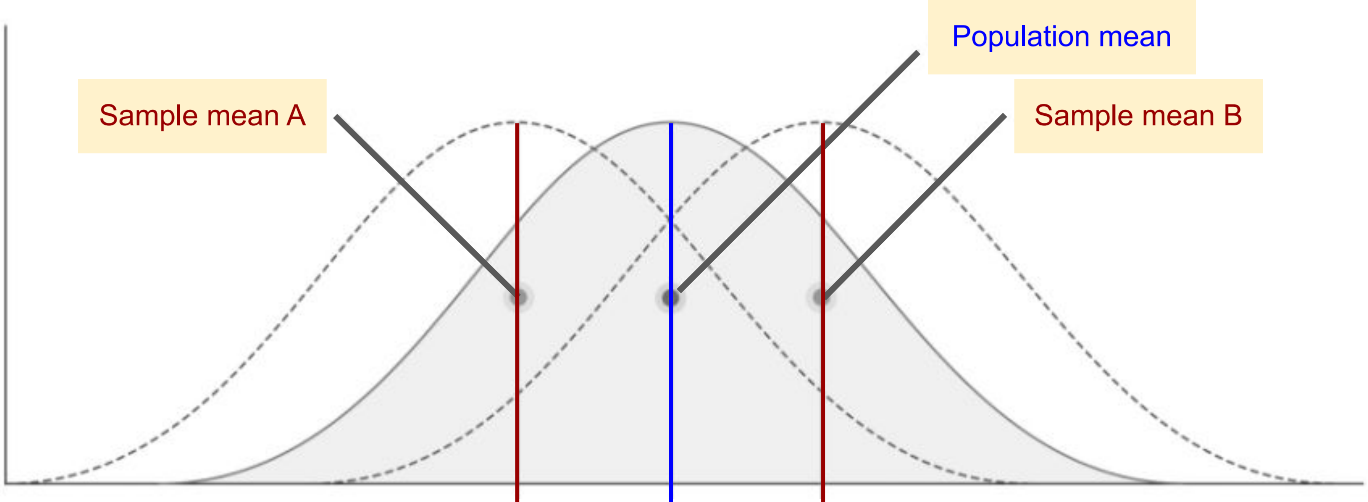


$\bar{x}_n$ = statistic

Example of a parameter: mean

- Consider a model which predicts the mean… i.e. $\hat{y} = \theta$
- Given a dataset $\{x_1, x_2, \ldots, x_n\}$, the estimate for this parameter is the sample mean:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$



The distribution of an estimator is called its **sampling distribution**.

Bias and Variance

- Bias = expected difference between estimator ($\hat{\theta}$) and parameter (θ)

In general: $$\text{Bias}(\hat{\theta}) = E\left[\hat{\theta} - \theta\right]$$

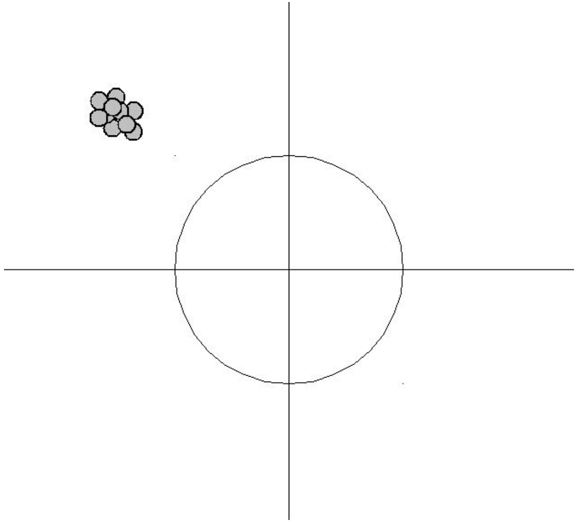
For example: $$E\left[\overline{X}_n - \mu_X\right]$$

- Variance = expected squared difference between estimator ($\hat{\theta}$) and E[estimator] (mean)

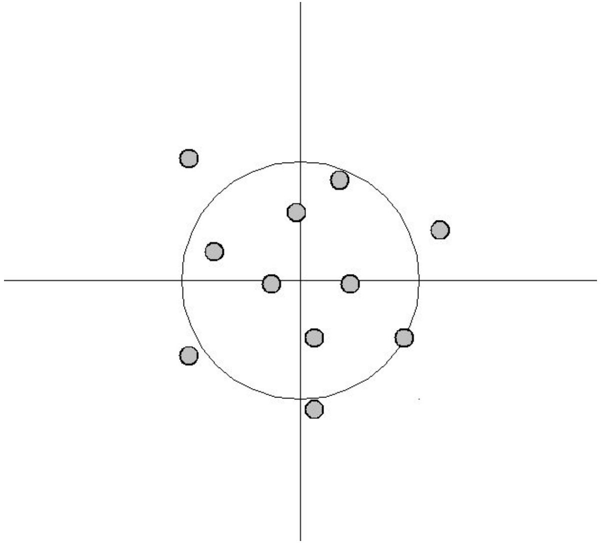
In general: $$E\left[(\hat{\theta} - E[\hat{\theta}])^2\right]$$

For example: $$E\left[\left(\overline{X}_n - E[\overline{X}_n]\right)^2\right]$$

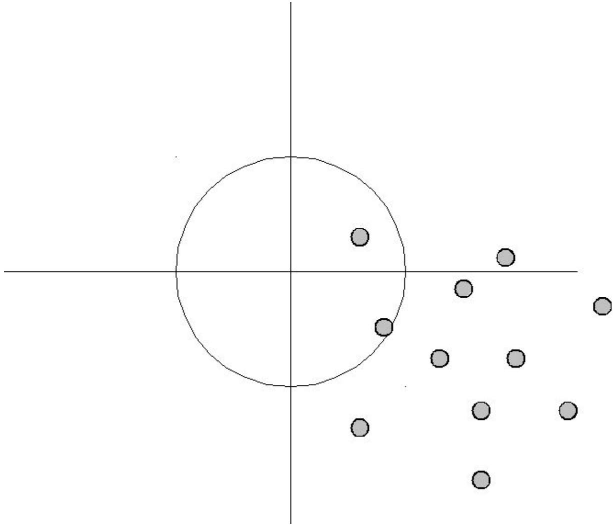
Bias and Variance



High bias, low variance



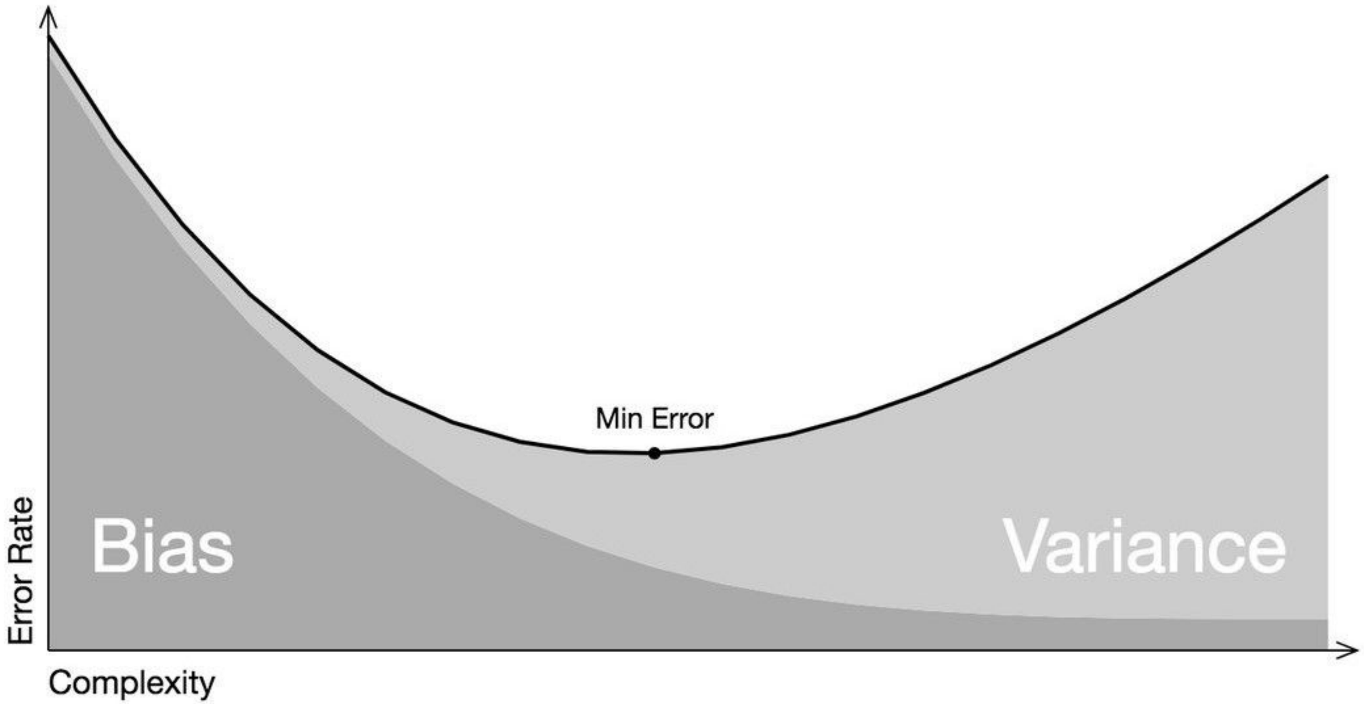
Low bias, high variance



High bias, high variance

Bias-Variance Tradeoff

Bias-Variance Tradeoff

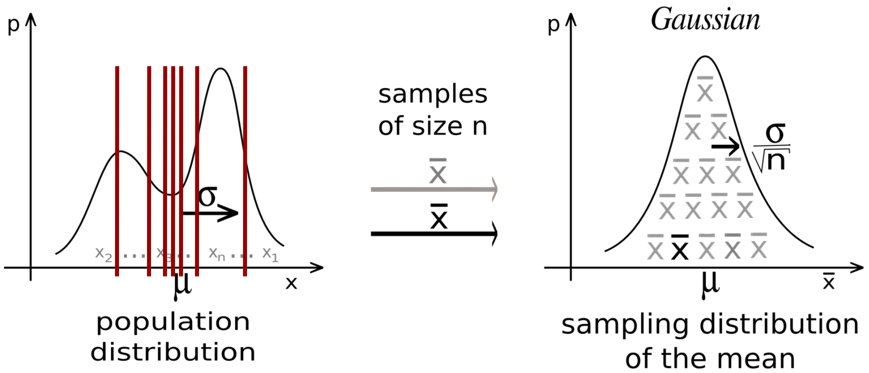


Central Limit Theorem (CLT)

- For large n, the sampling distribution of $\overline{X}_n$ is approximately normal.
- Formally, we can write:

$$\overline{x}_n \sim N\left(\mu,\ \sigma^2_{\overline{X}_n}\right),\ \text{where } \sigma_{\overline{X}_n} = \frac{\sigma_X}{\sqrt{n}}$$

Variance Standard error



Whatever the form of the population distribution, the sampling distribution tends to a Gaussian, and its dispersion is given by the central limit theorem [1]

Central Limit Theorem (CLT)

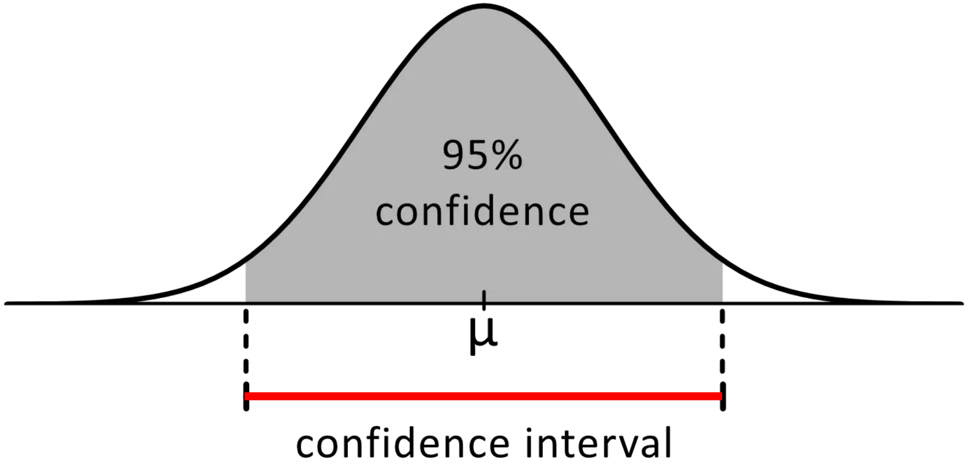
- We can use the CLT to construct **Confidence Intervals**

<u>Question</u>: What's a 95% confidence interval?

<u>Answer</u>: An interval which includes 95% of the sample means.

<u>Another Answer</u>: If we constructed this interval 100 times, it would contain the true mean in 95 of those instances.

Distribution of sample means ($\bar{x}$) around population mean (μ)



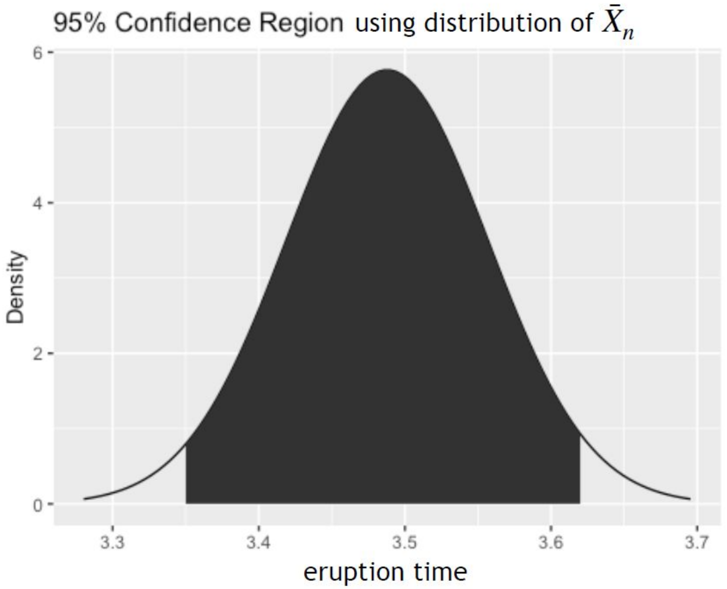
Central Limit Theorem (CLT)

- We can also say that 95% of the sample means are between μ - 1.96$\boldsymbol{\sigma}$ and μ + 1.96$\boldsymbol{\sigma}$
- Alternatively, 95% of the time the true mean μ will be between $\overline{x}_n$ - 1.96$\boldsymbol{\sigma}$ and $\overline{x}_n$ + 1.96$\boldsymbol{\sigma}$

Example: Old Faithful

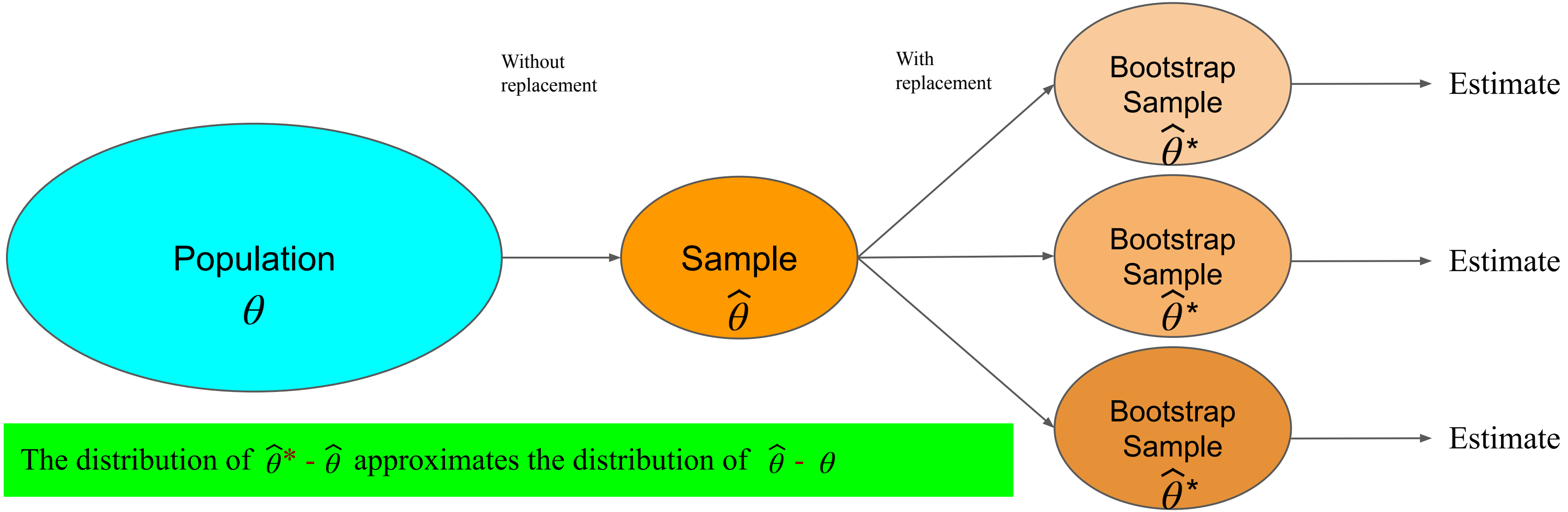
- Say we estimate that the mean value of eruption times is 3.4877831 (with n=272 observations)
- Is this a good estimate? How good is it?

- Mean = 3.49, Stdev = 1.14, SE = 0.07
- CI is therefore 3.49 +/- 1.96*(0.07) = 3.49 +/- 0.14



The Bootstrap

- CLT excellent for datasets with approximately gaussian noise, and does a good job getting a distribution of parameter estimates. What if standard errors not normal?
- Bootstrap = a powerful technique to construct confidence intervals using artificially drawn samples in addition to an originally-drawn sample



The distribution of $\hat{\theta}^* - \hat{\theta}$ approximates the distribution of $\hat{\theta} - \theta$

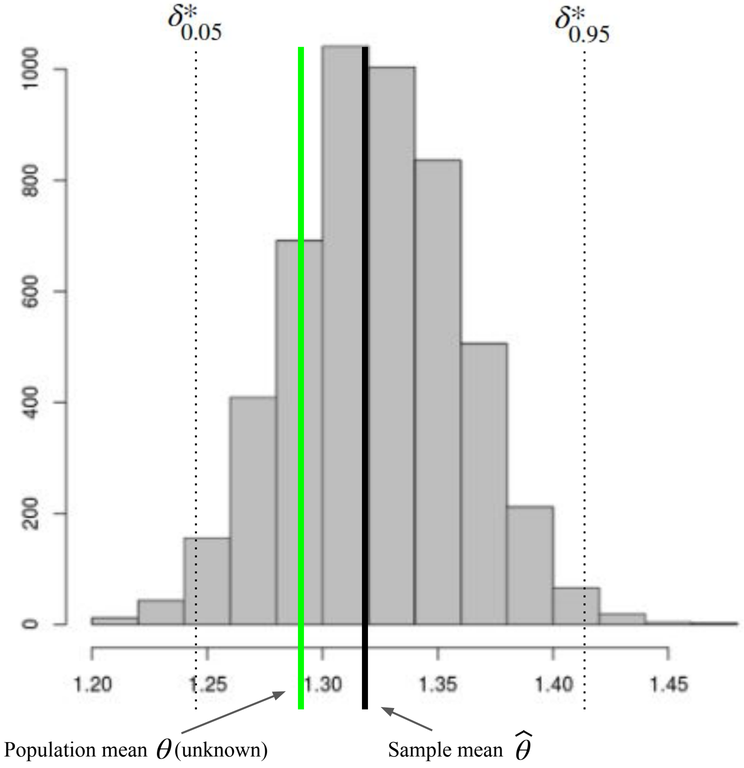
The Bootstrap

```
Your Sample S has N observations
For b in 1:numBootstrap:
     resample N from S with replacement -> S*
     Fit model to S* -> θ̂*(bootstrap statistics)
     Record your bootstrap statistics

Return the distribution of bootstrap statistics
```

The Bootstrap

- Now that we have a distribution of bootstrap statistics, we can construct a CI



- For example, a 90% confidence interval centred at the sample mean would be

$$CI = \left[\hat{\theta} - \delta^*_{0.95}, \hat{\theta}^* - \delta^*_{0.05}\right]$$

where $\delta^* = \hat{\theta}^* - \hat{\theta}$

and where $\delta^*_{0.95}$ is the 95% percentile of the bootstrap distribution

Prediction Uncertainty

Training data

$$\mathscr{D}=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$$

Model

$$\hat{y}_i=f(x_i,\theta)$$

$$L(y,f(x,\theta))$$

Parameter estimate

$$\hat{\theta}$$

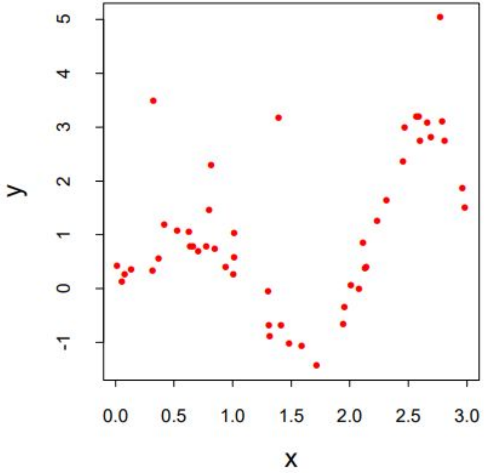
Prediction

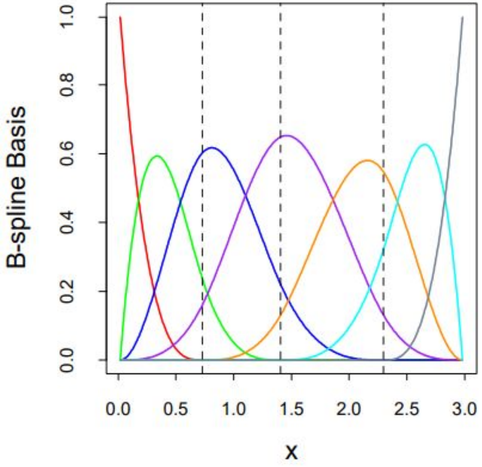
$$\hat{y}_i=f(x_i,\hat{\theta})$$

How does uncertainty in our parameter estimate influence the uncertainty of our prediction?

How much would the prediction change if we had used a different set of training data?

Prediction Uncertainty (Bootstrap)





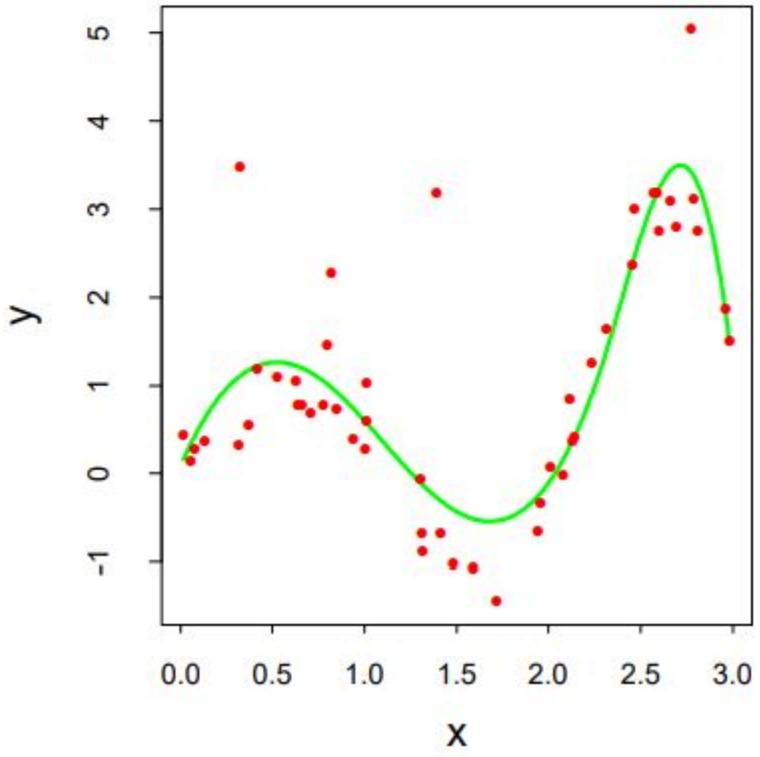
- Example: say we want to fit a cubic spline to this data. We can use a linear expansion of B-spline basis functions h_i(x).

- We store the B coefficients of these basis functions into a vector θ, and fit $\hat{y} = f(x) = X\theta$.

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_p(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_p(x_2) \\ & & & \\ h_1(x_n) & h_2(x_n) & \cdots & h_p(x_n) \end{pmatrix}_{n\times p} \begin{pmatrix} \beta_1 \\ \cdots \\ \beta_p \end{pmatrix}_{p\times 1}$$

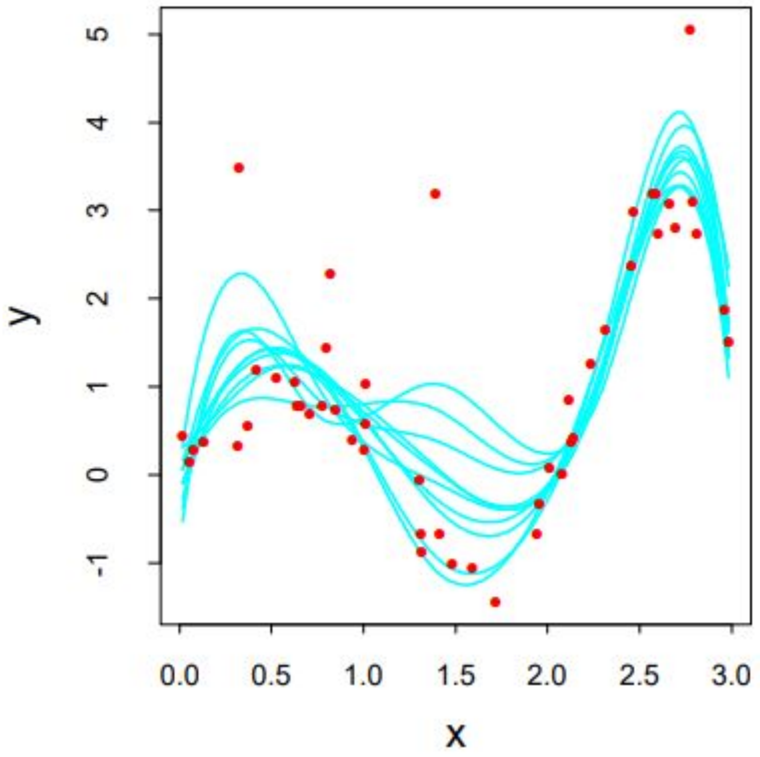
† See ESLII, page 262

Prediction Uncertainty (Bootstrap)



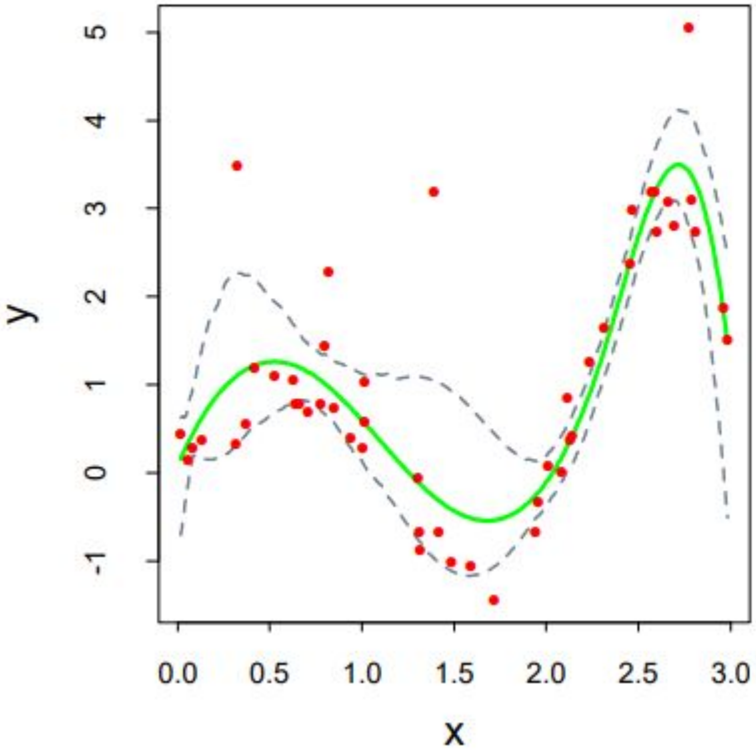
- Here is our fit $\hat{y} = \hat{f}(x) = X\hat{\theta}$
- Is it any good? Yes, but how confident can we be of this?
- **Let's use bootstrap**:
  - From our original sample, generate a new sample (with replacement)
  - For this new sample, get a new parameter estimate $\hat{\theta}_b^*$
  - Do this as many times as you can

Prediction Uncertainty (Bootstrap)



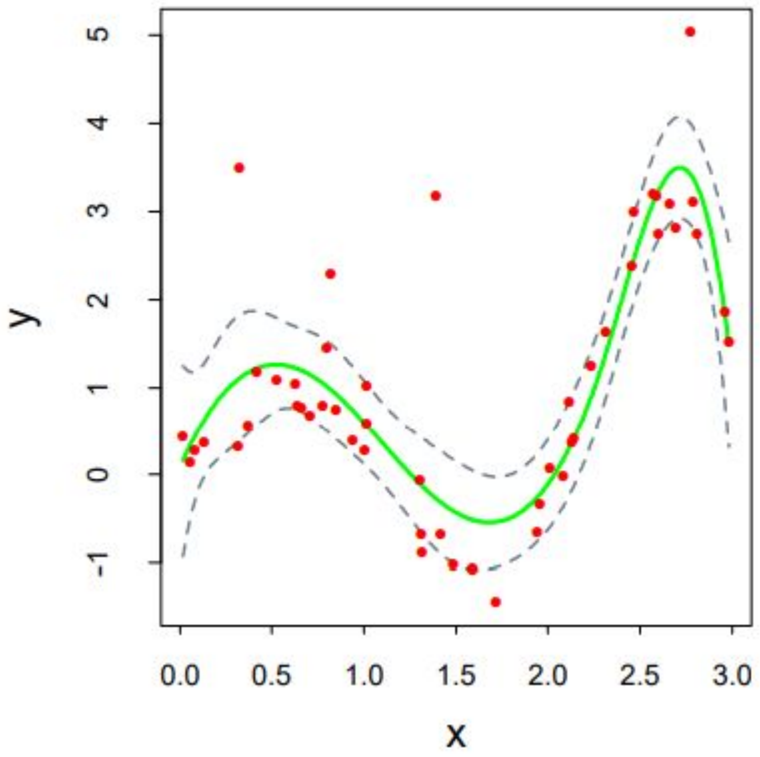
- We can plot each new prediction $\hat{y}^*_b = \hat{f}^*_b(x) = X\hat{\theta}^*_b$

Prediction Uncertainty (Bootstrap)



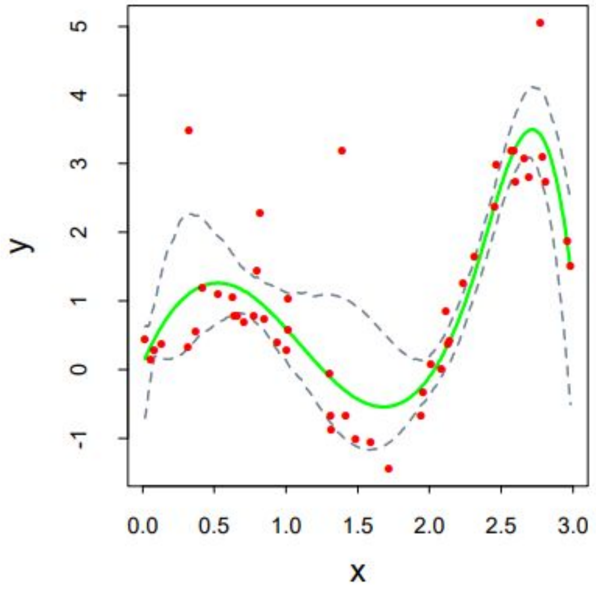
- And since we now have a distribution of samples for each x, we can compute a 95% Confidence Interval (CI)

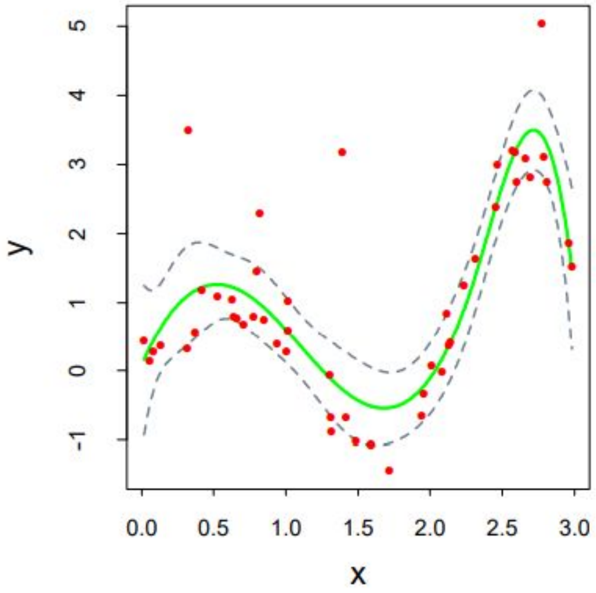
Prediction Uncertainty (Bootstrap)



- Note that we could have also used CLT to get the CIs

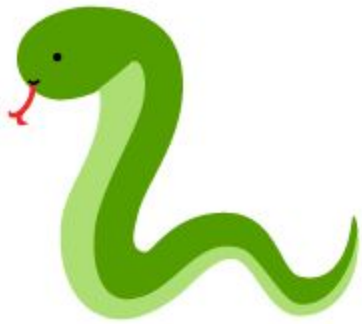
Prediction Uncertainty (Bootstrap)





- Warning: our Confidence Interval (via bootstrap [left] or CLT [right]) is for the true value of $f(x)$, not for new observations ($x_{new,}$ y_{new})
- Why? A CI for new data would need to also consider random variability (σ^2) between $f(x_n)$ and y_n.

Let’sss try it in Python...



Summary

- Parameter Uncertainty
  - Parameters, Statistics, Estimation
  - Example using Population/Sample Mean
  - Bias and variance
  - The Central Limit Theorem (CLT)
  - Constructing a Confidence Interval (CI)
  - Bootstrap
- Prediction Uncertainty
  - B-spline example (Bootstrap)
- Coding examples of CLT, Bootstrap